TAIL CALIBRATION OF PROBABILISTIC FORECASTS

Sam Allen, Jonathan Koh, Johan Segers, Johanna Ziegel

LIDAM Discussion Paper ISBA
2024 / 18

UCLouvain

LIDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA
Voie du Roman Pays 20 - L1.04.01
B-1348 Louvain-la-Neuve
Email : lidam-library@uclouvain.be
https://uclouvain.be/en/research-institutes/lidam/isba/publication.html

Tail calibration of probabilistic forecasts

Sam Allen*
Seminar for Statistics, ETH Zurich
and
Jonathan Koh
IMSV, University of Bern
and
Johan Segers
LIDAM/ISBA, UCLouvain
and
Johanna Ziegel
Seminar for Statistics, ETH Zurich

July 4, 2024

Abstract

Probabilistic forecasts comprehensively describe the uncertainty in the unknown future outcome, making them essential for decision making and risk management. While several methods have been introduced to evaluate probabilistic forecasts, existing evaluation techniques are ill-suited to the evaluation of tail properties of such forecasts. However, these tail properties are often of particular interest to forecast users due to the severe impacts caused by extreme outcomes. In this work, we introduce a general notion of tail calibration for probabilistic forecasts, which allows forecasters to assess the reliability of their predictions for extreme outcomes. We study the relationships between tail calibration and standard notions of forecast calibration, and discuss connections to peaks-over-threshold models in extreme value theory. Diagnostic tools are introduced and applied in a case study on European precipitation forecasts.

Keywords: extreme event, proper scoring rule, forecast evaluation, tail calibration diagnostic plot, precipitation forecast

*Corresponding author: sam.allen@stat.math.ethz.ch. Johan Segers gratefully acknowledges helpful questions and suggestions from participants, in particular Anja Janßen and Clément Dombry, of the EXSTA Kick-Off Workshop at Université Paris Cité, Laboratoire MAP5, June 5–7, 2024.

1 Introduction

Forecasts for future events should be reliable, or calibrated. For example, if an event is predicted to occur with a certain probability, this forecast will only be useful for decision making and risk assessment if the event does indeed transpire with the predicted probability. Since extreme events often have the largest impacts on forecast users, calibrated forecasts for these outcomes are particularly valuable. It is therefore essential that methods are available to assess the performance of probabilistic forecasts for extreme events. However, the evaluation of forecasts with regards to extreme events is difficult and prone to methodological pitfalls (Lerch et al., 2017; Bellier et al., 2017).

Proper scoring rules quantify the accuracy of probabilistic forecasts, thereby allowing forecasters to be ranked and compared objectively (see e.g. Gneiting and Raftery, 2007). However, despite their widespread use, no proper scoring rule can distinguish the true distribution of the observations from a forecast distribution with incorrect tail behavior (Brehmer and Strokorb, 2019). Weighted scoring rules, which are commonly employed to emphasize particular outcomes during forecast evaluation (Diks et al., 2011; Gneiting and Ranjan, 2011; Holzmann and Klar, 2017; Lerch et al., 2017; Allen et al., 2023; Olafsdottir et al., 2024), also suffer from these limitations. We extend the negative result of Brehmer and Strokorb (2019) by demonstrating that it cannot be circumvented by employing a class of scoring rules to assess forecast performance; see Appendix B. Alternative tools are therefore required to evaluate forecasts for extreme events.

In this paper, we focus on probabilistic forecasts for real-valued outcomes. A probabilistic forecast takes the form of a probability distribution F over all possible values of the outcome $Y \in \mathbb{R}$, and is typically stated as a cumulative distribution function (cdf). While scoring rules assess relative forecast performance, it is also important to determine whether probabilistic forecasts are calibrated. Forecast calibration is often assessed in practice using histograms or *pp*-plots of probability integral transform (PIT) values (Diebold et al., 1998; Gneiting et al., 2007). Such diagnostic tools additionally allow practitioners to ascertain what systematic biases manifest in the forecasts, thereby helping to improve future predictions.

While several definitions of calibration have been proposed (see Gneiting and Resin, 2023, for a recent review), all focus on overall forecast performance, and do not allow practitioners to evaluate forecasts made for extreme outcomes. In this paper, we introduce a general notion of forecast *tail calibration* that quantifies the reliability of predictions when forecasting extreme events. We derive theoretical implications of tail calibration, study the connection with peaks-over-threshold models in extreme value theory, and provide examples where standard calibration does and does not imply tail calibration. In general, tail calibration is not implied by classical definitions of calibration, highlighting that it provides additional information that is not available from existing forecast evaluation methods.

In Section 2, we revisit notions of calibration for probabilistic forecasts, and introduce our concepts that are suitable when evaluating forecasts for extreme events. Connections between tail calibration and well-known results in extreme value theory are explored in Section 3. Section 4 presents several examples and introduces diagnostic tools to assess tail calibration in practice, and these diagnostic tools are then applied to a canonical weather forecasting data set in Section 5. Section 6 summarizes our findings. The supplement contains sections about the following topics: additional proofs and calculations; a result about multiple scoring rules and tail behavior; a definition of and results on marginal tail calibration; and additional simulation results evaluating well-known forecasts in the literature with respect to their probabilistic tail calibration. R code to reproduce the results herein is available at `https://github.com/sallen12/TailCalibration`.

2 Tail calibration

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space on which a random pair (F, Y) is defined, where F is a random probabilistic forecast for the outcome variable Y. We restrict attention to real-valued outcomes and treat F as a random cumulative distribution function (cdf). We assume that F is almost surely continuous, as this case is most relevant with regards to extremes and leads to a simpler presentation. However, generalizations are possible, see Remark 1.

2.1 Classical notions of calibration

Tsyplakov (2011) refers to a probabilistic forecast F as *auto-calibrated* if $F(y) = \mathsf{P}(Y \leq y \mid F)$, $y \in \mathbb{R}$, almost surely. That is, given that the distribution F has been issued as the forecast, the observation will arise according to this distribution. This provides an intuitive requirement that a forecast must satisfy in order to be considered reliable. More generally, for some σ-algebra $\mathcal{B} \subseteq \mathcal{A}$, a forecast is called *(probabilistically) $\mathcal{B}$-calibrated* if

$$\mathsf{P}(Z_F \leq u \mid \mathcal{B}) = u \quad \text{almost surely, for all } u \in [0,1], \tag{1}$$

where $Z_F = F(Y)$ is the probability integral transform (PIT) of Y with respect to F. In other words, Z_F is required to have a standard uniform distribution, $Z_F \sim \mathrm{Unif}(0,1)$, and to be independent of $\mathcal{B}$. Modeste (2023, Proposition 3.4) shows that if $\sigma(F) \subseteq \mathcal{B}$, then $\mathcal{B}$-calibration of F is equivalent to $F(y) = \mathsf{P}(Y \leq y \mid \mathcal{B})$ almost surely, for all $y \in \mathbb{R}$. The case $\mathcal{B} = \sigma(F)$ yields auto-calibration. A closely related notion to $\mathcal{B}$-calibration of a forecast F has also been referred to as the forecast being ideal with respect to $\mathcal{B}$ by Gneiting and Ranjan (2013), while Tsyplakov (2014, Definition 3) refers to (1) as conditional probabilistic calibration.

Remark 1 (Discontinuous forecasts)**.** We assume that the forecast cdf F is continuous almost surely, but the notions of calibration that rely on Z_F can also be defined for possibly discontinuous F. To this end, let V be a standard uniform random variable that is independent of (F, Y) and define $Z_F = F(Y-) + V[F(Y) - F(Y-)]$, where $F(y-)$ is the left-hand limit of F at y. The result of Modeste (2023, Proposition 3.4) continues to hold with this definition of Z_F.

Auto-calibration is a strong and intuitive notion of calibration, and several statistical tests for auto-calibration have been proposed, see for example Strähl and Ziegel (2017), Modeste (2023), and related literature on testing forecast encompassing Harvey et al. (1998). However, it is typically difficult to assess auto-calibration in practice using diagnostic tools. Instead, it is common to identify relevant deficiencies of forecasts by employing powerful diagnostics that assess necessary criteria for auto-calibration.

Following Gneiting et al. (2007), the forecast distribution F is said to be *marginally calibrated* if $\mathsf{E}[F(y)] = \mathsf{P}(Y \leq y)$ for all $y \in \mathbb{R}$, and *probabilistically calibrated* if $\mathsf{P}(Z_F \leq u) = u$ for all $u \in [0,1]$, that is, if $Z_F \sim \text{Unif}(0,1)$. Marginal calibration implies that the average probability assigned to Y exceeding a threshold is equal to the true unconditional probability, for any threshold. Probabilistic calibration implies that central prediction intervals at all levels have the correct coverage. If a forecast is auto-calibrated, then it is both marginally and probabilistically calibrated, but the converse is not true. Furthermore, Gneiting et al. (2007) show that marginal and probabilistic calibration are neither necessary nor sufficient conditions for each other, in other words, for the trivial σ-algebra, the equivalence of $\mathcal{B}$-calibration and $\mathsf{E}[F(y)] = \mathsf{P}(Y \leq y)$, $y \in \mathbb{R}$, is generally false.

Marginal and probabilistic calibration are unconditional notions of calibration, whereas auto-calibration depends on the forecast distribution (Gneiting and Resin, 2023). Intermediate notions of conditional calibration are quantile and threshold calibration; see Appendix C in the supplement. Further, $\mathcal{B}$-calibration is possibly weaker than auto-calibration if $\mathcal{B} \subseteq \sigma(F)$, and can be stronger otherwise.

Example 2 (Unfocused forecaster)**.** To see that probabilistic calibration is strictly weaker than auto-calibration when F is not deterministic, suppose $Y \sim N(\mu, 1)$ follows a normal distribution with mean μ and variance one conditionally on μ, where $\mu \sim N(0,1)$, and consider the *unfocused* forecaster of Gneiting et al. (2007, Example 3): $F = (1/2)N(\mu, 1) + (1/2)N(\mu + \tau, 1)$ with τ equal to 1 and -1 with probability 0.5, independently of μ. Figure 1 shows an example of this forecast for a fixed μ. This forecaster is either positively or negatively biased, with equal probability. As a result, this forecaster is not auto-calibrated (and also not marginally calibrated). However, perhaps unintuitively, the biases cancel out such that the forecast is probabilistically calibrated.

2.2 Notions of tail calibration

To assess the reliability of forecasts for extreme events, we adopt the standard approach of extreme value theory and decompose $\mathsf{P}(Y > x + t) = \mathsf{P}(Y > t)\,\mathsf{P}(Y > x + t \mid Y > t)$, for $x \geq 0$ and large t, i.e., $\mathsf{P}(Y > t)$ positive but close to zero. Then, we separately consider the calibration of both factors on the right-hand side. Under auto-calibration, we have

$$\mathsf{P}(Y > x + t \mid F) = \mathsf{P}(Y > t \mid F)\,\frac{\mathsf{P}(Y > x + t \mid F)}{\mathsf{P}(Y > t \mid F)} = (1 - F(t))(1 - F_t(x))$$

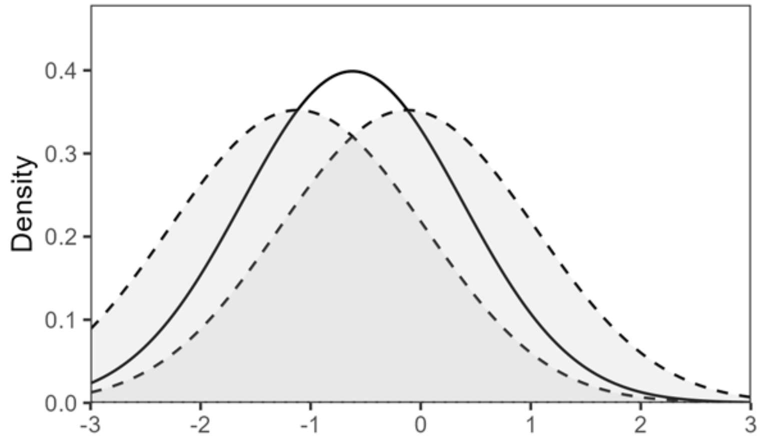

Figure 1: The unfocused forecaster in Example 2 issues each of the dashed densities with probability one half, while the true outcome distribution is the solid density.

with the *conditional forecast excess distribution* F_t defined as

$$F_t(x) = \frac{F(x+t) - F(t)}{1 - F(t)}, \quad x \geq 0, \tag{2}$$

if $F(t) < 1$, and $F_t(x) = 1$, otherwise. For large t, we are interested in assessing whether $\mathsf{P}(Y > t \mid F)$ is close to $1 - F(t)$ and whether $\mathsf{P}(Y > x + t \mid F)/\,\mathsf{P}(Y > t \mid F)$ is close to $1 - F_t(x)$ for all $x \geq 0$. In order to come up with non-trivial notions of calibration for large t, it is important to keep the following remark in mind.

Remark 3 (Scaling issue)**.** A direct comparison of the true and forecast conditional excess distributions $\mathsf{P}(Y - t \leq x \mid Y > t)$ and $F_t(x)$, for $x \geq 0$ and large t, would potentially yield uninformative results because of a scaling issue. To see this, suppose Y and F both have unbounded support. For fixed $x > 0$, if both Y and F are heavy-tailed, then both probabilities tend to 0 as $t \to \infty$, whereas if both have a tail which is lighter than exponential, then both probabilities tend to 1 almost surely. In both cases, the difference between $\mathsf{P}(Y - t \leq x \mid Y > t)$ and $F_t(x)$ tends to zero as $t \to \infty$, even though the tails of Y and F could be very different.

To avoid the scaling issue in Remark 3, we define the *excess probability integral transform (excess PIT)* as

$$Z_F^{(t)} = F_t(Y - t), \tag{3}$$

and we suggest the following definition of tail calibration.

Definition 4. Let $x_Y = \sup\{x \in \mathbb{R} : \mathsf{P}(Y \leq x) < 1\}$ be the upper endpoint of Y and let $\mathcal{B} \subseteq \mathcal{A}$ be a σ-algebra. Assume that the upper endpoint of the conditional distribution of Y given $\mathcal{B}$ is equal to x_Y, i.e., for all $t < x_Y$, we have, almost surely, both $\mathsf{P}(Y > t \mid \mathcal{B}) > 0$ and $\mathsf{P}(Y > x_Y \mid \mathcal{B}) = 0$. The forecast F is *tail $\mathcal{B}$-calibrated* for Y if $\mathsf{E}[1 - F(t) \mid \mathcal{B}] > 0$

almost surely for all $t < x_Y$ and if, for all $u \in [0,1]$, we have

$$\frac{\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} \to u, \qquad \text{almost surely as } t \to x_Y. \tag{4}$$

If $\mathcal{B} = \sigma(F)$, then F is called *tail auto-calibrated* for Y, and if $\mathcal{B} = \{\varnothing, \Omega\}$, then F is called *probabilistically tail calibrated* for Y.

Probabilistic tail calibration, that is, $\mathcal{B} = \{\varnothing, \Omega\}$, is similar in essence to the standard notion of probabilistic calibration, except that it relies on the conditional probability integral transform associated with the conditional forecast excess distribution and outcomes above a large threshold; a similar concept was informally introduced by Allen et al. (2023); see also Mitchell and Weale (2023). The definition of $\mathcal{B}$-calibration at (1) motivates the conditioning on $\mathcal{B}$ in the tail regime when defining tail $\mathcal{B}$-calibration.

The left-hand side of Eq. (4) can be written as a *combined ratio* via

$$\frac{\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} = \frac{\mathsf{P}\big(Y > t \mid \mathcal{B}\big)}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} \cdot \frac{\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)}{\mathsf{P}\big(Y > t \mid \mathcal{B}\big)}.$$

The single condition in Definition 4 is equivalent to two conditions together: For all $u \in [0,1]$,

$$\frac{\mathsf{P}(Y > t \mid \mathcal{B})}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} \to 1, \quad \text{almost surely as } t \to x_Y; \tag{5}$$

$$\frac{\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)}{\mathsf{P}(Y > t \mid \mathcal{B})} \to u, \quad \text{almost surely as } t \to x_Y. \tag{6}$$

Condition (5) regards the *occurrence* of excesses over a high threshold and can be justified as follows. The σ-algebra $\mathcal{B}$ encodes the information that the forecaster has at hand to predict the conditional tail distribution of Y, or with respect to which we would like to assess the forecaster's calibration. Therefore, we would like the exceedance probabilities over t conditionally on $\mathcal{B}$ to be forecast correctly, at least asymptotically. In other words, we would like the tails of Y and F to be equivalent, given $\mathcal{B}$.

Condition (6) concerns the *severity* of excesses over t, once they occur, and arguably needs more justification. The condition (6) is related to the characterization of $\mathcal{B}$-calibration in (1). Asymptotically, under tail $\mathcal{B}$-calibration, the conditional probability $\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)$ factorizes as $\mathsf{P}(Y > t \mid \mathcal{B}) \cdot u$, and hence the excess PIT is (approximately) uniformly distributed and independent of $\mathcal{B}$. The condition in Definition 4 has the following persuasive interpretation: Given the information $\mathcal{B}$, the conditional probability of $\{Y > t\} \cap \{Z_F^{(t)} \le u\}$ factorizes asymptotically as $(1 - F(t)) \cdot u$.

