Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, Raphaël de Fondeville

Extreme events evaluation using CRPS 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*

Abstract

Verification of probabilistic forecasts for extreme events has been a very active field of research, stirred by media and public opinions who naturally focus their attention on extreme events, and easily draw biased conclusions. In this context, classical verification methodologies tailored for extreme events, such as thresholded and weighted scoring rules, have undesirable properties that cannot be mitigated; the well-known Continuous Ranked Probability Score (CRPS) makes no exception.

In this paper, we define a formal framework to assess the behavior of forecast evaluation procedures with respect to extreme events, that we use to point out that assessment based on the expectation of a proper score is not suitable for extremes. As an alternative, we propose to study the properties of the CRPS as a random variable using extreme value theory to address extreme events verification. To compare calibrated forecasts, an index is introduced that summarizes the ability of probabilistic forecasts to predict extremes. Its strengths and limitations are discussed using both theoretical arguments and simulations.

Keywords: CRPS, Extreme events, Probabilistic forecasting, Scoring rules, Calibration, Verification.

*Corresponding author
Email address: maxime.taillardat@meteo.fr (Maxime Taillardat)

1. Introduction

By definition, the rarity of extreme events makes difficult to issue relevant forecasts, whose performance assessment is an even greater challenge. In particular, the scarcity of extremes imposes that verification schemes have to 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. For a better utilization of the forecasts, it is generally convenient, and even recommended (Ferro and Stephenson, 2011), to further assume that the forecast F is calibrated (Dawid, 1984; Diebold et al., 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 autocalibration (calibration in the following), we refer to the works of Tsyplakov (2011) and Strähl and Ziegel (2017) summarized in Appendix A.

Calibrated forecasts can be commonly evaluated based on their sharpness, also called refinement by Winkler et al. (1996), which usually refers to their spread. This leads to the paradigm of 'maximizing sharpness subject to calibration', introduced by Gneiting et al. (2007) and later formally justified by Tsyplakov (2011).

Probabilistic forecasting has become more and more popular over the last years in various fields such as economics and finance (Galbraith and Norden, 2012), demography and social science (Raftery and Ševčíková, 2021), health (Henzi et al., 2021), energy (Hong et al., 2016), hydrology and hydraulics (Tiberi-Wadier et al., 2021). In this work, we focus on weather probabilistic forecasts (Leutbecher and Palmer, 2008). Indeed, probabilistic forecasts are nowadays issued by most National Weather Services (NWS) and F is known through a sample of finite size called "ensemble" (see, e.g., Zamo and Naveau, 2017). In this context, forecast verification is performed by computing scoring rules such as the Continuous Ranked Probability Score (CRPS) (Epstein, 1969; Hersbach, 2000; Bröcker, 2012)

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

where $y \in \mathbb{R}$, and X and X' are independent random variables with common cdf F. The CRPS is attractive as it does not require predictive densities, is

inferred non-parametrically, and has simple interpretation. The right hand side of Equation (1) decomposes the CRPS into, in this order, a calibration and a sharpness term (Gneiting and Raftery, 2007). Alternative decompositions are also available; see Taillardat et al. (2016); Bessac and Naveau (2021) and Appendix B.

For the forecast evaluation of extreme events, proper weighted scoring rules were introduced by Gneiting and Ranjan (2011) and Diks et al. (2011). For a non-negative function $w(x)$, the weighted CRPS

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

with $W(x) = \int_{-\infty}^{x} w(t)dt$, aims to emphasize a region of interest, for instance distributional tails. When w is continuous, an alternative expression of the weighted CRPS is available and can be found in Appendix B. The choice of the weight function $w(x)$ is complex and depends on the different stakeholders, such as forecast users and forecasters; see, e.g., Ehm et al. (2016); Gneiting and Ranjan (2011); Patton (2014); Smith et al. (2015); Taillardat (2021b). Even in the hypothetical case where $w(x)$ could be objectively defined, it is essential that the verification process has to be made on the whole set of observations (Lerch et al., 2017) and one can wonder if the corresponding weighted CRPS correctly discriminates between two competitive forecasts with respect to extreme events.

In this work, we show that the expected weighted CRPS cannot discriminate forecasts with different extremal tail behaviors, a potentially redhibitory defect for extremal evaluation. To address this issue, we view the CRPS as a random variable. Its tail behavior is derived and compared to the tail regime of observations using Extreme Value Theory (EVT) (see, e.g. De Haan and Ferreira, 2007).

