Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions

Maxime Taillardat [a,b,*], Anne-Laure Fougères [c], Philippe Naveau [d], Raphaël de Fondeville [e]

[a] *CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France*
[b] *Météo-France, Toulouse, France*
[c] *Univ. Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France*
[d] *Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, IPSL & U Paris-Saclay, Gif-sur-Yvette, France*
[e] *Swiss Data Science Center, ETH Zürich and EPFL, Switzerland*

ARTICLE INFO

Keywords:
CRPS
Extreme events
Probabilistic forecasting
Scoring rules
Calibration
Verification

ABSTRACT

Verifying probabilistic forecasts for extreme events is a highly active research area because popular media and public opinions are naturally focused on extreme events, and biased conclusions are readily made. In this context, classical verification methods tailored for extreme events, such as thresholded and weighted scoring rules, have undesirable properties that cannot be mitigated, and the well-known continuous ranked probability score (CRPS) is no exception.

In this paper, we define a formal framework for assessing the behavior of forecast evaluation procedures with respect to extreme events, which we use to demonstrate that assessment based on the expectation of a proper score is not suitable for extremes. Alternatively, we propose studying the properties of the CRPS as a random variable by using extreme value theory to address extreme event verification. An index is introduced to compare calibrated forecasts, which summarizes the ability of probabilistic forecasts for predicting extremes. The strengths and limitations of this method are discussed using both theoretical arguments and simulations.

1. Introduction

By definition, the rarity of extreme events makes it difficult to issue relevant forecasts and performance assessments are even more challenging. In particular, the rarity of extreme events means that verification schemes must be built and understood in a probabilistic sense. The general framework for probabilistic forecast evaluation compares an observation y with a probabilistic forecast F represented by its cumulative distribution function (cdf). The framework also assumes that y is drawn from a random variable Y with cdf G. To better utilize forecasts, it is generally convenient and even recommended (Ferro & Stephenson, 2011) to further assume that the forecast F is calibrated (Dawid, 1984; Diebold, Gunther, & Tay, 1997), i.e., that the predictive distribution resembles the distribution of the observations given the information contained in the forecast. For a formal definition of auto-calibration (referred to as calibration in the following), we refer the reader to the studies by Tsyplakov (2011) and Strähl and Ziegel (2017) summarized in Appendix A.

Calibrated forecasts can generally be evaluated based on their sharpness, also called refinement by Winkler, Munoz, Cervera, Bernardo, Blattenberger, Kadane, et al.

[*] Corresponding author at: CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France.
E-mail address: maxime.taillardat@meteo.fr (M. Taillardat).

https://doi.org/10.1016/j.ijforecast.2022.07.003

(1996), which usually refers to their spread. This leads to the paradigm of "maximizing sharpness subject to calibration" introduced by Gneiting, Balabdaoui, and Raftery (2007) and later formally justified by Tsyplakov (2011).

Probabilistic forecasting has become increasingly popular in recent years in various fields, such as economics and finance (Galbraith & Norden, 2012), demography and social science (Raftery & Ševčíková, 2021), health (Henzi, Kleger, Hilty, Wendel Garcia, & Ziegel, 2021), energy (Hong, Pinson, Fan, Zareipour, Troccoli, & Hyndman, 2016), hydrology and hydraulics (Tiberi-Wadier et al., 2021). In this study, we focus on weather probabilistic forecasts (Leutbecher & Palmer, 2008). Indeed, probabilistic forecasts are now issued by most national weather services and F is known through a sample of finite size called an "ensemble" (e.g., see Zamo & Naveau, 2017). In this context, forecast verification is performed by computing scoring rules such as the continuous ranked probability score (CRPS) (Bröcker, 2012; Epstein, 1969; Hersbach, 2000)

$$\begin{aligned}\mathrm{CRPS}(F,y) &= \int_{-\infty}^{\infty}(F(x)-\mathbf{1}\{x\geq y\})^2\,dx,\\ &= \mathbb{E}_F|X-y|-\frac{1}{2}\mathbb{E}_F|X-X'|,\end{aligned}\qquad (1)$$

where $y \in \mathbb{R}$, and X and X' are independent random variables with the common cdf F. The CRPS is attractive because it does not require predictive densities, while it is inferred non-parametrically and it has a simple interpretation. The right-hand side of Eq. (1) decomposes the CRPS into a calibration and a sharpness term (Gneiting & Raftery, 2007) in order (alternative decompositions are also available; see Bessac and Naveau (2021), Taillardat, Mestre, Zamo, and Naveau (2016) and Appendix B).

Proper weighted scoring rules for extreme events forecast evaluations were introduced by Gneiting and Ranjan (2011) and Diks, Panchenko, and Van Dijk (2011). For a non-negative function $w(x)$, the weighted CRPS

$$\begin{aligned}\mathrm{wCRPS}(F,y) &= \int_{-\infty}^{\infty}(F(x)-\mathbf{1}\{x\geq y\})^2 w(x)\,dx, \qquad (2)\\ &= \mathbb{E}_F|W(X)-W(y)|-\frac{1}{2}\mathbb{E}_F|W(X)-W(X')|,\end{aligned}$$

with $W(x) = \int_{-\infty}^{x} w(t)dt$, aims to emphasize a region of interest, such as distributional tails. When w is continuous, an alternative expression of the weighted CRPS is available, as given by Appendix B. The choice of the weight function $w(x)$ is complex and it depends on the different stakeholders, such as forecast users and forecasters (e.g., see Ehm, Gneiting, Jordan, and Krüger (2016), Gneiting and Ranjan (2011), Patton (2014), Smith, Suckling, Thompson, Maynard, and Du (2015), Taillardat (2021b)). Even in the hypothetical case where $w(x)$ can be objectively defined, it is essential that the verification process is conducted based on the whole set of observations (Lerch, Thorarinsdottir, Ravazzolo, Gneiting, et al., 2017), but it is not clear whether the corresponding weighted CRPS correctly discriminates between two competitive forecasts with respect to extreme events.

In this study, we show that the expected weighted CRPS cannot discriminate forecasts with different extremal tail behaviors, which is a potentially redhibitory defect for extremal evaluation. To address this issue, we view the CRPS as a random variable, and its tail behavior is derived and compared to the tail regime of observations using extreme value theory (EVT) (e.g., see De Haan & Ferreira, 2007). This comparison is valid for forecasts that are known to be calibrated or that have been re-calibrated.

The remainder of this paper is organized as follows. In Section 2, we analyze the weighted CRPS with respect to the notion of tail equivalence, which is the main basis of EVT. In particular, we propose a benchmark for comparing the tail properties of forecast verification tools, thereby allowing us to identify the shortcomings of the CRPS and its weighted counterpart for scoring extreme events. In Section 3, we study the CRPS as a random variable and we make theoretical links between its tail behavior and the observational tail distribution. These mathematical connections allow us to propose and study a new index for assessing the capacity of calibrated probabilistic forecasts with respect to extreme events. The utility and possible disadvantages of this index and potential future research are discussed in Section 4. Index calculations were conducted with the R package extremeIndex (Taillardat, 2021a) in this study.

2. Limitations of the (w)CRPS as a proper scoring rule for extremes

2.1. Tail modeling using EVT

Based on the pioneering research of Gumbel (1935) and De Haan (1970), EVT provides a theoretically justified framework for modeling the tail of random variables, particularly excesses above a large threshold (e.g., see Beirlant, Goegebeur, Segers, Teugels, Waal, and Ferro (2004), Embrechts, Klüppelberg, and Mikosch (1997)). For any random variable X with cdf F, EVT models assume the existence of a domain of attraction, i.e., a positive auxiliary function b exists such that

$$\frac{\overline{F}\{u+xb(u)\}}{\overline{F}(u)} \longrightarrow \overline{H}(x)>0, \quad u\to x_F, \qquad (3)$$

where $\overline{F}=1-F$ corresponds to survival, also called the tail function, and $x_F = \sup\{x : F(x) < 1\}$ is the upper endpoint of F. Under condition (3), as noted $F \in \mathcal{D}(H)$, the Pickands–Balkema–de Haan theorem (De Haan, 1970; Pickands, 1975) establishes that H must belong to the family of generalized Pareto (GP) survival functions, i.e.,

$$\overline{H}_\gamma(x) = (1+\gamma x)^{-\frac{1}{\gamma}},$$

where $x \in \{x : 1+\gamma x > 0\}$. As a consequence, the GP tail, which is denoted by $\mathrm{GP}(\sigma,\gamma)$ in the following, appears to be the ideal candidate for approximating the survival function of excesses over a large threshold $u > 0$, i.e.,

$$\mathbb{P}(X-u\geq x|X>u) \approx \overline{H}_\gamma(x/\sigma) = \overline{H}_{\gamma,\sigma}(x) = \left(1+\frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}},$$

where $x \in \{x : 1+\gamma x/\sigma > 0\}$ and $\sigma > 0$. The GP family covers the three possible regimes of tail decay, which are determined by the value of the tail index γ: the decay is polynomial when $\gamma \neq 0$ and it has an upper bound when $\gamma < 0$. For $\gamma = 0$, the GP survival function becomes exponential, i.e., $\overline{H}_0(z) = e^{-z/\sigma}$.