For assessing the quality of forecasts in the tail, it is not sufficient to focus only on condition (6) regarding the severity ratio and forget about condition (5) on the occurrence ratio. Suppose F is a deterministic forecast such that there exists $t_0 < x_Y$ and $c > 0$ such that $1 - F(t) = c\,\mathsf{P}(Y > t)$ for all $t \ge t_0$. Then F and Y have the same conditional excess distributions for thresholds above t_0 and thus F satisfies condition (6), even though the tail probability ratio $\mathsf{P}(Y > t)/(1 - F(t))$ can be arbitrarily small or large.

Conversely, only considering condition (5) on the occurrence ratio without condition (6) on the severity ratio is insufficient, too, as the following example shows.

Example 5 (Misinformed forecaster)**.** For $\gamma > 0$, let Δ_1 and Δ_2 be two independent $\Gamma(1/\gamma, 1/\gamma)$ random variables. Conditionally on (Δ_1, Δ_2), let the outcome Y be an $\mathrm{Exp}(\Delta_1)$ random variable, $\mathsf{P}(Y \le x \mid \Delta_1, \Delta_2) = 1 - e^{-\Delta_1 x}$ for $x \ge 0$, while the *misinformed forecast* F is equal to $\mathrm{Exp}(\Delta_2)$, i.e., $F(x) = 1 - e^{-\Delta_2 x}$ for $x \ge 0$. For $t \ge 0$, we have

$$\mathsf{P}(Y \le t) = \mathsf{E}[F(t)] = \mathsf{E}[1 - e^{-\Delta_j t}] = 1 - (1 + \gamma t)^{-1/\gamma} = \mathrm{GPD}_{1,\gamma}(t),$$

the generalized Pareto distribution with scale parameter 1 and shape parameter γ (see also Section 3 below). For $\mathcal{B} = \{\varnothing, \Omega\}$, condition (5) is trivially satisfied: by construction, the marginal occurrence ratio $\mathsf{P}(Y > t)$ is forecast correctly on average by $\mathsf{E}[1 - F(t)]$. However, condition (6) is violated, the reason being that the excess PIT of the $\mathrm{Exp}(\Delta_1)$-distributed excess $Y - t$ is performed with respect to the conditional excess distribution $F_t = \mathrm{Exp}(\Delta_2)$, that is, at the wrong parameter value, Δ_2 rather than Δ_1. Intuitively, Y is large when the rate parameter Δ_1 is close to zero, but then the scale $1/\Delta_1$ of the excess $Y - t$ is usually underestimated to be $1/\Delta_2$ instead. As a consequence, the limit in (6) is 0 instead of $u \in (0, 1)$. Details are provided in Appendix A in the supplement.

The following lemma on the upper endpoint of the outcome Y is useful for further results. Its proof is given in Appendix A.

Lemma 6. *On the probability space* $(\Omega, \mathcal{A}, \mathsf{P})$*, let* F *be a random probabilistic forecast, let* Y *be a real-valued outcome variable with upper endpoint* $x_Y \in \mathbb{R} \cup \{+\infty\}$*, and let* $\mathcal{B} \subseteq \mathcal{A}$ *be a* σ*-algebra. Assume* F *is continuous a.s. and assume the upper endpoint of the conditional distribution of* Y *given* $\mathcal{B}$ *is* x_Y*. If* (6) *holds, and in particular if* F *is tail* $\mathcal{B}$*-calibrated for* Y*, then* $\mathsf{P}(Y = x_Y) = 0$.

For non-random probabilistic forecasts F, the situation simplifies considerably, as it is sufficient to evaluate the occurrence ratio in order to deduce tail $\mathcal{B}$-calibration.

Proposition 7. *If the continuous probabilistic forecast* F *is non-random, then* F *is tail* $\mathcal{B}$*-calibrated if and only if*

$$\frac{\mathsf{P}(Y > t \mid \mathcal{B})}{1 - F(t)} \to 1, \qquad \textit{almost surely as } t \to x_Y. \tag{7}$$

The proof of Proposition 7 is given in Appendix A in the supplement.

2.3 Implications of tail calibration

The unfocused forecaster in Example 2 shows that probabilistic calibration is a strictly weaker requirement than auto-calibration if the forecast distribution F is not deterministic. Similarly, the following example introduces a *tail unfocused* forecaster, demonstrating that probabilistic tail calibration does not imply tail auto-calibration; see also Example 19 in Section 4. In concert with Corollary 11, it follows that tail auto-calibration is a strictly stronger requirement than probabilistic tail calibration.

Example 8 (Tail unfocused forecaster)**.** Let ξ, τ be independent random variables, $\xi \sim$ Exp(1) unit-exponential and $\mathsf{P}(\tau = +1) = \mathsf{P}(\tau = -1) = 1/2$. Define $a_\tau = \log((2+\tau)/2)$. Let $Y = \xi$ and, given τ, let F be the cdf of $\xi + a_\tau$. The random variable Y is unit-exponential, while F is the conditional cdf of Y with a random location shift of a_+ or a_-. The *tail unfocused* forecaster F is probabilistically tail calibrated but not tail auto-calibrated for Y.

The following lemma explains that, for F to be tail $\mathcal{B}$-calibrated, the graph of

$$u \mapsto \frac{\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]}$$

should be close to the diagonal in the supremum distance, which motivates the diagnostic plots in Section 4. The proof of Lemma 9 is given in Appendix A in the supplement.

Lemma 9. *The convergence relation* (4) *(or* (6)*) holds pointwise for all* u *in a dense subset of* $[0,1]$ *that includes* 1 *if and only if it holds uniformly in* $u \in [0,1]$*, i.e.,*

$$\sup_{u \in [0,1]} \left| \frac{\mathsf{P}\big(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\big)}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} - u \right| \to 0, \qquad \textit{almost surely as } t \to x_Y.$$

If information sets are nested, we have the following relation between the corresponding notions of tail calibration.

Proposition 10. *Let* $\mathcal{C} \subseteq \mathcal{B} \subseteq \mathcal{A}$ *be* σ*-algebras. Suppose that* F *is tail* $\mathcal{B}$*-calibrated for* Y*, and that the convergence in Definition 4 is in* L^∞*. Then* F *is also tail* $\mathcal{C}$*-calibrated for* Y*.*

Proof. We have

$$\begin{aligned}
&\frac{\mathsf{P}(Z_F^{(t)} \le u, Y > t \mid \mathcal{C})}{\mathsf{E}[1 - F(t) \mid \mathcal{C}]} - u = \mathsf{E}\left[\frac{\mathsf{P}(Z_F^{(t)} \le u, Y > t \mid \mathcal{B})}{\mathsf{E}[1 - F(t) \mid \mathcal{C}]} \;\Big|\; \mathcal{C} \right] - u \\
&= \mathsf{E}\left[\frac{\mathsf{P}(Z_F^{(t)} \le u, Y > t \mid \mathcal{B})}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} \cdot \frac{\mathsf{E}[1 - F(t) \mid \mathcal{B}]}{\mathsf{E}[1 - F(t) \mid \mathcal{C}]} \;\Big|\; \mathcal{C}_t \right] - u\, \mathsf{E}\left[\frac{\mathsf{E}[1 - F(t) \mid \mathcal{B}]}{\mathsf{E}[1 - F(t) \mid \mathcal{C}]} \;\Big|\; \mathcal{C}_t \right] \\
&= \mathsf{E}\left[\left\{ \frac{\mathsf{P}(Z_F^{(t)} \le u, Y > t \mid \mathcal{B})}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]} - u \right\} \cdot \frac{\mathsf{E}[1 - F(t) \mid \mathcal{B}]}{\mathsf{E}[1 - F(t) \mid \mathcal{C}]} \;\Big|\; \mathcal{C} \right].
\end{aligned}$$

Since the term in curly braces tends to zero in L^∞, F is tail $\mathcal{C}$-calibrated for Y. □

Corollary 11. *If the convergence in Definition 4 is in* L^∞ *for* $\mathcal{B} = \sigma(F)$*, tail auto-calibration implies probabilistic tail calibration. The converse is false.*

Proof. This is a consequence of Proposition 10 with $\mathcal{B} = \sigma(F)$ and $\mathcal{C} = \{\varnothing, \Omega\}$. For the converse, see Examples 8 and 19. □

The larger the conditioning set $\mathcal{B}$ is, the stronger the notion of tail calibration. Since proper scoring rules are not available as an evaluation tool for tails of forecast distributions, one could alternatively term a forecast better the larger the set $\mathcal{B}$ with respect to which it is tail calibrated. In the case of classical notions of calibration, this has been formalized and studied by Strähl and Ziegel (2017), who considered cross-calibration of forecasters, that is, calibration conditional on the information of all other relevant forecasters. This approach translates to the evaluation of tail calibration, see also Section 4.2.

We can identify relationships between standard notions of calibration and their tail counterparts.

Proposition 12. *(a) If $\sigma(F) \subseteq \mathcal{B}$, then $\mathcal{B}$-calibration implies tail $\mathcal{B}$-calibration. In particular, auto-calibration implies tail auto-calibration but the converse is false.*

(b) Probabilistic calibration does not imply probabilistic tail calibration or vice versa.

Proof. Suppose that F is $\mathcal{B}$-calibrated with $\sigma(F) \subseteq \mathcal{B}$. For $B \in \mathcal{B}$, $u \in [0,1]$, and $t < x_Y$, we obtain

$$\begin{aligned}\mathsf{E}[\mathbb{1}\{F_t(Y-t) \leq u,\, Y > t\}\mathbb{1}_B] &= \mathsf{E}[\mathsf{E}[\mathbb{1}\{Y \leq F^{-1}(u(1-F(t)) + F(t))\}\mathbb{1}\{Y > t\} \mid \mathcal{B}]\mathbb{1}_B] \\ &= \mathsf{E}[(u(1-F(t)) + F(t) - F(t))\mathbb{1}_B] = u\,\mathsf{E}[(1-F(t))\mathbb{1}_B],\end{aligned}$$

since the conditional distribution of Y given $\mathcal{B}$ is F, and F is assumed to be continuous. This implies $\mathsf{E}\big[\mathbb{1}\{Z_F^{(t)} \leq u,\, Y > t\} - u(1-F(t)) \mid \mathcal{B}\big] = 0$ and thus $\mathsf{P}\big(Z_F^{(t)} \leq u,\, Y > t \mid \mathcal{B}\big) = u\,\mathsf{E}[1 - F(t) \mid \mathcal{B}]$, showing the sufficiency in part (a). For the converse, see Example 16.

Part (b) is shown by Examples 16 and 17. □

Remark 13 (Unconditional tail calibration)**.** Marginal and probabilistic calibration are two unconditional notions of forecast calibration. Due to its greater relevance in practice, we opt for probabilistic calibration as a starting point for defining notions of tail calibration. In Appendix C in the supplement, we explore an option based on marginal calibration.

While auto-calibration implies tail auto-calibration, probabilistic calibration does not imply probabilistic tail calibration, and similarly for marginal calibration and its tail counterpart, see Appendix C in the supplement. This highlights a shortcoming of prominent classical calibration definitions when interest is in the tails—in this sense, we extend the negative result regarding scoring rules and extremes to notions of calibration. The proposed notions of tail calibration therefore provide information that is not available from existing forecast verification tools.

Remark 14 (Endpoint)**.** In Definition 4, it is assumed that the upper endpoint of the conditional distribution of Y given $\mathcal{B}$ is not random and is equal to x_Y. In some scenarios, this assumption could be restrictive. For example, if the information in $\mathcal{B}$ determines whether an insurance-policy holder has a positive probability of filing a large claim Y or not, then the upper endpoint of the conditional distribution is random. One possible extension of the definition of tail $\mathcal{B}$-calibration would be to assume that there exists a $\mathcal{B}$-measurable random variable $x_{\mathcal{B}}$ such that the upper endpoint of Y given $\mathcal{B}$ is $x_{\mathcal{B}}$. Details for this added generality would have to be worked out.

3 Probabilistic tail calibration and peaks-over-threshold models

We argue that probabilistic tail calibration should routinely be checked to evaluate probabilistic forecasts with regards to their tails. Therefore, it is interesting to understand how stringent the requirement of probabilistic tail calibration is, and here we provide sufficient conditions. Consider the following assumption on the outcome variable Y:

(A1) The distribution of Y is in the max-domain of attraction of a non-degenerate distribution.

Assumption (A1) is standard in extreme value theory, and by Balkema and de Haan (1974), it is equivalent to the assumption that there exist $\xi \in \mathbb{R}$ and a scaling function $\sigma_Y(\,\cdot\,) > 0$ such that

$$\lim_{t \to x_Y} \mathsf{P}\big(Y > t + \sigma_Y(t)x \mid Y > t\big) = 1 - \mathrm{GPD}_{1,\xi}(x) = (1 + \xi x)_+^{-1/\xi}, \quad \text{for all } x \geq 0, \tag{8}$$

where $(y)_+ = \max(y, 0)$ and where $(1 + \xi x)_+^{-1/\xi}$ is to be understood as $\exp(-x)$ when $\xi = 0$. The parametric family in (8) specifies the generalized Pareto distribution (GPD) with shape parameter $\xi \in \mathbb{R}$ and unit scale. For a general scale parameter $\sigma > 0$, we define $\mathrm{GPD}_{\sigma,\xi}(x) = \mathrm{GPD}_{1,\xi}(x/\sigma)$, $x \geq 0$. The GPD is the basis of the peaks-over-threshold method in Davison and Smith (1990) used to approximate the tails of distributions satisfying (A1). The shape parameter determines the rate of tail decay, with power-law decay for $\xi > 0$, exponential decay for $\xi = 0$, and polynomial decay towards a finite upper bound for $\xi < 0$.

Next, consider a similar assumption on the random forecast distribution F, formulated in terms of the conditional forecast excess distribution F_t in (2):

(A2) The upper endpoint $x_F = \sup\{x \in \mathbb{R} : F(x) < 1\} \in \mathbb{R} \cup \{+\infty\}$ of F is deterministic and satisfies $x_F = x_Y$, and there exist $\eta \in \mathbb{R}$ and a deterministic scaling function $\sigma_F(\,\cdot\,) > 0$ so that

$$\lim_{t \to x_F} \mathsf{P}\left(\left|1 - F_t(\sigma_F(t)x) - (1 + \eta x)_+^{-1/\eta}\right| > \varepsilon \,\middle|\, Y > t\right) = 0, \quad \text{for all } x \geq 0,\, \varepsilon > 0. \tag{9}$$

In words, conditionally on $Y > t$, the rescaled conditional forecast excess distribution $F_t(\sigma_F(t)\,\cdot\,)$ converges in probability to a fixed GPD.

Proposition 15 (Sufficient condition for probabilistic tail calibration)**.** *Suppose that Assumptions (A1) and (A2) are fulfilled and that* $\lim_{t \to x_Y} \mathsf{P}(Y > t)/\,\mathsf{E}[1 - F(t)] = 1$. *The following conditions are equivalent:*

(a) $\xi = \eta$ *and* $\lim_{t \to x_Y} \sigma_F(t)/\sigma_Y(t) = 1$*;*

(b) F *is probabilistically tail calibrated for* Y*.*

For the proof of Proposition 15, see the more general Proposition 31 in Appendix C in the supplement. If the scale parameter $\sigma_F(t)$ in (9) is allowed to be random, then F can be probabilistically tail calibrated for Y even when the generalized Pareto shape parameters ξ and η in (8) and (9) differ. For instance, if $\Delta \sim \Gamma(1/\gamma, 1/\gamma)$ for some $\gamma > 0$ and if, conditionally on Δ, the distribution of Y is $F = \text{Exp}(\Delta)$, i.e., $\mathsf{P}(Y \leq x \mid \Delta) = 1 - e^{-\Delta x} = F(x)$ for $x \geq 0$, then the unconditional distribution of Y is $\text{GPD}_{1,\gamma}$, so that (8) holds with $\xi = \gamma$, while F_t is $\text{Exp}(\Delta)$ too, and thus $F_t(x/\Delta) = 1 - e^{-x}$, which is (9) with random scale parameter $\sigma_F(t) = 1/\Delta$ and with $\eta = 0$.

Assumptions (A1), (A2), and property (a) of Proposition 15 are sufficient but not necessary conditions for probabilistic tail calibration. Indeed, according to Proposition 7, for non-random probabilistic forecasts, it is sufficient that $\mathsf{P}(Y = x_Y) = 0$ and $\mathsf{P}(Y > t)/(1 - F(t)) \to 0$ as $t \to x_Y$.

4 Examples and diagnostic tools

In practice, we observe forecast-observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$ that are realizations of (F, Y), and evaluation must be performed using this finite sample. Let $\mathcal{I}_t = \{i \in \{1, \ldots, n\} : y_i > t\}$ be the set of indices corresponding to observations that exceed t, put $n_t = |\mathcal{I}_t|$, and let $F_{i,t}$ denote the conditional excess distribution of forecast F_i at threshold t.

4.1 Probabilistic tail calibration

First, consider the simplest case, $\mathcal{B} = \{\varnothing, \Omega\}$, corresponding to probabilistic (tail) calibration. Probabilistic calibration is easy to assess in practice by checking whether the PIT values $F_1(y_1), \ldots, F_n(y_n)$ resemble a sample from a standard uniform distribution.

To assess probabilistic tail calibration, consider, for some large threshold t, the n_t conditional PIT values

$$z_i^{(t)} = F_{i,t}(y_i - t) = \frac{F_i(y_i) - F_i(t)}{1 - F_i(t)}$$

for $i \in \mathcal{I}_t$, which correspond to realizations of the random variable $Z_F^{(t)}$. To assess (4) with $\mathcal{B} = \{\varnothing, \Omega\}$ using a finite sample, we can choose a large threshold t, calculate

$$\hat{R}_t(u) = \frac{\sum_{i \in \mathcal{I}_t} \mathbb{1}\{z_i^{(t)} \leq u\}}{\sum_{i=1}^{n} (1 - F_i(t))}, \tag{10}$$

and plot this quantity as a function of $u \in [0, 1]$. If the curve lies close to the diagonal, there is evidence to suggest that the forecasts are probabilistically tail calibrated. By including the same plot for $t = -\infty$, we recover a *pp*-plot of the PIT values $z_i = F_i(y_i)$, from which we can assess standard probabilistic calibration.

