# Relative scoring rules favour baseline models and distort forecasters' incentives

Johannes Bracher[1,5], Sebastian Lerch[2,5], Marc-Oliver Pohle[3,5], and **Johannes Resin**[4,5]

[1]Karlsruhe Institute of Technology
[2]University of Marburg
[3]University of Wuppertal
[4]Goethe University Frankfurt
[5] Heidelberg Institute for Theoretical Studies

SWIM Topic Meeting: Evaluating Epidemic Forecasts

April 21, 2026, Heidelberg University

# Comparative forecast evaluation

- Consider two competing forecasts $F$ and $G$:
  - both of the same type,
  - e.g., point predictions, distributions, or multiple predictive quantiles/intervals.
- Baseline forecast $B$:
  - e.g., marginal quantity, seasonal climatology, simple/standard model
- Observation $y$
- We want to assign scores $\mathrm{S}(F, y)$ and $\mathrm{S}(G, y)$ to the forecasts that summarize predictive performance and admit a meaningful comparison via a **scoring rule** S.
- Scoring rules should be **proper** or **consistent** for the prediction type T in that

$$\mathbb{E}[\mathrm{S}(F, Y)] \leq \mathbb{E}[\mathrm{S}(G, Y)]$$

  if $Y \sim D$ and $F = \mathrm{T}(D)$ (Gneiting and Raftery, 2007; Gneiting, 2011).
- Propriety/Consistency ensures truthful forecasting,
  - no incentive to hedge the forecast, i.e., deviate from true beliefs.

# Examples of scoring rules

- Proper scoring rules for probabilistic forecasts (Gneiting and Raftery, 2007):
  - Key example: Continuous ranked probability score (CRPS)

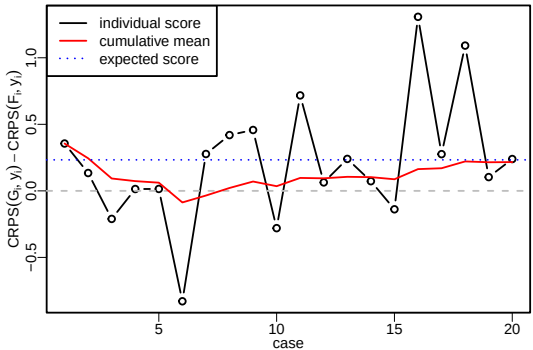
$$\mathrm{CRPS}(F, y) = \int (F(x) - \mathbb{1}\{x \geq y\})^2 \, \mathrm{d}x$$

  - Other examples: logarithmic score, Brier score, ...
- Consistent scoring functions for point forecasts (Gneiting, 2011):
  - The squared error $\mathrm{SE}(x, y) = (x - y)^2$ is consistent for the mean.
  - The absolute error $\mathrm{AE}(x, y) = |x - y|$ is consistent for the median.
  - The quantile score $\mathrm{QS}_\alpha(q, y) = 2(\mathbb{1}(y \leq q) - \alpha)(q - y)$
- Middle ground for multiple predictive quantiles/intervals (Bracher et al., 2021)
  - Quantile-weighted CRPS $\mathrm{qwCRPS}(F, y) = \frac{1}{k} \sum_{j=1}^{k} w_j \, \mathrm{QS}_{\alpha_j}(F^{-1}(\alpha_j), y)$ for $0 < \alpha_1 < \cdots < \alpha_k < 1$
  - Weighted interval score (WIS) arises if $\alpha_j = 1 - \alpha_{k+1-j}$

# Toy example

- Let $Y = X + \varepsilon$ with $X, \varepsilon \overset{\text{iid}}{\sim} \text{Norm}(0, 1)$.
- Consider two forecasts:
  - The ideal probabilistic forecast given $X$ is $F = \text{Norm}(X, 1)$.
  - The uninformed forecast $G = \text{Norm}(0, 2)$.
- Difference in expected CRPS:

$$\mathbb{E}[\text{CRPS}(G, Y)] - \mathbb{E}[\text{CRPS}(F, Y)] \approx 0.797 - 0.564 = 0.233$$



# Relative scores

- Scores often lack interpretability: What is a low (good) and a high (bad) score?
- Magnitude of the score may depend on outcome of interest.
- Individual scores may have different impact on the scores if we aggregate over outcomes of different scale.
- We may be tempted to scale scores by a reference score from a baseline forecast $B$ to obtain a **relative score**

$$\text{relS}_B(F, y) = \frac{\text{S}(F, y)}{\text{S}(B, y)}$$

  (assuming a non-negative scoring rule $\text{S} \geq 0$).
- Easy interpretation:
  - $\text{relS}_B(F, y) < 1$: $F$ is better.
  - $\text{relS}_B(F, y) > 1$: $B$ is better.