2.2. Tail equivalence and proper scoring rules

The comparison of the tail behavior of two random variables, or equivalently their respective cdfs F and G, can be framed using the notion of tail equivalence.

Definition 1 (*Embrechts et al., 1997, Section 3.3*). Two random variables X and Y with cdfs F and G, respectively, are *tail equivalent* if they have equal upper endpoint $x_F = x_G = x_*$ and if their survival functions $\overline{F}$ and $\overline{G}$ satisfy

$$\lim_{x \to x_*} \frac{\overline{F}(x)}{\overline{G}(x)} = c \in (0, +\infty).$$

Tail equivalence can also be simply expressed as the equality of tail indexes. In terms of extremal forecast, it is expect that we should favor one of two forecasters that is tail equivalent to the observations, but this may be difficult in practice. For instance, consider two GP distributed random variables X_1 and X_2 with survival functions $\overline{H}_1(x)$ and $\overline{H}_{1+\epsilon}(x/\sigma)$ with $\sigma = (1+\epsilon)/(2^{1+\epsilon}-1)$. By construction, the medians of X_1 and X_2 are both equal to one, but their tail behaviors can differ widely even for small ϵ. The 100 year return level for X_1 is 99 whereas it is equal to 138 for X_2 with $\epsilon = 0.1$. Thus, if the precedent random variables represent water levels, a small difference of 0.1 in the tail index implies a difference of 39 meters, which would most likely lead to massive and destructive flooding.

This example illustrates how issuing forecasts with the correct tail regime, i.e., as close as possible to that observed, is a priority for extreme events and that a verification method should yield a forecast with a close if not equal tail regime. Ideally, the measure of forecast performance should give the distance but also the "direction", i.e., if the forecast is more likely to over- or under-estimate the high quantiles. Indeed, let $\gamma_G \in \mathbb{R}$ be the tail index of observations. If the forecast satisfies $\gamma_F > \gamma_G$, the forecast over-estimates the risk of producing a pessimistic or risk averse scenario. By contrast, $\gamma_F < \gamma_G$ yields an optimistic forecast by under-estimating the likelihood of extreme events.

The classical methods for forecast evaluation do not conserve tail equivalence, even when they are designed to focus on extreme events. For instance, for any positive η and observation distribution G, it is always possible to construct a non-tail equivalent cdf F such that

$$|\mathbb{E}_G(\text{wCRPS}(G, Y)) - \mathbb{E}_G(\text{wCRPS}(F, Y))| \leq \eta, \tag{4}$$

where the proof can be found in Appendix C. In particular, if $G \in \mathcal{D}(H_{\gamma_G})$, then for any arbitrary $\gamma_F \in \mathbb{R}$, it is possible to find $F \in \mathcal{D}(H_{\gamma_F})$ that satisfies Eq. (4). Thus, the CRPS is unable to properly discriminate forecasts with different tail regime because non-tail equivalent forecasts can perform almost equally well as the ideal forecast G. A detailed illustration of this result for GP forecasts is given in Appendix D. We also refer to Brehmer and Strokorb (2019), who obtained a more general result by proving that proper scoring rule expectations are not suitable for distinguishing tail properties (see Theorem 5.4 in their study).

2.3. A benchmark for assessing forecasts of extremes

Following Gneiting et al. (2007) and Strähl and Ziegel (2017), we propose a benchmark for assessing the behavior of forecast evaluation procedures with respect to tail regimes. The design employs a hierarchical model based on Gamma–exponential mixtures,

$$\begin{cases} \Delta & \overset{d}{=} \Gamma(\gamma^{-1}, \gamma^{-1}) \\ Y|\Delta & \overset{d}{=} \text{Exp}(\Delta) \\ Y & \overset{d}{=} \text{GP}(1, \gamma), \end{cases} \tag{5}$$

where $\gamma > 0$, Exp(δ) refers to an exponential random variable with rate $\delta > 0$, $\Gamma(a, b)$ is the Gamma distribution with positive shape a and rate b, and $\overset{d}{=}$ denotes equality in distribution. In terms of densities, the density f_γ of Δ is:

$$f_\gamma(x) = \frac{x^{1/\gamma} e^{x/\gamma}}{\gamma^{1/\gamma} \Gamma(1/\gamma)},$$

and the density f_δ of $Y|\Delta$ is:

$$f_\delta(y) = \delta e^{-\delta y}.$$

The fact that Y follows a heavy tailed GP distribution (see relation (5)) can be proved using Laplace transforms. As an analogy with weather forecasting, we present the benchmark in a temporal setting. At each time $t = 1, \ldots, T > 1$, an observation y is drawn independently from an exponential distribution, where its rate δ is a realization of Δ. In this setting, Y has an exponential tail conditioned by the information due to its rate δ, which represents the *a priori* knowledge of the system, such as the weather at a previous time. Thus, the ideal forecast for each time step is Exp(δ), and knowledge of δ is required Using relation (5), we see that the *climatological* forecaster F_{clim} is a GP distribution with tail index γ and unit scale. Climatology is a commonly used forecast reference in meteorology. In other fields, it can be viewed as the unconditional distribution of the truth, and a climatological forecast can be estimated based on a sample of past and analogous observations. This setting is attractive because the ideal and climatological forecasters belong to two different tail decay regimes.

We introduce alternative competitors for modeling partial knowledge of the conditional state. The λ-*informed* forecaster F_λ, $\lambda \in [0, 1]$ is a mixture between the climatological and ideal forecasts, where a weight such as $\lambda \in [0, 1]$ indicates the contribution of each (see Table 1 for the definition).

Finally, the *extremist* forecaster $F_{\nu,\text{extr}}$ simply adds a multiplicative bias to the ideal forecaster. This forecast is not calibrated but it has the same tail behavior as the ideal forecaster; see Appendix A for detailed discussion on calibration. The benchmark is summarized in Table 1 and referred to as the "Model GE" in the following.

Closed forms of the CRPS are available for each forecast of the proposed benchmark. For instance, the extremist forecast $F_{\nu,\text{extr}}$ satisfies

$$\text{CRPS}(F_{\nu,\text{extr}}, y) = y + \frac{2\nu}{\delta} \exp\left(-\frac{\delta y}{\nu}\right) - \frac{3\nu}{2\delta}\ . \tag{6}$$

Table 1
Benchmark for assessing the behavior of the forecast evaluation procedure with respect to different tail regimes. The $F_{\nu,\mathrm{extr}}$ family is not calibrated. The other forecasts are all marginally and probabilistically calibrated. In addition, the λ-Informed forecasts are conditionally auto-calibrated with respect to Δ according to Tsyplakov (2020, Theorem 2).

Forecasts \ Truth	$Y \overset{d}{=} \mathrm{Exp}(\Delta)$ where $\Delta \overset{d}{=} \Gamma(1/\gamma, 1/\gamma)$, $1 > \gamma > 0$
Ideal F_{ideal}	$\mathrm{Exp}(\Delta)$
Climatological F_{clim}	$\mathrm{GP}(1,\gamma)$
λ-Informed F_λ	$\lambda\mathrm{Exp}(\Delta)+(1-\lambda)\mathrm{GP}(1,\gamma)$
Extremist $F_{\nu,\mathrm{extr}}$	$\mathrm{Exp}(\Delta/\nu)$, $\nu > 1$

Table 2
Relative ratios of the mean CRPS as percentages with respect to the ideal forecast for the model GE with $\gamma = 1/4$ based on $T = 10^6$ observation/forecast pairs.

Truth	$Y \overset{d}{=} \mathcal{E}xp(\Delta)$ where $\Delta \overset{d}{=} \Gamma(4, 4)$
Forecasts	% w.r.t. Ideal
Ideal F_{ideal}	100%
Extremist $\nu = 1.1$	100.48%
0.75-Informed $F_{0.75}$	100.90%
0.5-Informed $F_{0.5}$	103.58%
Extremist $\nu = 1.4$	106.68%
0.25-Informed $F_{0.25}$	108.06%
Climatological F_{clim}	114.33%
Extremist $\nu = 1.8$	122.89%

In addition, combining (D.1) and (6) yields the following formula for the λ-informed forecast, $\lambda \in [0, 1]$,

$$
\begin{aligned}
\mathrm{CRPS}(F_\lambda, y) &= y + \frac{\lambda^2}{2\delta} + \frac{2\lambda}{\delta}\{\exp(-\delta y) - 1\} \\
&- \frac{2(1-\lambda)}{1-\gamma}\left\{1 - (1+\gamma y)^{\frac{\gamma-1}{\gamma}}\right\} \\
&+ \frac{(1-\lambda)^2}{2-\gamma} + \frac{2\lambda(1-\lambda)\gamma^{\frac{-1}{\gamma}}}{\delta^{\frac{\gamma-1}{\gamma}}} \\
&\times \left\{\exp\left(\frac{\delta}{\gamma}\right)\Gamma\left(\frac{\gamma-1}{\gamma}, \frac{\delta}{\gamma}\right)\right\} ,
\end{aligned}
$$

where $\Gamma(s, x) = \int_x^{+\infty} e^{-t} t^{s-1}\, dt$. Table 2 shows the relative ratio of the empirical means of the CRPS for the benchmark with $\gamma = 1/4$. On average, the ideal forecast cannot be beaten in Table 2 because the CRPS is a proper score. Moreover, there are two different rankings, where the first is based on the extremist forecasters involving ν, and the second is in terms of λ based on the λ-informed forecasters. In the latter case, the information is represented by the pair (λ, δ) and the ideal information is $(0, \delta)$. Following the principle of tail equivalence presented in Section 2.2, the extremist forecast should be the forecast that is closest to the ideal because they both belong to the same tail decay regime; however, we observe that the performance of the CRPS average is between that of the least informed forecaster and the climatology.