This work is organized as follows: Section 2 provides an analysis of the weighted CRPS with respect to the notion of tail equivalence, the main backbone of EVT. In particular, we propose a benchmark to compare the tail properties of forecast verification tools allowing us to pinpoint 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 help us to propose and study a new index to assess

the skill of calibrated probabilistic forecasts with respect to extreme events. The paths and pitfalls of this index and potential future works are discussed in the Section 4.

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

2.1. Tail modelling using EVT

Thanks to the pioneering work of Gumbel (1935) and De Haan (1970), EVT provides a theoretically justified framework to model the tail of random variables, more precisely excesses above a large threshold; see, e.g., Embrechts et al. (1997); Beirlant et al. (2004). For any random variable X with cdf F, EVT models assume the existence of a domain of attraction, i.e., that there exists a positive auxiliary function b, such that

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

where $\overline{F} = 1 - F$ corresponds to the survival, also called tail function, and $x_F = \sup\{x : F(x) < 1\}$ is the upper endpoint of F. Under condition (3), noted $F \in \mathcal{D}(H)$, the Pickands-Balkema-de Haan's theorem (De Haan, 1970; Pickands, 1975) establishes that H has to 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 appears to be the ideal candidate to approximate the survival function of exceedances over a large threshold $u > 0$, i.e.,

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

where $x \in \{x : 1 + \gamma x/\sigma > 0\}$ and $\sigma > 0$. The GP family covers the three possible regimes of tail decay which is determined by the value of its tail index γ: when $\gamma \neq 0$ the decay is polynomial and 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 respective cdf F and G 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, we expect that, between two forecasters, one should favor the one that is tail equivalent to the observations. In practice, this may be difficult. 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. Still, their tail behavior widely differ even for small ϵ: The 100 year return level for X_1 is 99, while it is equal to 138 for X_2 with $\epsilon = 0.1$. In other words, if the precedent random variables were to represent water levels, a small difference of 0.1 in tail index, implied a difference of 39 meters which would most likely cause massive and destructive flooding.

This short example illustrates how issuing forecasts with the right tail regime, i.e., as close as possible to the observational one, is a priority for extreme events and that a verification methodology should reward forecast with close, if not equal, tail regime. Ideally, the measure of forecast performance should give not only 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 producing a pessimistic or risk averse scenario. On the contrary, $\gamma_F < \gamma_G$ falls on the optimistic side by under-estimating the likelihood of extreme events.

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

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

proof can be found in Appendix C. More precisely if $G \in \mathcal{D}(H_{\gamma_G})$, then it is possible for any arbitrary $\gamma_F \in \mathbb{R}$ to find $F \in \mathcal{D}(H_{\gamma_F})$ satisfying Equation (4). Thus the CRPS is unable to discriminate properly forecasts with different tail regime, as 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, proving that proper scoring rule expectations are not suitable to distinguish tail properties, see their Theorem 5.4.

2.3. A benchmark for assessing forecasts of extremes

Following Gneiting et al. (2007) and Strähl and Ziegel (2017), we propose a benchmark to assess the behavior of forecast evaluation procedures with respect to tail regimes. The design relies on a hierarchical model based on Gamma–exponential mixtures with $\gamma > 0$

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

where $\mathrm{Exp}(\delta)$ refers to an exponential random variable with scale $\delta > 0$. The fact that Y follows a heavy tailed GP distribution, see relation (5), can be proved using Laplace transforms. For 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 whose scale δ is a realization of Δ. In this setting, Y has an exponential tail which is conditioned by the information brought by its scale δ, representing the *a priori* knowledge of the system, for instance the weather at previous time. Thus the ideal forecast for each time step is $\mathrm{Exp}(\delta)$, and requires the knowledge of δ. 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 an estimation of a climatological forecast can be done based on a sample of past and analogs observations. This setting is attractive as the ideal and the climatological forecasters belong to two different regimes of tail decay.

We introduce alternative competitors modelling 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, say $\lambda \in [0; 1]$, indicates the contribution of each one, see Table 1 for the definition.

Finally, the *extremist* forecaster F_{extr} simply adds a multiplicative bias to the ideal forecaster: while it is not calibrated, such forecast 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 later referred to as the "Model GE".

Table 1: Benchmark to assess the behavior of forecast evaluation procedure with respect to different tail regimes. All forecasts but F_{extr} are calibrated.

