A Primer on Tail Calibration

Oliver Hannaoui

This document is inspired as an explainer for the topic of *Tail Calibration* described in Allen et al. [2025]. All errors are my own.

Table of Contents

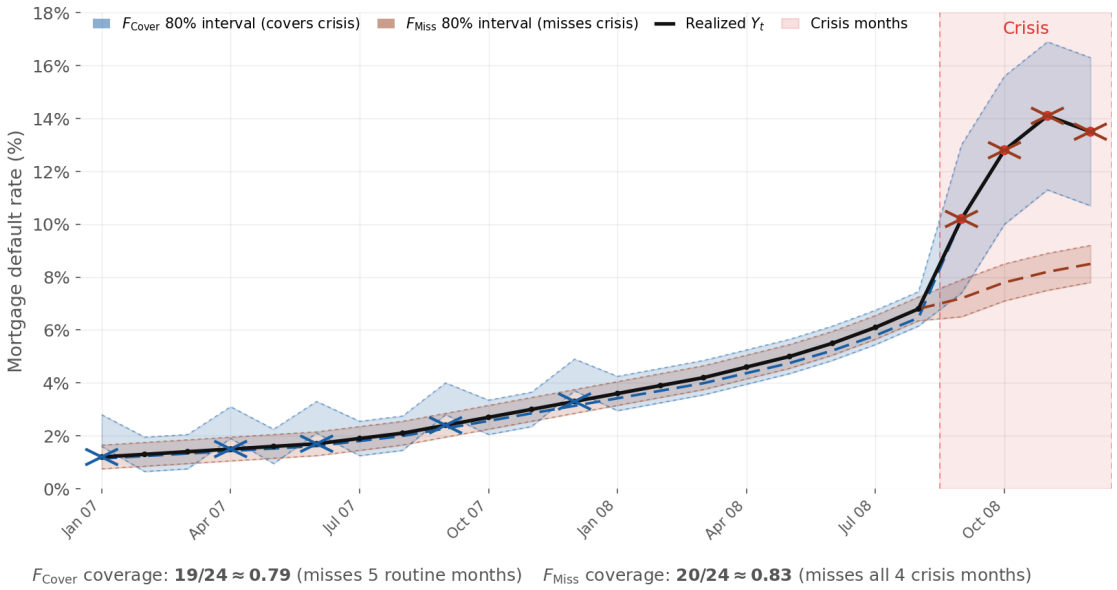
I Motivating the Need to Predict Tail Events

In many 'interesting' settings, extreme events hold an outsized importance on aggregate outcomes. Finance provides the most salient example for understanding this phenomenon: a study by JP Morgan found that the profit of a \$10,000 investment in the S&P 500 from 2003 to 2024 is *halved* if one missed out on the 10 best days over this whole period (i.e., less than 0.15% of days).

A natural next question: does evaluating a probabilistic forecast with our usual tools give us any insight on how well our forecasts can detect such extreme events?

One can remain in the domain of finance to provide an illustrative example of why the most natural measure of assessing forecast quality, the coverage metric, fails to do so.

Figure 1: Hypothetical example of two forecasters predicting mortgage default rates during the 2007 to 2009 period. The dotted red line signifies the onset of the crisis.



* Author's construction using Claude Sonnet 4.6.

As a quick reminder of the general setup, we observe periods $t \in \{1, \ldots, T\}$ and at each period t the forecaster provides a prediction of the distribution of Y_t. The forecasts $\hat{F}_t$ and $\tilde{F}_t$ supply α-confidence regions $\hat{R}_t^\alpha$ and $\tilde{R}_t^\alpha$, respectively, derived from their forecasted distributions which is used to assess the quality of their predictions.

The stylized example in Figure 1 showcases forecast of mortgage default rates during the 2007–2009 period at the onset of the Global Financial Crisis over $T = 24$ monthly forecasts. One observes that exactly 4 time periods correspond to the acute crisis months (e.g., September 2008 through December 2008, following the collapse of Lehman Brothers) during which the cost of forecast error was orders of magnitude larger than in any tranquil month.

The total coverage is calculated as,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t \in \hat{R}_t^{0.1}\} \approx 0.8, \quad \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{Y_t \in \tilde{R}_t^{0.1}\} \approx 0.83,$$

noting that the above calculation is agnostic to which of the time periods was covered, providing a semblance of parity between the two forecasts. A naive evaluator who fails to investigate the data may prefer forecast $\tilde{F}$ by noting its slight advantage in the coverage metric.

Yet investigating Figure 1 should make it evident which forecast the keen risk manager would overwhelmingly prefer. The practical consequences of choosing $\tilde{F}$ over $\hat{F}$ are asymmetric to a degree that defies comparison.

A final concluding remark is that outsized effect of extreme events is not limited to finance. Figure 2 depicts a typical setting of *infectious disease forecasting* which can be depicted as *heavy-tailed with concentration*. The plug-in estimator of the Gini coefficient shows that the average difference in the number of flu hospitalizations between two weeks is *more than 1.3 times the average* number of flu hospitalizations over the whole observed period.[1]

