Assessing the Tail Calibration of Probabilistic Forecasts

Delphi-Reich Joint Meeting, Spring 2026

Oliver Hannaoui

Carnegie Mellon University, Department of Statistics

April 3, 2026

The Setting of Infectious Disease Forecasting

- In infectious disease one typically sees that a small share of observed periods exercise an outsized effect on aggregate outcomes.
- Many other fields relevant to forecasting exhibit this exact phenomenon (e.g., meteorology, finance, seismology).
- Could be characterized as *heavy-tailed* with *strong concentration*.

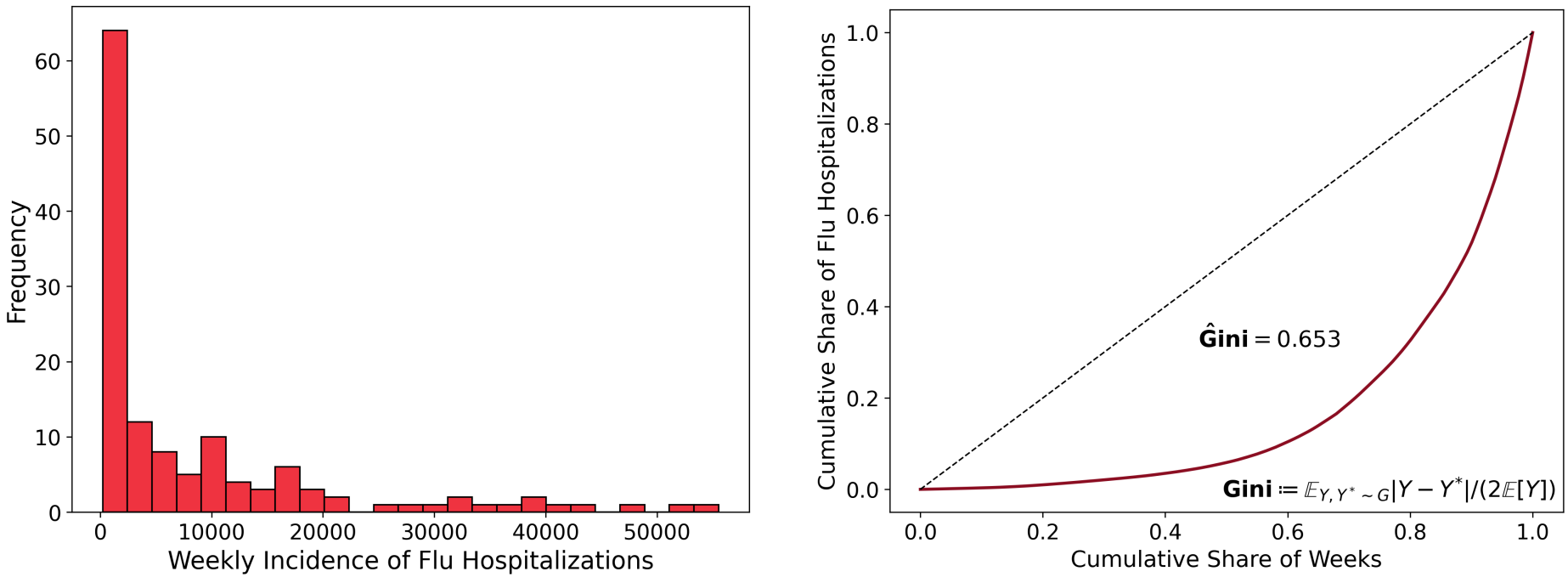
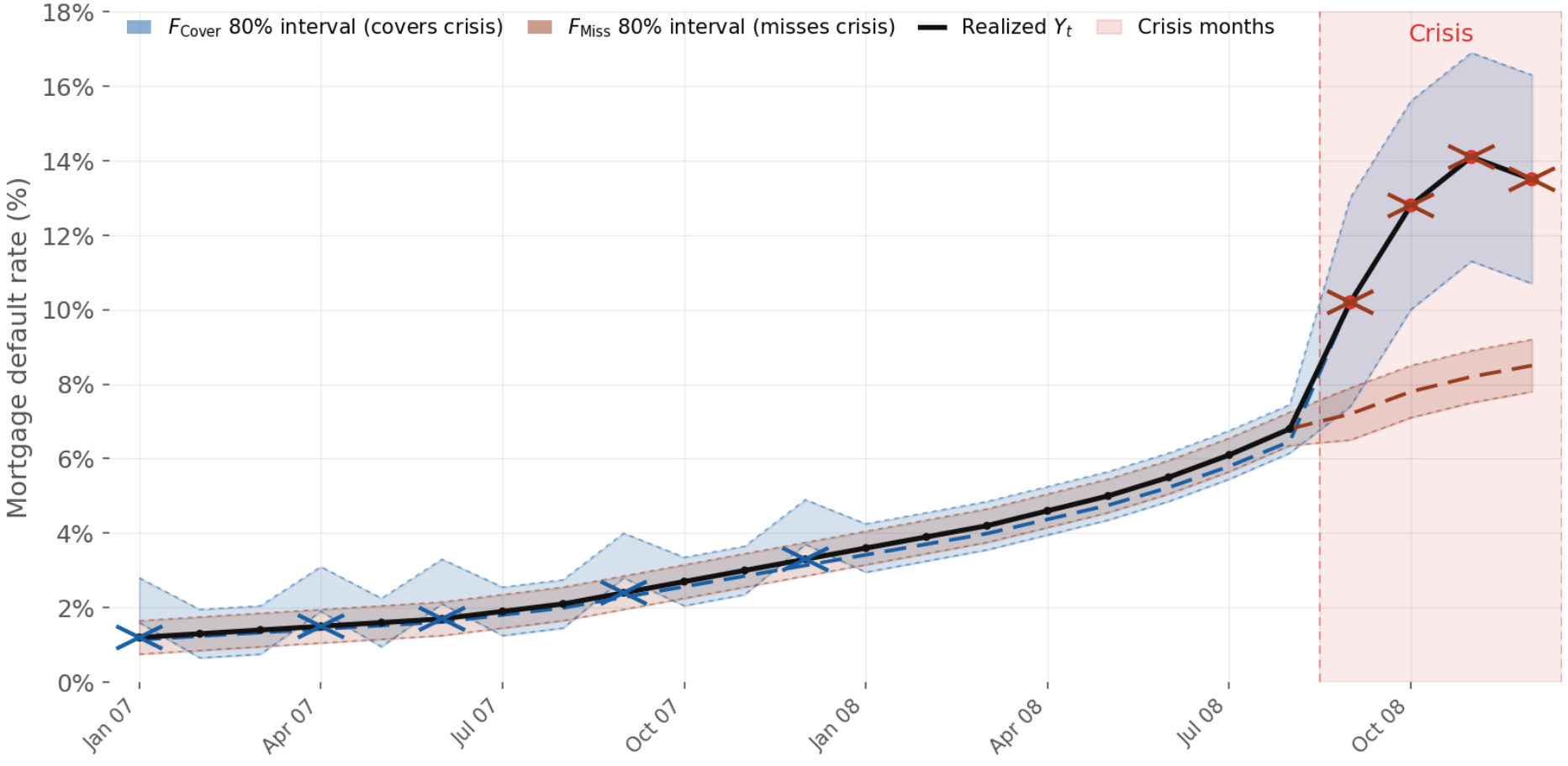


Figure: Histogram (left) and Lorenz curve (right) of *Weekly Incidence of Flu Hospitalizations* in US, Sep. 2023 to present.

Simple Forecast Evaluation Metrics Miss Tails



F_{Cover} coverage: **19/24 ≈ 0.79** (misses 5 routine months) F_{Miss} coverage: **20/24 ≈ 0.83** (misses all 4 crisis months)

Source: Author's construction via Claude Sonnet 4.6

Tail Accuracy in Practice

- Unsurprisingly, despite their importance, it is not easy to forecast extreme events.
- One natural question: do proper scoring rules capture tail behavior? That is, are they *tail aware*.