If the graph of the combined ratio $u \mapsto \hat{R}_t(u)$ does not lie along the diagonal, then it may be beneficial to consider additional diagnostics derived from (5) and (6) instead of (4). For example, if the graph is consistently below the diagonal, this could be either because the numerator is too small, suggesting a bias in the conditional distribution, or because the denominator is too large, suggesting a bias in the forecast exceedance probabilities.

The first condition, (5), requires that the forecasts issue reliable threshold exceedance probabilities. This can be assessed empirically by checking whether the occurrence ratio

$$\frac{\sum_{i=1}^{n} \mathbb{1}\{y_i > t\}}{\sum_{i=1}^{n}(1 - F_i(t))}, \tag{11}$$

tends towards one as the threshold t increases. A ratio larger (smaller) than one suggests that the forecasts under-estimate (over-estimate) threshold exceedance probabilities. Previously, other works have evaluated the performance of probabilistic forecasts with regards to tails by considering threshold exceedance probabilities, see for example Williams et al. (2014); Taillardat et al. (2019); Allen et al. (2023). When these are assessed using reliability diagrams, this assessment is closely related to our assessment of the occurence ratio. However, Taillardat et al. (2019) have looked at receiver operating characteristic (ROC) curves for a fixed threshold t, which assesses discrimination ability of the probability forecast for the extreme event rather than calibration of predicted tails (Dimitriadis et al., 2024).

The second condition, (6), considers the conditional distribution given that the threshold has been exceeded, allowing us to assess forecasts for the severity of an extreme event. Therefore, the conditional PIT values $z_i^{(t)}$ for $i \in \mathcal{I}_t$ should resemble a sample from a standard uniform distribution. This can again be examined using a histogram or a *pp*-plot, and the shape of the diagnostic plot can be used to identify whether the conditional distribution is heavy- or light-tailed (Allen et al., 2023).

Example 16 (Tail auto-calibration does not imply auto-calibration)**.** Let F be a deterministic forecast for which there exists a threshold t_0 such that $\mathsf{P}(Y > t_0) > 0$ and $F(y) = \mathsf{P}(Y \le y)$ for all $y \ge t_0$. Trivially, F is tail auto-calibrated, and hence probabilistically tail calibrated. However, $F(y)$ and $\mathsf{P}(Y \le y)$ can be very different for $y < t_0$, in which case F is not probabilistically calibrated. For example, consider the outcome $Y \sim \mathrm{GPD}_{1,1/4}$ together with the forecast distribution $F(x) = \mathrm{GPD}_{4/5,1/4}(x-1)$ for $x < 5$ and $F(x) = \mathrm{GPD}_{1,1/4}(x)$ for $x \ge 5$. Figure 2 shows the diagnostic plots for probabilistic tail calibration in this case.

Example 17 (Probabilistic calibration does not imply probabilistic tail calibration)**.** Consider a generalization of the unfocused forecaster in Gneiting et al. (2007), see Example 2. Let G be a distribution function supported on a compact interval with continuous positive density. Let $Y \sim G$, let τ be independent of Y and equal to $+1$ or -1 with probability $1/2$ and, given τ, let $F(y) = \frac{1}{2}\{G(y)+G(y+\tau)\}$ for $y \in \mathbb{R}$. Then F is probabilistically calibrated but not necessarily probabilistically tail calibrated; see Appendix A for details. Figure 3 illustrates the example for $G = \mathrm{Unif}(0,1)$: the curves for $t = -1$ confirm probabilistic calibration but those for $t > -1$ indicate the lack of probabilistic tail calibration.

Example 18 (Ideal, climatological and extremist forecasters)**.** The following example was introduced by Taillardat et al. (2023) for studying methods to evaluate forecast tails. Let

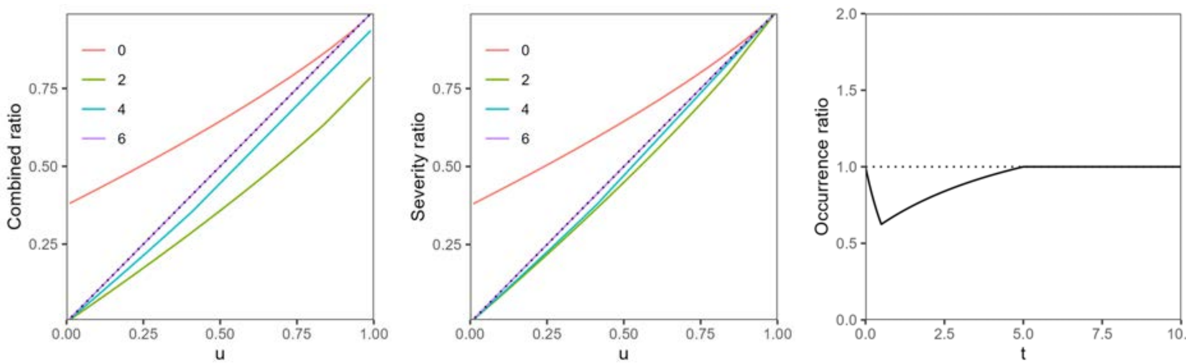

Figure 2: Probabilistic tail calibration diagnostic plots for Example 16. The different colors correspond to different thresholds t. Left: combined ratio (10); middle: pp-plot of $z_i^{(t)}$ for $i \in \mathcal{I}_t$; right: occurrence ratio (11) as a function of t.

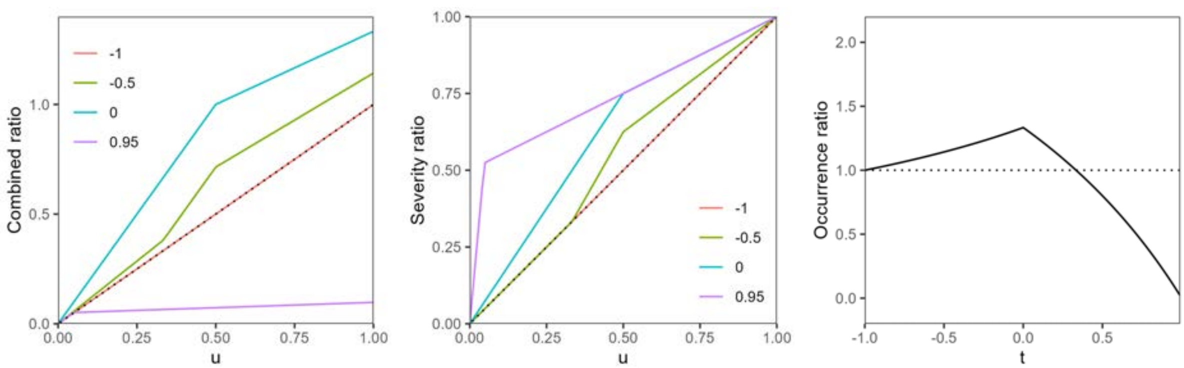

Figure 3: Probabilistic tail calibration diagnostic plots for the unfocused forecaster in Example 17 with G the uniform distribution on $[0, 1]$. The different colors correspond to different thresholds t. Left: combined ratio (10); middle: pp-plot of $z_i^{(t)}$ for $i \in \mathcal{I}_t$; right: occurrence ratio (11) as a function of t.

$\Delta \sim \Gamma(1/\gamma, 1/\gamma)$, where $\gamma > 0$ and $\Gamma(a, b)$ is the gamma distribution with shape $a > 0$ and scale $b > 0$. Conditionally on Δ, let $Y \sim \text{Exp}(\Delta)$ follow an exponential distribution with rate Δ. The unconditional distribution of Y is the heavy-tailed $\text{GPD}_{1,\gamma}$. As in Taillardat et al. (2023), results are presented for $\gamma = 1/4$.

Consider three different forecasters: the ideal forecaster, $F_{\text{id}} = \text{Exp}(\Delta)$; the climatological forecaster, $F_{\text{cl}} = \text{GPD}_{1,\gamma}$; and the extremist forecaster, $F_{\text{ex},\nu} = \text{Exp}(\Delta/\nu)$ for $\nu > 0$. The variable Δ represents a source of information that may or may not be available to the forecasters. The ideal forecaster correctly represents the conditional distribution of Y given Δ and is auto-calibrated, hence also probabilistically (tail) calibrated. The climatological forecaster does not use the information Δ, instead issuing the unconditional distribution of Y as a forecast, but is nevertheless also auto-calibrated. The extremist forecaster correctly predicts that, conditionally on Δ, the outcome Y follows an exponential distribution, but with a multiplicative bias on the rate (here, $\nu = 1.4$), and is not probabilistically (tail) calibrated. Given the information Δ, the climatological forecaster belongs to a different

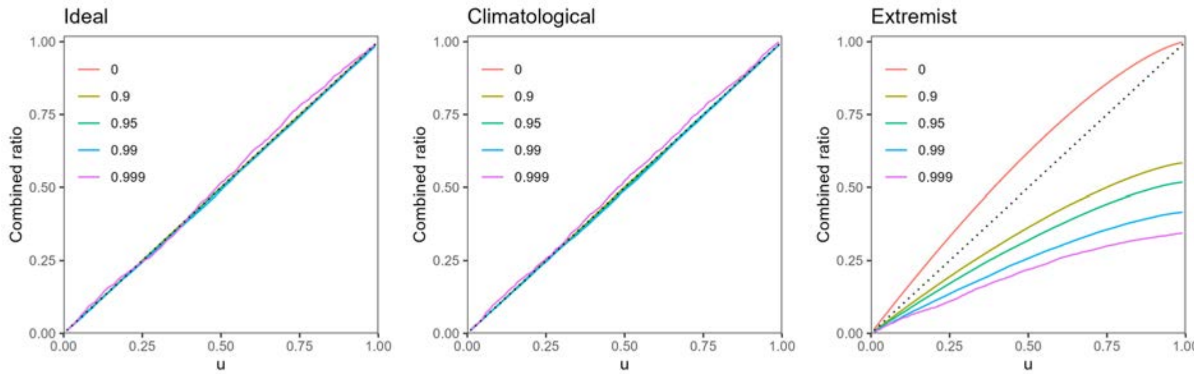

Figure 4: Combined probabilistic tail calibration diagnostic plots (10) for the ideal, climatological, and extremist forecasters in Example 18. Results are shown for five thresholds t, expressed in terms of quantiles of the 10^6 observations.

tail regime than the outcome (heavy-tailed rather than light-tailed), whereas the extremist forecaster belongs to the correct tail regime (light-tailed).

Figure 4 displays the diagnostic plots for the tail calibration of the three forecasters using $n = 10^6$ independent realizations of Δ and Y, see also Figure 9. The extremist forecast clearly has a heavier tail than the true distribution of the outcomes, and the nature of this miscalibration does not change as the threshold is increased.

In Appendix D, we examine the unfocused forecaster of Example 2 with regards to probabilistic (tail) calibration.

4.2 Tail $\mathcal{B}$-calibration

While it is most common in practice to assess probabilistic calibration, it is also useful to be able to assess calibration with respect to larger information sets. The same is true when interest is in extremes. For example, while the uninformative climatological forecaster in Example 18 is probabilistically tail calibrated (because it exhibits the same tail behavior as the unconditional distribution of the outcome), it is not $\sigma(\Delta)$-tail calibrated. In this section, we propose tools to assess tail $\mathcal{B}$-calibration for non-trivial σ-fields $\mathcal{B}$.

The forecast-observation pairs $(F_1, y_1), \dots, (F_n, y_n)$ can be stratified into $J \in \mathbb{N}$ mutually disjoint bins, $B_1, \dots, B_J$, and to assess tail $\mathcal{B}$-calibration, this binning can be performed using variables that represent the information contained in $\mathcal{B}$. The diagnostic plots for probabilistic tail calibration can then be applied separately to each bin of the forecasts and observations; if the forecasts appear probabilistically calibrated for all bins, then there is evidence to suggest that they are tail $\mathcal{B}$-calibrated. For example, the combined ratio in (10) can be calculated for each bin,

$$\hat{R}_{t,j}(u) = \frac{\sum_{i \in \mathcal{I}_t \cap \mathcal{J}_j} \mathbb{1}\{z_i^{(t)} \leq u\}}{\sum_{i \in \mathcal{J}_j} (1 - F_i(t))}, \tag{12}$$

where $\mathcal{J}_j = \{i \in \{1, \ldots, n\} : (F_i, y_i) \in B_j\}$ for $j = 1, \ldots, J$, and this ratio can be displayed as a function of $u \in [0, 1]$ for each of the J bins. Similarly, the conditional PIT values $z_i^{(t)}$ and the occurrence ratio (11) can be assessed for each bin separately.

Displaying the diagnostic plots for each forecast method and each bin leads to a large number of plots to analyze. To simplify visualization, (tail) mis-calibration can be summarized in a single value. This aligns with the general framework to assess conditional calibration proposed by Tsyplakov (2011) based on moment conditions and test functions, which is akin to how forecasts are compared in Giacomini and White (2006) and Nolde and Ziegel (2017). As a measure of tail calibration, consider the supremum distance $\sup_{u \in [0,1]} |\hat{R}_t(u) - u|$, which quantifies the maximum absolute difference between the combined ratio and the diagonal line in the diagnostic plot proposed above. Under probabilistic tail calibration, this difference should be close to zero for large t. To assess tail $\mathcal{B}$-calibration, $\hat{R}_t$ can be replaced by $\hat{R}_{t,j}$, and this distance can be calculated for $j = 1, \ldots, J$.

In Example 18, the random variable Δ controls the information governing the observations, and to assess $\sigma(\Delta)$-calibration, we can stratify the forecast-observation pairs depending on the realized Δ. In practice, where we do not know the true data generating process, Δ could be replaced by covariate information available to the forecasters, for example. The supremum distances

$$\sup_{u \in [0,1]} |\hat{R}_{t,j}(u) - u|, \qquad j = 1, \ldots, J, \tag{13}$$

for the ideal, climatological and extremist forecasters are shown as a function of the threshold in Figure 5, for $J = 3$ bins for Δ. For the ideal forecaster, the difference between $\hat{R}_{t,j}(u)$ and u is essentially zero for all bins, confirming that it is tail $\sigma(\Delta)$-calibrated. The climatological forecaster, on the other hand, is clearly not calibrated conditionally on Δ, despite being probabilistically tail and tail auto-calibrated. These plots for conditional calibration are shown for the empirical analogue (12) of the combined ratio in (4). As for probabilistic tail calibration in Section 4.1, an analogous approach can be employed to evaluate the occurrence ratio (5) or the severity ratio (6).

While the climatological forecaster in Example 18 is not tail $\sigma(\Delta)$-calibrated, it is tail auto-calibrated. However, the following example demonstrates that a forecast can be probabilistically tail calibrated but not tail auto-calibrated.

Example 19 (Probabilistic tail calibration does not imply tail auto-calibration)**.** Suppose that $\Delta \sim \Gamma(1/\gamma, 1/\gamma)$ for some $\gamma > 0$ and, conditionally on Δ, let $X \sim \mathrm{Exp}(\Delta)$. Let the random variable $L \sim \mathrm{GPD}_{1,\gamma/2}$ be independent of (Δ, X). Define Y as the second largest of $(X, 2X, L)$. Given Δ, the *optimistic forecaster* issues the random probabilistic forecast $F = \mathrm{Exp}(\Delta)$, the conditional distribution of X. This forecast is probabilistically tail calibrated, but not tail auto-calibrated; see Figure 6 for the diagnostic plots and see Appendix A for a formal analysis of a more general construction. An intuitive explanation is as follows: the marginal distribution of X is $\mathrm{GPD}(1, \gamma)$, which is heavier than the one of L; it follows that $\mathsf{P}(Y = X \mid Y > t) \to 1$ as $t \to \infty$ and the marginal tails of X and Y are equivalent, yielding probabilistic tail calibration. However, conditionally on Δ, the tail of F

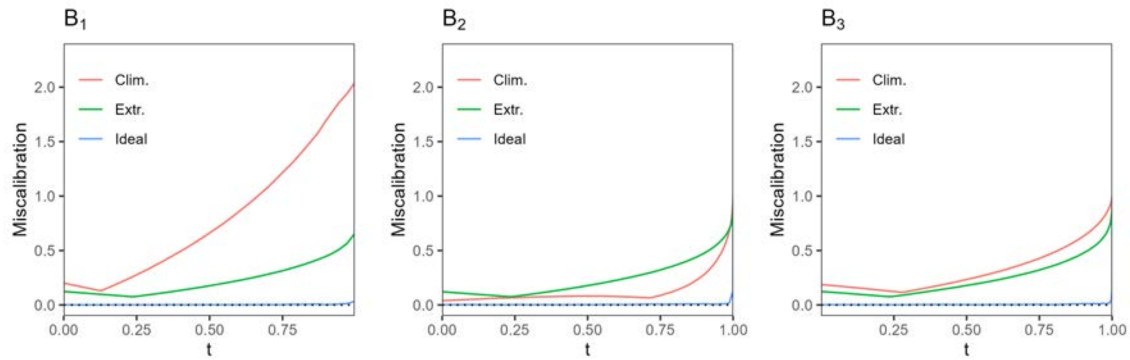

Figure 5: Maximal distance (13) of $\hat{R}_{t,j}$ from the diagonal as a function of t for the climatological (red), ideal (blue), and extremist (green) forecasters in Example 18 for $J = 3$ bins (left to right) of Δ. The threshold t is expressed as a quantile of the 10^6 observations.

