https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Forecasting threshold exceedance of atmospheric variables at a specific location

Roberta Baggio[a] and Jean-François Muzy[a]

[a]Laboratoire Sciences Pour L'Environnement, UMR 6134, CNRS Université de Corse, Avenue du 9 Septembre, Corte, France

Correspondence: Roberta Baggio (baggio_r@univ-corse.fr)

Abstract. Accurate short-term forecasting of extreme weather events is essential for early warning systems and disaster mitigation. This study compares two methodological approaches for predicting, at some given site, threshold exceedances of atmospheric variables such as temperature and wind speed: (i) direct probabilistic methods, which treat exceedance as a binary classification problem and (ii) full distribution probabilistic methods, which model the complete conditional probability law of the target variable. Using theoretical analysis and numerical simulations on a toy model, alongside real-world data from the MeteoNet dataset (2016–2018) for southeastern France, we demonstrate that the full distribution approach consistently outperforms the direct method for rare, extreme events. This advantage arises because the full distribution approach can effectively learn the parameters of the conditional distribution even from moderate and mild intensity events, thus achieving better calibration and discrimination in the tails. We find that the specific parametric shape of the chosen distribution plays a secondary role compared to accurately capturing predictable shifts in its bulk properties (i.e., mean and variance). This suggests that extreme exceedances are primarily driven by significant conditional displacements of the entire distribution, rather than by unpredictable, fat-tailed anomalies within a static climatology. Our results are validated for both strong surface wind speeds and intense hourly rainfall, with performance evaluated using proper scoring rules (Brier Score, logarithmic score) and deterministic skill scores (Peirce Skill Score, Critical Success Index, Heidke Skill Score). These findings highlight the critical importance of modeling the full probability distribution for rare-event forecasting and provide practical guidance for improving extreme weather prediction in operational meteorology.

1 Introduction

The accurate and timely prediction of extreme weather events is an important and difficult problem in operational meteorology (Seneviratne et al., 2023). Driven by climate change, the frequency and intensity of localized, high-impact phenomena are clearly increasing, posing severe risks to public safety, civil infrastructure and the stability of renewable energy grids. Despite remarkable progress in numerical weather prediction (NWP) (Bauer et al., 2015), including convection-permitting systems such as AROME from Météo-France, specifically designed to improve predictions at regional scale (Seity et al., 2011), pre-

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

dicting the precise timing and magnitude of localized extremes at the site level remains very challenging. This difficulty mainly stems from the highly non-linear and chaotic nature of atmospheric dynamics (Lorenz, 1963), compounded by the smoothing effects of grid-scale parameterizations, unresolved complex topography and sub-grid microphysical processes. Furthermore, for short-term forecasting purposes, the high computational cost of NWP models inherently limits their rapid-update capabilities. Consequently, at very short time scales, such as those required for nowcasting, prediction methods traditionally relied on statistical inference approaches that use historical data and past observed patterns to project future states. Early techniques range from the development of specific stochastic time-series models designed to account for observed localized fluctuations (see, e.g., Baïle et al. (2011); Tascikaraoglu and Uzunoglu (2014); Kaur et al. (2023)) to optical flow methods for radar tracking (Beauchemin and Barron, 1995; Ayzel et al., 2019). Building directly upon this foundation, modern machine learning (ML) leverages massive meteorological datasets to extract complex, nonlinear spatiotemporal patterns and enable hybrid approaches that combine in-situ observations with NWP outputs. The field of short-term weather prediction has been significantly transformed by deep learning approaches. For high-resolution prediction, architectures such as Deep Generative Models of Radar (DGMR) produce highly realistic probabilistic 90-minute rainfall forecasts (Ravuri et al., 2021), while attention-based models like MetNet and MetNet-2 deliver skillful, 1-km resolution predictions up to 12 hours ahead over continental domains (Sønderby et al., 2020; Espeholt et al., 2022). Concurrently, at the global scale, data-driven surrogates like GraphCast (Lam et al., 2023) and FourCastNet (Pathak et al., 2024) rapidly generate multi-day fields that can serve as boundary conditions for finer-scale models. A comprehensive review of these architectures falls beyond the scope of this paper and we refer to, e.g., Schultz et al. (2021); Bouallègue et al. (2024) for further details.

The current work specifically focuses on the application of ML and hybrid approaches to threshold exceedance predictions. Indeed, for early warning systems, disaster management or sectoral planning, predicting severe weather events is often formulated as forecasting threshold exceedances such as temperatures surpassing $30\,^{\circ}\mathrm{C}$, hourly rainfall exceeding $30\,\mathrm{mm}$, or wind speeds over $90\,\mathrm{km/h}$. In this study, we address the site-specific nowcasting of such exceedance events within a 0–6 h time window. Concretely, for a fixed location and atmospheric variable, our goal is to estimate the probability of exceedance in a form that can be rapidly updated and remains statistically well-calibrated (Bojinski et al., 2023). To achieve this task, existing methodologies can be split into two main categories. The first is *direct exceedance modeling*, which treats the exceedance (or non-exceedance) of a specific threshold as a Bernoulli outcome. This approach learns directly the probability $p \in [0, 1]$ of such Bernoulli event, framing the task as a standard binary classification problem optimized via Binary Cross-Entropy (BCE). The second category comprises *full-distribution* (or distributional) approaches. Instead of directly predicting the binary outcome, these methods estimate the complete conditional probability law of the target variable. The exceedance probability can then be computed directly from the cumulative distribution function (CDF) associated with the forecasted distribution.

Direct probabilistic forecasting for binary events is a well-established practice in the nowcasting of extreme meteorological conditions (Glahn and Lowry, 1972; Jolliffe, 2004; Wilks, 2009). This classification approach has been successfully applied to a wide variety of phenomena, including severe convective episodes (Pang et al., 2019), intense rainfall (Schaumann et al., 2021; Bouttier and Marchal, 2024; Pujol et al., 2025), and pollution peaks (Dutot et al., 2007). To generate these probabilistic forecasts, operational systems employ distinct methodological pathways. The first relies entirely on numerical weather predic-

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

tion (NWP) ensembles, estimating the probability of exceedance from the fraction of physical members that surpass a target threshold (Leutbecher and Palmer, 2008). A second class of methods is purely data-driven, treating threshold exceedance as a standard binary classification problem. ML classifiers, such as random forests or deep neural networks optimized via Binary Cross-Entropy, excel in this space by learning exceedance probabilities directly from large, labeled datasets of historical observations, radar imagery, or reanalysis (McGovern et al., 2017; Lagerquist et al., 2017; Agrawal et al., 2019). A third pathway consists of hybrid techniques that synthesize these two paradigms by post-processing and calibrating NWP forecasts to improve local accuracy and reliability. This includes statistical approaches such as logistic regression (Hess, 2020), as well as machine learning approaches such as neural networks combining NWP forecasts and observations (Pujol et al., 2025) or classifiers trained on NWP-derived predictors to directly estimate exceedance probabilities (McGovern et al., 2017).

Techniques designed to predict the full probability distribution provide a comprehensive characterization of predictive uncertainty and span a large range of statistical paradigms. Traditional parametric approaches, such as Ensemble Model Output Statistics and Generalized Additive Models for Location, Scale, and Shape, assume that the target atmospheric variable follows a pre-defined probability law (Gneiting et al., 2005; Schlosser et al., 2019). These models establish a mapping between atmospheric predictors and the distribution's parameters, typically optimizing a proper scoring rule such as the logarithmic score or the Continuous Ranked Probability Score (CRPS) (Jolliffe, 2004). In recent years, these parametric frameworks have been heavily augmented by deep learning (Salinas et al., 2020). The advent of "neural distributional regression" allows neural networks to non-linearly learn the predictor-to-parameter mapping, yielding significant improvements in forecast calibration and skill (Rasp and Lerch, 2018; Baggio and Muzy, 2024; Baggio et al., 2025). This concept has also been successfully extended to spatial domains through the use of gridded distributional U-Nets, particularly for the post-processing of precipitation fields (Pic et al., 2025).

Nonparametric and semiparametric alternatives, such as standard Quantile Regression Forests (QRF), circumvent rigid distributional assumptions for the bulk of the data by estimating conditional quantiles directly from the empirical distribution of decision tree leaves (Meinshausen and Ridgeway, 2006; Taillardat et al., 2016; Park et al., 2022). However, standard QRF exhibits a critical limitation for tail events: it cannot extrapolate beyond the maximum values observed in the training set. This extrapolation barrier is a fundamental challenge shared across the broader spectrum of statistical and deep learning architectures. To surmount this limitation, many extreme weather nowcasting approaches integrate Extreme Value Theory (EVT) (Coles et al., 2001) directly into modeling (Friederichs and Thorarinsdottir, 2012). This can typically be operationalized via the *Peaks Over Threshold* (POT) approach, which explicitly models excesses above a high threshold using the Generalized Pareto Distribution (GPD). While EVT-based methodologies are highly effective for calibrating early warnings of localized, severe events like flash floods, they introduce a notoriously difficult bias-variance trade-off: setting the POT threshold too low violates the asymptotic assumptions of the GPD, whereas setting it too high severely restricts the sample size available to robustly estimate the tail parameters (see, e.g., Bader et al., 2018)

The primary purpose of this paper is to systematically compare direct binary classification versus full-distribution parametric modeling for predicting the likelihood of extreme events. For both paradigms, this study relies upon a hybrid neural network architecture (Baggio et al., 2025; Pujol et al., 2025) that leverages high-resolution NWP predictions alongside local time-series

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

observations at the target site and its surrounding stations. We aim to investigate the intuitive idea that when exceedances are rare, direct binary classification suffers from extreme class imbalance, a scarcity of positive samples and a high sensitivity to threshold definitions. By contrast, distributional models can leverage abundant "non-extreme" outcomes to learn conditional scale and shape parameters. This allows them to produce better-calibrated exceedance probabilities, provided that the chosen family of probability distributions appropriately captures the conditional bulk-tail dependence. We first formalize this intuition within a simple theoretical framework, supported by both analytical and numerical evidence. We then validate these findings through two site-specific case studies: the exceedance of (i) strong near-surface winds and (ii) intense hourly rainfall in the Mediterranean region of southeastern France. For reproducibility and operational relevance, we use *MeteoNet* (2016–2018), an open Météo-France dataset aggregating co-registered ground-station and AROME/ARPEGE model outputs over two 550×550, km domains that encompass our study area (Larvor and Berthomier, 2021). Our verification methods follow established best practices, utilizing the Brier Score, Logarithmic Score, reliability diagrams, and ROC/AUC for probabilistic evaluation, alongside the Peirce Skill Score, Critical Success Index, and Heidke Skill Score for the deterministic prediction of binary outcomes. Finally, we discuss potential misspecification issues related to the specific choice of the parametric distribution.

The paper is organized as follows. Section 2 introduces the forecasting problem, formalizing the definitions and describing the two modeling approaches, direct probabilistic classification and full-distribution probabilistic modeling, along with the verification metrics used for evaluation. Section 3 provides a theoretical framework and illustrative experiments using a simplified generative model to compare the asymptotic behavior of both methods for extreme quantiles. Section 4 applies these approaches to real-world data from the MeteoNet dataset, detailing the dataset's characteristics, the neural network architecture used for predictions and presenting empirical results for wind speed and hourly cumulated rainfall forecasting. Section 5 concludes with a synthesis of the findings outlining questions for future research. Finally, the Appendix contains the technical material and detailed analytical computations notably for the toy model.

2 Statement of the problem

2.1 Extreme events as binary events

In this section, we set the main notations we use all along the paper and formally define the addressed problem. $Y(t)$ will stand for the value at time t of some atmospheric variable (i.e. $Y(t) = V(t)$ the wind speed, $Y(t) = T(t)$ the temperature, $Y(t) = R(t)$ the amount of precipitation during last hour, etc) at some given location. $Y(t)$ is considered as a stationary stochastic process taking value in $\mathbb{R}$.

At any time t, given a threshold Y_0 and a time horizon $h > 0$, our objective is to predict exceedance events of $Y(t+h)$. We formalize such events using a binary indicator:

$$I_{t+h}(Y_0) = \mathcal{H}(Y(t+h) - Y_0) = \begin{cases} 1 & \text{if } Y(t+h) \geq Y_0, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

where $\mathcal{H}$ denotes the Heaviside step function. Predicting extreme events thus reduces to forecasting $I_{t+h}(Y_0)$ for large Y_0 values, particularly those corresponding to high percentiles of the site's climatological distribution. For a probability level $1-p$ (with $p \ll 1$), we define the associated quantile Q_p as:

$$F_C(Q_p) = 1 - p, \tag{2}$$

where $F_C(z) = \int_{-\infty}^{z} f_C(u)\,du$ is the cumulative distribution function (CDF) derived from the climatological probability density function $f_C(u)$. When $p \ll 1$, $I_{t+h}(Q_p)$ indicates whether $Y(t+h)$ exceeds a threshold chosen in the distribution's upper tail. In the remainder of this paper, we drop the explicit threshold dependency and implicitly let I_{t+h} denote $I_{t+h}(Q_p)$ unless stated otherwise.

The previous prediction task constitutes a binary classification problem where the target is:

$$p_t \underset{\text{def}}{=} \text{Prob}\left(I_{t+h} = 1 \mid \mathcal{F}_t\right), \tag{3}$$

with $\mathcal{F}_t$ representing all information available at time t. A *probabilistic prediction*, namely an estimate of p_t, denoted as $\widehat{p}_t$, can then be converted to a *deterministic prediction* by thresholding at $p^\star$:

$$\widehat{I}_{t+h} = \begin{cases} 1 & \text{if } \widehat{p}_t > p^\star\,. \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

In the following sections, we introduce two distinct methodologies to estimate the conditional probability p_t and review the probabilistic and deterministic verification metrics used to assess and compare the quality of these threshold exceedance forecasts.

2.2 Modeling approaches

Our objective is to compare two distinct model classes for estimating the conditional exceedance probability p_t defined in Eq. (3). Both approaches utilize observable covariates X_t as input features representing all available information at time t. However, they fundamentally differ in how they process the target variable during training.

2.2.1 Class $\mathcal{M}_1$: Direct probability estimation

The first approach ($\mathcal{M}_1$) treats the task strictly as a binary classification problem. A model $M_1 \in \mathcal{M}_1$ maps the covariates directly to the estimated exceedance probability:

$$M_1(X_t; \widehat{\boldsymbol{\theta}}) = \widehat{p}_t^{(1)}, \tag{5}$$

where $\widehat{\boldsymbol{\theta}}$ represents the learned model parameters. This formulation corresponds to standard probabilistic classification, where parameters are typically optimized using the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[I_{t_i+h}\ln\left(\widehat{p}_{t_i}^{(1)}\right)(1 - I_{t_i+h})\ln\left(1-\widehat{p}_{t_i}^{(1)}\right)\right]. \tag{6}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

The BCE loss is particularly suitable for this task as it directly optimizes for probability calibration. However, it depends exclusively on the binary indicators I_{t_i+h} and consequently, during training, the model discards all continuous magnitude information of the underlying atmospheric variable, reacting only to whether the threshold was breached.

2.2.2 Class $\mathcal{M}_2$: Distribution-based probability estimation

The second approach ($\mathcal{M}_2$) adopts a two-stage procedure: (i) it estimates the full continuous conditional distribution of $Y(t+h)$ given X_t, and (ii) it derives the exceedance probability from this distribution. Following parametric deep learning frameworks (e.g., Salinas et al., 2020), the model outputs the time-dependent parameters Π_t of a chosen parametric family at each time step:

$$M_2\left(X_t;\widehat{\boldsymbol{\theta}}\right)=\widehat{\Pi}_t. \tag{7}$$

Let $f(y;\Pi_t):=\frac{d}{dy}\text{Prob}\left(Y(t+h)\le y\mid\mathcal{F}_t\right)$ denote the conditional probability density function (PDF) of $Y(t+h)$. The optimal parameters $\widehat{\boldsymbol{\theta}}$ are learned by minimizing the negative log-likelihood over the continuous observations:

$$\mathcal{L}_{\text{LL}}=-\frac{1}{N}\sum_{k=1}^{N}\ln f\left(Y(t_k+h);\widehat{\Pi}_{t_k}\right). \tag{8}$$

In contrast to $\mathcal{M}_1$, the training loss for $\mathcal{M}_2$ utilizes the exact continuous values of $Y(t+h)$, leveraging the entire dataset regardless of how extreme the threshold is. Once $\widehat{\Pi}_t$ is inferred, the exceedance probability is computed analytically or numerically as the tail probability:

$$\widehat{p}_t^{(2)}=\int_{Q_p}^{\infty}f(y;\widehat{\Pi}_t)\,dy. \tag{9}$$

This approach implicitly accounts for the full predictive distribution rather than focusing solely on the probability threshold, potentially stabilizing predictions when $p\to 0$.

2.3 Forecast verification of binary events

Evaluating threshold exceedance forecasts requires both deterministic metrics to assess binary decision-making and probabilistic scores to quantify calibration and resolution under heavy class imbalance (Jolliffe, 2004; Wilks, 2011).

2.3.1 Deterministic predictions

Binary forecasts $\widehat{I}_t(Y_0)$ are evaluated using standard metrics derived from the contingency table (TP, TN, FP, FN). We focus on three metrics suited to rare events ($p\ll 1$): the Peirce Skill Score (PSS), the Heidke Skill Score (HSS), and the Critical Success Index (CSI).