3. The CRPS as a random variable

3.1. The random CRPS and its properties

In Section 2, we highlighted the difficulty of summarizing the forecast performance in meaningful comparisons of extreme observations. In particular, we showed that using a single number, such as the mean of the CRPS or its weighted counterpart, fails to allow relevant comparisons. Alternatively, we propose studying the distribution of the CRPS when it is treated as a random variable (also see Bessac and Naveau (2021), Ferro (2017)).

For simplicity, we use the setting and corresponding notations for the benchmark presented in Section 2.3. Based on Eqs. (B.1) and (6), the climatological and ideal scores can be treated as random variables whenever y_t is replaced by Y_t. At this point, it is important to recall that a forecast is issued with only partial knowledge of the system. The exact value of δ_t and the distribution of Y_t are unknown, and only the observation y_t is available. Table 3 summarizes the quantities that are available to forecasters. Thus, to evaluate the performance of forecasts, it is only possible to compute $\mathrm{CRPS}(F_t, y_t)$ for each t. The climatological distribution referred to as G and the existence of which needs to be hypothesized in practice, is characterized by the observed sample $(y_1, \ldots, y_t)$, considered as a sample of independent realizations of the random variable Y.

For any set of forecasts $\{F_t\}_{t=1,\ldots,T}$ and sample $y_1, \ldots, y_T$, two types of sets of random variables can be defined:

$$
\begin{aligned}
&\mathcal{S}(F_T) = \{\mathrm{CRPS}(F_t, Y_t)\}_{t=1,\ldots,T} \quad \text{and} \\
&\mathcal{S}^*(F_T) = \{\mathrm{CRPS}(F_t, Y_{\pi(t)})\}_{t=1,\ldots,T} \ ,
\end{aligned}
\tag{7}
$$

where π is a random permutation of $\{1, \ldots, n\}$. Applying π breaks the conditional dependence between y_t and F_t, which is quantified by δ_t in the benchmark, thereby producing alternative less informative forecasts. Thus, for a given forecaster represented by the set $F_T = \{F_t\}_{i=1,\ldots,T}$ and permutation π, we introduce two random variables $\mathcal{S}(F_T)$ and $\mathcal{S}^*(F_T)$ characterized by their respective empirical cdfs.

The climatological forecaster is the only forecaster that satisfies

$$
\{\mathrm{CRPS}(G, Y_t)\}_{t=1,\ldots,T} = \mathcal{S}^*(G) \overset{d}{=} \mathcal{S}(G) \ , \tag{8}
$$

because by definition, it discards any information about system conditioning. The first equality in (8) is a direct consequence of auto-calibration (see Appendix A) and the second equality follows from the permutation invariance of the data from the viewpoint of the climatological forecaster.

The distributional properties of $\mathcal{S}(F_T)$, $\mathcal{S}^*(F_T)$, and $\mathcal{S}(G)$ give relevant insights into the behavior of the forecaster. For example, Fig. 1 shows qq-plots for the distributions of

Table 3
Availability status for the quantities of interest (a posteriori availability).

Object	Definition	Availability in practice
F_t	Distribution of the forecast for time t	Yes
y_t	Observed realization at time t	Yes
δ_t	Conditioning variable	No
Δ	Conditioning random variable	No
Y_t	Conditional random variable generating y_t	No
Y	Unconditional random variable of the observations	Yes
$\mathrm{CRPS}(F_t, y_t)$	CRPS of the couple for time t	Yes
$\mathrm{CRPS}(F_t, Y_t)$	Random variable associated with $\mathrm{CRPS}(F_t, y_t)$	No
$\mathrm{CRPS}_{\mathcal{S}}(F, Y)$	Random variable generated by the $(\mathrm{CRPS}(F_t, y_t))_t$	Yes
$\mathrm{CRPS}_{\mathcal{S}^*}(F, Y)$	Random variable generated by the $(\mathrm{CRPS}(F_t, y_{\pi(t)}))_t$	Yes

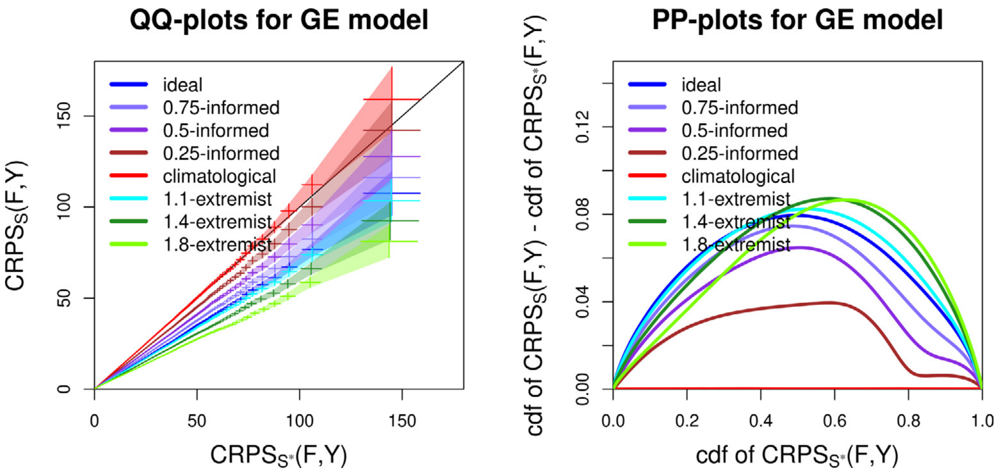

Fig. 1. Comparisons of the distributional properties between $\mathcal{S}$ and $\mathcal{S}^*$ for each forecast in model GE: qq-plots (left) and pp-plots (right panel). Each forecast is represented by samples with size $T = 10^6$. In the left panel, the points represent the average distributions over 100 independent samples, and the 95% confidence intervals are shown. In the right panel, the curves were averaged based on 100 independent samples. In practice, one must use a set of permutations π.

$\mathcal{S}^*(F_T)$ against $\mathcal{S}(F_T)$ for each forecast of the benchmark with $\gamma = 1/4$. We observe that the ideal, λ-informed, and extremist forecasts deviate from the diagonal, thereby demonstrating the influence of the loss of information caused by the permutation. This visual illustration summarizes how $\mathcal{S}(F_T)$ and $\mathcal{S}^*(F_T)$ capture relevant information from the modeled conditioned by the random variable Δ. The right panel in Fig. 1 displays these distributions on the probability scale and it highlights how the discrepancy of the λ-informed forecaster evolves with the parameter λ. Extremist forecasts with multiple values of the scale parameter ν are shown only to illustrate how the visual illustration behaves when the calibration is not satisfied. Fig. 1 also shows that the forecast dominance among forecasters can be inferred, as given by Ehm et al. (2016, Fig. 1,2,4,6, and pages 528–529) for point forecasts, but in the present case, higher is better compared with the Murphy diagrams. Under calibration, the discrepancy between the distributions can be interpreted in an appropriate manner as a direct measure of the forecaster's skill (the λ-informed curves never cross each other), thereby making this diagnostic method particularly relevant and compliant with the recommendations based on the extremal dependence indices established by Ferro and Stephenson (2011).

3.2. Tail properties of the random CRPS

We now study the upper tail behavior of the random CRPS by using EVT to develop meaningful forecast evaluations for extreme events. To reduce the technical contents of this section, all of the proofs are presented in Appendix E. For the notations with respect to any conditional model that depends on $\Delta = \delta$, we aim to emphasize the difference between a conditional forecast, such as F_δ, and an unconditional forecast F. It should be noted that δ depends on the time index t, but for notational simplicity, we omit this index. Δ may also change over time but we assume that it is invariant.