# Aggregating (relative) scores

Typically, individual scores $\mathrm{S}(F_i, y_i)$ from an evaluation sample $(F_1, y_1), \ldots, (F_n, y_n)$ are **averaged arithmetically** to estimate predictive performance:

$$\overline{\mathrm{relS}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{relS}_{B_i}(F_i, y_i)$$

With relative scores people tend to prefer **geometric averaging**:

$$\overline{\mathrm{relS}}_{\mathrm{geom}} = \left( \prod_{i=1}^{n} \mathrm{relS}_{B_i}(F_i, y_i) \right)^{\frac{1}{n}}$$

Alternatively, we can compute a **collective relative score** (or skill score)

$$\overline{\mathrm{relS}}_{\mathrm{coll}} = \frac{\frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(F_i, y_i)}{\frac{1}{n} \sum_{i=1}^{n} \mathrm{S}(B_i, y_i)}$$

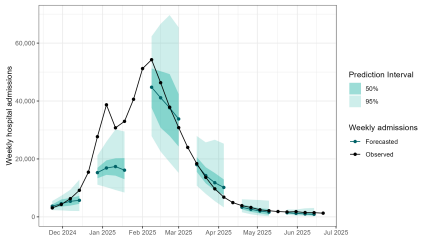
# CDC FluSight 2024–2025 Evaluation

https://www.cdc.gov/flu-forecasting/evaluation/2024-2025-report.html

*"Relative WIS was calculated using the geometric mean WIS of each model forecast compared to the geometric mean WIS of the corresponding FluSight baseline model forecast."*

$$\frac{\prod_i \text{WIS}(F_i, y_i)^{1/n}}{\prod_i \text{WIS}(B_i, y_i)^{1/n}} = \prod_i \left( \frac{\text{WIS}(F_i, y_i)}{\text{WIS}(B_i, y_i)} \right)^{1/n}$$

| Model | Relative WIS | 50% Coverage (%) | 95% Coverage (%) | % of Forecasts Submitted |
|---|---|---|---|---|
| FluSight-ensemble (ENS, STAT) | 0.62 | 52.9 | 82.92 | 100 |
| FluSight-lop_norm (ENS, STAT) | 0.62 | 61.03 | 94.23 | 100 |
| PSI-PROF_beta (MECH, STAT) | 0.62 | 58.28 | 89.73 | 93 |
| CU-ensemble (ENS, MECH, AI/ML, STAT) | 0.63 | 51.83 | 85.1 | 96 |



# ECDC RespiCast

(Gozzi et al., 2025)

Use another relative WIS:

$$\frac{1}{n}\sum_i \log_2\left(\frac{\mathrm{WIS}(F_i, y_i)}{\mathrm{WIS}(B_i, y_i)}\right) = \log_2\left(\prod_i \left(\frac{\mathrm{WIS}(F_i, y_i)}{\mathrm{WIS}(B_i, y_i)}\right)^{1/n}\right)$$


https://respicast.ecdc.europa.eu

# On displays of individual scores and inference

- Individual scores are very noisy and unreliable.
- Expected scores capture forecast accuracy (see definition of propriety/consistency).
- Averages over multiple observations are needed for scores to admit statistically sound inference.
- Formal inference for average scores via Diebold and Mariano (1995) type tests.

# Key literature on relative scores

- Comparison of point forecasts across multiple time series:
  - Armstrong and Collopy (1992) recommend geometric mean relative absolute error for point forecasts.
  - Hyndman and Koehler (2006) criticize that this and related measures of forecast accuracy "are not generally applicable, can be infinite or undefined, and can produce misleading results." They point out that the variance of individual scores becomes infinite if baseline errors can become arbitrarily small.
- Point forecasts and consistency:
  - Gneiting (2011) considers consistency of related scaled scoring functions (e.g., APE) and characterizes how these distort incentives.
- Probabilistic forecasts and hedging
  - Murphy (1973) demonstrates that individual skill scores incentivize hedging of forecast distributions in the case of the Brier score for categorical outcomes.