The PSS measures discrimination independently of class imbalance (Wilks, 2011):

$$\text{PSS}=\text{HR}-\text{FA}=\frac{\text{TP}}{\text{TP}+\text{FN}}-\frac{\text{FP}}{\text{TN}+\text{FP}}. \tag{10}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Conversely, the HSS measures accuracy relative to chance and remains sensitive to the base rate (Jolliffe, 2004):

$$\mathrm{HSS} = \frac{2(\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN})}{(\mathrm{TP}+\mathrm{FN})(\mathrm{FN}+\mathrm{TN}) + (\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})}. \tag{11}$$

The CSI isolates event detection by omitting true negatives (Wilks, 2011):

$$\mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}. \tag{12}$$

Binary decisions are obtained by thresholding probabilities at $p^\star$. Optimal thresholds depend strictly on the target metric (Mason, 1979; Jolliffe, 2004); analytically, $p^\star = p$ maximizes the PSS, whereas optimal thresholds for the CSI and HSS depend on the score values themselves and are computed numerically (Section 4). Further mathematical properties of these scores are detailed in Jolliffe (2004) and Wilks (2011).

2.3.2 Probabilistic predictions

Probabilistic forecasts are assessed using proper scoring rules and discrimination metrics to verify calibration, resolution, and sharpness (Wilks, 2011; Gneiting and Katzfuss, 2014). The Brier Score (BS) measures the mean squared error of the probabilities:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}\left(\widehat{p}_{t_i} - I_{t_i+h}\right)^2. \tag{13}$$

The Brier Skill Score corresponds to the comparison to $p(1-p)$, the expected BS obtained with "climatology" prediction:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{p(1-p)} = \frac{\mathrm{Res}-\mathrm{Rel}}{U}, \tag{14}$$

where Rel and Res denote the reliability and resolution components, and $U = p(1-p)$ represents the climatological uncertainty (see Murphy, 1973; Jolliffe, 2004; Wilks, 2011, for full algebraic decompositions). We also consider the logarithmic score (LS), which corresponds to the negative log-likelihood and perfectly matches the binary cross-entropy loss function used during training:

$$\mathrm{LS} = -\frac{1}{N}\sum_{i=1}^{N}\left[I_{t_i+h}\ln\widehat{p}_{t_i} + (1-I_{t_i+h})\ln(1-\widehat{p}_{t_i})\right]. \tag{15}$$

Both BS and LS are strictly proper scoring rules (Gneiting et al., 2006). Here they are complemented by the Area Under the ROC Curve (AUC) to assess threshold-independent ranking performance under severe class imbalance (Jolliffe, 2004).

3 Theoretical analysis and numerical experiments with a toy generative model

This section examines a simplified theoretical framework where the observable $Y(t+h)$ is generated from a covariate vector X_t through an underlying data-generating process. Our objective is to analytically compare the estimation errors of the two approaches ($\mathcal{M}_1$ and $\mathcal{M}_2$) introduced previously, focusing on their relative performance in predicting rare extreme events.

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Rather than aiming for a fully exhaustive and rigorous treatment, we provide analytical arguments that support the intuitive claim: for high thresholds where exceedances become increasingly rare, the distribution-based approach ($\mathcal{M}_2$) demonstrates superior sample efficiency compared to direct probability estimation ($\mathcal{M}_1$). This advantage stems from a critical distinction in information utilization: A model $M_2 \in \mathcal{M}_2$ leverages the complete continuous-valued observations, while model $M_1 \in \mathcal{M}_1$ effectively relies only on the sparse positive exceedance events, resulting in an effective sample size of approximately pN where p is the exceedance probability. Consequently, as $p \to 0$ (i.e., as events become increasingly rare), the performance advantage of M_2 over M_1 is expected to widen.

We begin by introducing our simplified modeling framework and derive analytical comparisons of M_1 and M_2 performance using three key metrics: the Brier Score (measuring L^2 error), the relative logarithmic score and the Peirce Skill Score. These theoretical findings are then validated through numerical experiments using an explicit generative model, with implementations of both estimation strategies ($\mathcal{M}_1$ and $\mathcal{M}_2$).

3.1 Estimation of the asymptotic errors for each model and their effects on performance scores.

We consider the following problem setup that is directly inspired from the simple example considered in Lerch et al. (2017). Let $(X_t)_{t\in\mathcal{T}}$ be a d-dimensional stationary random process of observable covariates. An unknown mapping $F : \mathbb{R}^d \to \mathbb{R}$ generates a latent mean signal:

$$\mu_t = F(X_t)\,. \tag{16}$$

By stationarity of X_t, the process μ_t is also stationary and we assume its marginal distribution is Gaussian with mean zero and variance s^2. At time $t+h$, we observe:

$$Y(t+h) = \mu_t + \nu_{t+h}, \tag{17}$$

where ν_t is a white noise process of variance σ^2 that is also assumed to be Gaussian. In that respect, the law of $Y(t)$ is Gaussian of variance $\sigma_Y^2 = s^2 + \sigma^2$. Throughout this paper, we will denote by $\phi(z)$ the standard normal density and $\Phi(z)$ the associated cumulative distribution function (CDF) and $\Phi^{-1}(q)$ the inverse cumulative distribution, namely, the reciprocal function of $\Phi(z)$. If $Z = (Z_1, \ldots, Z_n)$ is a random vector of law $f_Z(z_1, \ldots, z_n)$, the expectation of any function $G(z) = G(z_1, \ldots, z_n)$ with respect to Z, is denoted as

$$\mathbb{E}_Z\left[G(z)\right] = \int dz_1 \ldots dz_n \; G(z_1, \ldots, z_n)\; f_Z(z_1, \ldots, z_n) \tag{18}$$

For a fixed threshold $Y_0 \in \mathbb{R}$, the quantity of interest is the conditional exceedance probability:

$$P(X_t) \underset{\text{def}}{=} p_t = \Phi\left(\frac{\mu_t - Q_p}{\sigma}\right)\,. \tag{19}$$

where we noticed that $\text{Prob}\,(\nu > z) = 1 - \text{Prob}\,(\nu \leq z) = 1 - \Phi(\frac{z}{\sigma}) = \Phi(-\frac{z}{\sigma})$. We focus on small probability regime, namely $p \ll 1$ and we notice that for our model, the quantile Q_p reads:

$$Q_p = -\sqrt{s^2 + \sigma^2} \cdot \Phi^{-1}(p). \tag{20}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

A key parameter of the model is the noise-to-signal ratio:

$$\rho^2 = \frac{\sigma^2}{s^2}, \tag{21}$$

representing the relative magnitude of the idiosyncratic noise fluctuations ν_{t+h} in $Y(t+h)$ compared to its conditional mean signal μ_t. When ρ^2 is small, the predictability of I_{t+h} is high because p_t varies between values close to either $p_t = 0$ or $p_t = 1$ whereas when $\rho^2 \to \infty$, the predictability is small since p_t variation around its mean value p are small. In this context, it is natural to model ρ^2 as an increasing function of the forecasting horizon h, reflecting the fact that predictability inherently decreases over longer horizons. Consequently, when comparing model output to empirical results, a larger horizon must correspond to a higher effective value of ρ^2. Let us remark that, , with little algebra, one can easily compute the two first moments of p_t:

$$\mathbb{E}_{X_t}[p_t] = \mathbb{E}_{X_t}\left[\Phi\left(\frac{\mu_t - Q_p}{\sigma}\right)\right] = p. \tag{22}$$

and

$$\mathbb{E}_{X_t}[p_t^2] = \mathbb{E}_{X_t}\left[\Phi\left(\frac{\mu_t - Q_p}{\sigma}\right)^2\right] = \Phi_2\left(\Phi^{-1}(p), \Phi^{-1}(p); r\right) \tag{23}$$

where $r = \frac{1}{1+\rho^2}$ and $\Phi_2(x, y; r)$ stands for the cumulative distribution function of the standard bivariate normal distribution with correlation r. Notice that, when $\rho^2 \to \infty$, $r \to 0$. Since $\Phi_2(q, q, 0) = \Phi^2(q)$, we thus have, when $\rho^2 \to \infty$, $\mathrm{Var}(p_t) = \mathbb{E}(p_t^2) - \mathbb{E}(p_t)^2 \to p^2 - p^2 = 0$. Indeed, as mentioned above, when $\rho \to \infty$, $Y(t)$, is a pure, unpredictable, Gaussian white noise of variance σ^2 and therefore p_t is constant, $p_t = p$ independently of t. On the other hand, when $\rho^2 \to 0$, $r \to 1$ and since $\Phi_2(q, q, 1) = \Phi(q)$, one has $\mathrm{Var}(p_t) = \mathbb{E}(p_t^2) - \mathbb{E}(p_t)^2 \to p - p^2 = p(1-p)$ which is the variance of a Bernoulli process. This is simple to understand since, in that case, $Y(t+h)$ reduces to its predictable component μ_t and p_t becomes itself a Bernoulli process since $p_t = 1$ with probability p (if $\mu_t \geq Q_p$) and $p_t = 0$ otherwise.

Our goal is to estimate the function $P(\cdot)$ in Eq. (19) which maps X_t to the conditional probability p_t. This can be done using a model $M(., \boldsymbol{\theta})$, which parameters are learned over a training set $\{[X_t, Y(t+h)]\}_{t=1}^N$. We first consider a model $M_1(., \boldsymbol{\theta})$ that, following $\mathcal{M}_1$ approach, directly outputs an estimate $\widehat{p}_t^{(1)}$ of p_t. Its best parameters are obtained by minimizing the binary cross-entropy (BCE) loss associated with observed exceedances $I_{t+h}(Q_p)$. We also consider a model $M_2(., \boldsymbol{\theta})$ in the class $\mathcal{M}_2$ which provides an estimation of μ_t allowing one to compute the conditional probability estimate $\widehat{p}_t^{(2)}$ using Eq. (19). M_2 model's parameters are obtained by maximizing the log-likelihood which reduces, as σ^2 is known, to the Mean Squared Error.

In Appendix A, we show that, within this framework, in the regime $p \ll 1$ and when the number of observations N is large, under standard asymptotic regularity conditions (see, e.g., Vaart, 1998), one has the following estimation errors on $p_t^{(k)}$ of model M_k, $k = 1, 2$:

$$\mathcal{E}_1 = \mathbb{E}\left[(\widehat{p}_t^{(1)} - p_t)^2\right] \underset{p \to 0}{\sim} \frac{K_1(\rho)}{N} p^{\frac{\rho^2}{2+\rho^2}} \left[\ln(1/p)\right]^{-\frac{1}{2+\rho^2}}, \tag{24}$$

$$\mathcal{E}_2 = \mathbb{E}\left[(\widehat{p}_t^{(2)} - p_t)^2\right] \underset{p \to 0}{\sim} \frac{K_2(\rho)}{N} p^{\frac{2+2\rho^2}{2+\rho^2}} \left[\ln(\tfrac{1}{p})\right]^{\frac{1+\rho^2}{2+\rho^2}}, \tag{25}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

where the averages $\mathbb{E}$ are defined over all learned model's parameters (and over time t) and where the "noise-to-signal" ratio ρ^2 is defined in Eq. (21). This result first indicates that, at fixed p (small enough), as the noise-to-signal ratio increases, estimation error decreases. This counterintuitive result can be explained by the fact that the intrinsic predictability of the conditional probability p_t is limited by its variance, which can be computed from Eqs (22) and (23):

$$\mathrm{Var}(p_t) = \mathbb{E}[p_t^2] - p^2 = \Phi_2(\Phi^{-1}(p), \Phi^{-1}(p); r) - p^2 \,,$$

We have seen that as the ratio ρ^2 increases, the latent correlation r tends toward zero, physically implying that idiosyncratic noise dominates the systemic factor μ_t. Mathematically, this causes the bivariate distribution Φ_2 to factorize into the product of marginals p^2, driving $\mathrm{Var}(p_t)$ to zero and effectively turning p_t into a deterministic constant p, which is trivially predictable with zero error. From Eqs. (24) and (25) one can also see that, up to logarithmic corrections, we have

$$\frac{\mathcal{E}_2}{\mathcal{E}_1} = \mathcal{O}(p)$$

showing that, assuming that both M_1 and M_2 estimation methods are efficient and parameter estimation are asymptotically normal (see Appendix A), in the regime of large threshold Q_p (or $p \to 0$), one expects an error with approach $\mathcal{M}_2$ that is very small compared to the error using approach $\mathcal{M}_1$. This results originates from the fact that the effective amount of "information" used to calibrate M_1 parameters is pN instead of N resulting in a factor p in the asymptotic variance ratio of the two approaches.

These findings are confirmed when measuring the estimation performance in terms of skill scores, namely with PSS for deterministic predictions and Brier or logarithmic scores for probabilistic predictions. In Appendices B and C, we analyze the impact of parameter prediction errors on the performance as measured by BSS, LS and PSS for models M_1 and M_2. We notably show that BSS behavior is directly related to the behavior of errors $\mathcal{E}_2$ and $\mathcal{E}_1$ (see Eqs (B2)). When $p \ll 1$ we have:

$$BSS_k \approx p^{\frac{\rho 2}{2+\rho^2}} - \frac{\mathcal{E}_k}{p} \,. \tag{26}$$

Since, when $p \ll 1$, $\mathcal{E}_2 \ll \mathcal{E}_1$, this confirms that, in this regime, method M_2 provides better results than method M_1 since $BSS_2 > BSS_1$.

We can also compare the two methods in terms of LS. We demonstrate in Appendix B that LS difference reads when $p \downarrow 0$:

$$\Delta\mathrm{LS}_{1,2} \underset{\text{def}}{=} \mathrm{LS}_1 - \mathrm{LS}_2 \underset{p\to 0}{\sim} \frac{C_\rho}{2N}\left(K_\rho - p\ln\left(p^{-1}\right)\right) \tag{27}$$

where C_ρ and K_ρ are two positive constants defined in Appendix B. We see that, provided $p \ll 1$, $\Delta\mathrm{LS}_{1,2}$ is clearly positive meaning that M_2 approach outperforms M_1.

For PSS, we establish in Eqs. (C2) and (C3) of Appendix (C) explicit expressions in terms of p_t-averaged values for methods M_1 and M_2. Such integrals that can be evaluated numerically. In the regime $N \to \infty$, we obtains the the following asymptotic behavior for $p \to 0$:

$$\mathrm{PSS}_1 \quad \sim \quad 1 - K_\rho p^{\kappa^2} - \frac{C_1}{N} p^{-\gamma} \tag{28}$$

$$\mathrm{PSS}_2 \quad \sim \quad 1 - K_\rho p^{\kappa^2} - \frac{C_2}{N} p^{1-\gamma} \tag{29}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

where

$$\gamma = \frac{2\rho}{\sqrt{1+\rho^2}+\rho} \in (0,1) \text{ and } \kappa = \sqrt{1+\rho^2} - \rho$$

and K_ρ is a constant depending on ρ such that $1 - K_\rho p^{\kappa^2} \to 1$ when $\rho \to 0$ and $1 - K_\rho p^{\kappa^2} \to 0$ when $\rho \to \infty$. It results, as expected, that the maximum expected PSS cannot be positive in pure noise regime while can approach a maximum score (PSS = 1) in the perfectly predictable situation. According to Eq. (29), one expects the PSS to increase as $p \to 0$. This can be intuitively explained by the fact that, as $p \to 0$, the threshold Q_p becomes very large and the events $I_{t+h} = 1$ become "more predictable" since the idiosyncratic component plays a diminishing role. Indeed, exceedance for large thresholds can occur only when μ_t is very large and thus when $Y(t+h)$ is less dependent on ν_{t+h} in definition (17). Since μ_t represents the predictable part of the process, the predictability of $I_{t+h} = 1$ naturally improves in the small p regime. In contrast, this behavior is not observed for PSS_1 in Eq. (28). Although the intrinsic predictability of $I_{t+h} = 1$ increases as $p \to 0$, this benefit is entirely canceled by the estimation error of method M_1. Indeed, as the probability approaches zero, the effective number of positive cases pN drops, causing the variance of the estimator to explode and dominate the signal.

For N large enough, the PSS ratio is therefore expected to behave as:

$$\frac{\text{PSS}_1}{\text{PSS}_2} \sim 1 - \frac{K_p}{N} + \mathcal{O}(\frac{1}{N^2}) \text{ with } K_p \sim K p^{-\gamma}(1 - Cp) \, .$$

We thus recover that fact the M_2 has a better PSS than M_1 but both methods lead to the same PSS value as $N \to \infty$. We can also see that, at fixed N, the PSS ratio decreases as p becomes smaller, so that the smaller p, the better M_2 is with respect to M_1.

3.2 Numerical validation using a toy generative model

To empirically validate our analytical findings and provide illustrative examples, we implement the simple model described in Appendix D. Specifically, according to Eq. (D1), the latent process μ_t (Equation (16)) is constructed as a weighted sum of harmonic functions applied to the components of a d-dimensional Gaussian white noise input X_t. The so-obtained process μ_t is zero mean, approximately normal and the weights chosen such that its variance is s^2.

Both estimation models $M_1(X_t, \boldsymbol{\theta})$ and $M_2(X_t, \boldsymbol{\theta})$ employ identical multi-layer perceptron (MLP) architectures, each with 3 layers featuring:

- Input dimension matching the d-dimensional covariates X_t
- Two hidden layers with 32 ReLU-activated units
- Linear output layers representing logits for M_1 and regression for M_2

The $\mathcal{M}_1$-type model M_1 directly estimates exceedance probabilities $\text{Prob}(Y(t+h) > Y_0 \mid X_t)$ using binary cross-entropy with logits loss to optimize $\boldsymbol{\theta}$, while the $\mathcal{M}_2$-type model M_2 predicts the latent process μ_t by minimizing the mean squared error (MSE) between predicted and observed $Y(t+h)$ values. Both models are trained using the Adam optimizer with a

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

batch size of $2^{12} = 4096$, learning rate of 10^{-3} and early stopping based on validation loss with a patience of 20 epochs. The validation set comprises a separate 10% split of the original training data.