Let X and Y be two random variables with absolutely continuous cdfs F and G with a common upper bound $x_F = x_G$. Suppose that $\gamma < 1$ exists such that $G \in \mathcal{D}(H_\gamma)$ and that $c_F = 2\mathbb{E}_F(XF(X))$ is finite. Then, conditionally on $\Delta = \delta$, we have

$$\mathbb{P}\left(\frac{\mathrm{CRPS}(F_\delta, Y_\delta) + c_{F_\delta} - u_\delta}{b_\delta(u_\delta)} > x \;\middle|\; Y_\delta > u_\delta\right) \longrightarrow (1 + \gamma_\delta x)^{-1/\gamma_\delta}\,, \tag{9}$$

as u_δ tends to x_{G_δ}, with $1 + \gamma_\delta x > 0$. Thus, in any fixed state δ (such as a state of the atmosphere for a weather forecast), the CRPS upper tail behavior (conditionally on

$\Delta = \delta$) is equivalent to the observed tail behavior, thereby formalizing what we observe intuitively from (B.1).

Now, unconditionally, we can also obtain a result for the climatological forecast due to its property of invariance under permutation (see Section 3.1). If $\gamma < 1$ exists such that $G \in \mathcal{D}(H_\gamma)$, then

$$\mathbb{P}\left\{ \frac{\text{CRPS}(G, Y) + c_G - u}{b(u)} > x \middle| Y > u \right\} \longrightarrow (1 + \gamma x)^{-1/\gamma}, \quad u \to x_G, \tag{10}$$

for any x such that $1 + \gamma x > 0$. In the case where $\gamma > 0$, convergence in Eq. (10) also holds for $c_G = 0$ because the latter vanishes due to the linear behavior of the auxiliary function b in Eq. (3) (e.g., see Embrechts et al. (1997)).

The benchmark presented in Table 1 illustrates these results. The choice to work with a time indexed couple (F_t, Y_t) or with an invariant (G, Y) significantly affects the tail behavior of the CRPS random variables, where according to Table 1, the former implies that the limit in (9) exhibits an exponential tail, whereas the climatological tail given by (10) is heavy, i.e., $\gamma > 0$.

3.3. Assessment of the forecaster tail behavior

In this section, we propose a tail-equivalent forecast performance index inspired by Eqs. (9) and (10), and Fig. 1. We only aim to provide the intuitive basis of the index and formal theoretical analysis is left for future work. We assume that the forecasts lie in the domain of attraction for some distribution $H_{\gamma,\sigma}$. For a sufficiently large u, the null hypothesis $H_0 : \forall t = 1, \ldots, T$, $\text{CRPS}(F_t, Y_t)|Y_t > u \sim H_{\gamma,\sigma_u}$ should be rejected for any calibrated forecast with tail behavior closer to the ideal forecast than the climatological reference.

Furthermore, we assume that the variables in $\mathcal{S}(F_T)$ are iid. This assumption might not always be satisfied, e.g., temperature measures on two consecutive days are likely to be dependent, but it can be satisfied in a reasonable manner for measurements that are sufficiently far apart. For each forecast, we can compute a Cramér–von Mises criterion

$$\omega_u{}^2\{\mathcal{S}(F_T)\} = \int_{-\infty}^{+\infty} [\hat{K}^{(m)}_{\mathcal{S},u}(v) - H_{\gamma,\sigma_u}(v)]^2 dH_{\gamma,\sigma_u}(v),$$

where $\hat{K}^{(m)}_{\mathcal{S},u}$ is the empirical distribution of the observations in $\mathcal{S}(F_T)$ that exceed the threshold u. The empirical nature of $\hat{K}^{(m)}_{\mathcal{S},u}$ allows us to simplify $\omega_u{}^2\{\mathcal{S}(F_T)\}$ to

$$\Omega_u^F = m \times \widehat{\omega_u}^2\{\mathcal{S}(F_T)\} = \frac{1}{12m} + \sum_{i=1}^{m} \left[\frac{2i-1}{2m} - H_{\gamma,\sigma_u}(s_i) \right]^2,$$

where m denotes the number of observations exceeding u and $s_1, \ldots, s_m$ are the ordered values of $\mathcal{S}(F_T)$. A detailed algorithm for computing Ω_u^F is provided in Table F.4 in Appendix F.

As suggested by Fig. 1, we assume that $\Omega_u^F > \Omega_u^G$ for any calibrated forecasts and climatology G. In addition, for two calibrated forecasts F^1 and F^2, we conjecture that $\Omega_u^{F^2} \geq \Omega_u^{F^1}$ if F^2 has a tail behavior closer to the ideal forecast than F^1. Under these assumptions, we can summarize the comparison between Ω_u^F and Ω_u^G simply as

$$T_u(F, G) = 1 - \frac{\Omega_u^G}{\Omega_u^F}. \tag{11}$$

The behavior of the index T_u is illustrated with the help of model GE. Fig. 2 shows the changes in T_u as a function of the threshold u for $T = 10^6$ and $\gamma = 1/4$. Clearly, the behavior of the index is consistent with our conjecture, where the ideal forecast performs best whereas the climatology has the lowest index. The performance rankings among the calibrated forecasters are stable as the threshold increases, where the ideal forecast always obtains the largest index. The extremist forecasters shown to illustrate the behavior of the index for a non-calibrated forecast obtain a high index, and even larger than the ideal forecast, thereby highlighting the importance of calibration, which must be carefully assessed before any interpretation of T_u.

In practice, a threshold choice must be made and numerous methods have been developed for this purpose (e.g., see Beirlant et al. (2004), Naveau, Huser, Ribereau, and Hannart (2016), Papastathopoulos and Tawn (2013)).

4. Discussion

In this study, we used a carefully designed benchmark to argue that the mean of the CRPS, or its weighted counterparts, are unable to successfully discriminate a forecast upper tail regime, as demonstrated by Brehmer and Strokorb (2019). Ehm et al. (2016) introduced "Murphy diagrams" for assessing dominance in point forecasts. This original approach allows us to appreciate the dominance of different forecasts and to estimate their skill area, and a similar visual illustration is presented in Fig. 1 for calibrated forecasts.

Inspired by Friederichs and Thorarinsdottir (2012), we applied EVT directly based on common verification measures. By considering the CRPS as a random variable (also see Bessac and Naveau (2021) for non-extreme cases), one can view this contribution as a first step toward considering other functionals of score distributions rather than their means. The new index introduced in Section 3.3 can be considered as a probabilistic alternative to the scores introduced by Ferro (2007), Ferro and Stephenson (2011). We link the paradigm of *maximizing the sharpness subject to calibration* proposed by Gneiting et al. (2007) and the paradigm of *maximizing the information for extreme events subject to calibration*. Similarly, Murphy (1993) explained the differences between the forecast quality (accordance between forecasts and observations) and forecast value (ability to use information to achieve a benefit by choosing a forecast), and the forecast value seems to be the most important for extreme events, where decision making is crucial. Well-known tools are available for deterministic weather forecasts (e.g., see Richardson (2000), Zhu, Toth, Wobus, Richardson, and Mylne (2002)). Other widely used scores based on the dependence between forecasts and observed events were considered by Ferro and Stephenson (2011), Stephenson, Casati, Ferro, and Wilson (2008).

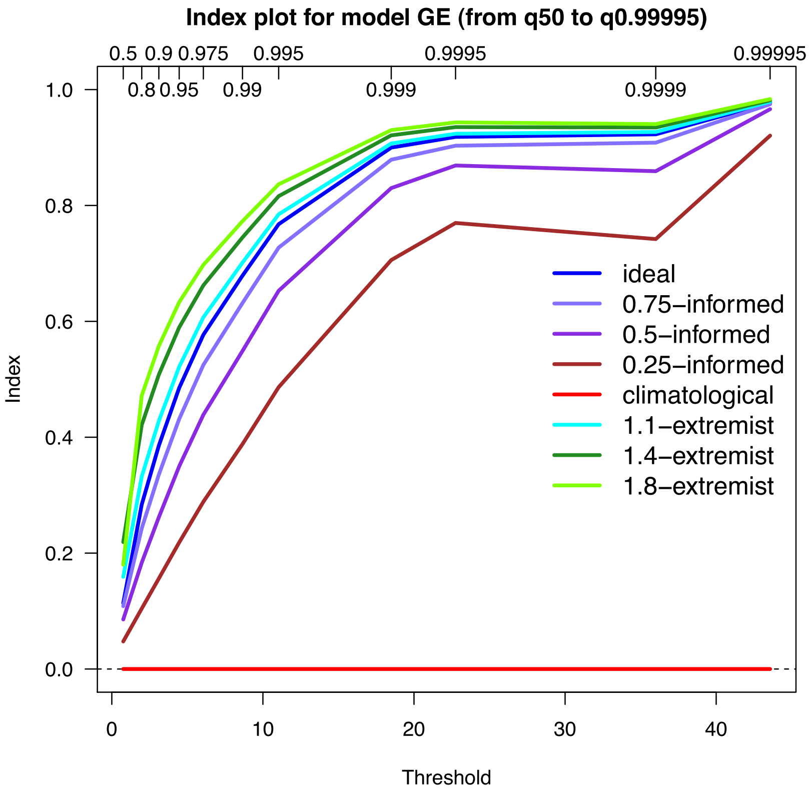

Fig. 2. Cramér–von Mises' criterion-based index as a function of the threshold for the different forecasts using model GE with parameters $T = 10^6$ and $\gamma = 1/4$. Indexes were computed for thresholds ranging from the 0.5 to 0.99995 empirical quantiles. Higher index values are assumed to reflect tail behavior closer to the ideal forecaster. The validity of the index is limited to calibrated forecasts and the calibration must be carefully checked for non-calibrated extremist forecasts before interpreting the results.