Our contribution: We generalize and substantiate arguments on the pitfalls of relative scores providing a general treatment of proper scoring rules and consistent scoring functions.

# Infinite expectations

**Proposition**. Let $S$ be a non-negative scoring function for point predictions such that $S(x, y) = 0$ if and only if $x = y$. Suppose that either

(a) $Y \sim D$ follows a discrete distribution with positive pmf $p_D(b) > 0$ at the baseline $b$, or

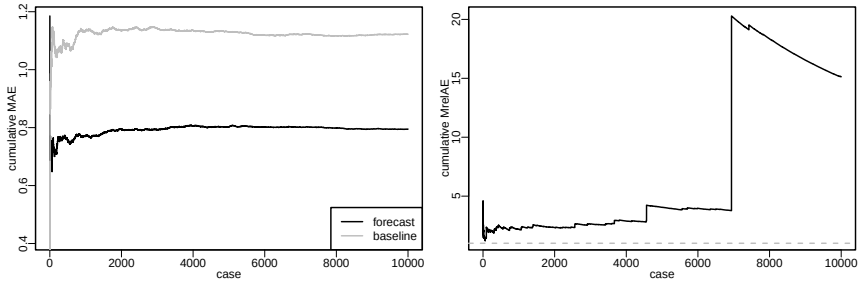
(b) $Y \sim D$ follows a continuous distribution with positive density $f_D(b) > 0$ at the baseline $b$, the function $y \mapsto S(x, y)$ is Lipschitz-continuous on $\mathbb{R}$, monotonically decreasing on $(-\infty, x)$ and monotonically increasing on $(x, \infty)$ for any $x$, and the density $f$ is continuous.

Then the expected relative score of a point prediction $x$ w.r.t. $b$ is given by

$$\mathbb{E}\,\mathrm{relS}_b(x, Y) = \begin{cases} 1, & \text{if } x = b, \\ \infty, & \text{otherwise.} \end{cases}$$

# Toy example: Point forecasts and infinite expectations

- Let $Y = X + \varepsilon$ with $X, \varepsilon \overset{\text{iid}}{\sim} \text{Norm}(0, 1)$.
- The absolute error $\text{AE}(x, y) = |x - y|$ is consistent for the median.
- Forecasts:
  - Ideal forecast given $X$: $F = \text{Median}(\text{Norm}(X, 1)) = X$
  - Baseline: $B = \text{Median}(\text{Norm}(0, 2)) = 0$
- The relative absolute error $\text{relAE}_B(F, Y) = \frac{|Y-X|}{|Y-0|}$ follows a folded Cauchy distribution with infinite expectation.



# Advantage of the baseline

**Proposition**. Consider two forecasters who provide predictions $F$ and $G$, respectively (either distributional or point forecasts). Suppose that $F, G, Y$ are random quantities on some shared probability space, and let $S$ be a non-negative, negatively oriented scoring rule.

(a) Under the natural regularity condition that $\mathbb{E}[S(F, Y)] > 0, \mathbb{E}[S(G, Y)] > 0$, it holds that

$$\mathbb{E}[\text{relS}_G(F, Y)] \geq \frac{\mathbb{E}[S(F, Y)]}{\mathbb{E}[S(G, Y)]} \geq \frac{1}{\mathbb{E}[\text{relS}_F(G, Y)]},$$

with equality only if $S(F, Y) = S(G, Y)$ almost surely.

(b) If $F$ and $G$ have equal predictive ability, i.e., $\mathbb{E}[S(F, Y)] = \mathbb{E}[S(G, Y)]$, then

$$\mathbb{E}[\text{relS}_G(F, Y)] \geq 1 \quad \text{and} \quad \mathbb{E}[\text{relS}_F(G, Y)] \geq 1.$$

Equality again holds only if $S(F, Y) = S(G, Y)$ almost surely.

