Detection of bow echoes in kilometer-scale forecasts using a convolutional neural network

Arnaud Mounier, Laure Raynaud, Lucie Rottner, Matthieu Plu, Philippe Arbogast, Michaël Kreitz, Léo Mignan, Benoît Touzé

Detection of bow echoes in kilometer-scale forecasts using a convolutional neural network

Arnaud MOUNIER *, Laure RAYNAUD, Lucie ROTTNER, Matthieu PLU

CNRM, University of Toulouse, Météo-France, CNRS, Toulouse, France

Philippe ARBOGAST

Forecasting department, Météo-France, Toulouse, France

Michaël KREITZ

National meteorological school, Météo-France, Toulouse, France

Léo MIGNAN and Benoît TOUZÉ

Forecasting department, Météo-France, Rennes and Toulouse, France

ABSTRACT

Bow echoes (BEs) are bow-shaped lines of convective cells that are often associated with swaths of damaging straight-line winds and small tornadoes. This paper describes a convolutional neural network (CNN) able to detect BEs directly from French kilometer-scale model outputs in order to facilitate and accelerate the operational forecasting of BEs. The detections are only based on the maximum pseudo-reflectivity field predictor (pseudo because expressed in $mm \cdot h^{-1}$ and not in dBZ). A pre-processing of the training database is carried out in order to reduce imbalance issues between the two classes (inside or outside bow echoes). A CNN sensitivity analysis against a set of hyperparameters is done. The selected CNN configuration has a hit rate of 86% and a false alarm rate of 39%. The strengths and weaknesses of this CNN are then emphasized with an object-oriented evaluation. The BE largest pseudo-reflectivities are correctly detected by the CNN which tends to underestimate the size of BEs. Detected BE objects have wind gusts similar to the hand labeled BE. Most of the time, false alarm objects and missed objects are rather small (e.g., < 1500 km²). Based on a cooperation with forecasters, synthesis plots are proposed that summarize the BE detections in French kilometer-scale models. A subjective evaluation of the CNN performances is also reported. The overall positive feedback from forecasters is in good agreement with the object-oriented evaluation. Forecasters perceive these products as relevant and potentially useful to handle the large amount of available data from numerical weather prediction models.

1. Introduction

Mesoscale convective systems (MCSs, Houze Jr. 2004) are organized thunderstorms with a linear or round shape lasting several hours, whereas single-cell thunderstorms last 20-30 minutes on average. MCSs can cause considerable wind, hail or flood damage. In an operational context, particular care must be taken to closely monitor the different types of MCSs such as squall lines (Trapp et al. 2005) or mesoscale convective complexes (Laing and Fritsch 1997). Among MCSs, bow echoes (BEs) can occur all year round in North America as well as in Europe (Goulet 2015). Even though they are relatively rare in Europe (around 5 BEs per year in France between 2006 and 2020), BEs can have significant meteorological impact because they can generate strong wind gusts and tornadoes. The size and genesis of BEs can differ from one case to another but a general common feature is a bow shape in reflectivity fields. This aspect may be explained by a rear inflow jet in the mid levels of the atmosphere (Przybylinski 1995; French and Parker 2014). This mid-level jet causes a strong downdraft and it is also responsible for strong winds under bow echoes (Fujita 1978; Atkins and Laurent 2009; Markowski and Richardson 2010).

Since the beginning of the 21st century, Numerical Weather Prediction (NWP) models at convection-permitting scale have been developed (Done et al. 2004; Seity et al. 2011). These models were some of the first ones to explicitly simulate MCSs. Due to the limited predictability at such scales (Hohenegger and Schär 2007), convection-permitting ensemble prediction systems (EPSs) are necessary. Systems such as

*Corresponding author: Arnaud MOUNIER, arnaud.mounier@meteo.fr

SREF[1] (Du et al. 2003), MOGREPS[2] (Bowler et al. 2008), COSMO-DE-EPS[3] (Peralta et al. 2012) or AROME-EPS[4] (Bouttier et al. 2016) have been developed by several national weather services to supplement deterministic forecasts.

However, leveraging the huge information provided by EPSs to forecast occurrences and trajectories of MCSs, or more generally, of meteorological objects, still remains a challenge. A visual examination of each member is time consuming, and usual products such as point-based probabilities or percentiles are not appropriate to recognize the different MCSs simulated in EPS members. A possible way to overcome these limitations is to automatically detect MCSs in NWP outputs and to develop probabilistic diagnostics from detected objects. For operational forecasting, Updraft Helicity (UH, Kain et al. 2008) has been developed to detect potential severe convective storms such as supercells (Moller et al. 1994). UH has been used to define severe weather indices (Sobash et al. 2016; Gallo et al. 2019; Sobash et al. 2020) and can be combined with reflectivity fields to recognize severe thunderstorms in a "member viewer" approach (Roberts et al. 2019). MCS detection algorithms have been also used on observed data (Patil et al. 2019). Concerning BEs, an automated detection based on computer vision with skeletonization and shape matching approaches has been developed by Kamani et al. (2016). Existing detection algorithms mostly rely on a threshold of predictor fields that dictates BE identification in model outputs. Such approaches require fine tuning and they are rarely designed to recognize specific shapes, which is a key aspect for the detection of objects such as BEs.

In meteorology, machine learning (ML) and deep learning (DL) methods have recently proved their ability to detect patterns and objects in observational and modeling datasets. One of the first DL method applications was to solve classification problems with the aim of predicting a label (e.g., sunny or cloudy) given an image from a large and varied dataset (Elhoseiny et al. 2015). Liu et al. (2016) were among the first to detect features in NWP outputs using Convolutional Neural Networks (CNNs, LeCun and Bengio 1995) in order to track down tropical cyclones, atmospheric rivers and fronts using classification systems. Then, ML and DL methods have been used for segmentation problems in order to detect object contours in forecast outputs. Segmentation algorithms have been applied to extend the previous work on tropical cyclones and atmospheric rivers (Kurth et al. 2018) and to detect fronts (Biard and Kunkel 2019; Matsuoka et al. 2019; Lagerquist et al. 2019).

Regarding the applications to convection, McGovern et al. (2017) and Gagne II et al. (2019) have shown that CNNs can discriminate severe hailstorms according to spatial structures of storms. These CNNs can differentiate a BE from a supercell and a pulse storm to quantify the risk of hail. Using radar data and radio-soundings, Jergensen et al. (2020) applied ML methods to classify convective storms in three categories, viz disorganised, quasi-linear convective system (QLCS) or supercell. ML and DL have also been used for short-term predictions of strong convective wind gusts (Lagerquist et al. 2017) or tornado occurrences (Lagerquist et al. 2020).

In previously mentioned works, CNN input data were manually pre-selected or obtained from heuristic algorithms for training and prediction processes, and ML/DL methods were used to classify these sets of inputs. In this work, we present a segmentation model that can directly detect and identify BEs in deterministic and ensemble model outputs to provide a comprehensive daily production of BE risk. The purpose of the present study is to train a CNN-based detection algorithm for BEs with a database consisting of forecasts from the French kilometer-scale ensemble prediction system, hereafter called AROME-EPS. A transfer learning approach is tested to apply the CNN to the French deterministic AROME model. This CNN is first assessed with object-oriented scores. The advantages of this CNN-based detection for operational purposes are evaluated by ten forecasters. To assess the practical benefit of CNN detections, methods for synthesizing information are also discussed with forecasters, following the suggestions by Demuth et al. (2020).

The CNN architecture, AROME-EPS databases, and scores used to evaluate CNN skills are detailed in section 2. In section 3, a hyperparameter search is presented, highlighting the roles of some important parameters in the CNN setting and in the training dataset design. The optimal CNN configuration derived from section 3 is further detailed and evaluated in section 4 to highlight its strengths and weaknesses. Section 5 discusses the application of the CNN to the deterministic AROME model without re-training. In a dedicated end-user section 6, different visualization products are proposed to summarize the BE risk in AROME-EPS and AROME models. In the last part of section 6 forecasters' feedback on CNN skills and

[1] Short Range Ensemble Forecast

[2] Met Office Global & Regional Ensemble Prediction System

[3] COnsortium for Small scale MOdelling-DEutschland-Ensemble Prediction System

[4] Application of Research to Operations at MEsoscale-Ensemble Prediction System

advantages in an operational context are gathered. Conclusions and perspectives are given in section 7.

2. Methods and Data

a. AROME and AROME-EPS

AROME is the non-hydrostatic high resolution model of Météo-France, operational since 2008 (Seity et al. 2011; Brousseau et al. 2016). The model covers mostly Western Europe (12°W-16°E and 37.5°N-55.4°N, with a size of 2000×2000km approximately, cf Fig.S1 in the online supplement for a figure of the domain). The current operational AROME model has a horizontal grid spacing of 1.3 km and 90 vertical levels, some statistics concerning the simulated convective cells in AROME are available in the section 3 of Brousseau et al. (2016). AROME is initialized five times a day at 00, 03, 06, 12 and 1800UTC with lead times up to 48h. Operational at Météo-France since 2016, AROME-EPS is the convection-permitting ensemble prediction system based on the non-hydrostatic AROME model. Its domain is similar to the one from the deterministic AROME model. AROME-EPS is a 16-member ensemble (since summer 2019) with a horizontal grid spacing of 2.5 km and 90 vertical levels. It is perturbed with four different sources of uncertainties : lateral boundary conditions (Bouttier and Raynaud 2018), surface conditions (Bouttier et al. 2016), initial conditions (Montmerle et al. 2018; Raynaud and Bouttier 2017) and model errors (Bouttier et al. 2012). AROME-EPS is initialized four times a day at 03, 09, 15 and 2100UTC with lead times up to 51h. AROME-EPS has been developed to improve the prediction of high-impact phenomena such as convective systems.

b. Input data

1) Bow Echo (BE) labeling

The training and validation datasets for the CNN segmentation model are built using pseudo-reflectivity forecasts from AROME-EPS members. The reflectivity field available in AROME-EPS members outputs is calculated with a radar simulator (Caumont et al. 2006) in $mm^6 \cdot m^{-3}$. The Marshall-Palmer Z-R relationship (Marshall and Palmer 1948) is applied to convert the reflectivity fields into pseudo-reflectivity fields in $mm \cdot h^{-1}$. The choice of $mm \cdot h^{-1}$ rather dBZ is a historic choice when the rainfall accumulation was updated hourly. Even if the fields in $mm \cdot h^{-1}$ and dBZ are now produced in operations, only pseudo-reflectivity fields are available for the oldest dates of the training and validation databases. The maximum value in the grid column of pseudo-reflectivities is used as an input for the CNNs. To train the CNNs and compute the validation database scores, a corresponding ground-truth hand-labeled dataset is produced. This dataset is constructed from the contours of hand-labeled BEs plotted by one expert and using the VIA software (VGG Image Annotator, Dutta and Zisserman 2019). After a postprocessing step, each hand-labeled field has the same size as the pseudo-reflectivity input, with a value of 1 for every grid point inside a BE and 0 outside. The labeling process utilizes three variables: maximum pseudo-reflectivity, mean sea level pressure (MSLP), and wind speed at 10m. From these three predictors, the expert uses three conditions to hand-label a BE. The first one is to observe a bow shape in the pseudo-reflectivity field. The second one is to notice a gradient in the MSLP field corresponding to the bow shape. According to Markowski and Richardson's (2010) conceptual model, a specific pressure pattern is observed with BEs, consisting of a minimum in pressure leading a BE, a maximum in pressure beneath a BE, and another relative pressure minimum in the trailing stratiform rain region. The last condition is to observe the bow shape in the wind speed field and especially the strong increase of wind speed in front of the BE.

In practice, the labeling process is highly time-consuming. In order to facilitate the analysis of the predictor fields, only two figures are provided to the expert (Fig. 1), showing the maximum pseudo-reflectivity field, and the contours of $\|\nabla(MSLP)\|_1$ equal to 2 hPa for 10km overlaid with the areas where the divergence of horizontal wind is below $1.5 \times 10^{-3} s^{-1}$. The contours of hand-labeled BEs are plotted over the pseudo-reflectivity fields. An example of fields used for the labeling process is shown in Fig. 1 where a BE is labeled. Note that the wind gust field is not considered because it is not instantaneous, only the maximum over the last hour is computed.