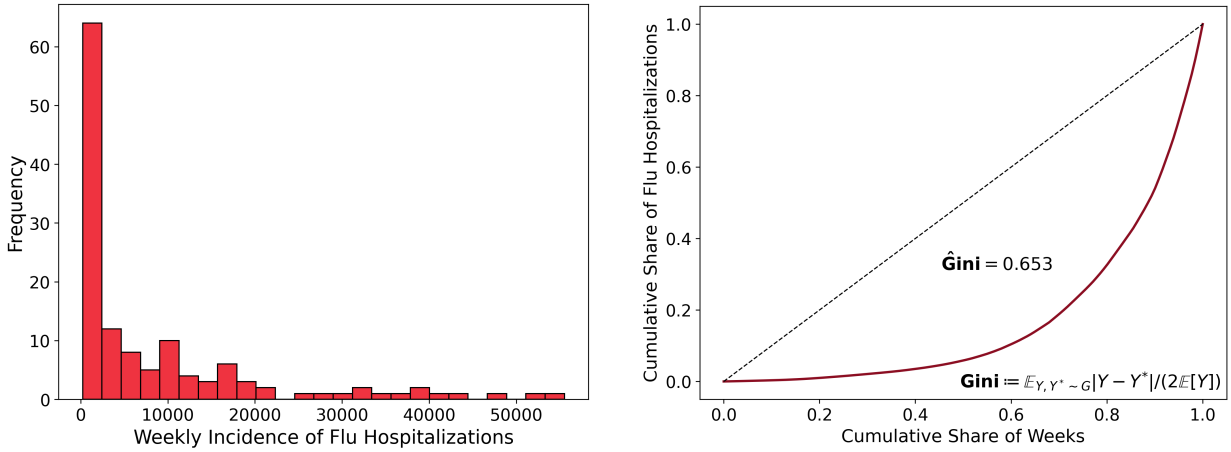


Figure 2: Histogram (left) and Lorenz curve (right) of *Weekly Incidence of Flu Hospitalizations* in US, Sep. 2023 to present.

II Why Proper Scoring Rules Don't Capture Extreme Events

The natural next question may be: do proper scoring rules capture difference of prediction forecasts at the tails? That is, are scoring rules *tail aware*. The title of this section already provides a hint to the answer of this question.

Definition 1 (Proper Scoring Rule). Let $\{Y_t\}_{t\in[T]} \subseteq \mathbb{R}$, with $Y \sim G$, $G \in \mathcal{P}(\mathbb{R})$. A *scoring rule* is a map:

$$S : \mathbb{R} \times \mathcal{P}(\mathbb{R}) \to \mathbb{R} \cup \{+\infty\}.$$

A scoring rule is said to be *proper* if

$$\mathbb{E}[S(F, Y)] \geq \mathbb{E}[S(G, Y)], \quad \forall F \in \mathcal{P}(\mathbb{R}).$$

Finally, a scoring rule is *strictly proper* if the above inequality is strict for all $F \neq G$.

A common scoring rule that the reader is no doubt familiar with is the Continuous Rank Probability Score (CRPS).

[1]For reference, this is roughly equal to the Gini index of the distribution of wealth in South Africa - the highest of any country in the world according to the World Bank's estimates.

Example 2 (CRPS). Let F be a random forecast and y a draw from the random target variable Y. The *Continuous Ranked Probability Score (CRPS)* is defined

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 dx.$$

It also admits the following decomposition (Gneiting and Raftery [2007]):

$$\text{CRPS}(F, y) = \mathbb{E}_{X \sim F}|X - y| - \frac{1}{2}\mathbb{E}_{X, X' \sim F}|X - X'|,$$

where $X \perp X'$.

The above scoring rule is of interest due to its widespread use in climate forecast evaluation. A priori, one may not be surprised that the CRPS does not differentiate tail behavior of forecasts - it is a score about average behavior after all. To remedy this, many have proposed a *weighted* CRPS, *wCRPS*:

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

with $W(x) := \int_{-\infty}^{x} w(t)dt$ and $w(x)$ non-negative. In order to emphasize the right-tail, a common choice is $w_q(x) = \mathbf{1}\{x \geq q\}$ for some threshold $q \in \mathbb{R}$.

Let us directly state the main, unfortunately negative, result of Taillardat et al. [2023].

Lemma 3 (wCRPS is not tail-aware [Taillardat et al., 2023]). Let $Y \sim G$. For all $\epsilon > 0$ it is always possible to construct a non-tail equivalent, F, such that

$$\mathbb{E}\left|\text{wCRPS}(G, Y) - \text{wCRPS}(F, Y)\right| \leq \epsilon.$$

Proof.

Appendix C of Taillardat et al. [2023] contains the proof. Appendix D also provides a useful illustrative example of two forecasts with a stark difference in their tail behavior - measured via the generalized Pareto distribution tail index - achieving the same CRPS score.

The intuition is that one can take an F that looks like G up to a change-point u. By 'pushing' u far enough and controlling its mass and decay, we can break the tail equivalence while maintaining an average score that is arbitrarily close.

One should not forget that one recovers the standard $CRPS$ score when letting $w(x) = 1$. The takeaway of the above is that the weighted CRPS is unable to differentiate forecasts

with different tail regimes due to the fact that a non-tail equivalent forecast can perform equally well, to an arbitrary degree, relative to the ideal forecast.

Before concluding this section by citing a more general result from Brehmer and Strokorb [2019], it is informative to introduce a few notions from Extreme Value Theory (EVT). Thus far, we have been using the term 'tail-aware' rather informally. We can make this notion precise.

Definition 4 (Tail Heaviness). Let F, G two distribution functions, $\bar{F} := 1 - F$, the *survival function*, $x^{\cdot} := \sup\{x \in \mathbb{R} : F(x) < 1\}$ the *upper endpoint* of F. Then, G is said to have a *heavier tail* than F, denoted $F <_t G$, if

$$\text{either } x^F < x^G, \text{ or, } \quad x^F = x^G = x^*, \text{ and, } \lim_{x \to x^*} \frac{\bar{F}(x)}{\bar{G}(x)} = 0,$$

for, $x^F, x^G \in \mathbb{R} \cup \{\infty\}$.

Two quick examples. If $X \sim Ber(p),\ Y = 2X$, then $F_X <_t F_Y$ since $x^{F_X} = 1 < 2 = x^{F_Y}$.

For two distributions with the same upper endpoint, one can take $Z \sim \mathcal{N}(0,1),\ S \sim \text{Exp}(\lambda)$. Clearly $x^{F_Z} = x^{F_S} = \infty$. Based on a smart guess one would say that $F_Z <_t F_S$ which we can verify via the survival function ratio, with Φ the CDF of Z:

$$\lim_{x \to \infty} \frac{1 - \Phi(x)}{e^{-\lambda x}}.$$

By L'Hopital's Rule we obtain,

$$\lim_{x \to \infty} \frac{-(2\pi)^{-1/2} e^{-\frac{1}{2}x^2}}{-\lambda e^{-\lambda x}} = (2\pi\lambda^2)^{-1/2} \lim_{x \to \infty} e^{-\frac{1}{2}x^2 + \lambda x} = 0.$$

Definition 5 (Tail Equivalence). Assume an identical setup as in 4. We say that F and G are *tail equivalent*, denoted $F \sim_t G$, if

$$x^F = x^G = x^* \in \mathbb{R} \cup \{\infty\}, \text{ and, } \lim_{x \to x^*} \frac{\bar{F}(x)}{\bar{G}(x)} \in (0, \infty).$$

We now conclude with the main resulting from Brehmer and Strokorb [2019].

