Extreme events evaluation using CRPS distributions

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

Extreme events evaluation using CRPS distributions

Maxime Taillardat[1], Anne-Laure Fougères[2], Philippe Naveau[3], Raphaël de Fondeville[4]

[1] CNRM UMR 3589, Météo-France/CNRS, Toulouse, France

[2] Univ. Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France

[3] Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212, CEA-CNRS-UVSQ, IPSL & U Paris-Saclay, Gif-sur-Yvette, France

[4] Institute of Mathematics, Ecole Polytechnique Fdrale de Lausanne, Station 8, Lausanne, Switzerland

Abstract

Verification of ensemble forecasts for extreme events remains a challenging question. The general public as well as the media naturely pay particular attention on extreme events and conclude about the global predictive performance of ensembles, which are often unskillful when they are needed. Ashing classical verification tools to focus on such events can lead to unexpected behaviors. To square up these effects, thresholded and weighted scoring rules have been developed. Most of them use derivations of the Continuous Ranked Probability Score (CRPS). However, some properties of the CRPS for extreme events generate undesirable effects on the quality of verification. Using theoretical arguments and simulation examples, we illustrate some pitfalls of conventional verification tools and propose a different direction to assess ensemble forecasts using extreme value theory, considering proper scores as random variables.

Keywords: Scoring rules; verification, ensemble forecasts; CRPS ; extreme events ; calibration.

1. Introduction

In a pioneering paper on forecast verification, Murphy (1993) distinguished three types of forecast "goodness": the quality that quantifies the

adequacy with what actually happened, the consistency based on the fidelity to an expert knowledge, and the value that describes how the forecast helps the decision maker to proceed efficiently. The quality of a forecast is often summarized by one scalar. For example, to identify the best forecast, one classically takes the mean on a validation period of proper scoring rules (see, e.g., Matheson and Winkler, 1976; Gneiting and Raftery, 2007; Schervish et al., 2009; Tsyplakov, 2013). Proper scoring rules can be decomposed in terms of reliability, uncertainty and resolution. Several examples of such decompositions can be found in Hersbach (2000) and Candille and Talagrand (2005). Bröcker (2015) showed that resolution is strongly linked with discrimination. Resolution and reliability can also be merged into the term calibration, and Gneiting et al. (2007) suggested to *maximize the sharpness subject to calibration*. Note that the sharpness is the spread of the forecast, and it is a property of the forecast only. In ensemble forecasts' verification, the most popular scoring rule is the Continuous Ranked Probability Score (CRPS) (see, e.g., Epstein, 1969; Hersbach, 2000; Bröcker, 2012) and it can be defined as

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

where $y \in \mathbb{R}$, and X and X' are two independent random copies coming from a given continuous cumulative distribution function (cdf) F. Hence, the CRPS is a proper[1] score that makes the link between the observed value y and the forecast distribution F. The second line in Equality (1) highlights the two terms of calibration and sharpness.

Regarding extremes verification, it is important to counteract some cognitive biases bounding to discredit skillful forecasters (examples of cognitive biases can be found in Kahneman and Tversky (1979); Morel (2014)). That is what is called in Lerch et al. (2017) the "Forecaster's dilemma". Citing these authors, "In the public, forecast evaluation often only takes place once an extreme event has been observed, in particular, if forecasters have failed to predict an event with high economic or societal impact". Indeed, the

[1] A proper score like the CRPS satisfies: $\mathbb{E}_{Y\sim G}(\text{CRPS}(G,Y)) \leq \mathbb{E}_{Y\sim G}(\text{CRPS}(F,Y))$.

only remedy is to consider all available cases when evaluating predictive performance. Proper weighted scoring rules (Gneiting and Ranjan, 2011; Diks et al., 2011) attempt to emphasize predefined regions of interest. Following the notation of Equation (1), the weighted CRPS can be defined as

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

where $w(x)$ is a non negative function and $W(x) = \int_{-\infty}^{x} w(t)dt$. Alternatively, the weighted CRPS can also be expressed as the following:

$$\begin{aligned} wCRPS(F,y) &= W(y) + 2\overline{F}(y)\mathbb{E}_F(W(X) - W(y)|X > y) \\ &\quad -2\mathbb{E}_F(W(X)F(X)) \qquad (3) \end{aligned}$$

as soon as the weight function $w(\cdot)$ is continuous (see proof in Appendix 6.1). Who among different users (e.g., forecast users and forecasters) should choose this weight function remains a complex issue (see, e.g. Ehm et al., 2016; Gneiting and Ranjan, 2011; Patton, 2014). Even in this case where $w(x)$ can be objectively chosen with respect to an application at hand, one can wonder if the corresponding weighted CRPS captures well the extreme behavior of the observational records, i.e discriminating between two competitive forecasts with respect to extreme events. This leads to the question of how to model accurately the distributional features of the forecast and observational vectors.

In this work, we move away from looking at averages like the properness of a score. Instead we propose a different framework that considers the observational vector, not as a realization, but as a random variable with a specific extreme value behavior. This change of view with regards to scoring rules brings us to study the distribution of the CRPS itself. It naturally suggests to bridge the *random variable CRPS* with the field of Extreme Value Theory (EVT).

The foundations of this theory are laid in De Haan (1970), see e.g. the books of Embrechts et al. (1997); Beirlant et al. (2004); De Haan and Ferreira (2007). EVT provides probabilistic models to represent the distributional behavior of large values, i.e. excesses above a large threshold. Roughly speaking, it is based on the survival function of the so-called Generalized

Pareto (GP) distribution with shape parameter $\gamma \in \mathbb{R}$ defined by

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

for each x such that $1+\gamma x/\sigma > 0$ and $\sigma > 0$ represents a scale parameter. For $\gamma = 0$, this can be viewed as the classical exponential tail. The fundamental property of this GP family is its stability with respect to thresholding (see, e.g., Embrechts et al., 1997, Theorem 3.4.13 (c)). As pinpointed by Friederichs and Thorarinsdottir (2012), it is possible to express explicitly the $CRPS(F, y)$ whenever F is a GP distribution and the real y is fixed. Starting from this link between CRPS and extreme values, it would be of interest to know what would be the behavior of $CRPS(F, y)$ when the observational vector y becomes a random variable that takes very large values.

This work is organized as follows. In Section 2 we point out some undesirable properties of the CRPS and its weighted counterpart. We derive the non-tail equivalence of the wCRPS and highlight potential difficulties of using the wCRPS for extreme weather evaluation. Section 3 sets the theoretical framework in which the temporal aspects in scoring rules are stressed. Section 4 links the observational tail behavior with the CRPS tail behavior. An index which captures some quality w.r.t. extremes forecast is introduced. The work closes with a discussion in Section 5.

2. Tail equivalence, weighted CRPS and choice of a weight function

2.1. Tail equivalence and (weighted) CRPS

Focusing on the upper tail behavior study, it is convenient to recall the definition of tail equivalence between two cdf F and G (see, e.g., Embrechts et al., 1997, Section 3.3). They are said to be *tail equivalent* if they share the same upper endpoint $x_F = x_G$ and if their survival functions $\overline{F} = 1 - F$ and $\overline{G}$ satisfy $\lim_{x \to x_F} \overline{F}/\overline{G} = c \in (0, +\infty)$. Another useful EVT concept is the notion of domain of attraction: a distribution F is said to belong to the domain of attraction of the GP distribution H_γ, denoted $F \in \mathcal{D}(H_\gamma)$, if for some γ, some positive auxiliary function b, the rescaled survival function converges as u tends to x_F to a GP with shape parameter γ, i.e.

$$\frac{\overline{F}(u + zb(u))}{\overline{F}(u)} \longrightarrow \overline{H}_\gamma(z), \tag{4}$$

for each z such that $1+\gamma z > 0$. Even if the weighted $CRPS$ is proper, there is no guarantee of tail equivalence[2]. More precisely, for any positive ϵ, it is always possible to construct a cdf F that is *not* tail equivalent to G such that

$$|\mathbb{E}_G(w\text{CRPS}(G,Y)) - \mathbb{E}_G(w\text{CRPS}(F,Y))| \leq \epsilon. \tag{5}$$

This result is proven in Appendix 6.2. The implication of (5) is that a forecast F could have a misspecified tail (one which is not tail equivalent to the "true" forecast G) and still score almost as well as the true forecast. To illustrate the inability to distinguish among different tail behaviors, the upcoming section investigates the case of a particular choice of weight function. This choice will also be considered later in Section 3 in a forecasts comparison perspective.

2.2. The quantile weight function

Since our interest concerns jointly the CRPS and the upper tail behavior, the quantile weight function appears as a natural candidate to highlight upper tail features. It is simply defined as $w_q(x) = \mathbf{1}\{x \geq q\}$ for any real q. The following lemma, see Appendix 6.3. for its proof, links the CRPS and its weighted version.

Lemma 1. *The weighted quantile CRPS defined by*

$$wCRPS(F,y;q) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 \mathbf{1}\{x \geq q\}, dx,$$

where q represents any real number can be written as

$$wCRPS(F,y;q) \;=\; \int_q^{\infty} \overline{F}^2(x)dx + \begin{cases} 0, & \text{if } q > y, \\ CRPS(F,y) - CRPS(F,q), & \text{if } y \geq q. \end{cases}$$

The following conditional equality in distribution holds for any random variable Y:

$$[CRPS(F,Y)|Y \geq q] \overset{d}{=} [wCRPS(F,Y;q) - c_F(q)|Y \geq q], \tag{6}$$

[2]It may explain the weakness in estimating the shape parameter of a GP distribution via the CRPS based inference, see Friederichs and Thorarinsdottir (2012).

where the constant

$$c_F(q) = CRPS(F, q) - \int_q^\infty \overline{F}^2(x)dx$$

only depends on the forecast distribution F and the constant q.

