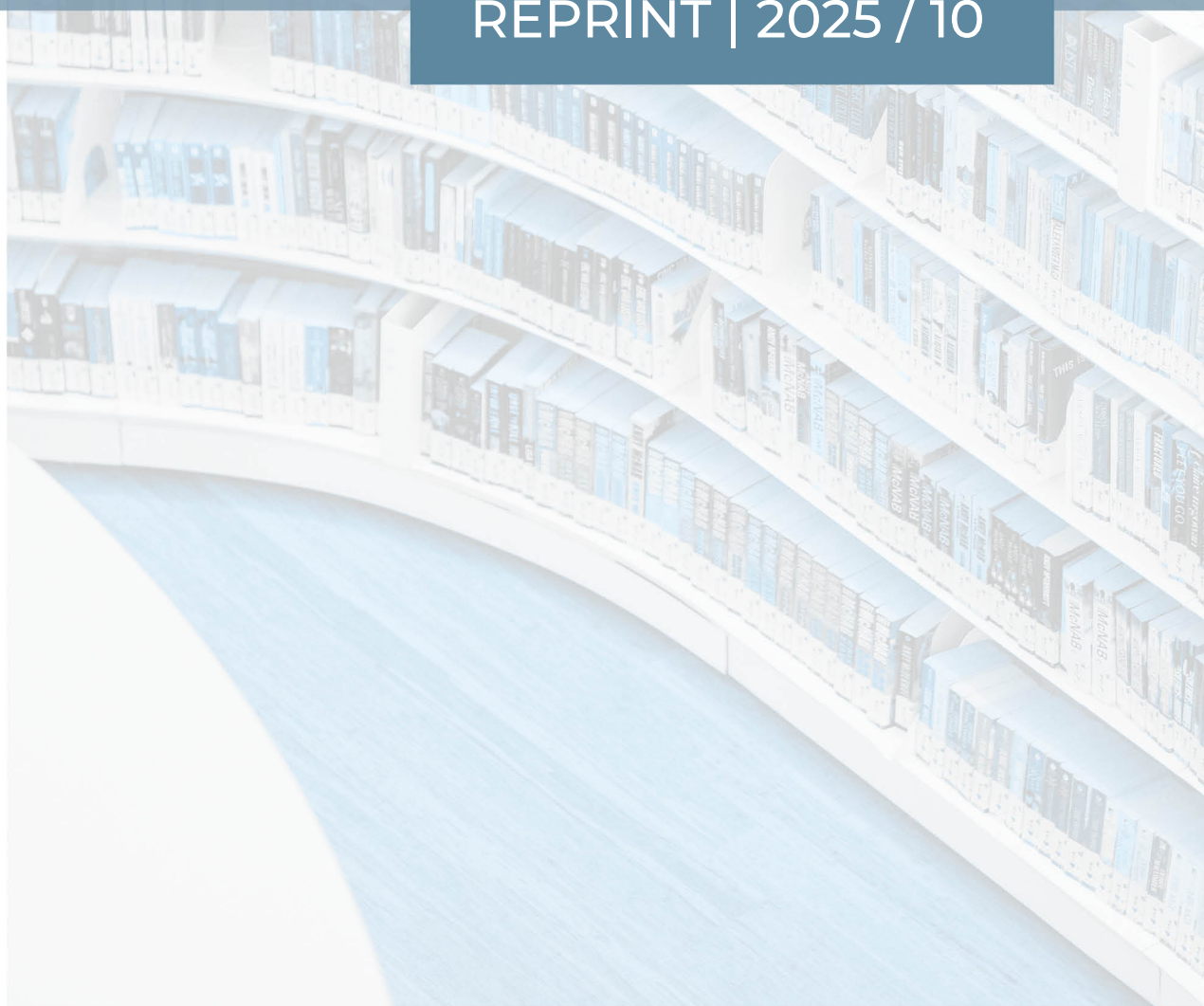
TAIL CALIBRATION OF PROBABILISTIC FORECASTS

Sam Allen, Jonathan Koh, Johan Segers, Johanna Ziegel

REPRINT | 2025 / 10



UCLouvain

ISBA | LIDAM
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

Tail calibration of probabilistic forecasts

Sam Allen[a,*] 0000-0003-1971-8277, Jonathan Koh[b] 0000-0002-7287-0588, Johan Segers[c] 0000-0002-0444-689X, Johanna Ziegel[a] 0000-0002-5916-9746

[a]Seminar for Statistics, ETH Zurich

[b]IMSV, University of Bern

[c]Dept. of Mathematics, KU Leuven, and LIDAM/ISBA, UCLouvain

*Corresponding author: sam.allen@stat.math.ethz.ch

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

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 in the supplement. 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 some fields, *forecast calibration* refers to re-calibration methods that adapt forecasts so that they are calibrated. This paper does not concern re-calibration, though it facilitates the introduction of *tail re-calibration* methods. Tail re-calibration methods currently do not exist, largely due to a lack of tools to evaluate the resulting forecasts. Our definitions of tail calibration remedy this by allowing forecasters to assess whether their forecasts for extreme

outcomes are reliable, and if not, to understand how the forecasts are miscalibrated, for example, if the tails of the predictive distribution are too light or too heavy.