Our experimental setup uses $d = 12$ input dimensions with training and test sets containing $N = 2^{15}$ and $N' = 2^{14}$ samples respectively. All simulations, model training and predictions were implemented in Python using the PyTorch framework, ensuring efficient GPU acceleration and reproducible results. To robustly evaluate estimation errors, we employ a kind of cross-validation with Monte-Carlo resampling approach: While keeping the set of test pairs $(Y(t+h), X_t)$ constant, we train both models on 30 independent realizations of the training set. This methodology provides stable estimates of model performance (notably their bias and variance) while accounting for the inherent variability in training process.

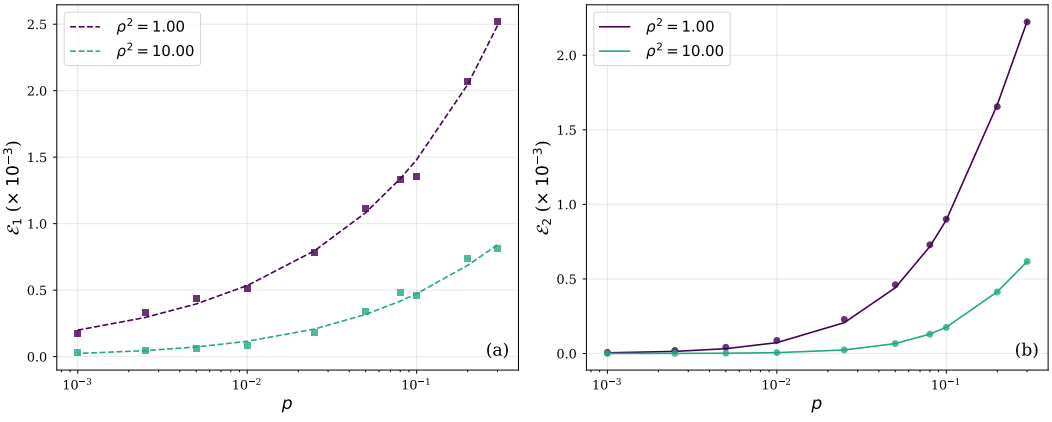

Figure 1. Comparison of the mean squared error of the M_1 and M_2 model predictions. Empirical estimates of $\mathcal{E}_1$ defined in Eq. (A16) (■) in panel (a)) and $\mathcal{E}_2$ (symbols (•) in panel (b)) defined in Eq. (A7) are displayed as a function of p for $\rho^2 = 1$ (dark blue) and $\rho^2 = 10$ (green). Dashed and continuous lines represent the analytical expressions expected from respectively Eqs. (A19) and (A9) (see text for details on numerical experiments).

Figure 1 presents empirical estimates of the prediction errors for models M_1 and M_2 across different exceedance probabilities p. Panel (a) displays $\mathcal{E}_1$ (squares, ■), as defined in Equation (A16), while panel (b) shows $\mathcal{E}_2$ (circles, •), defined in Eq. (A7). The results cover a range of threshold probabilities from $p = 10^{-3}$ to $p = 3 \times 10^{-1}$, corresponding to rare events. In accordance with our theoretical framework, these empirical estimates focus exclusively on the variance components of the prediction errors. We have verified that squared bias terms are negligible in the regime we consider, thereby validating the variance-dominated error assumption in Appendix A for the considered range of exceedance probabilities. We also consider two different noise-to-signal ratio, $\rho^2 = 1$ and $\rho^2 = 10$ while keeping the variance of $Y(t+h)$, $\sigma_Y^2 = s^2 + \sigma^2 = 2$ fixed.

As anticipated by the discussion after Eqs. (24) and (25), we clearly see that as the noise-to-signal ratio increases the variance of $\widehat{p}$ decreases, reflecting the fact that as ρ^2 increases p_t becomes more and more predictable (it converges to the climatology value p when $\rho^2 \to \infty$) and therefore the prediction error decreases. The dashed curves in panel (a) and solid curves in panel (b) represent our analytical predictions derived from Eqs (A19) and (A9) respectively. To achieve optimal alignment between theory and empirical results, we calibrated the constant terms in these analytical expressions. It is noteworthy that, for the $\mathcal{E}_1$ case when $\rho^2 = 1$, incorporating a quadratic correction term $V_1'(\mu) = V_1' + V_1''\mu^2$ in Eq. (A19) provides marginally better

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

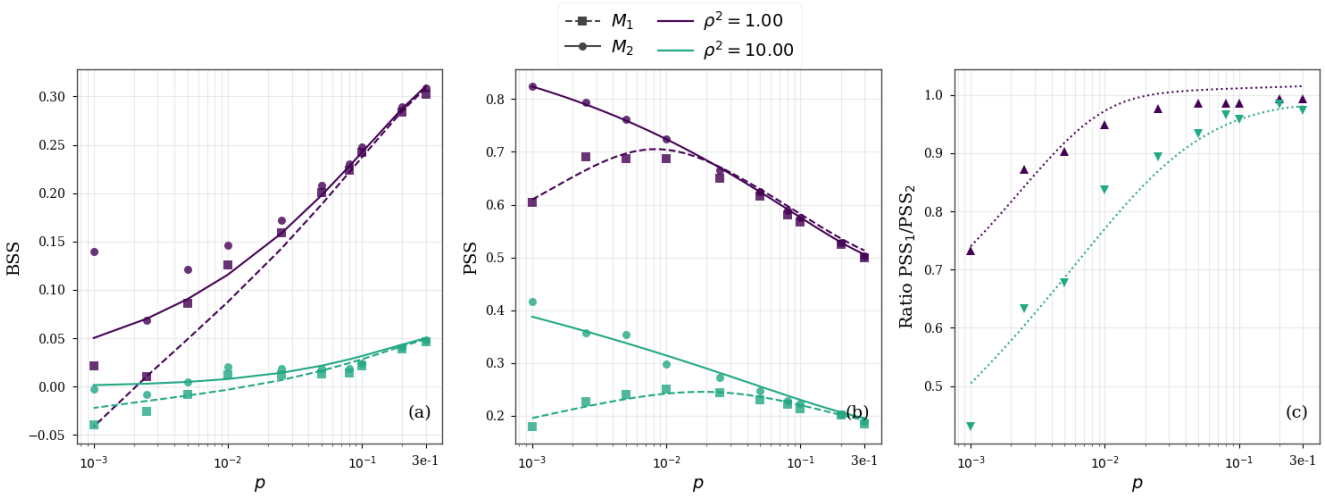

Figure 2. Comparison of Brier Skill Score (BSS) and Peirce Skill Score (PSS) for models $M_1 \in \mathcal{M}_1$ (symbols (■) and dashed lines) and $M_2 \in \mathcal{M}_2$ (symbols (●) and solid lines). Dark violet represent data for $\rho^2 = 1$ and while green represent data for $\rho^2 = 10$. Panel (a) shows empirically estimated BSS (Eq. (14)) a function of exceedance probability p. Panel (b) presents analogous PSS results and panel (c) illustrates the PSS performance ratios $\mathrm{PSS}_1/\mathrm{PSS}_2$ against p. Dashed and solid lines in all panels show theoretical predictions from Appendices B for BSS (Eq. (B2)) and C for PSS (Eqs. (C2) and (C3).

agreement than a simple constant adjustment. The results demonstrate excellent concordance between the estimated data and our analytical expressions, thereby providing empirical validation for theoretical hypotheses of Appendix A.

Figure 2 compares the performance of models M_1 and M_2 using the Brier Skill Score (Eq. (14)) and Peirce Skill Score (Eq. (10)). Panel (a) shows that, despite the superior predictability of p_t highlighted in Figure 1, the BSS falls as ρ^2 increases. This behavior stems from the degraded predictability of I_t, captured by the first term in Eq. (26). Indeed, the conditional probability distribution of p_t is sharper for small noise-to-signal ratio: p_t is more often closer to $p_t = 1$ or $p_t = 0$ when ρ^2 is small than when ρ^2 is large (in the limit $\rho \to 0$, p_t is either 0 or 1 and its conditional distribution is infinitely sharp). It also reveals that relative performance improves with increasing p, despite the increase in absolute error observed in Figure 1. This indicates that model performance relative to climatology deteriorates for rarer events. For moderate p, this behavior is mainly due to the term $p^{\frac{\rho^2}{2+\rho^2}}$ in Eq. (26) that does not depend on the the prediction method (see also Eq. (B2) for a more precise behavior). At smaller p, the contribution of $-\mathcal{E}_2$ is negligible for method M_2 while, since $\frac{\mathcal{E}_1}{p} \sim p^{-\frac{2}{2+\rho^2}}$, its contribution to BSS_1 becomes strongly negative. In Figure 2(b), we see that very much like BSS, PSS decreases with the noise-to-signal ratio ρ^2. Again, this stems from a better quality of I_t prediction for smaller ρ^2. The figure further demonstrates that as the exceedance probability p decreases, predictions from model M_2 become increasingly accurate, as evidenced by the monotonic increase in PSS_2. This trend confirms our theoretical analysis presented in Section 3.1 following Eqs (28) and (29). A similar pattern is observed for model M_1, though only for moderate p values. For very small p values, PSS_1 reaches a maximum and then declines, in full consistency with our theoretical predictions. In Figure 2(c) which examines the relative PSS of M_1 and M_2, one also clearly sees that method M_2 has larger PSS than method M_1 and this is all the more true when p is small. This

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

confirms that the two methods perform comparably for common events (largest p values), M_2 progressively outperforms M_1 as $p \to 0$. This observed advantage of M_2 for rare events aligns with our theoretical analysis in Section 3.1, thereby providing empirical validation of our analytical predictions regarding the superior sample efficiency of the distribution-based approach for extreme event prediction. Finally, we can notice that in all cases, the analytical curves derived in Appendices B (Eq. (B2) for BSS) and C (Eqs. (C2) and (C3) for PSS) provide a quite fair fit to the empirical data.

4 Application to rainfall and wind speed data

In this section, the problem introduced previously is examined in the context of forecasting the extreme occurrences of surface wind speed and hourly cumulated rainfall, respectively. First, the meteorological data used for this purpose are presented. The forecasting task, in its specific formulation, is then described in detail. This is followed by a description of the the structure of the input data, of the ANN architecture employed, and the main characteristics of the training procedure.

4.1 The MeteoNet dataset

The meteorological data used in this study were sourced from MeteoNet (Larvor and Berthomier, 2021), a comprehensive dataset curated and made publicly available by Météo-France to support researchers and data scientists. The dataset covers two regions, south-eastern and north-western France, for the three year period 2016–2018. It includes various type of observations measures and NWP forecasts gridded data. Following what done in previous work (Baggio et al., 2025), we focus on south-eastern France and use a subset of the available data, that is, only NWP forecasts and ground-station observations are considered in this study. The retained weather variables for each of the data type considered, which have been selected differently for wind speed and accumulated rainfall, are reported in Table 1. Although the original ground-station observations are available at 6-min resolution, we aggregate them to hourly time series to limit the number of model parameters, as detailed in Subsection 4.3.1. Concerning NWP forecasts, we consider two type of data: 2-D surface fields from the high-resolution AROME model (0.025°) and 3-D fields from the lower-resolution ARPEGE model (0.1°). For each day, the 24-h forecasts come from the 00 UTC run; AROME fields are provided hourly, while ARPEGE fields are available at 1-h or 3-h intervals depending on the lead time.

Starting from the raw dataset files, a two-step post-processing procedure is applied to prepare the model inputs. First, data is organized on a per-station basis by creating one file per selected ground station. Each file stores, in NetCDF format, the station's hourly time series together with those of neighboring sites, and includes a local subgrid of the 2-D and 3-D NWP fields centered on the station, retaining all available forecast times. During this first stage the integrity of data is preserved and no quality check is applied. Then, in a second phase, these station-based files are processed to form model-ready inputs. Samples containing invalid or missing data are discarded. To ensure statistical significance, for every station file it is checked that available keys are above a given threshold and discarded otherwise. When considering the weather variables reported on Table 1 and by using a threshold of 2000 and 500 for wind and rainfall respectively, 278 and 268 station are retained among the ones available in the original dataset (see Figure 3). Features tensors in valid samples of feature-label pairs are normalized

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

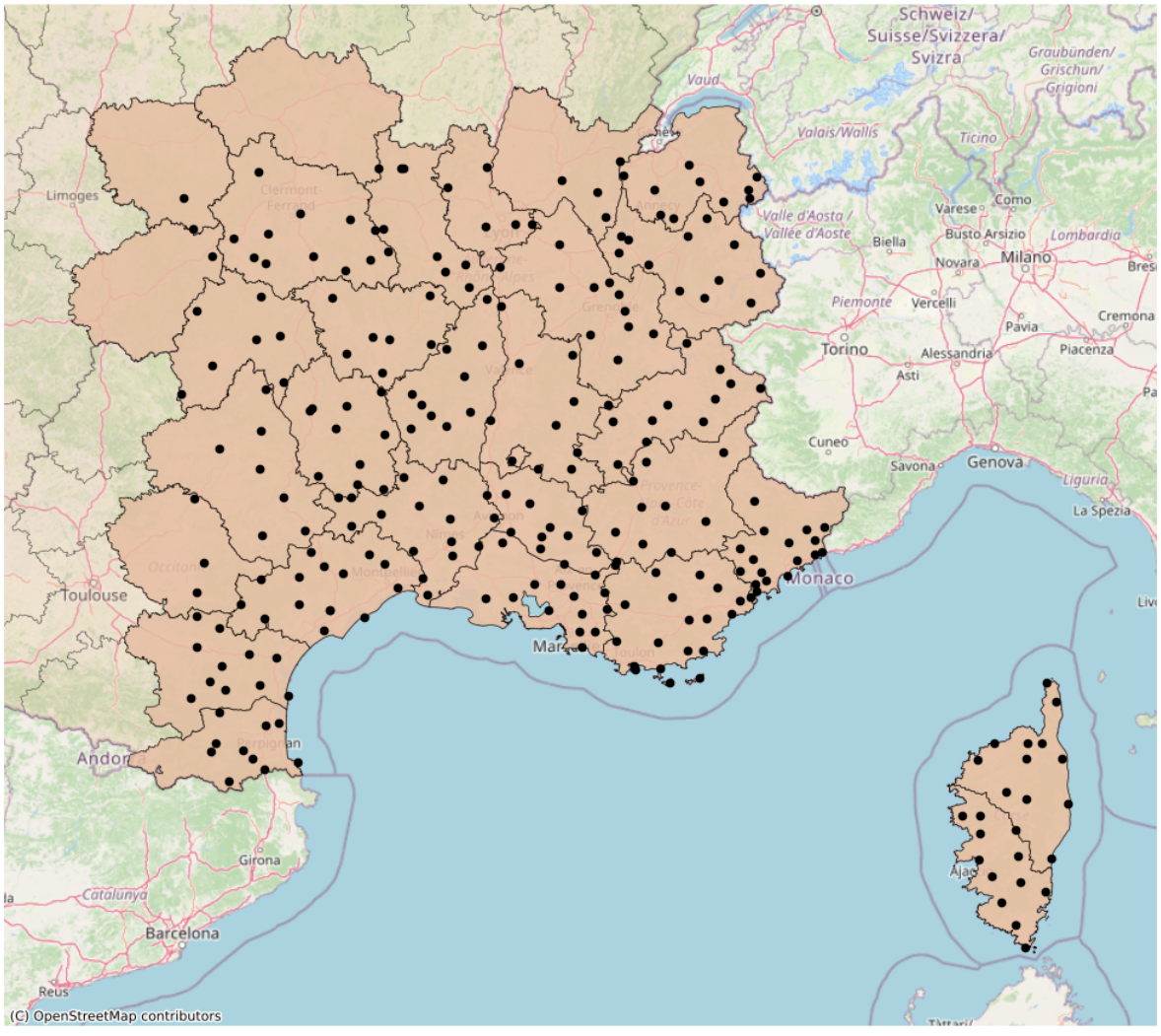

Figure 3. Geographical extent of the MeteoNet Southeast database, with the localization of the 278 ground stations (●)

385 before storing, then these samples are exported in an unified format compatible with Pytorch data generators, either as a large binary file or in-memory arrays.

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Table 1. Summary of the MeteoNet input data used for model training.

Input type	Space and time grids	Wind speed target	Cumulative rainfall target
Stations	Spatial grid: 11 locations (station + 10 neighbours) Time grid: current + 6 past hourly values	Wind components u, v ($\mathrm{m\,s^{-1}}$) Temperature (K)	Wind components u, v ($\mathrm{m\,s^{-1}}$) Temperature (K) Relative humidity (%) Precipitation ($\mathrm{mm\,h^{-1}}$)
AROME	Spatial grid: 11×11 Time grid: all 6 h-ahead predictions	2 m temperature (K) 2 m relative humidity (%) Wind components u, v ($\mathrm{m\,s^{-1}}$) MSLP (Pa)	Same as wind-speed target, plus 2 m dew-point temperature (K) and total precipitation (mm)
ARPEGE	Spatial grid: $7 \times 5 \times 5$ Time grid: all 6 h-ahead predictions	Temperature (K) Wind components u, v ($\mathrm{m\,s^{-1}}$) Pressure (Pa)	Same as wind-speed target, plus vertical velocity ($\mathrm{Pa\,s^{-1}}$)

4.2 Statement of the forecasting problem for wind and cumulative rainfall

Building on the discussion in Section 2.1, we now adapt the framework to our specific case study. Returning to the forecasting problem introduced in Eq. (1), we focus on threshold exceedance forecasts for a weather variable $Y(t)$ across multiple forecast horizons $\mathbf{h} = (h_1, h_2, \ldots, h_H)$. At a given initial time t and recording site S, the objective is to predict the vector of future exceedances $\mathbf{I}_{d,t+\mathbf{h}}$ that is, the H-dimensional vector

$$\mathbf{I}_{d,t+\mathbf{h}} = \begin{pmatrix} I_{d,t+h_1} \\ I_{d,t+h_2} \\ \vdots \\ I_{d,t+h_H} \end{pmatrix}, \tag{30}$$

representing threshold exceedances at site S over multiple future lead times. Following (Baggio et al., 2025), we set $H = 6$ with an hourly frequency, so that $\mathbf{I}_{d,t+\mathbf{h}}$ contains six components corresponding to exceedance predictions from 1 up to 6 hours ahead. Prediction vector $\widehat{\mathbf{I}}_{d,t+\mathbf{h}}$ is obtained by minimizing the losses defined in Eqs. (6) and (8) for methods M_1 and M_2, respectively. Notice that for each method, the loss function is extended to multiple horizons by summing over $h = 1, \ldots, 6$. This formulation implicitly treats the different forecast horizons as conditionally independent, an assumption adopted for tractability. Assessing and potentially relaxing this independence assumption constitutes a direction for future research.