In terms of extremes, Equation (6) tells that the distributional behavior of this weighted CRPS above the threshold q is identical to the classical CRPS conditional on $Y \geq q$, up to the constant $c_F(q)$.

2.3. The CRPS for the Pareto case

When both the forecast (with cdf F) and observational vectors (with cdf G) are Pareto distributed, explicit computations of the CRPS can be made, see the following lemma and its proof in Appendix 6.4.

Lemma 2. *Assume that $X \overset{d}{=} Pareto(\beta, \xi)$ and $Y \overset{d}{=} Pareto(\sigma, \gamma)$ with $0 \leq \xi < 1$ and $0 \leq \gamma < 1$, with respective survival functions $\overline{F}(x) = (1+\xi x/\beta)^{-1/\xi}$ and $\overline{G}(x) = (1+\gamma x/\sigma)^{-1/\gamma}$ for any $x > 0$. If $\gamma/\sigma = \xi/\beta$, with $\gamma \neq 0$, then*

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

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

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

In particular, this lemma tells us that if the GP forecast parameter is proportional to the GP ideal parameter, i.e. $\beta = a\sigma$ and $\xi = a\gamma$ for some $a > 0$, then we can study the effect of changing the forecast tail behavior captured by ξ and the spread forecast encapsulated in β. In this case, the CRPS can be written as a function of a, say

$$\mathbb{E}_G\left[CRPS(F, Y)\right] = \phi_\gamma(a) = \frac{\sigma}{1-\gamma} + 2a\sigma\left[\frac{1}{2(2-a\gamma)} - \frac{1}{1+a-a\gamma}\right]. \quad (7)$$

Figure 1 shows how this CRPS varies in function of a (x-axis) when we fix $\sigma = 1$ and $\gamma = 0.1$ (left panel) or $\gamma = 0.4$ (right panel). The important point is that no meaningful conclusions about the upper tail previsions from two competing forecasters can be made within the blueish "cup" region. Inside

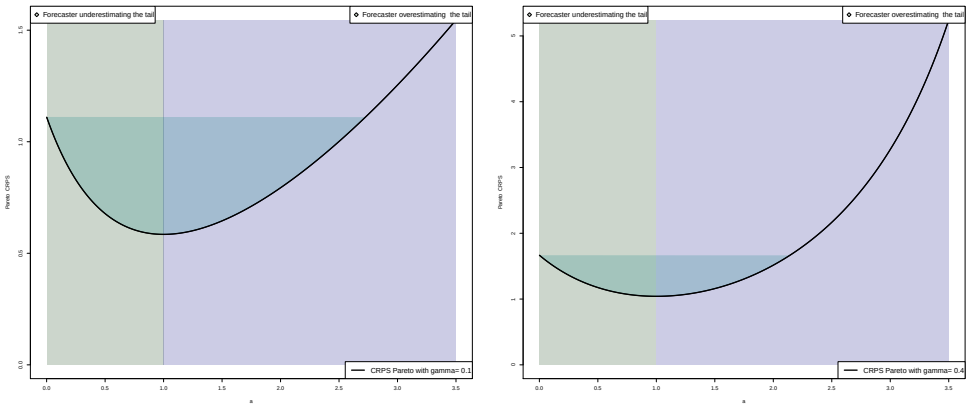

Figure 1: Function $\phi_\gamma(a)$ defined by (7) computed between $X \overset{d}{=} \text{Pareto}(a, a\gamma)$ and $Y \overset{d}{=} \text{Pareto}(1, \gamma)$ with $\gamma = .1$ (left panel) and $\gamma = .4$ (right panel). Whenever $a > 1$, the forecaster overestimates the true upper tail behavior. The opposite can be said when $a < 1$. The ideal forecast corresponds to $a = 1$. The blueish "cup" corresponds to the region where a same value of $\mathbb{E}_G\left[CRPS(F, Y)\right]$ (y-axis) provides an ambiguity, i.e. two forecasters can issue two very different values of a (x-axis), either greater or lower than one, for the same score value.

this ambiguous region, the same value of $\mathbb{E}_G\left[CRPS(F, Y)\right]$, y-axis, can be given by two different values of a. One a is greater than one, corresponding to a forecaster prone to a risky strategy, while another a is smaller than one, made by a forecaster risk averse. The spread of this region over a is not small, and spans from $a = 0$ to a around 2. The left panel in Figure 1 implies that two hypothetical forecasters, one multiplying by 2 the true extreme quantiles and the other dividing them by two, cannot be differentiated, both have $\phi_\gamma(a)$ around .7. In other words, the difference between two inaccurate forecasters with opposite risk views have to be very pronounced to be correctly ranked by the CRPS. To numerically assess this drawback, we can compute the area of this ambiguous region. The area delimited by the blueish cup can be written as

$$\begin{aligned} A(\gamma) &= \phi(0) \times a_0 - \int_0^{a_0} \phi_\gamma(a) da, \text{ with } a_0 = \frac{3}{1+\gamma} \text{ and } \phi(a_0) = \phi(0), \\ &= 2\sigma \left[\int_0^{a_0} \frac{a}{1 + a - a\gamma} da - \int_0^{a_0} \frac{a}{2(2 - a\gamma)} da \right], \end{aligned}$$

$$
\begin{aligned}
&= 2\sigma \left[\frac{1}{1-\gamma} \int_0^{a_0} \left(1 - \frac{1}{1+(1-\gamma)a} \right) da + \frac{1}{2\gamma} \int_0^{a_0} \left(1 - \frac{2}{(2-a\gamma)} \right) da \right], \\
&= 2\sigma \left[\frac{a_0(1+\gamma)}{2\gamma(1-\gamma)} - \frac{1}{(1-\gamma)^2} \log(1+(1-\gamma)a_0) + \frac{1}{\gamma^2} \log(1-a_0\gamma/2) \right].
\end{aligned}
$$

By plotting this function (graph not shown here but available upon request), one can notice that this integral varies from between .9 and 1 for $\gamma \in [0, .5]$ and $\sigma = 1$. So, despite that the blueish "cup" region appears smaller in the right panel than in the left one, this impression is just due to the y-axis scale. The area does not vary much and the problem aforementioned still remains over a wide range of γ.

3. Framework, examples and forecast comparison

3.1. Theoretical framework and examples

Our theoretical framework is the now classical *prediction space* already introduced by Murphy and Winkler (1987); Gneiting et al. (2013); Ehm et al. (2016). A probabilistic forecast for a real-valued outcome Y is identified with its cdf F. The elements of the associated sample space Ω can be identified with tuples of the form

$$(F_1, \ldots, F_t, Y)$$

where the predictive cdfs $F_1, \ldots, F_t$ use information sets $\Delta_1, \ldots, \Delta_t \subseteq \Delta$, respectively, where Δ is a general set with nice theoretical properties[3]. See e.g. Ehm et al. (2016, section 3.1) for details. As an illustration, if Y is a weather variable of interest, with (unconditional) distribution function G, Δ_t can for example depict the underlying atmospheric state at time t, see e.g. tables 1 and 2 for two examples of conditioning. Following Gneiting et al. (2007), the outcome Y can be viewed as a true data-generating process coming from $G_t := G|\Delta_t$ for each time step t. The associated sequence of probabilistic forecasts is F_t for each t, and a forecast F_t is *ideal* relative to Δ_t if $F_t = G_t$.

The asymptotic compatibility between the data-generating process and the predictive distributions can be divided in the three following modes of calibration, where an arrow denotes almost sure convergence as T tends to

[3]More precisely, Δ is a sigma-field on Ω.

infinity:

$$(\mathcal{P}): \qquad \frac{1}{T}\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) \to p \quad \text{for all } p \in (0,1)\ .$$

$$(\mathcal{E}): \qquad \frac{1}{T}\sum_{t=1}^{T} G_t^{-1} \circ F_t(x) \to x \quad \text{for all } x \in \mathbb{R}\ .$$

The third type of calibration $(\mathcal{M})$ can be summed up by the existence of the limit distributions F and G such as $F = G$. The existence of G is a natural assumption in meteorology and corresponds to the existence of a stable climate. Hence, $(\mathcal{M})$ can be interpreted in terms of the equality of observed and forecast climatology.[4]

For seek of illustration, let us now consider two different designs of experiments. The first one is the design introduced by Gneiting et al. (2007), also considered by Straehl and Ziegel (2017), and described here in Table 1. At times $t = 1, 2, ...$, the cdf of the truth (observation) is modeled by a normal distribution with mean Δ_t and unit variance, where Δ_t is a random draw from the standard normal distribution. This model is denoted in the following by Model NN. The *ideal* forecaster provides a forecast of $\mathcal{N}(\Delta_t, 1)$. The *climatological* forecaster corresponds to the unconditional distribution, i.e. a centered normal distribution with variance equal to 2. The *unfocused* forecaster adds a Rademacher-type bias in his/her forecast. Finally, the *extremist* forecaster adds an additive bias in his forecast. The Bayesian reader can notice that this design of experiments relies on the conjugate prior of the normal distribution. For extreme events concerns, this example can be limiting, since it involves light tails and exhibits a slow convergence of the large values. As a consequence, we introduce a second design of experiment in Table 2, based on Gamma-exponential mixtures and denoted by Model GE. More precisely, suppose that $[Y_t|\Delta_t]$ follows an exponential distribution with mean $1/\Delta_t$, and Δ_t follows a Gamma pdf $f_{\Delta_t}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}\exp(-\beta x)$, with $x > 0$, $\alpha > 0$ and $\beta > 0$. As the Gamma Laplace transform can be written as $\mathbb{E}\left[\exp\left(-x\Delta_t\right)\right] = \left(1 + \frac{x}{\beta}\right)^{-\alpha} = \overline{H}_\xi(x)$ whenever $\alpha = \beta = 1/\xi$,