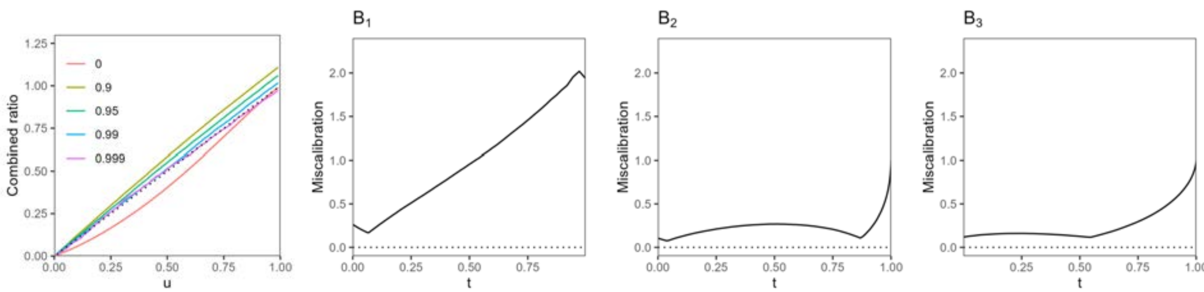

Figure 6: First plot: Combined probabilistic tail calibration diagnostic plot in (10) for the random forecast in Example 19, for five thresholds t. Second, third, and fourth plots: Distance (13) of $\hat{R}_{t,j}$, $j = 1, 2, 3$, from the diagonal as a function of t for $J = 3$ bins of Δ. In all plots, thresholds are expressed in terms of quantiles of 10^6 observations.

is *lighter* than the one of L, and the forecast is too optimistic, i.e., it assesses the tail of Y lighter than it really is, violating tail auto-calibration.

5 Application

Reliable forecasts for extreme precipitation are highly valuable for mitigating the impact of heavy rain and snowfall. We consider precipitation data from the European Meteorological Network's (EUMETNET) post-processing benchmark dataset (EUPPBench; Demaeyer et al., 2023), which contains ensemble forecasts issued by the European Centre for Medium-range Weather Forecasts' (ECMWF) Integrated Forecast System (IFS). Forecasts are available daily for 2017 and 2018, at 112 stations across central Europe. We restrict attention to forecasts issued one day in advance, for which 81548 forecast-observation pairs are available. The forecasts are evaluated using precipitation measurements at the stations.

Three forecasting methods are compared. Firstly, we consider the 51-member ensemble forecasts generated by the ECMWF's IFS ensemble. Ensemble forecasts are collections of point forecasts, forming an empirical predictive distribution function. An empirical distribution comprised of 51 sample values is unlikely to be useful when predicting extremes; in particular, the resulting forecast will often assert that high thresholds will be exceeded with probability zero. We therefore also assess the performance of a smoothed version of the discrete ensemble forecast, which converts the IFS ensemble to a continuous forecast distribution by assuming that precipitation observations follow a censored logistic distribution, with location and scale parameters that are equal, respectively, to the mean and standard deviation of the corresponding IFS ensemble members. Finally, we present results for a statistical post-processing model that aims to remove systematic biases in the IFS ensemble to yield calibrated forecasts. The post-processing method is a simple ensemble model output statistics approach (Gneiting et al., 2005), which also assumes that precipitation follows a censored logistic distribution. The location and scale parameters of the censored logistic distribution are affine functions of the IFS ensemble mean and standard deviation, respectively. Such a post-processing model is commonly employed in studies of precipitation forecasts (see e.g. Messner et al., 2014). The model parameters are estimated by minimizing the continuous ranked probability score (CRPS) over 20 years of twice-weekly reforecasts at the corresponding station and lead time; further details regarding the data can be found in Demaeyer et al. (2023).

Figure 7 displays the probabilistic (tail) calibration diagnostic plots for the three forecast methods. For the discrete ensemble forecasts, the diagnostics for tail calibration are adapted as described in Remark 1. Results are shown at thresholds of 5, 10, and 15mm, roughly corresponding to the 97^{th}, 99^{th}, and 99.5^{th} percentiles of the previously observed precipitation measurements (across all stations). The raw IFS ensemble forecasts are under-dispersed and positively biased, whereas post-processing yields considerably more reliable forecasts. However, all forecasts are miscalibrated when focus is placed on extreme precipitation; the diagnostic plots indicate that all forecasts have a tail that is too light. This becomes more severe as the threshold of interest increases, suggesting that the forecasts are not probabilistically tail calibrated. Smoothing the IFS ensemble unsurprisingly improves tail calibration, but appears to worsen the overall calibration of the forecasts in this example. While post-processing improves upon the calibration of the raw ensembles, the resulting forecasts are still acutely miscalibrated when predicting extreme outcomes.

6 Discussion

Brehmer and Strokorb (2019) demonstrate that for any proper scoring rule, it is possible to find a forecast distribution that exhibits incorrect tail behavior, but receives an expected score that is arbitrarily close to that of the true distribution. We demonstrate that this negative result cannot be circumvented by evaluating forecasts using multiple scoring rules. As an alternative, we introduce a general framework of forecast tail calibration that allows the reliability of forecasts to be assessed when interest is in extreme outcomes.

Tail calibration is generally not implied by standard calibration, and therefore provides additional information that is not available from existing checks for forecast calibration.

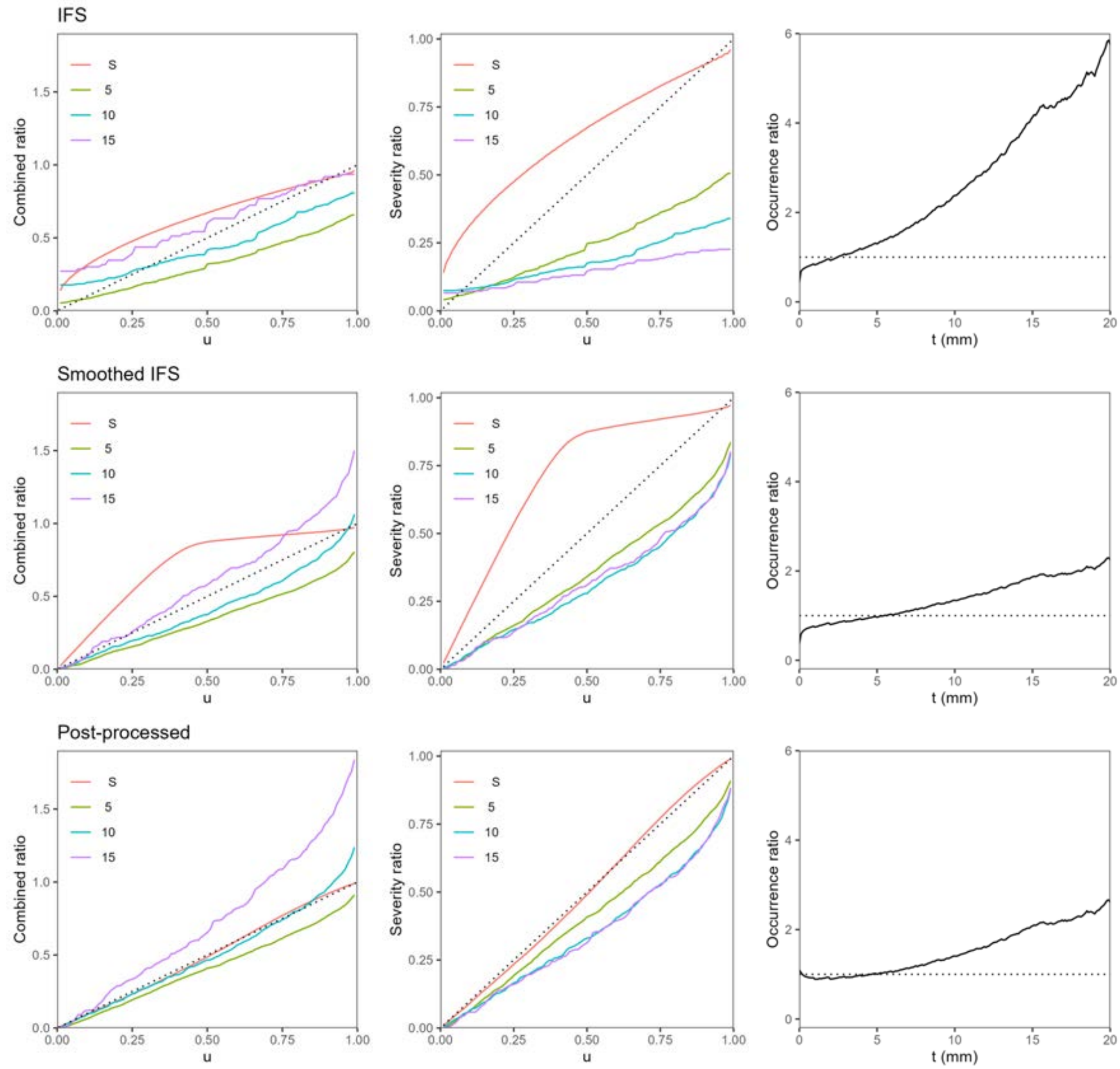

Figure 7: Probabilistic tail calibration plots for the IFS and statistically post-processed forecast distributions in Section 5. Left: combined ratio (10); middle: pp-plot of z_i^t for $i \in \mathcal{I}_t$; right: occurrence ratio (11) as a function of t. Results are shown for three thresholds, 5, 10, and 15mm, roughly corresponding to the 97^{th}, 99^{th}, and 99.5^{th} percentiles of the previously observed precipitation measurements. The symbol 'S' denotes the standard notion of probabilistic calibration (i.e. for $t = -\infty$).

This is illustrated in a case study on European precipitation forecasts, whereby statistical post-processing methods are found to produce forecasts that are calibrated despite issuing unreliable predictions for extreme outcomes. By introducing a method to evaluate forecasts for extreme outcomes, it becomes straightforward to identify deficiencies in predictions of these events, thereby facilitating the development of forecasting methods that can more reliably predict extreme outcomes.

While tail calibration can be assessed with respect to different information sets, we imagine the unconditional case will be of most interest in practical applications. To assess probabilistic tail calibration, we introduce graphical diagnostic tools that not only demonstrate whether or not a forecast is tail calibrated, but also elucidate what types of biases occur in the tail; for example, whether the predictive distributions are too light or heavy tailed. Formal testing procedures for tail calibration are an interesting avenue for future work, where the crux will be to choose a data-dependent threshold that is representative of the asymptotic regime. Such tests could either be classical tests for fixed pre-specified sample sizes or monitoring procedures as introduced by Arnold and Ziegel (2023), who propose a sequential procedure for probabilistic calibration based on e-processes, in recognition of the fact that forecasting is often an inherently sequential task.

Tail calibration may partially address the open question how best to compare probabilistic forecasts for large outcomes. While calibration is principally an absolute facet of probabilistic forecast performance, tail $\mathcal{B}$-calibration can be used for forecast comparison by finding the largest information set $\mathcal{B}$ on which forecasts are tail calibrated, with larger sets being better. However, for imperfect forecasts and non-nested maximal information sets, this does still not allow for comparison.

When the unconditional distribution of the outcome variable is in the max-domain of attraction of a non-degenerate distribution, i.e., satisfying Assumption (A1), probabilistic tail calibration is equivalent to accurately forecasting the shape and scale parameters of the GPD. Checks for probabilistic tail calibration therefore serve as a more general framework with which to validate tail models, since they are also applicable to distributions not satisfying Assumption (A1). Future work may investigate whether notions of tail calibration can conversely be used to estimate quantities relevant for extreme value analysis, such as tail indices.

Finally, while we focus here on univariate extremes, future work could consider an extension to the multivariate case. In recent reviews of probabilistic forecasting, Gneiting and Katzfuss (2014), Kneib et al. (2023), and Petropoulos et al. (2022) all remark that developing more theoretically-principled techniques to generate and evaluate probabilistic forecasts for multivariate outcomes is one of the most pressing issues in the field. General notions of multivariate probabilistic calibration have been proposed (e.g. Gneiting et al., 2008; Allen et al., 2024), but these typically project the multivariate forecasts and observations onto the univariate domain, and extremes in the transformed space may not be of practical interest. Instead, a more relevant definition of multivariate forecast calibration may be obtained by leveraging results in multivariate extreme value theory.

References

Allen, S., J. Bhend, O. Martius, and J. Ziegel (2023). Weighted verification tools to evaluate univariate and multivariate probabilistic forecasts for high-impact weather events. *Weather and Forecasting 38*, 499–516.

Allen, S., D. Ginsbourger, and J. Ziegel (2023). Evaluating forecasts for high-impact events using transformed kernel scores. *SIAM/ASA Journal on Uncertainty Quantification 11*, 906–940.

Allen, S., J. Ziegel, and D. Ginsbourger (2024). Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society 150*(760), 1315–1335.

Arnold, S. and J. Ziegel (2023). Isotonic conditional laws. *Preprint,* `arXiv:2307.09032`.

Balkema, A. A. and L. de Haan (1974). Residual life time at great age. *The Annals of Probability 2*(5), 792–804.

Bellier, J., I. Zin, and G. Bontron (2017). Sample stratification in verification of ensemble forecasts of continuous scalar variables: Potential benefits and pitfalls. *Monthly Weather Review 145*, 3529–3544.

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

Davison, A. C. and R. L. Smith (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society. Series B (Methodological) 52*(3), 393–442.

Demaeyer, J., J. Bhend, S. Lerch, C. Primo, B. Van Schaeybroeck, A. Atencia, Z. Ben Bouallègue, J. Chen, M. Dabernig, G. Evans, et al. (2023). The EUPPBench postprocessing benchmark dataset v1.0. *Earth System Science Data 15*, 2635–2653.

Diebold, F. X., T. A. Gunther, and A. S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review 39*, 863–883.

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

Dimitriadis, T., T. Gneiting, A. I. Jordan, and P. Vogel (2024). Evaluating probabilistic classifiers: The triptych. *International Journal of Forecasting 40*, 1101–1122.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*, 1545–1578.

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

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

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

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review 133*, 1098–1118.

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

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

Gneiting, T. and J. Resin (2023). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics 17*, 3226–3286.

Gneiting, T., L. I. Stanberry, E. P. Grimit, L. Held, and N. A. Johnson (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test 17*, 211–235.

Harvey, D. I., S. J. Leybourne, and P. Newbold (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics 16*, 254–259.

Holzmann, H. and B. Klar (2017). Focusing on regions of interest in forecast evaluation. *The Annals of Applied Statistics 11*, 2404–2431.

Kallenberg, O. (2002). *Foundations of Modern Probability* (second ed.). Springer-Verlag.

Kneib, T., A. Silbersdorff, and B. Säfken (2023). Rage against the mean–a review of distributional regression approaches. *Econometrics and Statistics 26*, 99–123.

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

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks (2014). Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Monthly Weather Review 142*, 448–456.

Mitchell, J. and M. Weale (2023). Censored density forecasts: Production and evaluation. *Journal of Applied Econometrics 38*, 714–734.

Modeste, T. (2023). *Évaluation et construction des prévisions probabilistes : score et calibration dans un cadre dynamique.* Ph. D. thesis, Université Claude Bernard Lyon 1.

Nolde, N. and J. F. Ziegel (2017). Elicitability and backtesting: Perspectives for banking regulation. *The Annals of Applied Statistics 11*, 1833–1874.

Olafsdottir, H. K., H. Rootzén, and D. Bolin (2024). Locally tail-scale invariant scoring rules for evaluation of extreme value forecasts. *International Journal of Forecasting (in press).*

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. B. Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, et al. (2022). Forecasting: Theory and practice. *International Journal of Forecasting 38*, 705–871.

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

Taillardat, M., A.-L. Fougères, P. Naveau, and R. de Fondeville (2023). Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting 39*(3), 1448–1459.

Taillardat, M., A.-L. Fougères, P. Naveau, and O. Mestre (2019). Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting. *Weather and Forecasting 34*, 617–634.

Tsyplakov, A. (2011). Evaluating density forecasts: a comment. Available at SSRN 1907799, `https://dx.doi.org/10.2139/ssrn.1907799`.

Tsyplakov, A. (2014). Theoretical guidelines for a partially informed forecast examiner. MPRA paper 67333, `https://mpra.ub.uni-muenchen.de/67333/`.

Williams, R., C. Ferro, and F. Kwasniok (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society 140*, 1112–1120.