For the two weather variables we consider, namely hourly wind speed (m/s) and 1-hour accumulated rainfall (mm), thresholds Q_p are calculated *station-wise*, that is, the climatological densities are station-specific: $f_C(y) = f_C^S(y)$. The quantile selection is done differently for wind speed and for cumulative rainfall. For wind speed, we simply define every Q_p using Eq. (2). Then results are computed for a list of 8 probabilities p, namely more specifically we use

$$p_W \in \{0.2, 0.1, 0.08, 0.05, 0.03, 0.01, 0.005, 0.002\}\,. \tag{31}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

In the case of rainfall, a different definition is adopted in order to ensure that the detected extreme quantiles remain meaningful despite the large number of dry days. Let $F^S_{C,+}$ denote the station-wise climatological CDF conditional on rainfall occurrence $F^S_{C,+}(y) = P(Y \leq y \mid Y > 0)$. The quantiles Q_p are defined with respect to this conditional distribution, i.e. $F^S_{C,+}(Q_p) = p_{+,R}$, where $p_{+,R}$ is such that:

$$p_{+,R} \in \{0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.025\} \, . \tag{32}$$

When displaying results, metrics are plotted as a function of the normalized p, recovered as $p = \bar{p}_{wet}\, p_{+,R}$, where $\bar{p}_{wet} = P(X > 0)$ is the probability of a rainy episode occurrence. This baseline probability is defined as the average probability of rainfall across all considered stations, yielding $\bar{p}_{wet} \approx 0.08$ for the present dataset. Since the rainfall occurrence probability is station-dependent, this formulation introduces a small approximation. However, it allows for a more consistent comparison of the results with theory, as the resulting unconditional probability levels p are substantially smaller than the corresponding conditional levels $p_{+,R}$.

4.2.1 Probabilistic models for surface wind speed and rainfalls

When using a probabilistic model of type $\mathcal{M}_2$, different parametric forms for the implied conditional density function $f(y)$ are adopted to model wind speed and accumulated rainfall. It is worth emphasizing that, although these distributions have been selected with care, the differences among alternative parametric families remain limited once fundamental physical constraints of the target variable are properly enforced, as discussed later in Subsection 4.5.2.

4.2.2 Wind speed

For wind speed, we adopt the so-called *Multifractal-Rice* (M-Rice) distribution, following previous work in Baggio and Muzy (2024), where it was shown to outperform classical alternatives such as the Weibull and Gamma distributions in forecasting applications. The M-Rice distribution, introduced in Baïle et al. (2011), is motivated by the random cascade framework used to describe fully developed turbulence. It generalizes the classical Rice distribution by allowing its scale parameter to be random, typically modeled as a log-normal variable, thereby incorporating intermittency effects. The resulting distribution is characterized by three parameters $(\nu, \sigma^2, \lambda^2)$. The parameter ν controls the mean level, σ^2 governs dispersion, and λ^2, often referred to as the *intermittency parameter*, regulates the tail behavior. In particular, larger values of λ^2 produce heavier tails, increasing the probability assigned to extreme wind speeds. The formal definition of the M-Rice distribution, together with a detailed interpretation of its parameters and numerical implementation, is provided in Appendix E1.

4.2.3 Accumulated rainfall

For hourly rainfall accumulation, we adopt a mixed lognormal (zero-inflated) distribution in order to account for the mixed discrete-continuous nature of precipitation. Rainfall data are characterized by a substantial probability mass at zero (dry events), together with a positively skewed continuous distribution for positive amounts. The mixed lognormal model explicitly captures this structure by combining a point mass at zero with a lognormal distribution for strictly positive values (Cho et al., 2004;

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Kedem et al., 1990). Formally, the distribution is governed by three parameters: the probability of rainfall occurrence p_{wet} and the lognormal parameters (μ, σ^2) controlling the mean and dispersion of positive rainfall amounts. The mathematical formulation of such "zero-inflated" lognormal distribution is provided in Appendix E2. In addition to the mixed lognormal, other mixed distributions have been tested, without substantial differences in the model results. These distributions are also defined in the Appendix E2.

4.3 Hybrid neural network for predicting cumulative precipitation or surface wind speed

In this subsection, the artificial neural network (ANN) used to forecast weather variables is presented. The overall architecture follows that proposed in a previous study (Baggio et al., 2025) and was re-implemented from scratch within the PyTorch framework. The data and the preprocessing steps required to prepare the model input are first described. Subsequently, the network architecture and the training procedure are briefly discussed.

4.3.1 Input design

The heterogeneous data sources described above are combined into a feature tensor used as input to the ANN, with each data type processed by a dedicated branch (see Subsection 4.3.2). More specifically, a feature-label couple $(\mathbf{X}_\mathrm{k}, \mathbf{Y}_\mathrm{k})$ is defined for each *key* $\mathrm{k} = (S, d, t)$ enconding station S, day d, and time t. More specifically, $(\mathbf{X}_k, \mathbf{Y}_\mathrm{k})$ denotes the input tensor containing all the variables used for training while $\mathbf{Y}_\mathrm{k}$ contains the target variables, consisting of the six future values (one for each forecast hour) of either wind speed or accumulated rainfall. The input $\mathbf{X}_k$ is defined as $\mathbf{X}_\mathrm{k} = \left[\mathbf{GS}_\mathrm{k}, \mathbf{AR}_\mathrm{k}, \mathbf{AP}_\mathrm{k}, \mathbf{C}_\mathrm{k}, \mathbf{D}_\mathrm{k}\right]$, where $\mathbf{GS}_\mathrm{k}$, $\mathbf{AR}_\mathrm{k}$, and $\mathbf{AP}_\mathrm{k}$ respectively denote features from ground stations, AROME, and ARPEGE, while $\mathbf{C}_\mathrm{k}$ and $\mathbf{D}_S$ encode temporal and spatial metadata. The ground station tensor $\mathbf{GS}_\mathrm{k}$ includes observations at the target site S and its 10 nearest neighboring stations. For n_S variables, this yields vectors of dimension $11\,n_S$ (33 for wind, 55 for rainfall; see Table 1). Using the current time and the six preceding hourly time steps, the resulting tensor has dimensions (7×33) for wind and (7×55) for rainfall. The AROME tensor $\mathbf{AR}_\mathrm{k}$ is constructed from a local spatial patch of 11×11 grid points centered on the station, corresponding to a spatial extent of $\pm 0.125°$. Forecasts at horizons $t+1$ to $t+6$ are included, leading to tensors of shape $(6 \times n_{AR} \times 11 \times 11)$, with $n_{AR} = 5$ for wind and 7 for rainfall. The ARPEGE tensor $\mathbf{AP}_\mathrm{k}$ is defined similarly, using a larger spatial extent ($\pm 0.2°$) but a coarser grid (5×5). Forecast horizons are matched to the closest available times (multiples of 3 hours when needed). The selected variables ($n_{AP} = 5$) are extracted at 7 vertical levels, yielding tensors of shape $(6 \times 4 \times 7 \times 5 \times 5)$ for wind and $(6 \times 5 \times 7 \times 5 \times 5)$ for rainfall. Temporal features in $\mathbf{C}_\mathrm{k}$ include cyclic encodings of hour and day along with station metadata (latitude, longitude, altitude) as explained in Baggio et al. (2025). The vector $\mathbf{D}_\mathrm{k}$ encodes the relative positions of the ten neighboring stations.

4.3.2 Neural network model architecture

The proposed ANN architecture follows the design introduced in Baggio et al. (2025), with minor adaptations. Each subtensor of $\mathbf{X}^S_{d,t}$ is processed by a dedicated branch tailored to its structure. Station-level time series $\mathbf{GS}^s_{d,t}$ are modeled through stacked

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

LSTM layers to capture temporal dependencies, while the spatiotemporal tensors $\mathbf{AR}^{s}_{d,t}$ and $\mathbf{AP}^{s}_{d,t}$ are first processed by convolutional layers to extract spatial features and subsequently passed to LSTM layers to encode their temporal evolution. In all branches, encoding of contextual features by fully connected layers are concatenated to the predictors prior to the recurrent layers. These context representations are produced by shallow fully connected networks (ContextEncoder) consisting of two layers of size $n_{enc} = 16$. The hidden dimension of all LSTM layers is controlled by parameter u_{LSTM}, set to $u_{LSTM} = 64$. Convolutional layers use kernel size 2 and stride equal to 1 (2D branch) or 2 (3D branch), without padding. For full architectural details we refer to Baggio et al. (2025). The outputs of the three branches are concatenated and passed through an additional dense block before the final prediction layers. For the classification model M_1, a sigmoid activation is applied to the final layer to produce occurrence probabilities. As predictions are issued simultaneously for six lead times, the output lies in $\mathbb{R}^6$. For the probabilistic model M_2, the dense representation is mapped to three distributional parameters via separate linear layers with suitable activation functions to enforce parameter constraints. Since each of the six lead times is associated with three parameters, the output lies in $\mathbb{R}^{6\times 3}$.

4.3.3 Dataset split

Time-series dataset splitting requires balancing sample independence with representative seasonal coverage to prevent data leakage from autocorrelated samples, while ensuring subsets share similar probability distributions (Schultz et al., 2021). Given our limited three-year data span (2016–2018), we adopted a balanced approach: all forecasting tasks and labels corresponding to the same calendar day, which exhibit the strongest autocorrelation, were strictly assigned to the same subset. This same-day constraint, combined with a data cutoff at 17:00 UTC, provides a natural temporal separation between consecutive days that mitigates leakage while preserving seasonal variability. To implement this, the complete pool of potential calendar days was randomly partitioned into training (85%), validation (10%), and test (5%) subsets, which were then intersected with the effectively available data. As reported in Table 1, the final dataset comprises approximately 4.1×10^6 total keys $\mathtt{k}$, distributed as 3.5×10^6 keys for training, 4×10^5 keys for validation, and 2×10^5 keys for testing.

4.3.4 Hyperparameter selection and model training

The network presented above contains a total of around 2.7×10^5 trainable parameters, which is relatively small by modern standards. The model was trained using the Adam optimizer. To mitigate potential overfitting, we adopted an early stopping strategy based on the validation loss (with a patience parameter of 10 for wind speed, 20 for cumulative rainfalls). The remaining hyperparameters were selected based on prior experience and are reported in Table 2. With this setup, model training takes about 3 to 4 minutes for epoch on a single Nvidia Tesla V100 GPU. Considering that the very first epoch takes approximately twice this time due to initialization overhead and that models takes between 11 to 25 epochs to converge, training takes less than 2 hours long. It is important to emphasize a fundamental difference between the two approaches: while M_2, which models the full predictive distribution, relies on a single model for all considered thresholds (Eqs. (31), (32)), strategy M_1, based on binary classification, requires training a separate model for each threshold. After training, inference from the ANN is very fast, so that forecasts can be issued in a matter of seconds (see Baggio et al. (2025) for details). Since we did not perform an extensive

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Table 2. training hyperparameters used in the ANN.

Training hyperparameters	
Learning rate	0.001
LSTM dropout level	0.02
Batch size (training)	512
Early stopping patience	20 (wind), 15 (rainfalls)

optimization of the hyperparameter space, the model parameters were kept fixed at standard baseline values, leading to highly consistent results between the validation and test splits. Accordingly, the scores presented in Subsection 4.4 are computed over both datasets simultaneously. This choice is motivated by our primary interest in the structural behavior of the curves as $p \to 0$, rather than in the absolute metric values themselves. In practice, evaluating solely on the test set produces noisier estimates because of its smaller sample size, while the relative ranking of the models remains unchanged across both subsets.

4.4 Applicaton results

In this Section we discuss the results obtained with the two modeling strategies $\mathcal{M}_1$ and $\mathcal{M}_2$ and discuss the reported evaluation metrics in light of the model introduced earlier. All the presented deterministic metrics and scores have been evaluated by using their own optimal threshold $p^\star$. This is known a priori for PSS, while for HSS and CSI it was obtained by evaluating a regularly spaced set of possible thresholds using a dedicated automated procedure. We report results for two forecast horizons, $h = 1$ and $h = 6$; intermediate horizons exhibit similar behaviour and tend to fall between these two cases. As expected, forecast skill progressively degrades as the lead time increases. This is consistent with our hypothesis in Section 3 presenting the noise-to-signal ratio ρ as an increasing function of forecast horizon. For the sake of clarity, all plots are showcased using a logarithmic scale on p.

4.4.1 Hourly wind speed

In Figure 4 values of BSS, PSS and the PSS ratio are shown. In line with what observed for the toy generative model, metrics associated with model M_2 are consistently better than the ones obtained with the classification approach M_1. Moreover, the overall trend of BSS and PSS for $p \to 0$ reflects what shown in Figure 2. In particular PSS_2 increases steadily with decreasing p, while PSS_1, though behaving similarly for intermediate values of p, sharply deteriorates when $p \to 0$ (Figure 4(c)). This behaviour is reflected in the trend displayed by ratio $\text{PSS}_1/\text{PSS}_2$, whose value is near to 1 for intermediate values of p but decreases as $p \to 0$, thus reflecting what predicted by the model curve (Figure 2(c)). This means that model M_2 maintains substantially higher discrimination ability in the rare-event regime, whereas M_1 exhibits a marked degradation. Similar behaviour is observed at all forecast horizons, though skill degrades with increasing h. The trend of Brier Skill Score BSS with decreasing p is in agreement with the model behaviour and shows a faster degradation of BSS_1 with respect to BSS_2 (Figures 4(a) and 2(a)). Moreover, when considering the BSS decomposition (14) (not shown) we observed that model M_2 performs

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

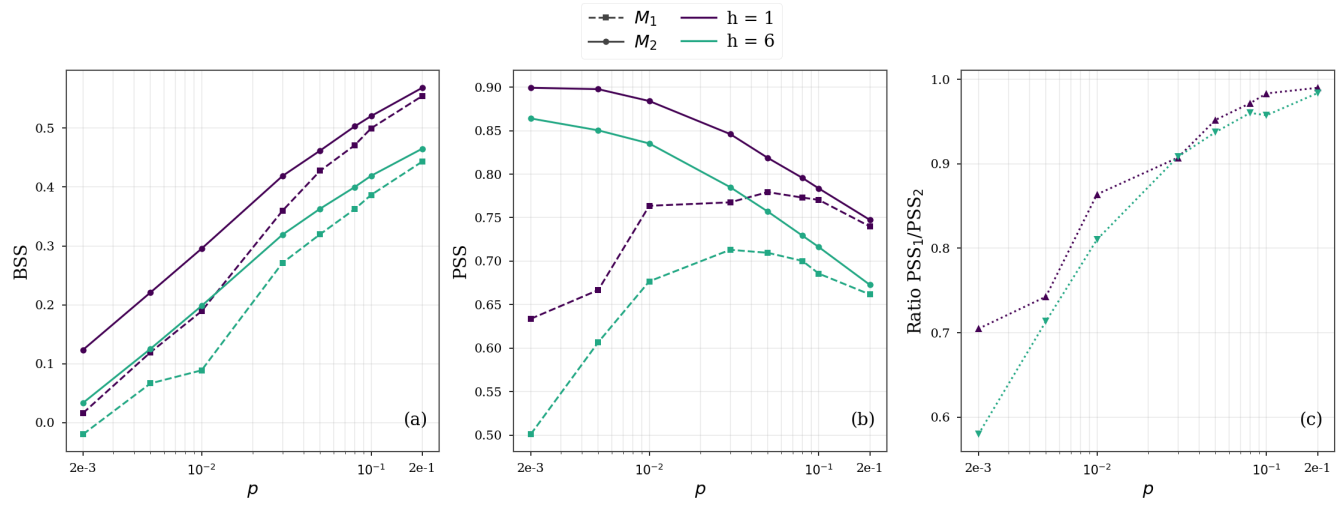

Figure 4. BSS (panel (a)) PSS (panel (b)) and its ratio $\mathrm{PSS}_1/\mathrm{PSS}_2$ (panel (c)) for **hourly wind speed** forecasts are shown for the two models M_1 (symbols (■) and dashed lines) and M_2 (symbols (●) and solid lines). Two different forecast horizons are highlighted: $h = 1$ h (in violet) and $h = 6$ h (green).

better both in terms of reliability and resolution. Despite the superiority of M_2, both models present good levels of calibration, as the term $\frac{Rel}{U}$, even if increasing rapidly with $p \to 0$, remains relatively well controlled. This is likely due to the use of proper scoring rules during training (binary cross-entropy for M_1 and negative log-likelihood for M_2). For completeness, we also