Lemma 6 (*All* scoring rules are not tail-ware Brehmer and Strokorb [2019]). *Let* $S : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$ a proper scoring rule, $Y \sim G$ with $G \in \mathcal{F}$. Then if there is an $F \in \mathcal{F}$ with *heaver tail* than G for all $\epsilon > 0$ there is an $F_\epsilon \in \mathcal{F}$ that is **not** tail equivalent to G and such that

$$|\mathbb{E}_{Y\sim G}[S(F_\epsilon, Y)] - \mathbb{E}_{Y\sim G}[S(G, Y)]| \leq \epsilon.$$

Proof.

See Lemma 5.3 of Brehmer and Strokorb [2019], which also provides an equivalent result formulation using *max-functionals*.

A more general result is established by leveraging the fact that proper scoring rules are *diagonal continuous*[2]: the expected score is stable under small perturbations. One then constructs a mixture $\lambda F_\epsilon + (1-\lambda)G$ via a convex combination of the perturbed forecast F_ϵ and the true G. Taking $\lambda \to 0$ breaks the tail equivalence while keeping the expected score arbitrarily close.

The above demonstrate that a lower forecast score does *not* necessarily imply a forecast is capable of forecasting extremes.

III Revisiting Classical Calibration

The results of section II suggest that in order to assess how well a forecast is calibrated to predicting extreme events, one must circumvent scoring rules. Indeed, Allen et al. [2025] propose combining classical notions of calibration with tools from extreme value theory to assess tail calibration.

Before, it is useful to remind the general setting we are in and how we define the usual notion of forecast calibration.

In the background both the random forecast F and the target Y live on a joint probability space $(\Omega, \mathcal{A}, P)$. [3]. Furthermore, one can think of F being a $\mathcal{G}$ measurable *random distribution function* and Y being a $\mathcal{A}$ measurable random variable, with $\mathcal{G} \subseteq \mathcal{A}$.

Intuitively, this indicates that the random forecast F 'sees' less of the information available that determines the underlying target random variable Y.[4] This notion of conditioning on available information is a useful framework for thinking about calibration.

[2]That is, if S is a *proper* scoring rule, $\mathbb{E}_{Y\sim G}[\lambda F + (1-\lambda)G, Y] \overset{\lambda\to 0}{\to} \mathbb{E}_{Y\sim G}[G, Y]$ for all distributions F.

[3]One should be careful not to conflate the random forecast F with the true distribution function of the random variable Y, which we will denote $G(y),\ \forall y \in \mathbb{R}$

[4]If F saw all the information that Y saw, there would be no need to forecast

In addition to this conditioning set, the *Probability Integral Transform* (PIT), $Z_F := F(Y)$ proves as an invaluable tool to characterize calibration. One should recall that if Y has *CDF* G, then, assuming Y is continuous with a strictly monotonic non-decreasing CDF, a heuristic proof[5]

$$\mathcal{L}\text{aw}(G(Y)) = P_Y(G(Y) \leq t) = P_Y(Y \leq G^{-1}(t)) = G \circ G^{-1}(t) = t.$$

Thus, a typical procedure to assess the quality of a forecast in predicting Y is to assess if the empirical PIT resembles a uniform on the unit interval. What makes this transformation so useful in practice, is the fact that we are inspecting a composition of the Forecast with the target, $F \circ Y$.

Definition 7 (Classical Notions of Calibration). Let (F, Y) a random forecast and target random variable, respectively. Furthermore, let $(\Omega, \mathcal{A}, P)$ the underlying joint probability space and $\mathcal{B} \subseteq \mathcal{A}$ a σ-algebra with $Z_F := F(Y)$.

i. F is *probabilistically $\mathcal{B}$-calibrated* for Y if:

$$P(Z_F \leq u |\ \mathcal{B}) = u, \quad \text{almost surely, } \forall u \in [0, 1].$$

ii. F is *auto-calibrated* for Y if F is $\sigma(F)$-calibrated for Y, that is:

$$P(Z_F \leq u |\ F) = u, \quad \text{almost surely, } \forall u \in [0, 1].$$

iii. F is *probabilistically* calibrated for Y if F is $\{\emptyset, \Omega\}$-calibrated for Y, that is:

$$P(Z_F \leq u) = u, \quad \forall u \in [0, 1].$$

The interpretation is straightforward here. Given the forecast has been issued, the law of the target Y must be *entirely* determined through the (now conditioned on, hence deterministic) forecast. Note that, by definition, conditioning on F is equivalent to condtioning on $\sigma(F)$. This motivates generalizing calibration based on the conditioning set.

In other words, the PIT transform of Y via F, after conditioning on all the information in $\mathcal{B}$ must follow the law of $U \sim \text{Uniform}[0, 1]$ which is also independent of $\mathcal{B}$. The case where $\sigma(F) = \mathcal{B}$ yields auto-calibration. Probabilistic calibration is the special case where $\mathcal{B} = \{\emptyset, \Omega\}$, and is naturally, a weaker notion.

Another useful notion of calibration is presented next.

[5] If the CDF is just monotonic, the result holds via the pseudo-inverse $F^{-1}(y) = \inf\{x : F(x) \geq y\}$. If Y is discrete, one must leverage a *randomized* PIT. More on this later.