In Section 2, we revisit notions of calibration for probabilistic forecasts, and introduce 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 appendices in the supplement contain sections about the following topics: additional proofs and calculations; a result about the evaluation of tail behavior using multiple scoring rules; a definition of, and results on, marginal tail calibration; additional simulation results evaluating well-known forecasts in the literature with respect to their probabilistic tail calibration; and explanations on the construction of confidence intervals in our diagnostic tools. An R package and R code to reproduce the results herein are available at github.com/sallen12/TailCalibration and github.com/sallen12/TailCalibration_rep respectively.

2 Tail calibration

Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space supporting a random pair (F, Y), 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, for all $y \in \mathbb{R}$, $F(y) = \mathsf{P}(Y \leq y \mid F)$ 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

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

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, $Z_F \sim \text{Unif}(0,1)$, 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. Tsyplakov (2014, Definition 3) refers to (1) as conditional probabilistic calibration, and for $\sigma(F) \subseteq \mathcal{B}$, $\mathcal{B}$-calibration is equivalent to the forecast F being ideal with respect to $\mathcal{B}$ (Gneiting and Ranjan, 2013).

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. In this case, 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 (Gneiting and Ranjan, 2013, Definition 2.5). 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 $\mathsf{E}[F(y)] = \mathsf{P}(Y \le y)$ for all $y \in \mathbb{R}$, and *probabilistically calibrated* if $\mathsf{P}(Z_F \le u) = u$ for all $u \in [0,1]$, i.e. if $Z_F \sim \mathrm{Unif}(0,1)$. 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 corresponds to (1) when $\mathcal{B}$ is the trivial σ-algebra, $\{\varnothing, \Omega\}$, and implies that all central prediction intervals of F have the correct coverage.

Marginal and probabilistic calibration are unconditional notions of calibration, whereas auto-calibration conditions on the forecast distribution (Gneiting and Resin, 2023). Intermediate notions of forecast calibration exist, including quantile and threshold calibration, which condition on functionals of the forecast distribution; see Appendix C in the supplement. Gneiting et al. (2007) show that marginal and probabilistic calibration are neither necessary nor sufficient conditions for each other. However, if a forecast is auto-calibrated, then it is both marginally and probabilistically calibrated, whereas the converse is not true. More generally, for $\mathcal{B} \subseteq \sigma(F)$, $\mathcal{B}$-calibration is generally weaker than auto-calibration, whereas it is generally stronger if $\sigma(F) \subseteq \mathcal{B}$.

Example 2 (Unfocused forecaster).

To see that probabilistic calibration is strictly weaker than auto-calibration when F is not deterministic, consider the *unfocused* forecaster of Gneiting et al. (2007, Example 3). Suppose $Y \sim N(\mu, 1)$ follows a normal distribution with mean μ and variance 1 conditionally on μ, where $\mu \sim N(0,1)$, and let $F = (1/2)N(\mu, 1) + (1/2)N(\mu + \tau, 1)$ 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 calibration 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 $x \ge 0$ and

large *t*, i.e., $\mathsf{P}(Y > t)$ positive but close to zero. Then, we separately consider calibration with respect to both factors on the right-hand side. Under auto-calibration,

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

with the *conditional forecast excess distribution* F_t defined as

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

if $F(t) < 1$, and $F_t(x) = 1$, otherwise. To evaluate the calibration of the forecast's tails, we propose checking two things for large *t*: first, whether the forecast survival function is close to the conditional survival function of the outcome variable, that is, whether $1-F(t)$ is close to $\mathsf{P}(Y > t \mid F)$; and second, whether the conditional forecast excess distribution is close to the conditional excess distribution of *Y*, that is, whether $1-F_t(x)$ is close to $\mathsf{P}(Y > x+t \mid F) / \mathsf{P}(Y > t \mid F)$ for all $x \geq 0$. We will call these two aspects the *occurrence* and the *severity* of excesses, respectively.

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 that is lighter than exponential, then both probabilities tend to 1 almost surely. In the two 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 this scaling issue , we define the *excess probability integral transform (PIT)* as

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

and propose the following definitions 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 $\mathsf{P}(Y > x_Y \mid \mathcal{B}) = 0$.

(a) 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

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

(b) The forecast F is *tail auto-calibrated* for Y if it is tail $\sigma(F)$-calibrated, that is, $\mathbf{1} - F(t) > 0$ almost surely for all $t < x_Y$, and

$$\frac{\mathsf{P}\left(Z_F^{(t)} \le u, Y > t \mid F\right)}{1-F(t)} \to u, \qquad \text{almost surely as } t \to x_Y.$$

(c) The forecast F is *probabilistically tail calibrated* for Y if it is tail $\{\varnothing, \Omega\}$-calibrated, that is, $\mathsf{E}[1-F(t)] > 0$ for all $t < x_Y$, and

$$\frac{\mathsf{P}\left(Z_F^{(t)} \le u, Y > t\right)}{\mathsf{E}[1-F(t)]} \to u, \qquad \text{as } t \to x_Y.$$

Remark 5 (Marginal 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.

The definition of $\mathcal{B}$-calibration at (1) motivates the definition of tail calibration with respect to an arbitrary information set $\mathcal{B}$. The σ-algebra $\mathcal{B}$ encodes the information with respect to which we would like to assess the forecaster's calibration; tail auto-calibration and probabilistic tail calibration emerge as special cases. Probabilistic tail calibration is similar in essence to the standard notion of probabilistic calibration, except that it relies on the excess PIT associated with the conditional forecast excess distribution and outcomes above a large threshold. Allen et al. (2023) and Mitchell and Weale (2023) consider a similar framework to assess forecast calibration when interest is in exceedances of a fixed threshold. By taking the limit as t tends to the upper endpoint, we provide a theoretically attractive framework with which to evaluate forecasts made for extreme events, addressing the lack of existing methods to achieve this.