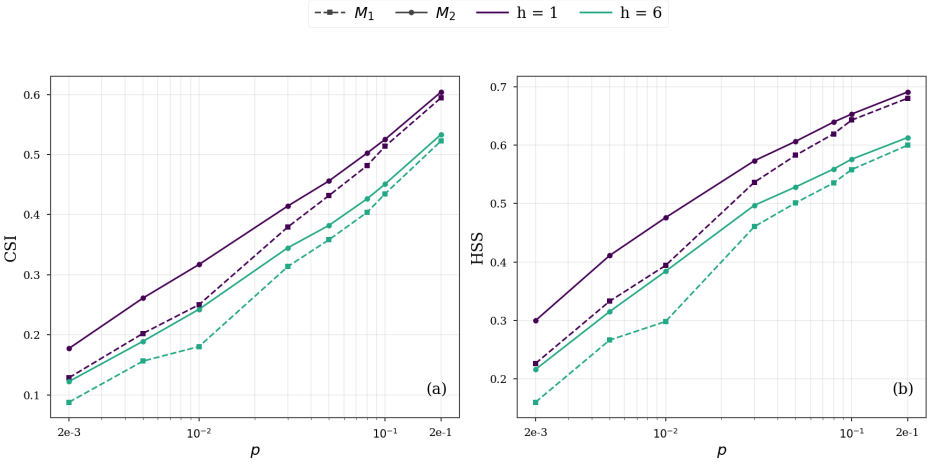

Figure 5. CSI (panel (a)) and HSS (panel (b)) relative to **hourly wind speed** forecasts are displayed for models M_1 (symbols (■) and dashed lines) and M_2 (symbols (●) and solid lines). Two different forecast horizons are highlighted: $h = 1$ h (in violet) and $h = 6$ h (green).

report the values of the CSI (Eq. (12)) and HSS (Eq. (11)), which confirm that the probabilistic model M_2 outperforms M_1 in all considered cases, (Figure 5). The evolution of CSI and HSS with respect to p is shown to provide an overall view of their

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

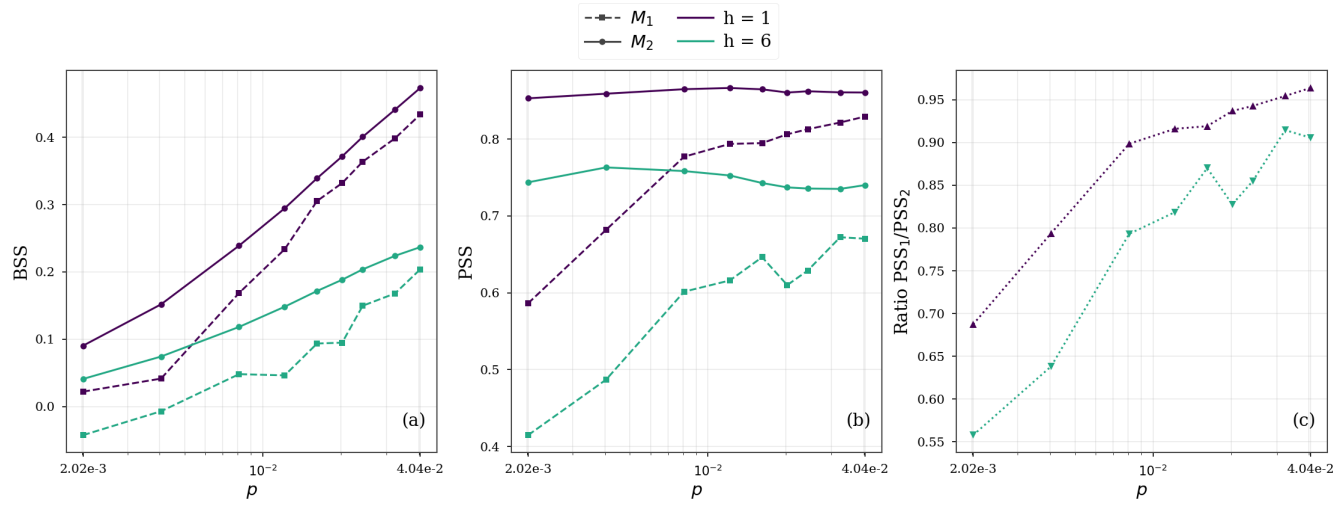

Figure 6. BSS (panel (a)) PSS (panel (b) and ratio $\mathrm{PSS}_1/\mathrm{PSS}_2$ (panel (c)) for **hourly accumulated rainfall** are shown for the two models M_1 (symbols (■) and dashed lines) and M_2 (symbols (●) and solid lines). Two different forecast horizons are highlighted: $h = 1$ h (in violet) and $h = 6$ h (green).

Table 3. AUC and LS scores for **hourly wind speed** forecasts at lead times $h = 1$ and 6 h, and probabilities $p = 0.05$ and 0.005.

Model	p	AUC		LS	
		1 h	6 h	1 h	6 h
$\mathcal{M}_1$	0.05	0.958	0.934	0.095	0.129
	0.005	0.963	0.955	0.019	0.024
$\mathcal{M}_2$	0.05	0.965	0.943	0.085	0.111
	0.005	0.982	0.969	0.010	0.014

behaviour. However, these metrics are not suitable for objective comparison across different base rates, as changes in event frequency affect the attainable range of these scores independently of the intrinsic discrimination ability of the model. Finally values of the AUC and LS (Eq. (15)) are reported in Table 3 for $p = 0.05, 0.005$ (mind that values of the logarithmic score are directly comparable only for equal values of p). It is possible to notice than for all the metrics considered, model M_2 performs better than M_1 and that this relative advantage tends to become more pronounced for decreasing values of p.

4.4.2 Hourly accumulated rainfall

The analysis presented in the case of hourly wind speed is now extended to accumulated rainfall.

The behaviour of BSS, PSS and ratio $\mathrm{PSS}_1/\mathrm{PSS}_2$ with $p \to 0$ are displayed in Figure 6. Looking at the PSS, showcased in 6(b), it is possible to note that while PSS_1 always decreases as $p \to 0$, PSS_2 remains almost constant. This differs from

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Table 4. AUC and LS for **hourly accumulated rainfall** forecasts at lead times $h = 1$ and 6 h, and probabilities $p \approx 0.04$ and 0.004.

Model	p	AUC		LS	
		1 h	6 h	1 h	6 h
$\mathcal{M}_1$	0.04	0.971	0.934	0.075	0.104
	0.004	0.954	0.872	0.025	0.031
$\mathcal{M}_2$	0.04	0.975	0.942	0.069	0.093
	0.004	0.972	0.941	0.018	0.019

what observed in the case of wind (Figure 4(b)), but is somewhat expected since the considered values of p span a smaller range near $p \to 0$ where the increasing trend of PSS_2 suggested by the theoretical model becomes less pronounced.The ratio $\mathrm{PSS}_1/\mathrm{PSS}_2$ (panel (c) in Figure 6) confirms that PSS_1 is always worse than PSS_2. Moreover, as suggested by the model and already observed for wind, this gap becomes and more pronounced with decreasing p. The overall decreasing behaviour observed for the BSS curves (panel (a) in Figure 6) resembles what already seen for wind speed and is in line with model predictions of Section 3. That is, the probabilistic model M_2 consistently outperforms the classification approach M_1. Results are in line with what described for wind even in terms of the reliability and resolution component, which we do not show, as model M_2 exhibits superior performance at a given p.

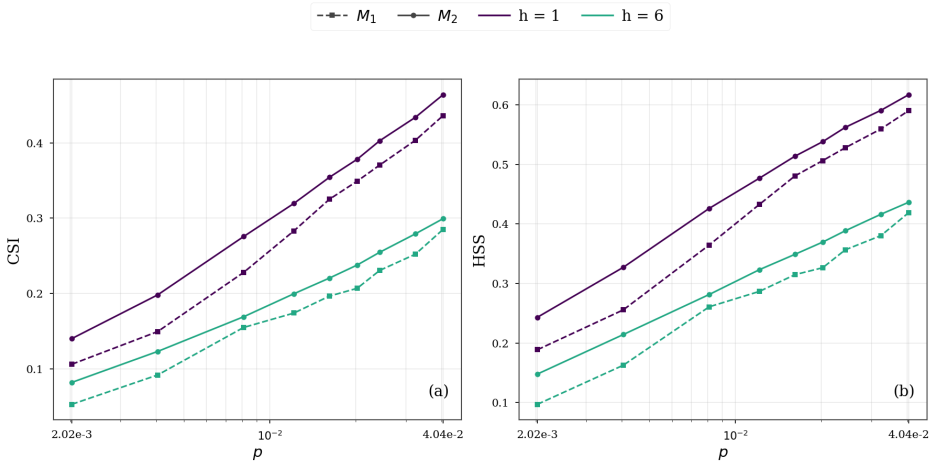

Figure 7. CSI (panel (a)) and HSS (panel (b)) relative to **hourly accumulated rainfall** forecasts are displayed for the two models M_1 (symbols (■) and dashed lines) and M_2 (symbols (●) and solid lines). Two different forecast horizons are highlighted: $h = 1$ h (in violet) and $h = 6$ h (green).

CSI and HSS are displayed in Figure 7. It is possible to remark that, at a given probability level p, M_2 is always better than M_1, consistently and for all considered thresholds. Moreover the results reported in Table 4, consisting in the values of AUC and LS for two forecasting horizons and two probability levels, also support superiority of M_2 over M_1. Finally, since one can legitimately question the effect of the distribution choice on the presented results for model M_2, different mixed

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

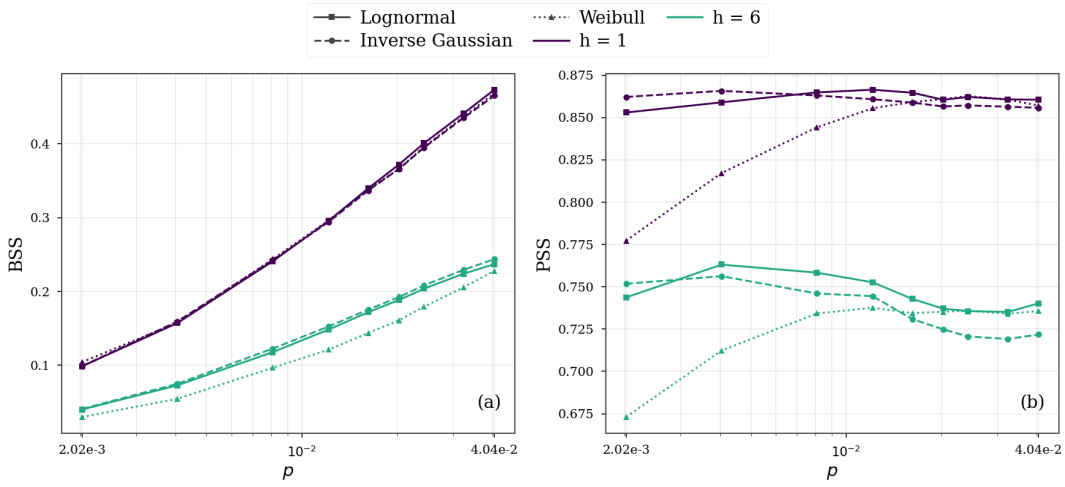

Figure 8. BSS (panel (a)), and PSS (panel (b)) for **hourly accumulated rainfall** are shown for different choices of the parametric distribution in model $\mathcal{M}_2$. All the three displayed families are mixed distributions of type (E3) with three parameters. More specifically, a mixed lognormal (E4) (symbols (■) and solid lines), a mixed inverse Gaussian (E5) (•) and dashed lines) and a mixed Weibull distribution (E6) (▲) and dotted lines) have been tested. Two different forecast horizons are highlighted: $h = 1$ h (in violet) and $h = 6$ h (green).

distributions have been tested. More specifically, the sensitivity of M_2's performance to the choice of parametric family has been investigated by replacing the mixed lognormal distribution with two alternative three-parameter families, namely the mixed inverse Gaussian (Eq. (E5)) and mixed Weibull (Eq. (E6)) distributions (see Appendix E2 for more details). The results, shown in Figure 8, indicate that the choice of parametric family does not impact significantly the results, even if the Weibull distribution appears somewhat less suitable.

4.5 Discussion

4.5.1 Comparative performance of wind and accumulated rainfall predictions

For both wind speed and hourly cumulative rainfall, the empirical results are broadly consistent with the theoretical model. However, rainfall forecasting appears intrinsically more challenging, yielding systematically lower classification and probabilistic scores. This suggests a higher level of intrinsic noise in the rainfall dataset for the chosen predictors and problem formulation. Consistently, rainfall validation performance saturates after only a few training epochs before deteriorating, indicating rapid exhaustion of the generalizable signal followed by overfitting to non-transferable variability. The dominant role of intrinsic noise is also clearly reflected in the forecast-horizon dependence. In the model of Section 3, the noise-to-signal ratio ρ acts as a proxy for the forecasting horizon. We observe that 'effective" ρ is increasing much more strongly for rainfall than for wind speed. Specifically, for rainfall, the transition from $h = 1$ to $h = 6$ roughly corresponds to a jump from $\rho = 1$ to $\rho = 10$ in the theoretical model, whereas for wind speed the observed difference between $h = 1$ and $h = 6$ is less pronounced than in the numerical experiments reported in Figure 2. Notably, for rainfall, the degradation of BSS as $p \to 0$ is slower for $h = 6$ than

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

for $h = 1$, which closely agrees with the theoretical model (Figures 6(a) and 2(a)), an effect not observed for wind. Likewise, the decline of PSS with increasing h is more pronounced for accumulated rainfall.

4.5.2 On the choice of the parametric distribution within the $\mathcal{M}_2$ approach

In forecasting extreme weather events, selecting a parametric family $\mathcal{F}$ (e.g., Weibull) over $\mathcal{G}$ (e.g., Gamma) presents a fundamental statistical challenge. Because underlying meteorological conditions continuously evolve, we only ever observe a single outcome for any specific atmospheric state. Identifying the "true" data-generating distribution is therefore an ill-posed problem; the actual conditional probability at a given time step is an inaccessible abstraction that can neither be directly observed nor asymptotically approached. Without access to this ground truth, the only operational proxy for reality is the aggregated evaluation of proper scoring rules such as the Negative Log-Likelihood or the Continuous Ranked Probability Score (CRPS), averaged over a heterogeneous test set. Consequently, model selection is characterized by empirical indistinguishability rather than a unique best-fit model.This structural equivalence allows some freedom in parametric family selection that can be based on theoretically desirable properties (like e.g. the range of the target variable or a spectific tail behavior compatible with the unconditional law) without sacrificing empirical accuracy in the bulk of the data. For sample sizes typical of climatological records, distinct parametric families can yield comparable results. In the context of surface wind speeds, for instance, Baggio and Muzy (2024) demonstrated that various predictive distributions (M-Rice, Weibull, and Gamma) display nearly indistinguishable probabilistic and deterministic scores when applied to meteorological series from the Netherlands and Corsica. Similarly, in our evaluation of intense hourly rainfall, Fig. 8 shows that three distinct statistical laws (Log-Normal, Inverse Gaussian, and Weibull) produce comparable performances across all standard scoring metrics. This empirical indistinguishability also provides valuable insight into the physical mechanics driving atmospheric extremes. Their predictability typically stems from large, resolved shifts in the "bulk" (the conditional mean and variance) of the distribution, rather than from atypical fluctuations drawn from the tail of a static climatology. When the dominant predictive signal is a massive displacement of the core probability mass, the specific parametric shape of the tail becomes a secondary factor. This dominance of bulk-shifting mechanisms naturally explains why the Peirce Skill Score (PSS) increases with the threshold, a structural behavior perfectly in line with our empirical observations for both intense rainfall and strong wind speeds.

5 Conclusions

In this study, we systematically compared two distinct paradigms for short-term probabilistic forecasting of atmospheric threshold exceedances at some location: direct binary classification, which frames exceedance as a Bernoulli outcome and full-distribution modeling, which estimates the conditional probability law of the target variable. On the theoretical ground, we considered a generative toy model inspired by the simple Gaussian model proposed in Lerch et al. (2017). By leveraging standard asymptotic theory, we derived analytical expressions for key evaluation metrics, including the Brier Score and the Peirce Skill Score, as functions of the extreme quantile probability level p. Our analysis, supported by both theoretical derivations and numerical simulations, reveals a striking contrast in behavior as $p \to 0$: for the full-distribution approach, predictive

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

skill, as measured by the PSS, is mathematically expected to improve. This is reminiscent of the fact that extreme events are driven by large excursions in the predictable component of the process. Conversely, the direct binary classification approach exhibits a maximum performance before inevitably declining as the threshold becomes more extreme, a limitation arising from the scarcity of positive examples in the training data for very high quantiles. These theoretically derived asymptotic behaviors were explicitly corroborated by our empirical validation using the MeteoNet dataset for southeastern France. The fundamental advantage of the full-distribution approach lies in its ability to mitigate the severe class imbalance that degrades the efficacy of direct binary classifiers in the deep tails. By modeling the complete conditional distribution, the framework successfully leverages abundant moderate and non-extreme observations to effectively learn the underlying scale and shape parameters. As validated on strong surface wind speeds and intense hourly rainfall, this approach translates to significantly sharper discrimination and better calibration for rare events.