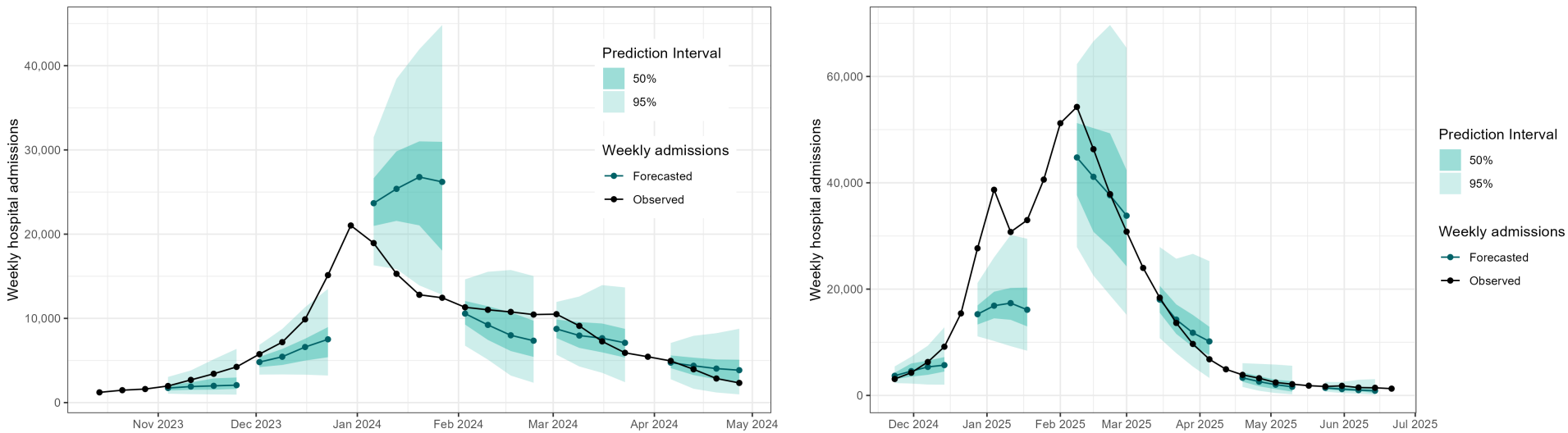


Figure: National median ensemble with 50% and 95% prediction intervals alongside observed weekly influenza hospital admissions for 2023-24 (left) and 2024-25 (right) seasons. Source: [2023-24], [2024-25].

Do Proper Scoring Rules Capture Tail Calibration?

- Taillardat et al. (2023) show that for any $\epsilon > 0$ and it is always possible to construct an F that is not *[tail equivalent]* to Y such that

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

- Brehmer and Strokorb (2019) extend Taillardat et al. (2023) to **all integrable proper scoring rules**. *[formal theorem statement]*
- In English: for any proper scoring rule, it is possible to find a forecast F that exhibits *incorrect tail behavior* while achieving an expected score that is *arbitrarily close* to the true one, G.
- To circumvent the use proper of scoring rules, Allen et al. (2025) introduce a *tail calibration* framework that combines classical notions of forecast calibration with tools from Extreme Value Theory (EVT).

A Quick Refresher on Classical Calibration

– Recall that a random forecast F is **auto-calibrated** for the target $Y \sim G$ if

$$P(Z_F \leq u | F) = u, \quad \text{almost surely.}$$

– While F is **probabilistically calibrated** for Y if

$$P(Z_F \leq u) = u, \quad \text{for all } u \in [0, 1]$$

where $Z_F := F(Y)$ the probability integral transform of Y via F.

– The notions of *probabilistic* and *auto* calibration above are equivalent to (respectively)

$$P(Y \leq y) = \mathbb{E}[F(y)], \quad \text{and, } P(Y \leq y |\ F) = F(y).$$

– Probabilistic calibration is an unconditional, and weaker, notion than auto-calibration. Also, AC $\rightarrow$ PC.

Example: The Unfocused Forecaster

(**Example 3 Gneiting et al. (2007)**). Let the target $Y \sim \mathcal{N}(\mu, 1)$ with $\mu \sim \mathcal{N}(0, 1)$, and the forecast $F = \frac{1}{2}\mathcal{N}(\mu, 1) + \frac{1}{2}\mathcal{N}(\mu + \tau, 1)$, with $\tau \sim \text{Rad}(\frac{1}{2}), \tau \perp \mu$.

- Then F is *probabilistically calibrated* for Y but <u>not</u> *auto-calibrated*.

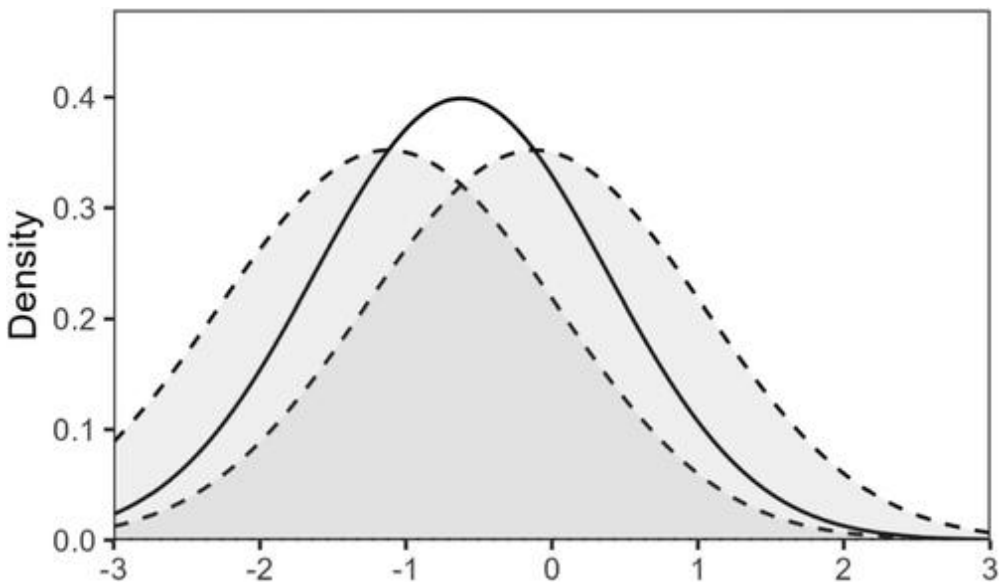


Figure: The 'unfocused' forecaster, F, flips a coin to choose one of the dashed forecasts, while the true distribution of Y (for a fixed μ) is the solid line.

Deriving a Notion of Tail Calibration

- A natural starting point from EVT is a decomposition of the survival function:

$$P(Y > x + t) = P(Y > t)\frac{P(Y > t + x)}{P(Y > t)}.$$

- Under auto calibration this can be written in terms of the forecast predicted forecast F

$$P(Y > x + t|F) = (1 - F(t))(1 - F_{(t)}(x)),$$

where $F_{(t)}(x)$ is the *conditional forecast excess distribution*[1]:

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

[1] *A [quick derivation] for clarity.*

The Excess Probability Integral Transform

- A direct comparison of the LHS and RHS of

$$P(Y - t > x|F) = 1 - F_{(t)}(x)$$

  is still not sufficient to distinguish tail behavior between Y and F.
- Some intuition, take $X \sim \log N(\mu, \sigma^2)$, and $Y \sim \mathcal{C}(x_0, \gamma)$, for fixed x

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

- This motivates a composition via the *excess PIT*,[2]

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

- As $t \to -\infty$ we recover the usual probability integral transform.

[2] If Y is discrete, the above is defined via the *[randomized PIT]*.

A Definition of Tail Calibration

Define $x_Y = \sup\{x \in \mathbb{R} : P(Y \leq x) < 1\}$ the *upper endpoint* of Y. Then,

- F is **tail auto-calibrated** if:

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

- F is **tail probabilistically calibrated** if:

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

- Tail calibration is defined more generally via the conditioning σ-algebra $\mathcal{B} \subseteq \mathcal{A}$. The above are special cases.