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

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

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

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

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

That (5) and (6) imply (4) is easy to see. For the converse, set $u = 1$ in (4) to conclude (5), and then combine (4) with (5) to obtain (6). Condition (5) concerns the *occurrence* of excesses over a high threshold: Conditionally on $\mathcal{B}$, we would like the exceedance probabilities over *t* 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. Condition (6) is related to the characterization of $\mathcal{B}$-calibration in (1). Asymptotically, under tail $\mathcal{B}$-calibration, the conditional probability $\mathsf{P}\left(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\right)$ factorizes as $\mathsf{P}(Y > t \mid \mathcal{B}) \cdot u$, and hence the excess PIT is (approximately) uniformly distributed and independent of $\mathcal{B}$; when $\mathcal{B} = \{\varnothing, \Omega\}$, condition (6) can be simplified to $\mathsf{P}\left(Z_F^{(t)} \le u \mid Y > t\right) \to u$ almost surely as $t \to x_Y$. The single condition at (4) 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 $(1 - F(t)) \cdot u$.

When assessing the quality of forecasts in the tail, it is not sufficient to focus only on condition (6), the *severity ratio*, and forget about condition (5), 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 the occurrence ratio (5) without the severity ratio (6) is also insufficient, as the following example shows.

Example 6 (Misinformed forecaster).

For $\gamma > 0$, let Δ_1 and Δ_2 be two independent $\Gamma(1/\gamma, 1/\gamma)$ random variables, with $\Gamma(a, b)$ the gamma distribution with shape $a > 0$ and scale $b > 0$. 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 $x \ge 0$, and consider the *misinformed forecaster* $F = \mathrm{Exp}(\Delta_2)$, 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, since the excess PIT is calculated with respect to the conditional excess distribution $F_t = \mathrm{Exp}(\Delta_2)$, while the conditional distribution of $Y - t$ is $\mathrm{Exp}(\Delta_1)$. 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 (under-)estimated by $1/\Delta_2$ 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.

For non-random probabilistic forecasts *F*, the situation simplifies considerably, as it is sufficient to evaluate the occurrence ratio in order to assess 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

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

The proof of Proposition 7 is given in Appendix A in the supplement. It makes use of the following lemma on the upper endpoint of the outcome *Y*, which is also useful for further results. Its proof is also given in Appendix A.

Lemma 8 .

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 $\mathsf{P}(Y = x_Y) = 0$.

Remark 9 (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.

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, which is probabilistically tail calibration but not tail auto-calibrated; see also Example 18 in Section 4.

Example 10 (Tail unfocused forecaster).

Let ξ, τ be independent random variables, $\xi \sim \text{Exp}(1)$ unit-exponential and $\mathsf{P}(\tau = +1) = \mathsf{P}(\tau = -1) = 1/2$. Define $a_\tau = \log\big((2+\tau)/2\big)$. 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.

In concert with the following proposition, it follows that tail auto-calibration is a strictly stronger requirement than probabilistic tail calibration.

Proposition 11 .

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

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

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

Corollary 12 .

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 11 with $\mathcal{B} = \sigma(F)$ and $\mathcal{C} = \{\varnothing, \Omega\}$. For the converse, see Examples 10 and 18. ▪

Proposition 11 shows that the larger the conditioning set $\mathcal{B}$, 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 rank forecasts for extreme events by the size of the σ-algebra $\mathcal{B}$ with respect to which they are tail calibrated. Strähl and Ziegel (2017) formalize a similar concept for classical notions of calibration: a forecast is *cross-calibrated* if it is calibrated with respect to the information of all other relevant forecasters. This approach could equivalently be employed using notions of tail calibration, see Section 4.2.

We can also identify relationships between standard notions of calibration and their tail counterparts. These relationships are summarized in Figure 2.

Proposition 13 .

(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

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

since the conditional distribution of Y given $\mathcal{B}$ is F, and F is assumed to be continuous. This implies $\mathsf{E}[1\{Z_F^{(t)} \le u, Y > t\} - u(1-F(t)) \mid \mathcal{B}] = 0$ and thus $\mathsf{P}\left(Z_F^{(t)} \le u, Y > t \mid \mathcal{B}\right) = u\mathsf{E}[1-F(t) \mid \mathcal{B}]$, showing the sufficiency in part (a). For the converse, see Example 15.

Part (b) is shown by Examples 15 and 16. ▪

While auto-calibration implies tail auto-calibration, probabilistic calibration does not imply probabilistic tail calibration. Marginal calibration also does not imply its tail counterpart, see Appendix C in the supplement. This highlights a shortcoming of classical definitions of calibration when interest is in the tails—in this sense, we extend the negative result of Brehmer and Strokorb (2019) 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.

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. In this section, 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 nondegenerate distribution.