A Omitted proofs and calculations

Proof that the misinformed forecaster in Example 5 does not satisfy (6). On the event $Y > t$, the excess PIT of $Y - t$ by F_t is

$$Z_F^{(t)} = \frac{F(Y) - F(t)}{1 - F(t)} = \frac{e^{-\Delta_2 t} - e^{-\Delta_2 Y}}{e^{-\Delta_2 t}} = 1 - e^{-\Delta_2 (Y - t)}$$

so that

$$\begin{aligned} \mathsf{P}\big(Z_F^{(t)} \leq u, Y > t\big) &= \mathsf{P}\big(t < Y \leq t - \Delta_2^{-1} \log(1-u)\big) \\ &= \mathsf{E}\big[e^{-\Delta_1 t} - e^{-\Delta_1 (t - \Delta_2^{-1} \log(1-u))}\big] \\ &= (1 + \gamma t)^{-1/\gamma} - \mathsf{E}\big[\big\{1 + \gamma(t - \Delta_2^{-1} \log(1-u))\big\}^{-1/\gamma}\big]. \end{aligned}$$

For $u \in (0,1)$, we thus find, by the dominated convergence theorem, as $t \to \infty$,

$$\frac{\mathsf{P}\big(Z_F^{(t)} \leq u, Y > t\big)}{\mathsf{P}(Y > t)} = 1 - \mathsf{E}\left[\left(\frac{1 + \gamma(t - \Delta_2^{-1} \log(1-u))}{1 + \gamma t}\right)^{-1/\gamma}\right] \to 1 - 1 = 0,$$

violating (6). □

Proof of Lemma 6. If $x_Y = +\infty$ there is nothing to show, so assume $x_Y \in \mathbb{R}$. We proceed by contraposition. Suppose $\mathsf{P}(Y = x_Y) > 0$. We will show that (6) fails.

Write $t_n = x_Y - 1/n$. We have $\mathbf{1}\{Y > t_n\} \to \mathbf{1}\{Y \geq x_Y\} = \mathbf{1}\{Y = x_Y\}$. By the conditional dominated convergence theorem (e.g. Kallenberg, 2002, Exercise 8 on page 117), $\mathsf{P}(Y > t_n \mid \mathcal{B}) \to \mathsf{P}(Y = x_Y \mid \mathcal{B})$ as $n \to \infty$ a.s. Since $\mathsf{E}[\mathsf{P}(Y = x_Y \mid \mathcal{B})] = \mathsf{P}(Y = x_Y) > 0$, we have $\mathsf{P}(B) > 0$ where $B = \{\mathsf{P}(Y = x_Y \mid \mathcal{B}) > 0\}$.

Since $B \in \mathcal{B}$, we have

$$\mathsf{E}\left[\mathsf{P}\left(F(x_Y) < 1, Y = x_Y \mid \mathcal{B}\right) \mathbf{1}_B\right] = \mathsf{P}\left(\{F(x_Y) < 1, Y = x_Y\} \cap B\right). \tag{14}$$

We consider two cases, whether $\mathsf{P}\left(\{F(x_Y) < 1, Y = x_Y\} \cap B\right)$ is positive or zero.

1. Case $\mathsf{P}\left(\{F(x_Y) < 1, Y = x_Y\} \cap B\right) > 0$. Eq. (14) implies that, on B, with positive probability, $\mathsf{P}\left(F(x_Y) < 1, Y = x_Y \mid \mathcal{B}\right) > 0$. Let $u \in (0,1]$; note that we assume $u > 0$. On the event $\{F(t) < 1\}$, we have $Z_F^{(t)} = (F(Y) - F(t))/(1 - F(t))$ and thus

$$Z_F^{(t)} \leq u \iff F(Y) \leq (1-u)F(t) + u.$$

If $F(x_Y) < 1$, then $F(x_Y) < (1-u)F(x_Y) + u$, since $u > 0$. Further, $Y \leq x_Y$ almost surely and thus $F(Y) \leq F(x_Y)$ almost surely. This implies that, on the event $\{F(x_Y) < 1\}$, we have $F(Y) < (1-u)F(x_Y) + u$ a.s. Since F is continuous a.s., the inequality $F(Y) \leq (1-u)F(t_n) + u$ is thus eventually true on the event $\{F(x_Y) < 1\}$, a.s. In view of the conditional dominated convergence theorem, we have

$$\begin{aligned} \mathsf{P}\left(Z_F^{(t_n)} \leq u, Y > t_n \mid \mathcal{B}\right) &\geq \mathsf{P}\left(Z_F^{(t_n)} \leq u, F(x_Y) < 1, Y = x_Y \mid \mathcal{B}\right) \\ &= \mathsf{P}\left(F(Y) \leq (1-u)F(t_n) + u,\ F(x_Y) < 1,\ Y = x_Y \mid \mathcal{B}\right) \\ &\to \mathsf{P}\left(F(x_Y) < 1, Y = x_Y \mid \mathcal{B}\right), \qquad n \to \infty, \text{ a.s.} \end{aligned}$$

As explained at the beginning of this case, the limit is positive on B with positive probability. Since the limit does not depend on u, Eq. (6) fails on an event with positive probability for $u > 0$ close to 0.

2. Case $\mathsf{P}(\{F(x_Y) < 1, Y = x_Y\} \cap B) = 0$*.* Eq. (14) then implies

$$\mathsf{E}\left[\mathsf{P}\left(F(x_Y) < 1, Y = x_Y \mid \mathcal{B}\right) \mathbf{1}_B\right] = 0,$$

so that $\mathsf{P}(F(x_Y) < 1, Y = x_Y \mid \mathcal{B}) = 0$ and thus $\mathsf{P}(F(x_Y) = 1, Y = x_Y \mid \mathcal{B}) = 1$ on B. Let $u \in [0, 1)$; note that $u < 1$. By definition, $F_t(x) = 1$ for $x \geq 0$ in case $F(t) = 1$. The inequality $Z_F^{(t)} \leq u$ can thus hold only if $F(t) < 1$. On B, as $\{t_n < Y < x_Y\} \downarrow \varnothing$, we have

$$\begin{aligned}
\mathsf{P}\left(Z_F^{(t_n)} \leq u,\, Y > t_n \mid \mathcal{B}\right) &= \mathsf{P}\left(Z_F^{(t_n)} \leq u,\, F(t_n) < 1,\, Y > t_n \mid \mathcal{B}\right) \\
&= \mathsf{P}\left(F(Y) \leq (1-u)F(t_n) + u,\, F(t_n) < 1,\, Y > t_n \mid \mathcal{B}\right) \\
&= \mathsf{P}\left(F(x_Y) \leq (1-u)F(t_n) + u,\, F(t_n) < 1,\, Y = x_Y \mid \mathcal{B}\right) + o(1) \\
&= \mathsf{P}\left(1 \leq (1-u)F(t_n) + u,\, F(t_n) < 1,\, Y = x_Y \mid \mathcal{B}\right) + o(1) \\
&= o(1),
\end{aligned}$$

as the two requirements on $F(t_n)$ are contradictory. We find that on B, which has positive probability, $\mathsf{P}\left(Z_F^{(t_n)} \leq u,\, Y > t_n \mid \mathcal{B}\right)$ converges to zero for all $u < 1$. But then (6) fails. $\square$

Proof of Proposition 7. We check the convergence relations (5) and (6). The first one is true by assumption. To verify the second one, let $u \in (0, 1)$ [for $u = 1$, condition (6) is trivial] and write $\overline{F}(t) = 1 - F(t)$, $G_{\mathcal{B}}(t) = \mathsf{P}(Y \leq t \mid \mathcal{B})$ and $\overline{G}_{\mathcal{B}}(t) = 1 - G_{\mathcal{B}}(t)$. Eq. (7) implies that $x_F = x_Y$. Here, we use that $\mathsf{P}(Y = x_Y) = 0$ and therefore also $\mathsf{P}(Y = x_Y \mid \mathcal{B}) = 0$ almost surely; see Lemma 6. For all $t \in \mathbb{R}$ such that $\overline{F}(t) > 0$, we have

$$\begin{aligned}
Z_F^{(t)} < u &\iff \frac{F(Y) - F(t)}{\overline{F}(t)} < u \\
&\iff F(Y) < F(t) + u\,\overline{F}(t) \\
&\iff Y < F^{-1}\left(F(t) + u\,\overline{F}(t)\right)
\end{aligned}$$

where $F^{-1}(v) = \inf\{y \in \mathbb{R} : F(y) \geq v\}$ for $v \in (0, 1]$. It follows that, for every $\varepsilon \in (0, \min(u, 1-u))$,

$$\begin{aligned}
\mathsf{P}\big(Z_F^{(t)} \leq u,\, Y > t \mid \mathcal{B}\big) &\leq \mathsf{P}\big(Z_F^{(t)} < u + \varepsilon,\, Y > t \mid \mathcal{B}\big) \\
&= \mathsf{P}\big(Y < F^{-1}\left(F(t) + (u+\varepsilon)\,\overline{F}(t)\right),\, Y > t \mid \mathcal{B}\big) \\
&\leq \mathsf{P}\big(Y \leq F^{-1}\left(F(t) + (u+\varepsilon)\,\overline{F}(t)\right),\, Y > t \mid \mathcal{B}\big) \\
&= G_{\mathcal{B}}\big(F^{-1}\left(F(t) + (u+\varepsilon)\,\overline{F}(t)\right)\big) - G_{\mathcal{B}}(t).
\end{aligned}$$

In the same way,

$$\begin{aligned}
\mathsf{P}\big(Z_F^{(t)} \leq u,\, Y > t \mid \mathcal{B}\big) &\geq \mathsf{P}\big(Z_F^{(t)} < u,\, Y > t \mid \mathcal{B}\big) \\
&= \mathsf{P}\big(Y < F^{-1}\left(F(t) + u\,\overline{F}(t)\right), Y > t \mid \mathcal{B}\big) \\
&\geq \mathsf{P}\big(Y \leq F^{-1}\left(F(t) + (u-\varepsilon)\,\overline{F}(t)\right) - \varepsilon,\, Y > t \mid \mathcal{B}\big) \\
&= G_{\mathcal{B}}\big(F^{-1}\left(F(t) + (u-\varepsilon)\,\overline{F}(t)\right)\big) - G_{\mathcal{B}}(t),
\end{aligned}$$

where we used that F is continuous by assumption and thus F^{-1} is strictly increasing. Writing $y_\pm(u,t) = F^{-1}\big(F(t) + (u \pm \varepsilon)\overline{F}(t)\big)$, we have

$$\begin{aligned}\frac{G_{\mathcal{B}}(y_\pm(u,t)) - G_{\mathcal{B}}(t)}{\overline{G}_{\mathcal{B}}(t)} &= 1 - \frac{\overline{G}_{\mathcal{B}}(y_\pm(u,t))}{\overline{G}_{\mathcal{B}}(t)} \\ &= 1 - \frac{\overline{F}(y_\pm(u,t))}{\overline{F}(t)} \cdot \frac{\overline{G}_{\mathcal{B}}(y_\pm(u,t))/\overline{F}(y_\pm(u,t))}{\overline{G}_{\mathcal{B}}(t)/\overline{F}(t)}.\end{aligned}$$

As $F(t) + (u \pm \varepsilon)\overline{F}(t)$ is less than 1 and converges to 1 as $t \to x_F$, we have $y_\pm(u,t) \to x_F$ as $t \to x_F$. Further, since F is continuous, we have $F(F^{-1}(v)) = v$ for all $v \in (0,1]$, and thus

$$\overline{F}(y_\pm(u,t)) = 1 - \big(F(t) + (u \pm \varepsilon)\overline{F}(t)\big) = (1 - (u \pm \varepsilon))\overline{F}(t).$$

We find that

$$\frac{G_{\mathcal{B}}(y_\pm(u,t)) - G_{\mathcal{B}}(t)}{\overline{G}_{\mathcal{B}}(t)} \to 1 - (1 - (u \pm \varepsilon)) \cdot \frac{1}{1} = u \pm \varepsilon, \qquad \text{almost surely as } t \to x_Y.$$

In view of the bounds earlier in the proof, we obtain

$$u - \varepsilon \le \liminf_{t \to x_Y} \frac{\mathsf{P}\big(Z_F^{(t)} \le u,\, Y > t \mid \mathcal{B}\big)}{\mathsf{P}(Y > t \mid \mathcal{B})} \le \limsup_{t \to x_Y} \frac{\mathsf{P}\big(Z_F^{(t)} \le u,\, Y > t \mid \mathcal{B}\big)}{\mathsf{P}(Y > t \mid \mathcal{B})} \le u + \varepsilon$$

almost surely. Setting $\varepsilon = 1/k$ for sufficiently large integer k and taking the intersection over all such k, we conclude that (6) is true, as required. □

Details for Example 8. Let $t \ge a_+$, $x \ge 0$, and $u \in [0,1]$. Then,

$$1 - F(t) = \mathsf{P}(\xi + a_\tau > t \mid \tau) = e^{-(t - a_\tau)} = \frac{2+\tau}{2} e^{-t}, \tag{15}$$

and thus $F_t(x) = (F(t+x) - F(t))/(1 - F(t)) = 1 - e^{-x}$, independently of τ, since the factors $(2+\tau)/2$ in the numerator and denominator cancel out. Furthermore,

$$\mathsf{P}\big(Z_F^{(t)} \le u,\, Y > t\big) = \mathsf{P}\big(1 - e^{-(Y-t)} \le u,\, Y > t\big) = \mathsf{P}\big(t < Y \le t - \log(1-u)\big) = e^{-t}u.$$

By taking expectations in (15), we obtain $\mathsf{E}[1 - F(y)] = e^{-y}$, since $\mathsf{E}[\tau] = 0$. It follows that F is probabilistically tail calibrated, since $\mathsf{P}\big(Z_F^{(t)} \le u,\, Y > t\big)/\,\mathsf{E}[1 - F(t)] = e^{-t}u/e^{-t} = u$. However, F is not tail auto-calibrated, since $\mathsf{P}(Y > t \mid \tau) = e^{-t}$, so that

$$\frac{\mathsf{P}(Y > t \mid \tau)}{1 - F(t)} = \frac{2}{2+\tau} = \begin{cases} 2/3 & \text{if } \tau = 1, \\ 2 & \text{if } \tau = -1, \end{cases}$$

which does not converge almost surely to 1 as $t \to \infty$. □

Proof of Lemma 9. We show the claim for (4). The arguments for (6) are identical except for the obvious modification of the definition of $R_t(u)$ below.

Suppose (4) holds for $u \in D$, where D is a dense subset of $[0,1]$ that includes 1. Fix $\varepsilon > 0$ and choose points $u_1, \ldots, u_k \in D$ such that $u_1 < \ldots < u_k = 1$ and moreover $u_1 \leq \varepsilon$ and $u_\ell - u_{\ell-1} \leq \varepsilon$ for all $\ell = 2, \ldots, k$. For $u \in [0,1]$, write

$$R_t(u) = \frac{\mathsf{P}\big(Z_F^{(t)} \leq u, Y > t \mid \mathcal{B}\big)}{\mathsf{E}[1 - F(t) \mid \mathcal{B}]}.$$

Note that $R_t(u)$ is monotone in $u \in [0,1]$. For $u \in [0,1]$, let $\ell(u) = \min\{\ell = 1, \ldots, k : u \leq u_\ell\}$. Then $u_{\ell(u)-1} \leq u \leq u_{\ell(u)}$ (with $u_0 = 0$) and thus

$$\begin{aligned} R_t(u) - u &\leq R_t(u_{\ell(u)}) - u_{\ell(u)-1} \leq R_t(u_{\ell(u)}) - u_{\ell(u)} + \varepsilon, \\ R_t(u) - u &\geq R_t(u_{\ell(u)-1}) - u_{\ell(u)} \geq R_t(u_{\ell(u)-1}) - u_{\ell(u)-1} - \varepsilon. \end{aligned}$$

It follows that $\sup_{u\in[0,1]} |R_t(u) - u| \leq \max_{\ell=1,\ldots,k} |R_t(u_\ell) - u_\ell| + \varepsilon$. By assumption, $R_t(u_\ell) \to u_\ell$ almost surely as $t \to x_Y$. As a consequence, almost surely,

$$\limsup_{t\to x_Y} \sup_{u\in[0,1]} |R_t(u) - u| \leq \limsup_{t\to x_Y} \max_{\ell=1,\ldots,k} |R_t(u_\ell) - u_\ell| + \varepsilon = \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we conclude $\limsup_{t\to x_Y} \sup_{u\in[0,1]} |R_t(u) - u| = 0$, almost surely. □

Details for Example 17. Consider a generalization of the unfocused forecaster in Gneiting et al. (2007). Let G be a distribution function supported on a (possibly unbounded) interval $[a,b]$ with continuous density $g > 0$ on (a,b), let $Y \sim G$, let τ be independent of Y and equal to $+1$ or -1 with probability 1/2 and, given τ, let $F(y) = \frac{1}{2}\{G(y) + G(y+\tau)\}$ for all real y. Then $Z_F = F(Y)$ is uniformly distributed on $[0,1]$. Indeed, writing $G_\pm(y) = \frac{1}{2}\{G(y) + G(y \pm 1)\}$ and $z_\pm = G_\pm^{-1}(u)$, we have, for $u \in (0,1)$,

$$\begin{aligned} \frac{d}{du}\mathsf{P}(Z \leq u) &= \frac{d}{du}\frac{1}{2}\{\mathsf{P}(G_+(Y) \leq u) + \mathsf{P}(G_-(Y) \leq u)\} \\ &= \frac{d}{du}\frac{1}{2}\{G(G_+^{-1}(u)) + G(G_-^{-1}(u))\} \\ &= \frac{1}{2}\left[\frac{g(z_+)}{\frac{1}{2}\{g(z_+) + g(z_+ + 1)\}} + \frac{g(z_-)}{\frac{1}{2}\{g(z_-) + g(z_- - 1)\}}\right]. \end{aligned}$$

But since

$$\begin{aligned} u &= G_+(z_+) = \frac{1}{2}(G(z_+) + G(z_+ + 1)), \\ u &= G_-(z_-) = \frac{1}{2}(G(z_-) + G(z_- - 1)) = \frac{1}{2}\{G(z_- - 1) + G((z_- - 1) + 1)\}, \end{aligned}$$

it follows that $z_- - 1 = z_+$, since $z \mapsto G(z) + G(z+1)$, $z \in [a-1, b]$ is strictly increasing, and $z_+, z_- - 1 \in [a-1, b]$. Inserting this equality above yields $(d/du)\,\mathsf{P}(Z \leq u) = 1$, as required.