2) BE datasets

The labeling process is applied to specific cases where BEs are observed (in radar data) in France. Studying every AROME-EPS run over a long period would be too time-consuming and simulated BEs can be preferentially identified among the 16 members on these specific cases. A list of these observed BEs in 2017, 2018 and 2019 is subjectively created with the help of ten forecasters to be sure that no observed BE is omitted. The previous four AROME-EPS initializations before the beginning of BE events are used and only the lead times around BE events are retained to save time during the labeling process. Some other severe convective cases (supercells,

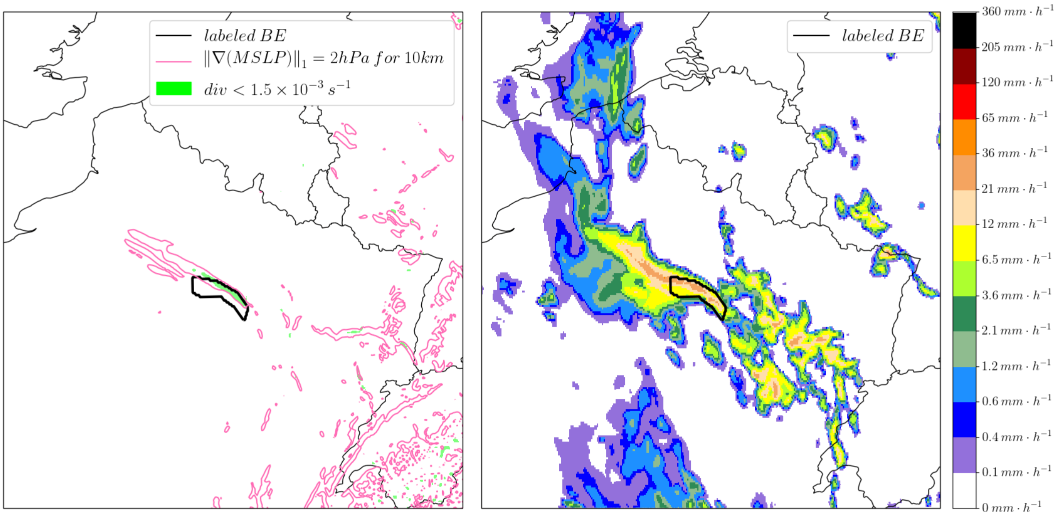

FIG. 1. Example of fields used during the labeling process. The fields are zoomed over the North-East of France. On the left, the contours of $\|\nabla(MSLP)\|_1$ equal to 2 hPa for 10km are in pink, the areas where the divergence of horizontal wind (div in legend) is below $1.5 \times 10^{-3}\ s^{-1}$ in green, and the contour of the labeled BE in black. On the right, the corresponding pseudo-reflectivity field.

squall lines and quasi-stationary convective systems) are also integrated in the validation dataset. The ability of CNNs to learn pseudo-reflectivity signatures not associated with BEs needs to be verified. Some ($\sim$ 10) hand-labeled BEs were plotted in those other severe convective cases. Indeed, some members can simulate BEs in the AROME domain even if a quasi-stationary convective system is observed for instance. Concerning the training dataset, other organized thunderstorms (especially supercells) are indirectly integrated because they are simultaneously simulated over the same areas in similar supporting environments.

Finally, the training dataset is based on a total of 11 observed BE cases (18 May 2017, 27 June 2017, 10 June 2018, 8 August 2018, 6 July 2019, 6-9-11 August 2019, 21 September 2019, and 14-23 October 2019), leading to 6206 pseudo-reflectivity fields and 556 hand-labeled BEs. During the design of the training dataset, a half of the AROME-EPS samples corresponding to one specific BE case (21 september 2021) is randomly selected. This subsampling is necessary because all the simulated BEs in the AROME-EPS members are very similar with same pseudo-reflectivity intensities and same sizes for this case (all BE contours are available in the online supplement to this paper, figure S2). If this subsampling were not made, this BE case could create a large number of very similar training samples and consequently could degrade the training process. Concerning the validation dataset, 4 different observed BE cases (8 August 2017, 26 May 2018, 4 and 5 June 2019) and 3 other convective cases (squall lines : 2 January 2018 and 3 December 2018, quasi-stationary convective system : 5 August 2019) are integrated, leading to 2440 pseudo-reflectivity fields and 264 hand-labeled BEs. This corresponds to approximately 70% of data used for the training and 30% for the validation dataset. The training and validation databases are composed of independent cases to limit as far as possible correlation between the two databases.

3) BE predictors

The selection of CNN inputs is crucial and affects the CNN ability to correctly detect BEs. Due to limited computing power, we can only chose one predictor as a first step toward building a BE identification system. The three variables utilized during the labeling process (maximum pseudo-reflectivity, MSLP and wind speed at 10m) are examined. The MSLP field is too noisy around mountainous areas (especially the Alps and the Pyrenees) with noise similar in magnitude to BE-induced perturbations. An example around the Alps at the bottom right corner of Fig. 1 is shown with pink contours but no precipitation. The wind speed field has sometimes an unrealistic behavior in the AROME models and is rejected at this stage. This behavior is observed in convective and dry conditions or in showers at the rear of cold fronts. The mean wind speeds can be above 120 $km \cdot h^{-1}$ (wind gusts above 210 $km \cdot h^{-1}$) which are unreasonable values in France. The AROME models can also create strong downdrafts in convective cases without associated rainfall. Even if this behavior is occasionally realistic (i.e dry microbursts), forecasters tend to use wind speed with care in convective cases (same conclusion applies to the wind gust parameter). The CNNs in this paper are univariate models with maximum pseudo-reflectivity forecasts as input for that reason. However, wind speed and especially wind gust speed at 10m (FFgust) remain crucial parameters for BEs. The main threat is strong wind gusts under BEs and therefore forecasters need to know magnitudes of these wind gusts to convey information on hazards. In this paper, the FFgusts inside BEs will be considered during the object-oriented evaluation in section 4 and 5 to evaluate the selected CNN in the section 3. Using also wind speed as another predictor could be tested in future work.

c. *CNN architecture*

For BE detection, CNNs are used as segmentation models. The selected CNN should be suitable for small training datasets because BEs are rare events (in space and time). A U-Net architecture (Ronneberger et al. 2015) has been chosen because past studies have shown that it gives satisfactory results when few data are available. This U-Net architecture (Fig. 2) is divided into two parts which correspond to the contracting and expanding paths. The contracting path consists of 2 convolutional layers (32 3×3 filters) followed by a rectified linear unit (ReLU) activation function and a dropout rate of 0.2. A 2×2 maxpooling operation is applied to divide the patch size by 2. The same sequence is repeated two more times with respectively 64 and 128 filters. The expanding path first performs a 2×2 upsampling ("upconvolution") operation. The upsampled maps are then concatenated with the corresponding maps in the contracting path, and 2 convolutional layers followed by a ReLU activation are applied (with the same number of filters as in the corresponding contracting path layer). Upsampling, concatenation and convolution are repeated another time. At this stage, the outputs have the same size as the inputs. A final convolutional layer with a 1x1 filter is added to obtain the desired number of classes (two classes here). In order to get a class probability as an output, a softmax function is applied. The U-Net architecture is presented here with an input size of $N \times M$ grid points. Several sizes of inputs are tested in the next section. Other classical parameters for the U-Net architecture such as the number and size of filters, the dropout rate and the choice of the ReLU activation are not discussed in the following sections.

d. *CNN training*

Limited computing power does not allow to take the entire AROME domain as a CNN input. To train the U-Nets, patches of $N \times M$ grid points are extracted from the original pseudo-reflectivity fields (717×1121 grid points). In addition, because the size of a BE is much smaller than the domain size, there is a strong imbalance between the BE and no BE classes in the training dataset. Extracting smaller patches allows us to design a more balanced training database by selecting the most informative patches. Fig. 3 shows the different steps undertaken to extract patches (from 1 to 6):

- (step 1) Patches from the pseudo-reflectivity fields are randomly selected. $N_{patches}$ are extracted from each field (values in Table 1). The associated groundtruth patches are also extracted.

- (step 2) The pairs of patches (pseudo-reflectivity/groundtruth) are split into two groups : the patches with BEs in which at least one grid point is labeled as a BE and the patches without any BE.

- (step 3) The number of patches with BEs is very limited compared to patches without any BE (1 patch with a BE for every 1600 patches without a BE). Initially, BEs with moderate pseudo-reflectivities were not correctly predicted because they were under-represented in the dataset. A data augmentation technique is proposed to solve this problem: in a given patch, the pseudo-reflectivities are multiplied by a coefficient of 0.75 if a BE is within this patch and the maximum pseudo-reflectivity is above a given threshold (mentioned in the next section). This new patch is added in the ones with BEs (step2). The pseudo-reflectivity field must remain physically consistent as much as possible. That is why lower coefficients are not investigated and only patches with large magnitude in pseudo-reflectivities are taken into account.

- (step 4) To reduce the number of patches without a BE, the patches without precipitation (pseudo-reflectivity maximum $< 0.1\ mm \cdot h^{-1}$) are deleted.

- (step 5) Even after the filtering procedure of step 4, the number of patches without a BE remains high (1 patch with a BE for every 400 patches without a BE). To limit the number of patches without a BE, the ratio between the patches with and without a BE (Ratio_noBE/BE) is

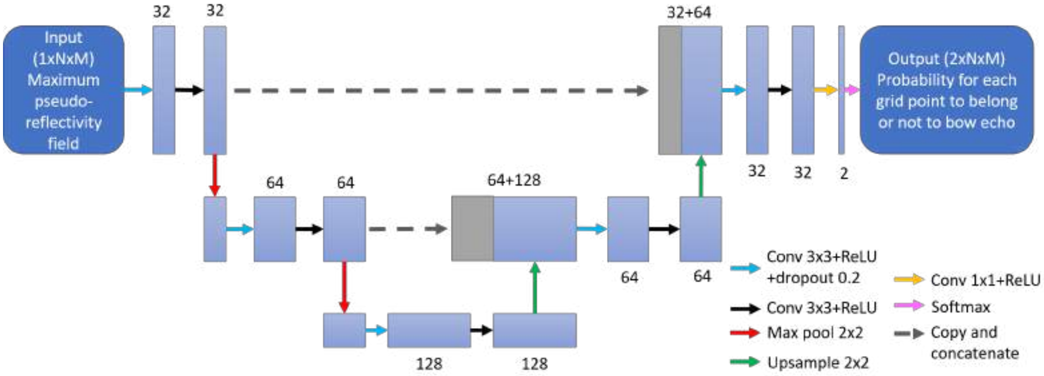
Input (1xNxM) Maximum pseudo-reflectivity field
32
32
64
64
128
128
64+128
64
64
32+64
32
32
2
Output (2xNxM) Probability for each grid point to belong or not to bow echo
Conv 3x3+ReLU +dropout 0.2
Conv 3x3+ReLU
Max pool 2x2
Upsample 2x2
Conv 1x1+ReLU
Softmax
Copy and concatenate

FIG. 2. U-Net architecture for bow echo detection

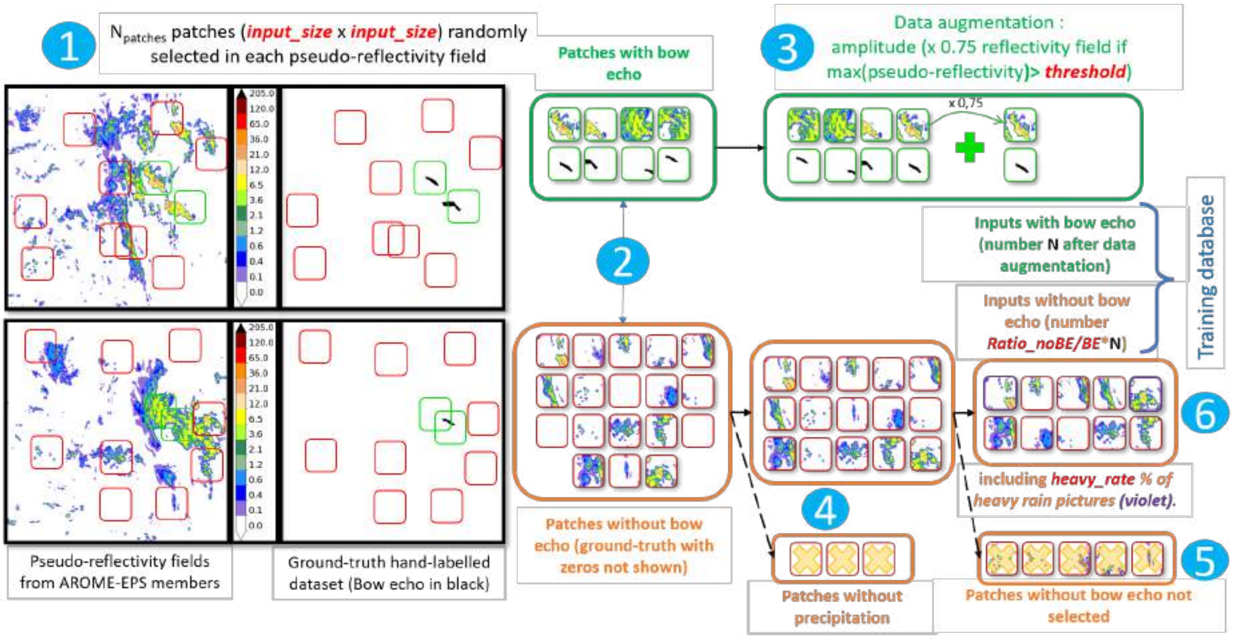
1
N_patches patches (input_size x input_size) randomly selected in each pseudo-reflectivity field
Patches with bow echo
2
3
Data augmentation : amplitude (x 0.75 reflectivity field if max(pseudo-reflectivity)> threshold)
x 0,75
Inputs with bow echo (number N after data augmentation)
Inputs without bow echo (number Ratio_noBE/BE*N)
Training database
6
including heavy_rate % of heavy rain pictures (violet).
4
5
Patches without bow echo (ground-truth with zeros not shown)
Patches without precipitation
Patches without bow echo not selected
Pseudo-reflectivity fields from AROME-EPS members
Ground-truth hand-labelled dataset (Bow echo in black)
205.0
120.0
65.0
36.0
21.0
12.0
6.5
3.6
2.1
1.2
0.6
0.4
0.1
0.0

FIG. 3. Training database design. Steps 1-6 are denoted by blue circles and are described in the section 2.d.

fixed to a lower value. The patches retained are randomly selected and the unnecessary patches without a BE are deleted. This ratio is discussed in the next section.