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

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

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 $\mathrm{GPD}_{\sigma,\xi}(x) = \mathrm{GPD}_{1,\xi}(x/\sigma)$, $x \ge 0$. The GPD is the basis of the peaks-over-threshold method

in Davison and Smith (1990) used to approximate the tails of distributions satisfying 1. 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

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

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 14 (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 14, see the more general Proposition C.9 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 GPD 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 $F = \mathrm{Exp}(\Delta)$, 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 $\mathrm{GPD}_{1,\gamma}$, so that (8) holds with $\xi = \gamma$, while F_t is $\mathrm{Exp}(\Delta)$ 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 14 are sufficient but not necessary conditions for probabilistic tail calibration. Indeed, according to Proposition 7, for non-random probabilistic forecasts, it suffices that $\mathsf{P}(Y > t) / (1 - F(t)) \to 1$ 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

The convergence relations (4) and (6) hold uniformly in $u \in (0,1]$; a precise statement is given in Lemma A.1 in Appendix A in the supplement. Hence, for F to be tail $\mathcal{B}$-calibrated, the graph of

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

should be close to the diagonal in the supremum distance. This motivates the following diagnostic plots.

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 using a finite sample, consider, for some large threshold t, the n_t excess 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 $Z_F^{(t)}$.

We can then calculate the combined ratio

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

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. Results in Allen et al. (2023) suggest that this diagnostic plot is also valid for non-extreme choices of t, and we recover a *pp*-plot of the PIT values $z_i = F_i(y_i)$ when $t = -\infty$, 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 either be 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

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

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 e.g. Williams et al., 2014; Taillardat et al., 2019; Allen et al., 2023). Assessing these forecasts using reliability diagrams is closely related to our assessment of the occurrence ratio using (11).

The second condition, (6), considers the conditional excess distribution given that the threshold has been exceeded, allowing us to evaluate forecasts for the severity of an extreme event. Under probabilistic tail calibration, the excess 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).

While (10) and (11) are defined in terms of a fixed threshold t, the threshold could also be chosen adaptively. For example, we may wish to assess forecasts made at different points in time or space, for which the notion of an extreme event varies. In this case, the diagnostic tools described above could be employed with a threshold that depends on the forecast index i, corresponding to high conditional quantiles of the outcome, for example. Choosing the threshold t in practice is often a compromise between t being large enough to represent the tail regime of the data, whilst also permitting sufficient data from which robust results can be obtained. Formal tests for tail calibration could be derived, for example based on the uniformity of the excess PIT values (e.g. Mitchell and Weale, 2023), though these are not considered in depth here; this is discussed further in Section 5.

Example 15 (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 $F(y) = \mathsf{P}(Y \leq y)$ for all $y \geq t_0$. Trivially, F is tail auto-calibrated, and hence probabilistically tail calibrated. However, $F(y)$ and $\mathsf{P}(Y \leq y)$ 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 $x \geq 5$. Figure 3 shows the diagnostic plots for probabilistic tail calibration in this case.

Example 16 (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 4 illustrates the example for $G = \text{Unif}(0,1)$: the curves for $t = -1$ confirm probabilistic calibration but those for $t > -1$ indicate the lack of probabilistic tail calibration.

Example 17 (Ideal, climatological and extremist forecasters).

The following example was introduced by Taillardat et al. (2023) for studying methods to evaluate forecast tails. Let $\Delta \sim \Gamma(1/\gamma, 1/\gamma)$, 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 tail regime than the outcome (heavy-tailed rather than light-tailed), whereas the extremist forecaster belongs to the correct tail regime (light-tailed).

Figure 5 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 D.2 in the supplement. 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 assess calibration with respect to larger information sets. The same is true when interest is in extremes. For example, the climatological forecaster in Example 17 is probabilistically tail calibrated because it exhibits the same tail behavior as the unconditional distribution of the outcome. However, it does not exhibit the correct tail behavior conditional on Δ, and is therefore not $\sigma(\Delta)$-tail calibrated. In this section, we propose tools to assess tail $\mathcal{B}$-calibration for non-trivial σ-algebras $\mathcal{B}$.

To assess tail $\mathcal{B}$-calibration, the forecast-observation pairs $(F_1, y_1), \ldots, (F_n, y_n)$ can be stratified into $J \in \mathbb{N}$ mutually disjoint bins, $B_1, \ldots, B_J$, depending on variables representing

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 per bin,

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

where $\mathcal{J}_j = \left\{ i \in \{1,\ldots,n\} : (F_i, y_i) \in B_j \right\}$ 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 excess 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) miscalibration 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 17, 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

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

for the ideal, climatological and extremist forecasters in Example 17 are shown as a function of the threshold in Figure 6, 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), though an analogous approach can serve to evaluate the occurrence ratio (5) or severity ratio (6).

While the climatological forecaster in Example 17 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 18 (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 $X \sim \text{Exp}(\Delta)$. Let the random variable $L \sim \text{GPD}_{1,\gamma/2}$ 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 $F = \text{Exp}(\Delta)$, the conditional distribution of X. This forecast is probabilistically tail calibrated, but not tail auto-calibrated; see Figure 7 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 $\text{GPD}(1, \gamma)$, which is heavier than the marginal distribution 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 is *lighter* than that of L, so that F is optimistic, i.e., it predicts that less severe outcomes will occur than reality, violating tail auto-calibration.