Corollary 8 (Alternative Characterization of Calibration). Assume an identical setting to Definition 7, then, if F is *probabilistically calibrated* for Y one has

$$\mathbb{E}[F(y)] = P(Y \leq y) \quad y \in \mathbb{R}.$$

And if F is *tail-calibrated* for Y we have

$$F(y) = P(Y \leq y|\ F) \quad y \in \mathbb{R}.$$

Probabilistic calibration is a weak and unconditional notion of calibration as it only requires calibration to hold on average over the whole observed period. Meanwhile, auto-calibraton requires probabilistic calibration to hold conditioned on all the available information of the forecast.

To be specific, for auto-calibration to hold we require probabilistic calibration when focusing on specific subsets of the observed period. Naturally then, auto-calibration *implies* probabilistic calibration (see Tsyplakov [2011] for more on this).

Referring back to Figure 1, the forecast $\tilde{F}$ which missed the crisis months would likely appear *probabilistically calibrated* in standard diagnostics. However, conditioning on just the second half of the observed period would reveal a clear failure of auto-calibration.

In practice, to rigorously verify auto-calibration one would have to verify that probabilistic calibration holds after conditioning on all $B \in \sigma(Y)$. It is thus more feasible to *rule out* auto-calibration by identifying a conditioning set where probabilistic calibration fails.

IV Tail Calibration: The Setting

Typically, in extreme value theory (EVT)[6], we are interested in the survival function $S(x) = P(Y > x + t)$ for t fixed. This can be decomposed as:[7]

$$P(Y > x + t) = \underbrace{P(Y > t)}_{\substack{\text{survival} \\ \text{probability}}} \underbrace{\frac{P(Y > t + x)}{P(Y > t)}}_{\substack{\text{conditional excess} \\ \text{distribution}}}.$$

Now, recall the characterization of calibration introduced in Corollary 8: under *auto-calibration* of the forecast F with respect to the target Y, we have that $F(y) = P(Y \leq y|F)$ for all $y \in \mathbb{R}$.

So if we want to relate the above decomposition to our random forecast, we condition on it

[6] A branch of statistics focused on modeling and predicting rare, high-impact events located in the tails of probability distributions, rather than the average. See

[7] One should think of t to be chosen such that $P(Y > t) = \epsilon$ for some small $\epsilon > 0$

[8] and see that under auto calibration the first term simplifies and we have:

$$P(Y > x + t|\ F) = (1 - F(t))\frac{P(Y > t + x|\ F)}{P(Y > t|\ F)}.$$

Now note, $\{Y > t + x\} = \{Y > t\} \setminus \{t < Y < t + x\}$. Hence focusing on the second term,

$$\begin{aligned}\frac{P(Y > t + x|F)}{P(Y > t|\ F)} &= \frac{P(Y > t|\ F) - P(t < Y < t + x)}{P(Y > t|\ F)} \\ &= 1 - \frac{P(t < Y < t + x|\ F)}{P(Y > t|\ F)} \\ &= 1 - \frac{F(t + x) - F(t)}{1 - F(t)} \qquad \text{(By auto-calibration).}\end{aligned}$$

This motivates defining the *conditional forecast excess distribution*[9],

$$F_{(t)}(x) = \frac{F(t + x) - F(t)}{1 - F(t)}.$$

So taken together, we have a neat and interpretable decomposition of our EVT quantity of interest,

$$P(Y > x + t|\ F) = \underbrace{(1 - \ F(t))}_{(\star)}\underbrace{(1 - \ F_{(t)}(x))}_{(\diamond)} \tag{1}$$

The first component, denoted $(\star)$ in Equation 1 is the *occurrence*. Ideally we would like $1 - \ F(t)$ to be close $P(Y > t|\ F)$. In English: does our forecast predict the occurrence of extreme events as frequently as they occur? This is the most straightforward and intuitive idea of what tail calibration may signify, a priori.

The second component, denoted $(\diamond)$, is interpreted as the *severity of excess*. In English once again: conditional on an extreme event $\{Y > t\}$ occurring, how well does our forecast F predict Y. In other words, we require the conditional excess forecast F_t to be calibrated to $Y - t$, conditional on $Y > t$.

Note, a direct comparison of $1 - F_t(x)$ with the true quantity $P(Y \geq t + x|Y > t) = \frac{1-P(Y<t+x)}{1-P(Y<t)}$ would be rather uninformative. For some intuition of why this is the case, take $X \sim \log N(\mu, \sigma^2)$, and $Y \sim \mathcal{C}(x_0, \gamma)$, for fixed x

$$\lim_{t\to\infty} \frac{1 - G_X(t + x)}{1 - G_X(t)} = 1, \text{ and, } \lim_{t\to\infty} \frac{1 - G_Y(t + x)}{1 - G_Y(t)} = 1.$$

[8] A priori one might wonder why the law of Y has any dependence on the random forecast F (or maybe I am the only one). When conditioning on F, we condition on the mutual information that both F and Y share as they live on the same probability space. In many applications, this is all the information available to F until the forecast is issued.

[9] Warning of notation overload: this definition is not to be confused with a forecast F made at time t which is also traditionally denoted F_t.

The Cauchy distribution has a *significantly* heavier tail than the log-Normal, but for large t, a crude comparison of $F_{(t)}(x)$ and $P(Y - t < x|Y > t)$ would not suffice to assess how tail calibrated F is with respect to Y as both quantities will be close to 1. Intuitively, this motivates investigating the composition $F_t \circ (Y - t)$ via the PIT which we visit next.

V Defining Tail Calibration

The above signifies composing the excess conditional distribution $F_{(t)}$ with the distribution $Y - t|Y > t$ via an *excess probability integral transform*[10],

$$Z_F^{(t)} := F_{(t)}(Y - t).$$

Taking a step back, note that this implies,

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

Note the above assumes that F is continuous. If F were discrete - as is the case for typical distribution forecasts - the results in the above, and that follow, remain valid under the *randomized* excess probability integral transform

$$\overset{*}{Z}_F^{(t)} := (1 - V) \cdot F_t(Y^- - t) + V \cdot F_t(Y - t), \quad V \sim U[0, 1].$$

Hence, as $t \to -\infty$ we expect to recover the usual notion of PIT calibration. Now to introduce the three main notions of tail calibration from the paper.