Returning to the Prior Decomposition

– Applying the previous decomposition to tail probabilistic calibration:

$$\frac{P(Z_F^{(t)} \leq u, Y > t)}{\mathbb{E}[1 - F(t)]} = \underbrace{\frac{P(Y > t)}{\mathbb{E}[1 - F(t)]}}_{(\diamond)} \cdot \underbrace{\frac{P(Z_F^{(t)} \leq u, Y > t)}{P(Y > t)}}_{(\star)}$$

– Hence, checking for *tail probabilistic calibration* is equivalent to checking two conditions:

$$\frac{P(Y > t)}{\mathbb{E}[1 - F(t)]} \overset{t \to x_Y}{\to} 1 \quad \textbf{and}, \quad P(Z_F^{(t)} \leq u | Y > t) \overset{t \to x_Y}{\to} u.$$

– The *occurrence ratio* ($\diamond$) checks if the rate the forecast F predicts extreme events agrees with the outcome Y.

– The *severity ratio* ($\star$) asks, *given* an extreme event has occurred, is the forecast of the tail well specified?

Assessing Tail Probabilistic Calibration in Practice

– As implied by the definition, we expect the following graph to be close to the diagonal:

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

– Assume we observe $\{(y_i, F_i) : i \in [n]\}$, consider some t large and $\mathcal{I}_t = \{y_i > t : i \in [n]\}$. The excess PIT realization reads:

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

– Leading to the map of the plug-in estimator over $u \in [0, 1]$,

$$u \mapsto \frac{\sum_{i \in \mathcal{I}_t} \mathbf{1}\{z_i^{(t)} \leq u\}}{\sum_{i=1}^n (1 - F_i(t))} = \underbrace{\frac{\sum_{i \in \mathcal{I}_t} \mathbf{1}\{z_i^{(t)} \leq u\}}{|\mathcal{I}_t|}}_{(\star)} \cdot \underbrace{\frac{|\mathcal{I}_t|}{\sum_{i=1}^n (1 - F_i(t))}}_{(\diamond)}.$$

More Intuition

- The expression $P(Z_F^{(t)} \leq u | Y > t) = u$ is equivalent to saying the PIT transform of the random variable $Y - t | Y > t$, via the excess conditional distribution of the forecast $F_{(t)}$, is uniform.
- Recall again that, if

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

recalling that we condition on $Y > t$. Hence,

$$Z_F^{(t)} = 1 \iff F(Y) = 1, \text{ and, } Z_F^{(t)} = 0 \iff F(Y) = F(t).$$

- In the plug-in estimator the parameter u shows up in the term $\mathbf{1}\{z_i^{(t)} \leq u\}$: u serves as a parameter that varies across the *conditional excess distribution*.

Final Word on Tail Calibration: Theory

- When Y is in the max-domain of attraction of a non-degenerate distribution, PTC is equivalent to F forecasting the shape and scale parameters of the limiting GPD of Y.
- In general, probabilistic calibration does **not** imply probabilistic *tail* calibration. A full picture of the hierarchy:

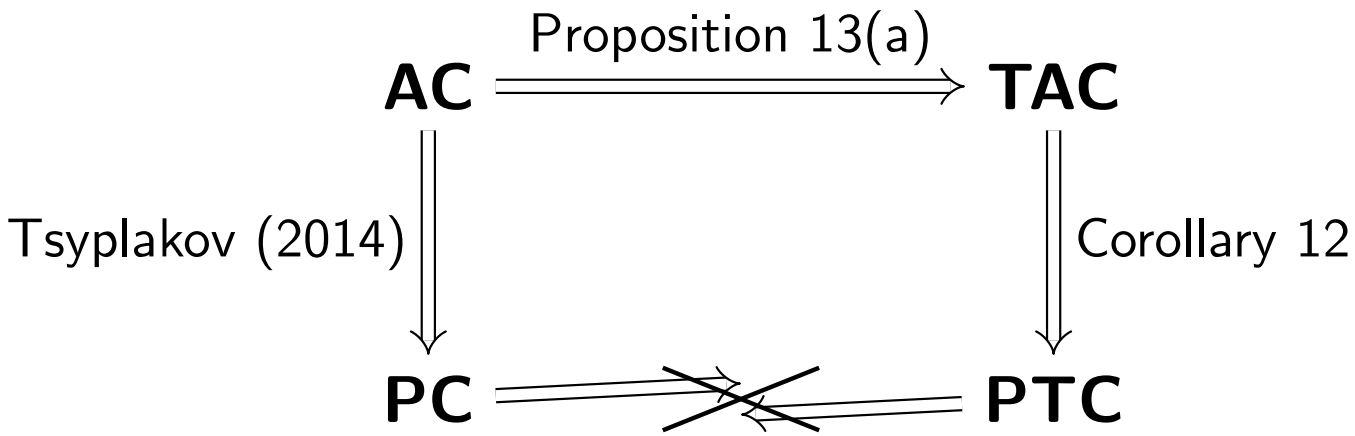


Figure 2 from Allen et al. (2025) presenting hierarchies of the relationships between classical notions of calibration and tail calibration.

Application: FluSight Forecast Hub

- If $t = Q(p)$ tail calibration is assessed over just $n(1-p)$ data points. This can be problematic: 56 weeks $\rightarrow$ **one** $p = 99$ evaluation point.
- Naturally, we *pool* the data across $\mathcal{L}$ locations, yielding the updated plug-in estimator, with $\mathcal{I}_{t_l} := \{y_{i,l} > t_l : i \in [n_l]\}$,

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

FluSight Ensemble: 2023-24 and 2024-25 seasons - same-week forecasts.

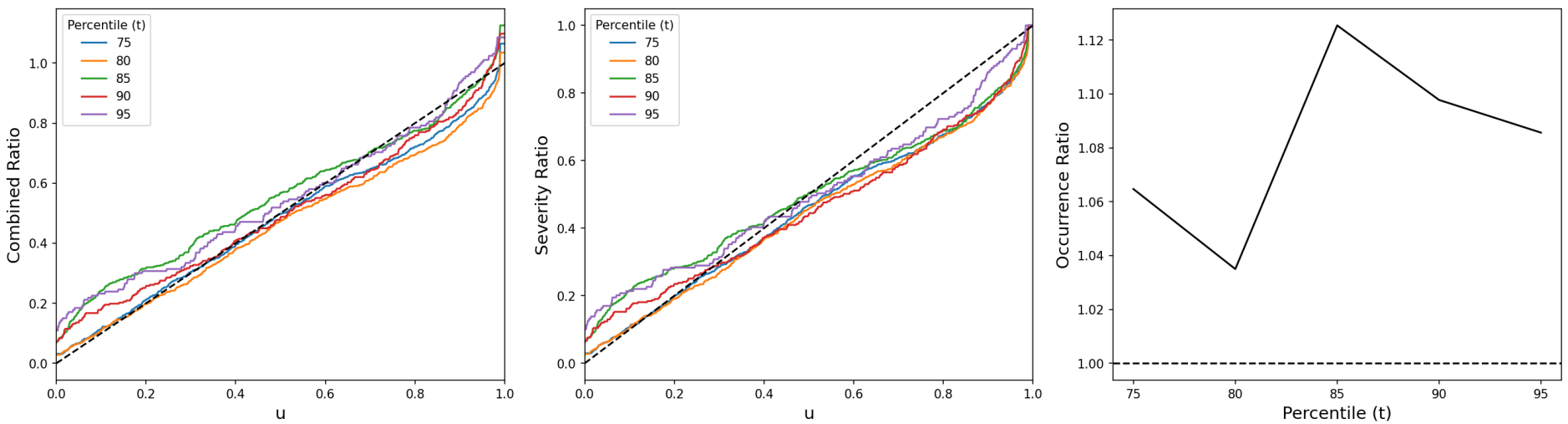
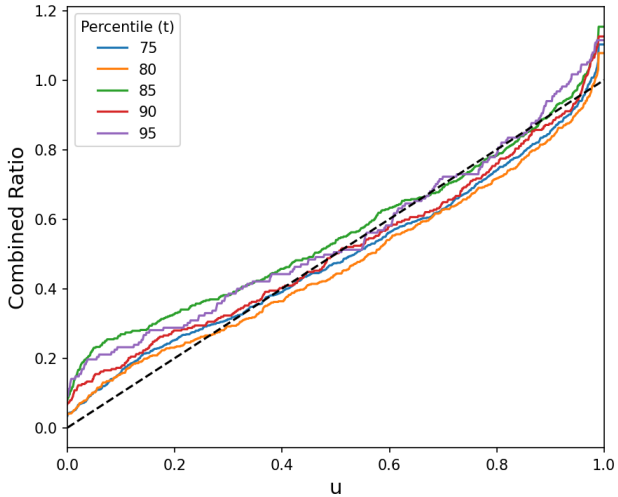
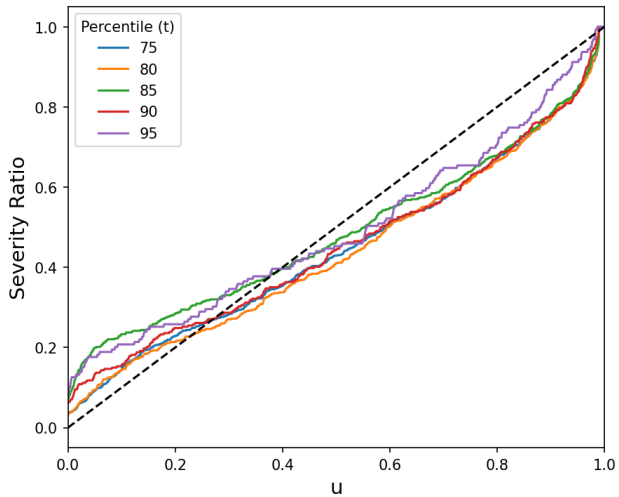