5 Application

Reliable forecasts for extreme precipitation are highly valuable for mitigating the impacts 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 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 assumes that precipitation follows a certain parametric distribution, with location and scale parameters that are affine functions of the IFS ensemble mean and standard deviation, respectively. This simple approach is generally referred to as ensemble model output statistics in the post-processing literature (Gneiting et al., 2005). We implement this post-processing model with two choices of parametric distribution: a logistic distribution, which is commonly employed in studies of precipitation forecasts (see e.g. Messner et al., 2014); and a generalized extreme value (GEV) distribution, which has been proposed as a means to better forecast extreme outcomes (Lerch and Thorarinsdottir, 2013; Scheuerer, 2014). Both distributions are censored below at zero. The parameters of the post-processing models 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 8 displays the probabilistic (tail) calibration diagnostic plots for the four 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 15 mm, roughly corresponding to the 97^{th}, 99^{th}, and 99.5^{th} percentiles of the previously observed precipitation measurements (across all stations). Similar conclusions are drawn when separate thresholds are employed at each station, corresponding to high quantiles of the local climatology, see Appendix E in the supplement. The raw IFS ensemble forecasts are under-dispersed and positively biased, whereas both post-processing methods yield considerably more reliable forecasts. The IFS forecasts are additionally strongly miscalibrated in the tails. 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 the censored logistic post-processing method improves upon the calibration of the raw ensembles, the resulting forecasts are still acutely miscalibrated when predicting extreme outcomes. Post-processing with a censored GEV distribution, on the other hand, yields forecasts that are both probabilistically calibrated and tail calibrated. Despite exhibiting contrasting tail behavior, the two distributions yield forecasts with very similar average CRPS; the logistic forecasts are roughly 1% more accurate than the GEV forecasts.

The poor tail calibration of the logistic post-processed forecasts is a deficiency of the chosen distribution. This is irrespective of the number of ensemble members: Even if we had no covariate information, we could still issue the climatological distribution of the outcome, which is auto-calibrated and therefore tail auto-calibrated. However, when employing existing forecast evaluation techniques, practitioners would generally conclude that this post-processing method generates reliable forecasts, which is misleading. In contrast, our diagnostic tools for tail calibration inform us that the forecast tails are too light, resulting in unreliable forecasts for extreme events. This information can be used to improve forecasts, for example by encouraging practitioners to experiment with alternative post-processing methods, based on different distributional assumptions.

It is not trivial to derive a unified test for tail calibration based on the combined ratio at (10), and this becomes yet more difficult considering the limit as $t \to x_Y$. When assessing calibration with respect to a finite threshold, Mitchell and Weale (2023) propose separately testing whether the predicted threshold exceedance is correct, and whether the excess PIT values resemble a sample from a standard uniform distribution. Here, we implement a binomial test for the occurrence ratio, and a Kolmogorov–Smirnov test for uniformity of the excess PIT values, assuming the forecast-observation pairs are independent and identically distributed. For thresholds 10 mm and 15 mm, all methods other than the GEV post-processing approach return a p-value of zero (up to numerical precision) on both tests, providing evidence against these forecasts being tail calibrated. The p-values for the GEV forecasts are 0.00, 0.01, and 0.83 when testing the severity ratio using a Kolmogorov–Smirnov test, and 0.59, 0.13, and 0.04 when testing the occurrence ratio using a binomial test, at thresholds 5, 10, and 15 mm, respectively. Both a nonparametric bootstrap that resamples forecast-observation pairs or the central limit theorem in combination with the delta method can be used to obtain pointwise confidence intervals for the occurrence, severity, and combined ratios; the latter method is discussed in detail in Appendix F in the Supplement.

6 Discussion

Brehmer and Strokorb (2019) prove 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 probabilistic 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. This is illustrated in a case study on European precipitation forecasts, whereby statistical post-processing methods are found to produce calibrated forecasts 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 we do not consider forecast re-calibration methods in this paper, tail re-calibration methods could be designed that modify forecast distributions so that they exhibit calibrated tails. For example, conformal prediction allows us to generate probabilistic forecasts with calibration guarantees (see e.g. Vovk et al., 2022), and there may be potential for such methods to be applied to extreme observations to construct forecasts with tail calibration guarantees.

Tail calibration can be assessed with respect to different information sets, though 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 heavy- or light-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 et al. (2023), who propose a sequential procedure for testing probabilistic calibration based on *e*-processes, recognizing that forecasting is often an inherently sequential task.

Tail calibration may partially address the open question how best to compare probabilistic forecasts for extreme 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. Tail calibration with respect to a larger information set corresponds to a more informed forecast for extremes, which aligns with the idea that forecasts should be as sharp, or informative, as possible, subject to being calibrated (Gneiting et al., 2007). Zamo et al. (2021) similarly suggest using this calibration-sharpness principle to select the most informative forecast among a set of reliable candidates when aggregating probabilistic forecasts, and such an approach could also be applied when forecasting extreme events. However, for miscalibrated forecasts and non-nested maximal information sets, this framework still does not allow for comparison between different forecasts.

When the unconditional distribution of the outcome variable is in the max-domain of attraction of a non-degenerate distribution (Assumption (A1)), probabilistic tail calibration is

equivalent to correctly 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 tail calibration may be obtained by leveraging results in multivariate extreme value theory.

Acknowledgments

We would like to thank two anonymous referees for their constructive feedback. 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.

Conflicts of interest

The authors have no conflicts of interests to declare.

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*, 1315–1335.

Arnold, S., A. Henzi, and J. F. Ziegel (2023). Sequentially valid tests for forecast calibration. *Annals of Applied Statistics 17*, 1909–1935.

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

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.