[4] In Gneiting et al. (2007), the sequence $(F_t)_t$ is then respectively said *probabilistically calibrated* (in short, *calibrated*), *exceedance calibrated*, and *marginally calibrated*. Besides, the sequence $(F_t)_t$ is said *auto-calibrated* when it is jointly probabilistically and marginally calibrated.

the unconditional Y has a GP tail with parameter ξ. Hence, the *climatological* forecaster corresponds to the unconditional distribution, i.e. a Pareto distribution, while the ideal forecaster should follow $[Y_t|\Delta_t]$ an exponential distribution with mean $1/\Delta_t$. The *unfocused* forecaster adds a uniform-type bias in his forecast, whereas the *extremist* forecaster adds a multiplicative bias in his forecast.

Table 1: Model NN (Normal-normal). Both samples $(\Delta_t)_t$ and the $(\tau_t)_t$ are mutually independent of each other.

Truth	$Y_t \overset{d}{=} \mathcal{N}(\Delta_t, 1)$ where $\Delta_t \overset{d}{=} \mathcal{N}(0,1)$	
Forecasts		Properties
Ideal forecaster	$\mathcal{N}(\Delta_t, 1)$	$\mathcal{PEM}$
Climatological forecaster	$\mathcal{N}(0, 2)$	$\mathcal{P}\overline{\mathcal{E}}\mathcal{M}$
Unfocused forecaster	$\frac{1}{2}(\mathcal{N}(\Delta_t, 1) + \mathcal{N}(\Delta_t + \tau_t, 1))$ with $\tau_t = \pm 2$ with $1/2$ probability each	$\mathcal{PE}\overline{\mathcal{M}}$
Extremist forecaster	$\mathcal{N}(\Delta_t + 5/2, 1)$	$\overline{\mathcal{PEM}}$

Table 2: Model GE (Gamma-Exponential). Both samples $(\Delta_t)_t$ and the $(\tau_t)_t$ are mutually independent of each other. For the unfocused forecaster, the $\mathcal{P}$ calibration gives $p + \varepsilon(p)$ with $|\varepsilon(p)| < 0.0051$.

Truth	$Y_t \overset{d}{=} \mathcal{E}\mathrm{xp}(\Delta_\mathrm{t})$ where $\Delta_t \overset{d}{=} \Gamma(4,4)$	
Forecasts		Properties
Ideal forecaster	$\mathcal{E}\mathrm{xp}(\Delta_\mathrm{t})$	$\mathcal{PEM}$
Climatological forecaster	Pareto cdf with $\sigma = 1$ and $\gamma = 1/4$	$\mathcal{P}\overline{\mathcal{E}}\mathcal{M}$
Unfocused forecaster	$\mathcal{E}\mathrm{xp}(\Delta_\mathrm{t}/\tau_\mathrm{t})$, with $\tau_t = 2/3 * U_1 + 1/3 * U_2$ where $U_1 \overset{d}{=} \mathcal{U}[1/2, 1]$ and $U_2 \overset{d}{=} \mathcal{U}[1,2]$	$\mathcal{PE}\overline{\mathcal{M}}$
Extremist forecaster	$\mathcal{E}\mathrm{xp}(\Delta_\mathrm{t}/1.5)$	$\overline{\mathcal{PEM}}$

3.2. Comparing forecasts with expected weighted CRPS

This subsection compares with the quantile weight function different forecasts based on the two designs presented in tables 1 and 2, (see also, Diebold et al., 1997; Dawid, 1984). In Figure 2, the expected weighted CRPS (at log scale) for each model (model GE: top, model NN: bottom) is plotted for each forecast according to q (left) and the corresponding ranking between forecasts (right). Properness of the wCRPS ensures that the ideal forecaster has the lowest score expectation. The dotted horizontal lines correspond to the unweighted CRPS expectation. Concerning the latter, the $CRPS(F_t, y)$ for the ideal forecaster can be expressed explicitly (see, e.g., Friederichs and Thorarinsdottir, 2012, and Equation (20) in Appendix 6.4) as

$$CRPS([G|\Delta_t], y) = y + \frac{2}{\Delta_t}\exp(-\Delta_t y) - \frac{3}{2\Delta_t}, \text{ where } [G|\Delta_t] \stackrel{d}{=} \mathcal{E}\mathrm{xp}(\Delta_t), \quad (8)$$

and for the climatological forecaster, as

$$CRPS(G, y) = y - \frac{8}{3}[1 - (\overline{H}_{0.25}(y))^{0.75}] + \frac{4}{7}. \quad (9)$$

From Figure 2, two important features can be identified: (1) the loss of information for large values, well-known by Brier score users and clearly pointed out by Stephenson et al. (2008), and (2) the fact that forecast rankings can switch order. A change in the forecast ranking is observed for the model GE in Figure 2. This can, artificially, favor a forecaster over another one. For example, the climatological forecaster could be tempted to only show his/her results for a weight function with $q > 0.96$ in order to outperform the extremist forecaster.

One can argue that the weight function should not be chosen by forecast makers but rather by forecast users (see e.g. Ehm et al., 2016). This opinion was also expressed by Gneiting and Ranjan (2011) who wrote that "the scoring function [must] be specified ex ante" and relayed by Patton (2014) "forecast consumers or survey designers should specify the single specific loss function that will be used to evaluate forecasts". This is a well-known fact for decision makers in economics. In contrast, users of weather forecasts may be less aware of potential negative impacts of poor forecasts. A recent survey (Hagedorn, 2017) concludes that 81% of probabilistic information users relies on heuristic/experience rather than on cost/loss models in order to take decisions. Last but not least, given the proximity between the wCRPS values it

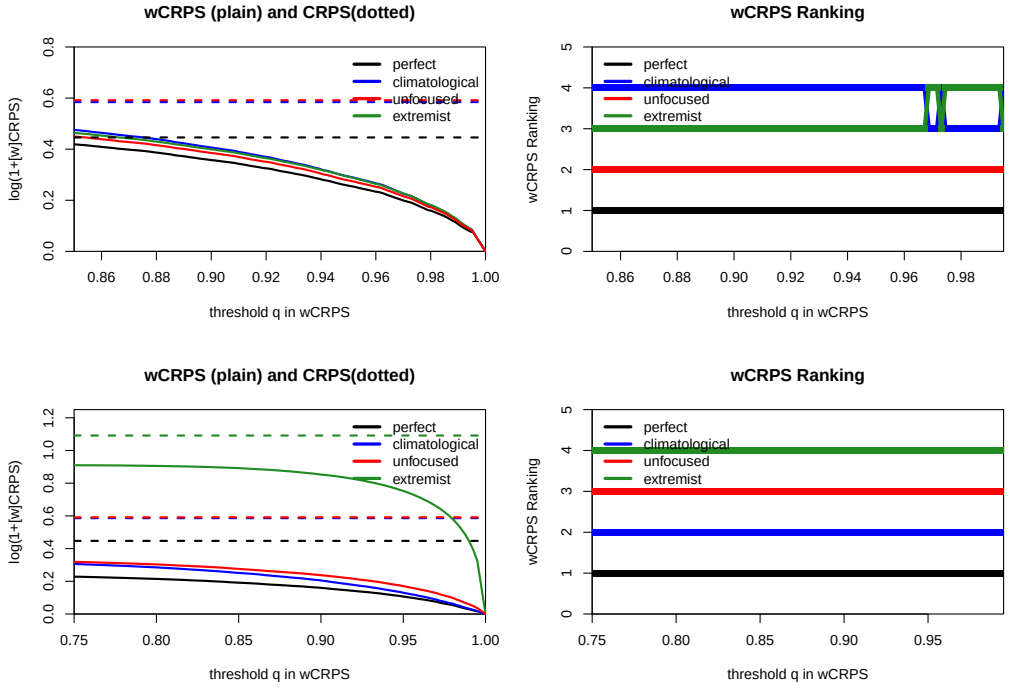

Figure 2: Time-averaged values of $\log(1 + \mathbb{E}_{G_t}[\mathrm{CRPS}(F_t, Y_t)])$ (dotted) and $\log(1 + \mathbb{E}_{G_t}[\mathrm{w_qCRPS}(F_t, Y_t)])$ (plain) among with $W_q(x) = x\mathbf{1}\{x \geq q\}$ as functions of quantiles q (left). Time t varies between 1 and 10^7. Ranking of the different forecasts according to the w_qCRPS (right). The model GE is on top, the model NN is on bottom. The forecast ranking can switch between very close weight functions. For high thresholds, it is not possible to distinguish a clear forecast ranking among forecasters.

seems highly valuable to compute tests of equal predictive performance. For example, Table 3 presents the two-sided pairwise Diebold-Mariano statistic values on the wCRPS among forecasters of the model GE for 2 different quantiles $(0.875, 0.975)$. These quantiles are before and after the ranking switch between two forecasters. Table 3 clearly shows that the ranking of Figure 2 can be challenged. From this section and Section 2.3, one can conclude that weighting scoring rules have to be handled with care, especially for forecast makers (Gilleland et al., 2018; Lerch et al., 2017).