It would useful to further investigate the theoretical properties of this CRPS-based tool. Another potentially interesting approach could involve extending this procedure to other scores, such as the mean absolute difference, Dawid–Sebastiani score (Dawid & Sebastiani, 1999), or ignorance score (Diks et al., 2011; Smith et al., 2015). Classical tools for verification rely on a verification period, and thus evaluation is always conducted a posteriori as a consequence. Therefore, it would be interesting to consider the sequential evaluation of rare events, such as by using the e-values (Vovk & Wang, 2021) introduced for continuously assessing and monitoring calibration (Arnold, Henzi, & Ziegel, 2021). Eventually, we invite scientists to work on a new scoring rule theory that is not based on average scores.

Declaration of competing interest

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

Acknowledgments

The authors would like to sincerely thank the Associate Editor and reviewers for their careful review, comments, and suggestions throughout the peer review process.

Part of this work was supported by the French National Research Agency (ANR) project T-REX (ANR-20-CE40-0025) and by Energy oriented Centre of Excellence-II (EoCoE-II), Grant Agreement 824158, funded within the Horizon2020 framework of the European Union. Part of this work was also supported by the ExtremesLearning grant from 80 PRIME CNRS-INSU, France and the ANR project Melody, France (ANR-19-CE46-0011). This work was partially supported by the ANR LABEX MILYON, France (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007).

Appendix A. Prediction framework and calibration

The theoretical framework considered in this study is the now classical *prediction space* introduced by Ehm et al. (2016), Gneiting and Ranjan (2013), Murphy and Winkler (1987), and generalized in a serial context by Strähl and Ziegel (2017). The framework starts formally with a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ and a collection of sub-σ-algebras $\mathcal{A}_1, \dots, \mathcal{A}_k \subset \mathcal{A}$, where $\mathcal{A}_i$ represents the information available to forecaster i. In a meteorological context, this framework can be seen as the representation of the atmosphere by each forecaster. In the benchmark considered in Section 2.3, for simplicity, we consider that the information set is generated by a random variable Δ.

A real-valued outcome Y is observed and seen as a (real-valued) random variable. A probabilistic forecast i for Y is identified with its so-called "predictive distribution" with cdf F_i. Rigorously, $F_i : \Omega \times \mathcal{B}(\mathbb{R}) \to [0,1]$ is a kernel[1] from $(\Omega, \mathcal{A}_i)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, but as in previous studies, we identify the kernels with random cumulative cdf (e.g., see Strähl and Ziegel (2017) for more details). In particular, for each $x \in \mathbb{R}$, we might use the notation $F_i(x)$ denoting the random element $\omega \mapsto F_i(\omega, (-\infty, x])$.

In this framework, a forecast F_i is termed *ideal* with respect to $\mathcal{A}_i$ if $F_i = \mathcal{L}(Y|\mathcal{A}_i)$ almost surely. Tsyplakov (2011) also refers to this property by stating that F_i is *calibrated* with respect to $\mathcal{A}_i$. He also defined *auto-calibration* as the property required for F_i to satisfy $F_i = \mathcal{L}(Y|\sigma(F_i))$ almost surely. $\sigma(F_i)$ denotes the σ-algebra generated by F_i, i.e., the smallest σ-algebra such that $\omega \mapsto F_i(\omega, x)$ is measurable for all $x \in \mathbb{R}$. It should be noted that if a forecast is calibrated with respect to $\mathcal{A}_i$, then it is auto-calibrated, but the converse does not hold in general. For the particular case considered in Section 2.3, the *climatological* forecaster is ideal with respect to the trivial σ-algebra.

In practice, we are not only concerned with predictions for an outcome Y at a single time point. The framework introduced above also allows us to deal with independent replicates at times $t = 1, 2, \ldots,$ as in Section 2.3. If this assumption of independence seems unrealistic in several situations, as argued by Strähl and Ziegel (2017), it can still provide a first step and lead to less technical complexity. Thus, we made this choice for simplicity in the present study.

Appendix B. An alternative expression for the weighted CRPS

The weighted CRPS defined by (2) can be reformulated as follows provided that the weight function $w(.)$ is continuous,

$$\mathrm{wCRPS}(F, y) = W(y) + 2\mathbb{E}_F[\{W(X) - W(y)\}\mathbf{1}_{X>y}] - 2\mathbb{E}_F[W(X)F(X)]. \tag{B.1}$$

Assume that the weight function $w(.)$ is continuous. By integrating by parts $\int_{-\infty}^{y} F^2(x)w(x)\,dx$ and $\int_{y}^{\infty} \overline{F}^2(x)w(x)\,dx$ and using $W(x) = \int_{-\infty}^{x} w(z)dz$, the weighted CRPS defined by (2) can be rewritten as

$$\mathrm{wCRPS}(F, y) = \mathbb{E}_F|W(X) - W(y)| - \frac{1}{2}\mathbb{E}_F|W(X) - W(X')|.$$

The equality $|a - b| = 2\max(a, b) - (a + b)$ gives

$$\begin{aligned}
&\mathbb{E}_F|W(X) - W(y)| \\
&\quad = 2\mathbb{E}_F \max(W(X), W(y)) - \mathbb{E}_F W(X) - W(y), \\
&\quad = W(y) - \mathbb{E}_F W(X) \\
&\qquad +2\mathbb{E}_F\, (W(X) - W(y)I[W(X) > W(y)])\,,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}_F|W(X) - W(X')| \\
&\quad = 2\mathbb{E}_F \max(W(X), W(X')) - 2\mathbb{E}_F W(X), \\
&\quad = 4\mathbb{E}(W(X)F_{W(X)}(W(X))) - 2\mathbb{E}_F W(X), \\
&\quad = 4\mathbb{E}(W(X)F(X)) - 2\mathbb{E}_F W(X)\,,
\end{aligned}$$

where the last line follows from the fact that $F_{W(X)}(W(X))$ and $F(X)$ have the same distribution, which is uniform on $(0, 1)$. We recall that $F_{W(X)}(x) = \mathbb{P}(W(X) \le x)$. As $W(x)$ is non-decreasing, so we have $\{W(X) > W(y)\} = \{X > y\}$, and it follows that

$$\begin{aligned}
\mathrm{wCRPS}(F, y) &= W(y) - \mathbb{E}_F W(X) \\
&\quad + 2\mathbb{E}_F\left[\{W(X) - W(y)\}\mathbf{1}_{W(X)>W(y)}\right] \\
&\quad - 2\mathbb{E}_F[W(X)F(X)] + \mathbb{E}_F W(X), \\
&= W(y) + 2\mathbb{E}_F[\{W(X) - W(y)\}\mathbf{1}_{X>y}] \\
&\quad - 2\mathbb{E}_F[W(X)F(X)]\,,
\end{aligned}$$

as stated in (B.1).

Appendix C. Proof of the inequality (4)

Let u be positive real. Denote Z as a non-negative random variable with finite mean and cdf H. Assume that Z and Y (with cdf J) are independent with the same right end point. We introduce the new random variable

$$X_u = Y\mathbf{1}\{u \ge Y\} + (Z + u)\mathbf{1}\{Y > u\}\,, \tag{C.1}$$

with the survival function $\overline{F_u}$ defined by

$$\overline{F_u}(x) = \begin{cases} \bar{J}(x), & \text{if } x \le u \\ \bar{H}(x - u)\bar{J}(u), & \text{otherwise.} \end{cases} \tag{C.2}$$

In particular, we note that for all x, the decreases in $\overline{F_u}$ show that

$$\overline{F_u}(x) \le \bar{J}(x). \tag{C.3}$$

In addition, for any $x \le u$, Eq. (C.2) and the monotonicity of W show that

$$\mathbb{E}[W(Y)\mathbf{1}\{Y < x\}] = \mathbb{E}[W(X_u)\mathbf{1}\{X_u < x\}]. \tag{C.4}$$

Equality (B.1) implies that

$$\begin{aligned}
&\frac{1}{2}[\mathrm{wCRPS}(F_u, x) - \mathrm{wCRPS}(J, x)] \\
&= \mathbb{E}_{F_u}[(W(X_u) - W(x))\mathbf{1}\{X_u > x\}] \\
&\quad - \mathbb{E}_J[(W(Y) - W(x))\mathbf{1}\{Y > x\}] \\
&\quad + \mathbb{E}_J[W(Y)J(Y)] - \mathbb{E}_{F_u}[W(X_u)F_u(X_u)], \\
&= \mathbb{E}_{F_u}[W(X_u)\overline{F_u}(X_u)] - \mathbb{E}_J[W(Y)\bar{J}(Y)] \\
&\quad - \mathbb{E}_{F_u}[(W(X_u) - W(x))\mathbf{1}\{X_u \le x\}] \\
&\quad + \mathbb{E}_J[(W(Y) - W(x))\mathbf{1}\{Y \le x\}] \\
&= \mathbb{E}_{F_u}[W(X_u)\overline{F_u}(X_u)] - \mathbb{E}_J[W(Y)\bar{J}(Y)] + \Delta(x)\,,
\end{aligned}$$

where

$$\begin{aligned}
\Delta(x) &= \mathbb{E}_J[(W(Y) - W(x))\mathbf{1}\{Y \le x\}] \\
&\quad - \mathbb{E}_{F_u}[(W(X_u) - W(x))\mathbf{1}\{X_u \le x\}].
\end{aligned}$$