Furthermore, if F where *tail calibrated* with respect to Y, then we would require $Z_F^{(t)}|Y > t \sim U[0, 1]$. And again, this would be equivalent to $F_{(t)}$ capturing $Y - t|\ Y > t$. We are now ready to rigorously define tail calibration.

Definition 9 (Tail Calibration). Define $x_Y := \sup\{x \in \mathbb{R} : P(Y \leq x) < 1\}$ the *upper endpoint* of Y. Let $\mathcal{B} \subseteq \mathcal{A}$ a σ-algebra.

Furthermore, assume x_Y is the upper endpoint of the conditional distribution of Y given $\mathcal{B}$. That is, $P(Y > x_Y|\mathcal{B}) = 0$ almost surely, and $P(Y > t|\mathcal{B}) > 0$ for all $t < x_Y$ almost surely. The following are defined for al $t < x_Y$ and for all $u \in [0, 1]$.

i. **Tail $\mathcal{B}$-calibration**. A random forecast F is *tail $\mathcal{B}$-calibrated* if $\mathbb{E}[1 - F(t)|\mathcal{B}] > 0$ almost surely and:

$$\frac{P(Z_F^{(t)} \leq u, Y > t|\ \mathcal{B})}{\mathbb{E}[1 - F(t)|\ \mathcal{B}]} \to u \quad \text{almost surely as } t \to x_Y. \tag{2}$$

[10] Perhaps we should coin it ePIT?

ii. **Tail auto-calibration**. Given $1 - F(t) > 0$, a random forecast F is *tail auto-calibrated* if it is $\sigma(F)$-calibrated. That is,

$$\frac{P(Z_F^{(t)} \leq u, Y > t|\ F)}{1 - F(t)} \to u \quad \text{almost surely as } t \to x_Y. \tag{3}$$

iii. **Tail probabilistic calibration**. Given $\mathbb{E}[1 - F(t)] > 0$, a random forecast $\hat{F}$ is *tail probabilistically calibrated* if it is $\{\emptyset, \Omega\}$-calibrated. That is,

$$\frac{P(Z_F^{(t)} \leq u, Y > t)}{\mathbb{E}[1 - F(t)]} \to u \quad \text{as } t \to x_Y. \tag{4}$$

Besides the fact that both 3 and 4 arise as special cases of 2, one should immediately note that 4 is an *unconditional* notion of tail calibration.

We would now like to return to a similar formulation as we saw in 1. We can apply the same decomposition 'trick' we saw in Equation 1 and see:

$$\frac{P(Y > t|\mathcal{B})}{\mathbb{E}[1 - F(t)|\ \mathcal{B}]} \cdot \frac{P(Z_F^{(t)} \leq u, Y > t|\ \mathcal{B})}{P(Y > t|\mathcal{B})}.$$

The above decomposition implies that condition 2 is equivalent to the to verifying the following two conditions:

$$\frac{P(Y > t|\ \mathcal{B})}{\mathbb{E}[1 - \hat{F}(t)|\ \mathcal{B}]} \to 1 \quad \text{as } t \to x_Y \ \text{ almost surely.} \tag{5}$$

$$\frac{P(Z_{\hat{F}}^{(t)} \leq u, Y > t|\ \mathcal{B})}{P(Y > t|\ \mathcal{B})} \to u \quad \text{as } t \to x_Y \ \text{ almost surely.} \tag{6}$$

The above signifies that we can verify two intuitive conditions of our forecast with respective to the target distribtion to verify tail calibration:

- Via Equation 5 we can verify that the *occurrence ratio* is 1 as we take t to approach the endpoint of Y, x_Y. In English:
- Via Equation 6 we can verify that the *severity ratio* is uniform as $t \to x_Y$. In English: conditional on an extreme event, does the PIT the random variable $Y - t$ our *forecasted conditional excess distribution*

We close this section on a few theoretical results. It turns out that auto-calibration (AC) implies *both* probabilistic calibration (PC) and tail auto-calibration (TAC). However, importantly, PC does not necessarily imply probabilistic tail calibration (PTC), and vice versa. The implication of this is that a calibrated forecast need not be tail calibrated. See section 3 from Allen et al. [2025] for more details on some theoretical results on the matter.

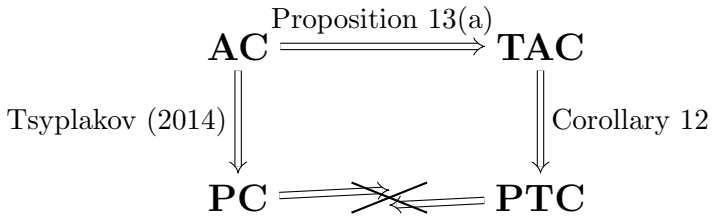


Figure 2 from Allen et al. [2025] presenting hierarchies of the relationships between classical notions of calibration and tail calibration.

One more neat result, it turns out that if Y is in the *max-domain of attraction* of a non-degenerate distribution[11], (i.e., suitable scaling $\max_i X_i$ and taking $n \to \infty$ yields a non-degenerate distribution) probabilistic tail calibration is equivalent to correctly forecasting the shape and scale parameters of the resulting Generalized Pareto Distribution.

One should note that this is purely a theoretical result. From an empirical risk minimization standpoint, *eliciting* the shape parameter of $\{Y_i\}_{i\in\mathbb{N}}$ via the minimization of an empirical loss function is not possible (this is shown in section 4 of Brehmer and Strokorb [2019]).

VI Measuring Tail Calibration in Practice

In practice, we observe a squence of forecast-observation pairs $(F_1, Y_1), \ldots, (F_n, y_n)$ drawn from (F, Y). For a threshold t we can define the set of observations observations that exceed it with the set

$$\mathcal{I}_t = \{i \in [n] : y_i > t\}, \quad n_t = |\mathcal{I}_t|.$$

Also of interest is the conditional excess distribution of forecast $\bar{F}_i$ at fixed threshold t,

$$\bar{F}_{i,t}(x) = \frac{F_i(t+x) - F_i(t)}{1 - F_i(t)}, \quad x \geq 0.$$

Now the above indicates that for a random forecast F to be $\mathcal{B}$-tail calibrated, the graph

$$u \mapsto \frac{P(Z_F^{(t)} \leq u, Y > t|\ \mathcal{B})}{\mathbb{E}[1 - \hat{F}(t)|\ \mathcal{B}]}$$

should be close to the diagonal. Let us first consider the simple unconditional case where $\mathcal{B} = \{\emptyset, \Omega\}$ which corresponds to tail probabilistic calibration.