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

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

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

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

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

where $\mathbb{\Gamma}(s, x) = \int_x^{+\infty} e^{-t} t^{s-1}\, dt$. Table 2 gives the relative ratio of the empirical means of the CRPS for the benchmark with $\gamma = 1/4$. The CRPS being a proper score, the ideal forecast cannot be beaten in average in the Table 2. Moreover, there is a clear ranking among calibrated forecasts, based on the nested information sets (Holzmann and Eulert, 2014). Following the principle of tail equivalence presented in Section 2.2, the extremist forecast should be the forecast the closest to the ideal as they both belong to the same regime of tail decay; however, we observe that the CRPS average gives

Table 2: Relative ratio of the mean CRPS, in percent, with respect to the ideal forecast for the model GE with $\gamma = 1/4$, based on $T = 10^6$ observation/forecast pairs.

Truth	$Y \overset{d}{=} \mathcal{E}\mathrm{xp}(\Delta)$ 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%

a performance in between the least informed forecaster and the climatology. An alternative measure for forecast evaluation, satisfying the tail equivalence principle is thus required. A good candidate commonly used in forecast science is the ROC curve (Gneiting and Vogel, 2018). However, in the case of Model GE, all the ROC curves, except the climatological one, coincide whatever the event, which illustrates its invariance under calibration (Kharin and Zwiers, 2003). Further alternatives should thus be investigated.

3. The CRPS as a random variable

3.1. The random CRPS and its properties

Section 2 pointed out the difficulty of summarizing forecast performance for meaningful comparisons for extreme observations. We illustrated in particular that a single number such as the mean of the CRPS, or its weighted counterpart, fails to deliver relevant comparisons. As an alternative, we propose to study the distribution of the CRPS when treated as a random variable, see also Ferro (2017); Bessac and Naveau (2021).

For simplicity, we use the setting and corresponding notations of the benchmark presented in Section 2.3. From equations (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 stage, it is important to remind that a forecast is issue with only a 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 quantities that are available to forecasters. Thus, to evaluate forecasts performance, it is only possible to compute $\mathrm{CRPS}(F_t, y_t)$ for each t. The climatological distribution, that we now note G and whose existence needs to be hypothesised 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:

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

where π is a random permutation of $\{1, \ldots, n\}$. Applying π breaks the conditional dependence between y_t and F_t, quantified by δ_t in the benchmark, creating 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 cdf.

The climatological forecaster is the only forecaster satisfying

$$\mathrm{CRPS}(G, Y) \overset{d}{=} \mathcal{S}^*(G) \overset{d}{=} \mathcal{S}(G) . \tag{8}$$

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

The distributional properties of $\mathcal{S}(F_T)$, $\mathcal{S}^*(F_T)$, and $\mathcal{S}(G)$ give relevant insights on the behavior of the forecaster. For illustration, Figure 1 gives qq-plots of the distributions of $\mathcal{S}^*(F_T)$ 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, illustrating the influence of the loss of information caused by the permutation: such a visual diagnostic summarizes how $\mathcal{S}(F_T)$ and $\mathcal{S}^*(F_T)$ capture relevant information from the conditioning modelled here by the random variable Δ. The right panel of Figure 1 displays these distributions on the probability scale and 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 displayed here for the sole purpose to illustrate how such visual diagnostics behave when calibration is not satisfied. In Figure 1, we can also see that forecast dominance

Table 3: Availability status of the quantities of interest. It can be an a posteriori availability.

Object	Definition	Availability in practice
F_t	Distribution of the forecast for time t	yes
y_t	Observed realisation 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 to $CRPS(F_t, y_t)$	no
$\mathrm{CRPS}_{\mathcal{S}}(F, Y)$	Random variable generated by the $(CRPS(F_t, y_t))_t$	yes
$\mathrm{CRPS}_{\mathcal{S}^*}(F, Y)$	Random variable generated by the $(CRPS(F_t, y_{\pi(t)}))_t$	yes