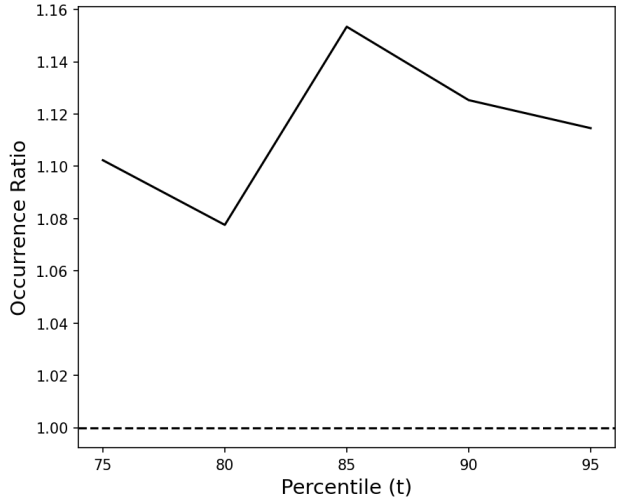
Figure: Combined ratio (left), severity ratio (middle), occurrence ratio (right).

Two Tail *Probabilistically* Calibrated Forecasts

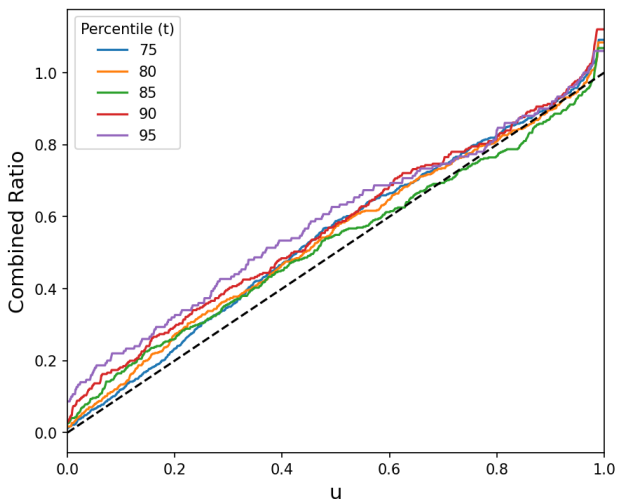
UMass Flusion - same-week forecasts for 2023-24 and 2024-25 seasons.

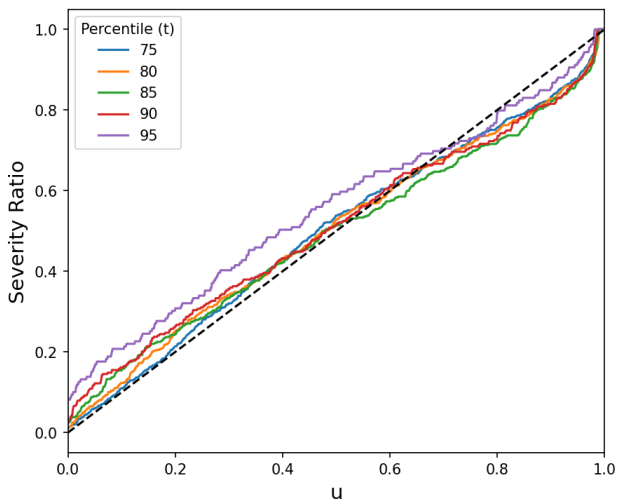


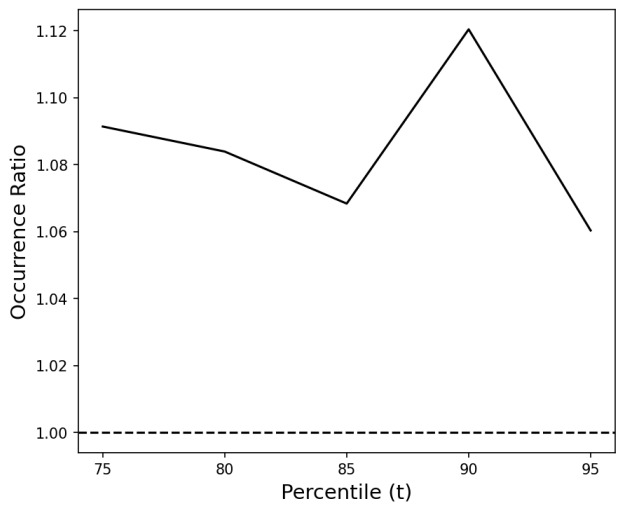




CMU Time Series - same-week forecasts for 2023-24 and 2024-25 seasons.