The stochastic ordering that holds between X_u and Y implies that the quantity $\mathbb{E}_{F_u}[W(X_u)\overline{F_u}(X_u)] - \mathbb{E}_J[W(Y)\bar{J}(Y)]$ is negative. Combined with (C.4), this leads to

$$\frac{1}{2}\left|\mathbb{E}_J[\mathrm{wCRPS}(F_u, Y)] - \mathbb{E}_J[\mathrm{wCRPS}(J, Y)]\right| \le \int_u^{x_J} \Delta(x)dJ(x). \tag{C.5}$$

[1] This means that for each fixed $\omega \in \Omega$, $F_i(\omega, \cdot)$ is a probability measure, and for each fixed $x \in \mathbb{R}$, $F_i(\cdot, (-\infty, x])$ is $\mathcal{A}_i$-measurable.

For $x > u$, we can write that

$$\begin{aligned}\Delta(x) &= \mathbb{E}_J[(W(Y) - W(x))\mathbf{1}\{u < Y \le x\}] \\ &\quad - \mathbb{E}_{F_u}[(W(X_u) - W(x))\mathbf{1}\{u < X_u \le x\}], \\ &\le \mathbb{E}_{F_u}[(W(x) - W(u))\mathbf{1}\{u < X_u \le x\}],\end{aligned}$$

since $W(Y) - W(x) \le 0$ in the first expectation, whereas $0 \le W(x) - W(X_u) \le W(x) - W(u)$ in the second. As a consequence, we obtain

$$\begin{aligned}\Delta(x) &\le (W(x) - W(u))[F_u(x) - F_u(u)], \\ &\le (W(x) - W(u))\overline{F_u}(u), \\ &= (W(x) - W(u))\bar{J}(u).\end{aligned}$$

Finally, combining the latter expression with (C.5) leads to

$$\begin{aligned}&\left|\mathbb{E}_G[\mathrm{wCRPS}(F_u, Y)] - \mathbb{E}_J[\mathrm{wCRPS}(J, Y)]\right| \\ &\quad \le 2\bar{J}(u)\int_u^{x_J}(W(x) - W(u))dJ(x).\end{aligned}$$

It should be noted that this inequality is true for any u and H, and its right-hand side does not depend on $\overline{H}(x)$. Thus, the tail behavior of the random variables Y and Z can be completely different, although the CRPS of J and F_u can be as closed as we require. The right-hand side tends to 0 due to the finite mean of $W(Y)$.

Appendix D. A detailed example related to Section 2.2

In this appendix, we demonstrate the fact that the CRPS fails to discriminate forecasts with different tails. We consider GP distributed forecasts and observations. In this case, the closed forms of the CRPS are available, as described in the following.

Lemma 1. *Consider $X \overset{d}{=} GP(\beta, \xi)$ and $Y \overset{d}{=} GP(\sigma, \gamma)$ with $0 \le \xi < 1$ and $0 \le \gamma < 1$, with the respective survival functions $\overline{F}(x) = (1 + \xi x/\beta)^{-1/\xi}$ (for $x > -\beta/\xi$) and $\overline{G}(x) = (1 + \gamma x/\sigma)^{-1/\gamma}$ (for $x > -\sigma/\gamma$). If $\gamma/\sigma = \xi/\beta$, with $\gamma \neq 0$, then*

$$\mathbb{E}_G\left[\mathrm{CRPS}(F, Y)\right] = \frac{\sigma}{1-\gamma} + 2\beta\left[\frac{1}{2(2-\xi)} - \frac{\gamma}{\gamma + \xi - \gamma\xi}\right],$$

which gives the minimum CRPS value for $\xi = \gamma$ and $\sigma = \beta$,

$$\mathbb{E}_G\left[\mathrm{CRPS}(G, Y)\right] = \frac{\sigma}{(2-\gamma)(1-\gamma)}.$$

Proof. By applying (B.1) with $W(y) = y$ and making use of the classical properties of the Pareto distribution (e.g., see Embrechts et al. (1997, Theorem 3.4.13)), we obtain

$$\begin{aligned}\mathrm{CRPS}(F, y) &= y + 2(1 + \xi y/\beta)^{-1/\xi}\frac{\beta + \xi y}{1-\xi} \\ &\quad - 2\beta\left(\frac{1}{1-\xi} - \frac{1}{2(2-\xi)}\right). \qquad \text{(D.1)}\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{E}\left[\mathrm{CRPS}(F, Y)\right] &= \frac{\sigma}{1-\gamma} + 2\frac{\beta}{1-\xi}m_0 + 2\frac{\xi}{1-\xi}m_1 \\ &\quad - 2\beta\left(\frac{1}{1-\xi} - \frac{1}{2(2-\xi)}\right),\end{aligned}$$

with

$$m_0 = \mathbb{E}\left[\left(1 + \frac{\xi}{\beta}Y\right)^{-1/\xi}\right], \text{ and}$$

$$m_1 = \mathbb{E}\left[Y\left(1 + \frac{\xi}{\beta}Y\right)^{-1/\xi}\right].$$

Since

$$\left(1 + \frac{\xi}{\beta}y\right)^{-1/\xi} = \overline{G}^s(cy), \text{ with } c = \frac{\xi\sigma}{\beta\gamma} \text{ and } s = \frac{\gamma}{\xi},$$

we can write

$$m_r = \mathbb{E}\left[Y^r\overline{G}^s(cY)\right] \text{ for } r = 0, 1.$$

In addition, $G^{-1}(v) = \frac{\sigma}{\gamma}\left((1-v)^{-\gamma} - 1\right)$, so we can rewrite by denoting U as a random variable that is uniformly distributed on $(0, 1)$,

$$\begin{aligned}m_r &= \mathbb{E}\left[G^{-1}(U)^r\overline{G}^s\left(cG^{-1}(U)\right)\right], \\ &= \mathbb{E}\left[\left(\frac{\sigma}{\gamma}\left((1-U)^{-\gamma} - 1\right)\right)^r\right. \\ &\qquad \left.\left(1 + \frac{\gamma}{\sigma}\left(c\frac{\sigma}{\gamma}\left((1-U)^{-\gamma} - 1\right)\right)\right)^{-s/\gamma}\right], \\ &= \left(\frac{\sigma}{\gamma}\right)^r\mathbb{E}\left[\left(U^{-\gamma} - 1\right)^r\left((1-c) + cU^{-\gamma}\right)^{-s/\gamma}\right], \\ &= \left(\frac{\sigma}{\gamma}\right)^r\mathbb{E}\left[\left(\frac{B}{1-B}\right)^r\left(\frac{1-(1-c)B}{1-B}\right)^{-s/\gamma}\right], \\ &\qquad \text{with } B = 1 - U^\gamma \\ &= \left(\frac{\sigma}{\gamma}\right)^r\mathbb{E}\left[B^r(1-B)^{-r+s/\gamma}\left(1-(1-c)B\right)^{-s/\gamma}\right], \\ &\qquad \text{with } B \sim \mathrm{Beta}(1, 1/\gamma) \\ &= \left(\frac{\sigma}{\gamma}\right)^r\mathbb{E}\left[B^r(1-B)^{-r+1/\xi}\left(1-(1-c)B\right)^{-1/\xi}\right], \\ &\qquad \text{because } s/\gamma = 1/\xi.\end{aligned}$$

If $c = \dfrac{\xi\sigma}{\beta\gamma} = 1$, then this can be simplified as

$$\begin{aligned}m_r &= \left(\frac{\sigma}{\gamma}\right)^r\frac{1}{\gamma}\int_0^1 u^r(1-u)^{-r+1/\xi+1/\gamma-1}du \\ &= \left(\frac{\sigma}{\gamma}\right)^r\frac{1}{\gamma}B(r+1, -r+1/\xi+1/\gamma), \\ &= \left(\frac{\sigma}{\gamma}\right)^r\frac{1}{\gamma}\frac{\Gamma(r+1)\Gamma(-r+1/\xi+1/\gamma)}{\Gamma(1+1/\xi+1/\gamma)}.\end{aligned}$$

In particular, $m_0 = \frac{1}{\gamma}B(1, 1/\xi + 1/\gamma) = \left(1 + \frac{\gamma}{\xi}\right)^{-1}$ and

$$m_1 = \frac{\sigma}{\gamma}\left(1 + \frac{\gamma}{\xi}\right)^{-1}\left(\frac{1}{\xi} + \frac{1}{\gamma} - 1\right)^{-1}.$$