Table 3: Diebold-Mariano statistic values for two wCRPS functions among forecasters for the model GE. The equal predictive performance is not rejected at the level 0.05 for the values in bold. A positive value means that the row forecast is better than the column one. For the quantile 0.975, the value between the climatological and the extremist forecasters are in contradiction with Figure 2, but non significant. This Table is built from 10^5 observation/forecast pairs which are independent with those used for Figure 2.

Quantile 0.875				
	Ideal	Climatological	Unfocused	Extremist
Ideal		44.7	36.9	48.3
Climatological	-44.7		-23.1	-9.94
Unfocused	-36.9	23.1		12.0
Extremist	-48.3	9.94	-12.0	
Quantile 0.975				
	Ideal	Climatological	Unfocused	Extremist
Ideal		21.7	21.7	27.7
Climatological	-21.7		-9.24	**2.61**
Unfocused	-21.7	9.24		7.19
Extremist	-27.7	**-2.61**	-7.19	

4. A CRPS-based tool using extreme value theory

4.1. CRPS behavior and calibration-information diagram

Ferro (2007) proposed to link EVT with forecast verification of extreme events by giving a theoretical framework to characterize the joint distribution of forecasts and rare events. This concerned deterministic forecasts and other tools are needed, see the quote by Stephenson et al. (2008) "development of verification methods for probability forecasts of extreme events is an important area that clearly requires attention". A difficulty is to summarize forecast knowledge in an informative way in order to make meaningful comparisons for extreme observations.

In sections 2 and 3, we saw that a single number as the mean of the CRPS (or the mean of the weighted CRPS) has many drawbacks to compare forecasts with extreme observations. Instead of focusing on one number, we propose in this section to study the distribution generated by these CRPS as random variables. To fix ideas, consider the model GE. From equations (8) and (9), it is possible to view these ideal and climatological scores as

random variables whenever Y is itself considered as a random variable. The properness of the CRPS ensures here that for each t

$$\mathbb{E}_{[G|\Delta_t]}[CRPS([G|\Delta_t], Y_t)] \leq \mathbb{E}_{[G|\Delta_t]}[CRPS(G, Y_t)].$$

At this stage, it is important to recall what can be considered as observed in applications. In practice, Δ_t is not given to the forecaster at time t, we just have access to one realization of y_t, and we can compute $CRPS(F_t, y_t)$ for each t. Under the assumption of the existence of G, we obtain that the sample $(y_1, \ldots, y_t)$ is a mixture on the information sets, and so is a sample of Y. The underlying distributions of Y_t from where the y_t are drawn are unavailable in practice.

For any forecast F_t described in tables 1 and 2, two types of samples of size T can be defined and then simulated :

$$\mathcal{S}_1 = \{CRPS(F_t, y_t)\}_{t=1,\ldots,T} \quad \text{and} \quad \mathcal{S}_2 = \{CRPS(F_t, y_{\pi(t)})\}_{t=1,\ldots,T}\ , \quad (10)$$

where the later is obtained by shuffling the observational vector $\{y_t\}_t$ into $\{y_{\pi(t)}\}_t$ with the random shuffling operator $\pi(t)$. This shuffling breaks the conditional dependence between y_t and F_t produced by the hidden variable Δ_t. Let then consider two random variables, respectively denoted by $CRPS^{\nabla}(F, Y)$ and $CRPS^{\otimes}(F, Y)$, with respective cdf equal to the empirical cdf associated to $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively.

By sorting separately the values of each sample $\mathcal{S}_1$ and $\mathcal{S}_2$, a qq-plot type can be obtained to compare the empirical distributions of these two samples. The left panels of Figure 3 (top panel for the model GE and low panel for the model NN) indicate that the climatological forecast (blue color) provides the same distributional features for the two samples $\mathcal{S}_1$ and $\mathcal{S}_2$. The same type of plots, see right panels, can be made at the uniform scale, see the pp-plots. The important message from Figure 3 is that the ideal, unfocused and extremes forecasts (black, red and green, respectively) move away from the diagonal. This behavior is linked to the notion of auto-calibration[5].

[5]The auto-calibration condition ensures a good interpretation of the discrepancy between distributions as a mean to evaluate the skill of the forecast. This is in accordance with the recommendations on the extremal dependence indices (EDI) of Ferro and Stephenson (2011), quoting that the forecasts should be calibrated in order to get meaningful EDIs. In the same vein, we corroborate the Gneiting et al. (2007) paradigm of *maximizing the sharpness subject to calibration* by *maximizing the information subject to auto-calibration.*

Besides the climatological forecaster, the samples $\mathcal{S}_1$ and $\mathcal{S}_2$ capture relevant information about the discrepancy generated by the hidden conditioning, i.e. the unobservable random variable Δ_t.

Table 4: Availability status of the quantities of interest.

Object	Definition	Availability in practice
F_t	Distribution of the forecast for time t	yes
y_t	Oberved realisation at time t	yes
Δ_t	Conditioning random variable	no
Y_t	Conditional random variable generating y_t	no
Y	Unconditional random variable of the observations	yes
$CRPS(F_t, y_t)$	CRPS of the couple for time t	yes
$CRPS(F_t, Y_t)$	Random variable associated to $CRPS(F_t, y_t)$	no
$CRPS^{\nabla}(F, Y)$	Random variable generated by the $(CRPS(F_t, y_t))_t$	yes
$CRPS^{\otimes}(F, Y)$	Random variable generated by the $(CRPS(F_t, y_{\pi(t)}))_t$	yes

One way to understand the difference in information between $CRPS^{\nabla}(F, Y)$ and $CRPS^{\nabla}(G, Y)$ is to study the integrated difference between the associated cdfs. Indeed, assuming auto-calibration, evaluating the general amount of information brought by a forecast boils down to measure its expected sharpness. This is also equivalent to evaluate its expected score, see Tsyplakov (2013, Section 2.3), where the "expected score" has to be understood with our notation as $\mathbb{E}\{CRPS^{\nabla}(F, Y)\}$.

The lemma presented below shows that the amount of information brought by an auto-calibrated forecast F is also related to the discrepancy $d(F)$, defined as the L^1-distance between the cdfs of $CRPS^{\nabla}(F, Y)$ and $CRPS^{\nabla}(G, Y)$:

$$d(F) := \int_0^{\infty} \left\{ F_{CRPS^{\nabla}(F,Y)}(t) - F_{CRPS^{\nabla}(G,Y)}(t) \right\} \mathrm{d}t \ .$$

Note that this quantity is non negative. This follows from the fact that the climatological forecast has the lowest sharpness, and $d(G) = 0$.

Lemma 3. *Assume that G, F_1, F_2 are auto-calibrated forecasts, and that the observation Y has cdf G; then the following statements hold:*

(i) $CRPS(G,Y) \overset{d}{=} CRPS^{\otimes}(G,Y) \overset{d}{=} CRPS^{\nabla}(G,Y)$ *in distribution;*

(ii) $\mathbb{E}\{CRPS^{\nabla}(F_1,Y)\} \leq \mathbb{E}\{CRPS^{\nabla}(F_2,Y)\} \Leftrightarrow d(F_1) \geq d(F_2)$.

The proof of this lemma is relegated to Appendix 6.5.

To summarize, see also Table 4, the available distributional objects are: the empirical distribution function associated to the unconditional distribution of the observations; the forecasts' distributions; and the empirical distribution made by the $(CRPS(F_t,y_t))_t$. This leaves the practitioners with the three possible empirical distributions, one obtained from $CRPS^{\otimes}(F,Y)$, the one from $CRPS^{\nabla}(F,Y)$, and $CRPS^{\nabla}(G,Y)$. The remaining question is to determine if the empirical distributions associated with these samples can bring relevant information concerning the extremal behavior.

Before addressing this question, we close this section by Figure 4 that summarizes a calibration-information diagram about the trade-off between calibration and information. Subject to auto-calibration, sharpness and information represent the same attributes of a forecast. This diagram can be seen as a natural extension of the idea of Bentzien and Friederichs (2014).

4.2. CRPS behavior for extreme events

4.2.1. Pareto approximation of the CRPS

So far, our examples were based on very specific parametric forms, see tables 1 and 2. In this section, we will see how the study of the upper tail behavior of the CRPS can be moved from these specific examples towards less stringent conditions based on EVT (see Appendix 6.6 for the proofs).

Let X_t and Y_t be two random variables with absolutely continuous cdfs F_t and G_t, respectively. If these two cdfs have identical upper bounds, $x_{F_t} = x_{G_t}$, G_t belongs to $\mathcal{D}(H_{\gamma_t})$ and $c_{F_t} = 2\mathbb{E}_{F_t}(XF_t(X))$ is finite, then conditionally to Δ_t, one has for s such that $1+\gamma_t s > 0$, as $u \to x_{G_t}$,

$$\mathbb{P}\left(\frac{\mathrm{CRPS}(F_t,Y_t) + c_{F_t} - u}{b(u)} > s | Y_t > u, \Delta_t\right) \longrightarrow (1+\gamma_t s)^{-1/\gamma_t}\,. \qquad (11)$$

Equation (11) tells us that, given Δ_t and for large observational values, the CRPS upper tail behavior is equivalent to the one provided by the observation. This generalises the ideas seen throughout the specific case of (8). It is also interesting to point out that $\mathrm{CRPS}^{\nabla}(G,Y)$ has a different tail behavior.