[11] If you are unfamiliar with EVT, this is similar to a Central Theorem result but for the maximum of a sequence of random variables rather than a sample mean.

If we assume that an LLN type of result holds, the natural plug-in estimator, which we can define is,

$$\hat{R}_t(u) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_F^{(t)} \leq u, Y_i > t\}}{\frac{1}{n}\sum_{i=1}^{n}(1 - F_i(t))}.$$

Recall this is the *combined ratio*, we can apply the familiar decomposition to derive the plug-in *occurrence ratio* and *severity ratios*:

$$\hat{R}_t(u) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_F^{(t)} \leq u, Y_i > t\}}{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i > t\}} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i > t\}}{\frac{1}{n}\sum_{i=1}^{n}(1 - F_i(t))}.$$

Which can be simplified as,

$$\hat{R}_t(u) = \frac{\sum_{i \in \mathcal{I}_t} \mathbf{1}\{Z_F^{(t)} \leq u\}}{|\mathcal{I}_t|} \cdot \frac{|\mathcal{I}_t|}{\sum_{i=1}^{n}(1 - F_i(t))}. \tag{7}$$

In practice, we might take $t = Q(p)$, the p-th empirical quantile of Y. Naturally, this indicates that $n_t \approx n(1-p)$. Thus, as we take $p \to 1$, $n \to 0$. There is thus a fundamental trade-off between choosing an appropriate threshold t to evaluate tail calibration, and having enough data to compute it reliably.

A few more key points to conclude. First, the structure of $\hat{R}_t(u)$ makes asymptotic confidence intervals computable via the Delta Method. The appendix of Allen et al. [2025] considers their computation along with bootstrapping procedure for inference. Alternatively, one can apply a Kolmogorov-Smirnoff test for the severity ratio, and a binomial test for the combined ratio. These are not considered in this exposition but are found in section 5 of Allen et al. [2025].

Finally, consider that we instead observe a panel of l locations (or identifiers) each with n_l observations. One could estimate the combined ratio via a pooled procedure, the updated plug-in estimator, with $\mathcal{I}_{t_l} := \{y_{i,l} > t_l : i \in [n_l]\}$,

$$u \mapsto \frac{\sum_{l=1}^{\mathcal{L}} \sum_{i \in \mathcal{I}_{t_l}} \mathbf{1}\{z_{i,l}^{(t_l)} \leq u\}}{\sum_l |\mathcal{I}_{t_l}|} \cdot \frac{\sum_{l=1} |\mathcal{I}_{t_l}|}{\sum_{l=1}^{\mathcal{L}} \sum_{i=1}^{n_l}(1 - F_{i,l}(t_l))}.$$

As an illustrative example, we can check the tail calibration of the FluSight Ensemble for the 2023-24 and 2024-25 Flu seasons. The full range of diagnostics is outputted in **??**

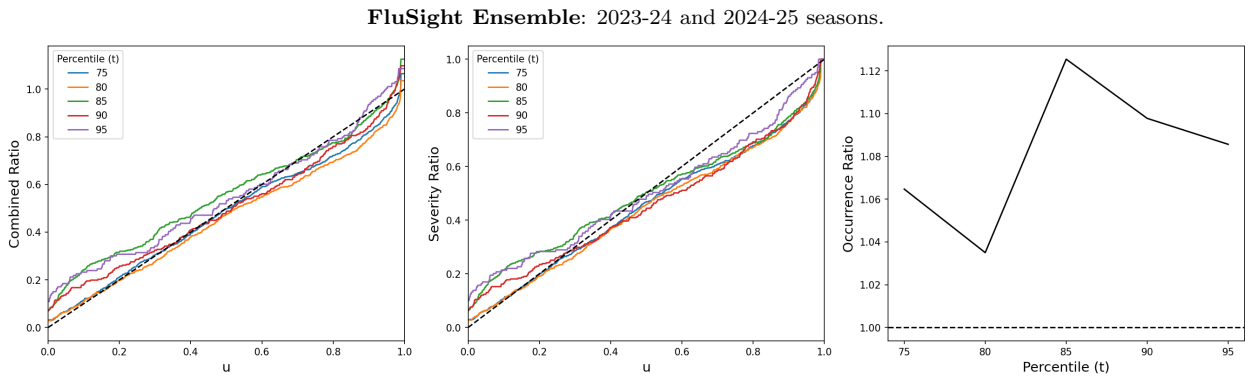


Figure 3: Combined ratio (left), severity ratio (middle), occurrence ratio (right).

The results pass the eye test. Over the 2023-24 and 2024-25 seasons the FluSight ensemble appears to be tail probabilistically calibrated. However, if we condition on each season individually and check for probabilistic calibration, the corresponding diagnostics provide evidence that ensemble is *not* tail auto-calibrated.