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. and T. L. Thorarinsdottir (2013). Comparison of non-homogeneous regression models for probabilistic wind speed forecasting. *Tellus A: Dynamic Meteorology and Oceanography 65*, 21206.

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 40*, 1701–1720.

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.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society 140*, 1086–1096.

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*, 1448–1459.

Taillardat, M., A.-L. Fougères, P. Naveau, and O. Mestre (2019). Forest-based and semiparametric 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/.

Vovk, V., A. Gammerman, and G. Shafer (2022). *Algorithmic learning in a random world* (second ed.), Volume 29. Springer.

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.

Zamo, M., L. Bel, and O. Mestre (2021). Sequential aggregation of probabilistic forecasts — Application to wind speed ensemble forecasts. *Journal of the Royal Statistical Society Series C: Applied Statistics 70*, 202–225.

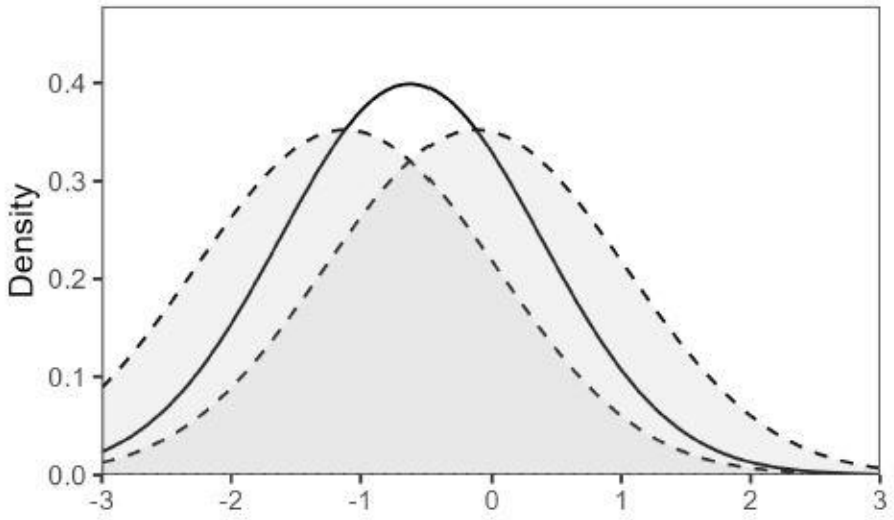


Figure 1: The unfocused forecaster in Example 2 issues each of the dashed densities with probability one half, while the true outcome distribution is the solid density.

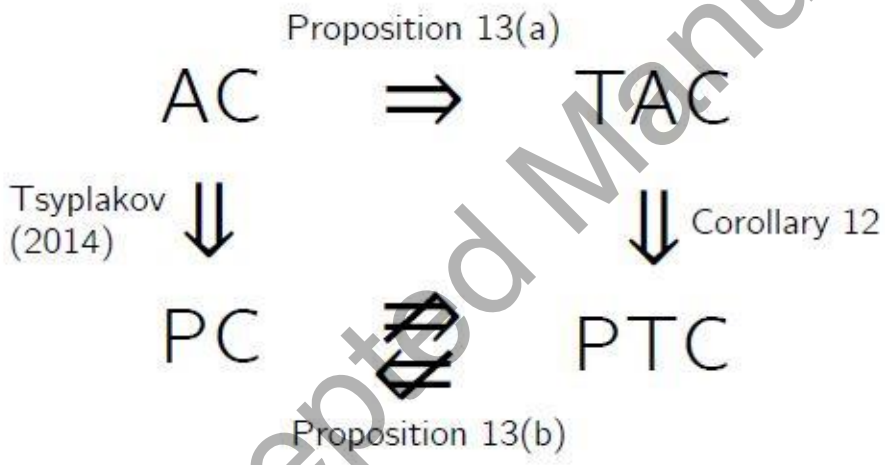


Figure 2: Hierarchies of the notions of calibration and tail calibration. Auto-calibration (AC) implies probabilistic calibration (PC), as well as tail auto-calibration (TAC), which in turn implies probabilistic tail calibration (PTC) (under technical conditions).

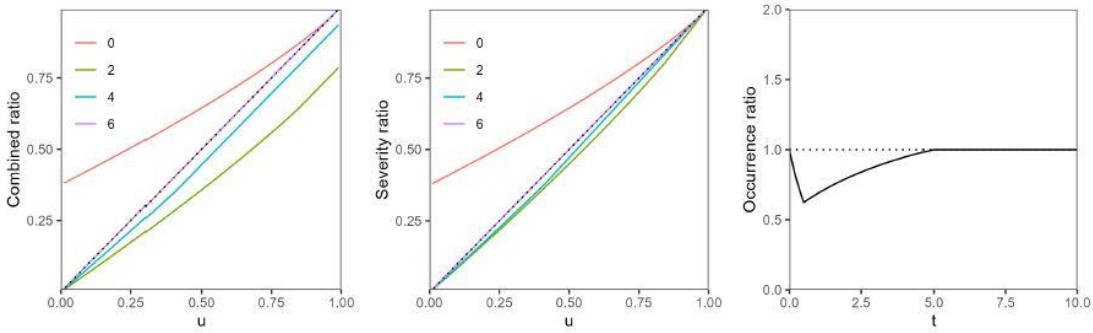


Figure 3: Probabilistic tail calibration diagnostic plots for Example 15. The different colors correspond to different thresholds *t*. Left: combined ratio (10); middle: *pp*-plot of $z_i^{(t)}$ for $i \in \mathcal{I}_t$; right: occurrence ratio (11) as a function of *t*.