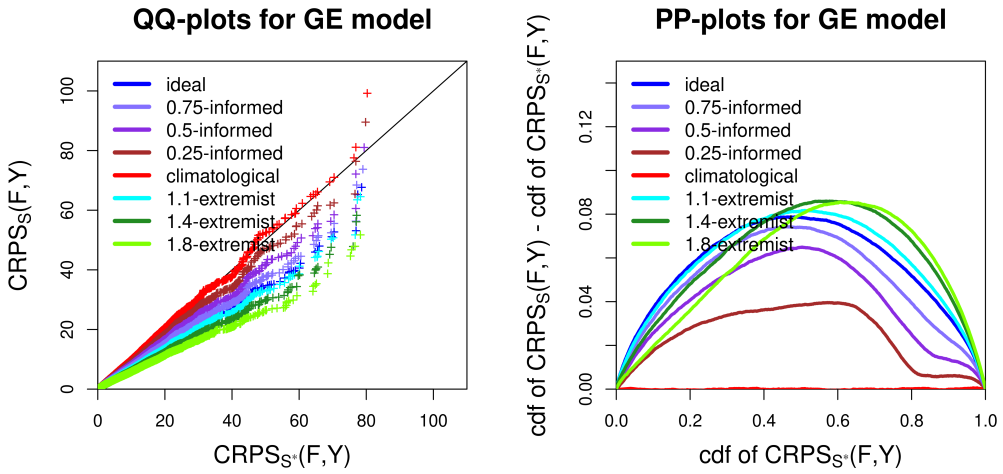

Figure 1: Comparisons of the distributional properties between $\mathcal{S}$ and $\mathcal{S}^*$ for each forecast in model GE: qq-plots (left) and a pp-plots (right panel). Each forecasts are represented by a sample of size $T = 10^6$.

among forecasters could be inferred, as in Ehm et al. (2016, Fig. 1,2,4,6) for point forecasts. Under calibration, discrepancy between distributions can be appropriately interpreted as a direct measure of the forecaster skill (the λ-informed curves never cross each other), making such diagnosis particularly relevant and compliant with the recommendations 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, using EVT to develop a meaningful forecast evaluation for extreme events. To lighten the technicality of this section, all proofs are relegated to Appendix E. In terms of notations with respect to any conditional model that depends on $\Delta = \delta$, we want to emphasize the difference between a conditional forecast, say F_δ, and an unconditional forecast F. Note that δ depends on the time index t, but for notation simplicity, we drop this index; Δ might also change over time but here assumed invariant.

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

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

as u_δ tends to x_{G_δ}, with $1 + \gamma_\delta x > 0$. So at any fixed state δ (state of the atmosphere for a weather forecast, say), the CRPS upper tail behavior (conditionally on $\Delta = \delta$) is equivalent to the observation tail behavior and formalizes what could be intuited from (B.1).

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

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

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

The benchmark presented in Table 1 illustrates these results. The choice of working with a time indexed couple (F_t, Y_t) or with an invariant (G, Y) impacts significantly the tail behavior of the CRPS random variables: according to Table 1, the former case 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. Assessing the forecaster tail behavior

In this section, we propose a tail-equivalent forecast performance index inspired from equations (9), (10), and Figure 1. We aim only to provide the intuition behind the index and leave formal theoretical analysis for future work. We assume that the forecasts lie in the domain of attraction of some distribution $H_{\gamma,\sigma}$. For sufficiently large u, the null hypothesis $H_0 : \mathcal{S}(F_T)|Y > u \overset{d}{=} H_{\gamma,\sigma_u}$ should be rejected for any calibrated forecast with tail behaviour closer to the ideal forecast than the climatological reference.

To go further, assume that the variables in $\mathcal{S}(F_T)$ are iid. This assumption may not be always satisfied, as for instance temperature measures of two consecutive days are likely to be dependent, but can be reasonably satisfied for measurements from 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 $\mathcal{S}(F_T)$ exceeding the threshold u. The empirical nature of $\hat{K}^{(m)}_{\mathcal{S},u}$ allows to simplify $\omega_u{}^2\{\mathcal{S}(F_T)\}$ to

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

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 the computation of Ω^F_u is provided in Table F.4 of Appendix F.

As suggested by Figure 1, we assume that $\Omega^F_u > \Omega^G_u$, for any calibrated forecasts and climatology G. Also, for two calibrated forecasts F^1 and F^2, we conjecture that $\Omega^{F^2}_u \geq \Omega^{F^1}_u$ if F^2 has a tail behaviour closer to the ideal

forecast than F^1. Under these assumptions, we can summarize simply the comparison between Ω_u^F and Ω_u^G through

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

The behaviour of the index T_u is illustrated with the help of model GE; Figure 2 displays the evolution of T_u as a function of the threshold u for $T = 10^6$ and $\gamma = 1/4$. The behaviour of the index is shown to be consistent with our conjecture: first, the ideal forecast performs best, while the climatology has the lowest index. Performance ranking among calibrated forecasters is stable as the threshold increases, with the ideal forecast always obtaining the largest index. The extremist forecasters, displayed here to illustrate the behaviour of the index for non-calibrated forecast, obtain a high index, even larger than the ideal forecast, stressing the importance of calibration which must be carefully assessed before any interpretation of T_u.

