Improving Prediction of Heavy Rainfall in the Mediterranean with Neural Networks Using Both Observation and Numerical Weather Prediction Data

Killian Pujol, Roberta Baggio, Dominique Lambert, Jean-François Muzy, Jean-Baptiste Filippi, Florian Pantillon

Improving Prediction of Heavy Rainfall in the Mediterranean with Neural Networks Using Both Observation and Numerical Weather Prediction Data

KILLIAN PUJOL,[a] ROBERTA BAGGIO,[b] DOMINIQUE LAMBERT,[a] JEAN-FRANÇOIS MUZY,[b] JEAN-BAPTISTE FILIPPI,[b] AND FLORIAN PANTILLON[a]

[a] *Laboratoire d'Aérologie, Université de Toulouse, CNRS, IRD, Toulouse, France*
[b] *Laboratoire Sciences Pour L'Environnement (SPE), Université de Corse, CNRS, Corte, France*

(Manuscript received 18 April 2025, in final form 1 December 2025, accepted 17 February 2026)

ABSTRACT: Forecasting heavy precipitation events (HPEs) in the Mediterranean is crucial but challenging due to the complexity of the processes involved. In this context, artificial intelligence methods have recently proven to be competitive with state-of-the-art numerical weather prediction (NWP). This work focuses on improving the prediction of the occurrence of HPEs over periods from 1 to 24 h using neural network (NN) models. The proposed method uses both ground station observations and data from Météo France's AROME and ARPEGE NWP models, on two regions with oceanic and Mediterranean climates for the period 2016–18. The verification metric is the Peirce skill score. Results show that the NN model using only observations or NWP data performs better for shorter and longer rainfall accumulation periods, respectively. In contrast, a hybrid method combining both observations and NWP data offers the best performance and remains stable with the rainfall accumulation period. The hybrid method also improves the performance in predicting increasingly intense rainfall, from the 5% to the 0.1% rarest events. The choice of the loss function is found to be an important aspect of this work, where only balanced loss functions provide results insensitive to rare event frequency. Finally, the hybrid method is particularly well suited for the prediction of HPEs in the Mediterranean climate, especially during the fall season, the period during which most HPEs occur.

SIGNIFICANCE STATEMENT: Heavy precipitation events can be deadly, particularly in the Mediterranean, and are difficult to predict. Traditionally, numerical weather prediction models have been used for forecasting such events but have shown limitations. This study explores the use of artificial intelligence methods to forecast the occurrence of heavy precipitation, using numerical weather prediction and observational data. The combination of both types of data, the so-called hybrid method, offers the best results and is stable with the time period and intensity of precipitation events. The hybrid method is found to be particularly well suited in the Mediterranean climate, especially during the fall season, the period during which most heavy precipitation events occur.

KEYWORDS: Mediterranean Sea; Extreme events; Numerical weather prediction/forecasting; Probabilistic Quantitative Precipitation Forecasting (PQPF); Classification; Deep learning

1. Introduction

The Mediterranean is characterized by hot, dry summers and mild, wet winters. It presents significant meteorological variability due to interactions between tropical and polar air masses, together with a complex topography characterized by high mountain ranges. These peculiarities can lead to extreme weather events such as intense rainfall and storms. Heavy precipitation events (HPEs) in particular are a cause of major concern as they can generate flash floods that significantly impact the economy and society and often result in human fatalities. The list of catastrophic cases is long, from historic floods caused by an Aiguat in Catalonia in 1940 (Pardé 1941) to cyclone Daniel that brought heavy rainfall in Greece and Libya (Flaounas et al. 2025) or the more recent flooding in the Valencia region of Spain (Mandement and Kreitz 2025). As the Mediterranean is densely populated, understanding and forecasting the meteorological dynamics of these events is crucial, even more so with a changing climate (IPCC 2023).

Knowledge about the processes responsible for the occurrence of HPEs in the Mediterranean has improved over the years, especially with the HyMeX research program conducted between 2010 and 2020 (Ducrocq et al. 2014). After summer, the warm Mediterranean fills the lower atmosphere with moisture and heat. The air near the sea surface becomes conditionally unstable, i.e., it will ascend vertically if triggered by a lifting mechanism. The most common mechanism is lifting by orography: If a flow of moist and conditionally unstable air faces a mountain range, the air may start to ascend and trigger a thunderstorm. This scenario is common in the Mediterranean, as it is surrounded by steep orography, e.g., the Pyrenees between Spain and France, the Atlas Mountains throughout Morocco, Algeria, and Tunisia, the Taurus Mountains in Turkey, the Alps, as well as mountainous islands like Corsica or Crete. This broad orography in proximity to a warm Mediterranean is an important ingredient in the formation of rainfall.

Moreover, the triggering of HPEs is partly determined by the synoptic-scale context that guides the moist airflow toward

Corresponding author: Killian Pujol, killian.pujol--nicolas@utoulouse.fr

DOI: 10.1175/AIES-D-25-0031.1 e250031

the orography. For instance, Nuissier et al. (2011) and Ricard et al. (2012) showed the synoptic environment before and during HPEs in the French Mediterranean: a midtropospheric trough deepening over the Bay of Biscay, whose orientation and intensity influence the orientation of moisture fluxes, and thus the location of HPEs. This is due to the influence of upper-tropospheric instabilities on surface disturbances, as shown by previous studies in the Mediterranean (Argence et al. 2006, 2009). Depending on the synoptic-scale context, the moisture origin of HPEs can differ as well as the quantity of precipitation. Duffourg and Ducrocq (2011) further showed that both local and remote sources of moisture can feed HPEs in cyclonic conditions, with moisture coming from the Atlantic Ocean or tropical Africa, and the added moisture greatly influences the rainfall intensity.

Although the necessary conditions to trigger HPEs in the Mediterranean are well known, challenges persist in fully understanding their location, intensity, and duration. Bresson et al. (2009) studied different lifting effects and their impact on the precise time and location of convective rainfall. Although orographic lifting is the main mechanism, vertical updraft can also be triggered by low-level convergence and cold pools—cold air formed during a thunderstorm by rain evaporation—and the location of the events can vary depending on the lifting mechanism. The authors also emphasized that the location depends on several features, such as the speed and instability of the upstream flow. The duration and stationarity of HPEs are other important aspects: the longer the event lasts, the more precipitation might accumulate at a specific location. Ducrocq et al. (2008) studied three cases of stationary HPEs and demonstrated that stationarity depends on the ability of atmospheric and topographic features, through blocking and convergence processes, to focus moist and unstable flow over a specific location for an extended period.

Forecasting HPEs traditionally relies on numerical weather prediction (NWP) models. These models simulate the evolution of the atmosphere by solving the equations of fluid mechanics and thermodynamics. The initial state is formed by gathering multiple observations from several sources, such as weather stations, radiosondes, precipitation radars, and satellites, through a data assimilation process to determine the complete state of the atmosphere. NWP models fall into two types: global NWP models that simulate atmospheric circulation over the entire Earth, with horizontal resolution of about 10 km or more, depending on the model, and local NWP models that simulate the weather in smaller, country-sized areas, with horizontal resolution down to about 1 km. At such resolution, the representation of the atmosphere becomes fine enough to solve deep convection without using parameterization schemes and allows better representation of orography. In general, high-resolution NWP models have been a great advance for HPE forecasts in the Mediterranean, along with improved parameterization schemes and data assimilation (see Khodayar et al. 2021, for a review of outcomes of the HyMeX program).

However, despite these advances for HPEs in the Mediterranean, NWP models still face limitations in predicting their precise characteristics. A first source of uncertainties arises from the initial conditions, as the observations are sparse, especially over the sea, and can contain flaws. The uncertainties can then be propagated during simulations (Scheffknecht et al. 2016). Moreover, even if NWP model resolution has increased over time, small-scale processes such as turbulence cannot be explicitly resolved. In addition, other small-scale processes, such as air–sea interaction or cloud microphysics (Hally et al. 2014), are not yet well represented in NWP model physics. Finally, even though available computational capacity has increased over the years, the computational cost increases exponentially with the resolution of NWP models (Bauer et al. 2015), such that the use of higher resolutions remains a technological challenge.

Recently, the development of artificial intelligence (AI) has had a major impact on the world of science. Among AI approaches, machine learning and deep learning are research fields that focus on developing statistical algorithms that are capable of learning a specific task from data without explicit instructions to perform that task (Murphy 2022). In particular, neural networks (NNs) are highly flexible learning algorithms that have proven to be successful on a wide variety of tasks (Goodfellow et al. 2016). As NNs are a very active field of research, new architectures and paradigms continue to emerge, yet even more classical models, such as convolutional and recurrent networks, are particularly well suited for learning complex spatiotemporal patterns from data. The use of these algorithms in science has grown alongside the increasing availability of data across many areas of research. As weather data have been extensively collected in the past, it is not surprising that AI techniques have been increasingly applied to weather forecasts and that many data-driven meteorological models for weather forecasting have been developed in the last few years (Reichstein et al. 2019; Schultz et al. 2021). The performance of these approaches has risen rapidly, even in a task as challenging as medium-range global weather forecasts, where several high-performing models based on cutting-edge NN architectures, such as vision transformers (Bi et al. 2023; Pathak et al. 2022) and graph NNs (Lam et al. 2023), have demonstrated skill comparable to that of operational NWP models, at least for some variables and metrics (Ben Bouallègue et al. 2024).

Among the earliest approaches exploring the use of deep learning for rainfall prediction, many focused on nowcasting, that is, the forecasting of weather evolution in the very short term, typically up to 6 h in the future. Most of these approaches use only observational data as input, especially radar data (e.g., Shi et al. 2015; Agrawal et al. 2019; Ayzel et al. 2020). By combining convolutional and recurrent NNs for images and time series of observed radar echoes, respectively, these studies outperform NWP forecasts and classical time extrapolation methods, especially for very short lead times of less than 1 h. However, for longer lead times and for increasingly intense rainfall, the model performance drops significantly. Alternatively, a dense network of weather stations can also be used as a reliable data source for rainfall prediction (Wang et al. 2017). For instance, Pirone et al. (2023) proposed a method to forecast the precipitation rate up to 3 h at a given station by using a fully connected NN with present and past observations from the same station as well as neighboring stations as input. This method performs well for the forecast of light to

moderate rain but faces a drop in performance with increasing lead times and for heavy rainfall coming from convective motions. As weather stations also record meteorological variables other than precipitation, similar approaches that consider neighboring stations to leverage spatiotemporal dependencies have been used to predict temperature or wind speed from observed fields (e.g., Baïle and Muzy 2023; Baggio and Muzy 2024). Some studies combine different sources of observational data. For instance, the latest developments of the Google DeepMind MetNet model, MetNet-2 (Espeholt et al. 2022) and MetNet-3 (Andrychowicz et al. 2023), take as input observational data from radar and weather stations as well as satellite and data assimilation products from a high-resolution NWP model. The prediction of precipitation issued by MetNet-3 drastically outperforms the NWP model for short lead times up to about 6 h, but this performance gap progressively reduces for longer lead times. Moreover, the model performance tends to be lower for higher precipitation intensities.

As NNs are highly effective at learning complex nonlinear relationships and can integrate diverse types of data, they are natural candidates for the postprocessing of NWP forecasts. Such forecasts are known to suffer from systematic biases and errors, particularly for variables near the surface (Haiden et al. 2015). To address these issues, operational centers make use of postprocessing methods to improve weather forecasts. Recently, machine learning (ML) methods have been increasingly incorporated into postprocessing methods, and in particular, NNs have been proven particularly effective (Vannitsem et al. 2021). With respect to more traditional postprocessing methods, such as model output statistics (MOS), which rely on relatively simple forms of regression, NNs can leverage heterogeneous data sources and extract more complex patterns, making them a very promising tool for postprocessing. For example, Rasp and Lerch (2018) compare several postprocessing techniques on global ensemble forecasts to predict 2-m temperature at weather stations in Germany, using several forecasted fields. They found that the NN postprocessing provides the best results for most stations. In Schulz and Lerch (2022), the authors make a systematic comparison between several postprocessing methods for wind gusts and find that the methods based on NNs are the most performing overall. Other studies have showed that using NNs for NWP model postprocessing is relevant for precipitation forecasting. For instance, Frnda et al. (2022) and Liu et al. (2023) improved global forecasts up to a lead time of 3 days, showing better results than with other postprocessing methods and reducing the loss of predictability with the lead time of the NWP.

As discussed above, NNs which leverage observational data for rainfall predictions perform well for short lead times, but the prediction performance drops rapidly when considering longer forecast horizons. Rather than fully replacing NWP models with deep learning, an alternative approach, motivated by the success of NNs for postprocessing NWPs, is to integrate NWP outputs into the NN input alongside observations. Such an approach is defined here as a hybrid. Among the few studies that have applied this approach, Espeholt et al. (2022) used MetNet-2 for rainfall prediction for lead times up to 12 h. Using only observations or postprocessing performs well at either short or long lead times only. In contrast, a hybrid approach using both observed and NWP data gives the best performance for all lead times.