- (step 6) During the first tests, all patterns with strong pseudo-reflectivities were detected as BEs and consequently the number of false alarms was very high in the validation database. Strong pseudo-reflectivities are rare in space and time and the majority of patches without a BE contains no or weak precipitation whereas BEs are frequently associated with heavy precipitation. Only 0.7% of patches without BE are associated with heavy precipitation (i.e above 60 $mm \cdot h^{-1}$) whereas, after the data augmentation in amplitude (step 3), around 55% of patches with BEs are associated with heavy precipitation. In this case, the pseudo-reflectivity magnitude is relied on too heavily to detect the BEs in the pseudo-reflectivity fields. Patches with large magnitude pseudo-reflectivities but without a BE are forced in the training database to solve this problem and the rate of large magnitude pseudo-reflectivity patches without a BE in the total number of patches without a BE is defined (heavy_rate). A patch is considered with large magnitude pseudo-reflectivities if the maximum is above 60 $mm \cdot h^{-1}$. Another way to address

TABLE 1. Values of N_{stride} and N_{patch} according to input size

Input size	N_{stride}	N_{patch}
$24 \cdot 24$	7	4000
$48 \cdot 48$	15	1000
$96 \cdot 96$	30	250

this problem could be to add more input predictors.

e. CNN prediction

The U-Net inputs and outputs are composed of $N \times M$ grid point patches. The aim of this study is to predict the BE risk over the entire original AROME domain. The same U-Net is applied on overlapping patches as depicted in Fig. 4. The entire grid is divided in ordered patches with a stride of N_{stride} grid points along the longitude and latitude axes (Fig. 4 step 1 : blue, black, and orange squares). The values of N_{stride} depend on the input size of the U-Net and are given in Table 1. A prediction for each patch is computed. The prediction results are patches representing the probability of each class (Fig. 4 step 2 : two patches after the U-Net architecture for each pseudo-reflectivity patch, $p_{n,i}$ is the probability of the ith grid point according to the nth patch). The mean probability for each class is computed (Fig. 4 step 3) considering all the patches where the grid point is included. Finally, the probabilities are used (>50% defines a categorical prediction) to define a detected BE (Fig. 4 step 4). Following the training, Fig. 4 depicts the U-Net application in the prediction process from the input pseudo-reflectivity field of 717 × 1121 grid points to the segmentation mask (same shape of 717 × 1121).

f. CNN implementation

For the implementation of this CNN, the TensorFlow/Keras software is used (Abadi et al. 2015). Using only pseudo-reflectivity fields as inputs, the U-Net is a univariate model. The training and validation data do not need to be normalized. After some tests and loss curve verifications, the number of epochs is fixed at 50. Different batch sizes have been tested (16, 32 or 64): results are similar with 16 and 32, but with 64 an overfitting is noticed. The batch size is fixed at 32. The weights of neural networks are randomly initialized and updated using the Adam optimizer (Kingma and Ba 2014), a default learning rate value of 10^{-3} is applied. The weights are iteratively optimized during the training to minimize a weighted cross-entropy loss function (L, Eq. 1). Contrary to the U-Net weights, the loss function weights are fixed. This weighted loss function has been chosen because BEs are rare events and neural networks tend to overestimate the "no bow echo" class with a non-weighted loss function.

$$L = -\frac{1}{N_{gp}} \sum_{k=1}^{N_{gp}} w_0\,(1-Y_k)\,log(1-p_k) + w_1\,Y_k\,log(p_k). \tag{1}$$

In Eq. 1, N_{gp} is the number of grid points in the training database, p_k the BE probability and Y_k the true label for the k^{th} grid point. w_0 and w_1 represent respectively the weights of class 0 (no BE) and class 1 (BE). The value of w_0 is fixed to 1. The value of w_1 is discussed in the next section with a hyperparameter selection. Finally, Table 1 shows the values of N_{stride} and N_{patch} presented in the above subsections.

g. Object-oriented evaluation

The capacity of this CNN to detect BEs is evaluated with an object-oriented approach. A grid point approach is not suitable for assessing the BE detections because BEs are small objects and rare events. As a consequence, a small shift between groundtruth and predicted labels may be responsible for bad scores, which is known in the literature as double penalty problem (Davis et al. 2006; Rossa et al. 2008; Ebert 2008). An object-oriented approach has already been used to evaluate AROME-EPS precipitation forecasts (Raynaud et al. 2019). Object attributes used in this study are respectively the 0.25-th (Q25) and the 0.9-th (Q90) percentiles of the pseudo-reflectivity field within bow echo objects, the position of the object mass center, the object area and the FFgust maximum within the object. The attributes are presented in Fig. 5 except for the last one. The Q25 and Q90 percentiles allow differentiation of strong from moderate bow echoes. Another use of Q90 and Q25 is to separate the most active part of BEs (center) from the least active part (edge). The object area is used to distinguish large and small BEs. The maximum FFgust reveals the intensity of wind around the BE, which is a complementary signature of a BE. The distribution of object attributes is computed using all labeled and detected objects. An attribute comparison for matching objects is also performed. The labeled and detected objects are matched if the distance between the centers of their mass is lower than a given threshold in the remainder of the paper.

h. Scores

Contingency-based scores are used to quantify the ability of different U-Net configurations to correctly

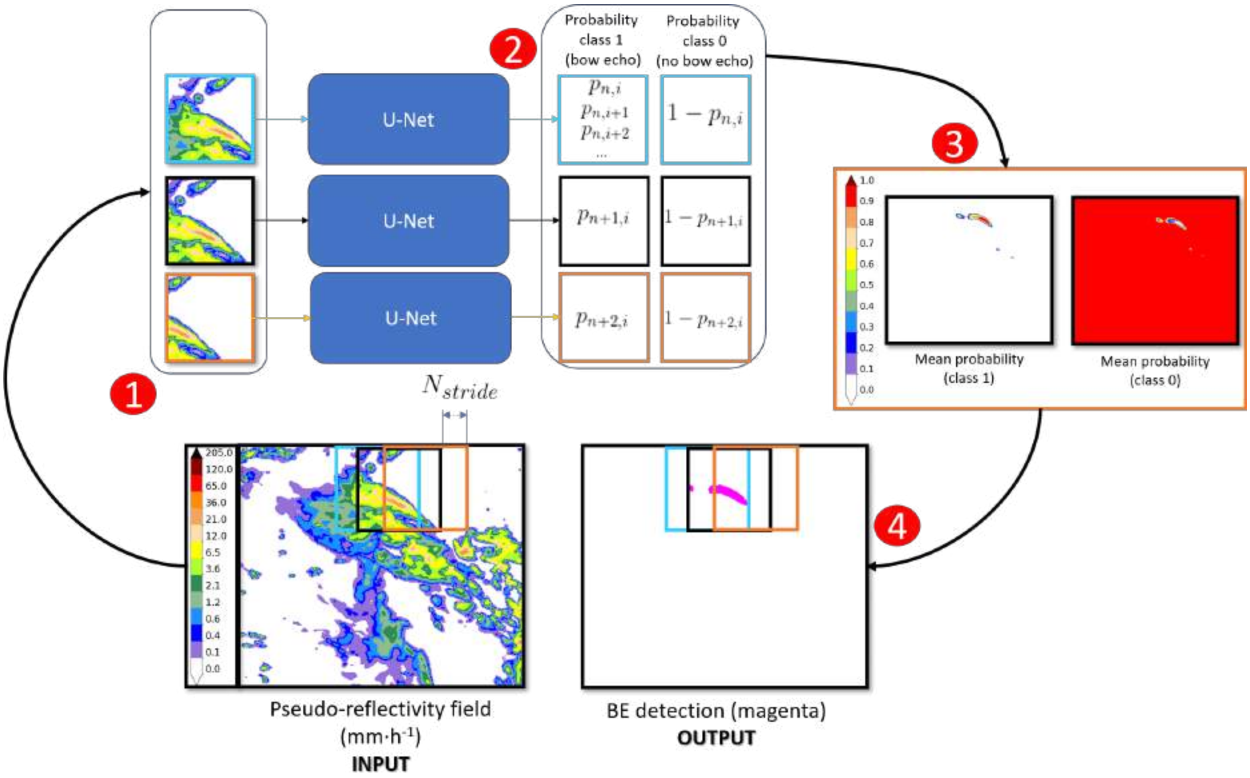

FIG. 4. The whole **prediction** process from pseudo-reflectivity field (INPUT) to BE segmentation mask (OUTPUT) is represented. Steps 1-4 are denoted by red circles and are described in the section 2.e.

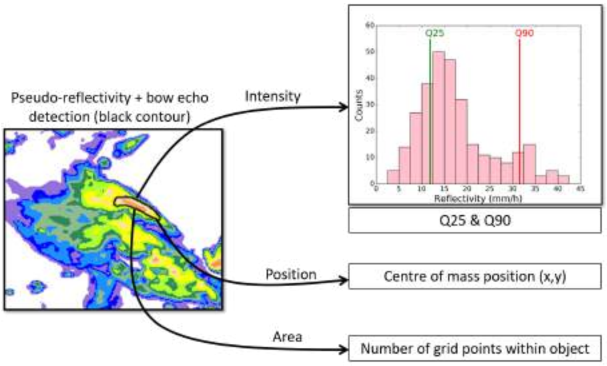
Pseudo-reflectivity + bow echo detection (black contour)
Intensity
Position
Area
Q25
Q90
Counts
Reflectivity (mm/h)
0
5
10
15
20
25
30
35
40
45
10
20
30
40
50
60
Q25 & Q90
Centre of mass position (x,y)
Number of grid points within object

FIG. 5. Object attributes. A bow echo is characterized by 3 attributes: intensity, position and area. Bow echo intensity is described with Q25 and Q90 of pseudo-reflectivities within the object (pink histogram), position with the center of mass and area is the number of grid points within object.

detect the occurrence or non-occurrence of BEs in addition to the comparison of labeled and detected object attributes. A contingency table is computed for the validation database as described in Table 2. The risk of having two or more BEs in the same field is very low because BEs are rare events. The contingency table is based on the detection of at least one labeled or detected BE over the entire grid for each field. From this contingency table, the classical hit rate (HR, Eq. 2) and false alarm rate (FAR, Eq. 3) are used to evaluate the U-Net skill:

TABLE 2. Contingency table. Hits (a) correspond to the number of pictures with at least one labeled bow echo and one detected bow echo. False alarms (b) are the number of pictures with detected bow echo(es) but not labeled. Misses (c) are the number of pictures with labeled bow echo(es) but not detected. Correct negatives (d) are pictures with no labeled and detected bow echo.

Contingency Table		BE label(s)	
		YES	NO
BE detection on picture	YES	a	b
	NO	c	d

$$HR = \frac{a}{a + c} \tag{2}$$

$$FAR = \frac{b}{a + b} \tag{3}$$

We use the Critical Success Index (or Jaccard Index, Eq. 4) to combine the false alarms and missed

TABLE 3. Values tested for the five considered hyperparameters.