In practice, a threshold choice has to be made, for which numerous methodologies have been developed, see, e.g., Beirlant et al. (2004); Papastathopoulos and Tawn (2013); Naveau et al. (2016).

4. Discussion

In this work, we have argued with the help of a carefully designed benchmark 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) have introduced the so-called "Murphy diagrams" for assessing dominance in point forecasts. This original approach allows to appreciate dominance among different forecasts and anticipate their skill area; a similar visual diagnostic is presented in Figure 1 for calibrated forecasts.

Inspired by Friederichs and Thorarinsdottir (2012), we apply EVT directly on common verification measures. By considering the CRPS as a random variable, see also Bessac and Naveau (2021) for non-extreme cases, one can view this contribution as a first step in considering other functionals of the scores 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) and Ferro and Stephenson (2011). We make a link between the paradigm of *maximizing the sharpness subject to calibration* from Gneiting et al. (2007) and the paradigm of *maximizing the information*

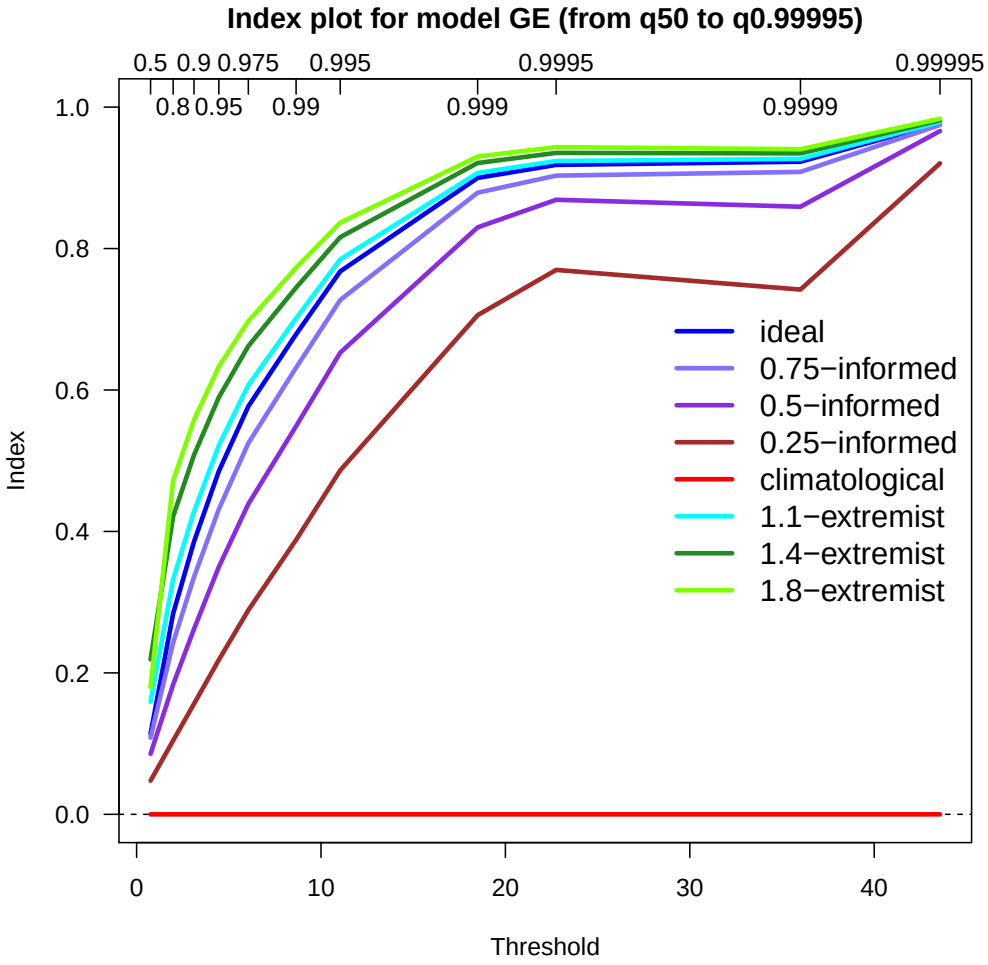

Figure 2: Cramér-von Mises' criterion-based index as a function of the threshold for the different forecasts in model GE with parameters $T = 10^6$ and $\gamma = 1/4$. Indexes are computed for thresholds ranging from the 0.5 to the 0.99995 empirical quantile. Higher index values are assumed to reflect a tail behaviour closer to the ideal forecaster. Validity of the index is limited to calibrated forecast and Non-calibrated extremists forecast are shown to recall that calibration must be first carefully checked before interpreting such graphics.