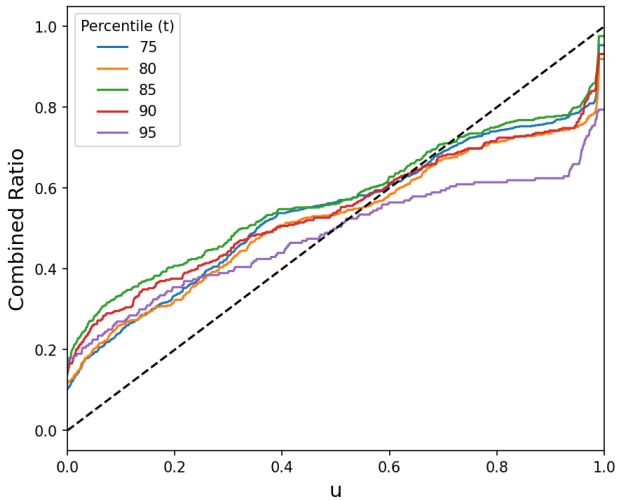


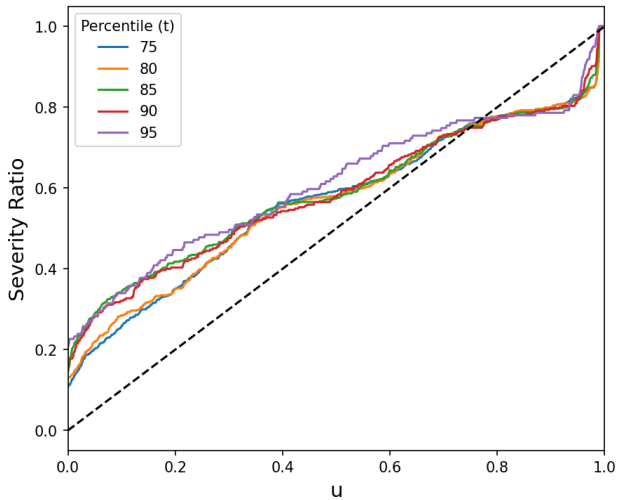


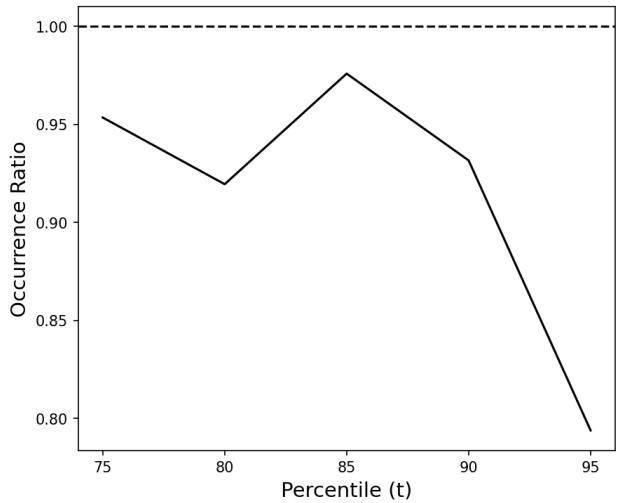


Two Tail Probabilistically *Miscalibrated* Forecasts

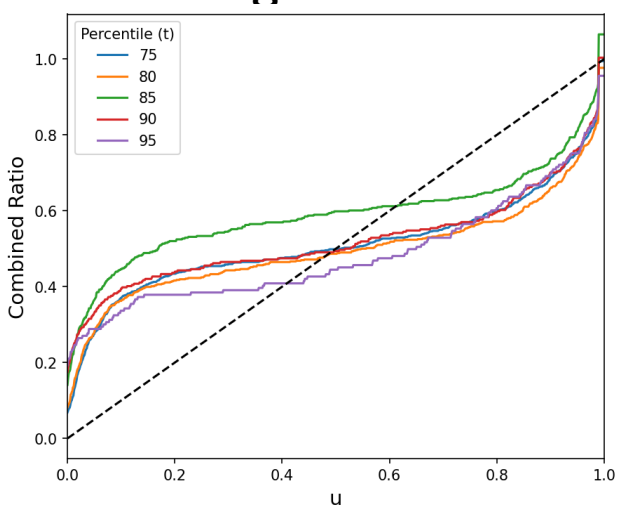
UM-DeepOutbreak - same-week forecasts for 2023-24 and 2024-25 seasons.

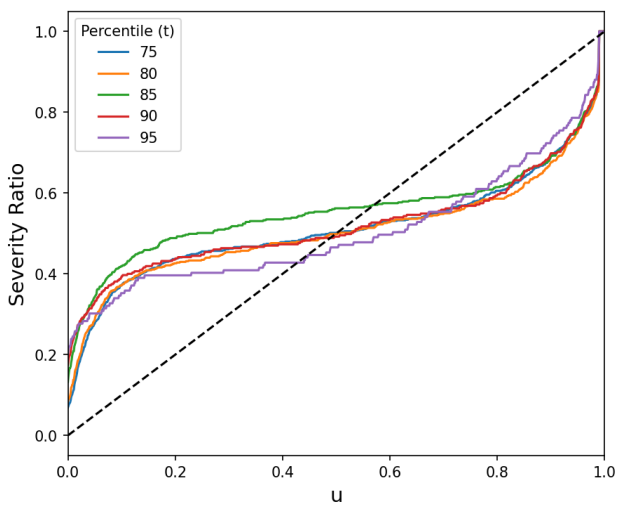


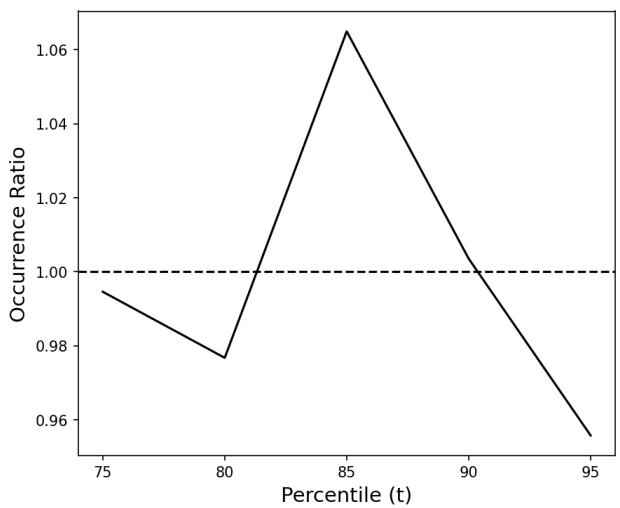




FluSight Baseline - same-week forecasts for 2023-24 and 2024-25 seasons.







Assessing Tail Auto-Calibration in Practice

Return to Slide 12

- To rigorously verify tail auto-calibration one would have to verify probabilistic tail calibration holds conditioned over all $B \in \sigma(F)$.
- We can check a weaker, but necessary, condition that it holds over J disjoint partitions of $\{(F_i, Y_i)\}_{i\in[n]}$,

$$\mathcal{J}_j = \{i \in [n] : (F_i, y_i) \in B_j\}, \quad j \in [J].$$

and for each j and threshold t the plug-in estimator,

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

- This would require $3 \times T \times J$ diagnostic plots.[3] However, under tail calibration as $t \to x_Y$ we must observe

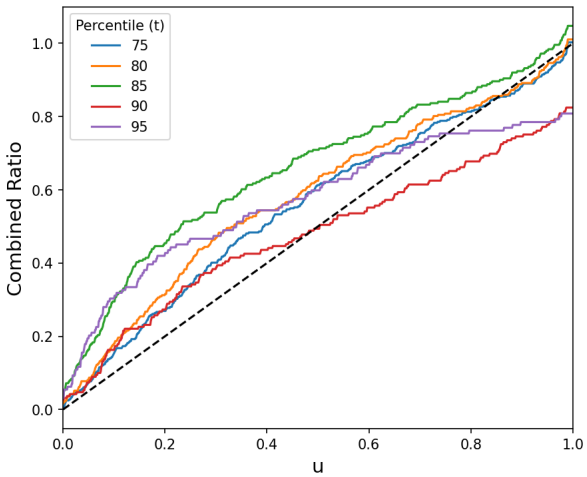
$$\sup_{u\in[0,1]} |\hat{R}_{j,t}(u) - u| \to 0, \quad \forall j \in [J].$$

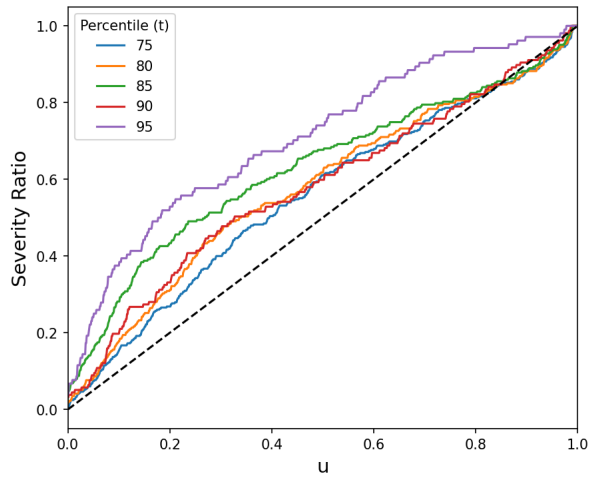
[3] The three plots correspond to the severity, occurrence, and combined ratio plots, for each t of the T total threshold values under consideration.

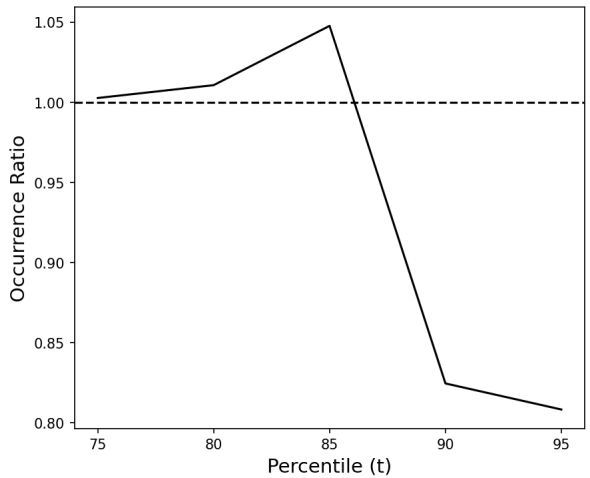
Application: Tail *Auto-Calibration* of FluSight Ensemble

– Due to data constraints, we don't have the luxury to coarsen our partitions too much.