The special case where $Y \sim \mathrm{Unif}(0,1)$ yields, for $y \in [0,1]$,

$$F(y) = \begin{cases} y/2 & \text{with probability } 1/2 \text{ (if } \tau = -1), \\ (y+1)/2 & \text{with probability } 1/2 \text{ (if } \tau = +1), \end{cases}$$

and thus $Z_F = Y/2$ or $Z_F = (Y+1)/2$, both with probability $1/2$, so that Z_F is indeed uniformly distributed on $(0,1)$, too, and thus F is probabilistically calibrated. In this case, the forecast $F = \frac{1}{2}(G + G(\,\cdot\, + \tau))$ is either the uniform distribution on $[-1,1]$ or the uniform distribution on $[0,2]$. However, F is not probabilistically tail calibrated. To see this, note that for $0 < t < y < 1$, we have

$$F_t(y-t) = \frac{F(y)-F(t)}{1-F(t)} = \begin{cases} \dfrac{y/2 - t/2}{1-t/2} = \dfrac{y-t}{2-t} = \dfrac{1-t}{2-t} \cdot \dfrac{y-t}{1-t} & \text{with probability } 1/2, \\ \dfrac{(y+1)/2 - (t+1)/2}{1-(t+1)/2} = \dfrac{y-t}{1-t} & \text{with probability } 1/2. \end{cases}$$

For $0 < t < 1$, the distribution of $(Y-t)/(1-t) \mid Y > t$ is uniform on $[0,1]$. Recalling from (3) the excess PIT $Z_F^{(t)} = F_t(Y-t)$ given $Y > t$, we get, for $u \in (0,1)$,

$$\lim_{t \uparrow 1} \mathsf{P}\big(Z_F^{(t)} \le u \mid Y > t\big) = \frac{1}{2} + \frac{1}{2}u.$$

As a consequence, F is not probabilistically tail calibrated. □

Example 20 (More general version of Example 19)**.** Let G_1 and G_2 be two random cdfs such that G_1 is stochastically smaller than G_2 almost surely, and let λ be a deterministic cdf. We will specify relevant conditions on G_1, G_2, and λ below, but for a concrete example, suppose that $\Delta \sim \Gamma(1/\gamma, 1/\gamma)$ for some $\gamma > 0$ and define $G_1 = \mathrm{Exp}(\Delta)$, $G_2 = \mathrm{Exp}(\Delta/2)$, that is, $G_1(x) = 1 - e^{-\Delta x}$ for $x \ge 0$. Then $\mathsf{E}[G_1] = \mathrm{GPD}_{1,\gamma}$, that is, $\mathsf{E}[G_1(x)] = 1 - \mathsf{E}[e^{-\Delta x}] = 1 - (1+\gamma x)^{-1/\gamma}$ for $x \ge 0$, and we take $\lambda = \mathrm{GPD}_{1,\gamma/2}$.

The conditional distribution of Y given G_1, G_2 is defined as

$$G(y) = \lambda(y) G_1(y) + (1 - \lambda(y))\, G_2(y), \quad y \in \mathbb{R}, \tag{16}$$

and the random forecast is $F = G_1$. Note that G is automatically increasing due to the stochastic ordering of G_1 and G_2. A stochastic construction yielding the conditional distribution at (16) goes as follows. Let (U, L) be independent random variables, independent also of (G_1, G_2), with U uniform on $(0,1)$ and L with distribution function λ. Put $X_j = G_j^{-1}(U)$ for $j \in \{1,2\}$, with G_j^{-1} the quantile function of G_j. Conditionally on (G_1, G_2), the random variables X_1 and X_2 have distribution functions G_1 and G_2 respectively, and $X_1 \le X_2$ almost surely since $G_1^{-1}(u) \le G_2^{-1}(u)$ for all $u \in (0,1)$, as G_1 is stochastically smaller than G_2 almost surely. Define

$$Y = \text{second largest of } (X_1, X_2, L). \tag{17}$$

For $y \in \mathbb{R}$, we have $Y \le y$ if and only if at least two out of three variables in (X_1, X_2, L) are bounded by y. This yields the following event decomposition:

$$\{Y \le y, X_1 \le X_2\} = \{L \le y, X_1 \le y, X_1 \le X_2\} \cup \{L > y, X_1 \le X_2 \le y\}.$$

Indeed, provided $X_1 \le X_2$, the event $Y \le y$ occurs in two cases: either $L \le y$, and then it is necessary and sufficient that $X_1 \le y$; or $L > y$, and then it is necessary and sufficient that $X_2 \le y$, forcing $X_1 \le y$. Since $\mathsf{P}(X_1 \le X_2 \mid G_1, G_2) = 1$ almost surely, we obtain (16):

$$\begin{aligned}\mathsf{P}(Y \le y \mid G_1, G_2) &= \mathsf{P}(L \le y, X_1 \le y \mid G_1, G_2) + \mathsf{P}(L > y, X_2 \le y \mid G_1, G_2) \\ &= \lambda(y)\, G_1(y) + (1 - \lambda(y))\, G_2(y) = G(y).\end{aligned}$$

The random forecast is $F = G_1$, the conditional distribution of X_1. Intuitively, since the marginal tail of X_1 is heavier than the one of L and since $X_1 \le X_2$, the marginal tail of Y is the same as the one of X_1, so that F should be probabilistically tail calibrated. However, conditionally on (G_1, G_2), the tail of L is heavier than the one of X_1, and F will not be tail auto-calibrated if the conditional tail of the minimum of X_2 and L is not lighter than the one of X_1 almost surely. This intuition underlies the formal arguments in the following paragraph.

The conditional tail of Y given (G_1, G_2) can be computed as follows: we have $Y > y$ if and only if at least two out of three variables in (X_1, X_2, L) exceed y. As a consequence,

$$\begin{aligned}\mathsf{P}(Y > y \mid G_1, G_2) &= \mathsf{P}\left(\{L > y, X_2 > y\} \cup \{L \le y, X_1 > y\} \mid G_1, G_2\right) \\ &= (1 - \lambda(y))\,(1 - G_2(y)) + \lambda(y)\,(1 - G_1(y))\,. \end{aligned} \tag{18}$$

The random forecast $F = G_1$ is not tail auto-calibrated if, with positive probability, the conditional tail of $L \wedge X_2 := \min(L, X_2)$ is at least as heavy as the one as X_1, that is, if

$$\mathsf{P}\left(\limsup_{y\to\infty} \frac{(1 - \lambda(y))(1 - G_2(y))}{1 - G_1(y)} > 0\right) > 0. \tag{19}$$

Indeed, if (19) holds, then, by (18), we have

$$\frac{\mathsf{P}(Y > y \mid G_1, G_2)}{1 - F(y)} = \frac{(1 - \lambda(y))(1 - G_2(y))}{1 - G_1(y)} + \lambda(y),$$

which does *not* converge to 1 almost surely as $y \to \infty$, since already $\lim_{y\to\infty} \lambda(y) = 1$. The concrete example given above satisfies (19).

Next, we show that F is probabilistically tail calibrated provided the marginal tail of G_1 is heavier than the one of λ. We assume

$$\lim_{y\to\infty} \frac{1 - \lambda(y)}{\mathsf{E}[1 - G_1(y)]} = 0. \tag{20}$$

The excess probability integral transform $Z_F^{(t)}$ of $Y - t$ given $Y > t$ by the forecast $F = G_1$ satisfies

$$1 - Z_F^{(t)} = 1 - F_t(Y - t) = \frac{1 - G_1(Y)}{1 - G_1(t)}.$$

For $u \in [0, 1]$, we have by definition of Y in (17), since $X_1 \le X_2$ almost surely,

$$\mathsf{P}(Z_F^{(t)} \le u, Y > t) = \mathsf{P}(Z_F^{(t)} \le u, L \le t < X_1) + \mathsf{P}(Z_F^{(t)} \le u, L > t, X_2 > t).$$

The second term is bounded by $\mathsf{P}(L > t) = 1 - \lambda(t)$, which is of smaller order than $\mathsf{E}[1 - F(t)] = \mathsf{E}[1 - G_1(t)]$ because of the assumption (20). On the event $\{L < t \leq X_1\}$, we have $Y = X_1$ by the definition (17), so that, by independence of L of everything else,

$$\mathsf{P}\left(Z_F^{(t)} \leq u, L \leq t < X_1\right) = \lambda(t)\,\mathsf{P}\left(1 - \frac{1 - G_1(X_1)}{1 - G_1(t)} \leq u, X_1 > t\right) = \lambda(t)\,u\,\mathsf{E}[1 - G_1(t)].$$

The last identity follows from the fact that the conditional distribution of X_1 is G_1 together with the (implicit) assumption that G_1 is continuous almost surely. Together, we deduce that, for $u \in [0, 1]$,

$$\lim_{t\to\infty} \frac{\mathsf{P}\left(Z_F^{(t)} \leq u, Y > t\right)}{\mathsf{E}[1 - F(t)]} = \lim_{t\to\infty} \left(\lambda(t)u + \frac{\mathsf{P}(Z_F^{(t)} \leq u, L > t, X_2 > t)}{\mathsf{E}[1 - G_1(t)]}\right) = u + 0 = u,$$

meaning that F is probabilistically tail calibrated.

B Scoring rules and extremes

A scoring rule is a function of the form

$$S : \mathcal{P} \times \mathbb{R} \to [-\infty, \infty],$$

where $\mathcal{P}$ is a class of probability measures on $\mathbb{R}$. Scoring rules take a probabilistic forecast $F \in \mathcal{P}$ and an observation $y \in \mathbb{R}$ as inputs, and output a real (possibly infinite) value that measures the forecast's accuracy. It is assumed that $\bar{S}(F, G) = \mathsf{E}_G[S(F, Y)]$ exists for all $F, G \in \mathcal{P}$, that is, $S(F, \cdot)$ is assumed to be $\mathcal{P}$-quasiintegrable for all $F \in \mathcal{P}$. A scoring rule S is called *proper* (relative to $\mathcal{P}$) if

$$\bar{S}(G, G) \leq \bar{S}(F, G) \quad \text{for all} \quad F, G \in \mathcal{P}. \tag{21}$$

The scoring rule is strictly proper (relative to $\mathcal{P}$) if equality in (21) implies $F = G$. Hence, proper scoring rules encourage forecasters to issue what they truly believe will occur.

While proper scoring rules have become the standard approach with which to rank and compare competing forecast systems, they have undesirable properties when interest is on extreme events. Taillardat et al. (2023) demonstrate that the continuous ranked probability score (CRPS), arguably the most popular scoring rule for forecasts of real-valued outcomes, cannot distinguish between forecasts with different tail behavior. Brehmer and Strokorb (2019) generalize this result to all proper scoring rules. Specifically, let S be a proper scoring rule relative to a convex class $\mathcal{P}$. Given $G \in \mathcal{P}$ and assuming that there exists a distribution $H \in \mathcal{P}$ with a heavier tail than G, then, for any $\epsilon > 0$, it is possible to construct a distribution $F \in \mathcal{P}$ that is not tail equivalent to G but such that

$$\left|\bar{S}(F, G) - \bar{S}(G, G)\right| \leq \epsilon.$$

That is, the tail of the distribution can be modified while keeping the expected score ϵ-close to its minimum. Recall that the upper endpoint x^F of a distribution with cdf F is defined as

$x^F = \sup\{x \in \mathbb{R} \mid F(x) < 1\}$. For two distributions with cdfs F and G, we say that G has a *heavier tail* than F if $x^F < x^G$, or $x^F = x^G = x^*$ and $\lim_{x \to x^*}(1 - F(x))/(1 - G(x)) = 0$. We denote this by $F <_t G$. The distributions are *tail equivalent* if $x^F = x^G = x^*$ and $\lim_{x \to x^*}(1 - F(x))/(1 - G(x)) \in (0, \infty)$.

That is, given a proper scoring rule S, it is possible to construct a forecast distribution F whose expected score under S is ϵ-close to that of the true distribution G, despite F and G having different tail behavior. However, F's expected score may not be ϵ-close to G's expected score under *all* (or several) proper scoring rules. Hence, this result could potentially be circumvented by evaluating forecasts using multiple scoring rules, and determining whether the score difference is sufficiently close to zero for all scores. The following theorem shows that even if such a class would exist, it is "all or nothing", and would therefore not be interesting for forecast ranking. For the interpretation of the next theorem, it is helpful to keep in mind that for the class of kernel scores, which includes the CRPS, the divergence is symmetric, that is $\bar{S}(H, G) - \bar{S}(G, G) = \bar{S}(G, H) - \bar{S}(H, H)$, $G, H \in \mathcal{P}$ (Gneiting and Raftery, 2007, Section 5.1).

Theorem 21. *Let $\mathcal{S}$ be a collection of quasi-integrable, proper scoring rules on a convex set $\mathcal{P}$ of probability measures on $\mathbb{R}$. Let $G \in \mathcal{P}$ and assume that the set $\{H \in \mathcal{P} : G <_t H\}$ is not empty. If*

$$\inf_{H \in \mathcal{P}: G <_t H} \sup_{S \in \mathcal{S}} \left| \bar{S}(H, G) - \bar{S}(G, G) \right| > 0 \tag{22}$$

then actually

$$\forall H \in \mathcal{P},\ G <_t H : \sup_{S \in \mathcal{S}} \left| \bar{S}(G, H) - \bar{S}(H, H) \right| = +\infty.$$

Proof. Let ϵ be equal to half the infimum in (22). Let $H \in \mathcal{P}$ be such that $G <_t H$. Define $F_\lambda = \lambda H + (1 - \lambda) G$ for $\lambda \in (0, 1)$. For all $\lambda \in (0, 1)$, we have $G <_t F_\lambda$, too. By assumption, there exists $S_\lambda \in \mathcal{S}$ such that $\left| \bar{S}_\lambda(F_\lambda, G) - \bar{S}_\lambda(G, G) \right| \geq \epsilon$. But, by the proof of Lemma 5.3 in Brehmer and Strokorb (2019), we have

$$\left| \bar{S}_\lambda(F_\lambda, G) - \bar{S}_\lambda(G, G) \right| \leq \frac{\lambda}{1 - \lambda} \left| \bar{S}_\lambda(G, H) - \bar{S}_\lambda(H, H) \right|,$$

so that

$$\left| \bar{S}_\lambda(G, H) - \bar{S}_\lambda(H, H) \right| \geq \frac{1 - \lambda}{\lambda} \epsilon.$$

Since this is true for all $\lambda \in (0, 1)$, it follows that

$$\sup_{S \in \mathcal{S}} \left| \bar{S}(G, H) - \bar{S}(H, H) \right| = +\infty. \qquad \square$$

Theorem 21 further substantiates the doubts concerning the adequacy of proper scoring rules for assessing tail behavior of probabilistic forecasts, and strengthens the results of Brehmer and Strokorb (2019). A class of scoring rules would be uninformative when comparing two forecasters that do not have the same tail behavior as the true distribution. However, a key quality of proper scoring rules is that they allow different imperfect forecasts to be ranked and compared. If the assessment of tail behavior with proper scoring rules becomes a binary decision of whether or not the tail behavior is correctly captured, we see no gain in using proper scoring rules for this purpose.

The result of Brehmer and Strokorb (2019) also holds for max-functionals, and similarly, Theorem 21 also extends to this case. A functional $\mathrm{T} : \mathcal{P} \to \mathbb{R}$ is a *max-functional* if it satisfies $\mathrm{T}(\lambda F_1 + (1-\lambda)F_2) = \max(\mathrm{T}(F_1), \mathrm{T}(F_2))$ for all $F_1, F_2 \in \mathcal{P}$ and $\lambda \in (0,1)$, with examples including the upper end point, the tail index, and the extreme value index.

Theorem 22. *Let $\mathcal{S}$ be a collection of quasi-integrable, proper scoring rules on a convex set $\mathcal{P}$ of probability measures on $\mathbb{R}$, and let $T : \mathcal{P} \to \mathbb{R}$ be a max-functional. Let $G \in \mathcal{P}$ and assume that the set $\{H \in \mathcal{P} : T(G) < T(H)\}$ is not empty. If*

$$\inf_{H \in \mathcal{P}: T(G) < T(H)} \sup_{S \in \mathcal{S}} \left| \bar{S}(H, G) - \bar{S}(G, G) \right| > 0$$

then actually

$$\forall H \in \mathcal{P},\ T(G) < T(H) : \sup_{S \in \mathcal{S}} \left| \bar{S}(G, H) - \bar{S}(H, H) \right| = +\infty.$$

Proof. Let $H \in \mathcal{P}$ be such that $T(G) < T(H)$. Define $F_\lambda = \lambda H + (1-\lambda)G$ for $\lambda \in (0,1)$. For all $\lambda \in (0,1)$, we have $T(F_\lambda) = T(H) > T(G)$, since T is a max-functional. The rest of the proof is completely analogous to that of Theorem 21. □

C Marginal tail calibration

C.1 General construction of notions of (tail) calibration

Auto-calibration is a strong requirement and generally hard to assess empirically. There are statistical tests for auto-calibration, but natural diagnostic tools that identify reasons for a lack of auto-calibration are not available. Therefore, one typically resorts to checking and improving necessary unconditional criteria for auto-calibration such as probabilistic and marginal calibration, with the former much more popular than the latter.