Parameters	Values
input size	24×24,48×48,96×96
data augmentation threshold	30,40,60 $mm \cdot h^{-1}$
ratio noBE/BE patch	2,3,4 noBE for one BE
heavy pseudo-reflectivity patch percentage	15%,25%,35%
class 1 weight (w_1)	2.5,3,3.5,4,4.5

detections:

$$CSI = \frac{a}{a+b+c} \quad (4)$$

3. Hyperparameter Search

a. Hyperparameter configuration

The aim of this part is to identify the most influential hyperparameters and to find the optimal combination for the U-Net. We apply the tuning to the following parameters (hyperparameters): input patch size, data augmentation threshold, ratio noBE/BE, heavy pseudo-reflectivity patch percentage and w_1 in the loss function. The selection is based on the scores presented in the previous section and computed on the validation database. The number of values that can be tested is restricted because of limited computing resources. The hyperparameter values tested are presented in Table 3. The 405 combinations are tested. Only the results for input size, ratio noBE/BE and w_1 are presented in the following paragraphs because the sensitivity to data augmentation and heavy pseudo-reflectivity patch percentage turned out to be weak.

b. Filtering threshold

The UNets occasionally predict BEs whose size is only a few grid points, on account of the prediction process described in Fig. 4 and especially the choice of a specific probability threshold to split the U-Net outputs into two classes. These detections were mostly false detections which affected the object-oriented scores in a significant way. In order to remove these very small objetcs, the detected BEs with an area below a certain threshold are converted to null events(i.e the class was switched from 1 to 0). The hand-labeled BEs are not affected and they are all incorporated to calculate the object-oriented scores. The optimal threshold is selected with the aid of these object-oriented scores. This threshold is set from the scores of the 405 configurations (Fig. 6). The configurations with HRs and FARs equal to 0 are not taken into account in this figure because they are associated with unskillful U-Nets (cf section 3.c.1). The hit rate naturally decreases when the filtering threshold increases. The same behavior is observed for the false alarm rate. According to the CSI, the optimum threshold is around 150 grid points. With a 150-grid-point threshold, the HR is around 70%. However, forecasters prefer a higher HR even if the FAR increases simultaneously (see section 6.c, the second item). Hence, a filtering threshold of 100 grid points is preferred, since it leads to a HR median above 80% and a CSI close to the optimal value. Moreover, this threshold approximately corresponds to the size of the smallest hand-labeled bow echoes.

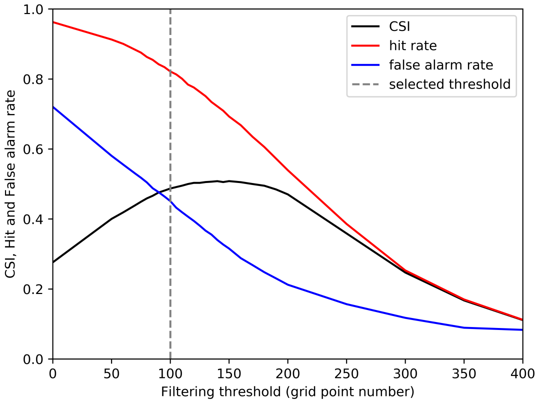

FIG. 6. Scores as a function of filtering threshold. Scores are computed for thresholds from 0 to 400 grid points. Considering the 405 configurations, the CSI, HR and FAR median are plotted in respectively black, red and blue.

c. Results

The sensitivity test results are divided into two parts. The first part focuses on the U-Nets that always predict the same probability value for the class 1 in each grid point. This probability value is equal to 0 or 0.5, depending on the U-Net. This issue is well known for cases of unbalanced datasets (Chawla et al. 2004). For these U-Nets, the loss function remains unchanged according to the epochs. They converge to an unskillful solution where the class 1 is completely missing, or unable to differentiate the class 0 and 1 when the probabilities are equal to 0.5 in every grid point. These U-Nets are called unskillful U-Nets hereafter. The second part examines the performances of other skillful U-Nets able to predict bow echoes.

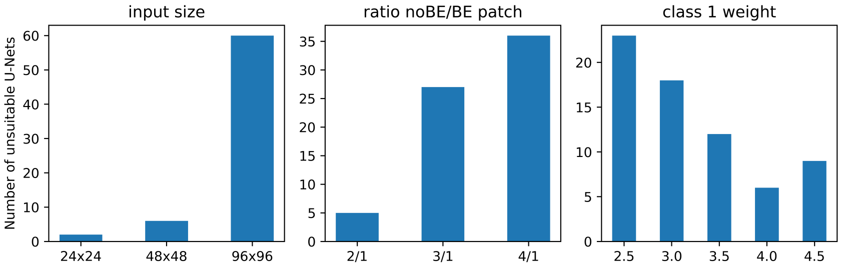

Fig. 7. The number of unskillful U-Nets is counted according to each parameter: input size (left), ratio_noBE/BE patch (center), weight of class 1 in loss function (right).

1) unskillful U-Nets

BEs are rare events in space and in time as previously mentioned. The training database must be carefully set up so that a U-Net can properly detect bow echoes. Fig. 7 shows the number of unskillful U-Nets as a function of the tested parameters. Sixty-eight among the 405 U-Nets are found to be unskillful (17%). The number of unskillful U-Nets increases with input size, especially from 48×48 grid points to 96×96 grid points. This result can be explained by the typical length of BEs which is around 100km and corresponds to 40 grid points. Bow echoes occupy relatively less space in a patch of 96×96 grid points than in one of 48×48 grid points. The classes 0 and 1 are more unbalanced in the first case than in the second one and the risk of unskillful U-Nets is higher. Risk of an unskillful U-Net is less frequent when the ratio noBE/BE patch is 2/1 than when it is 3/1 because the classes are less unbalanced. Finally when the weight of class 1 increases, the number of unskillful U-Nets decreases. The main role of a weighted loss function is to avoid unbalanced data issues (Kurth et al. 2018). Those unskillful U-Nets with a HR and FAR equal to 0 are removed in the next sub-subsection.

The intuition behind why a smaller patch size and larger w1 weight produces more skillfull Unets is because the contribution of BE pixels to the overall DL loss function are increased for a smaller patch and a larger w1 weight. In other words, when considering a smaller patch the ratio of pixels labeled 1 (i.e., BE) to pixels labeled 0 (i.e., noBE) is larger. Thus the influence of BE pixels on the loss (i.e., what the DL model learns) is larger. Similarly, the larger w1 weight for the BE class accomplishes the same increased influence.

2) configuration scores

HR, FAR and CSI are analyzed for each hyperparameter. Only the skillful U-Nets are considered in this sub-subsection. The major results are presented in Fig. 8. The size of patches is the main parameter that explains the fluctuation of scores. HR and FAR tend to decrease with an increasing input size. The CSI is lower for 24×24 input because of a high FAR. These results can be explained by the small input size (24×24) compared to the typical length of BEs (40 grid points). The size of U-Net inputs should be able to get all information of bow echo objects, otherwise all BEs in the preprocessing step (Fig. 3, step 1) are split into different CNN inputs and U-Nets can not properly learn the BE features. Focusing on false alarms, they mainly correspond to objects with moderate pseudo-reflectivities (Q90 maximum around 30 $mm \cdot h^{-1}$). The false alarm distribution slightly changes with input size, with a shift towards weaker pseudo-reflectivities for small input sizes. This result is also consistent with the previous remark concerning the BEs split into different CNN inputs. Considering bow echo class weight (w_1), HR tends to increase with higher weights (same tendency for FAR). Concerning CSI, optimum weights depend on input size, but no trend is found for 48×48 and 96×96 inputs. However, for 24×24 input, CSI is higher for the lowest weights thanks to much lower FAR. Ratio noBE/BE does not have any significant influence on the different scores contrary to what has been mentioned in the above paragraph about unskillful U-Nets (not shown). The role of the different parameters can be summarized as follows:

- Input size must be carefully selected because of its influence on unskillful U-Nets (many more with large-size input) and scores (HR and FAR decrease with input size).
- Bow echo class weight (w_1) is also an important parameter which influences unskillful U-Nets and global scores (higher HR and FAR with higher w_1). Optimum weights depend also on input size.

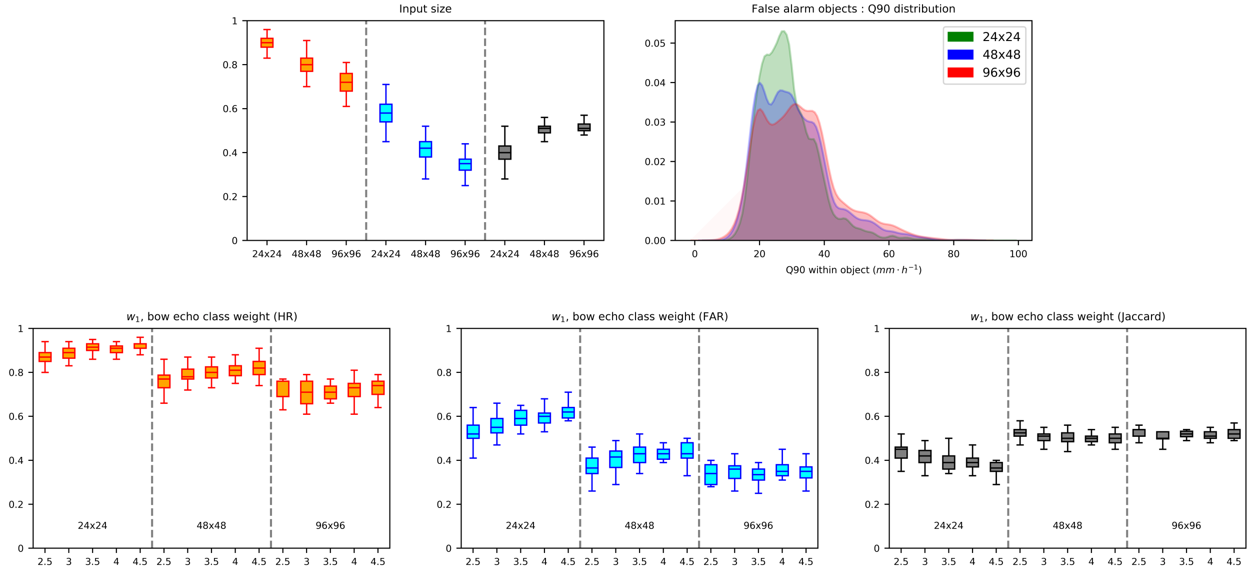

FIG. 8. Performance scores for skillful U-Nets. On the first line, boxplots for HR (orange), FAR (blue) and CSI (black) are represented according to input size (left). False alarm distributions for Q90 attribute in $mm \cdot h^{-1}$ are plotted for the three input sizes (right). On the last line, a focus on bow echo class weight (w_1) and links to input size is proposed. For HR (left), FAR (center) and CSI (right), scores are plotted according to weights and input sizes separated by dashed gray lines.

- Though ratio noBE/BE has an impact on the number of unskillful U-Nets, it does not influence HR and FAR.

d. *Optimal configuration*

In an operational context, we have to chose a configuration among the 405 ones presented in the section 3.a. To make a choice, the fifteen best configurations according to CSI are presented in Table 4. Only input sizes of 48×48 or 96×96 are included in these fifteen configurations. Even if the CSIs of these configurations are very close, different "strategies" are possible. The HR/FAR pairs for each configuration can have high (8^{th} : 0.8/0.36) or low (3^{rd} : 0.71/0.25) values and consequently the total number of detections varies greatly from one U-Net to another. A way to decide which U-Net is optimal is to consider how this CNN will be used. Since BEs are severe but rare events, misses should be avoided. A U-Net with a high HR is preferable. The optimal configuration on this specific point is the 10^{th} setup with a HR of 0.86. This configuration is now considered the optimal one and will be described in more details in the next section.

TABLE 4. Fifteen best configurations according to the CSI. The values of the five tested hyperparameters are indicated in the following order : input size, data augmentation threshold, ratio noBE/BE, heavy pseudo-reflectivity patch percentage and bow echo class weight. HR and FAR are also shown for each configuration. The selected configuration is in bold.