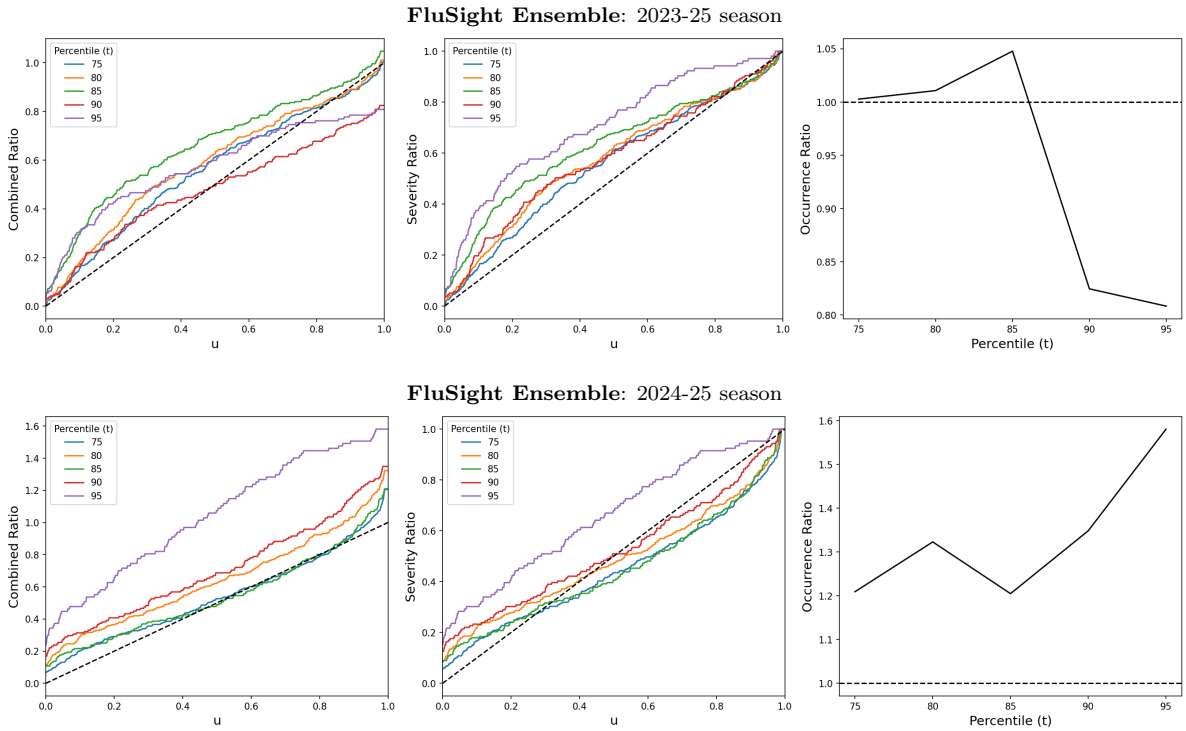


Figure 4: Tail calibration diagnostics of FluSight ensemble for 2023-24 and 2024-25 seasons.

When interpreting the diagnostic plots it is important to step back and remind the original plug-in estimator,

$$\hat{R}_t(u) = \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Z_F^{(t)} \leq u, Y_i > t\}}{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i > t\}} \cdot \frac{\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{Y_i > t\}}{\frac{1}{n}\sum_{i=1}^{n} (1 - F_i(t))}.$$

The u parameter only appears via the excess PIT transform which we can recall one more time,

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

Let us take $u = 1$ and $u = 0$ as two extreme examples, recalling that we condition on $Y > t$,

$$\begin{aligned} u = 1 \Rightarrow Z_F^{(t)} = u &\iff F(Y) = 1 \\ u = 0 \Rightarrow Z_F^{(t)} = u &\iff F(Y) = F(t). \end{aligned}$$

Hence, when we fix a point u and observe the *vertical* distance between the diagonal and the estimated diagnostic value, we are interpreting how tail calibrated the forecast is along the distribution of $Y > t$. For example, at $u = 0.5$ we would be interpreting the diagnostic with respect to the median values of $Y|\ Y > t$. Furthermore, at $u = 1$ we are observing calibration at the most extreme observed value of Y.

VII Concluding Remarks

In conclusion, we began by motivating the need for assessing the ability of a forecast to predict extreme events. As it turns out, the conventional probabilistic forecast methods do not provide us sufficient information to assess this.

To remedy this tension, Allen et al. [2025] introduce a *tail calibration* framework that combines traditional notions of forecast calibration with extreme value theory to provide an elegant and interpretable suite of diagnostics to assess how well calibrated a forecast is at predicting *tail events*.

I recommend the reader to read the Discussion section 6 Allen et al. [2025] for some proposed extensions. These include, but are not limited to,

- Developing a formal testing procedure to arrive at how to choose t in a data-driven fashion.
- Developing post-processing techniques that generate *tail calibration guarantees* via conformal prediction. One could use a fixed sample size to leverage traditional hypothesis testing or consider a sequential testing framework via e-values.
- The literature on multivariate scoring rules is still rather thin, but it is of interest in this setting as well. The author's mention a toolkit of multivariate extreme value theory that can be pulled form.

One natural question: can we improve tail calibration when training forecasts? A natural next step would be to add a regularization term during training that penalizes models that

are not tail calibrated. This is exactly what is considered in Wessel et al. [2025] who considers

$$\hat{F} \in \arg\min_{F} \frac{1}{n} \sum_{i=1}^{n} S(F_i, y_i) + \gamma W_1(\hat{R}_t, U), \quad \gamma > 0, U \sim \text{Unif}[0,1]$$

with, $W_1(\hat{R}_t, U) = \int_0^1 |\hat{R}_t(u) - u| du$. As one would expect, out-of-sample tail calibration is improved at the expense of probabilistic calibration and forecast skill.

References

Sam Allen, Jonathan Koh, Johan Segers, and Johanna Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 120(552):2796–2808, 2025.

Jonas R Brehmer and Kirstin Strokorb. Why scoring functions cannot assess tail properties. 2019.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, and Raphaël De Fondeville. Evaluating probabilistic forecasts of extremes using continuous ranked probability score distributions. *International Journal of Forecasting*, 39(3):1448–1459, 2023.

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

Jakob Benjamin Wessel, Maybritt Schillinger, Frank Kwasniok, and Sam Allen. Enforcing tail calibration when training probabilistic forecast models. *arXiv preprint arXiv:2506.13687*, 2025.

Appendix

Theorem 10 (Worst Case Error of Tail Calibration Plug-In Estimator). Assume $\{F_i, Y_i\}_{i\in[n]}$ independent and identically distributed and define the worst case *occurrence* estimation error:

$$\Delta_t := \inf_{i\in[n]} 1 - F_i(t) - \frac{1}{n}\mathcal{I}_t.$$

Then, with probability at least $1-\delta$, for $\delta > 0$,

$$\sup_{u\in[0,1]} \left|\hat{R}_t(u) - u\right| \leq \frac{(1-p)\sqrt{\frac{\log(2/\delta)}{2n(1-p)}} + |\Delta_t|}{|\Delta_t| + (1-p)},$$

where $\hat{R}_t(u)$ is defined as in Equation 7, and t is chosen as the p-th empirical quantile of Y, that is, $t = n(1-p)$.