# Toy example: Advantage of the baseline

- Let $Y = X_1 + X_2$ with $X_1, X_2 \overset{\text{iid}}{\sim} \text{Norm}(0, 1)$.
- Forecasts:
  - Ideal forecast given $X_1$: $F = \text{Norm}(X_1, 1)$
  - Ideal forecast given $X_2$: $G = \text{Norm}(X_2, 1)$
- Both forecasts are of equal skill with equal expected CRPS

$$\mathbb{E}[\text{CRPS}(F, Y)] = \mathbb{E}[\text{CRPS}(B, Y)] \approx 0.564$$

- However, the relative CRPS clearly favors the baseline:

$$\mathbb{E}\left[\frac{\text{CRPS}(F, Y)}{\text{CRPS}(B, Y)}\right] = \mathbb{E}\left[\frac{\text{CRPS}(B, Y)}{\text{CRPS}(F, Y)}\right] \approx 1.408$$

# Hedging individual relative scores

- Suppose the forecaster beliefs $Y \sim P$ and the benchmark $B$ is known.
- The forecaster is incentivized to issue a forecast that optimizes the expected score:

$$H = \arg\min_G \mathbb{E}_{Y\sim P}[\mathrm{relS}_B(G, Y)]$$

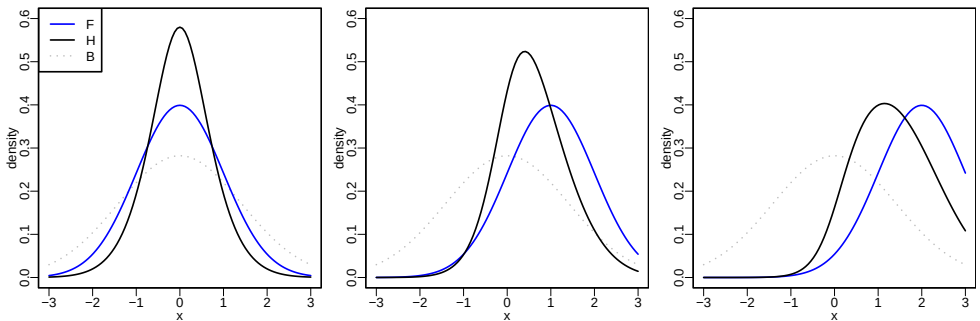
- If $P$ is continuous with density $f_P$, then

$$\mathbb{E}_{Y\sim P}[\mathrm{relS}_B(G, y)] = \int \frac{\mathrm{S}(G, y)}{\mathrm{S}(B, y)} f(y)\mathrm{d}y = c \int S(G, y) f_Q(y)\mathrm{d}y = c\mathbb{E}_{Y\sim Q}[\mathrm{S}(G, Y)],$$

  where $f_Q(y) \propto \frac{f_P(y)}{S(B,y)}$ and $c = \int \frac{f_P(y)}{S(B,y)}\,\mathrm{d}y$.
- Thus, instead of predicting $F = \mathrm{T}(P)$, the forecaster can hedge the score by predicting $H = \mathrm{T}(Q)$.

# Toy example: Hedging individual relative scores

- Let $Y = X + \varepsilon$ with $X, \varepsilon \overset{\text{iid}}{\sim} \text{Norm}(0, 1)$.
- Forecasts:
  - Ideal probabilistic given $X$: $F = \text{Norm}(X, 1)$.
  - Baseline: $B = \text{Norm}(0, 2)$.
- The hedged distribution $H$ tends to be more concentrated than $F$ and shifted towards the center of the baseline $B$:



# What about geometric averages of relative scores?

Geometric averages are improper:

- Suppose we take a geometric average over relative scores of $n$ iid pairs $(F_1, y_1), \ldots, (F_n, y_n)$:

$$\overline{\mathrm{S}}_n = \prod_{i=1}^{n} \mathrm{relS}(F_i, y_i)^{\frac{1}{n}}.$$

- The forecasts can be hedged by solving $H_i = \arg\min_G \mathbb{E}_{Y \sim F_i}[relS_{B_i}(G, Y)]$ for each $i$ (due to independence).
- Asymptotically, we have

$$\mathbb{E}\overline{\mathrm{S}}_n \longrightarrow \exp(\mathbb{E} \log \mathrm{S}(F_1, Y_1) - \mathbb{E} \log \mathrm{S}(B_1, Y_1)) \qquad \text{as} \quad n \to \infty.$$

- Thus, for large $n$, the forecaster can hedge by optimizing the logarithm of the score, $\log \circ \mathrm{S}$, which is known to be improper (Bosse et al., 2023).