U-Net configurations					HR	FAR	CSI
96x96	30	3/1	15 %	4	0.73	0.25	0.58
48x48	60	4/1	35 %	2.5	0.76	0.29	0.58
96x96	60	2/1	35 %	3.5	0.71	0.25	0.57
96x96	60	2/1	25 %	4.5	0.78	0.32	0.56
96x96	40	3/1	25 %	4.5	0.77	0.32	0.56
96x96	40	2/1	15 %	2.5	0.77	0.34	0.56
48x48	60	3/1	35 %	2.5	0.73	0.29	0.56
48x48	40	3/1	35 %	3.5	0.8	0.36	0.56
48x48	40	3/1	25 %	4	0.77	0.33	0.56
48x48	**40**	**2/1**	**15 %**	**2.5**	**0.86**	**0.39**	**0.56**
96x96	60	4/1	25 %	4.5	0.72	0.29	0.55
96x96	60	2/1	35 %	4.5	0.78	0.35	0.55
96x96	30	2/1	15 %	4	0.75	0.33	0.55
48x48	60	4/1	35 %	4.5	0.82	0.38	0.55
48x48	40	4/1	25 %	3	0.72	0.29	0.55

4. Optimal U-Net : an object-oriented evaluation

Global scores are not sufficient to precisely evaluate the U-Net skills. To go further, attributes of BE objects must be analyzed to support the HR, FAR and CSI scores. Those scores described in the section 2.h provide no information about the detection quality but only about the BE detection and the BE labeling simultaneously in the same field. A com-

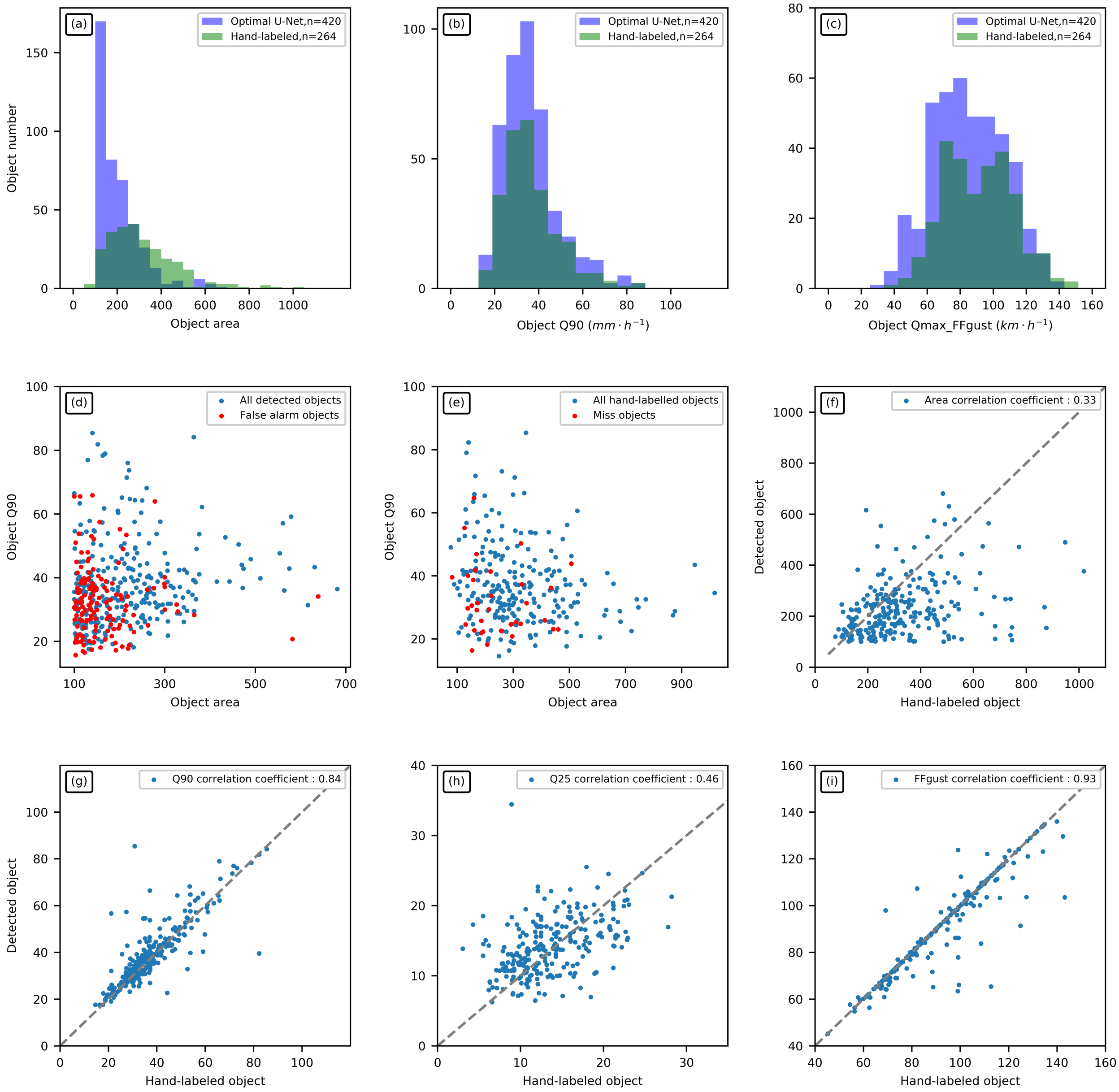

FIG. 9. Evaluation of the U-Net optimal configuration with an object-oriented approach. At the top left, area histograms for the optimal configuration (blue) and the hand-labeled dataset (green) are compared (a). The total number of BEs is visible in the legend. The same is done for Q90 attribute (b) and maximum FFgust (c). Then, the characteristics of false alarm (d) and miss (e) features are presented. All detected objects are plotted according to area and Q90 in blue, the false alarms are in red. The same is done for the hand-labeled dataset (blue) and the misses (red). The correlation between matching pairs of hand-labeled objects (x-axis) and detected objects (y-axis) is studied for four attributes : area (f), Q90 (g), Q25 (h) and Qmax_FFgust (i). The gray dashed line represents a perfect match (y=x). The correlation coefficient is also mentioned in the legend.

prehensive object-oriented evaluation is done on the validation dataset. Results are presented in Fig. 9. Examples of correct detections (Fig.S3), false alarms (Fig.S4) and misses (Fig.S5) are available in the online supplement to this paper.

a. Attribute histograms

The global distributions of bow echoes areas and Q90 are first studied (Fig. 9a and b). The area distribution is different between detected and hand-labeled datasets. The total number of detected objects is rather high with a hit rate of 0.86 and a false alarm rate of 0.39. No detected object has an area below 100 grid points thanks to the filtering threshold. Detected objects have mostly an area below 250 grid points with several between 100 and 150 grid points. The number of detected objects decreases exponentially with area size. This behavior is not observed for the hand-labeled dataset, with a maximum around 300 grid points and objects more normally distributed. Detected objects are more frequent but they tend to be smaller than hand-labeled objects. The distributions of pseudo-reflectivity Q90 for detected and hand-labeled objects are quite similar even if the total number of detected objects is higher. The maximum of the Q90 distribution is around 35 $mm{\cdot}h^{-1}$ for both detected or hand-labeled datasets.

The hand-labeled distribution for maximum FFgust (Fig. 9c) has a range from 50 $km \cdot h^{-1}$ to 150 $km \cdot h^{-1}$ with two maximums around 70-80 $km \cdot h^{-1}$ and 110 $km \cdot h^{-1}$. The hand-labeled BEs associated with weak FFgusts are rare. They are examined one by one and they are mainly BEs at the beginning of their life cycle with weaker wind gusts. The distribution of detected objects is very similar with a single maximum around 80 $km{\cdot}h^{-1}$. Even if FFgust is not a predictor of U-Nets, the comparison of detected and hand-labeled objects shows that the detected BEs are associated with the same FFgust distribution as the hand-labeled BEs. The wide range of FFgust advocates the addition of a severity attribute for each BE detection to differentiate 'moderate' BEs without significant damage from BEs with potentially significant damages.

b. False alarm and missed attributes

To extract false alarm and missed features, each object is represented in area/Q90 graphs (Fig. 9d and e). BEs with high (respectively low) Q90 are usually associated with small (respectively large) areas at first glance. This characteristic is physically consistent with the life cycle of BEs and the hand-labeled objects (strong and small at the beginning and then weaker and larger (Goulet 2015)). This is a reassuring aspect concerning the general behavior of the U-Net. The false alarms mainly correspond to small and weak objects. On the other hand, the large or strong objects are usually correctly detected. In an operational context, this information may prove useful when discussing the relevance of object detections. The number of misses is limited but their distribution is more homogeneous. They are slightly more concentrated on small objects but this result is not as obvious as for false alarms. Two characteristics are nevertheless common : the areas of missed objects are always below 500 grid points and the maximum wind gusts are constantly below 110 $km \cdot h^{-1}$. The prediction of BEs is more difficult for small and weak objects. This may be related to the more difficult recognition of these BEs by the expert, leading to a less precise labeling. In such cases, the U-Net is not very accurate.

c. Attribute correlations

This section defines a pair of matching objects (detected/hand-labeled) in order to compute attribute correlations. A pair is formed when the distance between two mass centers is less than 100km. Object attributes are compared and a correlation coefficient is computed (Fig. 9f,g,h and i). The correlation for the areas is weak. The U-Net tends to underestimate the size of objects, especially for the large bow echoes. This conclusion is consistent with the comparison of the area histograms. Regarding Q90, the correlation is higher and points are located around the dashed gray line (y=x). Q90 is representative of the most active part (i.e heaviest pseudo-reflectivities) located in the center of BE objects. Q25 is more representative of the BE borders with lower pseudo-reflectivities, and the correlation is worse than that of Q90. These findings point to the ability of the U-Net to properly retrieve the most relevant part of BEs, whereas the BEs borders are less well estimated. The strong correlation of FFgust maximum (as well as Q90) supports the conclusion of the previous sentence since the strongest FFgusts are in the most relevant part. In subsection 6.c, the results about the false alarms, misses and correlation attributes will be further discussed and compared with comments from forecasters.

d. BE and SC confusion

The capacity of the U-Net to differentiate BEs from other types of severe convective storms is also verified. The main issue could be to mix up BE with other convective storms as intensity can be similar. Initially, the CNNs detected all patterns with strong pseudo-reflectivities as BEs (section 2.d, last item). To remove any doubt, it is verified that BEs are not mixed up with isolated supercells (SCs, most common severe convective storm in France) when the confusion between the two convective events is very unlikely for a human. Eleven convective situations that occurred in 2019 and in 2020 are examined (Table 5). These situations are covered by 5066 pseudo-reflectivity forecast fields from AROME-EPS, in

TABLE 5. BE and supercell (SC) overlaps. For each date, the number of fields with hand-labeled supercells (second column) and predicted bow echoes (third column) is reported. The last column, overlaps, shows the number of fields where predicted BE and hand-labeled SC objects have at least one grid point in common.

Date	SCs	BEs	Overlaps
2019-06-15	67	94	3
2019-06-18	37	27	0
2019-06-19	18	21	0
2019-06-24	47	88	6
2019-07-06	36	6	1
2019-07-15	172	12	1
2019-07-26	34	52	0
2019-08-09	90	38	1
2019-08-18	54	38	1
2020-04-17	25	17	1
2020-05-09	48	37	1
Total	628	430	15

which a same expert hand-labeled 628 SCs. The labeling of SCs rely on pseudo-reflectivity fields and UHmax between 800 and 500 hPa. We keep the maximum in absolute value of UH to correctly detect both right and left moving SCs. The specific contour of 50 $m^2 \cdot s^{-2}$ or (-50 $m^2 \cdot s^{-2}$ for negative values) is plotted over the pseudo-reflectivity fields to help the labeling process. Discrete and embedded SCs in lines of convection are hand-labeled. High reflectivties ($>$ 50 $mm \cdot h^{-1}$) and high UHmax values are required to label a SC. In this analysis, predicted BEs from the U-Net are compared with hand-labeled SC objects. The confusions between SCs and BEs are rare with only 15 occurrences for over 628 SCs and 430 BEs. As SCs can be embedded in quasi-linear convective systems, these overlaps can be understandable. A SC can also sometimes evolve into a BE (Klimowski et al. 2004) and during the transition phase, an overlap between BE and SC labels is possible. These 15 overlapping occurrences have been manually analyzed to check whether the overlaps occurred in one of the previously mentioned two cases. Only 3 of the 15 overlapping occurrences do not occur in one of the two cases and can be considered as abnormal overlaps after examination. To conclude, the risk of confusion between BEs and SCs is very limited with only 0.5% of SCs wrongly detected as BEs.

5. Extension to the AROME deterministic model

At this stage, the U-Net has been trained and tested using AROME-EPS forecasts. The possibility to detect BEs in the deterministic AROME forecasts, which has a higher resolution (1.3 km instead of 2.5 km), is also important for forecasters. Transfer learning from AROME-EPS to AROME is tested because data from deterministic models are limited and setting up another training database would be very time-consuming. We test the optimal configuration on the AROME deterministic model without re-training after encouraging results with AROME-EPS. Pseudo-reflectivity fields from AROME are available on the same grid as AROME-EPS (2.5 km) thanks to a quadratic interpolation method. The cases in the AROME and AROME-EPS validation databases should be the same to have comparable results. There are only 348 pseudo-reflectivity fields and 62 hand-labeled bow echoes in its validation database because deterministic AROME is a single forecast, which can limit the significance of results.