While full-distribution modeling offers a robust framework for extreme weather prediction, several complex challenges remain for future research in statistical learning. One major issue is parametric misspecification and the accurate modeling of heavy tails. As illustrated by former results for wind speed and the specific examples we considered for hourly rainfalls and as also discussed in section 4.5.2, it appears that different choices of the probability distribution class provide comparable results. The success of distributional models intrinsically does not relies so much on the suitability of the chosen parametric family. Instead, our findings indicate that predictive skill for extreme exceedances is primarily derived from accurately capturing large, predictable shifts in the bulk properties of the conditional distribution. Because these extreme occurrences are dominantly driven by strong displacements of the core probability mass rather than by atypical, unpredictable anomalies drawn from a "static" climatological tail, the precise parametric shape of the predictive tail does not play a primary role. This question will be considered with more details in a future work where we will notably explore the need for a dynamic integration of Extreme Value Theory into deep distributional frameworks. Another interesting prospect concerns the current site-specific modeling framework that could be extended to continuous spatial domains. Utilizing advanced architectures like distributional U-Nets or Graph Neural Networks could allow for the joint modeling of spatial dependencies and multivariate extremes, yielding physically coherent, high-resolution probabilistic fields rather than isolated point forecasts. Finally, while hybrid deep learning and distributional architectures significantly improve predictive skill, operational forecasters require interpretable outputs to confidently issue life-saving warnings. Adapting explainable artificial intelligence methods for distributional regression outputs is an appealing next step. Understanding exactly which atmospheric covariates drive structural shifts in the predicted tail behavior will foster greater trust and facilitate the integration of these advanced statistical learning models into operational decision-making pipelines.

Code and data availability. The meteorological data used in this study originate from the MeteoNet database (Larvor and Berthomier, 2021), originally developed by Météo-France. The reference version of the dataset utilized in this work is hosted on the Harvard Dataverse and can be accessed at https://doi.org/10.7910/DVN/NCKRZ2. The complete Python source code for data preprocessing, model implementation, and analysis scripts, along with a minimal self-contained example dataset and an interactive Jupyter Notebook illustrating all aspects of

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

code usage (including data preparation, model training and visualization), is publicly available under the MIT License on Zenodo at https://doi.org/10.5281/zenodo.20327672 (Muzy and Baggio, 2026).

Author contributions. Roberta Baggio contributed to code development, experiment design, model execution, result analysis and manuscript writing. Jean-François Muzy contributed to the theoretical analysis, model setup, experiment design, code set up and development and manuscript writing.

Competing interests. The author declare that they have no competing interests.

Financial support. Both authors were supported in their research by the ANR research grant SAPHIR (ANR-21-CE04-0014).

Appendix A: Computation of prediction error in asymptotic regime

Let us estimate the prediction error associated with each prediction method when the number of observations N is large enough and we assume standard asymptotic regularity conditions (see, e.g., Vaart, 1998).

Let $\boldsymbol{\theta}$ denote the vector of model parameters (e.g., the collection of weights and biases in a neural network). For a given input X_t, the model output is written as $\widehat{z}(X_t;\boldsymbol{\theta}) \in \mathbb{R}^d$. This means that we have $\widehat{z} = \widehat{p}$ for M_1 and $\widehat{z} = \widehat{\mu}$ for model M_2. Let $\ell(Y, X;\boldsymbol{\theta})$ denote the *per-sample loss function* (e.g. $[Y - \widehat{\mu}(Z,\boldsymbol{\theta})]^2$ in the case of Gaussian log-likelihood or MSE) from which the (population) loss is computed as empirically as:

$$\mathcal{L}(\boldsymbol{\theta})_N = \frac{1}{N}\sum_{t=1}^{N} \ell(Y_t, X_t;\boldsymbol{\theta})\,.$$

We assume that there exists a unique (pseudo-)true parameter value $\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}} \mathbb{E}\,(\mathcal{L}_N(\boldsymbol{\theta}))$, such that, $\widehat{z}(X_t;\boldsymbol{\theta}_0)$ recovers the "true," i.e. data-generating, function $z_0(X_t)$ (namely $\mu(X_t)$ or $P(X_t)$ according to the model one considers). Even this assumption is unrealistic in practical situation (notably when misspecification induces at non-zero bias) it is an helpful framework to compare approaches M_1 and M_2. We suppose that standard regularity conditions are met and the the estimator $\widehat{\boldsymbol{\theta}}$ minimizing $\mathcal{L}_N(\boldsymbol{\theta})$ satisfies the usual asymptotic normality property

$$\sqrt{N}\,\big(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\big) \xrightarrow{d} \mathcal{N}\big(0, I_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)^{-1}\big), \tag{A1}$$

where $I_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) = \mathbb{E}\big[-\nabla^2_{\boldsymbol{\theta}}\ell(Y, X;\boldsymbol{\theta}_0)\big]$ is the Fisher information matrix. Applying the Delta method to the smooth mapping $\widehat{z}(X;\boldsymbol{\theta})$ then yields, for each fixed input X,

$$\sqrt{N}\,\big(\widehat{z}(X) - z_0(X)\big) \xrightarrow{d} \mathcal{N}\Big(0, G(X)^\top I_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)^{-1} G(X)\Big), \tag{A2}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

where $G(X) = \nabla_{\boldsymbol{\theta}}\, \widehat{z}(X;\boldsymbol{\theta}_0)$ is the Jacobian of the model output with respect to the parameters, evaluated at $\boldsymbol{\theta}_0$. Under these conditions, we can estimate the error associated with each method.

A1 Model M_2 in $\mathcal{M}_2$ class

Let us start with model $M_2 \in \mathcal{M}_2$ and let us write $\widehat{z}(X_t, \boldsymbol{\theta}) = \hat{\mu}(X_t) = M_2(X_t;\boldsymbol{\theta})$ where $M_2(Z;\boldsymbol{\theta})$, is defined in Eq. (7) and represents the non-linear function of parameter vector $\boldsymbol{\theta}$ used to infer $\widehat{\mu}_t$ (the single varying parameter of the Gaussian law) for an observed covariate X_t. Since the loss function is the Gaussian log-likelihood, the Fisher information matrix simply reads:

$$I_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) = \frac{1}{\sigma^2} J_{\boldsymbol{\theta}_0}$$

where σ^2 is the conditional variance of observations (the variance of the noise term ν_t in Eq. (17)) and

$$J_{\boldsymbol{\theta}_0} = \mathbb{E}_{X_t}\left(\nabla_{\boldsymbol{\theta}} M_2(X_t,\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \cdot \nabla_{\boldsymbol{\theta}} M_2(X_t,\boldsymbol{\theta})^\top\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) . \tag{A3}$$

It follows, from Eq. (A2), that $V_2(X_t)$, the asymptotic variance of $\widehat{\mu}(X_t) = M_2(X_t, \boldsymbol{\theta}_0)$, is simply:

$$V_2(X_t) = \frac{\sigma^2}{N} V_2'(X_t) \tag{A4}$$

where we have defined

$$V_2'(X_t) = \nabla_{\boldsymbol{\theta}} M_2(X_t,\boldsymbol{\theta})^\top\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\, J_{\boldsymbol{\theta}_0}^{-1} \nabla_{\boldsymbol{\theta}} M_2(X_t,\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} . \tag{A5}$$

Let us use again the Delta method for estimating the asymptotic variance of $\widehat{p}^{(2)}(X_t) = \Phi(\frac{\widehat{\mu}(X_t)-Q_p}{\sigma})$. Since $\Phi'(z) = \phi(z)$, we have

$$\mathrm{Var}\left(\widehat{p}^{(2)}(X_t)\right) \approx \frac{V_2'(X_t)}{N} \phi^2\left(\frac{Q_p - \mu(X_t)}{\sigma}\right) \tag{A6}$$

It results that the unconditional error on $\widehat{p}^{(2)}(X_t)$ corresponds to:

$$\mathcal{E}_2 = \frac{1}{N}\mathbb{E}_{X_t}\left[V_2'(X_t)\, \phi^2\left(\frac{Q_p - \mu(X_t)}{\sigma}\right)\right] . \tag{A7}$$

If one denotes $V_2'(\mu_t) = \mathbb{E}_{X_t}(V_2'(X_t)|\mu(X_t) = \mu_t)$, previous equation can be rewritten as $\mathcal{E}_2 = \frac{1}{N}\mathbb{E}_{\mu_t}\left[V_2'(\mu_t)\, \phi^2\left(\frac{Q_p-\mu_t}{\sigma}\right)\right]$. Since μ_t is supposed to be Gaussian random variable of zero mean and variance s^2, considering the definition of Q_p provided in Eq. (20), we obtain:

$$\mathcal{E}_2 = \frac{1}{N(2\pi)^{3/2}} \int V_2'(u) e^{-\frac{u^2}{2}} e^{-\frac{\left(-\sqrt{1+\rho^2}\Phi^{-1}(p)-u\right)^2}{\rho^2}}\, du \tag{A8}$$

where ρ^2 is the "noise-to-signal" ratio defined in (21).

If one wants a closed-form expression of $\mathcal{E}_2$, one needs to know the function $V_2'(\mu)$. The gaussian integral (A8) can be exactly computed for a wide variety of shapes $V_2'(\mu)$, e.g. polynomial, exponential, etc. The simplest expression is obtained when one neglects correlations and one assumes that $\mathbb{E}(V_2'(\mu)) \approx V_2$. In that case $\mathcal{E}_2$ reads:

$$\mathcal{E}_2 = \frac{V_2 \rho}{2\pi N\sqrt{2+\rho^2}} \exp\left(-\frac{1+\rho^2}{2+\rho^2}\left[\Phi^{-1}(p)\right]^2\right) \tag{A9}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Since, as $p \downarrow 0.$, $e^{-[\Phi^{-1}(p)]^2} \sim 4\pi p^2 \ln(1/p)$, one has finally:

$$\mathcal{E}_2 \underset{p\to 0}{\approx} \frac{K_2(\rho)}{N}\, p^{\frac{2+2\rho^2}{2+\rho^2}} \left[\ln(\frac{1}{p})\right]^{\frac{1+\rho^2}{2+\rho^2}}, \tag{A10}$$

where $K_2(\rho)$ is a constant that depends on ρ. We notably see that, up to logarithmic corrections:

$$\mathcal{E}_2 \underset{p\to 0}{\sim} \begin{cases} \frac{\rho\, p}{N} & \text{if } \rho \ll 1 \\ \frac{p^2}{N} & \text{if } \rho \gg 1. \end{cases} \tag{A11}$$

A2 Model M_1 in class $\mathcal{M}_1$

In the case of model $M_1 \in \mathcal{M}_1$, we have $\widehat{z}(X_t;\boldsymbol{\theta}) = \widehat{p}^{(1)}(X_t;\boldsymbol{\theta}) = M_1(X_t,\boldsymbol{\theta})$ and the loss function is simply given by expression (6). In order to simply upcoming developments, let us remark that $\widehat{p}^{(1)}(X_t;\boldsymbol{\theta}) = \mathrm{Sig}\left[L(X_t,\boldsymbol{\theta})\right]$ where $L(X_t,\theta)$ denotes the logit output corresponding to the model output just before entering in the sigmoid function, $\mathrm{Sig}(u) = \frac{1}{1+e^{-u}}$. For example, one can choose $L(X_t,\boldsymbol{\theta}) = M_2(X_t,\boldsymbol{\theta})$. Since the sigmoid function satisfies $0 \le \mathrm{Sig}(u) \le 1$ and $\mathrm{Sig}'(u) = \mathrm{Sig}(u)(1-\mathrm{Sig}(u))$, we have:

$$\frac{d\widehat{p}^{(1)}}{dL} = \widehat{p}^{(1)}(1-\widehat{p}^{(1)})\,. \tag{A12}$$

One can thus estimate $\nabla_{\boldsymbol{\theta}}\widehat{p}^{(1)}(X_t,\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$, as:

$$\nabla_{\boldsymbol{\theta}}\widehat{p}^{(1)}(X_t,\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = p_t(1-p_t)\,\nabla_{\boldsymbol{\theta}}L(X_t,\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{A13}$$

In order to estimate the Fisher information Matrix behavior, let us remark that, the case of M_1, the per-sample loss function involved with the BCE is

$$\ell(I_{t+h},X_t,\boldsymbol{\theta}) = -\left[I_{t+h}\ln(\widehat{p}_t^{(1)}) + (1-I_{t+h})\ln(1-\widehat{p}_t^{(1)})\right]$$

where $\widehat{p}_t^{(1)}$ stands for $\widehat{p}^{(1)}(X_t;\boldsymbol{\theta})$ and $I_{t+h} = I_{t+h}(Q_p)$ is defined in Eq. (1). One thus has, thanks to Eq. (A12),

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}\ell(I_{t+h},X_t,\boldsymbol{\theta}) &= \frac{\partial \ell}{\partial \widehat{p}_t^{(1)}}\cdot\frac{\partial \widehat{p}_t^{(1)}}{\partial L}\cdot\nabla_{\boldsymbol{\theta}}L(X_t,\boldsymbol{\theta}) \\ &= \left(-\frac{I_{t+h}}{\widehat{p}_t^{(1)}} + \frac{1-I_{t+h}}{1-\widehat{p}_t^{(1)}}\right)\cdot\left(\widehat{p}_t^{(1)}(1-\widehat{p}_t^{(1)})\right)\cdot\nabla_{\boldsymbol{\theta}}L \\ &= (\widehat{p}_t^{(1)} - I_{t+h})\cdot\nabla_{\boldsymbol{\theta}}L(X_t,\boldsymbol{\theta})\end{aligned}$$

Hence, because when $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\widehat{p}_t^{(1)} = p_t$, the Fisher information matrix becomes:

$$I(\boldsymbol{\theta}_0) = \mathbb{E}_{X_t}\Big(p_t(1-p_t)V(X_t)\Big) \tag{A14}$$

where, in order handle simple expressions, we define the matrix

$$V(X_t) = \nabla_{\boldsymbol{\theta}}L(X_t,\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\,.\,\nabla_{\boldsymbol{\theta}}L(X_t,\boldsymbol{\theta})^\top\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\,.$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

From Eq. (A2), using Eq. (A13), we thus compute the error associated with the asymptotic variance of M_1 output:

$$\mathrm{Var}\left(\widehat{p}^{(1)}(X_t)\right) = \frac{p_t^2(1-p_t)^2}{N} V_1(X_t) \tag{A15}$$

with:

$$V_1(X_t) = \nabla_{\boldsymbol{\theta}} L(X_t, \boldsymbol{\theta})^\top \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} I_{\boldsymbol{\theta}_0}^{-1} \nabla_{\boldsymbol{\theta}} L(X_t, \boldsymbol{\theta}) \big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} .$$

The final expression of the error thus becomes:

$$\mathcal{E}_1 = \mathbb{E}_{X_t}\left[\mathrm{Var}\left(\widehat{p}^{(3)}(X_t)\right)\right] = \frac{1}{N}\mathbb{E}_{X_t}\left[p_t^2(1-p_t)^2 V_1(X_t)\right] . \tag{A16}$$

If one neglects correlations between $p_t(1-p_t)$ and $V(X_t)$ in $I(\boldsymbol{\theta}_0)$ as given in Eq. (A14), then because $\mathbb{E}_{X_t}(p_t(1-p_t)) \approx p$ when $p \ll 1$, one has:

$$V_1(X_t) \approx p^{-1} V_1'(X_t) \tag{A17}$$

where $V_1'(X_t)$ is a scalar defined similarly as in Eq. (A5) that does not depend on p. We then have:

$$\mathcal{E}_1 \approx \frac{1}{pN}\mathbb{E}_{X_t}\left[p_t^2(1-p_t)^2 V_1'(X_t)\right] \tag{A18}$$

that is the analog of Eq. (A7). By defining $V_1'(\mu) = \mathbb{E}(V_1'(X_t)|\mu(X_t) = \mu)$ and assuming that μ is Gaussian random variable of zero mean and variance s^2, one gets:

$$\mathcal{E}_1 = \frac{1}{pN(2\pi)^{1/2}s} \int V_1'(\mu) e^{-\frac{\mu^2}{2s^2}} \Phi^2\left(\frac{\mu - Q_p}{\sigma}\right) \left[1 - \Phi\left(\frac{\mu - Q_p}{\sigma}\right)\right]^2 d\mu \tag{A19}$$

with $Q_p = -\sqrt{s^2+\sigma^2}\Phi^{-1}(p)$. By supposing, as previously, that $V_1'(\mu)$ independent of μ (or more specifically of p_t), by setting $V_1 = \mathbb{E}(V_1(\mu))$ and considering that, when $p \to 0$ ($Q_p \to \infty$), $\left[1 - \Phi^2\left(\frac{\mu - Q_p}{\sigma}\right)\right] \simeq 1$ one finally gets:

$$\mathcal{E}_1 \approx \frac{V_1}{pN(2\pi)^{1/2}s} \int e^{-\frac{\mu^2}{2s^2}} \Phi^2\left(\frac{\mu - Q_p}{\sigma}\right) d\mu . \tag{A20}$$

Such an expression can be exactly computed by Gaussian integration. It reads:

$$\mathcal{E}_1 \approx \frac{V_1}{pN} \Phi_2(\Phi^{-1}(p), \Phi^{-1}(p); r) . \tag{A21}$$

where $\Phi_2(q, q, r)$ is the cdf of the bivariate normal standard normal distribution with correlation coefficient r ($0 < r < 1$):

$$r = \frac{s^2}{s^2+\sigma^2} = \frac{1}{1+\rho^2}, \tag{A22}$$

ρ^2 being the noise-to-signal ratio defined in (21). From asymptotic behavior when $p \to 0$:

$$\Phi_2\big(\Phi^{-1}(p), \Phi^{-1}(p); r\big) \sim \frac{(4\pi)^{-\frac{r}{1+r}}}{\sqrt{1-r^2}} p^{\frac{2}{1+r}} \left[\ln(1/p)\right]^{-\frac{r}{1+r}} , \tag{A23}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

one obtains the behavior of the error of method $\mathcal{M}_1$ as a function of p:

$$\mathcal{E}_1 \underset{p\to 0}{\approx} \frac{K_1(\rho)}{N} p^{\frac{\rho^2}{2+\rho^2}} \left[\ln(1/p)\right]^{-\frac{1}{2+\rho^2}}, \tag{A24}$$

where $K_1(\rho)$ is the constant that depends on ρ. This equation can be directly compared with Eq. (A10). We notably see that, up to logarithmic corrections:

$$\mathcal{E}_1 \underset{p\to 0}{\sim} \begin{cases} \frac{1}{N} \text{ if } \rho \ll 1 \\ \frac{p}{N} \text{ if } \rho \gg 1. \end{cases} \tag{A25}$$