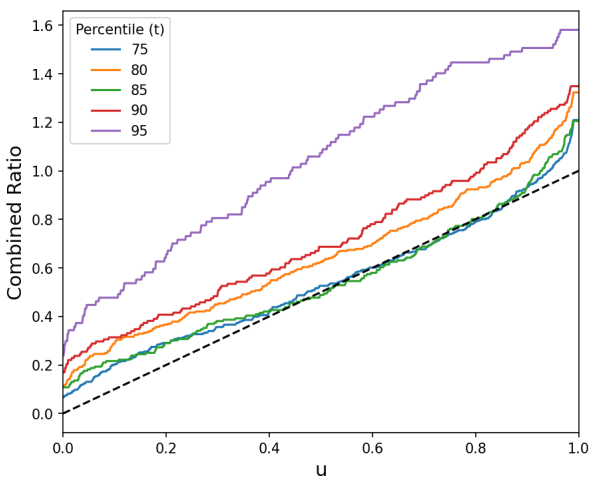
FluSight Ensemble: 2023-24 season - same-week forecasts.

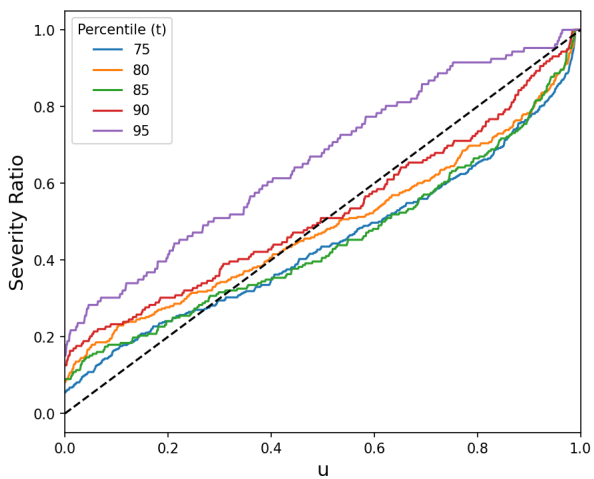


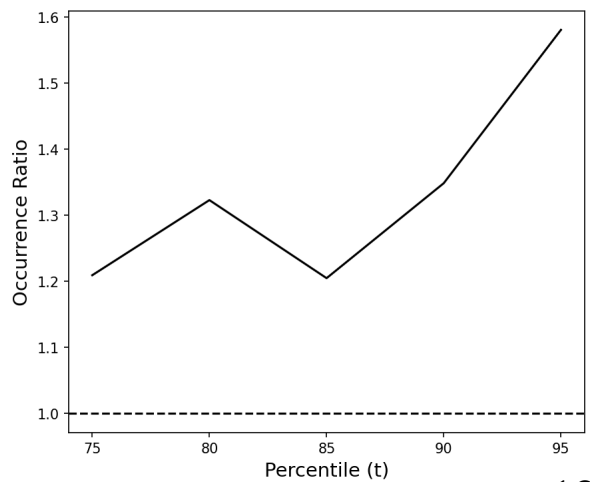




FluSight Ensemble: 2024-25 season - same-week forecasts.







Tail Probabilistic Calibration Variation by State

– A within-state decomposition proves to be a useful diagnostic.

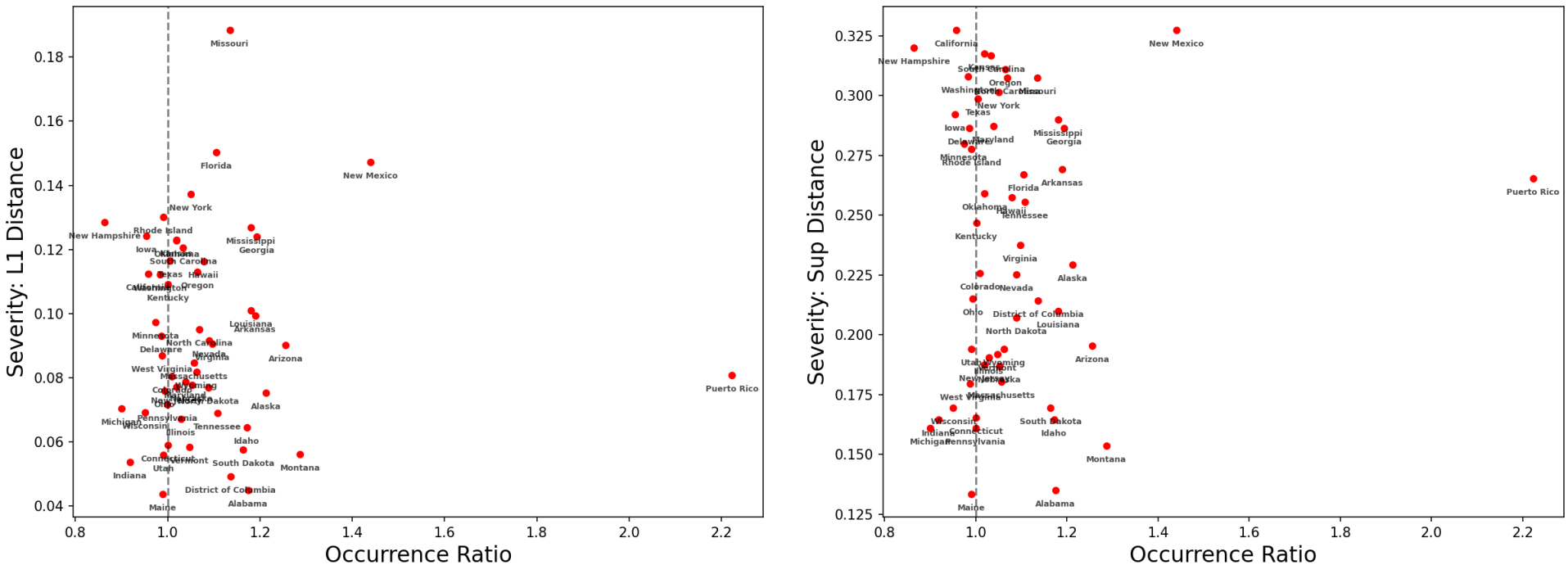


Figure: Occurrence Ratio versus L1 $= \int_0^1 |\hat{R}_t(u) - u| du$ (left) and supremum $= \sup_{u \in [0,1]} |\hat{R}_t(u) - u|$) (right) distances of severity ratio for *same-week FluSight Ensemble* forecasts, at $p = 75$.

Increasing the Forecast Horizon Degrades Tail Calibration

– As our intuition would lead us to believe, an increasing horizon degrades tail calibration.