If $\frac{\gamma}{\sigma} = \frac{\xi}{\beta}$, then it follows that we have

$$\mathbb{E}\left[\mathrm{CRPS}(F,Y)\right] = \frac{\sigma}{1-\gamma} + 2\beta\left[\frac{1}{2(2-\xi)} - \frac{\gamma}{\gamma+\xi-\gamma\xi}\right],$$

which gives the minimum CRPS value for $\xi = \gamma$ and $\sigma = \beta$,

$$\mathbb{E}\left[\mathrm{CRPS}(G,Y)\right] = \frac{\sigma}{(2-\gamma)(1-\gamma)},$$

and thus we conclude the proof of Lemma 1. □

Lemma 1 allows us to study the effect of changing the forecast's tail behavior captured by ξ and the spread forecast encapsulated in β when F and G have proportional parameters, i.e., $\beta = a\sigma$ and $\xi = a\gamma$ for some $a > 0$. In this case, the CRPS simplifies to

$$\mathbb{E}_G\left[\mathrm{CRPS}(F,Y)\right] = \frac{\sigma}{1-\gamma} + 2a\sigma\left[\frac{1}{2(2-a\gamma)} - \frac{1}{1+a-a\gamma}\right], \tag{D.2}$$

thereby leading to a forecaster with a heavier tail when $a > 1$ and overestimating the true upper tail behavior, and the opposite when $a < 1$.

Counter examples to the previous one can be found to illustrate how weighted scoring rules fail to compare tail behaviors. Thus, they should be handled with care, especially for forecast makers, as previously advocated by Gilleland, Hering, Fowler, and Brown (2018), Lerch et al. (2017).

Appendix E. Proof of convergence for (9) and (10)

The proof of (10) can be viewed as a particular case of (9), so we focus on proving (9). The following lemma helps us to obtain the result, and thus it is presented first with its proof. In the following, the mean excess function of any random variable Z with finite mean and cdf F is denoted by $M(F,z)$ such that $\overline{F}(z)M(F,z) = \mathbb{E}_F[(Z-z)\mathbb{1}_{Z>z}]$.

Lemma. *Consider a random variable Z with finite mean that belongs to the domain of attraction $\mathcal{D}(H_\gamma)$ with $\gamma < 1$. Non-negative real numbers α and β exist such that for each $z \in \mathbb{R}$,*

$$0 \le 2\mathbb{E}_F\left[(Z-z)\mathbb{1}_{Z>z}\right] \le \overline{F}(z)(\alpha z + \beta). \tag{E.1}$$

Proof of the Lemma. The indicator function $\mathbb{1}_{Z>z}$ implies that we always have $0 \le 2\mathbb{E}_F((Z-z)\mathbb{1}_{Z>z})$. To prove that $2\mathbb{E}_F((Z-z)\mathbb{1}_{Z>z})$ is smaller than $\overline{F}(z)(\alpha z+\beta)$, we first show that this inequality holds for large values of z. First, we note that if $z > x_F$, then (E.1) is trivially true. We then show the result when $z \overset{<}{\to} x_F$, and we decompose the proof depending on the sign of γ.

1. F belongs to $\mathcal{D}(H_\gamma)$ with $0 < \gamma < 1$: In this case, Embrechts et al. (1997) (Section 3.4) showed that $M(F,z) \sim \gamma z/(1-\gamma)$ as z tends to x_F, and thus we can reach the conclusion directly.
2. F belongs to $\mathcal{D}(H_\gamma)$ with $\gamma < 0$: In this case, the result also follows readily from Embrechts et al. (1997) since when z tends to x_F, $M(F,z) \sim \gamma(x_F - z)/(\gamma-1)$. This allows us to fix $\alpha = 0$ and $\beta = \sup_{z\in V(x_F)} \gamma(x_F - z)/(\gamma-1)$ for an appropriate neighborhood $V(x_F)$ of x_F.
3. F belongs to $\mathcal{D}(H_0)$: When F is in the Gumbel domain of attraction, $M(F,z)/z \to 0$ as z tends to x_F (e.g., see Theorem 3.9 in Ghosh and Resnick (2010)). If x_F is finite, then a positive β exists such that $2M(F,z) \le \beta$ and α can be fixed to 0, whereas if x_F is infinite, the fact that $2M(F,z) < z$ for a sufficiently large z enables us to conclude the proof.

Thus, we have shown that for some large z_0, non-negative α and β exist such that

$$2\mathbb{E}_F((Z-z)\mathbb{1}_{Z>z}) \le \overline{F}(z)(\alpha z+\beta), \text{ for all } z > z_0.$$

We still need to prove that this statement also holds for $z \le z_0$. Define

$$0 \le \beta_0 = 2\max_{z\le z_0} \mathbb{E}_F[(Z-z)\mathbb{1}_{Z>z}].$$

As $\gamma < 1$, β_0 is finite and, as $\overline{F}(z) \ge \overline{F}(z_0)$ for all $z \le z_0$, we have

$$0 \le \beta_0 \le \beta_0 \frac{\overline{F}(z)}{\overline{F}(z_0)}.$$

We now have two cases: either $\beta < \frac{\beta_0}{\overline{F}(z_0)}$ or $\beta \ge \frac{\beta_0}{\overline{F}(z_0)}$. In the latter case, we have $2\mathbb{E}_F((Z-z)\mathbb{1}_{Z>z}) \le \beta_0 \le \overline{F}(z)(\alpha z+\beta)$, and thus the required result is obtained. In the case of $\beta < \frac{\beta_0}{\overline{F}(z_0)}$, it is always possible to increase the β chosen when $z > z_0$, and bring it above $\frac{\beta_0}{\overline{F}(z_0)}$. □

We are now ready to prove (9) as stated above.

Proof of (9). Given the conditional forecast F_δ, the CRPS can be computed with respect to the conditional observation y_δ in the following manner

$$\mathrm{CRPS}(F_\delta, y_\delta) = y_\delta - c_\delta + 2\mathbb{E}_{F_\delta}\left[(X_\delta - y_\delta)1(X_\delta > y_\delta)\right],$$

where $c_\delta = 2\mathbb{E}_{F_\delta}\left[X_\delta F_\delta(X_\delta)\right]$. To simplify the notations, we omit the subscript δ in the remainder of the proof, but it is returned at the end. The previous lemma allows us to write

$$Y \le \mathrm{CRPS}(F,Y) + c \le (1+\alpha\overline{F}(Y))Y + \beta\overline{F}(Y) \quad a.s.$$

Let us work conditionally on $Y > u$, for a large u close to $x_F = x_Y$. We then obtain

$$Y \le \mathrm{CRPS}(F,Y) + c \le (1+\alpha\overline{F}(u))Y + \beta\overline{F}(u) \quad a.s.$$

This holds when the right end point of Y is non-negative. If this is not the case, then we can simply write $Y \le \mathrm{CRPS}(F,Y) + c \le Y + \beta\overline{F}(u)$ *a.s.*.

The main idea of the proof involves noting that $\overline{F}(u)$ tends to zero as u becomes large, and thus the inequalities above indicate that the thresholded random variable $Y[u] = [(Y-u)/b(u) \mid Y > u]$ and the thresholded CRPS $C[u] = [(\mathrm{CRPS}(F,Y)+c-u)/b(u) \mid Y > u]$ should behave in a similar manner for large u. The choice of the positive constant $b(u)$ depends on the domain of attraction of Y. In particular, we assume that the distribution of $Y[u]$ converges toward a Gaussian probability distribution (GPD)

Table F.4
Computation of Cramér–von Mises criterion from N couples forecsgts/observations. The computation can be conducted with the R package `extremeIndex` (Taillardat, 2021a).