Appendix B: Relative Brier Score and Log-score performances of $\mathcal{M}_1$ and $\mathcal{M}_2$

We can exploit the conditional variance formulas established in Appendix A to derive the asymptotic behavior Brier and logarithmic scores associated with predictions of models M_1 and M_2. As before, we pay particular attention to the rare-event regime ($p \to 0$). We denote $\widehat{p}_t^{(k)}$ denote the estimator of the exceedance probability $p_t = \text{Prob}(Y_{t+h} > Q_p \mid X_t)$ with model M_k where $k = 1$ or 2. According to Eq. (13), the Brier Skill score BS_k of M_k, reads (by replacing the average over observations by the mathematical expectation over the joint law of $(\widehat{\boldsymbol{\theta}}, X_t, \nu_{t+h})$ or equivalently of $(\widehat{\boldsymbol{\theta}}, p_t, I_{t+h})$

$$\begin{aligned} BS_k &= \mathbb{E}\left(\widehat{p}_t^{(k)} - I_{t+h}\right)^2 = \mathbb{E}\left(\widehat{p}_t^{(k)} - p_t + p_t - I_{t+h}\right)^2 \\ &= \mathbb{E}_{\widehat{\boldsymbol{\theta}}}\mathbb{E}_{p_t}\left(\widehat{p}_t^{(k)} - p_t\right)^2 + \mathbb{E}_{p_t}\mathbb{E}_{I_{t+h}|p_t}\left(p_t - I_{t+h}\right)^2 \\ &= \mathcal{E}_k + \mathbb{E}_{p_t}\left(p_t(1-p_t)\right) \end{aligned}$$

where we have considered zero correlations between errors $(\widehat{p}_t^{(k)} - p_t)$ and $(p_t - I_{t+h})$ and used the fact the $\mathbb{E}_{I_{t+h}|p_t}(I_{t+h}) = p_t$ with $I_{t+h}^2 = I_{t+h}$. Thanks to Eqs. (22) and (23) we finally have the exact relationship between the Brier score of M_k and the previously computed asymptotic error $\mathcal{E}_k$ that reads:

$$BS_k = \mathcal{E}_k + p - \Phi_2\left(\Phi^{-1}(p), \Phi^{-1}(p), r\right) \tag{B1}$$

with $r = \frac{1}{1+\rho^2}$. It results from (14), that the Brier Skill of each model M_k is :

$$BSS_k = 1 - \frac{\mathcal{E}_k + p - \Phi_2\left(\Phi^{-1}(p), \Phi^{-1}(p), r\right)}{p(1-p)} \tag{B2}$$

Given the asymptotic behavior of Φ_2 when $p \downarrow 0$ (Eq. (A23), we have:

$$BSS_k \approx p^{\frac{\rho 2}{2+\rho^2}} - \frac{\mathcal{E}_k}{p}. \tag{B3}$$

The logarithmic score corresponds to the binary cross-entropy defined in Eq. (6) which expectation gives

$$\text{LS}_k = -\mathbb{E}_{\widehat{\boldsymbol{\theta}}}\mathbb{E}_{X_t}\mathbb{E}_{I_{t+h}|X_t}\left[I_{t+h}\ln\left(\widehat{p}_t^{(k)}\right) + (1 - I_{t+h})\ln\left(1 - \widehat{p}_t^{(k)}\right)\right]$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

which leads to:

$$\mathrm{LS}_k = -\mathbb{E}_{\widehat{\boldsymbol{\theta}}}\mathbb{E}_{X_t}\left[p_t \ln\left(\widehat{p}_t^{(k)}\right) + (1-p_t)\ln\left(1-\widehat{p}_t^{(k)}\right)\right]$$

where, as before, $\widehat{p}_t^{(k)} = \widehat{p}^{(k)}(X_t;\widehat{\boldsymbol{\theta}})$ and $I_{t+h} = I_{t+h}(Q_p)$ is defined in Eq. (1). In this context, in order to compare methods M_1 and M_2, one can evaluate

$$\Delta\mathrm{LS}_{1,2} = \mathrm{LS}_1 - \mathrm{LS}_2$$

where a positive value means that M_2 performs better than M_1. $\Delta\mathrm{LS}_{1,2}$ can be conveniently expressed as: $\Delta\mathrm{LS}_{1,2} = \mathcal{R}_1 - \mathcal{R}_2$ where "excess risk" $\mathcal{R}_k = D_{KL}\left(p_t||\widehat{p}_t^{(k)}\right)$ represents the Kullback-Leibler divergence with respect to the true probability, namely:

$$\mathcal{R}_k = \mathbb{E}_{\widehat{\boldsymbol{\theta}}}\mathbb{E}_{X_t}\left[p_t \ln\frac{p_t}{\widehat{p}_t^{(k)}} + (1-p_t)\ln\frac{1-p_t}{1-\widehat{p}_t^{(k)}}\right]. \tag{B4}$$

Assuming consistency of the estimators, ie., that $\widehat{p}_t^{(k)} = p_t + \delta_t$ with $\delta_t \ll 1$, we can perform a second-order Taylor expansion of the KL divergence around p_t and then, taking the expectation over the sampling distribution of the parameters $\widehat{\boldsymbol{\theta}}$ (which governs the variance of $\widehat{p}_t^{(k)}$), we obtain:

$$\mathcal{R}_k \approx \frac{1}{2}\mathbb{E}_{X_t}\left(\frac{\mathrm{Var}\left(\widehat{p}_t^{(k)}\right)}{p_t(1-p_t)}\right).$$

Former expressions of the variance (A15) and (A6), thus entail respectively:

$$\mathcal{R}_1 \approx \frac{1}{2Np}\mathbb{E}_{X_t}\left(p_t(1-p_t)V_1'(X_t)\right)$$

$$\mathcal{R}_2 \approx \frac{1}{2N}\mathbb{E}_{X_t}\left(\frac{V_2'(X_t)\phi^2\left(\frac{Q_p-\mu(X_t)}{\sigma}\right)}{\Phi\left(\frac{\mu(X_t)-Q_p}{\sigma}\right)\left[1-\Phi\left(\frac{\mu(X_t)-Q_p}{\sigma}\right)\right]}\right)$$

where we used the the fact that $p_t = \Phi\left(\frac{\mu(X_t)-Q_p}{\sigma}\right)$. If one focuses on rare-event regime where $p_t \ll 1$ and one supposes that $V_1'(X_t)$ and $V_2'(X_t)$ are independent from terms in p_t and ϕ^2, we obtain the asymptotic approximations:

$$\mathcal{R}_1 \approx \frac{1}{2Np}\mathbb{E}_{X_t}\left(p_t V_1'(X_t)\right) = \frac{V_1}{2N} \tag{B5}$$

$$\mathcal{R}_2 \approx \frac{V_2}{2N}\mathbb{E}_{X_t}\left(\frac{\phi^2\left(\frac{Q_p-\mu(X_t)}{\sigma}\right)}{\Phi\left(\frac{\mu(X_t)-Q_p}{\sigma}\right)}\right) \tag{B6}$$

where we used the definition of p, namely $\mathbb{E}(p_t) = p$. $\mathcal{R}_2$ can be rewritten as $\mathcal{R}_2 = \mathbb{E}_\mu[\sigma^2\phi^2(z)/\Phi(z)]$ with $z = (\mu - Q_p)/\sigma$. In the rare-event regime ($p \to 0$), since $Q_p \to \infty$, the argument z tends to $-\infty$. Using the Mill's ratio approximation $\Phi(z) \sim \phi(z)/|z|$, the integrand simplifies to a linear-Gaussian form $\sigma|z|\phi(z)$. As before, the resulting integral is evaluated using the saddle-point method, dominated by the contribution at $\mu_* = Q_p s^2/(s^2+\sigma^2)$. This leads to a scaling proportional to $p[\Phi^{-1}(p)]^2$.

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

By incorporating the refined asymptotic expansion of the quantile function and setting $\rho^2 = \frac{\sigma^2}{s^2}$, we obtain the final behavior $\mathcal{R}_2 \approx \frac{V_2\rho^2}{(1+\rho^2)N}\, p\, \ln\left(\frac{1}{p}\right)$ leading to our final estimation:

$$\Delta \mathrm{LS}_{1,2} \underset{p\to 0}{\approx} \frac{C_\rho}{2N}\left(K_\rho - p\ln\left(p^{-1}\right)\right) \tag{B7}$$

where $C_\rho = \frac{2V_2\rho^2}{(1+\rho^2)}$ and $K_\rho = \frac{V_1}{C_\rho}$. We see, provided p is small enough, $\Delta S_{1,2}$ is clearly positive and M_2 outperforms M_1.

Appendix C: Asymptotic Analysis of the Peirce Skill Score (PSS)

Let us perform the same kind of analysis for PSS score. We can remark, from the definition of (10), the averaged PSS_k for method $\mathcal{M}_k$ can be written as:

$$\begin{aligned}\mathrm{PSS}_k = \mathbb{E}_{\widehat{\boldsymbol{\theta}}}\Big[&\mathrm{Prob}\left(\widehat{I}^{(k)}_{t+h} = 1 \mid I_{t+h} = 1\right)\\ &-\mathrm{Prob}\left(\widehat{I}^{(k)}_{t+h} = 1 \mid I_{t+h} = 0\right)\Big]\end{aligned} \tag{C1}$$

where $k = 1, 2$ and $\widehat{I}^{(k)}_{t+h}$ is defined as in Eq. (4), namely, $\widehat{I}^{(k)}_{t+h} = 1$ if $\widehat{p}^{(k)}_t > p$ and $\widehat{I}^{(k)}_t = 0$ otherwise with $\widehat{p}^{(k)}_t = \widehat{p}^{(k)}(X_t, \widehat{\boldsymbol{\theta}})$. It results that:

$$\mathrm{Prob}\left(\widehat{I}^{(k)}_{t+h} = 1 \mid I_{t+h} = 1\right) = \int dx f_X(x|I_{t+h}=1)\mathbb{I}_{\{\widehat{p}^{(k)}(x,\widehat{\boldsymbol{\theta}})>p\}} = \int f_X(x|I_{t+h}=1)\mathbb{I}_{\{P(x)-\widehat{p}^{(k)}(x,\widehat{\boldsymbol{\theta}})<P(x)-p\}}$$

where we denoted, at fixed t, $f_X(x)$ the pdf of X_t and $P(X_t) = p_t$ is defined in (19). By Bayes rule, we have

$$f_X(x|I_{t+h}=1) = \frac{\mathrm{Prob}(I_{t+h}=1|\, X_t = x) f_X(x)}{\mathrm{Prob}(I_{t+h}=1)} = \frac{P(x) f_X(x)}{p}$$

and in the same way we have

$$f_X(x|I_{t+h}=0) = \frac{(1-P(x)) f_X(x)}{1-p}\,.$$

This thus entails:

$$\mathrm{Prob}\left(\widehat{I}^{(k)}_{t+h} = 1 \mid I_{t+h} = 1\right) = p^{-1}\int P(z) f_X(x)\mathbb{I}_{\{P(x)-\widehat{p}^{(k)}(x,\widehat{\boldsymbol{\theta}})<P(x)-p\}} = p^{-1}\mathbb{E}_\mu\left(\Phi\left(\tfrac{\mu-Q_p}{\sigma^2}\right)\mathbb{I}_{\{P(x)-\widehat{p}^{(k)}(x,\widehat{\boldsymbol{\theta}})<\Phi(\frac{\mu-Q_p}{\sigma^2})-p\}}\right)$$

where we made the change of variable $x \to \mu(x)$ and used Eq. (19) for the relationship between $P(x)$ and $\mu(x)$, Φ standing for the standard normal CDF. Finally, taking the expectation as respect to the law of $\widehat{\boldsymbol{\theta}}$ that is supposed to be asymptotically normal, we obtain:

$$\mathbb{E}_{\widehat{\boldsymbol{\theta}}}\left[\mathrm{Prob}\left(\widehat{I}^{(k)}_{t+h} = 1 \mid I_{t+h} = 1\right)\right] = p^{-1}\mathbb{E}_\mu\left[\Phi\left(\tfrac{\mu-Q_p}{\sigma^2}\right)\Phi\left(\frac{\Phi(\frac{\mu-Q_p}{\sigma^2})-p}{\sqrt{\mathrm{Var}\big(\widehat{p}^{(k)}(X_t)\big)}}\right)\right] = p^{-1}\mathbb{E}_{p_t}\left[p_t\Phi\left(\frac{p_t-p}{\sqrt{\mathrm{Var}\big(\widehat{p}^{(k)}(X_t)\big)}}\right)\right]$$

Along the same line one can establish that

$$\mathbb{E}_{\widehat{\boldsymbol{\theta}}}\left[\mathrm{Prob}\left(\widehat{I}^{(k)}_{t+h} = 1 \mid I_{t+h} = 0\right)\right] = \quad (1-p)^{-1}\mathbb{E}_{p_t}\left[(1-p_t)\Phi\left(\frac{p_t-p}{\sqrt{\mathrm{Var}\big(\widehat{p}^{(k)}(X_t)\big)}}\right)\right]$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

and therefore, from (C1), we obtain the following expression of PSS_k:

$$\mathrm{PSS}_k = \frac{1}{p(1-p)}\mathbb{E}_{p_t}\left[(p_t-p)\Phi\left(\frac{p_t-p}{\sqrt{\mathrm{Var}\left(\widehat{p}^{(k)}(X_t)\right)}}\right)\right]$$

From Eqs. (A15) and (A6), when $1-p_t \approx 1$, we get:

$$\mathrm{PSS}_1 = \tfrac{1}{p(1-p)}\mathbb{E}_{p_t}\left[(p_t-p)\Phi\left(\tfrac{\sqrt{p}(p_t-p)}{p_t\sqrt{V_1'}N^{-\frac{1}{2}}}\right)\right] \tag{C2}$$

$$\mathrm{PSS}_2 = \tfrac{1}{p(1-p)}\mathbb{E}_{p_t}\left[(p_t-p)\Phi\left(\tfrac{p_t-p}{p_t\sqrt{2\ln(p_t^{-1})V_2'}N^{-\frac{1}{2}}}\right)\right]. \tag{C3}$$

When N is very large, $\sigma_N^{(k)} \to 0$, so we can use the expansion of $\Phi(\frac{x}{\sigma})$ as $\sigma \to 0$:

$$\Phi\left(\frac{x}{\sigma}\right) = \mathcal{H}(x) + \frac{\sigma^2}{2}\delta'(x) + \mathcal{O}(\sigma^4),$$

where $\mathcal{H}(x)$ is the Heaviside function and $\delta'(x)$ is the derivative of the Dirac distribution. It results that, at fixed p, for large N, we have:

$$\mathrm{PSS}_1 \approx \tfrac{1}{p(1-p)}\left(\mathbb{E}_{p_t}(p_t-p)^+ - \tfrac{V_1'}{2N}pf_{p_t}(p)\right) \tag{C4}$$

$$\mathrm{PSS}_2 \approx \tfrac{1}{p(1-p)}\left(\mathbb{E}_{p_t}(p_t-p)^+ - \tfrac{V_2'p^2}{N}\ln(\tfrac{1}{p})f_{p_t}(p)\right) \tag{C5}$$

To evaluate the behavior of PSS_k for $p \ll 1$, we thus have to analyze the asymptotic limits of both $\mathbb{E}_{p_t}[(p_t-p)^+]$ and $f_{p_t}(p)$. Since $p_t = \Phi((\mu_t - Q_p)/\sigma)$ where $\mu_t \sim \mathcal{N}(0,s^2)$, a simple change of variable leads to the density:

$$f_{p_t}(y) = \frac{\sigma}{s}\frac{\phi\left(\frac{\sigma\Phi^{-1}(y)+X_p}{s}\right)}{\phi(\Phi^{-1}(y))}. \tag{C6}$$

Evaluating this density at the boundary $y=p$ using the asymptotic relation $\Phi^{-1}(p)^2 \sim 2\ln(1/p)$ leads a power-law behavior:

$$f_{p_t}(p) \propto \exp\left(\gamma\frac{\Phi^{-1}(p)^2}{2}\right) \sim p^{-\gamma}, \tag{C7}$$

where the scaling exponent is bounded by $\gamma = \frac{2\sigma}{\sqrt{s^2+\sigma^2}+\sigma} \in (0,1)$. The expectation $\mathbb{E}[(p_t-p)^+] = \int_p^1 (y-p)f_{p_t}(y)dy$ can be mapped to the original distribution of $\mu_t \sim \mathcal{N}(0,s^2)$. By determining the positivity threshold of the integrand, $L = \Phi^{-1}(p)(\sigma - \sqrt{s^2+\sigma^2})$, and applying the bivariate identity $\int \phi(z)\Phi(az+b)dz = \Phi_2(z, \frac{b}{\sqrt{1+a^2}}; \frac{-a}{\sqrt{1+a^2}})$, the integral evaluates to:

$$\mathbb{E}[(p_t-p)^+] = p\Phi\left(\frac{L}{s}\right) - \Phi_2\left(\frac{L}{s}, \Phi^{-1}(p); \frac{-s}{\sqrt{s^2+\sigma^2}}\right).$$

In the limit $p \ll 1$, this yields the leading-order behavior $\mathbb{E}[(p_t-p)^+] \approx P_r p$, where P_r depends on the correlation parameter $r = -s/\sqrt{s^2+\sigma^2}$. Substituting these asymptotic limits back into Eqs. (C4), (C5) and dropping minor logarithmic corrections directly yields the final PSS scaling rules:

$$\mathrm{PSS}_1 \approx 1 - K_r p^{\kappa^2} - \frac{C_1}{N}p^{-\gamma} \tag{C8}$$

$$\mathrm{PSS}_2 \approx 1 - K_r p^{\kappa_2} - \frac{C_2}{N}p^{1-\gamma} \tag{C9}$$

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Consequently, when N is large relative to $Kp^{-\gamma}$, the performance ratio simplifies to:

$$\frac{\text{PSS1}}{\text{PSS2}} \approx 1 - \frac{C_p}{N} p^{-\gamma} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{C10}$$

with $C_p \sim C_1 p^{-\gamma}(1 - C_2 p)$. This formalizes why M_2 achieves a superior PSS over M_1 at small p, while both safely converge to 1 as $N \to \infty$.