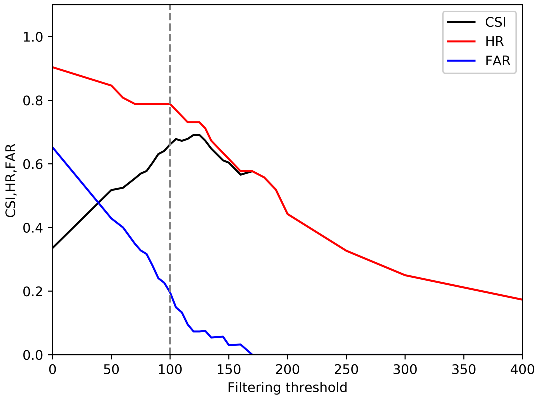

FIG. 10. Similar plot to Fig. 6, but only the optimal U-Net is considered and not all U-Net configurations.

a. Filtering threshold and global scores

Even if the pseudo-reflectivity fields have the same resolution (2.5km) as those of AROME-EPS, the native resolution of AROME is smaller (1.3km) and the pseudo-reflectivity fields in the AROME deterministic model are more realistic with stronger pseudo-reflectivity gradients and consequently we re-examined the filtering threshold applied to AROME data. A similar graph to Fig. 6 is plotted in Fig. 10. With only one setup, the curves are noisier. The maximum CSI is around 100-150 grid points which is in favor of retaining the same filtering threshold of 100 grid points as for AROME-EPS (dashed grey line). With this threshold, the HR (respectively FAR) is equal to 0.79 (respectively 0.2). The FAR is better than the one for the AROME-EPS outputs

(0.39) while the HR is slightly lower. There are several possible explanations besides the small size of the database. Bow echo recognition is in this case easier for experts with less ambiguous cases because pseudo-reflectivity fields are more realistic. A hand-labeled database with higher quality and stronger pseudo-reflectivity gradients can both explain the lower FAR.

b. Results

The same figures as in the previous section (Fig. 11 from a to i) are presented to study more specifically the U-Net behavior in AROME forecast outputs. The same U-Net behavior is observed despite the small database. Even if the number of hand-labeled and detected bow echoes is more even, undersized objects are still predicted with a correct Q90 distribution. The FFgust distributions are also comparable even if some U-Net detections, which are false alarm objects, are associated with weak FFgusts. The false alarm and missed attributes are also very similar to that of the AROME-EPS database. The false alarms correspond to weak and small objects. The misses concern mostly small and weak BEs, but a small and strong BE (Q90 around 85 $mm \cdot h^{-1}$) is also missed. The attribute correlation results are also very similar to AROME-EPS with a high correlation coefficient for the Q90 and FFgust attributes but a worse correlation for area and Q25. No significant change can be noticed when the comparison is made with the AROME-EPS database except a better FAR. We conclude that the same U-Net can be used to detect bow echoes in AROME deterministic model. The extension to AROME will be further discussed in subsection 6.c based on the feedback from the subjective comparison by forecasters. The possibility to extend U-Nets from EPS to deterministic models is an advantage of both DL and ML methods. EPS outputs are indeed well adapted to these methods which require large datasets (Schumacher et al. 2021). This is not always the case for deterministic models. However, this extension is realistic if both EPS and deterministic models are sufficiently similar. For instance, the grid of EPS and deterministic model outputs should be close to assume that spatial features learned during the training process of the EPS database are still valid for the deterministic models. Otherwise, interpolating model outputs could be tested.

6. Utilization of BE detections by end-users : synthesis plots and feedback

The satisfactory evaluation of the U-Net performances for both AROME-EPS and AROME outputs motivates the development of forecasting products about BE detections. For that purpose, synthesis plots are presented in this section. They have been designed in collaboration with forecasters, inspired by Demuth et al. (2020). Three synthesis plots are presented: trajectory and paintball plots in the first subsection and a probability map of BE occurrences in the second subsection. To facilitate the use of U-Net outputs, each AROME-EPS run is divided into three different time periods (Day 1, Night Day 1/2 and Day 2). Thanks to these time periods, the entire life cycle of a BE can be followed over only one of them. The three synthesis plots are computed for the three time periods and consequently, for each AROME-EPS run, 3×3 synthesis plots are produced. The three time periods are based on UTC hour to facilitate comparisons between AROME-EPS runs. The last subsection is a summary of forecasters' feedback concerning the U-Net performance and the synthesis plots.

a. Trajectory and paintball plots

One way to summarize U-Net outputs is to plot the overlapping of the detections of every member for a time period (Day 1, Night day1/2 or Day2). The trajectory plot is very useful to visualize the temporal evolution of a BE in AROME forecasts with a different color for each UTC hour. Fig. 12a shows the BE northward trajectory over the North-East of France. It gives useful information about the BE risk period which is during the afternoon and early evening in this case (14UTC and 18UTC from green to yellow).

The paintball plot (Fig. 12b) helps evaluate the number of different members that predict a BE with a different color for each member. Focusing on a specific BE object, we can assess the member and the UTC time of the detection by combining both the trajectory and paintball plots. The U-Net outputs can be quickly identified thanks to this information.

b. Probabilistic approach

Trajectory and paintball plots are very useful to evaluate BE risks. However, the interpretation of a large number of detections over a small area as shown in Figs. 12a and b can remain troublesome. A probabilistic approach is useful to quantify the risk of BEs in AROME-EPS.

1) Space-time tolerance

A probabilistic synthesis should allow for some space and time tolerances because of the rather low number of AROME-EPS members and the small size

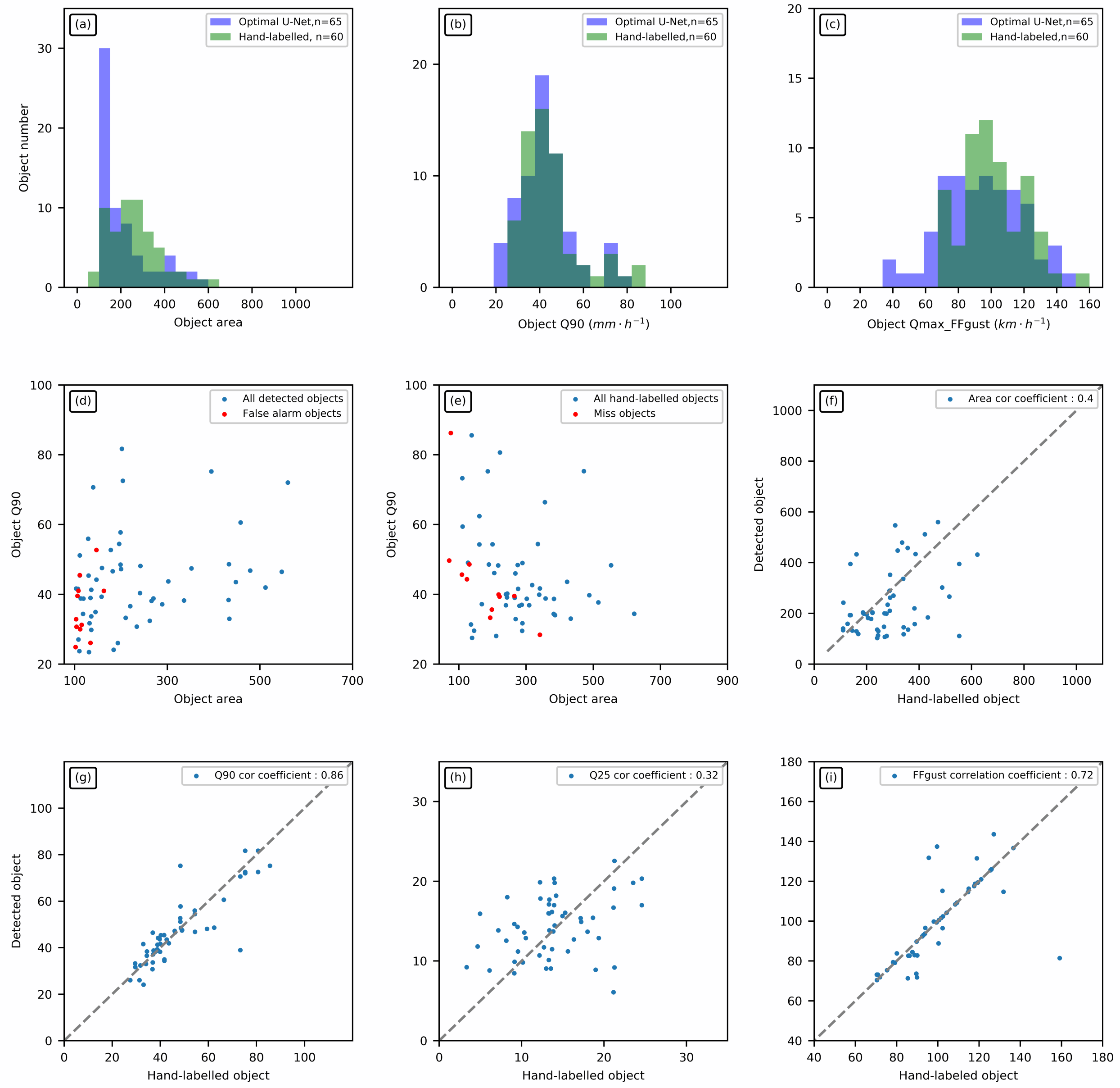

FIG. 11. Evaluation of the U-Net on the AROME deterministic model database. The subplots are the same as in Fig. 9.

of BEs. For that purpose, a neighborhood disc of radius εkm, denoted d_ε, is defined around each grid point x:

$$d_\varepsilon(x) = \{y \mid \|x - y\|_2 \leq \varepsilon\} \tag{5}$$

The value of ε will be discussed in the next paragraph. A neighborhood probability of BE ($P_{\varepsilon,l}$) is defined by :

$$\mathbb{P}^m_{\varepsilon,l}(x) = \begin{cases} 1 & if \ \sum_{y \in d_\varepsilon(x)} \mathbb{1}_{BE}(y) \ \geq l \\ 0 & \quad otherwise \end{cases} \tag{6}$$

$\mathbb{1}_{BE}(y)$ corresponds to the U-Net output for the grid point y (1 in a BE, 0 out). A grid point x for a member m is associated with a probability of 1 if at least l grid points in the neighborhood of x correspond to a BE detection. To include a time tolerance, the probability at time t is computed using forecasts valid at t, $t - 1h$ and $t + 1h$:

$$\mathbb{P}_{\varepsilon,l}(x,t) = \frac{1}{3 \times N_{mb}} \sum_{i=-1}^{1} \sum_{m=1}^{N_{mb}} \mathbb{P}^m_{\varepsilon,l}(x, t+i), \tag{7}$$