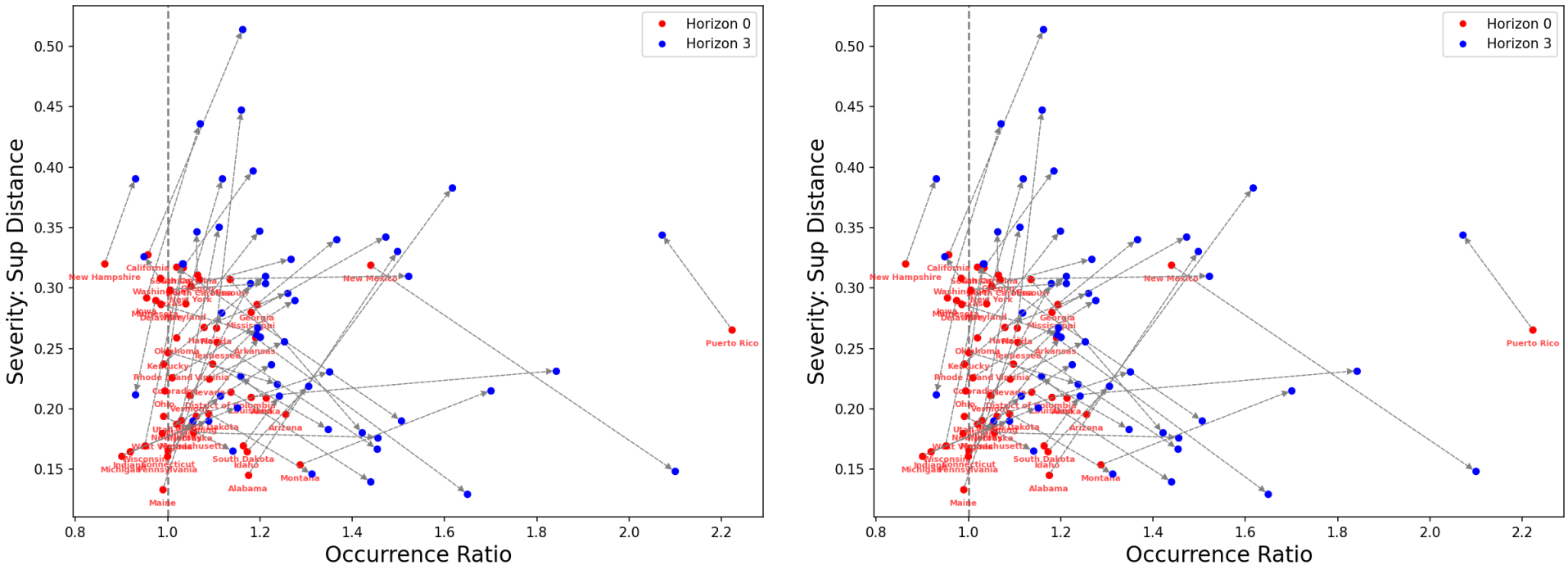


Figure: Occurrence Ratio versus L1 (left) and supremum (right) distances of severity ratio for *same-week* (red) to 1-month ahead (blue) *FluSight Ensemble* forecasts, at $p = 75$.

Tail Calibration by Horizon: Pooled

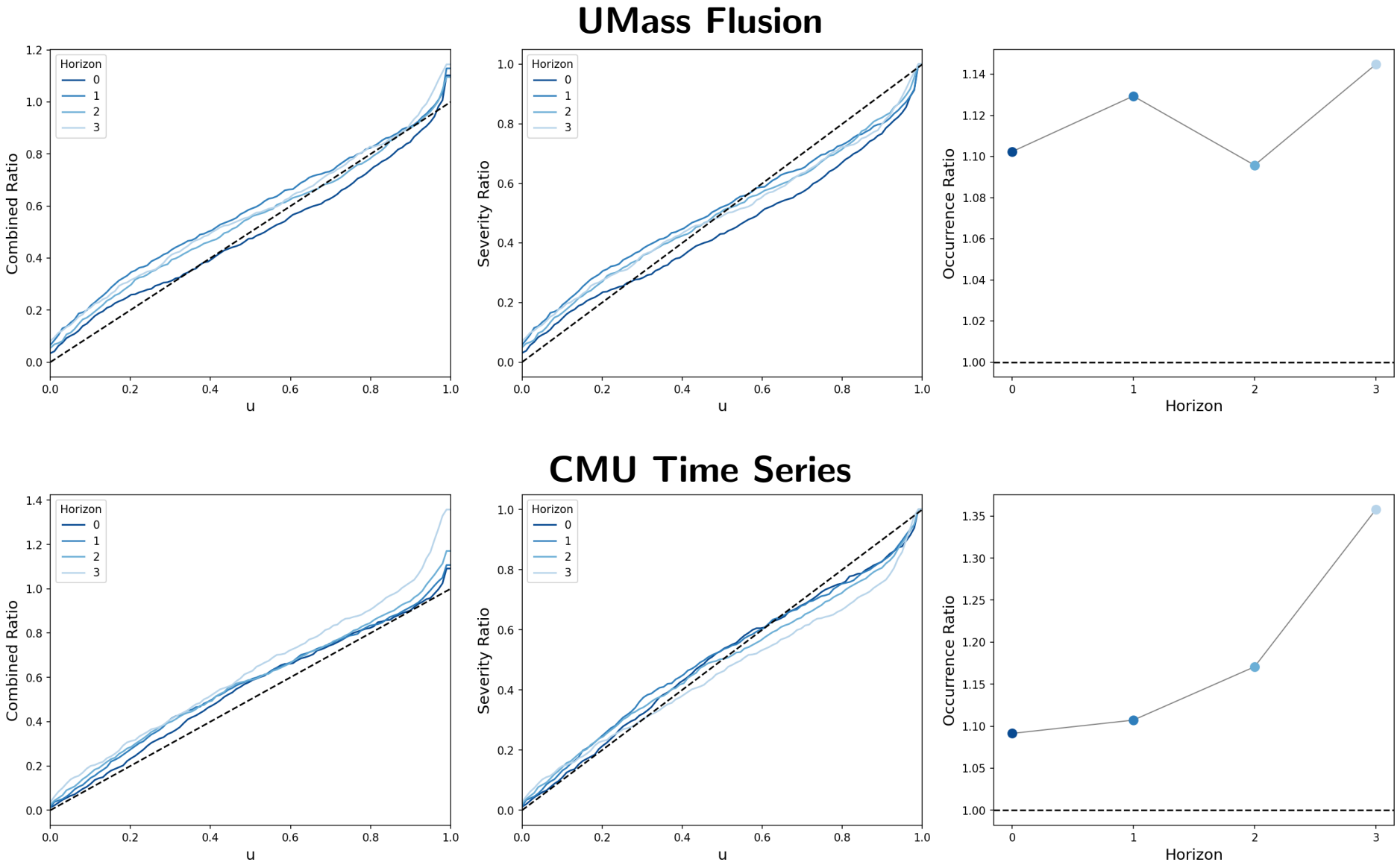


Figure: Tail Probabilistic Calibration diagnostics by horizon at $p = 75$.

Tail Calibration by Horizon: Pooled

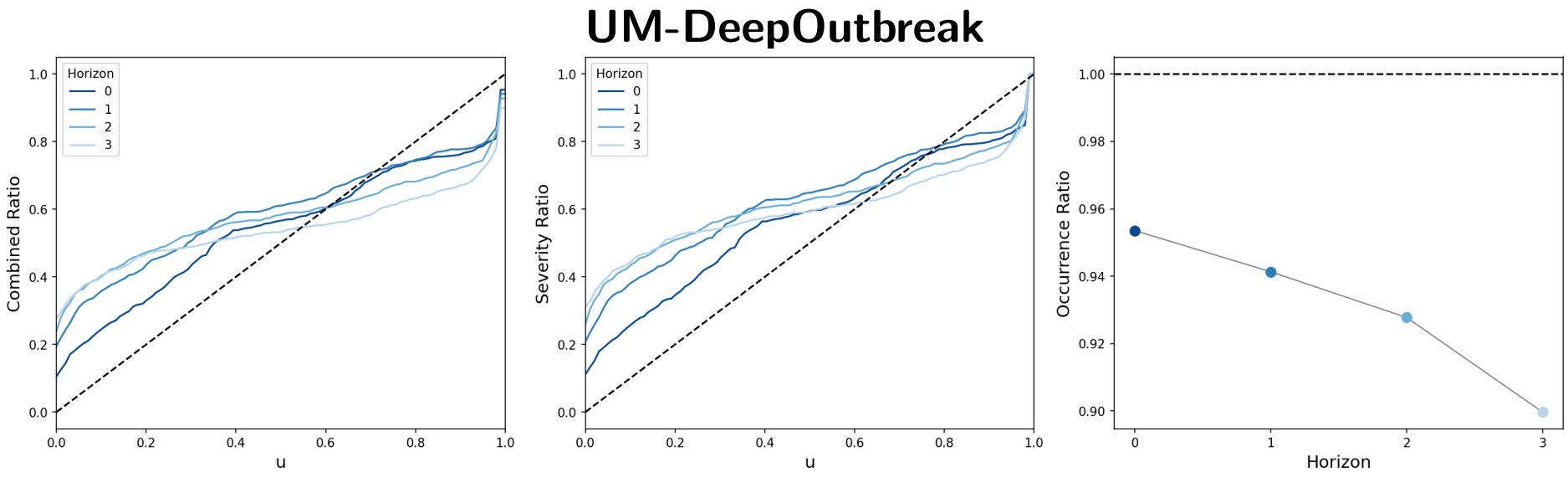


Figure: Tail Probabilistic Calibration diagnostics by horizon at $p = 75$.