Appendix D: A toy model: the weighted harmonic model

Let Z be a vector of random latent factors defined as:

$$X_t \in \mathbb{R}^{N \times d}, \quad X_t \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)\ .$$

This means that, at each time $t \in [0, N-1]$, $X_t = (X_{t,1}, \ldots, X_{t,d})$ is sampled independently from a standard normal distribution $\mathcal{N}(0, I_d)$. The model generates a scalar signal μ_t as a weighted sum of centered harmonic transformations of these factors:

$$\mu_t = \sum_{k=1}^{d} \left[w_k^{(c)} \left(\cos(X_{t,k}) - e^{-1/2} \right) + w_k^{(s)} \sin(X_{t,k}) \right]. \tag{D1}$$

Here, the constant $e^{-1/2}$ ensures that the cosine terms have zero mean. The coefficients $w_k^{(c)}$ and $w_k^{(s)}$ are fixed model parameters (frozen randomness). They are initialized by drawing from a standard normal distribution and then rescaled by a global factor λ to ensure that the theoretical variance of μ_t matches the target parameter s^2.

Appendix E: Probability distributions for rainfall and wind speed

E1 The M-Rice probability distribution

In (Baïle et al., 2011), the Rice probability distribution (which corresponds to the norm of a two dimensional random vector which components are 2 independent Gaussian random variables of mean μ_1 and μ_2 and of same variance σ^2) has been extended to "Multifractal Rice" (M-Rice) distribution that accounts for the situation when, as observed in turbulence models, this variance σ^2 is itself stochastic with a log-normal distribution. The M-Rice distribution involves 3 parameters, namely the two Rice parameters coming from from Gaussian law $\nu = \sqrt{\nu_1^2 + \nu_2^2}$ and σ^2 and a supplementary parameter, denoted as λ^2 associated with the variance of the log-normal law. This parameter is referred to, in the literature on turbulence, as the "intermittency coefficient" (Frisch, 1995). The M-Rice probability density function (PDF) is then:

$$f_{\text{MR}}(y) = \frac{1}{\sqrt{2\pi\lambda^2}} \int e^{-\frac{\omega^2}{2\lambda^2}} \frac{y}{e^{2\omega}\sigma^2} e^{-\frac{y^2+\nu^2}{2e^{2\omega}\sigma^2}} I_0\left(\frac{y\nu}{e^{2\omega}\sigma^2}\right) d\omega\ .$$

where $I_0(z)$ is the order zero modified Bessel function of the first kind. As advocated in (Baggio and Muzy, 2024), this last formula can be fastly evaluated using the a simple Gauss-Hermite quadrature. The M-Rice cumulative distribution function (CDF) or the mean value function can also be obtained along the same way. For the latter, since for a Rice law of parameter ν

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

and σ^2, the mean value is $\mu_{\mathrm{R}} = \sigma\sqrt{\frac{\pi}{2}}\, L_{\frac{1}{2}}\left(-\frac{\nu^2}{2\sigma^2}\right)$, where $L_{\frac{1}{2}}$ stands for the order $\frac{1}{2}$ Laguerre polynomial, the mean value of a M-Rice distribution reads:

$$\mu_{\mathrm{MR}}(\nu,\sigma,\lambda^2) \simeq \frac{\sigma}{\sqrt{2}} \sum_{i=1}^{n} w_i\, L_{\frac{1}{2}}\left(-\frac{e^{2\sqrt{2}\lambda y_i}\nu^2}{2\sigma^2}\right) \tag{E1}$$

with $w_i = \frac{2^{n-1}n!\sqrt{\pi}}{n^2[H_{n-1}(y_i)]^2}$ for a quadrature order n. For the purpose of this paper, we chose, $n = 11$.

E2 Mixed distributions for rainfall

Mixed distributions, that is, combining a discrete with a continuous part, are common in the statistical modelling of rainfalls (Kedem et al., 1990). Indeed, there is a finite probability mass concentrated in $X = 0$ (the probability that does not rain at all). The continuous part, which models distribution of rain event only, is observed to be highly non-symmetrical and skewed towards high intensity events, so that two common choices are the mixed lognormal and the mixed gamma distributions (Cho et al., 2004). Considering this, we consider mixed distributions of the general form

$$P(X = 0) = 1 - p_{wet}, \qquad P(X > 0) = p_{wet}, \tag{E2}$$

where, conditionally on $X > 0$, the rainfall intensity follows a continuous distribution with density $f_+(x;\theta)$. The resulting mixture distribution can be written as

$$f(x) = (1 - p_{wet})\,\delta_0(x) + p_{wet}\, f_+(x;\theta)\,\mathbf{1}_{\{x>0\}}, \tag{E3}$$

where δ_0 denotes the Dirac mass at zero and θ represents the parameters of the positive component. Three different distributions $f_+(x;\theta)$ have been tested throughout this work, all characterized by two parameters. Meaning that Eq. (E3) has 3 parameters in total (the probability of rainfall occurrence p_{wet} and two additional parameters). The tested distributions $f_+(x;\theta)$ are briefly discussed below:

Mixed lognormal distribution

$$f_+(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\log x - \mu)^2}{2\sigma^2}\right),\ x > 0. \tag{E4}$$

where μ controls the central tendency of positive rainfall amounts on the logarithmic scale and σ^2 governs dispersion and tail heaviness.

Mixed inverse Gaussian distribution

$$f_+(x) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right),\ x > 0. \tag{E5}$$

where $\mu > 0$ is the mean of the positive rainfall component and $\lambda > 0$ the shape parameter controlling dispersion and tail behavior. The inverse Gaussian distribution class has notably been shown to account very well for monthly cumulated rainfalls in Sukrutha et al. (2018).

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Mixed Weibull distribution

$$f_+(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right],\ x > 0, \tag{E6}$$

where λ is a scale parameter regulating the magnitude of rainfall and k controls the shape of the distribution and tail behavior. The Weibull distribution has been used to model rainfall accumulation in several research works, such as Wilks (1989); Olivera and Heard (2019) and more recently Marra et al. (2023).

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

References

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

Ayzel, G., Heistermann, M., and Winterrath, T.: Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainymotion v0. 1), Geoscientific Model Development, 12, 1387–1402, 2019.

Bader, B., Yan, J., and Zhang, X.: Automated threshold selection for extreme value analysis via ordered goodness-of-fit tests with adjustment for false discovery rate, 2018.

Baggio, R. and Muzy, J.-F.: Improving probabilistic wind speed forecasting using M-Rice distribution and spatial data integration, Applied Energy, 360, 122 840, https://doi.org/10.1016/j.apenergy.2024.122840, 2024.

Baggio, R., Pujol, K., Pantillon, F., Lambert, D., Filippi, J.-B., and Muzy, J.-F.: Local wind speed forecasting at short time horizons relying on both Numerical Weather Prediction and observations from surrounding station, arXiv preprint arXiv:2503.18797, 2025.

Baïle, R., Muzy, J. F., and Poggi, P.: An M-Rice wind speed frequency distribution, Wind Energy, 14, 735–748, https://doi.org/10.1002/we.454, 2011.

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

Baïle, R., Muzy, J. F., and Poggi, P.: Short-term forecasting of surface layer wind speed using a continuous random cascade model, Wind Energy, 14, 719–734, https://doi.org/10.1002/we.452, 2011.

Beauchemin, S. S. and Barron, J. L.: The computation of optical flow, ACM computing surveys (CSUR), 27, 433–466, 1995.

Bojinski, S., Blaauboer, D., Calbet, X., De Coning, E., Debie, F., Montmerle, T., Nietosvaara, V., Norman, K., Bañón Peregrín, L., Schmid, F., et al.: Towards nowcasting in Europe in 2030, Meteorological applications, 30, e2124, 2023.

Bouallègue, Z. B., Clare, M. C. A., Magnusson, L., Gascón, E., Maier-Gerber, M., Janoušek, M., Rodwell, M., Pinault, F., Dramsch, J. S., Lang, S. T. K., Raoult, B., Rabier, F., Chevallier, M., Sandu, I., Dueben, P., Chantry, M., and Pappenberger, F.: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context, Bulletin of the American Meteorological Society, 105, E864 – E883, https://doi.org/10.1175/BAMS-D-23-0162.1, 2024.

Bouttier, F. and Marchal, H.: Probabilistic short-range forecasts of high-precipitation events: optimal decision thresholds and predictability limits, Natural Hazards and Earth System Sciences, 24, 2793–2816, https://doi.org/10.5194/nhess-24-2793-2024, 2024.

Cho, H.-K., Bowman, K. P., and North, G. R.: A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission, Journal of Applied meteorology, 43, 1586–1597, 2004.

Coles, S., Bawa, J., Trenner, L., and Dorazio, P.: An introduction to statistical modeling of extreme values, vol. 208, Springer, 2001.

Dutot, A.-L., Rynkiewicz, J., Steiner, F. E., and Rude, J.: A 24-h forecast of ozone peaks and exceedance levels using neural classifiers and weather predictions, Environmental Modelling & Software, 22, 1261–1269, 2007.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Carver, R., Andrychowicz, M., Hickey, J., et al.: Deep learning for twelve hour precipitation forecasts, Nature communications, 13, 5145, 2022.

Friederichs, P. and Thorarinsdottir, T. L.: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction, Environmetrics, 23, 579–594, 2012.

Frisch, U.: Turbulence. The legacy of AN Kolmogorov, Turbulence. The legacy of AN Kolmogorov, 1995.

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Glahn, H. R. and Lowry, D. A.: The use of model output statistics (MOS) in objective weather forecasting, Journal of Applied Meteorology and Climatology, 11, 1203–1211, 1972.

Gneiting, T. and Katzfuss, M.: Probabilistic forecasting, Annual Review of Statistics and Its Application, 1, 125–151, 2014.

Gneiting, T., Raftery, A. E., III, A. H. W., and Goldman, T.: Calibrated Probabilistic Forecasting Using EMOS and Minimum CRPS Estimation, Monthly Weather Review, 133, 1098–1118, https://doi.org/10.1175/MWR2904.1, 2005.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., and Aldrich, E.: Calibrated probabilistic forecasting at the stateline wind energy center: The regime-switching space–time method, Journal of the American Statistical Association, 101, 968–979, 2006.

Hess, R.: Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst, Nonlinear Processes in Geophysics, 27, 473–487, 2020.

Jolliffe, I. T., ed.: Forecast verification: a practitioner's guide in atmospheric science, Wiley, Chichester, repr edn., ISBN 978-0-471-49759-2, 2004.

Kaur, J., Parmar, K. S., and Singh, S.: Autoregressive models in environmental forecasting time series: a theoretical and application review, Environmental Science and Pollution Research, 30, 19 617–19 641, 2023.

Kedem, B., Chiu, L. S., and North, G. R.: Estimation of mean rain rate: Application to satellite observations, Journal of Geophysical Research: Atmospheres, 95, 1965–1972, 1990.

Lagerquist, R., McGovern, A., and Smith, T.: Machine learning for real-time prediction of damaging straight-line convective wind, Weather and Forecasting, 32, 2175–2193, 2017.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Battaglia, P., Vinyals, O., Stott, D., Pritzel, A., Kavukcuoglu, K., and Brandstetter, J.: GraphCast: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Larvor, G. and Berthomier, L.: Meteonet: An open reference weather dataset for ai by météo-france, in: American Meteorological Society Meeting Abstracts, vol. 101, pp. 1–ii, 2021.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T.: Forecaster's dilemma: extreme events and forecast evaluation, Statistical Science, pp. 106–127, 2017.

Leutbecher, M. and Palmer, T. N.: Ensemble forecasting, Journal of Computational Physics, 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, 2008.

Lorenz, E. N.: Deterministic Nonperiodic Flow, Journal of the Atmospheric Sciences, 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963.

Marra, F., Amponsah, W., and Papalexiou, S. M.: Non-asymptotic Weibull tails explain the statistics of extreme daily precipitation, Advances in Water Resources, 173, 104 388, 2023.

Mason, I.: On reducing probability forecasts to yes/no forecasts, Monthly Weather Review, 107, 207–211, 1979.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., and Williams, J. K.: Using artificial intelligence to improve real-time decision-making for high-impact weather, Bulletin of the American Meteorological Society, 98, 2073–2090, 2017.

Meinshausen, N. and Ridgeway, G.: Quantile regression forests., Journal of machine learning research, 7, 2006.

Murphy, A. H.: A new vector partition of the probability score, Journal of Applied Meteorology, 12, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2, 1973.

Muzy, J.-F. and Baggio, R.: saphir_predict, https://doi.org/10.5281/zenodo.20327672, 2026.

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Olivera, S. and Heard, C.: Increases in the extreme rainfall events: Using the Weibull distribution, Environmetrics, 30, e2532, 2019.

Pang, G., He, J., Huang, Y., and Zhang, L.: A binary logistic regression model for severe convective weather with numerical model data, Advances in Meteorology, 2019, 6127 281, 2019.

Park, Y., Maddix, D., Aubet, F.-X., Kan, K., Gasthaus, J., and Wang, Y.: Learning quantile functions without quantile crossing for distribution-free time series forecasting, in: International conference on artificial intelligence and statistics, pp. 8127–8150, PMLR, 2022.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., Hassanzadeh, P., Kashinath, K., and Anand, A.: FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators, npj Climate and Atmospheric Science, 7, 245, https://doi.org/10.1038/s41612-024-00834-8, 2024.

Pic, R., Dombry, C., Naveau, P., and Taillardat, M.: Distributional regression u-nets for the postprocessing of precipitation ensemble forecasts, Artificial Intelligence for the Earth Systems, 4, 240 067, 2025.

Pujol, K., Baggio, R., Lambert, D., Muzy, J.-F., Filippi, J.-B., and Pantillon, F.: Improving prediction of heavy rainfall in the Mediterranean with Neural Networks using both observation and Numerical Weather Prediction data, arXiv preprint arXiv:2503.24216, 2025.

Rasp, S. and Lerch, S.: Neural networks for postprocessing ensemble weather forecasts, Monthly Weather Review, 146, 3885–3900, 2018.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A. S., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arenas, A., and Pritzel, A.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z, 2021.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T.: DeepAR: Probabilistic forecasting with autoregressive recurrent networks, International journal of forecasting, 36, 1181–1191, 2020.

Schaumann, P., Hess, R., Rempel, M., Blahak, U., and Schmidt, V.: A calibrated and consistent combination of probabilistic forecasts for the exceedance of several precipitation thresholds using neural networks, Weather and Forecasting, 36, 1079–1096, 2021.

Schlosser, L., Hothorn, T., Stauffer, R., and Zeileis, A.: Distributional regression forests for probabilistic precipitation forecasting in complex terrain, The Annals of Applied Statistics, 13, 1564–1589, https://doi.org/10.1214/19-AOAS1247, 2019.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A, 379, 20200 097, https://doi.org/10.1098/rsta.2020.0097, 2021.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France Convective-Scale Operational Model, Monthly Weather Review, 139, 976 – 991, https://doi.org/10.1175/2010MWR3425.1, 2011.

Seneviratne, S. I., Zhang, X., Adnan, M., Badi, W., Dereczynski, C., Di Luca, A., et al.: Weather and Climate Extreme Events in a Changing Climate, p. 1513–1766, Cambridge University Press, 2023.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N.: Metnet: A neural weather model for precipitation forecasting, arXiv preprint arXiv:2003.12140, 2020.

Sukrutha, A., Dyuthi, S. R., and Desai, S.: Multimodel response assessment for monthly rainfall distribution in some selected Indian cities using best-fit probability as a tool, Applied Water Science, 8, https://doi.org/10.1007/s13201-018-0789-4, 2018.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics, Monthly Weather Review, 144, 2375–2393, 2016.

Tascikaraoglu, A. and Uzunoglu, M.: A review of combined approaches for prediction of short-term wind speed and power, Renewable and Sustainable Energy Reviews, 34, 243–254, 2014.

Vaart, A. W. v. d.: Asymptotic Statistics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1998.

https://doi.org/10.5194/egusphere-2026-3111
Preprint. Discussion started: 11 June 2026

Wilks, D. S.: Rainfall intensity, the Weibull distribution, and estimation of daily surface runoff, Journal of Applied Meteorology and Climatology, 28, 52–58, 1989.

Wilks, D. S.: Extending logistic regression to provide full-probability-distribution MOS forecasts, Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling, 16, 361–368, 2009.

975 Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.