Both of these unconditional notions are a special case of the following approach. Choose a class of functions $(F, y) \mapsto h_\lambda(F, y)$ indexed by some parameter λ that take as inputs a cdf F and a real value y. The forecast F is then called calibrated with respect to $(h_\lambda)_\lambda$ if

$$\mathsf{E}[h_\lambda(F, Y)] = 0 \quad \text{for all } \lambda.$$

Probabilistic calibration uses $h_u(F, y) = \mathbb{1}\{F(y) \le u\} - u$, for $u \in [0,1]$, and marginal calibration arises with $h_z(F, y) = F(z) - \mathbb{1}\{y \le z\}$, for $z \in \mathbb{R}$. The following example of Gneiting et al. (2007) shows that marginal calibration is also strictly weaker than auto-calibration, and neither implies nor is implied by probabilistic calibration.

Example 23. Consider the sign-reversed forecaster of Gneiting et al. (2007). Again, assume that $Y \mid \mu \sim N(\mu, 1)$, with $\mu \sim N(0,1)$. The sign-reversed forecaster is $F = N(-\mu, 1)$. Although this is clearly not a good forecast, taking the expectation over μ results in the unconditional distribution of the observation, meaning this forecast is marginally calibrated. However, reassuringly, this forecast is neither probabilistically calibrated nor auto-calibrated.

Alongside any unconditional notion of calibration, we can define $\mathcal{B}$-calibration with respect to $(h_\lambda)_\lambda$ as

$$\mathsf{E}[h_\lambda(F,Y) \mid \mathcal{B}] = 0 \quad \text{almost surely, for all } \lambda, \tag{23}$$

for some σ-algebra $\mathcal{B} \subseteq \mathcal{A}$. If the class of functions $(h_\lambda)_\lambda$ is chosen such that it identifies distributions, that is, if for any deterministic cdf G it holds that

$$(\forall \lambda : \mathsf{E}[h_\lambda(G,Y)] = 0) \quad \text{if and only if} \quad Y \sim G,$$

then auto-calibration implies $\mathcal{B}$-calibration with respect to $(h_\lambda)_\lambda$ for $\mathcal{B} \subseteq \sigma(F)$. In some instances, conditional notions of calibration have also been defined for families of σ-algebras $(\mathcal{B}_\lambda)_\lambda$ indexed by λ, that is, F is calibrated with respect to $(h_\lambda)_\lambda$ and $(\mathcal{B}_\lambda)_\lambda$ if

$$\mathsf{E}[h_\lambda(F,Y) \mid \mathcal{B}_\lambda] = 0 \quad \text{almost surely, for all } \lambda.$$

Examples include threshold-calibration with $h_z(F,y) = F(z) - \mathbb{1}\{y \le z\}$, $\mathcal{B}_z = \sigma(F(z))$, $z \in \mathbb{R}$ and quantile calibration with $h_u(F,y) = \mathbb{1}\{F(y) \le u\} - u$, $\mathcal{B}_u = \sigma(F^{-1}(u))$, $u \in [0,1]$.

Starting from a notion of calibration as in (23), a corresponding notion of tail calibration can be derived. For example, one can request that

$$\lim_{t \to x_Y} \sup_\lambda |\, \mathsf{E}[h_\lambda(F_t, Y-t) \mid Y > t]| = 0,$$

and $\lim_{t \to x_Y} \mathsf{P}(Y > t)/\,\mathsf{E}[1-F(t)] = 1$. In the main paper, we have followed this path starting from probabilistic calibration. In the next section, we consider the corresponding notion of tail calibration that arises from marginal calibration.

C.2 The example of marginal tail calibration

Definition 24 (Marginal tail calibration)**.** Let $x_Y = \sup\{x \in \mathbb{R} : \mathsf{P}(Y \le x) < 1\}$ be the upper endpoint of Y. The forecast F is *marginally tail calibrated* for Y if $\mathsf{E}[1-F(t)] > 0$ for all $t < x_Y$ and it holds that

$$\lim_{t \to x_Y} \sup_{x \ge 0} \left| \frac{\mathsf{P}(t < Y \le t+x)}{\mathsf{E}[1-F(t)]} - \mathsf{E}[F_t(x) \mid Y > t] \right| = 0. \tag{24}$$

This notion of marginal tail calibration could be further refined by defining conditional versions similarly to Definition 4, where this framework has been pursued starting from probabilistic calibration.

As in the definition of $\mathcal{B}$-tail calibration at (4), the single requirement (24) is equivalent to two conditions together that separately concern the occurrence and severity of excesses over the threshold t: For all $u \in [0,1]$,

$$\lim_{t \to x_Y} \sup_{x \ge 0} |\mathsf{P}(Y - t \le x \mid Y > t) - \mathsf{E}[F_t(x) \mid Y > t]| = 0, \quad \lim_{t \to x_Y} \frac{\mathsf{P}(Y > t)}{\mathsf{E}[1-F(t)]} = 1. \tag{25}$$

Indeed, by taking the limit in (24) as $x \to \infty$, the difference $|\,\mathsf{P}(Y > t)/\,\mathsf{E}[1 - F(t)] - 1|$ is bounded by the supremum. This gives the second part of (25); the first part then follows too.

Akin to Proposition 7, for non-random probabilistic forecasts F, it is straightforward to deduce that marginal tail calibration is equivalent to marginal tail equivalence.

Proposition 25. *If the continuous probabilistic forecast F is non-random, then F is marginally tail calibrated if and only if $x_F = x_Y$ and $\lim_{t \to x_Y} \mathsf{P}(Y > t)/(1 - F(t)) = 1$.*

Proof. Showing that the second condition in (25) holds is trivial. For the first condition, write $G(t) = \mathsf{P}(Y \le t)$ and note that

$$G_t(x) = \frac{G(t+x) - G(t)}{1 - G(t)} = 1 - \frac{1 - F(t+x)}{1 - F(t)} \cdot \frac{(1 - G(t+x))/(1 - F(t+x))}{(1 - G(t))/(1 - F(t))}.$$

The second fraction is arbitrarily close to 1 for large t and $x \ge 0$ such that $t + x < x_F = x_G$, and $G_t(x)$ can be made arbitrarily close to $F_t(x) = 1 - (1 - F(t+x))/(1 - F(t))$. $\square$

Remark 26. We know that for non-random F, (standard) marginal and probabilistic calibration are equivalent. Propositions 7 and 25 show that marginal and probabilistic tail calibration are also equivalent in this case, and both correspond to marginal tail equivalence.

We can also derive implications for marginal tail calibration. The misinformed forecaster shows that ordinary marginal calibration does not imply marginal tail calibration.

Example 27 (Marginal calibration does not imply marginal tail calibration)**.** The *misinformed* forecast in Example 5 is marginally calibrated because the unconditional distribution of Y and the expectation of F are both $\mathrm{GPD}_{1,\gamma}$. However, the forecast is not marginally tail calibrated because the first condition in (25) is violated: For any $t > 0$ and $x > 0$, we have on the one hand

$$\mathsf{E}[F_t(x) \mid Y > t] = \mathsf{E}[F_t(x)] = \mathsf{E}[1 - e^{-\Delta_2 x}] = 1 - (1 + \gamma x)^{-1/\gamma},$$

independently of the value of t, while on the other hand

$$\begin{aligned}\mathsf{P}(Y - t \le x \mid Y > t) &= \frac{(1+\gamma t)^{-1/\gamma} - (1 + \gamma(t+x))^{-1/\gamma}}{(1+\gamma t)^{-1/\gamma}} \\ &= 1 - \left(1 + \gamma \frac{x}{1 + \gamma t}\right)^{-1/\gamma} \to 1 - 1 = 0, \qquad t \to \infty.\end{aligned}$$

The tail unfocused forecaster in Example 8 is marginally tail calibrated, which shows that marginal tail calibration (and joint probabilistic and marginal tail calibration) is a strictly weaker requirement than tail auto-calibration. A second example of this type is given by the forecaster in Example 19, see Example 29 below.

Example 28 (Example 8 continued)**.** We consider the same setting as in Example 8 and show that the tail unfocused forecaster F is marginally tail calibrated. The second condition in (25) holds since F is probabilistically tail calibrated. The first condition holds since $F_t(x) = 1 - e^{-x} = \mathsf{P}(Y \le x + t \mid Y > t)$ for $t \ge a_+$ and $x \ge 0$.

Example 29 (Examples 19 and 20 continued)**.** Consider the forecaster $F = G_1$ and outcome Y as defined in Example 20. Under condition (20), F is marginally tail calibrated. Indeed, the second condition in (25) holds since F is probabilistically tail calibrated. Furthermore, for $t \in \mathbb{R}$ and $x \geq 0$, we have, by (18) and the tower property,

$$
\begin{aligned}
&\mathsf{P}(Y > t + x \mid Y > t) - \mathsf{E}[1 - F_t(x) \mid Y > t] \\
&= \frac{1}{\mathsf{P}(Y > t)} \left\{ \mathsf{P}(Y > t + x) - \mathsf{E}\left[\frac{1 - G_1(t + x)}{1 - G_1(t)} \mathbf{1}(Y > t) \right] \right\} \\
&= \frac{\mathsf{E}[1 - G_1(t)]}{\mathsf{P}(Y > t)} \frac{1}{\mathsf{E}[1 - G_1(t)]} \left\{ \mathsf{P}(Y > t + x) - \mathsf{E}\left[\frac{1 - G_1(t + x)}{1 - G_1(t)} \mathsf{P}(Y > t \mid G_1, G_2) \right] \right\} \\
&= \frac{\mathsf{E}[1 - G_1(t)]}{\mathsf{P}(Y > t)} \left\{ \frac{\mathsf{P}(Y > t + x)}{\mathsf{E}[1 - G_1(t + x)]} \frac{\mathsf{E}[1 - G_1(t + x)]}{\mathsf{E}[1 - G_1(t)]} \right. \\
&\qquad\qquad \left. - \frac{1 - \lambda(t)}{\mathsf{E}[1 - G_1(t)]} \mathsf{E}\left[(1 - G_2(t)) \frac{1 - G_1(t + x)}{1 - G_1(t)} \right] - \lambda(t) \frac{\mathsf{E}[1 - G_1(t + x)]}{\mathsf{E}[1 - G_1(t)]} \right\}.
\end{aligned}
$$

The factor before the curly braces tends to one since the second condition in (25) holds; inside the curly braces, the middle term converges to zero as $t \to \infty$ uniformly in $x \geq 0$ by (20), while the first and the third term cancel out as $t \to \infty$ and uniformly in $x \geq 0$. Indeed, the difference between the first and third term is bounded uniformly in $x \geq 0$ by

$$
\begin{aligned}
&\sup_{x \geq 0} \left| \frac{\mathsf{P}(Y > t + x)}{\mathsf{E}[1 - G_1(t + x)]} \cdot \frac{\mathsf{E}[1 - G_1(t + x)]}{\mathsf{E}[1 - G_1(t)]} - \lambda(t) \cdot \frac{\mathsf{E}[1 - G_1(t + x)]}{\mathsf{E}[1 - G_1(t)]} \right| \\
&\quad \leq \sup_{x \geq 0} \left| \frac{\mathsf{P}(Y > t + x)}{\mathsf{E}[1 - G_1(t + x)]} - \lambda(t) \right| \cdot \frac{\mathsf{E}[1 - G_1(t + x)]}{\mathsf{E}[1 - G_1(t)]} \leq \sup_{x \geq 0} \left| \frac{\mathsf{P}(Y > t + x)}{\mathsf{E}[1 - G_1(t + x)]} - \lambda(t) \right| \\
&\quad \leq \sup_{x \geq 0} \left| \frac{\mathsf{P}(Y > t + x)}{\mathsf{E}[1 - G_1(t + x)]} - 1 \right| + |\lambda(t) - 1|,
\end{aligned}
$$

and the right-hand side converges to 0 as $t \to \infty$ by the second condition in (25) and the fact that λ is a cdf.

Proposition 30. *(a) If $\mathsf{P}(Y > t \mid F)/\,\mathsf{P}(Y > t)$ is asymptotically bounded in L^∞ for $t \to x_Y$, tail auto-calibration implies marginal tail calibration. The converse is false.*

(b) Marginal tail calibration does not imply marginal calibration or vice versa.

Proof. Concerning part (a), we have

$$
\begin{aligned}
&\sup_{x \geq 0} |\mathsf{P}(Y - t \leq x \mid Y > t) - \mathsf{E}[F_t(x) \mid Y > t]| \\
&\quad \leq \mathsf{E}\left[\sup_{x \geq 0} \left| \frac{\mathsf{P}(Y - t \leq x, Y > t \mid F)}{\mathsf{P}(Y > t)} - F_t(x) \frac{\mathsf{P}(Y > t \mid F)}{\mathsf{P}(Y > t)} \right| \right] \\
&\quad = \mathsf{E}\left[\frac{\mathsf{P}(Y > t \mid F)}{\mathsf{P}(Y > t)} \sup_{x \geq 0} \left| \frac{\mathsf{P}(Y - t \leq x, Y > t \mid F)}{\mathsf{P}(Y > t \mid F)} - F_t(x) \right| \right] \\
&\quad = \mathsf{E}\left[\frac{\mathsf{P}(Y > t \mid F)}{\mathsf{P}(Y > t)} \sup_{u \in [0,1]} \left| \frac{\mathsf{P}(Z_F^{(t)} \leq u, Y > t \mid F)}{\mathsf{P}(Y > t \mid F)} - u \right| \right].
\end{aligned}
$$

By Lemma 9, the supremum on the right-hand side converges to zero almost surely as $t \to x_Y$. Since it is bounded above by one, it also converges to zero in L^1. By assumption, the first factor is asymptotically bounded in L^∞, which yields the first claim of part (a).

For the converse and for part (b), see Examples 27, 28, and 29. $\square$

We conjecture that marginal and probabilistic tail calibration generally do not imply each other. However, the following result shows that some creativity will be required to construct counter-examples.

Proposition 31 (Sufficient condition for both probabilistic and marginal tail calibration)**.** *Suppose that Assumptions (A1) and (A2) are fulfilled and that* $\lim_{t\to x_Y} \mathsf{P}(Y > t)/\,\mathsf{E}[1 - F(t)] = 1$. *The following conditions are equivalent:*

(a) $\xi = \eta$ *and* $\lim_{t\to x_Y} \sigma_F(t)/\sigma_Y(t) = 1$*;*

(b) F *is probabilistically tail calibrated for* Y*;*

(c) F *is marginally tail calibrated for* Y*.*

Proof. By Lemma 32 below, the convergence statements in (A1) and (A2) are valid uniformly in $x \geq 0$. We show first that (a) and (b) are equivalent, and then that (a) and (c) are equivalent.

Proof that (a) and (b) are equivalent. For $\varepsilon > 0$ and $t < x_Y$, define the event

$$A(t,\varepsilon) = \left\{ \sup_{x\geq 0} \left| 1 - F_t(x) - (1 + \eta x/\sigma_F(t))_+^{-1/\eta} \right| \leq \varepsilon \right\}.$$

By Assumption (A2) and Lemma 32, we have $\lim_{t\to x_F} \mathsf{P}(A(t,\varepsilon) \mid Y > t) = 1$. Since $Z_F^{(t)} = F_t(Y - t)$, we have, on the event $A(t,\varepsilon) \cap \{Y > t\}$, the bound

$$\left| Z_F^{(t)} - \left\{ 1 - (1 + \eta(Y-t)/\sigma_F(t))_+^{-1/\eta} \right\} \right| \leq \varepsilon.$$

We deduce that, on $A(t,\varepsilon) \cap \{Y > t\}$, we have the implications

$$\begin{aligned} 1 - \{1 + \eta(Y-t)/\sigma_F(t)\}_+^{-1/\eta} \leq u - \varepsilon &\implies Z_F^{(t)} \leq u \\ &\implies 1 - \{1 + \eta(Y-t)/\sigma_F(t)\}_+^{-1/\eta} \leq u + \varepsilon. \end{aligned}$$

For $0 < p < 1$ and $v \geq 0$, we have the equivalence

$$1 - (1 + \eta v)_+^{-1/\eta} \leq p \iff v \leq \frac{(1-p)^{-\eta} - 1}{\eta},$$

to be interpreted as $1 - \exp(-v) \leq p \iff v \leq -\log(1-p)$ in case $\eta = 0$. Since $\lim_{t\to x_Y} \mathsf{P}(A(t,\varepsilon)^c \mid Y > t) = 0$, it follows that, for $0 < u < 1$ and $0 < \varepsilon < \min(u, 1-u)$,

$$\begin{aligned} \mathsf{P}\left(Y \leq t + \sigma_F(t) \frac{(1-u+\varepsilon)^{-\eta} - 1}{\eta} \mid Y > t \right) + o(1) &\leq \mathsf{P}\left(Z_F^{(t)} \leq u \mid Y > t \right) + o(1) \\ &\leq \mathsf{P}\left(Y \leq t + \sigma_F(t) \frac{(1-u-\varepsilon)^{-\eta} - 1}{\eta} \mid Y > t \right) + o(1) \end{aligned}$$

as $t \to x_Y$. Because of Assumption (A1) and Lemma 32, we have thus, as $t \to x_Y$,

$$1 - \left(1 + \xi \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u+\varepsilon)^{-\eta} - 1}{\eta}\right)_+^{-1/\xi} + o(1) \le \mathsf{P}\left(Z_F^{(t)} \le u \mid Y > t\right) + o(1)$$
$$\le 1 - \left(1 + \xi \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u-\varepsilon)^{-\eta} - 1}{\eta}\right)_+^{-1/\xi} + o(1).$$

On the one hand, if $\xi = \eta$ and $\lim_{t\to x_Y} \sigma_F(t)/\sigma_Y(t) = 1$, then the left- and right-hand sides converge to $u - \varepsilon$ and $u + \varepsilon$, respectively, and since $\varepsilon \in (0, \min(u, 1-u))$ can be taken arbitrarily small, this implies $\lim_{t\to x_Y} \mathsf{P}(Z_F^{(t)} \le u \mid Y > t) = u$, so that F is probabilistically tail calibrated.