This work contributes to the prediction of HPEs by introducing such a hybrid approach. Specifically, the proposed model forecasts rainfall threshold exceedance as a binary classification outcome, capturing events of varying severity across different rainfall accumulation periods. The proposed hybrid approach leverages both observational data and NWP outputs to produce stable forecasts with lead times of up to 24 h. The focus is on HPEs, which both NWP and data-driven models struggle to forecast. In the case of data-driven models, the challenge for HPEs is to learn to forecast events that are, by definition, rarely represented in the training dataset. To address this imbalance, several studies have proposed using weighted loss functions, functions that compare model predictions with observed outcomes during training, to give more importance to underrepresented events (Shi et al. 2017; Leinonen et al. 2022; Liu et al. 2023). In this paper, a custom loss function based on the Peirce skill score (PSS), an equitable evaluation metric, is presented and compared with the binary cross-entropy loss function and its weighted variant.

This article is organized as follows. Section 2 presents the data used for this work, the performed forecasting tasks, and their related NN models. The results are presented in section 3, and the conclusions of this work are provided in section 4.

2. Data and methods

a. *MeteoNet*

This study uses data from MeteoNet (Larvor et al. 2020), an open-source dataset provided by Météo France. This dataset includes many different types of weather data, both observations and NWP forecasts, that are available over two regions of France and over the 2016–18 three-year period. In this study, observation data from ground weather stations and NWP model forecasts are used. Figure 1 shows the extent of NWP data and the locations of the ground stations. The two available regions are northwest France, with 93 stations, and southeast France, with 159 stations. The number of ground stations has been reduced in this study, compared to the original dataset, due to the removal of missing values and meteorological fields. The northwest and southeast regions are here referred to as the "oceanic" region and "Mediterranean," respectively, in accordance with the Köppen climate classification (Beck et al. 2018). Figure 1 highlights the difference between the two regions in terms of orography, with the Mediterranean featuring high mountain ranges. Due to differences in climate and orography, the two regions are considered separately in the remainder of this document. Within each region, however, the weather stations are subject to similar meteorological conditions and are therefore treated consistently in the analysis, for example, when selecting the thresholds used to define extreme events.

Observed meteorological fields at ground stations are provided with a time step of 6 min. The available NWP data include deterministic forecasts from two different Météo France operational models, ARPEGE (Courtier et al. 1991) and AROME

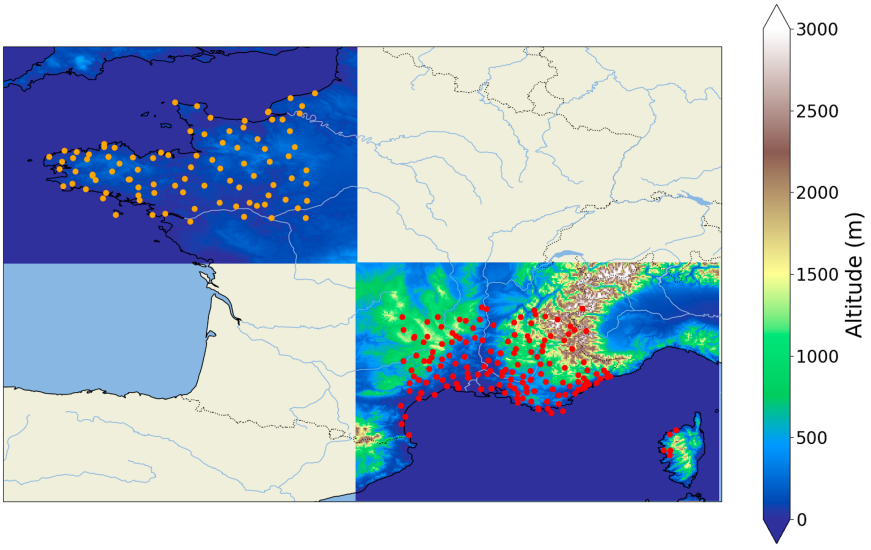

FIG. 1. Geographical extent of the MeteoNet database, with ground station localization (orange: northwest; red: southeast) and topography in shades of color.

(Seity et al. 2011). ARPEGE is a global model, with a horizontal resolution of 0.1° (10 km). At this resolution, deep convection is not explicitly solved and is therefore parameterized. AROME is a regional model that forecasts weather over France with a higher spatial resolution of 0.025° (2.5 km), which implies that AROME is a convection-permitting model. The data available in MeteoNet come from the operational model versions from 2016 to 2018 and are provided over the oceanic and Mediterranean regions, as portrayed in Fig. 1. Operational model versions are typically updated every year, but the relatively short period of 3 years ensures a sufficiently homogeneous configuration, in particular a stable resolution during that period. For each day, and up to 24 h ahead, only the forecast from the initialization at 0000 UTC of both models is provided. Near-surface fields from AROME are given with a time step of 1 h. ARPEGE forecasts are provided on seven vertical levels, with the same extent as the surface fields, with a time step of 1 h between 0000 and 1200 UTC and 3 h between 1200 and 2400 UTC. To simplify the data preprocessing, this three-dimensional ARPEGE data are selected with a constant time step of 3 h. In this study, all meteorological fields available for these three types of data are used, that is, ground station observation data, surface forecast fields from AROME, and three-dimensional forecast data from ARPEGE. The summary of the data used in this study is presented in Table 1.

b. *Forecasting task*

The forecasting task in this study is defined as follows: At each ground station, the objective is to determine whether the accumulated precipitation between t and $t + \Delta t$ [hereafter referred to as $R_{\Delta t}(t)$] will be lower or greater than a given rainfall threshold $Q_{\Delta t}$. The studied values of Δt, defined as the accumulation window size (hereafter referred to as the window size), are 1, 2, 3, 6, 12, and 24 h. The $Q_{\Delta t}$ values depend on Δt and the region and are presented in section 2c. As this task amounts to forecasting two classes of rain—below or above $Q_{\Delta t}$— let us define $Y_{t+\Delta t}$ as the corresponding rainfall class such that

$$\begin{cases} Y_{t+\Delta t} = 0 \text{ if } R_{\Delta t}(t) < Q_{\Delta t} \\ Y_{t+\Delta t} = 1 \text{ if } R_{\Delta t}(t) \geq Q_{\Delta t} \end{cases}. \tag{1}$$

A machine learning approach for a classification task predicts the probability $p_{t+\Delta t}$ that $Y_{t+\Delta t} = 1$. A critical probability P_C must then be considered, such that, with $\hat{Y}_{t+\Delta t}$, the predicted rain class:

$$\begin{cases} \hat{Y}_{t+\Delta t} = 0 \text{ if } p_{t+\Delta t} < P_C \\ \hat{Y}_{t+\Delta t} = 1 \text{ if } p_{t+\Delta t} \geq P_C \end{cases}. \tag{2}$$

As explained in Jolliffe and Stephenson (2003) and Wilks (2019), an optimal value $P_C^\star$ can be selected for P_C to maximize the score used to evaluate prediction quality (further details are provided in section 2e).

The prediction uses $\mathcal{X}$ which represents the NN model inputs. All data shown in Table 1 are used as inputs: Station inputs are defined as the observations at the given station from $t - \Delta t$ to t, with a time step of 6 min, and AROME and ARPEGE inputs are defined as the predictions of each NWP model with 10×10 grid points around the given station from t to $t + \Delta t$, the former with a time step of 1 h, the latter with a time step of 3 h. The learning target—i.e., the data on which the NN model learns—is calculated using the accumulated rainfall between t and $t + \Delta t$ observed at the ground station. More details regarding data management for training are given in section 2d, and the choice of the value of $Q_{\Delta t}$ is discussed in section 2c.

Figure 2 summarizes the forecasting task for different NN model initialization times t_j during a given day. Index $j = 1, \ldots, n$ runs over the number of tasks per day, which is $n = 24/\Delta t$

TABLE 1. Summary of the data used in this study.

Source	Spatial and temporal resolutions	Meteorological fields
Station observations	Spatial resolution: pointwise Time resolution: 6 min	Wind direction (°) Wind speed (m s^{-1}) Precipitation (mm) Humidity (%) Dewpoint (K) Temperature (K) Mean sea level pressure (Pa)
NWP AROME	Spatial resolution: 0.025° Time resolution: 1 h	2-m temperature (K) 2-m dewpoint (K) 2-m relative humidity (%) 10-m wind speed (m s^{-1}) 10-m wind direction (°) 10-m u and v wind components (m s^{-1}) Mean sea level pressure (Pa) Total precipitation since initialization (mm)
NWP ARPEGE isobar levels	Spatial resolution: 0.1° with seven isobar levels (1000, 950, 925, 850, 700, 600, and 500 hPa) Time resolution: 3 h	Temperature (K) Wet-bulb potential temperature (K) Relative humidity (%) Wind speed (m s^{-1}) Wind direction (°) u and v wind components (m s^{-1}) Vertical velocity (Pa s^{-1}) Geopotential (m^2 s^{-2})
NWP ARPEGE height levels	Spatial resolution: 0.1° with seven vertical levels (20, 100, 500, 875, 1375, 2000, and 3000 m) Time resolution: 3 h	Pressure (Pa)

for $\Delta t \geq 3$ h and $n = 8$ for $\Delta t < 3$ h to match the ARPEGE time step of 3 h and avoid overlapping experiments. For the NN model initialization time t_j, $R_{\Delta t}(t_j)$ (green arrows) is predicted using either station data observed between $t_j - \Delta t$ and t_j (blue arrows), NWP model data predicted between t_j and $t_j + \Delta t$ (dashed red arrows), or both. Depending on the accumulation window size Δt value (hereafter referred to as the window size), given that MeteoNet provides NWP forecast data only from the 0000 UTC initialization run (top red arrow), the NN model may use NWP data with a greater or lesser temporal distance from the NWP model initialization. This fact should be kept in mind when discussing the obtained results.

It is important to note that the number of experiments varies depending on the value of Δt, as Δt also represents, for $\Delta t \geq 3$ h, the temporal distance between two NN model initializations. One can estimate the number of experiments N_{exp} such that $N_{\text{exp}} = n \times (365 \times 2 + 364) \times N_{\text{station}}$, with N_{station} the number of weather stations considered, $365 \times 2 + 364$ representing the number of days during the three studied years (2016 being a leap year). Hence, as the proportion of experiments being discarded due to missing values is around

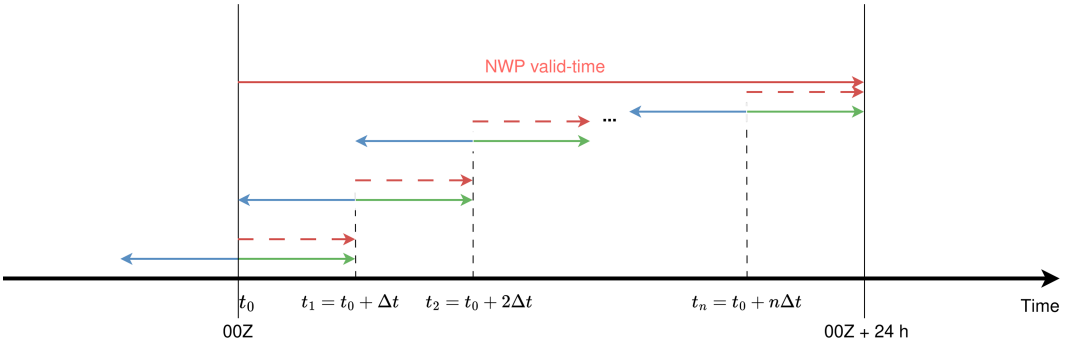

FIG. 2. Diagram of forecasting tasks during a day depending on the value of Δt. At each t_j, a prediction is made by the NN model. Blue arrows represent the time period of the observed data used for the NN model prediction. Dashed red arrows represent the time period of the NWP models predictions data used for the NN model prediction. The red arrow represents the time period of NWP models predictions from their initialization. Green arrows represent the time period of Δt.

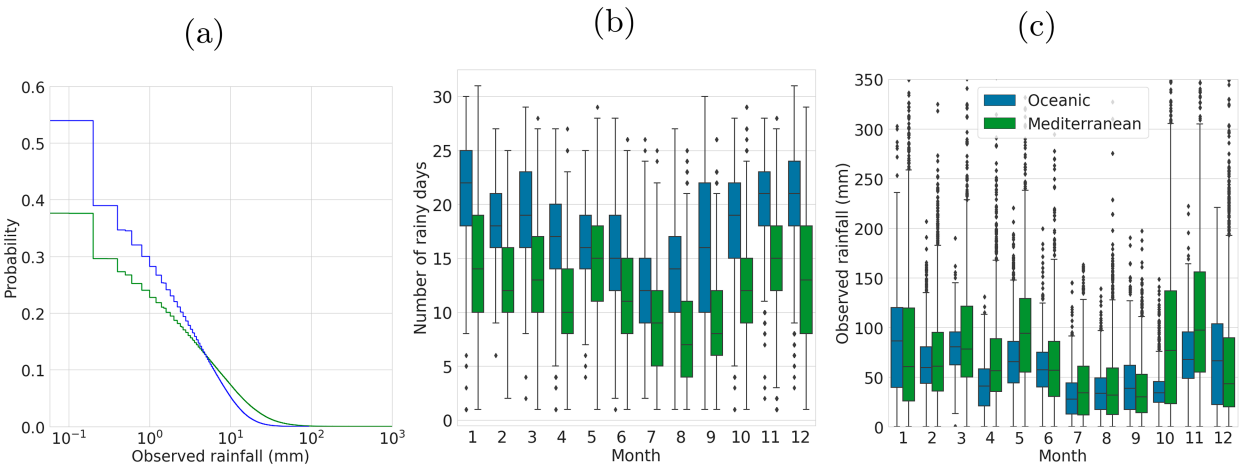

FIG. 3. (a) CCDF of the climatology for $\Delta t = 24$ h. (b) Boxplot of the number of rainy days by month and region. (c) Boxplot of monthly cumulated rainfall by month and region. In blue: oceanic region; in green: Mediterranean.