for extreme events subject to calibration. In a same vein, Murphy (1993) has presented the differences between forecast quality (accordance between

forecasts and observations) and forecast value (ability to bring information to realize a benefit by choosing a forecast), the forecast value seems to be the most important for extreme events, where decision making is crucial. For deterministic weather forecasts, such tools are well-known, see e.g. Richardson (2000); Zhu et al. (2002). Other widely-used scores based on the dependence between forecasts and observed events have been considered in Stephenson et al. (2008); Ferro and Stephenson (2011).

It would be worthwhile to further study the theoretical properties of this CRPS-based tool. Another potentially interesting investigation could be to extend this procedure to other scores like the mean absolute difference, the Dawid-Sebastiani score (Dawid and Sebastiani, 1999) or the ignorance score (Smith et al., 2015; Diks et al., 2011). Classical tools in verification relies on a verification period, as a consequence evaluation is always done a posteriori. Thus, an interesting manner to pursue this work would be to consider sequential evaluation of rare events, in the spirit of the e-values (Vovk and Wang, 2021) introduced to assess and monitor calibration continuously (Arnold et al., 2021). Eventually, we invite scientists to work on new theory of scoring rule departing from the score's averages.

Acknowledgments

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 and the ANR project Melody (ANR-19-CE46-0011). This work was partially supported by the ANR LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007).

Implementation details

The implementation of the index relies on the `extremeIndex` package (Taillardat, 2021a). The `R` code generating simulation data and Figures is available upon request.

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.
URL `https://doi.org/10.1214/19-EJS1622`

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 (Methodological) 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. Annals of Statistics, 65–81.

De Haan, L., Ferreira, A., 2007. Extreme value theory: an introduction. Springer Science & Business Media.

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

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. New York. Springer-Verlag, Berlin.

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

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

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.

Gneiting, T., Vogel, P., 2018. Receiver operating characteristic (roc) curves. arXiv preprint arXiv:1809.04808.

Gumbel, E. J., 1935. Les valeurs extrêmes des distributions statistiques. In: Annales de l'institut Henri Poincaré. Vol. 5. pp. 115–158.

Henzi, A., Kleger, G.-R., Hilty, M. P., Wendel Garcia, P. D., Ziegel, J. F., for Switzerland, R.-.-I. I., 2021. Probabilistic analysis of covid-19 patients' individual length of stay in swiss intensive care units. PloS one 16 (2), e0247265.

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

Holzmann, H., Eulert, M., 2014. The role of the information set for forecasting—with applications to risk management. The Annals of Applied Statistics 8 (1), 595–621.

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

Kharin, V. V., Zwiers, F. W., 2003. On the roc score of probability forecasts. Journal of Climate 16 (24), 4145–4150.

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 & 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., Bouttier, F., Monteil, C., 2021. Strategies for hydrologic ensemble generation and calibration: On the merits of using model-based predictors. Journal of Hydrology 599, 126233.

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

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., Lindley, D. V., Murphy, A. H., Oliver, R. M., Ríos-Insua, D., 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.

Appendix A. Prediction framework and calibration

The theoretical framework considered in this paper is the now classical *prediction space* already introduced by Murphy and Winkler (1987); Gneiting and Ranjan (2013); Ehm et al. (2016), and generalized in a serial context by Strähl and Ziegel (2017). It starts formally with a probability space $(\Omega, \mathcal{A}, \mathbb{Q})$ and a collection of sub-σ-algebras $\mathcal{A}_1, \ldots, \mathcal{A}_k \subset \mathcal{A}$, where $\mathcal{A}_i$ represents the information available to forecaster i. In a meteorological context, it can be seen as the representation of the atmosphere done by each forecaster. In the benchmark considered in Section 2.3, we will consider for simplicity 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 speaking, $F_i : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, 1]$ is a kernel[1] from $(\Omega, \mathcal{A}_i)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, but as done by previous authors, we will identify the kernels with random cumulative cdf, see e.g. Strähl and Ziegel (2017) for more details. For each $x \in \mathbb{R}$, we might in particular use the notation $F_i(x)$ meaning the random element $\omega \mapsto F_i(\omega, (-\infty, x])$.