0. CRPS estimates for each forecaster:	– For the N couples forecasts/observations, compute their corresponding instantaneous CRPS.
1. Estimation of γ based on the observations:	– Find a threshold u where the Pareto approximation is acceptable and estimate the Pareto shape parameter γ and σ .
2. For a threshold $w \geq u$:	– Compute the scale parameter $\sigma_w = \sigma + \gamma w$.
3. Computation of X_u For $i \in [1, m]$ End 3. End 2.	– Order the m CRPS values where the observations $y \geq w$ are in increasing order $s_1, \dots, s_m$. – Compute for each CRPS value s_i, $H_{\gamma,\sigma_w}(s_i)$. – Compute $\left[\frac{2i-1}{2m} - H_{\gamma,\sigma_w}(s_i)\right]^2$.

with finite mean. Thus, we have

$$
\begin{aligned}
&0 \leq \mathbb{P}\left(\frac{\mathrm{CRPS}(F, Y) + c - u}{b(u)} > t \mid Y > u\right) \\
&\quad -\mathbb{P}\left(\frac{Y-u}{b(u)} > t \mid Y > u\right) \\
&\leq \mathbb{P}([1+\alpha\overline{F}(Y)]Y + \beta\overline{F}(Y) > tb(u) + u \mid |Y > u) \\
&\quad -\mathbb{P}(Y > tb(u) + u \mid Y > u) \\
&\leq \mathbb{P}\left(Y > \frac{tb(u) + u - \beta\overline{F}(u)}{1+\alpha\overline{F}(u)} \mid Y > u\right) \\
&\quad -\mathbb{P}(Y > tb(u) + u \mid Y > u).
\end{aligned}
$$

We recognize that the probability (conditionally on $Y > u$) for Y is in an interval denoted by

$$
I_u = \left[\frac{tb(u) + u - \beta\overline{F}(u)}{1+\alpha\overline{F}(u)}, tb(u) + u\right].
$$

The remaining part of the proof involves showing that this conditional probability tends to 0 as $u \to x_F$. We can write

$$
\mathbb{P}\left(Y \in I_u \mid Y > u\right) = \mathbb{P}\left(Y \in u + J_u \mid Y > u\right) \ ,
$$

where $J_u = \left[\frac{tb(u) - \overline{F}(u)(\alpha+\beta)}{1+\alpha\overline{F}(u)}, tb(u)\right]$. For a sufficiently large u, the latter probability can be approximated by a GPD such that

$$
\begin{aligned}
\mathbb{P}\left(Y \in I_u \mid Y > u\right) &\sim |J_u| \sup_{v \in J_u} g_{GP}(v) \\
&= \frac{\overline{F}(u)[\alpha + \beta + \alpha tb(u)]}{1+\alpha\overline{F}(u)} \sup_{v \in J_u} g_{GP}(v) \ ,
\end{aligned}
$$

where g_{GP} denotes the probability density function associated with the GPD. This implies the convergence of the latter probability to 0. This is true conditionally on $\Delta = \delta$, so it can be rewritten after reintroducing the subscript δ as

$$
\begin{aligned}
&\mathbb{P}\left(\frac{\mathrm{CRPS}(F_\delta, Y_\delta) + c_\delta - u_\delta}{b_\delta(u_\delta)} > x \mid Y_\delta > u_\delta\right) \\
&\quad \longrightarrow (1+\gamma_\delta x)^{-1/\gamma_\delta} \ ,
\end{aligned}
$$

as u tends to x_{G_δ}, with $1 + \gamma_\delta x > 0$. □

Appendix F. Algorithm for computing the Cramer–von Mises criterion

Note that for large values of u, under the null hypothesis, the statistic Ω_u^F follows a Cramér-von Mises distribution. The associated p-values $p_u^F \in [0, 1]$ could be computed but they are actually subject to numerical instabilities (Csörgő & Faraway, 1996; Prokhorov, 1968). Furthermore, Ω_u^F is sufficient to compare the effect size of the deviation.

Appendix G. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2022.07.003. The index was implemented using the `extremeIndex` package (Taillardat, 2021a). The R code for generating simulation data and figures is available with this article.

References

Arnold, S., Henzi, A., & Ziegel, J. F. (2021). Sequentially valid tests for forecast calibration. arXiv preprint arXiv:2109.11761.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D., & Ferro, C. (2004). Statistics of extremes: theory and applications.

Bessac, J., & Naveau, P. (2021). Forecast score distributions with imperfect observations. *Advances in Statistical Climatology, Meteorology and Oceanography*, *7*(2), 53–71.

Brehmer, J. R., & Strokorb, K. (2019). Why scoring functions cannot assess tail properties. *Electronic Journal of Statistics*, *13*(2), 4015–4034.

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, *138*(667), 1611–1617.

Csörgő, S., & Faraway, J. J. (1996). The exact and asymptotic distributions of Cramér-von Mises statistics. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 221–234.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 278–292.

Dawid, A. P., & Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *The Annals of Statistics*, 65–81.

De Haan, L. F. M. (1970). On regular variation and its application to the weak convergence of sample extremes.

De Haan, L., & Ferreira, A. (2007). *Extreme value theory: An introduction*. Springer Science & Business Media.

Diebold, F. X., Gunther, T. A., & Tay, A. S. (1997). Evaluating density forecasts.

Diks, C., Panchenko, V., & Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, *163*(2), 215–230.

Ehm, W., Gneiting, T., Jordan, A., & Krüger, F. (2016). Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *78*(3), 505–562.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events, volume 33 of applications of mathematics*. Berlin: New York. Springer-Verlag.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*(6), 985–987.

Ferro, C. A. (2007). A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting*, *22*(5), 1089–1100.

Ferro, C. A. T. (2017). Measuring forecast performance in the presence of observation error. *Quarterly Journal of the Royal Meteorological Society*, *143*(708), 2665–2676, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3115.

Ferro, C. A., & Stephenson, D. B. (2011). Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, *26*(5), 699–713.

Friederichs, P., & Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, *23*(7), 579–594.

Galbraith, J. W., & Norden, S. v. (2012). Assessing gross domestic product and inflation probability forecasts derived from Bank of England fan charts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(3), 713–727.

Ghosh, S., & Resnick, S. (2010). A discussion on mean excess plots. *Stochastic Processes and their Applications*, *120*(8), 1492–1517.

Gilleland, E., Hering, A. S., Fowler, T. L., & Brown, B. G. (2018). Testing the tests: What are the impacts of incorrect assumptions when applying confidence intervals or hypothesis tests to compare competing forecasts? *Monthly Weather Review*, *146*(6), 1685–1703.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *69*(2), 243–268.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411–422.

Gneiting, T., & Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, *7*, 1747–1782.

Gumbel, E. J. (1935). Les valeurs extrêmes des distributions statistiques. *Vol. 5*, In *Annales de l'institut henri poincaré* (2), (pp. 115–158).

Henzi, A., Kleger, G.-R., Hilty, M. P., Wendel Garcia, P. D., Ziegel, J. F., & RISC-19-ICU Investigators for Switzerland (2021). Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units. *PLoS One*, *16*(2), Article e0247265.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559–570.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., & Hyndman, R. J. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., Gneiting, T., et al. (2017). Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, *32*(1), 106–127.

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539.

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, *8*(2), 281–293.

Murphy, A. H., & Winkler, R. L. (1987). A general framework for forecast verification. *Monthly Weather Review*, *115*(7), 1330–1338.

Naveau, P., Huser, R., Ribereau, P., & Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, *52*(4), 2753–2769, URL http://dx.doi.org/10.1002/2015WR018552.

Papastathopoulos, I., & Tawn, J. A. (2013). Extended generalised Pareto models for tail estimation. *Journal of Statistical Planning and Inference*, *143*(1), 131–143.

Patton, A. J. (2014). Comparing possibly misspecified forecasts. *Tech. rep.*, Working paper, Duke University.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 119–131.

Prokhorov, Y. V. (1968). An extension of SN Bernstein's inequalities to multidimensional distributions. *Theory of Probability and its Applications*, *13*(2), 260–267.

Raftery, A. E., & Ševčíková, H. (2021). Probabilistic population forecasting: Short to very long-term. *International Journal of Forecasting*, URL https://www.sciencedirect.com/science/article/pii/S0169207021001394.

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649–667.

Smith, L. A., Suckling, E. B., Thompson, E. L., Maynard, T., & Du, H. (2015). Towards improving the framework for probabilistic forecast evaluation. *Climatic Change*, *132*(1), 31–45.

Stephenson, D., Casati, B., Ferro, C., & Wilson, C. (2008). The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorological Applications*, *15*(1), 41–50.

Strähl, C., & Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electronic Journal of Statistics*, *11*(1), 608–639.

Taillardat, M. (2021a). Extremeindex: Forecast verification for extreme events. R package version 0.0.3, URL https://CRAN.R-project.org/package=extremeIndex.

Taillardat, M. (2021b). Skewed and mixture of Gaussian distributions for ensemble postprocessing. *Atmosphere*, *12*(8), 966.

Taillardat, M., Mestre, O., Zamo, M., & Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, *144*(6), 2375–2393.

Tiberi-Wadier, A.-L., Goutal, N., Ricci, S., Sergent, P., Taillardat, M., Boutier, F., et al. (2021). Strategies for hydrologic ensemble generation and calibration: On the merits of using model-based predictors. *Journal of Hydrology*, *599*, Article 126233.

Tsyplakov, A. (2011). Evaluating density forecasts: a comment. Available at SSRN 1907799.

Tsyplakov, A. (2020). Evaluation of probabilistic forecasts: Conditional auto-calibration. https://www.sas.upenn.edu/~fdiebold/papers2/Tsyplakov_Auto_calibration_sent_eswc2020.pdf.

Vovk, V., & Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, *49*(3), 1736–1754.

Winkler, R. L., Munoz, J., Cervera, J. L., Bernardo, J. M., Blattenberger, G., Kadane, J. B., et al. (1996). Scoring rules and the evaluation of probabilities. *Test*, *5*(1), 1–60.

Zamo, M., & Naveau, P. (2017). Estimation of the continuous ranked probability score with limited information. *Mathematical Geosciences*.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., & Mylne, K. (2002). The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, *83*(1), 73–83.