10% for each experiment, in the Mediterranean, the number of experiments with $\Delta t = 24$ h and $\Delta t = 3$ h is, respectively, $N_{\text{exp}} \approx 150\,000$ and $N_{\text{exp}} \approx 1\,200\,000$. Hereafter, index $i = 1, \ldots, N_{\text{exp}}$ will be used to address the experiments, and p_i will denote the value $p_{t+\Delta t}$ of the model at index i.

c. *Rainfall threshold values*

For each region and each Δt, if $R_{\Delta t}(t)$ represents the cumulative precipitation observed at some weather station between t and $t + \Delta t$, the "climatological distribution" allows one to define the probability that $R_{\Delta t}$ is above or below a given threshold z. For instance, the so-called complementary cumulative distribution function (CCDF) of the climatology can be defined as the probability of exceedance of a given rainfall value:

$$P_{\Delta t}(z) = \text{Probability}[R_{\Delta t} \geq z] = \frac{\text{Number of events}\{R_{\Delta t}(t) \geq z\}}{N_{\text{exp}}}. \quad (3)$$

Figure 3a summarizes the differences between the two regions by illustrating the climatology CCDF for $\Delta t = 24$ h. This figure highlights the fact that low to moderate rain is more likely to occur in the oceanic region, while heavier rainfall is more frequent in the Mediterranean. Note that the value at $z < 10^{-1}$ directly indicates the probability of observing a rainy day. This result can also be seen in Fig. 3b that shows the number of rainy days (number of days for which the recorded daily precipitation accumulation was above 0 mm) for each month by region: One is more likely to observe rain in the oceanic region than in the Mediterranean. However, even though more rainy days are found in the oceanic region, Fig. 3c reveals that the monthly accumulated rainfall is generally higher in the Mediterranean. This is particularly true during fall (especially October and November) which is when most HPEs occur in the Mediterranean (Ricard et al. 2012).

Reference thresholds $Q_{\Delta t}$ for each region can be defined simply from $P_{\Delta t}(z)$ as the values of z associated with a given set of probability levels s. These are the so-called "quantiles." By definition, QL, the quantile of level L, is the value z such that

$$P_{\Delta t}(z) = s \text{ with } s = 1 - \frac{L}{100}. \quad (4)$$

For example, the 95th quantile, Q95, is associated with $s = 0.05$ and is such that $P_{\Delta t}(\text{Q95}) = 0.05$. Note that the smaller the value of s (i.e., the larger L), the rarer and more severe the considered events become.

Table 2 summarizes the values of the 95th, 99th, 99.5th, and 99.9th quantiles of the climatology for each region and Δt. These quantiles, which correspond to increasingly intense events, serve as the thresholds $Q_{\Delta t}$ used throughout this study to classify cumulative precipitation events into class 0 or 1

TABLE 2. Values of quantiles for each region and each Δt. Values are in millimeters.

	Q95		Q99		Q99.5		Q99.9	
Δt (h)	Oceanic	Mediterranean	Oceanic	Mediterranean	Oceanic	Mediterranean	Oceanic	Mediterranean
1	0.4	0.4	1.8	2.6	2.8	4.0	5.5	9.1
2	0.8	0.8	3.4	5.0	4.9	7.6	9.2	15.7
3	1.4	1.4	4.8	7.2	6.8	10.8	12.3	21.3
6	2.8	3.3	8.4	13.1	11.2	18.6	18.7	34.7
12	5.6	7.2	13.5	22.3	17.4	30.4	27.8	55.4
24	10.3	14.1	20.2	35.8	24.8	48.5	38.2	89.6

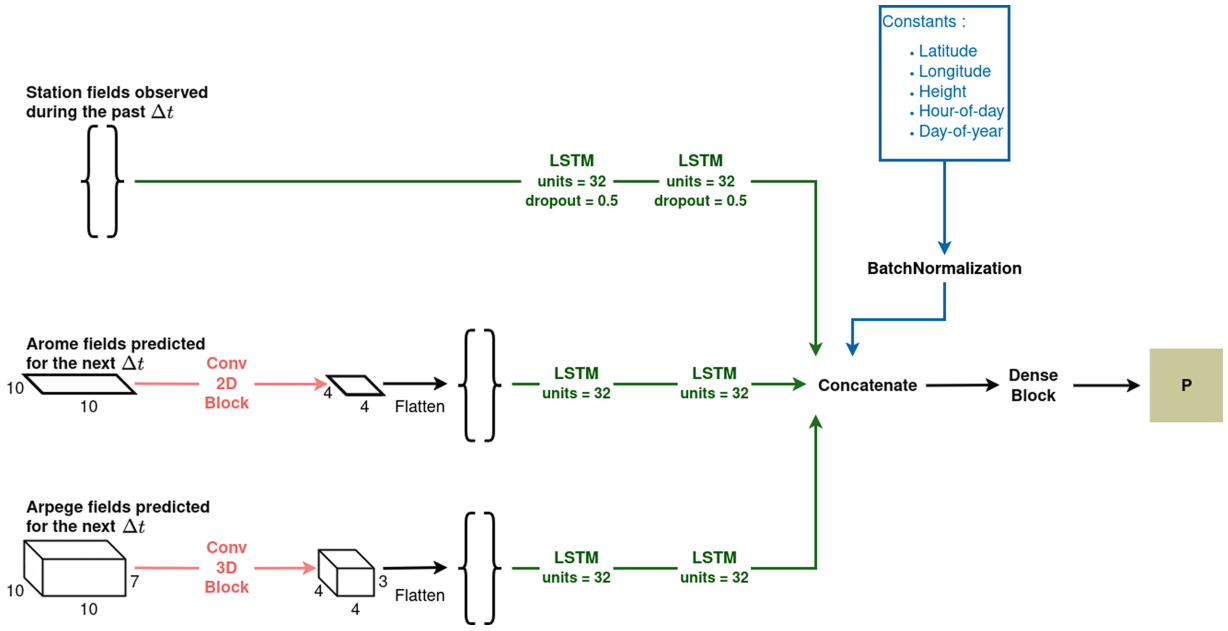

FIG. 4. NN architecture. The top branch represents how data from ground stations are processed. The middle branch shows the processing of the surface fields predicted by AROME, while the bottom branch shows how the three-dimensional field data predicted by ARPEGE are handled. Green arrows indicate temporal processing of data; red arrows represent spatial processing of data. When multiple data sources are used together, the outputs of the individual branches are concatenated and passed through a densely connected NN. The output of this NN architecture is the probability p of occurrence of class 1.

[i.e., $Y = 0$ or $Y = 1$, as defined in Eq. (1)]. By definition, the relative frequency of the class 1 event ($Y = 1$) is s; namely, it will be $s = 0.05$ when considering $Q_{\Delta t}$ = Q95, $s = 0.01$ for $Q_{\Delta t}$ = Q99, and so on.

Therefore, the classification task is performed for a highly unbalanced dataset, and the approach to this issue is discussed in sections 2e and 2f.

d. *Neural network to handle hybrid data*

Figure 4 illustrates the NN architecture used in this study and highlights how the different data sources are integrated by the NN model during training. This study uses the Python libraries TensorFlow (Abadi et al. 2015) and Keras (Chollet et al. 2015) to build the NNs. This architecture answers one of the following prediction tasks, with $\mathcal{M}$ denoting the NN model output:

$$\begin{cases} p_{t+\Delta t} = \mathcal{M}(\mathcal{X}^{\text{Station}}_{t,t-\Delta t}) \\ p_{t+\Delta t} = \mathcal{M}(\mathcal{X}^{\text{AROME}}_{t,t+\Delta t}) \\ p_{t+\Delta t} = \mathcal{M}(\mathcal{X}^{\text{ARPEGE}}_{t,t+\Delta t}) \\ p_{t+\Delta t} = \mathcal{M}(\mathcal{X}^{\text{Hybrid}}_{t-\Delta t,t+\Delta t}) \end{cases}, \tag{5}$$

with $\mathcal{X}^{\text{Station}}_{t,t-\Delta t}$ representing the seven recorded meteorological fields at the station during the past Δt hours, $\mathcal{X}^{\text{AROME}}_{t,t+\Delta t}$ representing the eight weather fields forecasted by AROME on 10×10 grid points around the ground station during the next Δt hours, $\mathcal{X}^{\text{ARPEGE}}_{t,t+\Delta t}$ representing the nine fields forecasted by ARPEGE on 10×10 grid points around the ground station with seven vertical levels during the next Δt hours, and $\mathcal{X}^{\text{Hybrid}}_{t-\Delta t,t+\Delta t} = \{\mathcal{X}^{\text{Station}}_{t,t-\Delta t}, \mathcal{X}^{\text{AROME}}_{t,t+\Delta t}, \mathcal{X}^{\text{ARPEGE}}_{t,t+\Delta t}\}$, which is defined as the hybrid dataset.

Given the time steps of respectively 6 min, 1 h, and 3 h, $\mathcal{X}^{\text{Station}}_{t,t-\Delta t}$ has shape $(10 \times \Delta t, 7)$, $\mathcal{X}^{\text{AROME}}_{t,t+\Delta t}$ has shape $(\Delta t, 10, 10, 8)$ (AROME forecasts at time t being excluded during data preprocessing), and $\mathcal{X}^{\text{ARPEGE}}_{t,t+\Delta t}$ has shape $(\Delta t/3 + 1, 10, 10, 7, 9)$ for $\Delta t \geq 3$ h and $(1, 10, 10, 7, 9)$ for $\Delta t < 3$ h.

Long short-term memory (LSTM) NNs are used to extract the temporal information from the feature data. AROME and ARPEGE data, which consist of 2D and 3D representations of atmospheric variables, respectively, are first processed by convolutional blocks (Conv2D and Conv3D blocks), which are displayed in Fig. 5, to extract spatial information before being passed to an LSTM. In alignment with the assigned forecasting task, the output layer is a densely connected NN—as defined by TensorFlow—with a Sigmoid activation function, whose output $p_{t+\Delta t}$ represents the probability of class 1 occurrence. A series of experiments is conducted in which each type of input data (Station, AROME, and ARPEGE) is considered and analyzed separately. Additionally, the combination of each input data in a hybrid dataset is considered, in which case the respective NN structures of each dataset are concatenated together. Finally, ground station spatial and temporal information data (latitude, longitude, and height of the ground station as well as hour of the day and day of the year) are concatenated with each structure before passing through a dense block which is displayed in Fig. 5. Hereafter, the NN model trained with

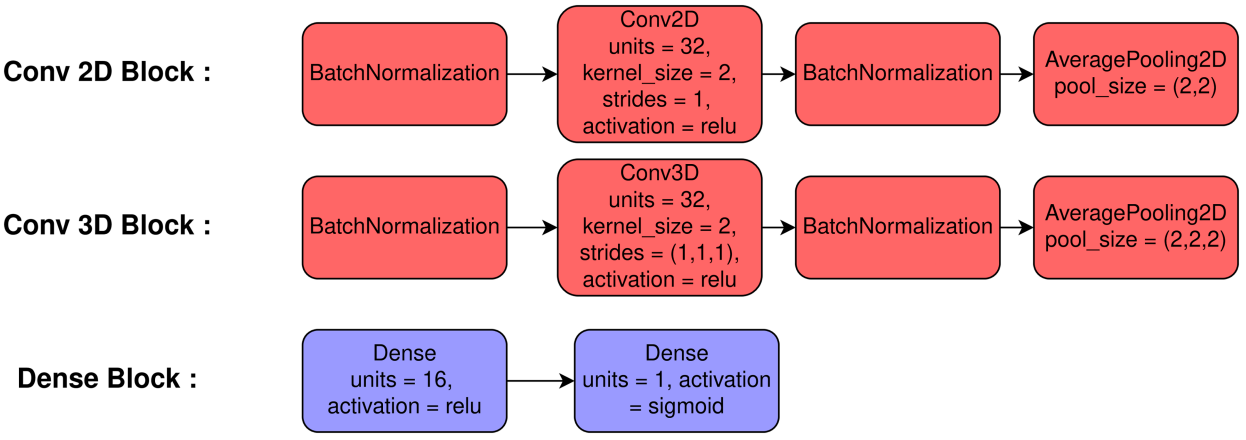

FIG. 5. Description of Conv2D, Conv3D, and dense blocks with their associated hyperparameters.

either the Station, AROME, ARPEGE, or Hybrid dataset will be called, respectively, NN-Station, NN-AROME, NN-ARPEGE, and NN-Hybrid.