- Interestingly, tail calibration appears to *imrpove* in this model.
- Worth noting it is a deep neural network with conformalized prediction intervals - RNNs have shown to be superior at long-horizon forecasting empirically.

How Well Does WIS Measure Tail Calibration?

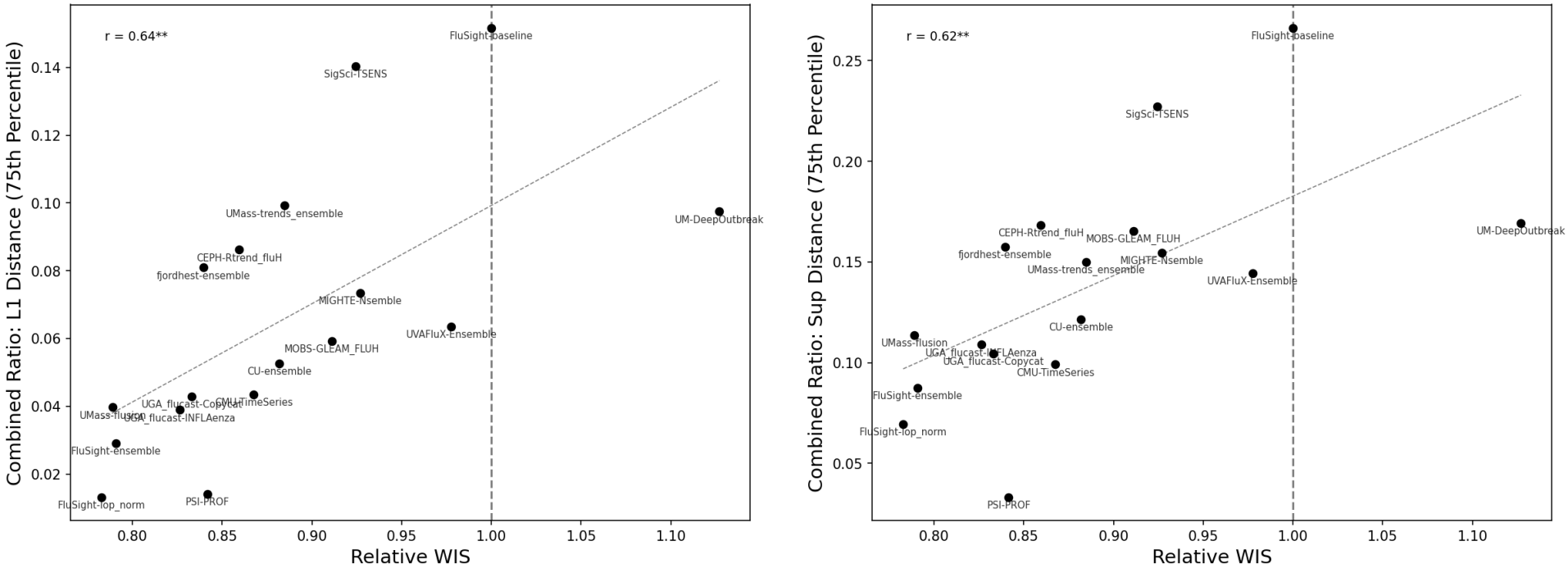


Figure: rWIS versus L1 $= \int_0^1 |\hat{R}_t(u) - u| du$ (left) and supremum $= \sup_{u \in [0,1]} |\hat{R}_t(u) - u|$) (right) distances of *combined* ratio for *same-week* forecasts, at $p = 75$.

- Note: large vertical distance between models reveals evidence between two models with similar scores and *different* tail behavior.

Extensions Proposed by Allen et al. (2025)

- One can consider post-hoc forecast calibration via conformal prediction to construct confidence sets with tail calibration guarantees. (i.e. tail *re-calibration* methods)
- Can extend framework to multivariate case, via multivariate EVT.
- How does one choose t? One must trade off taking $t \to x_Y$ with the resulting reduction in sample size.
- Simulations show that one needs $\approx 100,000$ observations for tail calibration of $p = 95+$ thresholds, even when $F = G$, for $Y \sim G$.

My Concluding Remarks

- Can we improve tail calibration at train time? Wessel et al. (2025) considers adding a tail miscalibration regularization term. [Details]
- We have already seen that scoring rules penalize forecasts asymmetrically (Buchweitz et al. (2025)).
- Tail calibration reveals another feature of a forecast to be aware of that a direct inspection of the proper score will miss.
- Still unclear to me if optimizing for tail calibration would improve detection of fast-changing environments - the diagnostics reveal that focusing on occurrence ratio could be promising.
- Formal and rigorous evaluation of tail calibration in the context of infectious disease forecasting context would be useful for future work.
- Still a tension between our ability to detect tail behavior empirically with the proposed definitions which are asymptotic.

Defining Tail Equivalence

Return to Slide 5

Definition (Tail Equivalence): Two random variables X and Y with CDF F and G, respectively, are *tail equivalent* if they have equal upper endpoints $x_F = x_G = x^*$ and their survival functions $\bar{F}$, $\bar{G}$ satisfy,

$$\lim_{x \to x^*} \frac{\bar{F}(x)}{\bar{G}(x)} \in (0, \infty).$$

Where $x_F := \sup\{x \in \mathbb{R} : F(x) < 1\}$, $\bar{F}(x) := 1 - F(x)$. Furthermore, G has a *heavier* tail than F if

$$\lim_{x \to x^*} \frac{\bar{F}(x)}{\bar{G}(x)} = 0.$$

Brehmer and Strokorb (2019) Theorem 5.4

Return to Slide 5

Let $S : \mathcal{F} \times \mathbb{R} \to \mathbb{R}$ a proper scoring rule, $Y \sim G$ with $G \in \mathcal{F}$. Then the following are true.

1. 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

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

2. Let $T : \mathcal{F} \to \mathbb{R}$ a *max-functional*. If there is an $F \in \mathcal{F}$ with $T(F) > T(G)$, then for all $\epsilon > 0$, there exists an $F_\epsilon \in \mathcal{F}$ such that $T(F_\epsilon) = T(F) > T(G)$, while

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

Quick Derivation: Conditional Forecast Excess Distribution

Return to Slide 8

We have:

$$\begin{aligned}\frac{P(Y > t+x|\ F)}{P(Y > t|\ F)} &= \frac{P(\{Y > t\} \setminus \{t < Y < t+x\}|\ F)}{P(Y > t|\ F)} \\ &= \frac{P(Y > t|\ F) - P(t < Y < t+x|\ F)}{P(Y > t|\ F)} \\ &= 1 - \frac{P(t < Y < t+x|\ F)}{P(Y > t|\ F)} \\ &= 1 - \frac{F(t+x) - F(t)}{1 - F(t)},\end{aligned}$$

recalling that under tail-calibration, $P(X \leq x|\ F) = F(x)$. This leads to,

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

Randomized Excess PIT

Return to Slide 9

- Let $Y \sim G$ a *discrete* random variable, then the *randomized PIT* of Y via the forecast F is defined

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

  with $\overset{*}{Z}_F \sim U[0, 1]$.
- The corresponding *excess PIT* of Y via F, at threshold t, reads

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

- The usual notions of tail calibration for discrete random variables hold with the above modification.

Training F to Enforce Tail Calibration

Return to slide 26

- Wessel et al. (2025) proposes a natural extension: during training time, add tail miscalibration regularization term such that:

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

  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 I

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

Brehmer, J. R. and Strokorb, K. (2019). Why scoring functions cannot assess tail properties.

Buchweitz, E., Romano, J. V., and Tibshirani, R. J. (2025). Asymmetric penalties underlie proper loss functions in probabilistic forecasting. *arXiv preprint arXiv:2505.00937*.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(2):243–268.

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

Wessel, J. B., Schillinger, M., Kwasniok, F., and Allen, S. (2025). Enforcing tail calibration when training probabilistic forecast models. *arXiv preprint arXiv:2506.13687*.