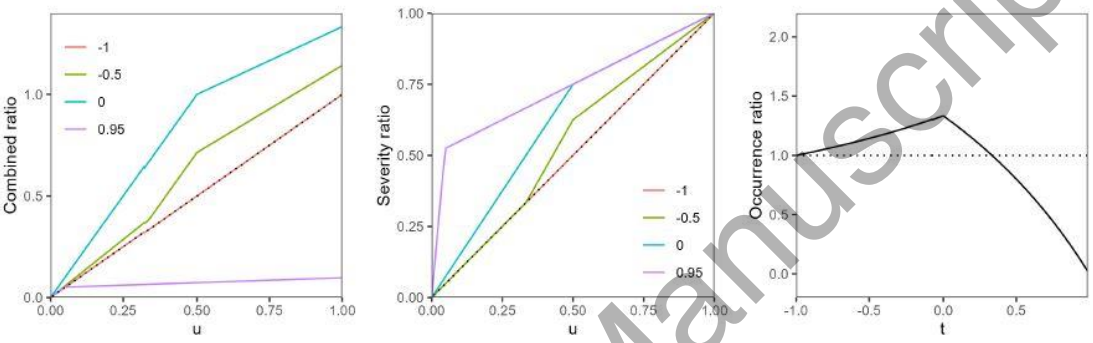


Figure 4: Probabilistic tail calibration diagnostic plots for the unfocused forecaster in Example 16 with *G* the uniform distribution on *[*0, 1*]*. The different colors correspond to different thresholds *t*. Left: combined ratio (10); middle: *pp*-plot of $z_i^{(t)}$ for $i \in \mathcal{I}_t$; right: occurrence ratio (11) as a function of *t*.

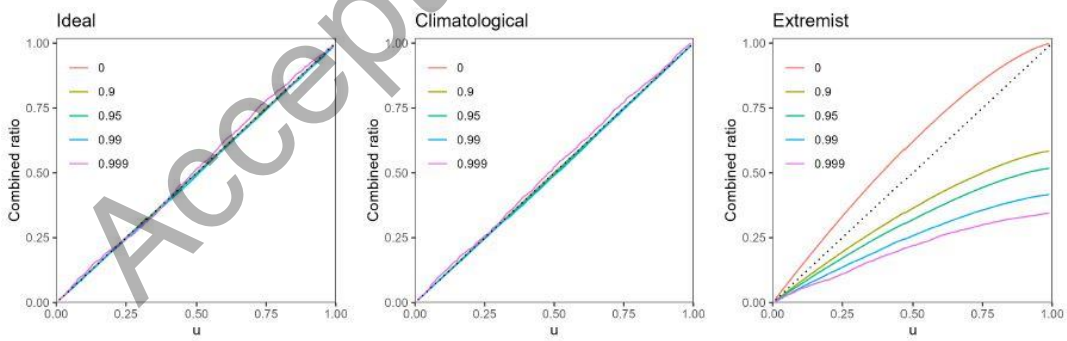


Figure 5: Combined probabilistic tail calibration diagnostic plots (10) for the ideal, climatological, and extremist forecasters in Example 17. Results are shown for five thresholds *t*, expressed in terms of quantiles of the 10^6 observations.

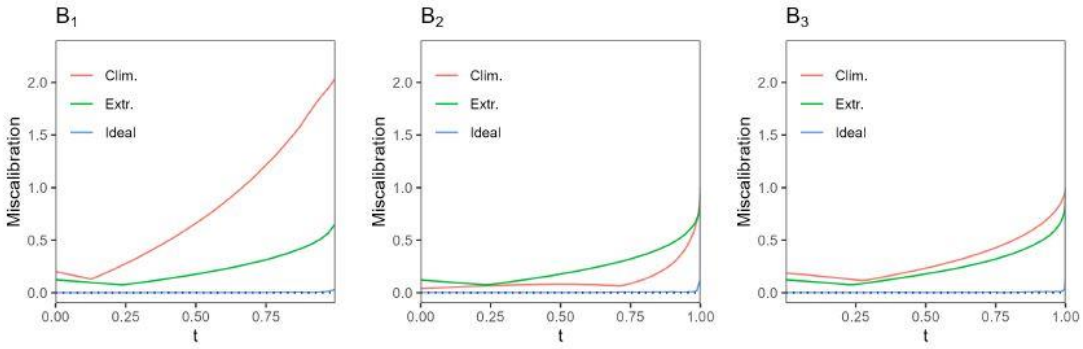


Figure 6: Maximal distance (13) of $\hat{R}_{t,j}$ from the diagonal as a function of t for the climatological (red), ideal (blue), and extremist (green) forecasters in Example 17 for $J = 3$ bins (left to right) of Δ. The threshold t is expressed as a quantile of the 10^6 observations.