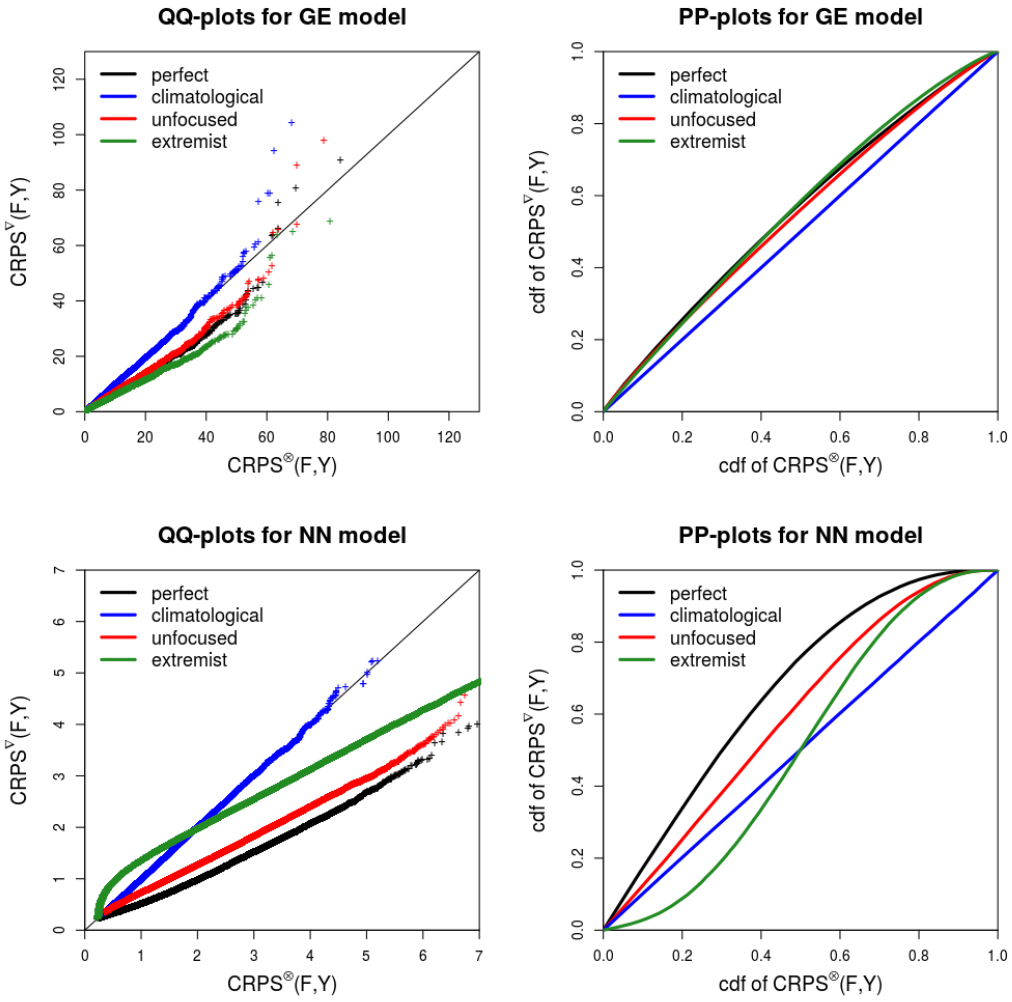

Figure 3: Distributional comparaison via qq-plots and pp-plots between the two samples $\mathcal{S}_1$ and $\mathcal{S}_2$ defined in (10), with respect to the forecasters described in tables 1 and 2.

More precisely, if G belongs to $\mathcal{D}(H_\gamma)$ with $\gamma > 0$, then, for s such that $1 + \gamma s > 0$, as $u \to x_G$:

$$\mathbb{P}\left(\frac{\mathrm{CRPS}^{\nabla}(G,Y) - u}{b(u)} > s | Y > u\right) \longrightarrow (1+\gamma s)^{-1/\gamma} . \tag{12}$$

The constant term in (11) vanishes in (12) due to the linear behavior of the

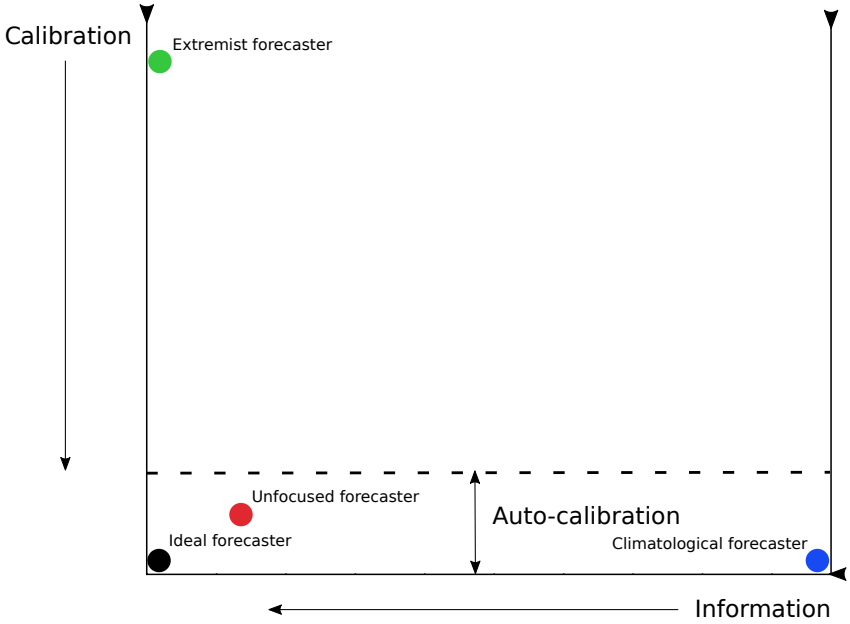

Figure 4: Theoretical calibration-information diagram. The positions of the different types of forecasts are presented. The ideal forecast is calibrated and the most informative.

auxiliary function $b(u)$ for $\gamma > 0$ (see, e.g. Von Mises, 1936; Embrechts et al., 1997). The GE model in Table 2 illustrates this. Working with a dependent couple (F_t, Y_t) or with an independent (G, Y) has an enormous impact on the tail behavior of the respective CRPS. For the GE model, the limit in (11) exhibits an exponential tail behavior, whereas the Pareto limit in (12) shows a heavy-tailed behavior. The difference of tail behaviors could be beneficial for partitioners.

4.2.2. Practical use of the CRPS Pareto tail behavior

To take advantage of the limiting behavior difference between (11) and (12), one needs to have access to samples from these two distributions. At this stage, we can use the convergence provided by (12) to write the following

approximations (as u gets large)

$$CRPS^{\nabla}(G,Y)|Y>u \;\sim\; Y|Y>u \;\sim\; H_{\gamma,\sigma_u}, \tag{13}$$

where H_{γ,σ_u} represents the Generalized Pareto distribution with scale parameter σ_u and shape parameter γ. The two approximations in (13) lead us to propose comparing, for extreme observations above u with u large, the empirical cdf generated by the $(\text{CRPS}(F_t,y_t))_t$ with the theoretical Pareto associated to $CRPS^{\nabla}(G,Y)$. To make this comparison, we rely on the Cramér-von Mises criterion (Cramér, 1928; Von Mises, 1928) :

$$\omega_u^2 = \int_{-\infty}^{+\infty} [K_{u,T}(v) - H_{\gamma,\sigma_u}(v)]^2 dH_{\gamma,\sigma_u}(v).$$

where $K_{u,T}$ corresponds to the empirical distribution based on the sample $CRPS^{\nabla}(F,Y)$ for Y_t above u. To compute this criterion, we order the $\text{CRPS}(F_t,y_t)$ values (in increasing order) and called them $v_1,\cdots,v_m$. The integer m represents the number of observations above u. Then, the Cramér-von Mises statistic[6] becomes

$$T_u = m \times \hat{\omega}_u^2 = \frac{1}{12m} + \sum_{i=1}^{m}\left[\frac{2i-1}{2m} - H_{\gamma,\sigma_u}(v_i)\right]^2.$$

Given an auto-calibrated forecast, a large value of T_u indicates an added value of this forecast for extremes.

4.2.3. An index for extremes skill subject to calibration

For large m and large u, under the null hypothesis the statistic T_u should approximatively follow a Cramér-von Mises distribution. Hence, under this distributional assumption, it is possible to compute the quantile (p-value) associated with T_u. This number in $[0,1]$ is denoted p_u^F here. Note that the auto-calibration condition is fundamental to avoid large type II error, as Ferro and Stephenson (2011) already recommended for EDIs. In order to get an asymmetric index, we propose to compute the following ratio

$$1 - \frac{p_u^F}{p_u^{clim}}. \tag{14}$$

[6] A description of the algorithm calculating T_u is provided in Table 5

This index ranges from zero to one[7], and equals to zero for the climatological forecast. The higher the better.