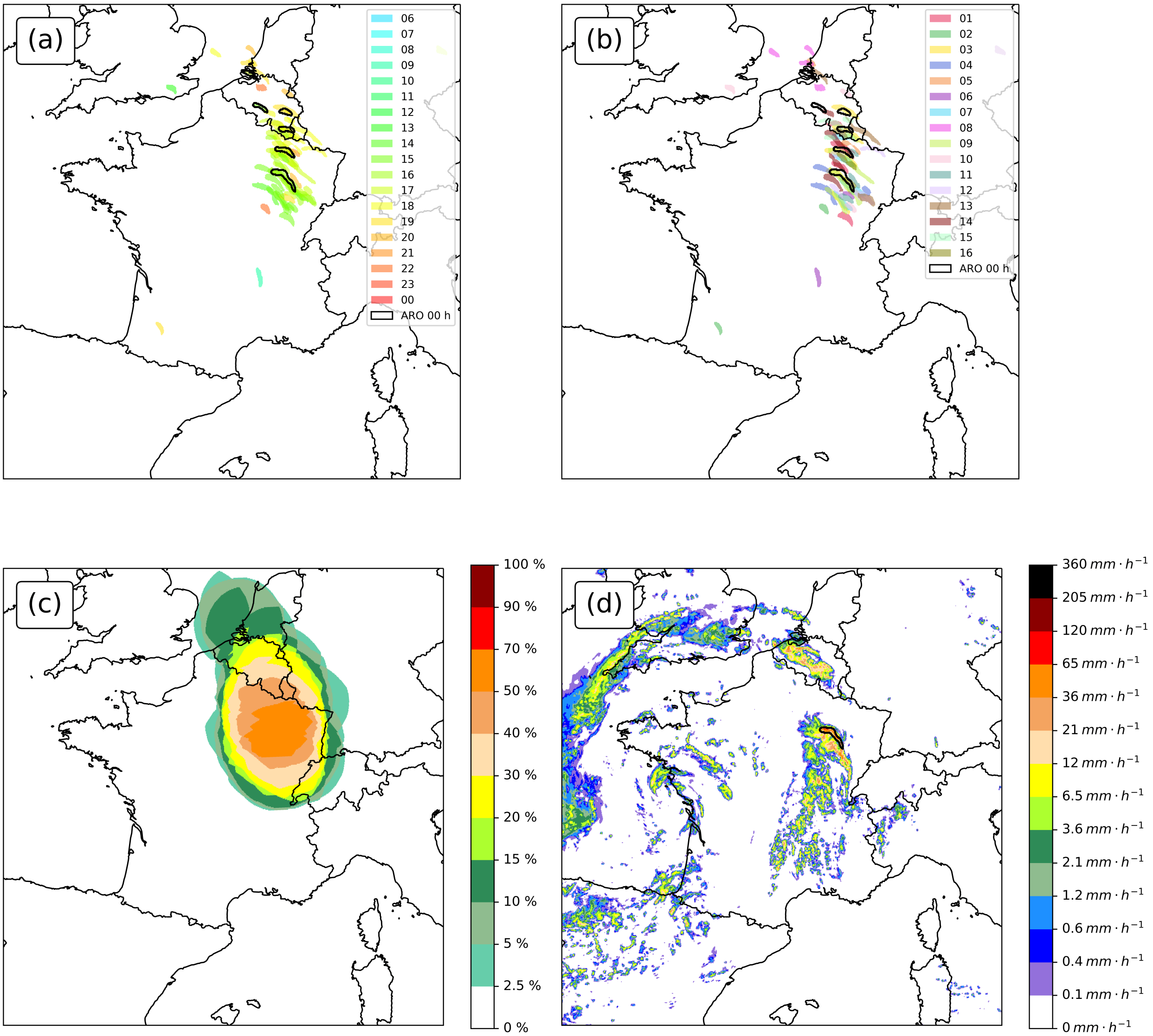

FIG. 12. Synthesis plots of U-Net detections. The forecasts from AROME-EPS launched on 12 June 2020 at 03UTC and from the deterministic AROME model launched on 12 June 2020 at 00UTC are taken as an example. The considered period of forecasts is Day 1, ranging from 12 June 2020 at 06UTC to 13 June 2020 at 00UTC. Trajectory plot is on the top left (a), paintball plot on the top right (b) and probabilistic plot (c) on the bottom left. A different color is used for each UTC hour of the BE detections for the trajectory plot (a). The paintball plot (b) represents the same detections as the trajectory ones, but a different color is used for each member (from 1 to 16) instead of UTC time. Black contours are added around deterministic AROME objects. On the bottom right (d), the predicted pseudo-reflectivity field and the corresponding BE detection (black contour) are overlaid, using the AROME outputs valid on 12 June 2020 at 16UTC. Forecasters examine this kind of superposition (d) in section 6.c for their feedback.

where N_{mb} is the total number of members in AROME-EPS and $\mathbb{P}^{m}_{\varepsilon,l}(x, t+i)$ indicates the BE probability of a grid point x at time $t+i$ for a member m. A probability is defined for each grid point at a specific time t by taking into account equally all AROME-EPS members. Over a given time period (Day 1, Night Day 1/2 or Day 2), only the maximum probability at each grid point is kept (Eq. 8).

$$\mathbb{P}^{max}_{\varepsilon,l}(x) = max\,\{t \in period,\ \mathbb{P}_{\varepsilon,l}(x,t)\} \qquad (8)$$

This maximum probability plot is presented in Fig. 12c. The value of the radius ε is equal to 150km and will be discussed hereafter. The value of l is fixed to 10 hereafter and will not be discussed in aid

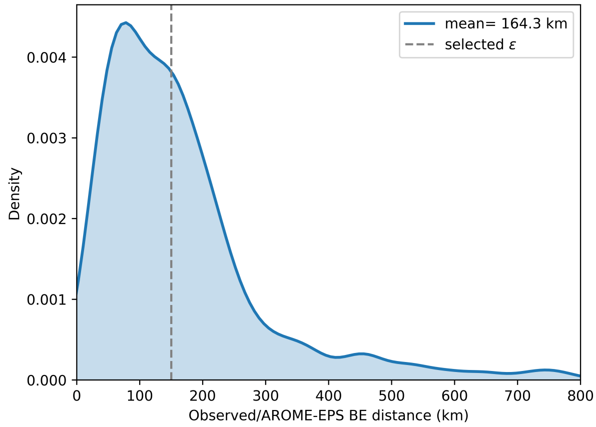

FIG. 13. Density plot of distances between observed BEs and AROME-EPS detections. The average distance is computed in legend (excluding errors above 500km, see text for explanation).

of the value of the radius ε. The maximum probability shown on Fig. 12c is above 50 % (high probability for a BE risk) over Northeastern France, which is consistent with the other synthesis plots. Moreover, isolated detections like the ones over the Southwest of France are filtered (risk below 2,5 %) because of time averaging.

2) NEIGHBORHOOD DISTANCE

An optimal value of the radius ε used for the probability plot should be defined. Position errors between observed BEs and AROME-EPS detections are examined as this ε value depends on AROME-EPS skill. Observed BEs have been hand-labeled on radar data, in the same way as for AROME training or validation databases. However, radar data grid covers only mainland France and near the borders. Only AROME-EPS detections over this smaller grid are kept. A density plot of position errors is presented in Fig. 13. The majority of position errors are under 300km. The average error is around 180km but this value seems to be overestimated because of some unrealistic distances of several hundred kilometers. In that case the detected object should be considered as a false alarm and the observed object as a miss. If distance values above 500km are ignored, the average error is around 150km. The value of ε is set to 150km based on this result.

c. Forecasters' feedback

In this study, a U-Net has been trained and tested to detect bow echoes in AROME-EPS and AROME forecasts. Only one expert has hand-labeled the training and validation databases. All findings and scores are based on these hand-labeled contours of bow echoes. In some cases the detection of BEs is highly subjective and there is a need for more experts to assess the U-Net performance. A subjective evaluation from future users was considered interesting in addition to object-oriented scores. Ten Météo-France forecasters contribute to this study. Each expert is given a superposition of pseudo-reflectivity field and U-Net detection such as that of Fig. 12d to subjectively evaluate the U-Net. The examined cases come from the validation AROME-EPS database and the AROME model. Since the number of pictures is very large (around 3000), experts are divided into three groups, and each group examines a third of the pictures. Each forecaster uses an evaluation form for their evaluation. Each forecaster counts the number of false alarms, misses or correct detections for each date and each member using the same criteria as in section 2.h. A detection rate and false alarm rate are calculated and compared with the HR and FAR of sections 4 and 5. This comparison is feasible because the HR and FAR calculated with the complete validation database are similar to HR and FAR computed with the three subsets one by one (more or less 3%). A last section named 'general comments' in the evaluation form allows for expressing an opinion concerning U-Net behavior. The results of the U-Net object-oriented evaluation are not communicated to the experts to avoid influencing their judgement. Concerning the general comment section, some guidelines are given to focus on miss, false alarm features (more frequent in case of large/small area or strong/weak intensity ?), comparison of U-Net skills between AROME-EPS and AROME but also concerning confusion with other MCSs. Each expert examines, independently of the others, the U-Net outputs. Feedback and conclusions have been gathered hereafter (more details concerning the forecasters' feedback are available in the online supplement to this paper). Based on a broad consensus, the main findings are reported in the following section :

- The feedback is overall positive. The pair of HR/FAR according to the experts is around 70%/20%. Most of the forecasters consider that these detections and this work can facilitate the use of EPS in an operational context. The detection of large BEs is acceptable (still based on the general comments). This point is consistent with findings of section 4.
- Only a few cases are identified as false alarms (mean FAR of 20%). This conclusion is very promising for an operational application, but it contradicts the higher false alarm rate of 39 %

obtained in section 4. Even if most forecasters have found less false alarms, some of them comply with the conclusions of sections 4 and 5 (higher false alarm rate that corresponds to small objects). Some experts consider small objects as cell bow echoes (Klimowski et al. 2004), which can explain these different judgments. This result highlights the subjective point of view of different experts concerning the exact BE definition and the fact that they only had the pseudo-reflectivity field at their disposal to assess it in the model, as the U-Net did. Forecasters tend to favor detection algorithm with high FAR because all detections are carefully examined. On the contrary, a synthesis plot without detection does not draw attention. This conclusion supports the selection of the optimal configuration as described in section 4, a high HR and high FAR are preferred.

- Some misses are also noticed (mean HR of 70%). There are three main reasons to explain these misses. First in case of weak pseudo-reflectivities, which is consistent with the object-oriented evaluation. In case of fragmented BE, the detection can be missed or incomplete. The last one highlights a lack of temporal robustness/stability of the detections. A BE is sometimes not detected at each lead time: for instance a BE is detected at the time t, not at $t + 1h$ and detected once again at $t + 2h$. In that case, misses are less problematic because they mixed together all the others in the synthesis plots.

- BE contours are most of the time too small and focus on the strongest part of BEs. This observation is consistent with results in sections 4 and 5.

- The detection quality in AROME-EPS and AROME is approximately the same according to the majority of experts (consistent with section 5). The risk of confusion between SCs and BEs in AROME is slightly more important according to some of them.

- A last comment concerns the poor U-Net behavior for a specific weather situation in 2018. In this case, the BE had an unusual propagation axis from South-East to North-West in comparison to BE climatology in France (South-West/North-East or West/East). Both the original curve and the fragmented characteristics of this BE can explain these very bad results of the U-Net. Fig. 14 is a good example with a BE correctly detected at 12 and 13UTC but not at 14UTC. Unusual cases are always difficult to handle for neural networks when these cases are not representative in the training samples.

- Trajectory plot (Fig. 12a) is the preferred visualization for the majority of the expert group, even though this conclusion should be confirmed in an operational context.

In summary, forecasters' feedback is overall positive and consistent with the object-oriented evaluation. This experiment is also a first step toward introducing AI-based products in the forecasters tools. The cooperation to design the synthesis plots in section 6 was strongly appreciated.

7. Conclusion and future work

In this paper a convolutional neural network inspired from a U-Net architecture has been developed to detect Bow Echoes (BEs), specific Mesoscale Convective Sytems (MCSs), in the outputs of the French kilometer-scale AROME and AROME-EPS models. This U-Net is trained and evaluated using a hand-labeled dataset of more than 9000 training samples. Pseudo-reflectivity (pseudo because expressed in $mm \cdot h^{-1}$ and not in dBZ) is the unique input of the neural network. Pseudo-reflectivity fields are cut into smaller patches and several pre-processing steps are implemented in order to design a balanced training database. During the prediction process, pseudo-reflectivity fields are also cut into smaller overlapping patches and fields are reconstructed to detect bow echoes over the whole AROME domain.

A filtering threshold has been defined and fixed to 100 grid points to optimize object-oriented scores. Different configurations are tested to understand the role of several hyperparameters involved in the training dataset design and in the U-Net. The input patch size is the most influential parameter. The size should be large enough to extract information from the whole BE and its surroundings. The proportions of training patches with and without BEs, and the weights in the loss function are also influential hyperparameters. The optimal configuration is determined from the CSI and the hit rate.

This optimal configuration was examined in details with an object-oriented approach. The false alarms and misses mainly correspond to small and weak objects. The U-Net underestimates the size of BEs but properly captures the most active part of BEs. Even if the wind gust speed at 10m (FFgust) is not a predictor for the U-Net, the detected BEs are associated with the same FFgust distribution as the hand-labeled BEs. An extension of the U-Net

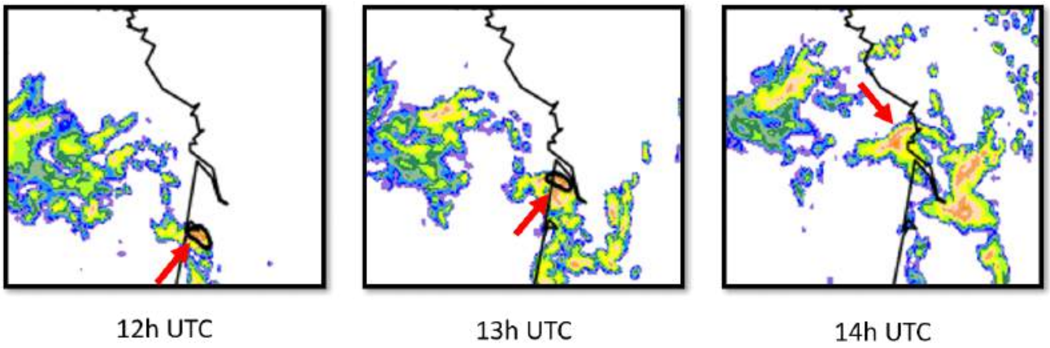

FIG. 14. Forecasts from AROME-EPS initialized on 25 May 2018 at 21UTC, zoomed over western France. Pseudo-reflectivity fields and BE detections (black contours) from member 2 between 26 May 2018 at 12UTC (left) to 14UTC (right) are overlaid. The red arrow indicates the BE.

to AROME output without re-training provided satisfactory results. The quality of the detection, as measured by the object-oriented evaluations, is similar for AROME-EPS and AROME. The success of transfer learning is very promising for future NWP model evolutions, suggesting that re-training the U-Net each time the AROME model changes may not be necessary.