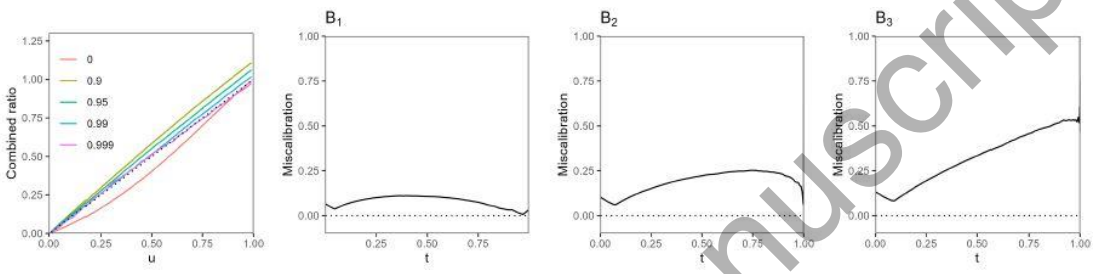


Figure 7: First plot: Combined probabilistic tail calibration diagnostic plot in (10) for the random forecast in Example 18, for five thresholds t. Second, third, and fourth plots: Distance (13) of $\hat{R}_{t,j}$, $j = 1,2,3$, from the diagonal as a function of t for $J = 3$ bins of Δ. In all plots, thresholds are expressed in terms of quantiles of 10^6 observations.

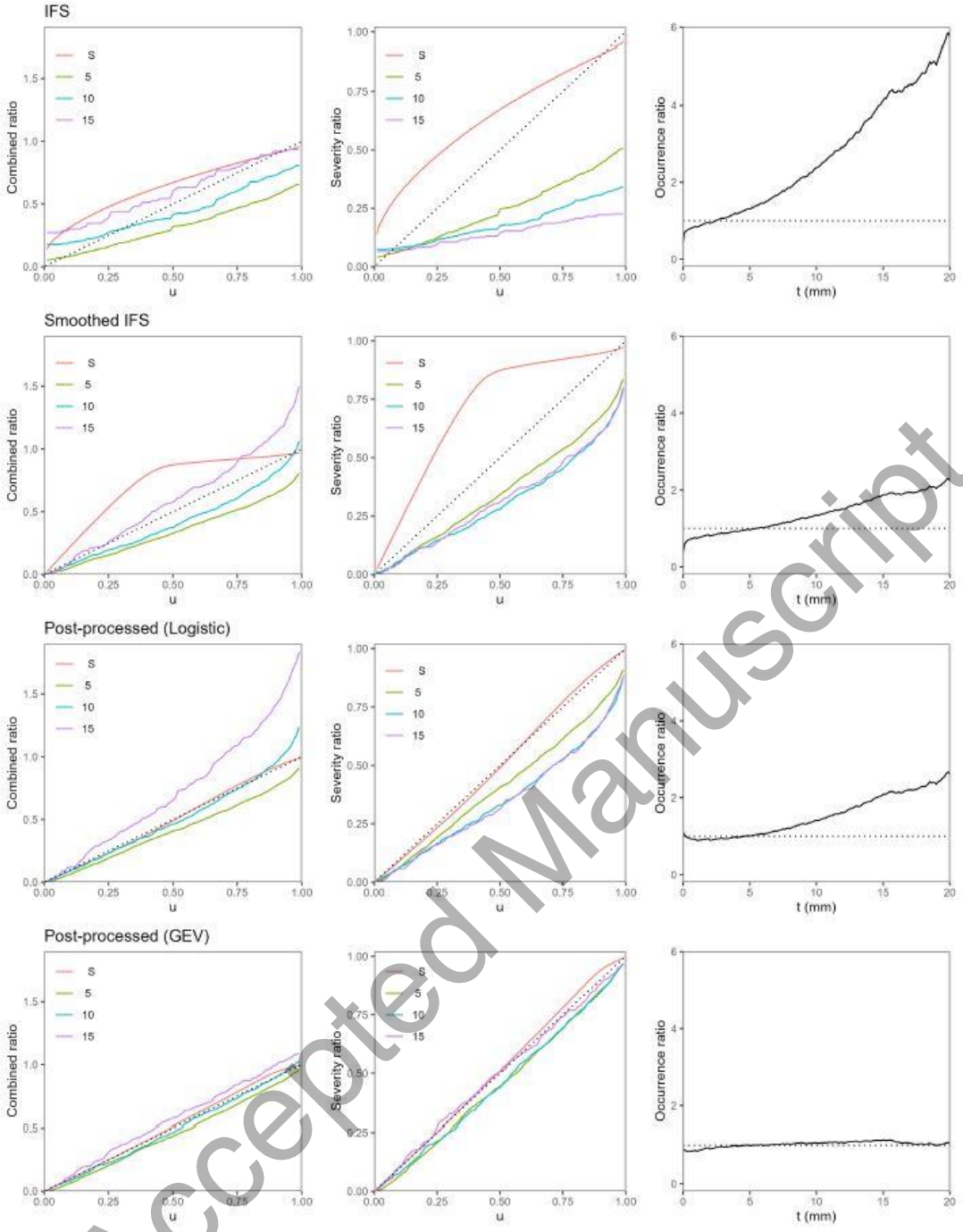


Figure 8: Probabilistic tail calibration plots for the IFS and statistically post-processed forecast distributions in Section 5. Left: combined ratio (10); middle: *pp*-plot of z_i^t for $i \in \mathcal{I}_t$; right: occurrence ratio (11) as a function of t. Results are shown for thresholds 5, 10, and 15 mm, roughly corresponding to the 97^{th}, 99^{th}, and 99.5^{th} percentiles of the observed precipitation measurements. The symbol 'S' denotes the standard notion of probabilistic calibration ($t = -\infty$).