This theorem aims to propose a formal testing procedure for tail calibration to choose a data-dependent threshold while maintaining finite sample error guarantees. It is still in progress and may contain errors.

Proof.

One immediately notices the empirical CDF function in the numerator of the plug-in estimator *Equation* 7. This motivates finding an upper bound that isolates the CDF to apply the *DKW inequality*. An initial bound reads:

$$\hat{R}_t(u) = \frac{\sum_{i\in\mathcal{I}_t} \mathbf{1}\{Z_i^{(t)} \leq u\}}{\sum_{i=1}^{n}(1-F_i(t))} \leq \frac{\sum_{i\in\mathcal{I}_t} \mathbf{1}\{Z_i^{(t)} \leq u\}}{n\inf_{i\in[n]}(1-F_i(t))}.$$

Let us introduce an auxiliary term, to be interpreted as the *worst case* error empirical occurrence error:

$$\Delta_t = \inf_i(1-F_i(t)) - \frac{1}{n}\mathcal{I}_t,$$

reminding $\mathcal{I}_t = i : y_i > t$. Furthermore, let us choose the threshold via the empirical quantile, $t = \lceil n(1-p)\rceil$ so that

$$\hat{R}_t(u) \leq \frac{\sum_{i\in\mathcal{I}_t} \mathbf{1}\{Z_i^{(t)} \leq u\}}{n(\Delta_t - \mathcal{I}_t)}.$$

This implies,

$$\sup_{u\in[0,1]}\left|\hat{R}_t(u)-u\right| \leq \sup_{u\in[0,1]}\left|\frac{\frac{1}{n_t}\sum_{i\in\mathcal{I}_t}\mathbf{1}\{Z_i^{(t)}\leq u\}}{\frac{n}{n_t}(\Delta_t+(1-p))}-u\right|.$$

Now take $C_t := \frac{n}{n_t}(\Delta_t + (1-p))$, then adding and subtracting a $u \cdot C_t$ cross-term, one observes that the upper bound of the previous expression can be factorized as

$$\sup_{u\in[0,1]}\left|\hat{R}_t(u)-u\right| \leq \sup_{u\in[0,1]}\left|C_t\left(\frac{1}{n_t}\sum_{i\in\mathcal{I}_t}\mathbf{1}\{Z_i^{(t)}\leq u\}-u\right)+u(C_t-1)\right|.$$

Applying the triangle inequality and taking the sup over u, we have,

$$\sup_{u\in[0,1]}\left|\hat{R}_t(u)-u\right| \leq |C_t|\sup_{u\in[0,1]}\left|\frac{1}{n_t}\sum_{i\in\mathcal{I}_t}\mathbf{1}\{Z_i^{(t)}\leq u\}-u\right|+|C_t-1|$$

A final decomposition is required to apply the *DKW inequality* and achieve a final bound. Notice,

$$\frac{1}{n}\frac{\sum_{i=1}^{n}\mathbf{1}\{Z_i^{(t)}\leq u, Y_i>t\}}{1-p} \to \frac{P(Z_i^{(t)}\leq u, Y>t)}{P(Y>t)} = P(Z_i^{(t)}\leq u|Y>t) =: H(u).$$

And clearly under tail calibration one would have $H(u) = u$. A final decomposition by adding and subtracting the cross term $(1-p)H(u)$,

$$\begin{aligned}\sup_{u\in[0,1]}&\left|(1-p)\left(\frac{1}{n}\sum_{i=1}\frac{\mathbf{1}\{Z_i^{(t)}\leq u, Y>t\}}{1-p}\right)-(1-p)H(u)+(1-p)H(u)-u\right| \\ &\leq (1-p)\sup_{u\in[0,1]}\left|\hat{F}_t(u)-H(u)\right|+(1-p)\sup_{u\in[0,1]}|H(u)-u|,\end{aligned}$$

with

$$\hat{F}_t(u) := \sum_{i=1}^{n}\frac{\mathbf{1}\{Z_i^{(t)}\leq u, Y>t\}}{1-p}$$

Under tail calibration we require

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{1}\{Z_i^{(t)}\leq u\}$$

Now we can introduce $\epsilon' := |C_t|\epsilon - |C_t - 1|$, yielding,

$$\epsilon' = \left|\frac{n(\Delta_t+(1-p))}{n_t}\right|\epsilon - \left|\frac{n(\Delta_t+(1-p))-n_t}{n_t}\right|.$$

Recall that $n_t = |\mathcal{I}_t| = \#\{i \in [n] : Y_i > t\} = \#\{i \in [n] : Y_i > Q(p)\} = \lfloor n(1-p) \rfloor \approx n(1-p)$.

With some algebra, and noting that $\Delta_t + (1-p) = \inf_i(1 - F_i(t)) \geq 0$, we arrive at a neat decomposition of the error term that provides a nice interpretation:

$$\epsilon' = \underbrace{\epsilon}_{\text{tolerance}} + \underbrace{\frac{\Delta_t(\epsilon - \text{sign}(\Delta_t))}{1-p}}_{\text{irreducible occurrence error}} .$$

And hence, by the *DKW inequality*,

$$P\left(\sup_{u \in [0,1]} \left|\hat{R}_t(u) - R(u)\right| > \epsilon'\right) \leq 2e^{-n(1-p)\epsilon'^2}.$$

And thus, with probability at least $1 - \delta$,

$$\sup_{u \in [0,1]} \left|\hat{R}_t(u) - u\right| \leq \frac{(1-p)\sqrt{\frac{\log(2/\delta)}{2n(1-p)}} + |\Delta_t|}{\Delta_t + (1-p)}.$$

(show what happens when miscalibration error is 0, and give an example of what to what kind of error is allowed for large n and small error term)