To see how this index behaves in practice, we revisit the experimental design described in Table 2. For the GE model, 10^6 CRPS values were computed for each type of forecast. In Figure 5, the index defined by (14) is plotted against increasing quantiles, ranging from .75 to 1. The extremist forecast (green curve) appears superior to the others, but being no calibrated, it has to be discarded. Subject to auto-calibration, the perfect forecast (black curve) with an exponential distribution, conditionally on the unobserved Δ_t, is the most rewarded. The climatological forecast, although heavy tailed, has the lowest index.

Table 5: Computation of Cramér-von Mises' statistic and p-value for a forecast F from T couples forecast/observation.

0. CRPS estimates:	- For the T couples forecast/observation, compute their corresponding instantaneous CRPS.
1. Estimation of γ:	- 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 T_u	- Order the m CRPS values in increasing order $v_1, \ldots, v_m$.
For $i \in [1, m]$	-Compute for each CRPS value v_i, $H_{\gamma,\sigma_w}(v_i)$.
	-Compute $\left[\frac{2i-1}{2m} - H_{\gamma,\sigma_w}(v_i)\right]^2$.
End 3.	
End 2.	

As expected, we can see for the GE model in Figure 5 that excluding the extremist forecaster which is not auto-calibrated, the index rewards the perfect forecast.

[7]It can happen that the computation makes this quantity non-positive (ie. $p_u^F > p_u^{clim}$), this pathological outcome is the result of an uncalibrated forecast F and therefore a meaningless quantity.

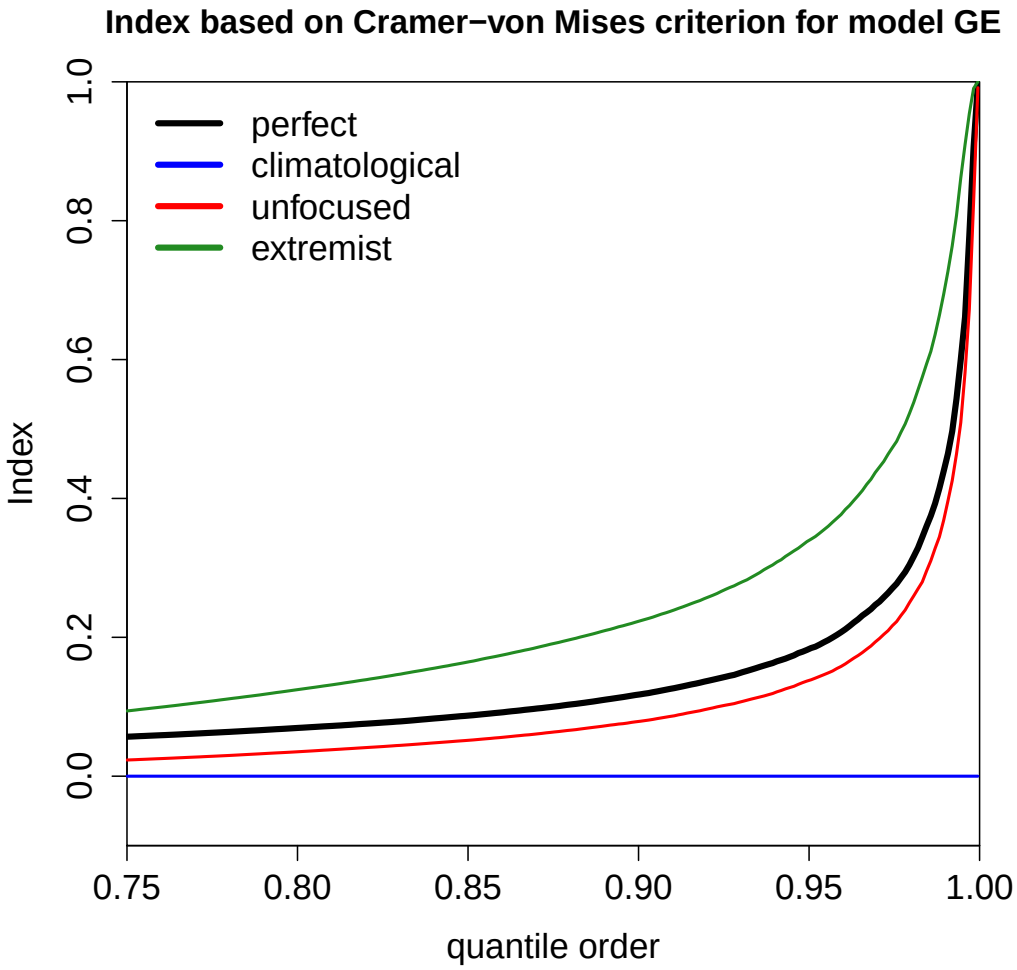

Figure 5: Cramér-von Mises' criterion-based index as a function of the order of the quantile kept as threshold. Subject to auto-calibration, the index rewards the perfect forecast and is information-sensitive.

Concerning the threshold choice, this is not an issue for the GE example because both the exponential and Pareto distributions are threshold invariant. This setup, by removing the threshold choice issue, is ideal to understand our approach. But, it is not realistic in practice. A threshold choice has to made and, many approaches could be used to handle this non-trivial

decision (see, e.g., Naveau et al., 2016; Beirlant et al., 2004). Such decision clearly depends on the application at hand. For example, the recent study by Taillardat et al. (2019) proposed and compared different post processing approaches to improve precipitation forecasts in France. As extreme rainfall can be heavy tailed distributed, we expect that the forecasting approaches coupled with EVT, referred as QRF EGP tail, EMOS GEV and EMOS EGP in Figure 6, should perform better than other non-EVT based approaches like Anologs, QF and QRF, see Taillardat et al. (2019) for a detailed description of each post-processing techniques. The behavior of our index (y-axis in Figure 6) defined by (14), independently of the threshold choice, clearly indicates the superiority of representing extremes for the forecasting methods that included a specific EVT treatment of heavy rainfall. This confirms the conclusions of Taillardat et al. (2019), especially in regards to their Figure 3.

5. Discussion

According to our analysis of the CRPS, the mean of the CRPS and its weighted derivations seem to be unable to discriminate forecasts with different tails behaviors, whatever the weighting scheme.

Coming back to the three types of "goodness" introduced by Murphy (1993), the forecast value seems to be the most important for extreme events. For example, severe weather warnings are still made by forecasters, and despite of a possible inaccurate prediction quantitatively speaking, the forecaster has to take the decision according to a threshold of interest. This approach is completely linked with the economic value of the forecast, or equivalently with the ability of standing out from the climatology. For deterministic 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). Recently, Ehm et al. (2016) have introduced the so-called "Murphy diagrams" for deterministic forecasts. This original approach allows to appreciate dominance among different forecasts and anticipate their skill area. In the same vein, we show that, subject to auto-calibration, relevant information about extremes can be represented by the discrepancy between unconditional and conditional score's distribution. An open question remains: subject to auto-calibration, do the score's distribution can locally beat the perfect forecast or be beaten by the climatological forecast ? We illustrate this conjecture in Figure 7.

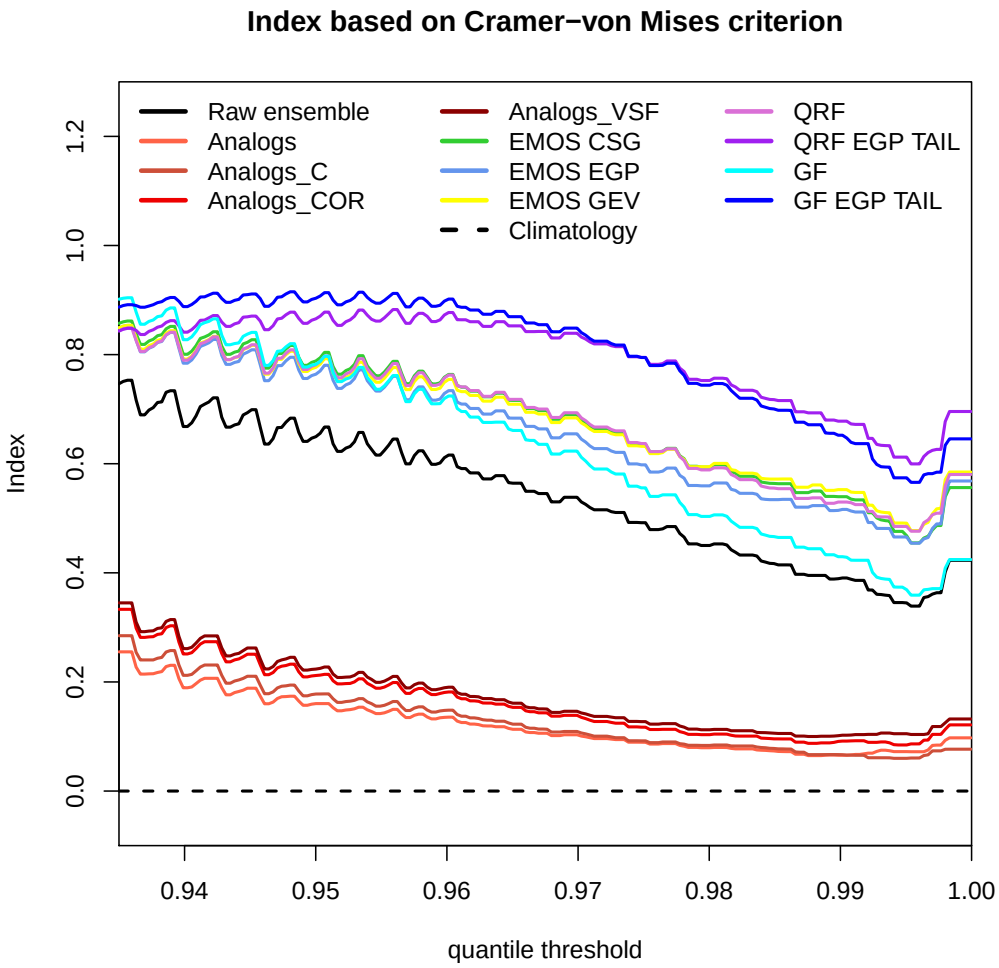

Figure 6: Cramér-von Mises' based indexes as a function of the threshold for the 6-h rainfall forecast.