Data are randomly split into training, validation, and test datasets, with respective ratios of 80%, 15%, and 5%, ensuring that separate days are used when building the three datasets to avoid biases in the predictions. Note that another split was tested by separating full months between datasets and gave similar results. However, given the low quantity of data, the split per day was preferred. All models were trained on CPUs, which was sufficient given the number of parameters to be optimized and the amount of data. The main training hyperparameters used in this study are displayed in Table 3. Note that no hyperparameter tuning has been performed for this work.

e. *Evaluation scores*

In this study, as the main benchmark is the raw AROME deterministic forecast, predictions are treated as deterministic. This means that model evaluation is based directly on binary outcomes $\hat{Y}_i$ corresponding to model outputs p_i by means of Eq. (2), rather than on the values of p_i themselves. These scores are derived from the components of the confusion matrix, which is a 2×2 table comparing predicted outcomes $\hat{Y}_i$ and actual outcomes Y_i using four key components: true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Considering this, the problem of evaluating binary forecast–observation pairs $(\hat{Y}_i, Y_i)$ is inherently a three-dimensional problem. When using a scalar score for forecast evaluation, some information is inevitably lost in the process. Since the classification task considered here concerns the prediction of rare events, this study argues that forecast performance should be evaluated using an equitable skill score. Indeed, following Gandin and Murphy (1992), equitable metrics not only give equal value to random and constant forecasts but also give greater weight to correct forecasts of rare events, which discourages artificial bias toward the more common event. Taking this into consideration, this study uses the Peirce skill score (Peirce 1884) (also called true skill statistic or Hanssen–Kuipers discriminant) to evaluate HPE prediction. This score is the only equitable metric for binary classification which strictly follows the equitability criteria introduced in Gandin and Murphy (1992) and is independent of the class relative frequency s and thus particularly suitable to evaluate predictions of rare events (Woodcock 1976; Rodwell 2011; Ebert and Milne 2022). PSS is positively oriented, and its possible values range from -1 to 1, where 0 indicates no skill and 1 indicates perfect skill. It is defined as the difference between the hit rate $H = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and the false alarm rate $F = \mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$:

$$\mathrm{PSS} = H - F = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) - \mathrm{FP}/(\mathrm{TN} + \mathrm{FP}). \quad (6)$$

The hit rate H, also known as recall, is often used together with precision, defined as $\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$ to evaluate classification performance on unbalanced datasets. Both precision and recall values range from 0 to 1, where 0 indicates no skill and 1 indicates perfect skill. Precision measures the likelihood that a predicted class 1 event is actually observed, while recall quantifies the probability that an observed class 1 event was correctly forecast. The comparison between these two metrics illustrates the trade-off between correctly identifying events and minimizing missed detections.

For completeness, results for two additional metrics are considered, the critical success index (CSI), which is also frequently used in cases of class imbalance, and the Heidke skill score (Jolliffe and Stephenson 2003; Wilks 2019). The definition of these scores, along with the corresponding model results, is reported in appendix B.

The scores presented here and displayed in section 3 are calculated on both the validation and test datasets. As no hyperparameter tuning has been performed in this study, the performance

TABLE 3. NN model hyperparameter.

Parameter	Value
Optimizer	Adam
Learning rate	0.001
Batch size	512
Epoch	512

on both datasets is expected to be similar, as confirmed by the results.

f. Loss function

The choice of the loss function is a crucial aspect of the learning process. The loss function guides the NN by comparing its predictions with the ground truth, and the focus of the NN training depends on the loss function. This study focuses on classification prediction; therefore, the loss functions used are classification losses. A similar approach can be followed through regression losses but shows lower performance, as discussed by Baggio et al. (2025). In this study, a comparison of three different loss functions is performed. The first one, the binary cross entropy (BCE), is the standard loss function used for binary classification tasks, and it is defined as

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N} Y_i \log(p_i) + (1 - Y_i)\log(1 - p_i), \tag{7}$$

where Y_i is the ground truth, p_i is the predicted probability at prediction task index i, and N is the number of events in the considered sample for metric computation.

As this study focuses on rare events, two loss functions that explicitly account for the unbalanced dataset are also tested. These are the weighted BCE (w-BCE) and the Peirce loss (PL). The w-BCE is a variant of BCE in which weights w_1 and w_0 are introduced to handle the class imbalance:

$$\text{w-BCE} = -\frac{1}{N}\sum_{i=1}^{N} w_1 Y_i \log(p_i) + w_0(1 - Y_i)\log(1 - p_i). \tag{8}$$

The value of w_i is determined by the frequency of the corresponding classes, that is $w_1 = 1/s$ and $w_0 = 1/(1 - s)$, s being the mean observed rate of the HPEs.The PL is a custom loss function which is directly derived from the PSS score [(6)]:

$$\mathrm{PL} = \sum_{i=1}^{N} p_i \frac{1 - Y_i}{\sum_{j=1}^{N}(1 - Y_j)} - p_i \frac{Y_i}{\sum_{j=1}^{N} Y_j}. \tag{9}$$

Indeed, by expressing the components of the confusion matrix in terms of forecasts $\hat{Y}_i$ and observations Y_i, H, and F are given by

$$H = \frac{\sum_{i}^{N} \hat{Y}_i Y_i}{\sum_{j}^{N} Y_j} \qquad F = \frac{\sum_{i}^{N} \hat{Y}_i (1 - Y_i)}{\sum_{j}^{N}(1 - Y_j)}, \tag{10}$$

and then Eq. (9) is obtained by setting $\hat{Y}_i \approx p_i$. It should be noted that when using a class-balanced loss function, more weight is assigned to the underrepresented class 1, leading to an expected increase in FPs and a decrease in FNs. This effect is not considered to be problematic, as the societal cost of overforecasting HPEs is considered significantly lower than that of underforecasting such events.

A discussion of the relationship between the "true probability" q_i that $Y_i = 1$ occurs and p_i, the output of the NN model, according to the optimized loss, can be found in appendix A.

3. Results

a. The benefits of the hybrid dataset

The relative performance, in terms of PSS, obtained using NN-Station, NN-AROME, NN-ARPEGE, as well as the NN-Hybrid model is illustrated in Fig. 6. Results based on alternative scores are presented in appendix B. The loss function chosen here is the PL, and further comparisons between loss functions are discussed in section 3b. Performance obtained using a benchmark forecast, given by raw AROME predictions, is also shown. This benchmark is computed by taking the AROME grid point closest to the station and calculating the accumulated rainfall forecast over Δt. Persistence is also displayed as a natural benchmark for rainfall forecasts.

Figure 6a displays the PSS for the 95% quantile threshold exceedance relative to the window size Δt. The benchmark (dashed gray line) shows increasing PSS with Δt, with PSS values between 0.4 and 0.5 for $\Delta t = 1$ h to PSS values around 0.6 for $\Delta t = 24$ h. This means that the benchmark performs better at predicting rainfall threshold exceedance over a longer window size than over a shorter one. For NN-AROME (dashed red line), the behavior is similar to that of the benchmark in the sense that performance improves for longer Δt, and it is significantly better overall: e.g., for $\Delta t = 24$ h, PSS = 0.8 instead of PSS = 0.6 for the benchmark. NN-ARPEGE (dotted green line) shows PSS comparable to NN-AROME, although it performs better for $\Delta t \leq 3$ h, thus providing a more stable performance with Δt overall. In contrast, the NN-Station model (solid blue line) better predicts rainfall threshold exceedance for short Δt than for longer ones. It shows a rapid decrease in PSS with Δt, with PSS values higher than 0.6 for $\Delta t \leq 6$ h—then outperforming the benchmark over this range of Δt values—and decreasing below 0.6 thereafter. Persistence performance portrays a similar behavior to NN-Station with significantly lower values, which makes NN-Station appear to be an improved persistence forecast. The NN-Hybrid model (dotted black line) offers the best overall performance, with the highest PSS values and the most stability with Δt. Indeed, PSS values for short Δt are similar to values for longer Δt, although a slight decrease can be seen for $\Delta t = 24$ h. This emphasizes the complementarity of observed and predicted data. Finally, except with NN-Station and for high values of Δt, the value dispersion between the PSS computed on the validation and test datasets (shading) is overall low, which highlights the relevance of using both datasets for calculating scores.

Figure 6b shows the sensitivity of NN models and benchmark predictions to the time of day at which predictions are made for $\Delta t = 3$ h. When examining the predictions of NN models, the time of day represents their initialization time. However, for the benchmark, it represents the valid time. The benchmark depicts a PSS slightly sensitive to the time of day, with a PSS increase between 0000 and 0300 UTC, followed by

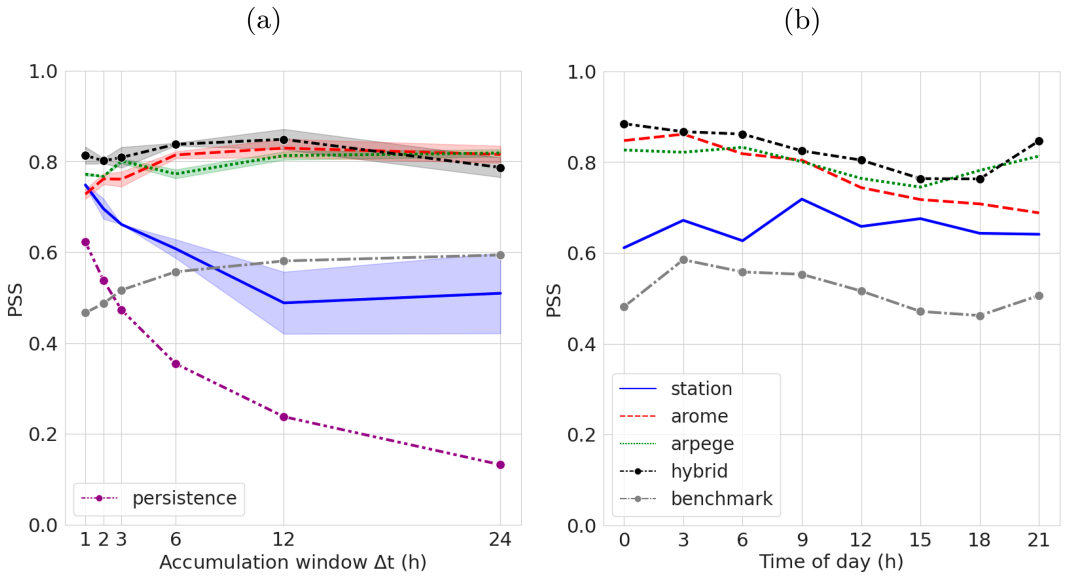

FIG. 6. PSS performance of the NN model trained on different types of input data from the Mediterranean, for the 5% rarest rainfall events (95th quantile) prediction, with the PL as the loss function and $P_C = 0.5$. (a) Sensitivity to the window size Δt. The shading represents the dispersion between the PSS computed on the validation and test datasets. (b) Sensitivity to the time of the day, with $\Delta t = 3$ h, at which NN models make their predictions. In solid blue: NN-Station. In dashed red: NN-AROME. In dashed green: NN-ARPEGE. In dotted black: NN-Hybrid. In dotted and dashed gray: benchmark. In dotted and dashed purple: persistence.

a slow decrease between 0300 and 1800 UTC. Though a similar trend can be seen for NN-AROME, PSS values are much higher overall, as seen in Fig. 6a. NN-ARPEGE provides a slightly decreasing PSS with the time of day, although slower than NN-AROME and with a slight increase for the time of day greater than 15 h. NN-ARPEGE outperforms NN-AROME after 6 h. On the other hand, apart from a few variations, NN-Station shows low sensitivity to the time of day. The differences in PSS sensitivity with the time of day between NN-Station, NN-AROME, and NN-ARPEGE can be explained by the fact that NN-Station uses observed information that is continuously updated, whereas NN-AROME and NN-ARPEGE use data derived solely from the NWPs initialization at 0000 UTC. NN-Hybrid again delivers the best PSS performance. A slight sensitivity to the time of day is observed, with a slowly decreasing PSS followed by an increase at 2100 UTC. This dataset improves the NN model's performance in PSS overall and mitigates the time of day sensitivity. Results for CSI and Heidke skill score (HSS), which are displayed in appendix B, exhibit similar patterns to those discussed here, thus confirming the results for PSS.

b. *About the choice of the loss function*

This subsection discusses how the use of the three loss functions previously defined in section 2f—namely, BCE, w-BCE, and PL—affects the obtained results. In particular, the choice of the loss function significantly impacts the probability distribution of the issued probabilistic forecasts, and therefore performance once the critical probability P_C is fixed. Results are shown in Fig. 7 for $\Delta t = 3$ h and the 95th percentile as the rainfall threshold.