In such a 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 saying that F_i is *calibrated* with respect to $\mathcal{A}_i$. He additionally defines the *auto-calibration* as the property for F_i to satisfy $F_i = \mathcal{L}(Y|\sigma(F_i))$ almost surely. Here, $\sigma(F_i)$ denotes the σ-algebra generated by F_i, that is to say the smallest σ-algebra such that $\omega \mapsto F_i(\omega, x)$ is measurable for all $x \in \mathbb{R}$. Note 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. As a particular case considered in Section 2.3, the *climatological* forecaster is ideal with respect to the trivial σ-algebra.

In practice, one is not only concerned with predictions for an outcome Y at a single time point. The framework introduced above also allows to deal with independent replicates at times $t = 1, 2, \ldots$, as is done in Section 2.3. If

[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. See e.g. Kallenberg (2017).

such an assumption of independence sounds unrealistic in several situations, as argued by Strähl and Ziegel (2017), it can nevertheless provide a first step and takes advantage of a lighter context. We chose therefore to keep it in this paper for simplicity.

Appendix B. An alternative expression of the weighted CRPS

The weighted CRPS defined by (2) can be reformulated in the following way, as soon as the weight function $w(.)$ is continuous,

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

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

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

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

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

and

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

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 $(0,1)$. As $W(x)$ is non-decreasing, one has $\{W(X) > W(y)\} = \{X > y\}$, and it follows that

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

as announced in (B.1).

Appendix C. Proof of the inequality (4)

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

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

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

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

Note that the decreasingness of $\overline{F_u}$ yields in particular that for all x,

$$\overline{F_u}(x) \leq \overline{G}(x)\ . \tag{C.3}$$

Besides, equation (C.2) and the monotonicity of W allows to write that for any $x \leq u$

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

Equality (B.1) implies that

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

where

$$\Delta(x) = \mathbb{E}_G[(W(Y) - W(x))\mathbf{1}\{Y \leq x\}] - \mathbb{E}_{F_u}[(W(X_u) - W(x))\mathbf{1}\{X_u \leq x\}].$$

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

$$\frac{1}{2}\left|\mathbb{E}_G[\mathrm{wCRPS}(F_u, Y)] - \mathbb{E}_G[\mathrm{wCRPS}(G, Y)]\right| \leq \int_u^{x_G} \Delta(x) dG(x). \tag{C.5}$$

For $x > u$ we can write that

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

since $W(Y) - W(x) \leq 0$ in the first expectation, whereas $0 \leq W(x) - W(X_u) \leq W(x) - W(u)$ in the second one. As a consequence, one gets

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

This last expression combined with (C.5) leads finally to

$$|\mathbb{E}_G[\text{wCRPS}(F_u, Y)] - \mathbb{E}_G[\text{wCRPS}(G, Y)]| \leq 2\overline{G}(u) \int_u^{x_G} (W(x) - W(u)) dG(x).$$

Note 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 G and G can be as closed as one wishes. The right hand side goes to 0 due to the finite mean of $W(Y)$.

Appendix D. A detailed example related to Section 2.2

In this appendix, we illustrate the fact that the CRPS fails at discriminating forecasts with different tails. We consider GP distributed forecasts and observations. In this case, closed form of the CRPS are available, as detailed in the following.

Lemma 1. *Consider $X \overset{d}{=} \text{GP}(\beta, \xi)$ and $Y \overset{d}{=} \text{GP}(\sigma, \gamma)$ with $0 \leq \xi < 1$ and $0 \leq \gamma < 1$, with 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*

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

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

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

Proof: Applying (B.1) with $W(y) = y$, and making use of classical properties of the Pareto distribution (see e.g. (Embrechts et al., 1997, Theorem 3.4.13)), one gets

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

It follows that

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

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

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

one can write

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

Besides, as $G^{-1}(v) = \frac{\sigma}{\gamma}\left((1 - v)^{-\gamma} - 1 \right)$, one can thus rewrite, denoting by U a random variable uniformly distributed on $(0, 1)$,

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

If $c = \dfrac{\xi\sigma}{\beta\gamma} = 1$, then this simplifies to

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

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

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

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

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

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

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

concluding the proof of Lemma 1. □

Lemma 1 allows 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

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

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

Counter examples as the previous one can thus be found, illustrating how weighted scoring rules fail to compare tail behaviors. They should therefore be handled with a particular care, especially for forecast makers, as already advocated by Gilleland et al. (2018); Lerch et al. (2017).