Inspired by Friederichs (2010), our work consists of applying extreme value theory on common verification measures itselves. We therefore consider the score as a random variable. Relying on some properties of the CRPS for large observed events we put a theoretical framework concerning the score's behavior for extremes. As a result, we obtain a bounded index in $[0, 1]$ to assess the nexus between forecasts and observations. One can view

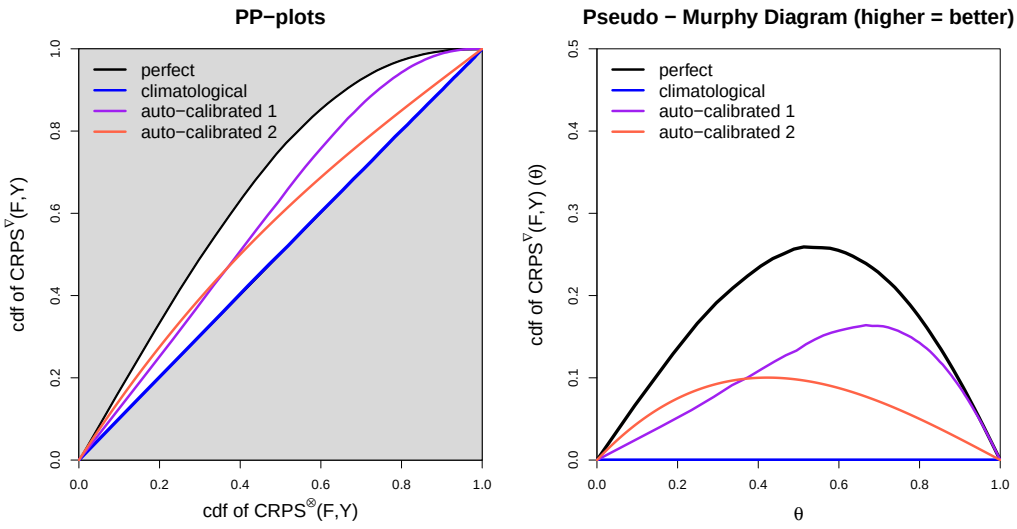

Figure 7: Illustration of the conjecture of the Discussion Section. On the left, we think that subject to auto-calibration the distribution score's are bounded by respectively the best and the poorest auto-calibrated forecasts (grey areas can never be visited). In the right, a pseudo-Murphy diagram (Ehm et al., 2016) is presented. In our conjcture, for a given value of θ (corresponding to a quantile order of the observations) a ranking can be made among forecasters.

this contribution as an additional step in bridging the gap in the field of ensemble verification and extreme events, (see, e.g. Ferro, 2007; Friederichs and Thorarinsdottir, 2012; Ferro and Stephenson, 2011). The ensemble forecast information is kept by the use of the CRPS. The index introduced in Section 4.2.3 can be considered as the probabilistic alternative to the deterministic scores introduced by Ferro (2007) and Ferro and Stephenson (2011). We would say that the paradigm of *maximizing the sharpness subject to calibration* can be associated with the paradigm of *maximizing the information for extreme events subject to auto-calibration*. It would be convenient to study the specific properties of this CRPS-based tool and its potential paths and pitfalls. Another potentially interesting investigation could be to extend this procedure to other scores like the mean absolute difference, the igno-

rance[8] score (Diks et al., 2011) or the Dawid-Sebastiani score (Dawid and Sebastiani, 1999).

6. Appendix

6.1. Proof of the inequality (3)

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

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

as announced in (3).

[8] Indeed, the Neyman-Pearson lemma described in Lerch et al. (2017) let us think that this score could be a natural candidate.

6.2. Proof of the inequality (5)

Let u be a positive real. Denote Z a non-negative random variable with finite mean and cdf H. We introduce the new random variable

$$Y = X\mathbf{1}\{u \geq X\} + (Z+u)\mathbf{1}\{X > u\} \tag{15}$$

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

$$\overline{G}(x) \;=\; \begin{cases} \overline{F}(x), & \text{if } x \leq u \\ \overline{H}(x-u)\overline{F}(u), & \text{otherwise ,} \end{cases} \tag{16}$$

where Z has the same end point than X, $\overline{H}(0) = 1$ and

$$\overline{H}(x-u) \leq \overline{F}(x)/\overline{F}(u), \text{ for any } x \geq u. \tag{17}$$

This latter condition implies that

$$\overline{G}(x) \leq \overline{F}(x), \text{ for all } x. \tag{18}$$

Because of $W(x)$ is strictly increasing, the equation (16) allows to write almost surely that

$$\mathbb{E}(W(X)\mathbf{1}\{X < x\}) = \mathbb{E}(W(Y)\mathbf{1}\{Y < x\}) \text{ , for any } x \leq u. \tag{19}$$

Equality (19) combined with the expression of the CRPS implies that

$$\begin{aligned} & \frac{1}{2}[\text{wCRPS}(G,x) - \text{wCRPS}(F,x)] \\ = \;& \mathbb{E}_Y[(W(Y) - W(x))\mathbf{1}\{Y > x\}] - \mathbb{E}_X[(W(X) - W(x))\mathbf{1}\{X > x\}] \\ & + \mathbb{E}_X(W(X)F(X)) - \mathbb{E}_Y(W(Y)G(Y)), \\ = \;& \mathbb{E}_Y(W(Y)\overline{G}(Y)) - \mathbb{E}_X(W(X)\overline{F}(X)) \\ & - \mathbb{E}_Y[(W(Y) - W(x))\mathbf{1}\{Y \leq x\}] + \mathbb{E}_X[(W(X) - W(x))\mathbf{1}\{X \leq x\}] \\ = \;& \mathbb{E}_Y(W(Y)\overline{G}(Y)) - \mathbb{E}_X(W(X)\overline{F}(X)) + \int_u^{x_F} \Delta(x)dF(x), \end{aligned}$$

where

$$\Delta(x) = \mathbb{E}_X[(W(X) - W(x))\mathbf{1}\{X \leq x\}] - \mathbb{E}_Y[(W(Y) - W(x))\mathbf{1}\{Y \leq x\}].$$

The stochastic ordering between Y and X implies that $\mathbb{E}_Y(W(Y)\overline{G}(Y)) - \mathbb{E}_X(W(X)\overline{F}(X)) \leq 0$. This leads to

$$\frac{1}{2}\left|\mathbb{E}_X[\text{wCRPS}(G,X)] - \mathbb{E}_X[\text{wCRPS}(F,X)]\right| \leq \int_u^{x_F} \Delta(x) dF(x).$$

For $x > u$ we can write that

$$\begin{aligned}
&\Delta(x) \\
&= \mathbb{E}_X[(W(X) - W(x))\mathbf{1}\{u < X \leq x\}] - \mathbb{E}_Y[(W(Y) - W(x))\mathbf{1}\{u < Y \leq x\}], \\
&\leq \mathbb{E}_Y[(W(x) - W(u))\mathbf{1}\{u < Y \leq x\}], \\
&\qquad \text{since } W(X) - W(x) \leq 0 \text{ and } 0 \leq W(x) - W(Y) \leq W(x) - W(u), \\
&= (W(x) - W(u))[G(x) - G(u)], \\
&\leq (W(x) - W(u))\overline{G}(u), \\
&= (W(x) - W(u))\overline{F}(u).
\end{aligned}$$

Hence, we can write that

$$\begin{aligned}
\left|\mathbb{E}_X[\text{wCRPS}(G,X)] - \mathbb{E}_X[\text{wCRPS}(F,X)]\right| &\leq 2\overline{F}(u)\int_u^{x_F} (W(x) - W(u)) dF(x), \\
&\leq 2\overline{F}^2(u)\mathbb{E}_X[W(X) - W(u)|X > u].
\end{aligned}$$

This inequality is true for any u and H. The right hand side of the last inequality does not depend on $\overline{H}(x)$. Thus, the tail behavior of the random variables X and Z can be completely different, although the CRPS of F and G can be as closed as one wishes. The right hand side goes to 0 due to the finite mean of $W(X)$.