The empirical distribution of HPE outcome probabilities p_i obtained according to the three losses is displayed in Fig. 7a. The BCE (purple bars) produces most predicted probabilities in the range between 0 and 0.1, with few exceeding 0.1. This implies that, with the "natural" choice $P_C = 0.5$, most predicted events would not exceed the rainfall threshold. The w-BCE (pink bars) leads to a more homogeneous distribution of predicted probabilities, thus allowing to predict many more threshold exceedances to be predicted when setting the same threshold $P_C = 0.5$. Finally, the PL (yellow bars) results in a distribution which is more binary, where the large majority of predicted probabilities are either between 0 and 0.1 or between 0.9 and 1, with no predicted probabilities in between. The consequence of such a binary behavior is that the predicted class would be the same for a large range of P_C, which frees up the choice of the critical probability. The differences in behavior between the three loss functions are further illustrated in Fig. 7b, which displays TP, FP, and FN values for $P_C = 0.5$. As previously mentioned, with the BCE, most predicted probabilities remain below P_C, resulting in a predominance of TN (not shown) and FN (red bars), while FP (orange bars) is extremely rare. When using the w-BCE or PL, TP (green bars) increases, meaning that rain events are correctly identified more frequently. However, this also leads to a significant rise in FP. This outcome is expected, as assigning greater importance to the underrepresented class implies that misclassifying a positive event as negative is considered far more costly than the reverse, which aligns with the objectives of this study.

Figures 7c and 7d offer a more in-depth view on each loss function's behavior. In particular, Fig. 7c shows the sensitivity

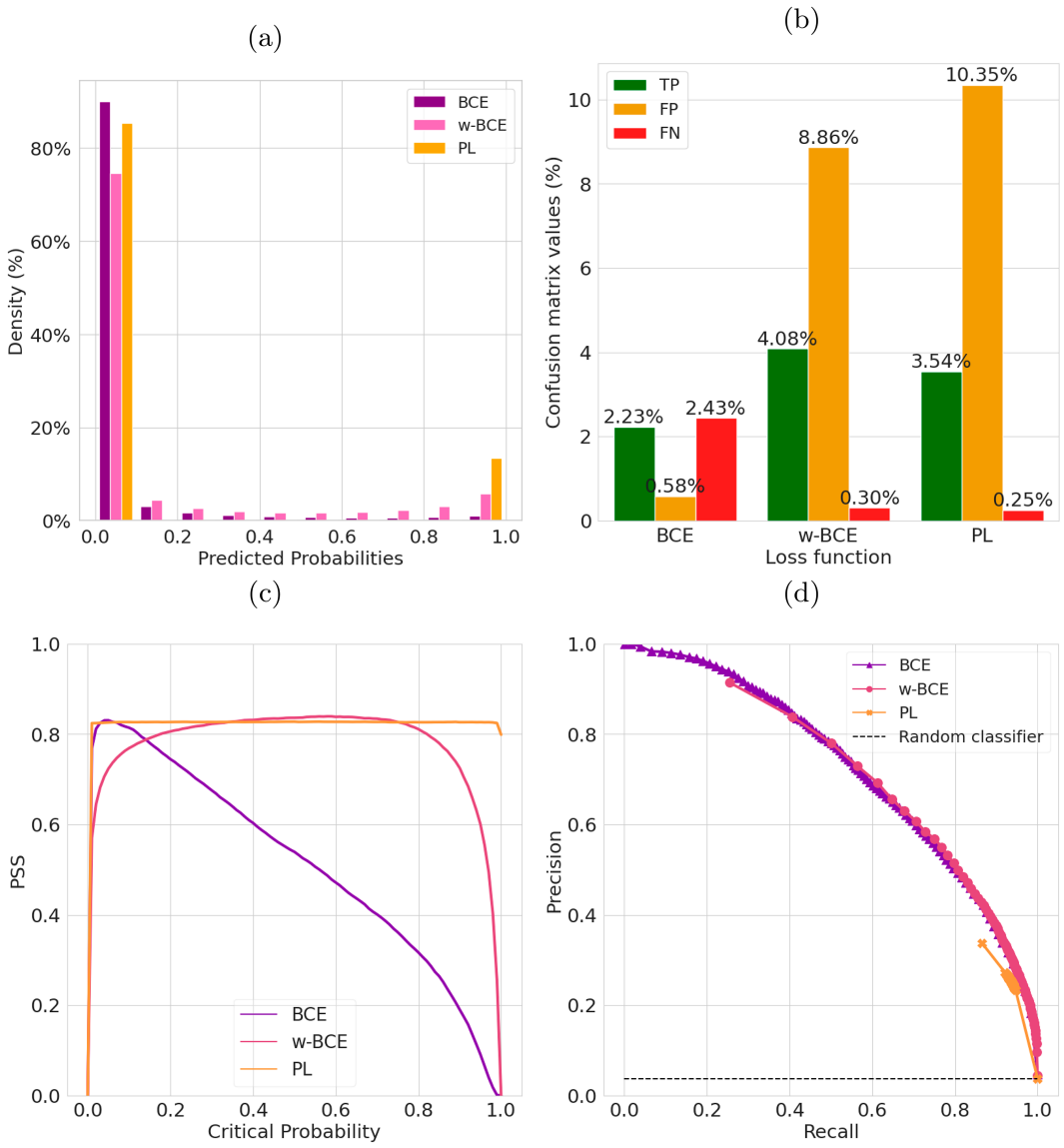

FIG. 7. Impact of loss function and critical probability on learning performance for the NN-Hybrid trained on the Mediterranean, for a window size of 3 h, with the 95th quantile as the rainfall threshold. (a) Predicted probabilities density by loss function. (b) TP, FP, and FN values in percentage by loss function. In green: TPs; in orange: FPs; in red: FNs. (c) Critical probability impact on the PSS for each loss function. (d) Precision–recall curve for each loss function. In purple: BCE; in pink: w-BCE; in orange: PL.

of the PSS to P_C for each loss function, while Fig. 7d shows the associated precision–recall curve. As shown by the purple line in Fig. 7c, the BCE is highly sensitive to the choice of P_C, and the PSS drops rapidly with increasing P_C values. A peak in PSS is observed for $P_C^\star = s = 0.05$, which corresponds to the relative frequency s of the class 1 event ($Y = 1$). This value corresponds to the theoretical value which optimizes PSS as shown by Mason (1979). Most BCE values are then located at high precision and low recall (purple symbols in Fig. 7d). When using the w-BCE (pink line in Fig. 7c), PSS reaches its maximum around $P_C^\star \approx 0.5$ and is more stable over a wide range of P_C (between 0.2 and 0.8). This observation is explained by simple arguments in appendix A. Most corresponding values in Fig. 7d are found for high recall and low precision. Finally, when using the PL (orange line in Fig. 7c), the PSS performance is almost constant for all values of P_C, with the exception of extreme values close to 0 or 1. This stability is reflected in the precision–recall curve (Fig. 7d), where a cluster of values is located at high recall and low precision.

In summary, as illustrated by the empirical results in Fig. 7 and explained in appendix a, both balanced loss functions (w-BCE and PL) provide a greater stability with P_C, making the standard choice of $P_C = 0.5$ appropriate for PSS evaluation. The BCE shows sensitivity to the choice of the threshold

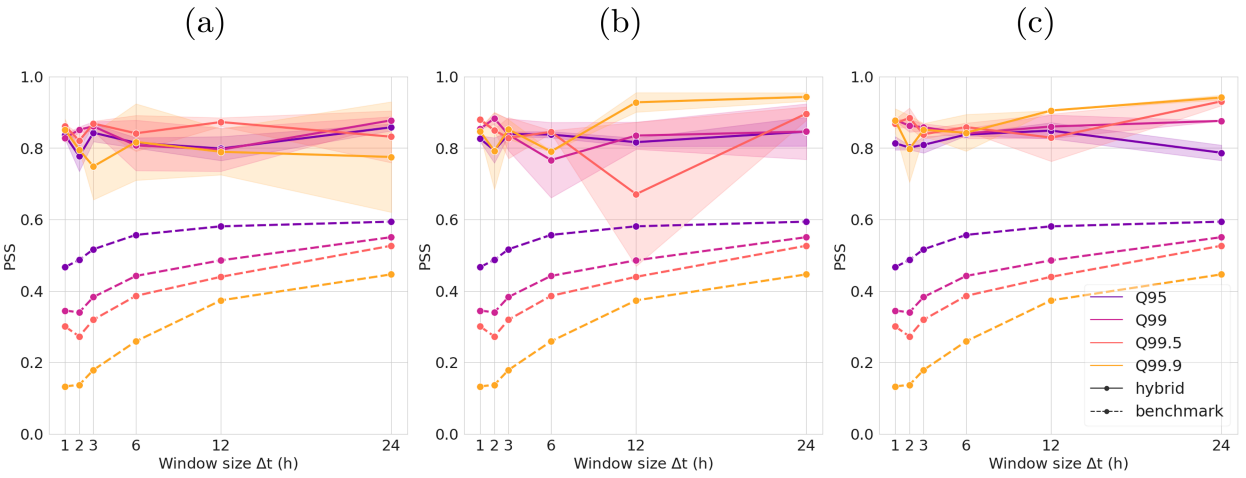

FIG. 8. Results for predicting increasingly intense rainfall events with NN-Hybrid trained on the Mediterranean hybrid dataset, with (a) the BCE, (b) the w-BCE, and (c) the PL as the loss function. For (a), $P_C = 0.05$, $P_C = 0.01$, $P_C = 0.005$, and $P_C = 0.001$, respectively, for the 95th, 99th, 99.5th, and 99.9th quantiles. For (b) and (c), $P_C = 0.5$. From coldest to warmest color: 95th quantile, 99th quantile, 99.5th quantile, and 99.9th quantile. Solid line: NN-Hybrid predictions. Dashed and dotted line: benchmark predictions. The shading represents the dispersion between the PSS computed on the validation and test datasets.

probability P_C and requires careful tuning of P_C to match the PSS and recall performances of the two balanced loss functions.

c. *NN-Hybrid performance for HPEs prediction*

Figure 8 summarizes the performance of NN-Hybrid on PSS compared with the benchmark for increasingly intense rainfall events and different loss functions, namely, the BCE (Fig. 8a), the w-BCE (Fig. 8b), and PL (Fig. 8c).

For all rainfall thresholds, the benchmark (in dashed lines) displays results that are consistent with what was discussed previously in section 3a and showcased in Fig. 6a, namely, better performance for longer Δt than shorter ones. However, performance decreases with increasing rainfall thresholds (from purple to pink, orange, and yellow). In other words, the more intense and rare the event being predicted, the more the benchmark struggles to correctly predict its occurrence. Figure 8a shows that, when using the BCE with P_C set equal to the event's probability frequency s, the PSS performance of NN-Hybrid (solid lines) is significantly better than the benchmark, reaching PSS values close to 0.8, with minimal sensitivity to rainfall thresholds. Figures 8b and 8c show similar results when using a balanced loss function with a standard choice of $P_C = 0.5$. First, with both the w-BCE and the PL, all NN-Hybrid predictions are reaching PSS values above 0.7, outperforming the benchmark for all Δt and rainfall threshold values. Moreover, the PSS performance of NN-Hybrid with both balanced loss functions is stable with increasing rainfall threshold.

Summing up, the results reported in Fig. 8 show that the NN-Hybrid model is suitable for the prediction of HPEs for all the considered quantiles. Indeed, all three methods demonstrated stable performance as the rainfall threshold increases, which means that NN-Hybrid can predict the occurrence of the rarest 5% of HPEs as effectively as the 0.1%. These results also confirm that the performance is comparable across the three tested loss functions, provided that P_C has been chosen in a relevant manner. It is important to note that the result variability between validation and test datasets (indicated by shading) can be significant (e.g., Fig. 8b with $\Delta t = 12$ h and the 99.5th quantile) but is overall low, especially when using the PL as a loss function.

d. *Difference in predictability between regions and case studies*

Figure 9 highlights the PSS performance of the NN-Hybrid model by region and month of the year, using the 95th quantile as the rainfall threshold, $P_C = 0.5$, and the PL as a loss function. Figure 9a first shows the sensitivity to Δt for the oceanic region (in blue) and the Mediterranean region (in green). These results reveal consistently better performance in the Mediterranean, for all Δt, for both NN-Hybrid and benchmark predictions. Moreover, Fig. 9b provides further detail by comparing monthly PSS performance across the two regions. While benchmark performance is stable with seasons for the two regions, a drop in PSS performance for NN-Hybrid predictions is seen during the summer in the Mediterranean, with better performance during fall and winter. In the oceanic region, however, no clear seasonal cycle is observed. This can be explained by the dry summers of the Mediterranean climate, where few days of rain are observed, while the oceanic climate is prone to more frequent rainy days as shown in Fig. 3b. It is noteworthy that the PSS performance in the Mediterranean peaks during the autumn season, which is when most HPEs occur in this region. Further analyses were performed to test a model on a region while trained on the other, but results showed poor performance. This suggests that the models learn specific meteorological patterns related to each region.

Examples of HPE predictions are shown in Fig. 10 for both the oceanic and Mediterranean regions to illustrate the NN-Hybrid forecast in comparison to the raw AROME forecast.