Appendix E. Proof of the convergences (9) and (10)

The proof of (10) can be seen as a particular case of (9), so that we will focus on proving (9). The following lemma will help to get the result, and is presented first with its proof. In what follows, the mean excess function of any random variable Z with finite mean and with cdf F will be denoted by $M(F,z)$, so 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 domain of attraction $\mathcal{D}(H_\gamma)$ with $\gamma < 1$. There exist non negative real numbers α and β such that for each $z \in \mathbb{R}$,

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

Proof of the lemma: The indicator function $\mathbb{1}_{Z>z}$ implies that we always have $0 \leq 2\mathbb{E}_F((Z-z)\mathbb{1}_{Z>z})$. 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. Note first that if $z > x_F$, then (E.1) is trivially true. Let then show the result when $z \overset{<}{\to} x_F$, and for this, let 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) show that $M(F,z) \sim \gamma z/(1-\gamma)$ as z tends to x_F, and we can conclude directly.
2. F belongs to $\mathcal{D}(H_\gamma)$ with $\gamma < 0$: In this case, the result also follows easily from Embrechts et al. (1997) since when z tends to x_F, $M(F,z) \sim \gamma(x_F - z)/(\gamma - 1)$. This allows 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 (see e.g. Theorem 3.9 in Ghosh and Resnick (2010)). If x_F is finite, then there exists a positive β such that $2M(F,z) \leq \beta$ and α can be fixed to 0, whereas if x_F is infinite, the fact that $2M(F,z) < z$ for z large enough enables to conclude.

So far, we have shown that, for some large z_0, there exist non negative α and β such that

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

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

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

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

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

We have now two cases: either $\beta < \frac{\beta_0}{\overline{F}(z_0)}$ or $\beta \geq \frac{\beta_0}{\overline{F}(z_0)}$. In the latter case, we have $2\mathbb{E}_F((Z-z)\mathbb{1}_{Z>z}) \leq \beta_0 \leq \overline{F}(z)(\alpha z + \beta)$, and so, the required result is obtained. In the case of $\beta < \frac{\beta_0}{\overline{F}(z_0)}$, it is always possible to increase 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 announced.

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 way

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

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

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

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

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

This holds when the right end point of Y is non-negative. If this was not the case, note that one can simply write $Y \leq \mathrm{CRPS}(F, Y) + c \leq Y + \beta\overline{F}(u) \quad a.s.$.

The main idea of the proof is to notice that $\overline{F}(u)$ goes to zero as u gets large, and consequently, the above inequalities 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 similarly for large u. The choice of positive constant $b(u)$ depends on the domain of attraction of Y. More precisely, we assume that $Y[u]$ converges in distribution towards a GPD with finite mean. So that

$$0 \leq \mathbb{P}\left(\frac{\mathrm{CRPS}(F, Y) + c - u}{b(u)} > t \mid Y > u\right) - \mathbb{P}\left(\frac{Y-u}{b(u)} > t \mid Y > u\right)$$

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

We recognize the probability (conditionally on $Y > u$) for Y to be 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 consists in 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 u large enough, the latter probability can be approximated by a GPD, so that

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

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

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

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

Appendix F. Algorithm for the computation of the Cramer-von-Mises criterion

Table F.4: Computation of Cramér-von Mises' statistic from N couples forecast/observation. It can be done with the `R` package `extremeIndex` (Taillardat, 2021a).

0. CRPS estimates for each forecaster:	- For the N couples forecast/observation, compute their corresponding instantaneous CRPS.
1. Estimation of γ 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	- Order the m CRPS values where the observation $y \geq w$ in increasing order $s_1, \ldots, s_m$.
For $i \in [1, m]$	-Compute for each CRPS value s_i, $H_{\gamma,\sigma_w}(s_i)$.
	-Compute $\left[\frac{2i-1}{2m} - H_{\gamma,\sigma_w}(s_i)\right]^2$.
End 3.	
End 2.	

Note that for large 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 have been computed, but they are actually subject to numerical instabilities (Prokhorov, 1968; Csörgő and Faraway, 1996). Furthermore, Ω_u^F is sufficient to compare the effect size of the deviation.