Forecasters have been involved to develop three kinds of synthesis plots to summarize BE detections in AROME-EPS and AROME. They sum up information of more than 700 fields and thus facilitate forecasters' expertise. These plots include a trajectory visualization that presents all detections as a function of their validity time, a paintball visualization with colors according to the member of AROME-EPS and a probability plot including space-time tolerances. To subjectively assess the quality of BE detections, forecasters examine several hundreds of pictures. Their feedback mainly supports the results of the previous sections but more formal evaluations are necessary to validate these findings. To continue the evaluation in a real-time context, a daily production in a research mode is deployed since summer 2021. This daily production is enabled thanks to a short computation time in operations mode. The U-Net predictions (for all AROME-EPS members) take around 10 minutes with a GPU[5].

Future work will explore the addition of FFgust, FF or other parameters inspired by Lagerquist et al. (2017, 2020) as predictors of a new U-Net. This new U-Net could be compared with the current optimal U-Net configuration. Adding a severity attribute based on FFgust in BE objects will be also a priority. Another research direction will explore how to improve the temporal stability of U-Net predictions. Adding pseudo-reflectivity fields at time $t - 1h$ and $t + 1h$ as inputs of the U-Net could be a possible solution. This study is largely driven by an expert's skill. More objectivity is needed and consequently a panel of experts could hand-label the same images. A probability map for the target value instead of a binary map (0 or 1) could be tested. A U-Net could also be tested and tuned to apply on radar data in order to improve the nowcasting of MCSs. These observed BEs could be used for evaluating the quality of AROME-EPS and AROME BE forecasts. Finally, the U-Net approach could be extended to detect other kinds of severe convective storms. Even if the methodology remains similar, the number of cases should be sufficient to train AI methods. The input size of the U-Net could be modified according to the size of the convective storms. Some additional predictors may be deemed necessary (Updraft Helicity for supercell detection for instance).

Acknowledgments. The authors thank Sébastien Baille, Romaric Cinotti, Pierre-Emmanuel Gallerand, Jean Imbert, Julien Léger and Martin Paccaud for their feedback about U-Net that used parts of their spare time. The authors thank also Fabrice Guillemot for his help to selecting relevant parameters for this study. The authors acknowledge Jean-Antoine Maziejewski and Gregory Roux for proofreading and for improving English in the manuscript.

Data availability statement. The three datasets (training, validation and deterministic AROME databases) and the weights of the optimal U-Net are available online in open access (Mounier (2021) , doi : `10.5281/zenodo.5534445`, license : Etalab Open License 2.0)

[5] NVIDIA Corporation GP102 (GeForce GTX 1080 Ti)

References

Abadi, M., and Coauthors, 2015: TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. URL https://www.tensorflow.org/, software available from tensorflow.org.

Atkins, N., and M. S. Laurent, 2009: Bow echo mesovortices. part i: Processes that influence their damaging potential. *Mon. Wea. Rev.*, **137**, 1497–1513, doi:10.1175/2008MWR2649.1.

Biard, J., and K. Kunkel, 2019: Automated detection of weather fronts using a deep learning neural network. *Advances in Statistical Climatology, Meteorology and Oceanography*, **5**, 147–160, doi:10.5194/ascmo-5-147-2019.

Bouttier, F., and L. Raynaud, 2018: Clustering and selection of boundary conditions for limited-area ensemble prediction. *Quart. J. Roy. Meteor. Soc.*, **144**, 2381–2391, doi:10.1002/qj.3304.

Bouttier, F., L. Raynaud, O. Nuissier, and B. Ménétrier, 2016: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quart. J. Roy. Meteor. Soc.*, **142**, 390–403, doi:10.1002/qj.2622.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, doi:10.1175/MWR-D-12-00031.1.

Bowler, N., A. Arribas, K. Mylne, K. Robertson, and S. Beare, 2008: The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **134**, 703–722, doi:10.1002/qj.234.

Brousseau, P., Y. Seity, D. Ricard, and J. Léger, 2016: Improvement of the forecast of convective activity from the arome-france system. *Quart. J. Roy. Meteor. Soc.*, **142**, 2231–2243, doi:10.1002/qj.2822.

Caumont, O., and Coauthors, 2006: A radar simulator for high-resolution nonhydrostatic models. *J. Atmos. Oceanic Technol.*, **23**, 1049–1067, doi:10.1175/JTECH1905.1.

Chawla, N., N. Japkowicz, and A. Kotcz, 2004: Editorial: Special issue on learning from imbalanced data sets. *SIGKDD Explorations Newsletter*, **6**, 1–6, doi:10.1145/1007730.1007733.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, doi:10.1175/MWR3145.1.

Demuth, J., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for nws forecasters. *Wea. Forecasting*, **35**, 1381–1406, doi:10.1175/WAF-D-19-0108.1.

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (wrf) model. *Atmos. Sc. Lett.*, **5**, 110–117, doi:10.1002/asl.72.

Du, J., G. DiMego, M. Tracton, and B. Zhou, 2003: NCEP short-range ensemble forecasting (SREF) system: multi-IC, multi-model and multi-physics approach. *Research activities in atmospheric and oceanic modelling*, **33**.

Dutta, A., and A. Zisserman, 2019: The VGG image annotator (VIA). *ArXiv*, **abs/1904.10699**.

Ebert, E., 2008: Fuzzy verification of high?resolution gridded forecasts: a review and proposed framework. *Meteor. Appl.*, **15**, 51–64, doi:10.1002/met.25.

Elhoseiny, M., S. Huang, and A. Elgammal, 2015: Weather classification with deep convolutional neural networks. 3349–3353, doi:10.1109/ICIP.2015.7351424.

French, A., and M. Parker, 2014: Numerical simulations of bow echo formation following a squall line supercell merger. *Mon. Wea. Rev.*, **142**, 4791–4822, doi:10.1175/MWR-D-13-00356.1.

Fujita, T., 1978: *Manual of downburst identification for project NIMROD*. Satellite and Mesometeorology Research Paper 156, 104 pp.

Gagne II, D., S. Haupt, D. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, doi:10.1175/MWR-D-18-0316.1.

Gallo, B., A. Clark, B. Smith, R. Thompson, I. Jirak, and S. Dembek, 2019: Incorporating UH Occurrence Time to Ensemble-Derived Tornado Probabilities. *Wea. Forecasting*, **34**, 151–164, doi:10.1175/WAF-D-18-0108.1.

Goulet, L., 2015: Bow echoes : Conceptual schemes and European relevance. *The European Forecaster*, URL http://www.euroforecaster.org/newsletter20/meteofr2.

Hohenegger, C., and C. Schär, 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1794, doi:10.1175/BAMS-88-11-1783.

Houze Jr., R., 2004: Mesoscale convective systems. *Reviews of Geophysics*, **42**, doi:10.1029/2004RG000150.

Jergensen, G., A. McGovern, R. Lagerquist, and T. Smith, 2020: Classifying convective storms using machine learning. *Wea. Forecasting*, **35**, 537–559, doi:10.1175/WAF-D-19-0170.1.

Kain, J., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, doi:10.1175/WAF2007106.1.

Kamani, M., F. Farhat, S. Wistar, and J. Wang, 2016: Shape matching using skeleton context for automated bow echo detection. doi:10.1109/BigData.2016.7840685.

Kingma, D., and J. Ba, 2014: Adam: A method for stochastic optimization. *International Conference on Learning Representations.*

Klimowski, B., M. Hjelmfelt, and M. Bunkers, 2004: Radar observations of the early evolution of bow echoes. *Wea. Forecasting*, **19**, 727–734, doi:10.1175/1520-0434(2004)019⟨0727:ROOTEE⟩2.0.CO;2.

Kurth, T., and Coauthors, 2018: Exascale deep learning for climate analytics. 649–660.

Lagerquist, R., A. McGovern, and D. Gagne II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Wea. Forecasting*, **34**, 1137–1160, doi:10.1175/WAF-D-18-0183.1.

Lagerquist, R., A. McGovern, C. Homeyer, D. Gagne II, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, doi:10.1175/MWR-D-19-0372.1.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, doi:10.1175/WAF-D-17-0038.1.

Laing, A., and M. Fritsch, 1997: The global population of mesoscale convective complexes. *Quart. J. Roy. Meteor. Soc.*, **123**, 389–405, doi:10.1002/qj.49712353807.

LeCun, Y., and Y. Bengio, 1995: Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, **3361**, 1995.

Liu, Y., E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, 2016: Application of deep convolutional neural networks for detecting extreme weather in climate datasets. *arXiv*.

Markowski, P., and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes*. John Wiley and Sons, Ltd, 249-265 pp., doi:10.1002/9780470682104.ch9.

Marshall, J., and W. Palmer, 1948: The distribution of raindrops with size. *J. Atmos. Sci.*, **5**, 165–166, doi:10.1175/1520-0469(1948)005⟨0165:TDORWS⟩2.0.CO;2.

Matsuoka, D., S. Sugimoto, Y. Nakagawa, S. Kawahara, F. Araki, Y. Onoue, M. Iiyama, and K. Koyamada, 2019: Automatic detection of stationary fronts around Japan using a deep convolutional neural network. *SOLA*, **15**, doi:10.2151/sola.2019-028.

McGovern, A., K. Elmore, D. Gagne II, S. Haupt, C. Karstens, R. Lagerquist, T. Smith, and J. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, doi:10.1175/BAMS-D-16-0123.1.

Moller, A., C. Doswell III, M. Foster, and G. Woodall, 1994: The operational recognition of supercell thunderstorm environments and storm structures. *Wea. Forecasting*, **9**, 327–347, doi:10.1175/1520-0434(1994)009⟨0327:TOROST⟩2.0.CO;2.

Montmerle, T., Y. Michel, E. Arbogast, B. Ménétrier, and P. Brousseau, 2018: A 3D ensemble variational data assimilation scheme for the limited-area AROME model: Formulation and preliminary results. *Quart. J. Roy. Meteor. Soc.*, **144**, 2196–2215, doi:10.1002/qj.3334.

Mounier, A., 2021: Detection of bow echoes in French kilometre-scale models (AROME-EPS & AROME models of Météo-France). Zenodo, doi:10.5281/zenodo.5534445.

Patil, V., S. Das, and A. Phadke, 2019: Methods for mesoscale convective systems detection and tracking: a survey. 1–7, doi:10.1109/ICCCNT45670.2019.8944656.

Peralta, C., Z. Ben Bouallègue, S. Theis, C. Gebhardt, and M. Buchhold, 2012: Accounting for initial condition uncertainties in COSMO-DE-EPS. *J. Geophys. Res.*, **117**, 7108–, doi:10.1029/2011JD016581.

Przybylinski, R., 1995: The bow echo: Observations, numerical simulations, and severe weather detection methods. *Wea. Forecasting*, **10**, 203–218, doi:10.1175/1520-0434(1995)010⟨0203:TBEONS⟩2.0.CO;2.

Raynaud, L., and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quart. J. Roy. Meteor. Soc.*, **143**, 3037–3047, doi:https://doi.org/10.1002/qj.3159.

Raynaud, L., I. Pechin, P. Arbogast, L. Rottner, and M. Destouches, 2019: Object-based verification metrics applied to the evaluation and weighting of convective-scale precipitation forecasts. *Quart. J. Roy. Meteor. Soc.*, **145**, 1992–2008, doi:10.1002/qj.3540.

Roberts, B., I. Jirak, A. Clark, S. Weiss, and J. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, doi:10.1175/BAMS-D-18-0041.1.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *LNCS*, **9351**, 234–241, doi:10.1007/978-3-319-24574-4_28.

Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. doi:10.1007/978-3-540-77655-0_16.

Schumacher, R., A. Hill, M. Klein, J. Nelson, M. Erickson, S. Trojniak, and G. Herman, 2021: From random forests to flood forecasts: A research to operations success story. *Bull. Amer. Meteor. Soc.*, **102**, 1742–1755, doi:10.1175/BAMS-D-20-0186.1.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Mon. Wea. Rev.*, **139**, 976–991, doi:10.1175/2010MWR3425.1.

Sobash, R., G. Romine, and C. 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, doi:10.1175/WAF-D-20-0036.1.

Sobash, R., C. Schwartz, G. Romine, K. Fossell, and M. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, doi:10.1175/WAF-D-15-0138.1.

Trapp, R., S. Tessendorf, E. Godfrey, and H. Brooks, 2005: Tornadoes from squall lines and bow echoes. part I: Climatological distribution. *Wea. Forecasting*, **20**, 23–34, doi:10.1175/WAF-835.1.