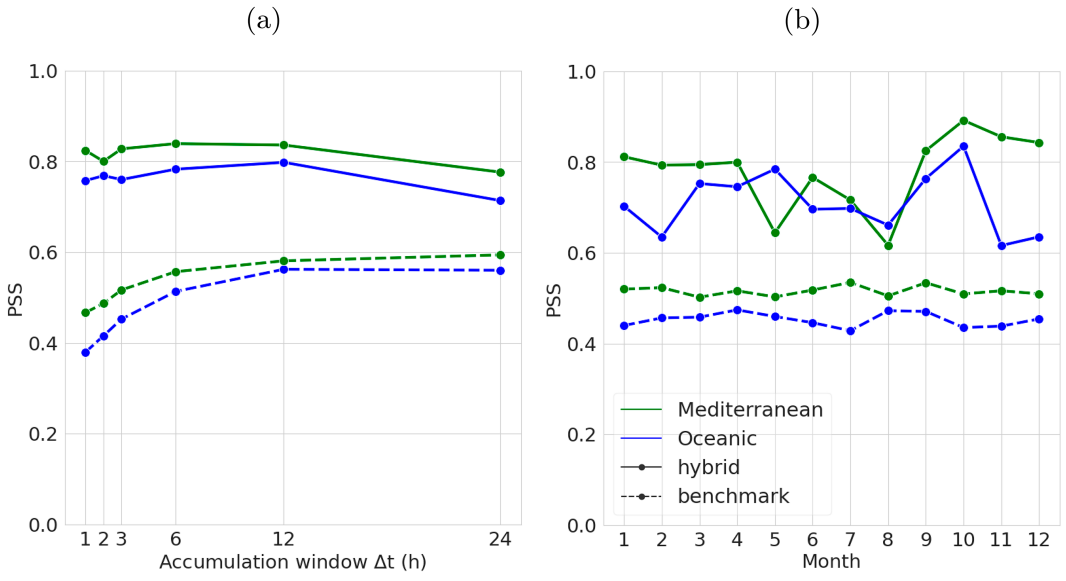

FIG. 9. Regional and seasonal impacts on the PSS performance. (a) Sensitivity to Δt and (b) sensitivity to the month of the year, for the oceanic (blue) and Mediterranean (green) regions, and NN-Hybrid (solid lines) and benchmark predictions (dashed lines). In both panels, $P_C = 0.5$, and the rainfall threshold is the 95th quantile, with the PL as the loss function.

The Mediterranean example (Figs. 10a,b) is an event with southeasterly moist airflow originating from the Mediterranean and impinging on the Massif Central's orography, leading to intense convection and thunderstorms. The maximum accumulated precipitation observed between 1200 and 2400 UTC reaches 167.1 mm at ground stations and 340 mm over the sea as measured by radars. The oceanic example (Figs. 10c,d) is associated with a surface low pressure system over Germany, which induced a southwesterly flow from the Atlantic. The maximum cumulated precipitation observed between 0000 and 1200 UTC reaches comparatively lower values of 51.9 mm from ground stations and 100 mm over the sea from radars. The forecast improvement of NN-Hybrid (Figs. 10b,d) compared to AROME (Figs. 10a,c) can be seen for both events. At a significant number of stations, AROME fails to predict rainfall threshold exceedance, as shown by many FN (red points) observed in locations where precipitation was intense. On the other hand, NN-Hybrid successfully forecasts rainfall threshold exceedance, and TPs (green) are observed where AROME fails, whereas no FN is present. However, as expected, the number of FPs (orange) increases with NN-Hybrid. TNs (white) are generally well predicted by both AROME and NN-Hybrid, with AROME performing slightly better, as expected.

Thus, the case studies illustrate the improvement provided by NN-Hybrid in both climates for reducing missed detections of actual and potentially high-impact events, albeit at the cost of more frequent false alarms.

4. Conclusions

Heavy precipitation events (HPEs) in the Mediterranean can cause significant damage and casualties; therefore, it is necessary to anticipate their occurrence. However, despite considerable improvement in the past decades, numerical weather prediction (NWP) models still face challenges in forecasting Mediterranean HPEs (Khodayar et al. 2021). This study proposes an approach to improve precipitation forecasting with neural networks (NNs) using data from both NWPs and station observations. It focuses on exceedance of high rainfall thresholds at specific locations for time windows from 1 to 24 h. The Peirce skill score is used for forecast verification, as it is an equitable metric suited for rare event predictions.

Results show that NN forecast using data from the global ARPEGE and regional AROME NWP models considerably improve precipitation forecasts compared to raw AROME NWP. When using only observed data, the NN offers similar performance but only for short time windows. This is consistent with the literature, as most studies which leverage only observed data face a rapid decline in performance with increasing lead times (e.g., Andrychowicz et al. 2023).

The hybrid dataset—NN-Hybrid, which combines observed and NWP data—offers the best performance overall. By combining the best of both types of data, a clear stability with the time window is found. An additional benefit of NN-Hybrid is that rainfall threshold exceedance is predicted equally well across high thresholds ranging from the 95th to the 99.9th quantiles, for each region and each time window. Results also show statistically better performance in the Mediterranean climate compared to the oceanic climate, with peak performance during fall and winter, the seasons during which most Mediterranean HPEs occur (Ricard et al. 2012). The approach of this study is similar to MetNet2 Hybrid from Espeholt et al. (2022), where the authors present a certain stability with increasing rainfall threshold for short lead times, although their performance drops at the highest rainfall threshold and lead times greater than 1 h.

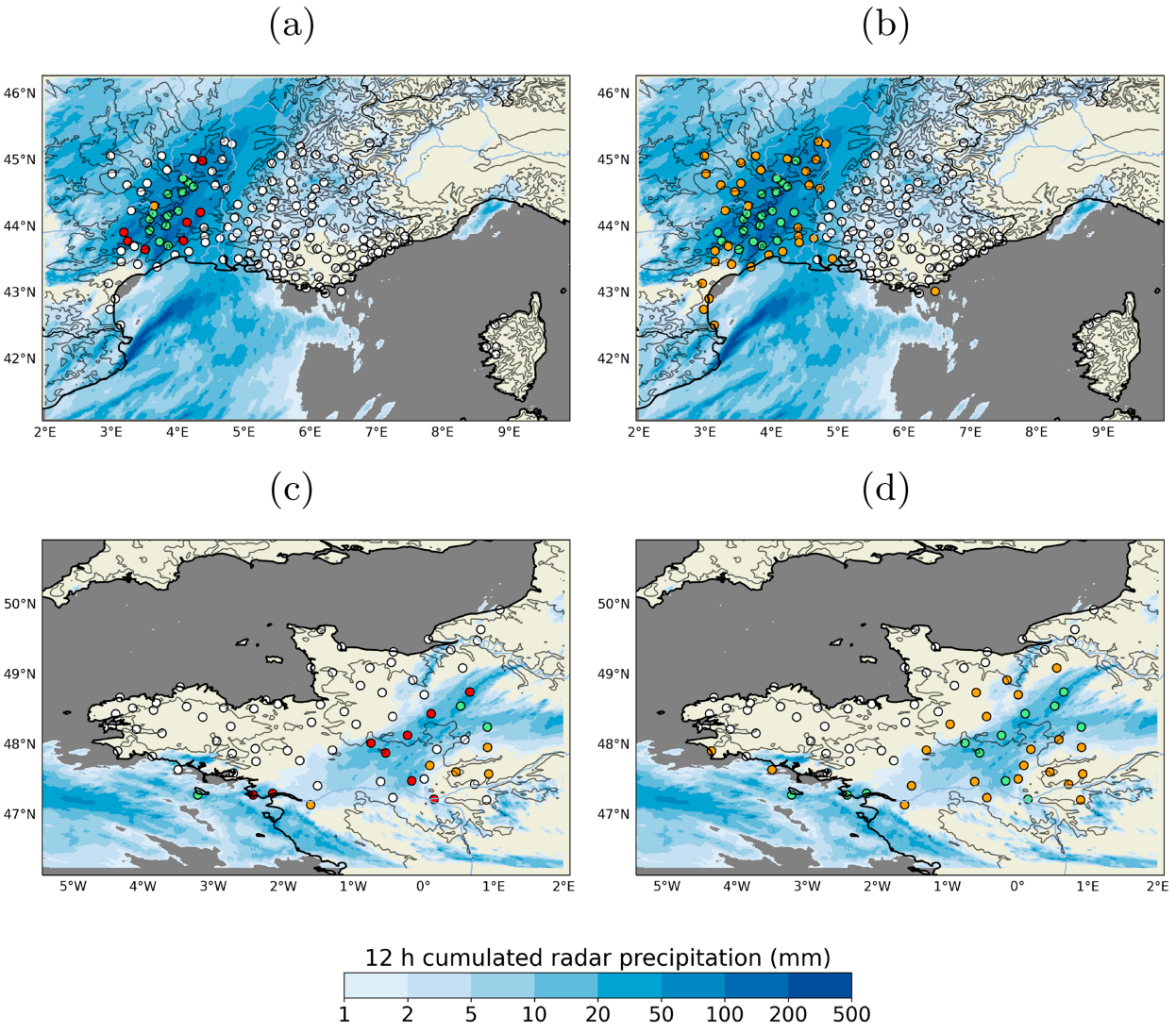

FIG. 10. Example case studies of rainfall threshold exceedance forecast: 12-h cumulated precipitation prediction from (a),(c) AROME and (b),(d) NN-Hybrid at (a),(b) 1200 to 2400 UTC 4 Nov 2017 in the Mediterranean and (c),(d) 0000 to 1200 UTC 29 May 2016 in the oceanic region. White, green, orange, and red dots represent TN, TP, FP, and FN, respectively, with the 99.9th quantile as the rainfall threshold for the Mediterranean event and Q95th quantile for the oceanic one, the w-BCE as the loss function, and $P_C = 0.5$. The color shading represents 12-h cumulated precipitation observed by radar, and gray contours represent altitude.

These results depend on the informed choice of the loss function and the associated critical probability P_C, which is guided by the trade-off between missing as few HPEs as possible and avoiding overpredicting the occurrence of HPEs. Three loss functions are compared: the classical binary cross entropy and both the balanced weighted binary cross entropy and the Peirce loss introduced here. The two balanced loss functions portray a binary behavior; thus, results are stable with regard to the choice of P_C. On the other hand, the binary cross entropy outputs predicted probabilities close to the actual frequencies of the events to be predicted, which requires careful selection of P_C. By adjusting the value of P_C to the corresponding probability of the event (e.g., $P_C = 0.5$ for the 95th quantile), similar results are obtained with the binary cross entropy as with the two balanced loss functions with a standard choice of $P_C = 0.5$. Applying this configuration across the three loss functions leads to a considerable decrease in the false negative rate at the cost of an increased false positive rate. Given the severity of HPEs, this trade-off is considered acceptable and aligns with the scope of this study.

It is important to acknowledge the limitations of this study. First, the developed NN models only provide forecasts for different rainfall accumulation window sizes, but the sensitivity to delays from NN model initialization is not investigated. Furthermore, the available NWP is always initialized at 0000 UTC, which influences lead time sensitivity. A more detailed study is needed to assess whether the stability of NN models (in

particular NN-Hybrid) is still valid with delays from initialization to better match the framework used in operational forecasts and in other studies (Espeholt et al. 2022).

Additionally, the period covered by the database (2016–18) is relatively short, as is the number of HPEs to learn from. A longer database would allow for learning and validation on more extreme cases, which also are the most destructive. However, NWP data are nonhomogeneous due to regular improvements in NWP models, and high-resolution NWP models such as AROME typically cover only one or two decades. Using reforecast data would allow us to alleviate the former limitation by offering homogeneous data over a long period and an up-to-date model version (e.g., Doiteau et al. 2024). Moreover, this study focuses exclusively on deterministic forecasts, while a probabilistic approach could also be highly relevant for predicting Mediterranean HPEs (Grazzini et al. 2024). Exploring such probabilistic methods will be an important direction for future research. We also remark that using NWP data with higher resolution, as well as fine-scale remote sensing data such as from a lightning mapping array, may be particularly useful, especially given the steep mountains in the Mediterranean region, as a higher resolution would enable convection to be fully represented and the impact of orography to be more accurate. Finally, future work will explore whether training across both regions yields significantly better performance than training exclusively on the Mediterranean.

On another note, even though the goal of this study is to compare different approaches rather than to propose an innovative NN architecture, we acknowledge that the NN models used in this study, i.e., a combination of recurrent and convolutional NNs, are somewhat outdated compared to the latest developments in the field. Using more recent architectures, such as transformers or generative adversarial networks, could help us to better extract spatial and temporal information from feature data. However, these methods are more computationally expensive than the NNs used here. Finally, an explainable NN model would help us better understand how the model operates, which could help meteorological research in understanding intense weather events (McGovern et al. 2019). For instance, preliminary work on feature importance, following the Rasp and Lerch (2018) method, suggests that ARPEGE fields are the most influential in the NNs considered here. This raises questions about the role of synoptic-scale information in predicting intense rainfall events, which will be investigated in future work.