# Asymptotic consistency of collective skill scores

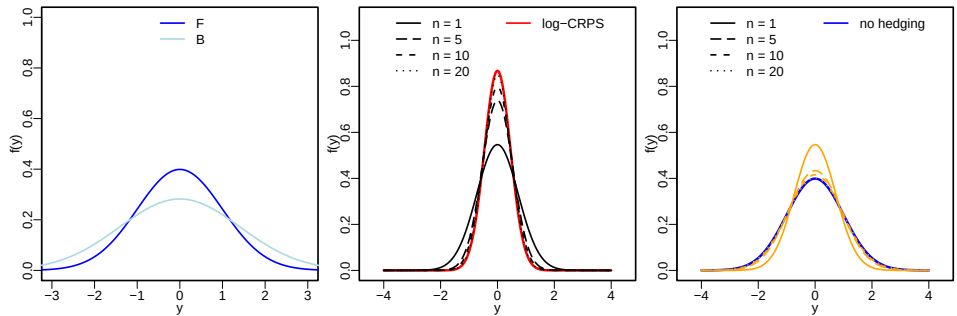
For the collective relative score, we have (subject to iid assumption)

$$\frac{\sum_i \mathrm{S}(F_i, y_i)}{\sum_i \mathrm{S}(B_i, y_i)} \longrightarrow \frac{\mathbb{E}\,\mathrm{S}(F_1, Y_1)}{\mathbb{E}\,\mathrm{S}(B_1, Y_1)} \qquad \text{as} \quad n \to \infty.$$

and thus asymptotic propriety (as the forecaster should optimize the numerator for sufficiently large samples).

# Toy example: Hedging geometric averages of relative score

- Let $Y = X + \varepsilon$ with $X, \varepsilon \overset{\text{iid}}{\sim} \text{Norm}(0, 1)$.
- Forecasts:
  - Ideal probabilistic given $X$: $F = \text{Norm}(X, 1)$.
  - Baseline: $B = \text{Norm}(0, 2)$.
- Figures:
  - Left: Forecast densities.
  - Middle: Hedged distributions for geometric average relative score.
  - Right: Hedged distributions for collective relative score.



# Conclusions

- Individual scores are misleading (very noisy). Averages over sufficiently large samples are needed to draw meaningful conclusions.
- Arithmetic averages of relative scores are very flawed.
- Geometric averages also distort incentives and encourage hedging (even in large samples).
- Collective relative scores (or skill scores) are asymptotically consistent and can be used to transform average scores to a nicely interpretable scale.
- Transforming outcomes (and predictions) with a logarithm can help to scale predictions such that scores are better comparable if variances grow with the outcome scale (Bosse et al., 2023).

# References


- Armstrong, J. and F. Collopy (1992). "Error measures for generalizing about forecasting methods: Empirical comparisons". *International Journal of Forecasting* 8.1, pp. 69–80.
- Bosse, N. I. et al. (2023). "Scoring epidemiological forecasts on transformed scales". *PLOS Computational Biology* 19.8, pp. 1–23.
- Bracher, J. et al. (2021). "Evaluating epidemic forecasts in an interval format". *PLOS Computational Biology* 17.2, pp. 1–15.
- Diebold, F. X. and R. S. Mariano (1995). "Comparing Predictive Accuracy". *Journal of Business and Economic Statistics* 13, pp. 253–263.
- Gneiting, T. (2011). "Making and evaluating point forecasts". *Journal of the American Statistical Association* 106.494, pp. 746–762.
- Gneiting, T. and A. E. Raftery (2007). "Strictly proper scoring rules, prediction, and estimation". *Journal of the American Statistical Association* 102.477, pp. 359–378.
- Gozzi, N. et al. (2025). "Performance evaluation of RespiCast ensemble forecasts for primary care syndromic indicators of viral respiratory disease in Europe during the 2023/24 winter season". *medRxiv*.
- Hyndman, R. J. and A. B. Koehler (2006). "Another look at measures of forecast accuracy". *International journal of forecasting* 22.4, pp. 679–688.
- Murphy, A. H. (1973). "Hedging and Skill Scores for Probability Forecasts". *Journal of Applied Meteorology and Climatology* 12.1, pp. 215 –223.