On the other hand, suppose F is probabilistically tail calibrated. Then the term in the middle of the display above converges to $u = 1 - (1 + \xi\frac{(1-u)^{-\xi}-1}{\xi})_+^{-1/\xi}$. By monotonicity, it follows that

$$\limsup_{t\to x_Y} \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u+\varepsilon)^{-\eta} - 1}{\eta} \le \frac{(1-u)^{-\xi} - 1}{\xi} \le \liminf_{t\to x_Y} \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u-\varepsilon)^{-\eta} - 1}{\eta}.$$

Since this is true for all $u \in (0,1)$ and all $\varepsilon \in (0, \min(u, 1-u))$, we can decrease ε to zero to find that, for all $u \in (0,1)$,

$$\limsup_{t\to x_Y} \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u)^{-\eta} - 1}{\eta} \le \frac{(1-u)^{-\xi} - 1}{\xi} \le \liminf_{t\to x_Y} \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u)^{-\eta} - 1}{\eta},$$

and thus, actually

$$\lim_{t\to x_Y} \frac{\sigma_F(t)}{\sigma_Y(t)} \frac{(1-u)^{-\eta} - 1}{\eta} = \frac{(1-u)^{-\xi} - 1}{\xi},$$

which is in turn equivalent to

$$\lim_{t\to x_Y} \frac{\sigma_F(t)}{\sigma_Y(t)} = \frac{\left((1-u)^{-\xi} - 1\right)/\xi}{\left((1-u)^{-\eta} - 1\right)/\eta}.$$

The left-hand side does not depend on $u \in (0,1)$. The only way for the right-hand side to be constant in $u \in (0,1)$, too, is when $\xi = \eta$, and then the ratio is equal to 1.

Proof that (a) and (c) are equivalent. As $t \to x_Y = x_F$, we have both

$$\sup_{x\ge 0} \left| \mathsf{P}(Y > t + x \mid Y > t) - (1 + \xi x/\sigma_Y(t))_+^{-1/\xi} \right|, \to 0$$

and, for every $\varepsilon > 0$,

$$\sup_{x\ge 0} \left| \mathsf{E}[1 - F_t(x) \mid Y > t] - (1 + \eta x/\sigma_F(t))_+^{-1/\eta} \right|$$
$$\le \mathsf{E}\left[\sup_{x\ge 0} \left| 1 - F_t(x) - (1 + \eta x/\sigma_F(t))_+^{-1/\eta} \right| \mid Y > t\right]$$
$$\le \varepsilon + \mathsf{P}\left(\sup_{x\ge 0} \left| 1 - F_t(x) - (1 + \eta x/\sigma_F(t))_+^{-1/\eta} \right| > \varepsilon \mid Y > t\right) \to \varepsilon,$$

since the supremum inside the expectation on the second line is bounded by one. Since $\varepsilon > 0$ can be taken arbitrarily small, we find that marginal tail calibration is equivalent to

$$\lim_{t \to x_Y} \sup_{x \geq 0} \left| (1 + \xi x/\sigma_Y(t))_+^{-1/\xi} - (1 + \eta x/\sigma_F(t))_+^{-1/\eta} \right| = 0.$$

Writing $y = x/\sigma_F(t)$ and taking the supremum over $y \geq 0$ yields the equivalent statement

$$\lim_{t \to x_Y} \sup_{y \geq 0} \left| \left(1 + \xi \frac{\sigma_F(t)}{\sigma_Y(t)} y\right)_+^{-1/\xi} - (1 + \eta y)_+^{-1/\eta} \right| = 0.$$

This is a statement about convergence of distribution functions, which is equivalent to convergence of their quantile functions (since the quantile functions are continuous), i.e., for all $u \in (0, 1)$,

$$\lim_{t \to x_Y} \frac{\sigma_Y(t)}{\sigma_F(t)} \frac{(1 - u)^{-\xi} - 1}{\xi} = \frac{(1 - u)^{-\eta} - 1}{\eta}.$$

The proof of the equivalence with (a) now proceeds as above. $\square$

Lemma 32 (Uniform convergence)**.** *The convergences in Equations* (8) *and* (9) *are necessarily valid uniformly in $x \geq 0$: Equation* (8) *implies*

$$\lim_{t \to x_Y} \sup_{x \geq 0} \left| \mathsf{P}\left(Y > t + \sigma_Y(t)x \mid Y > t\right) - (1 + \xi x)_+^{-1/\xi} \right| = 0 \tag{26}$$

while Equation (9) *implies, for all $\varepsilon > 0$,*

$$\lim_{t \to x_F} \mathsf{P}\left(\sup_{x \geq 0} \left| 1 - F_t(\sigma_F(t)x) - (1 + \eta x)_+^{-1/\eta} \right| > \varepsilon \,\middle|\, Y > t \right) = 0. \tag{27}$$

Proof. Let us prove that (9) implies (27); the proof that (8) implies (26) is similar but easier. Fix integer $n \geq 2/\varepsilon$ and let $0 = x_0 < x_1 < \ldots < x_{n-1} < x_n = \infty$ be such that

$$\forall k \in \{0, \ldots, n\}, \qquad (1 + \eta x_k)_+^{-1/\eta} = 1 - k/n.$$

Apply (A2) to each x_k individually and with ε replaced by $\varepsilon/2$: considering the event

$$A_k(t) = \left\{ \left| 1 - F_t(\sigma_F(t)x_k) - (1 + \eta x_k)_+^{-1/\eta} \right| > \varepsilon/2 \right\},$$

we have

$$\forall k \in \{0, 1, \ldots, n\}, \qquad \lim_{t \to x_F} \mathsf{P}\left(A_k(t) \mid Y > t\right) = 0.$$

(Note that the cases $k = 0$ and $k = n$ are trivial, since F is a continuous distribution function.) By the union bound applied to the event $A(t) = A_0(t) \cup \cdots \cup A_n(t)$, we deduce that

$$\mathsf{P}\left(A(t) \mid Y > t\right) \leq \sum_{k=0}^{n} \mathsf{P}\left(A_k(t) \mid Y > t\right) \to 0, \qquad t \to x_F.$$

For arbitrary $y \geq 0$, let $k(y) = \max\{k = 0, 1, \ldots, n-1 : x_k \leq y\}$, so that $x_{k(y)} \leq y < x_{k(y)+1}$. Then on the event $A(t)$, we have, by monotonicity of a distribution function,

$$1 - F_t(\sigma_F(t)y) \leq 1 - F_t(\sigma_F(t)x_{k(y)}) \leq (1 + \eta x_{k(y)})_+^{-1/\eta} + \frac{\varepsilon}{2} \leq (1 + \eta y)_+^{-1/\eta} + \frac{1}{n} + \frac{\varepsilon}{2},$$

and

$$1 - F_t(\sigma_F(t)y) \geq 1 - F_t(\sigma_F(t)x_{k(y)+1}) \geq (1 + \eta x_{k(y)+1})^{-1/\eta} - \frac{\varepsilon}{2} \geq (1 + \eta y)_+^{-1/\eta} - \frac{1}{n} - \frac{\varepsilon}{2}.$$

The two inequalities imply

$$\left|1 - F_t(\sigma_F(t)y) - (1 + \eta y)_+^{-1/\eta}\right| \leq \frac{1}{n} + \frac{\varepsilon}{2} \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

On the event $A(t)$, the previous inequality is true for every $y \geq 0$. We conclude that the uniform convergence statement (27) holds. $\square$

D Normal distribution simulation study

In the following, we consider forecasts that are well-known in the forecast verification literature, and evaluate them with respect to their probabilistic tail calibration. Suppose observations are drawn from $Y \mid \mu \sim N(\mu, 1)$, where $\mu \sim N(0, 1)$. The unconditional distribution of Y is $N(0, 2)$, and μ represents a source of information that may or may not be available to the forecasters.

Following Gneiting et al. (2007), consider four different forecasts for Y: the *ideal* forecaster, $F_{\text{id}} = N(\mu, 1)$; the *climatological* forecaster, $F_{\text{cl}} = N(0, 2)$; the *unfocused* forecaster, $F_{\text{un}} = \frac{1}{2}N(\mu, 1) + \frac{1}{2}N(\mu + \tau, 1)$, where τ is -1 or 1 with equal probability, independently of μ; and the *sign-reversed* forecaster, $F_{\text{sr}} = N(-\mu, 1)$. The unfocused and sign-reversed forecasters were introduced in Examples 2 and 23, respectively.

We draw $n = 10^6$ independent realizations of μ and τ to obtain the forecasts. Conditionally on μ, the outcome Y is then drawn for each instance. Figure 8 displays the probabilistic tail calibration of the four forecasters for five thresholds. The climatological and ideal forecasters are auto-calibrated, and are therefore probabilistically calibrated and probabilistically tail calibrated. The unfocused forecaster is probabilistically calibrated but not probabilistically tail calibrated, while the sign-reversed forecaster is neither probabilistically nor probabilistically tail calibrated (though it is marginally calibrated). Similarly to in Example 18, we could assess conditional notions of calibration, as described in Section 4.2, allowing us to confirm that the ideal forecaster is $\sigma(\mu)$-tail calibrated, whereas the climatological forecaster is not.

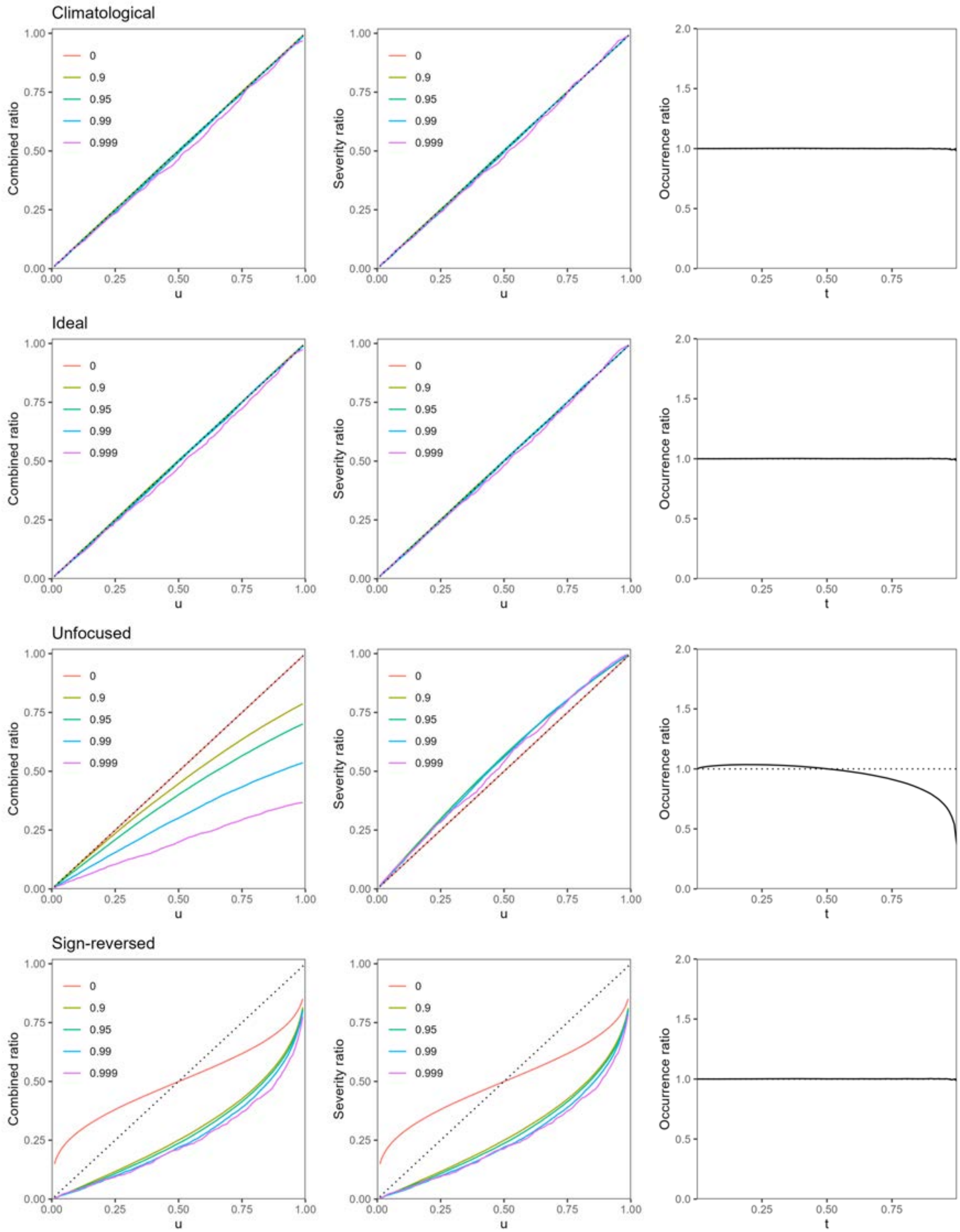

Figure 8: Probabilistic tail calibration diagnostic plots for the climatological, ideal, unfocused, and sign-reversed forecasters in Section D. Results are shown for five thresholds, expressed in terms of quantiles α of the 10^6 observations.

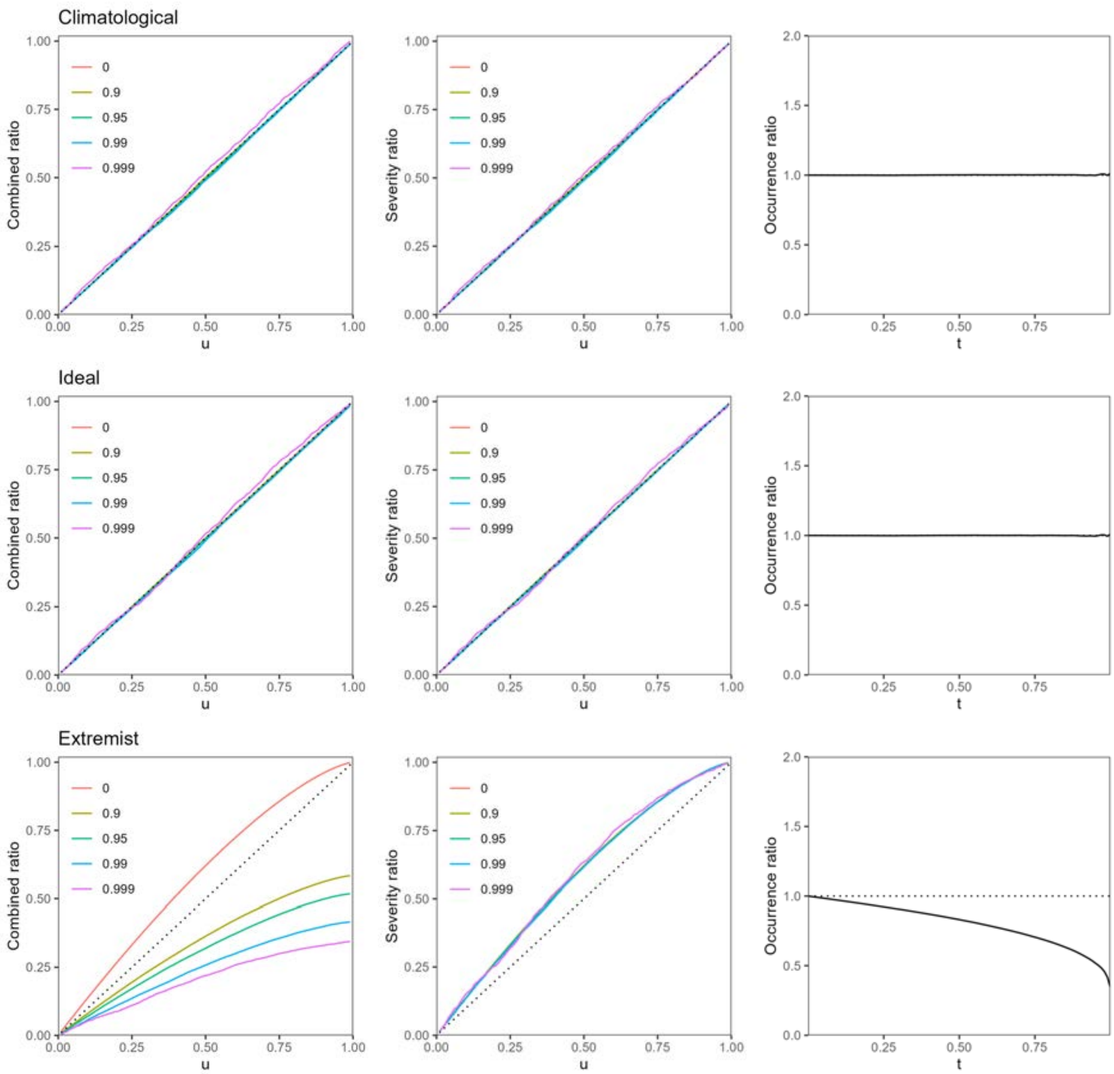

Figure 9: Probabilistic tail calibration plots for the climatological, ideal, and extremist forecasters in Example 18. Results are shown for five thresholds, expressed in terms of quantiles α of the 10^6 observations.