Acknowledgments. This work was supported by ANR, France, Grant SAPHIR project ANR-21-CE04-0014-03. We thank four anonymous reviewers for their constructive comments that helped improve the paper.

Data availability statement. The data used in this work come from the MeteoNet database (Larvor et al. 2020), which is publicly available and can be found at https://meteonet.umr-cnrm.fr/dataset/. The code used for this study is available at https://doi.org/10.5281/zenodo.16753452.

APPENDIX A

Optimal Threshold Value and Threshold Sensitivity for PSS

In this appendix, we examine the relationship between q_i, the probability that $Y_i = 1$, and p_i, the output of a model trained to minimize each of the three losses defined in section 2f. This problem is closely related to the calibration of the predicted distribution, a key aspect in probabilistic forecasting.

In practice, the exact relationship between q_i and p_i is difficult to establish, as q_i cannot be observed. Here, we consider an idealized scenario in which each q_i is assumed to be perfectly known. Under this assumption, we can replace Y_i with its probabilistic counterpart q_i in Eqs. (7)–(9). We also assume in Eq. (9) that N is large enough so that $N^{-1}\sum_{i=1}^{N} Y_i \simeq s$. In that case, the value that minimizes the BCE (that turns out to be the negative log likelihood) is simply:

$$p_i^{\mathrm{BCE}} = q_i. \tag{A1}$$

According to Eq. (8), p_i^{WBCE} that minimizes the w-BCE loss has to satisfy $p_i/(sp_i) = (1 - q_i)/[(1 - s)(1 - p_i)] = 0$, which leads to

$$p_i^{\text{w-BCE}} = \frac{q_i(1-s)}{q_i(1-s) + s(1-q_i)}. \tag{A2}$$

Note that if $s = 1/2$, w-BCE reduces to BCE, as expected, and we recover $p_i^{\text{w-BCE}} = p_i^{\mathrm{BCE}}$. Finally, it is straightforward to show that PL is optimized by

$$p_i^{\mathrm{PL}} = \begin{cases} 1 & \text{if } q_i \geq s \\ 0 & \text{otherwise} \end{cases}. \tag{A3}$$

This equation implies that the PL output directly corresponds to Y_i when $P_C = s$.

Although q_i is not observable, we can empirically verify that Eqs. (A2) and (A3) are meaningful since, thanks to Eq. (A1), they entail a specific relationship of $p_i^{\text{w-BCE}}$ and p_i^{PL} as functions of p_i^{BCE}. This is illustrated in Fig. A1 using prediction data from the hybrid model for Q95 and Q99 exceedance of Δt = 3-h cumulative precipitation at Mediterranean ground stations.

Equations (A1)–(A3) are also useful for determining the optimal threshold $P_C^\star$ for a given score. In particular, since the optimal threshold for the PSS score is given by $q^\star = s$, as shown by Mason (1979) (see also Jolliffe and Stephenson 2003), which corresponds to $P_C^\star = s$ in the case of the BCE loss. It results, from Eq. (A2), that the optimal value for w-BCE becomes $P_C^\star = 1/2$, independent of s. This latter value closely matches the empirically observed optimal threshold in Fig. 7c. For PL, as it can be seen from Eq. (A3), there is no well-defined $P_C^\star$, and regardless of the chosen P_C (e.g., $P_C = 1/2$), the predicted class Y_i remains unchanged and so the PSS value. This is also what we observe in Fig. 7c. Note that if the optimal threshold for a given metric is $q^\star = 1/2$, then the best threshold for w-BCE is $P_C^\star = 1 - s$, whereas PL is not well suited for this score.

Equations (A1)–(A3) can be further leveraged to analyze the behavior of the score $S(q)$ around its optimal value $q = q^\star$ when using different loss predictions. Specifically, if $S(q)$ corresponds to

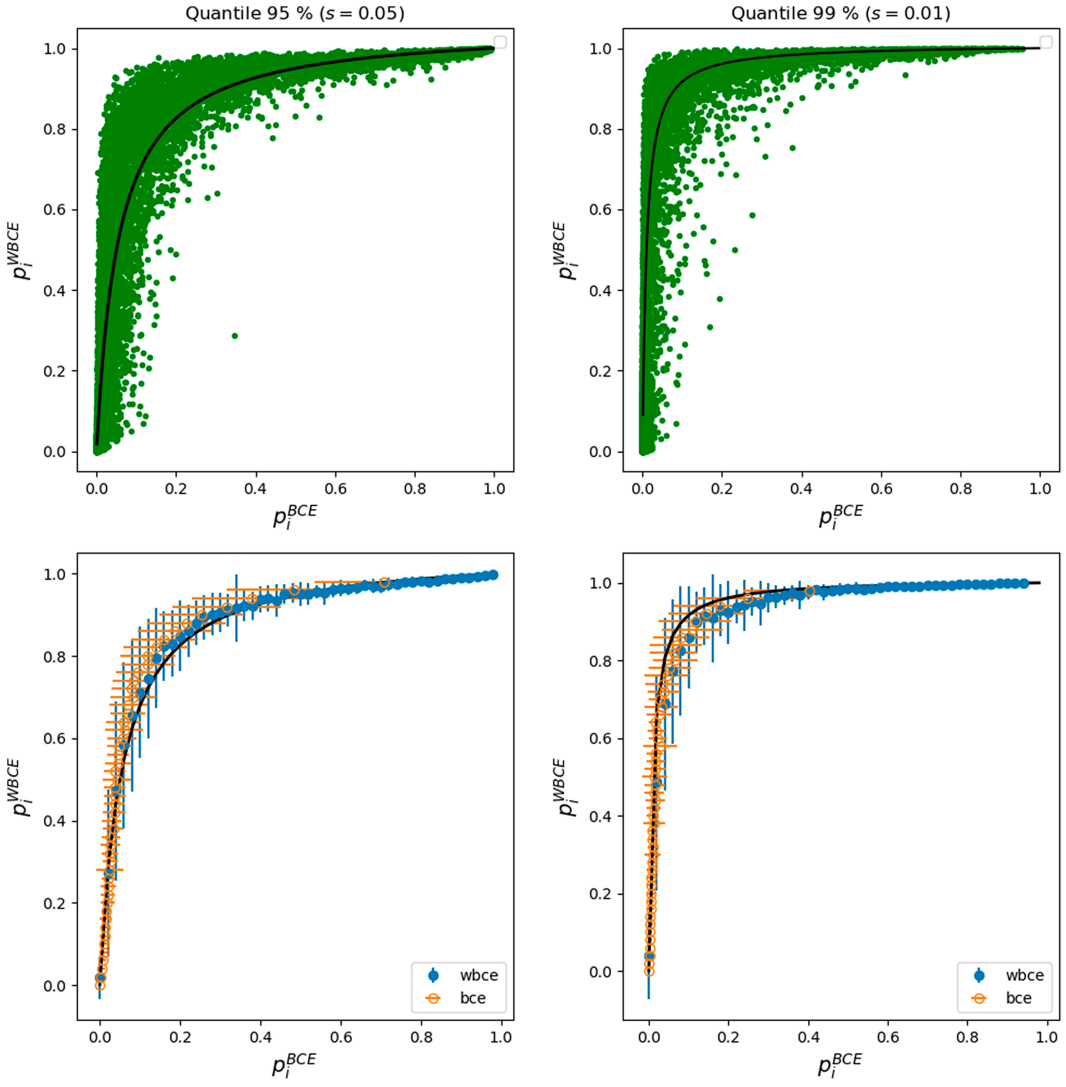

FIG. A1. Relationship between probabilistic predictions based on BCE (p_i^{BCE}) and w-BCE ($p_i^{\text{w-BCE}}$) losses. For each elements of the validation and test samples, the obtained values of $p_i^{\text{w-BCE}}$ vs p_i^{BCE} in the prediction of occurrence of threshold exceedance of the cumulative precipitation for $\Delta t = 3$ h in the case of (top left) Q95 and (top right) Q99 quantiles. (bottom) The corresponding conditional empirical mean and rms estimated either when p_i^{BCE} is fixed [blue (•) symbols] or when $p_i^{\text{w-BCE}}$ is fixed [orange (○) symbols]. In all panels, the solid black lines represent the expected relationship with respect to Eq. (A2).

$S(p^{\mathrm{BCE}})$, then the score computed for w-BCE, denoted as $S_1(p^{\mathrm{WBCE}})$, is given by

$$S_1(p^{\text{w-BCE}}) = S[q(p^{\text{w-BCE}})],$$

where $q(p)$ represents the inverse function of expression Eq. (A2). A straightforward algebraic derivation shows that at $q = q^\star$, where the first derivative of $S(q)$ vanishes, the ratio of the second derivatives of S and S_1 reads:

$$r^\star = \frac{S''_1(P_C^\star)}{S''_0(q^\star)} = \frac{s(1-s)}{[1-2P_C^\star(1-s)-P_C^\star]^2}. \tag{A4}$$

In the case of the PSS, where $P_C^\star = 1/2$, this ratio takes values close to $r^\star = 1/5$ and $r^\star = 1/25$ for the Q95 and Q99 thresholds, respectively. As observed in Fig. 7c, for high-quantile thresholds, the PSS score based on w-BCE exhibits significantly lower sensitivity around its maximum than the one obtained with BCE, indicating greater stability across a wider range of P_C values.

This effect is even more pronounced when using the PL loss, for which the second derivative is theoretically zero. Consequently, as emphasized earlier, the PSS score remains unaffected by the choice of P_C.

APPENDIX B

Alternative Classification Scores

Besides the PSS used throughout this paper, there are many scores that can be derived from the confusion matrix and that are commonly used to assess the performance of binary classification across various domains (Jolliffe and Stephenson 2003; Wilks 2019).

The Heidke skill score (HSS) is a widely used score that evaluates the performance of binary forecast performance by comparing its accuracy to that expected by "pure chance." It is defined as

$$\mathrm{HSS} = \frac{\frac{\mathrm{TP}+\mathrm{TN}}{N} - E}{1 - E},$$

where $(\mathrm{TP}+\mathrm{TN})/N$ is the rate of correct predictions (both negative and positive) by the method being evaluated and E is the expected rate of correct prediction obtained by chance. If FP and FN stand for the number of false positive and false negative forecasts, E, the probability of correctly predicting 0 or 1, is simply the sum of the product of probabilities of predicting one class (0 or 1):

$$E = \frac{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP}) + (\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}{N^2}.$$

The HSS yields a score ranging from -1 to 1. A perfect forecast achieves an HSS of 1, indicating perfect agreement between forecasts and observations beyond what is expected by chance. A score of 0 indicates that the forecast has no skill relative to the reference forecast (i.e., it is no better than chance or climatology), while negative HSS values imply that the forecast performs worse than the reference forecast. HSS is valuable for assessing overall forecast skill, but it is sensitive to class imbalance and may not be suitable for predicting HPEs. As advocated in Jolliffe and Stephenson (2003), the optimal threshold probability for HSS depends on HSS itself and of s, the rate of observed events, but when the latter is small enough, choosing the value $P_C = 0.5$ is appropriate.

The critical success index (CSI), also known as the threat score (TS), is another metric that can be particularly useful for imbalanced datasets. It measures the accuracy of positive predictions, ignoring true negatives, making it useful for assessing rare event forecasting, such as HPEs. Unlike PSS or HSS, CSI does not incorporate all four elements of the confusion matrix. It is defined as

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{B1}$$

Note that the CSI ranges from 0 (worst) to 1 (best). As with HSS, the optimal threshold value for CSI depends on the score itself and must be estimated empirically. However, when s is small, $P_C = 0.5$ is generally close to the optimal value.

The performance of all NN models optimized using the binary cross-entropy loss, compared to the benchmark in terms of HSS and CSI scores, is reported in Fig. B1. In both cases, we used a threshold probability $P_C = 0.5$, as it is expected to approximate the optimal value. We observe that, even for these scores, the hybrid model outperforms the other variants. Moreover, even if P_C is suboptimal, the hybrid model still outperforms the benchmark across all window sizes.