6.3. Proof of Lemma 1

As the weight function is discontinuous at $x = z$, (3) cannot be applied. Instead, if $z > y$, the condition $\{x \geq z\}$ implies $\{x \geq y\}$, and consequently

$$wCRPS(F, y; z) = \int_z^{\infty} \overline{F}^2(x)\, dx, \text{if } z > y.$$

If $y \geq z$, then

$$wCRPS(F, y; z) = \int_z^{y} F^2(x) dx + \int_y^{\infty} \overline{F}^2(x) dx = CRPS(F, y) - c_z$$

where $c_z = \int_{-\infty}^{z} F^2(x)dx$. In summary, we write with the notation $c_z^* = \int_z^{\infty} \overline{F}^2(x)\,dx$,

$$wCRPS(F,y;z) \;=\; \begin{cases} c_z^*, & \text{if } z > y, \\ CRPS(F,y) - c_z, & \text{if } y \geq z. \end{cases}$$

As $CRPS(F,z) = \int_{-\infty}^{z} F^2(x)dx + \int_z^{\infty} \overline{F}^2(x)dx = c_z + c_z^*$, In addition, it follows that

$$\begin{aligned} &\mathbb{E}\left[wCRPS(F,Y;z)\right] = c_z^* G(z) + \mathbb{E}\left[CRPS(F,Y)\mathbf{1}\{Y \geq z\}\right] - c_z\overline{G}(z), \\ = \;& \int_z^{\infty} \overline{F}^2(x)\,dx + \mathbb{E}\left[CRPS(F,Y)|Y \geq z\right]\overline{G}(z) - CRPS(F,z)\overline{G}(z). \end{aligned}$$

6.4. Proof of Lemma 2

In order to simplify the notation, and since $t > 0$ is fixed, the dependence in time t is omitted along the lines of this proof. Assume that $X \overset{d}{=}$ Pareto(β,ξ) and $Y \overset{d}{=}$ Pareto(σ,γ) with $0 \leq \xi < 1$ and $0 \leq \gamma < 1$, i.e. with respective survival functions $\overline{F}(x) = (1+\xi x/\beta)^{-1/\xi}$ and $\overline{G}(x) = (1+\gamma x/\sigma)^{-1/\gamma}$ for any $x > 0$. Applying (3) 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

$$CRPS(F,y) = y + 2(1+\xi y/\beta)^{-1/\xi}\frac{\beta+\xi y}{1-\xi} - 2\beta\left(\frac{1}{1-\xi} - \frac{1}{2(2-\xi)}\right). \quad (20)$$

It follows that

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

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 $(0,1)$,

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

If $c = \frac{\xi\sigma}{\beta\gamma} = 1$, 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

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

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

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

6.5. Proof of Lemma 3

(i) The first equality on law follows from the definition of auto-calibration.The second equality in law comes from the fact that the data are invariant by shuffling when the forecast is the climatology. This has been in particular illustrated on Figure 3, see comments following (10).

(ii) Consider two forecasts F_1 and F_2, and assume that F_1 contains more information than F_2, that is to say

$$\mathbb{E}\{CRPS^{\nabla}(F_1, Y)\} \leq \mathbb{E}\{CRPS^{\nabla}(F_2, Y)\} \ .$$

Since the latter random variables are non negative, this is equivalent to

$$\int_0^{\infty} \{1 - F_{CRPS^{\nabla}(F_1,Y)}(t)\}\mathrm{d}t \leq \int_0^{\infty} \{1 - F_{CRPS^{\nabla}(F_2,Y)}(t)\}\mathrm{d}t \ . \tag{21}$$

Noticing that

$$\begin{aligned}\int_0^{\infty} \{1 - F_{CRPS^{\nabla}(G,Y)}(t)\}\mathrm{d}t &= \int_0^{\infty} \{1 - F_{CRPS(G,Y)}(t)\}\mathrm{d}t \\ &= \mathbb{E}\{CRPS^{\nabla}(G, Y)\} \ ,\end{aligned}$$

one gets that (21) is equivalent to

$$\int_0^{\infty} \{F_{CRPS^{\nabla}(F_1,Y)} - F_{CRPS^{\nabla}(G,Y)}\}(t)\mathrm{d}t \geq \int_0^{\infty} \{F_{CRPS^{\nabla}(F_2,Y)} - F_{CRPS^{\nabla}(G,Y)}\}(t)\mathrm{d}t \ ,$$

which gives the expected result.

6.6. Proof of the convergences (11) and (12)

The proof of (11) and (12) is similar, so that we will focus on proving (12). 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 X with finite mean and with cdf F will be denoted by $M(F, u)$, so that $M(F, u) = \mathbb{E}_F(X - u|X > u)$.

Lemma : Consider a random variable X with finite mean that belongs to domain of attraction $\mathcal{D}(H_\gamma)$ for $\gamma < 1$. There exist real positive numbers α and β such that for large u tending to $\to x_F$,

$$0 \leq 2\mathbb{E}_F(X - u|X > u) \leq \alpha u + \beta \ . \tag{22}$$

Proof of the lemma: Let decompose the proof depending on the sign of γ :

1. First case : F belongs to $\mathcal{D}(H_\gamma)$ with $0 < \gamma < 1$:
   In this case, Embrechts et al. (1997) (Section 3.4) show that $M(F,u) \sim \gamma u/(1-\gamma)$ as u tends to x_F, and we can conclude directly.
2. Second case : 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 u tends to x_F, $M(F,u) \sim \gamma(x_F - u)/(\gamma - 1)$.
3. Third case : F belongs to $\mathcal{D}(H_0)$:
   When F is in the Gumbel domain of attraction, Theorem 3.9 in Ghosh and Resnick (2010) ensures that $M(F,u)/u \to 0$ as u tends to x_F.

Proof of (12):
According to the formula (1) of the CRPS, we can write that

$$\mathrm{CRPS}(F,Y) \stackrel{a.s.}{=} Y - c_F + 2\overline{F}(Y)\mathbb{E}_F(X - Y | X > Y),$$

in terms of $c_F = 2\mathbb{E}_F(XF(X))$. Fix a large u conditionally to Y, one gets thanks to the previous lemma:

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

So that

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

We recognize the probability for Y to be in an interval denoted by $[\delta_u, \Delta_u]$:

$$\frac{1}{\overline{F}(u)}\mathbb{P}\left(Y \in \left[\frac{tb(u) + u - \beta\overline{F}(u)}{1 + \alpha\overline{F}(u)}, tb(u) + u\right]\right) = \frac{\mathbb{P}(Y \in [\delta_u, \Delta_u])}{\overline{F}(u)}.$$

Note then that

$$\begin{aligned}
\frac{\mathbb{P}(Y \in [\delta_u, \Delta_u])}{\overline{F}(u)} &\leq \sup_{v \in [\delta_u, \Delta_u]} g(v)\, \frac{\Delta_u - \delta_u}{\overline{F}(u)} \\
&= \sup_{v \in [\delta_u, \Delta_u]} g(v)\, \frac{\alpha(tb(u) + u) + \beta}{1 + \alpha\overline{F}(u)} \\
&= O(ug(u)) \longrightarrow 0 \quad \text{as} \quad u \to x_F\ .
\end{aligned}$$

The dominance in $ug(u)$ is provided by the sublinear/linear behavior of b (Von Mises condition in Von Mises (1936)). Indeed, Von Mises (1936) noticed that a possible choice for $b(u)$ can be the mean excess function of Y which is (sub)linear. The limit to 0 is due to the finite first moment of the random variable Y, because in this case $ug(u) \sim 1 - G(u) \to 0$ as $u \to x_F$.

Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., Waal, D., Ferro, C., 2004. Statistics of extremes: Theory and applications.

Bentzien, S., Friederichs, P., 2014. Decomposition and graphical portrayal of the quantile score. Quarterly Journal of the Royal Meteorological Society 140 (683), 1924–1934.

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

Bröcker, J., 2015. Resolution and discrimination–two sides of the same coin. Quarterly Journal of the Royal Meteorological Society 141 (689), 1277–1282.

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography 131 (609), 2131–2150.

Cramér, H., 1928. On the composition of elementary errors: First paper: Mathematical deductions. Scandinavian Actuarial Journal 1928 (1), 13–74.

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.

Friederichs, P., 2010. Statistical downscaling of extreme precipitation events using extreme value theory. Extremes 13 (2), 109–132.

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.

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., et al., 2013. Combining predictive distributions. Electronic Journal of Statistics 7, 1747–1782.

Hagedorn, R., 2017. Slowly but surely: Observing and supporting the growing use of ensemble forecasts. In: Annual Seminar, ECMWF 2017, Reading, United Kingdom. pp. http://www.ecmwf.int/sites/default/files/elibrary/2017/17625–slowly–surely–observing–and–supporting–growing–use–ensemble–products.pdf.

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

Kahneman, D., Tversky, A., 1979. Prospect theory: An analysis of decision under risk. Econometrica: Journal of the econometric society, 263–291.

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.

Matheson, J. E., Winkler, R. L., 1976. Scoring rules for continuous probability distributions. Management science 22 (10), 1087–1096.

Morel, C., 2014. Les décisions absurdes. Vol. 1. Editions Gallimard.

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

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

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.

Schervish, M. J., Seidenfeld, T., Kadane, J. B., 2009. Proper scoring rules, dominated forecasts, and coherence. Decision Analysis 6 (4), 202–221.

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.

Straehl, C., Ziegel, J., 2017. Cross-calibration of probabilistic forecasts. Electronic Journal of Statistics 11, 608–639.

Taillardat, M., Fougères, A.-L., Naveau, P., Mestre, O., 2019. Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting. Weather and Forecasting in press.

Tsyplakov, A., 2013. Evaluation of probabilistic forecasts: proper scoring rules and moments. Tech. rep., http://dx.doi.org/10.2139/ssrn.2236605.

Von Mises, R., 1928. Statistik und wahrheit. Julius Springer.

Von Mises, R., 1936. La distribution de la plus grande de n valeurs. Rev. math. Union interbalcanique 1 (1).

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.