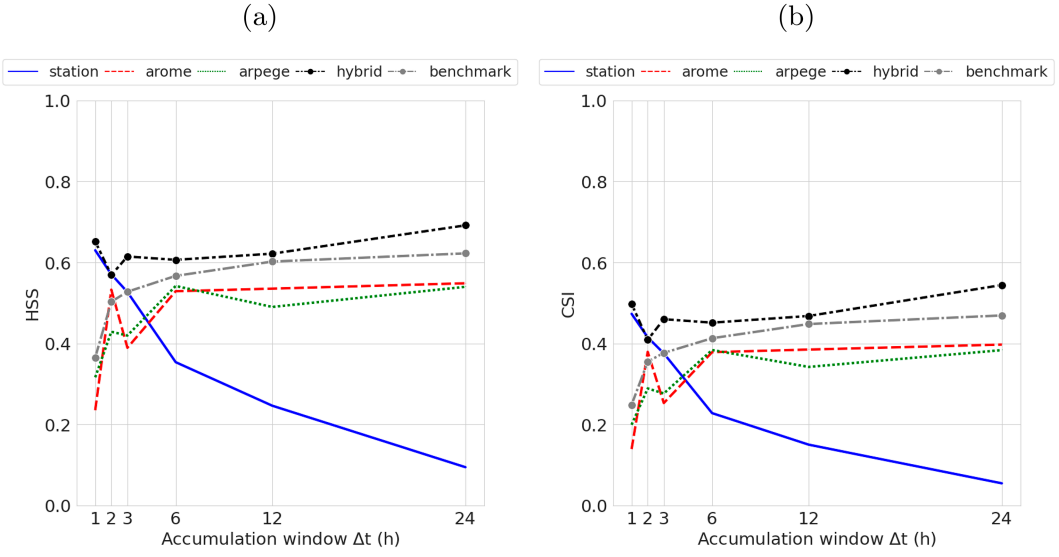

FIG. B1. (a) HSS and (b) CSI scores as the function of the window size Δt for various NN models and the benchmark. The HPEs considered are the 5% ($s = 0.05$) rarest rainfall events (95th quantile), and the loss is the BCE loss with $P_C = 0.5$. In solid blue: NN-Station. In dashed red: NN-AROME. In dotted black: NN-Hybrid. In dotted and dashed gray: benchmark performance. Even if $P_C = 0.5$ can be suboptimal for both metrics, we see that the hybrid model provides the best results at all Δt.

REFERENCES

Abadi, M., and Coauthors, 2015: TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org.

Agrawal, S., L. Barrington, C. Bromberg, J. Burge, C. Gazen, and J. Hickey, 2019: Machine learning for precipitation nowcasting from radar images. arXiv, 1912.12132v1, https://arxiv.org/abs/1912.12132.

Andrychowicz, M., L. Espeholt, D. Li, S. Merchant, A. Merose, F. Zyda, S. Agrawal, and N. Kalchbrenner, 2023: Deep learning for day forecasts from sparse observations. arXiv, 2306.06079v3, https://doi.org/10.48550/arXiv.2306.06079.

Argence, S., D. Lambert, E. Richard, N. Söhne, J.-P. Chaboureau, F. Crépin, and P. Arbogast, 2006: High resolution numerical study of the Algiers 2001 flash flood: Sensitivity to the upper-level potential vorticity anomaly. *Adv. Geosci.*, **7**, 251–257, https://doi.org/10.5194/adgeo-7-251-2006.

——, ——, ——, J. Pierre Chaboureau, J. Philippe Arbogast, and K. Maynard, 2009: Improving the numerical prediction of a cyclone in the Mediterranean by local potential vorticity modifications. *Quart. J. Roy. Meteor. Soc.*, **135**, 865–879, https://doi.org/10.1002/qj.422.

Ayzel, G., T. Scheffer, and M. Heistermann, 2020: RainNet v1.0: A convolutional neural network for radar-based precipitation nowcasting. *Geosci. Model Dev.*, **13**, 2631–2644, https://doi.org/10.5194/gmd-13-2631-2020.

Baggio, R., and J.-F. Muzy, 2024: Improving probabilistic wind speed forecasting using M-Rice distribution and spatial data integration. *Appl. Energy*, **360**, 122840, https://doi.org/10.1016/j.apenergy.2024.122840.

——, K. Pujol, F. Pantillon, D. Lambert, J.-B. Filippi, and J.-F. Muzy, 2025: Local wind speed forecasting at short time horizons based on numerical weather prediction and observations from surrounding stations. *J. Geophys. Res. Mach. Lear. Comput.*, **2**, e2025JH000709, https://doi.org/10.1029/2025JH000709.

Baïle, R., and J.-F. Muzy, 2023: Leveraging data from nearby stations to improve short-term wind speed forecasts. *Energy*, **263**, 125644, https://doi.org/10.1016/j.energy.2022.125644.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Beck, H. E., N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, 2018: Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data*, **5**, 180214, https://doi.org/10.1038/sdata.2018.214.

Ben Bouallègue, Z., and Coauthors, 2024: The rise of data-driven weather forecasting: A first statistical assessment of machine learning–based weather forecasts in an operational-like context. *Bull. Amer. Meteor. Soc.*, **105**, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1.

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619**, 533–538, https://doi.org/10.1038/s41586-023-06185-3.

Bresson, R., D. Ricard, and V. Ducrocq, 2009: Idealized mesoscale numerical study of Mediterranean heavy precipitating convective systems. *Meteor. Atmos. Phys.*, **103**, 45–55, https://doi.org/10.1007/s00703-008-0338-z.

Chollet, F., and Coauthors, 2015: Keras. https://keras.io.

Courtier, P., C. Freydier, J.-F. Geleyn, F. Rabier, and M. Rochas, 1991: The Arpege project at Meteo France. *Seminar on Numerical Methods in Atmospheric Models, 9-13 September 1991*, Reading, United Kingdom, ECMWF, 193–232, https://www.ecmwf.int/en/elibrary/74049-arpege-project-meteo-france.

Doiteau, B., F. Pantillon, M. Plu, L. Descamps, and T. Rieutord, 2024: Systematic evaluation of the predictability of different Mediterranean cyclone categories. *Wea. Climate Dyn.*, **5**, 1409–1427, https://doi.org/10.5194/wcd-5-1409-2024.

Ducrocq, V., O. Nuissier, D. Ricard, C. Lebeaupin, and T. Thouvenin, 2008: A numerical study of three catastrophic precipitating events over southern France. II: Mesoscale triggering and stationarity factors. *Quart. J. Roy. Meteor. Soc.*, **134**, 131–145, https://doi.org/10.1002/qj.199.

——, and Coauthors, 2014: HyMeX-SOP1: The field campaign dedicated to heavy precipitation and flash flooding in the northwestern Mediterranean. *Bull. Amer. Meteor. Soc.*, **95**, 1083–1100, https://doi.org/10.1175/BAMS-D-12-00244.1.

Duffourg, F., and V. Ducrocq, 2011: Origin of the moisture feeding the heavy precipitating systems over southeastern France. *Nat. Hazards Earth Syst. Sci.*, **11**, 1163–1178, https://doi.org/10.5194/nhess-11-1163-2011.

Ebert, P. A., and P. Milne, 2022: Methodological and conceptual challenges in rare and severe event forecast verification. *Nat. Hazards Earth Syst. Sci.*, **22**, 539–557, https://doi.org/10.5194/nhess-22-539-2022.

Espeholt, L., and Coauthors, 2022: Deep learning for twelve hour precipitation forecasts. *Nat. Commun.*, **13**, 5145, https://doi.org/10.1038/s41467-022-32483-x.

Flaounas, E., and Coauthors, 2025: Dynamics, predictability, impacts and climate change considerations of the catastrophic Mediterranean Storm Daniel (2023). *Wea. Climate Dyn.*, **6**, 1515–1538, https://doi.org/10.5194/wcd-6-1515-2025.

Frnda, J., M. Durica, J. Rozhon, M. Vojtekova, J. Nedoma, and R. Martinek, 2022: ECMWF short-term prediction accuracy improvement by deep learning. *Sci. Rep.*, **12**, 7898, https://doi.org/10.1038/s41598-022-11936-9.

Gandin, L. S., and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361–370, https://doi.org/10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.

Goodfellow, I., Y. Bengio, and A. Courville, 2016: *Deep Learning.* The MIT Press, 800 pp.

Grazzini, F., J. Dorrington, C. M. Grams, G. C. Craig, L. Magnusson, and F. Vitart, 2024: Improving forecasts of precipitation extremes over northern and central Italy using machine learning. *Quart. J. Roy. Meteor. Soc.*, **150**, 3167–3181, https://doi.org/10.1002/qj.4755.

Haiden, T., and Coauthors, 2015: Evaluation of ECMWF forecasts, including 2014-2015 upgrades. ECMWF Tech. Memo 765, 53 pp., https://www.ecmwf.int/sites/default/files/elibrary/2015/15275-evaluation-ecmwf-forecasts-including-2014-2015-upgrades.pdf.

Hally, A., E. Richard, S. Fresnay, and D. Lambert, 2014: Ensemble simulations with perturbed physical parametrizations: Pre-HyMeX case studies. *Quart. J. Roy. Meteor. Soc.*, **140**, 1900–1916, https://doi.org/10.1002/qj.2257.

IPCC, 2023: Mediterranean region. *Climate Change 2022—Impacts, Adaptation and Vulnerability.* Cambridge University Press, 2233–2272.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide* in *Atmospheric Science.* John Wiley and Sons, 254 pp.

Khodayar, S., and Coauthors, 2021: Overview towards improved understanding of the mechanisms leading to heavy precipitation in the western Mediterranean: Lessons learned from

HyMeX. *Atmos. Chem. Phys.*, **21**, 17 051–17 078, https://doi.org/10.5194/acp-21-17051-2021.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Larvor, G., L. Berthomier, V. Chabot, B. L. Pape, B. Pradel, and L. Perez, 2020: Meteonet, an open reference weather dataset by Meteo France. https://github.com/meteofrance/meteonet.

Leinonen, J., U. Hamann, and U. Germann, 2022: Seamless lightning nowcasting with recurrent-convolutional deep learning. *Artif. Intell. Earth Syst.*, **1**, e220043, https://doi.org/10.1175/AIES-D-22-0043.1.

Liu, Q., and Coauthors, 2023: Deep-learning post-processing of short-term station precipitation based on NWP forecasts. *Atmos. Res.*, **295**, 107032, https://doi.org/10.1016/j.atmosres.2023.107032.

Mandement, M., and M. Kreitz, 2025: Précipitations et inondations exceptionnelles autour de valence. *Météorologie*, **128**, 7–10, https://doi.org/10.37053/lameteorologie-2025-0003.

Mason, I., 1979: On reducing probability forecasts to yes/no forecasts. *Mon. Wea. Rev.*, **107**, 207–211, https://doi.org/10.1175/1520-0493(1979)107<0207:ORPFTY>2.0.CO;2.

McGovern, A., R. Lagerquist, D. John Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Murphy, K. P., 2022: *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning Series, Vol. 1, The MIT Press, 864 pp.

Nuissier, O., B. Joly, A. Joly, V. Ducrocq, and P. Arbogast, 2011: A statistical downscaling to identify the large-scale circulation patterns associated with heavy precipitation events over southern France. *Quart. J. Roy. Meteor. Soc.*, **137**, 1812–1827, https://doi.org/10.1002/qj.866.

Pardé, M., 1941: La formidable crue d'octobre 1940 dans les pyrénées-orientales. *Revue Géographique des Pyrénées et du Sud-Ouest. Sud-Ouest Européen*, **12**, 237–279, https://doi.org/10.3406/rgpso.1941.4493.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arXiv.2202.11214.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454, https://doi.org/10.1126/science.ns-4.93.453.b.

Pirone, D., L. Cimorelli, G. Del Giudice, and D. Pianese, 2023: Short-term rainfall forecasting using cumulative precipitation fields from station data: A probabilistic machine learning approach. *J. Hydrol.*, **617**, 128949, https://doi.org/10.1016/j.jhydrol.2022.128949.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Ricard, D., V. Ducrocq, and L. Auger, 2012: A climatology of the mesoscale environment associated with heavily precipitating events over a northwestern Mediterranean area. *J. Appl. Meteor. Climatol.*, **51**, 468–488, https://doi.org/10.1175/JAMC-D-11-017.1.

Rodwell, M. J., 2011: On Peirce's motivation for equitability in forecast verification. *Mon. Wea. Rev.*, **139**, 3667–3669, https://doi.org/10.1175/MWR-D-11-00167.1.

Scheffknecht, P., E. Richard, and D. Lambert, 2016: A highly localized high-precipitation event over Corsica. *Quart. J. Roy. Meteor. Soc.*, **142**, 206–221, https://doi.org/10.1002/qj.2795.

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Philos. Trans. Roy. Soc.*, **A379**, 20200097, https://doi.org/10.1098/rsta.2020.0097.

Schulz, B., and S. Lerch, 2022: Machine learning methods for postprocessing ensemble forecasts of wind gusts: A systematic comparison. *Mon. Wea. Rev.*, **150**, 235–257, https://doi.org/10.1175/MWR-D-21-0150.1.

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, https://doi.org/10.1175/2010MWR3425.1.

Shi, X., Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, 2015: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. arXiv, 1506.04214v2, https://doi.org/10.48550/arXiv.1506.04214.

——, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, 2017: Deep learning for precipitation nowcasting: A benchmark and a new model. arXiv, 1706.03458v2, https://doi.org/10.48550/arXiv.1706.03458.

Vannitsem, S., and Coauthors, 2021: Statistical postprocessing for weather forecasts: Review, challenges, and avenues in a big data world. *Bull. Amer. Meteor. Soc.*, **102**, E681–E699, https://doi.org/10.1175/BAMS-D-19-0308.1.

Wang, Y., and Coauthors, 2017: Guidelines for Nowcasting Techniques. WMO/TD-1198, 82 pp., https://library.wmo.int/records/item/55666-guidelines-for-nowcasting-techniques?offset=112.

Wilks, D. S., 2019: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, 4th ed. Elsevier, 369–483, https://doi.org/10.1016/B978-0-12-815823-4.00009-2.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209–1214, https://doi.org/10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2.