Metric	Explanation
CRPS (Continuous) ranked probability score	Proper scoring rule for continuous and integer forecasts that measures the distance between the CDF of the predictive distribution and the data-generating distribution. The CRPS is given as $$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(x) - 1(x \geq y)\right)^2 dx,$$ where y is the true observed value an F the predictive distribution. Often An alternative representation is used: $$\text{CRPS}(F, y) = \frac{1}{2}\mathbb{E}_F|X - X'| - \mathbb{E}_P|X - y|,$$ where X and X' are independent realisations from the predictive distributions F with finite first moment and y is the true value. In this represenation we can simply replace X and X' by samples sum over all possible combinations to obtain the CRPS. For integer-valued forecasts, the RPS is given as $$\text{RPS}(F, y) = \sum_{x=0}^{\infty} (F(x) - 1(x \geq y))^2.$$ **Usage**: Proper scoring rule, recommended in most instances, smaller values are better
Log score	The Log score is a proper scoring rule and is simply the log of the predictive density evaluated at the true observed value. It is given as $$\text{log score} = \log f(y),$$ where f is the predictive density function and y is the true value. **Usage and caveats**: Larger values are better, but sometimes the sign is reversed. Sensitive to outliers, as individual negative log score contributions can become very large if $f(y)$ is close to zero. In practice the log score is also hard to use for integer values, as a predictive density is required.

(continued)

Metric	Explanation
WIS (Weighted) interval score	Proper scoring rule for quantile forecasts. converges to crps for increasing number of interval. The score can be decomposed into a sharpness contribution and penalties for over- and underprediction. For a single interval, the score is computed as $$IS_\alpha(F, y) = (u - l) + \frac{2}{\alpha} \cdot (l - y) \cdot 1(y \leq l) + \frac{2}{\alpha} \cdot (y - u) \cdot 1(y \geq u)$$, where 1() is the indicator function, y is the true value, and l and u are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantile of F, i.e. the lower and upper bound of a single prediction interval. For a set of K prediction intervals and the median m, the score is computed as a weighted sum, $WIS = \frac{1}{K+0.5} \cdot (w_0 \cdot |y - m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F, y))$. w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$. **Caveat**: The wis is based on measures of absolute error. When averaging across multiple targets, it will therefore be dominated by targets with higher absolute values.
DSS Dawid-Sebastiani score	proper scoring rule for continuous and integer forecasts. The dss has a slightly simpler formula that only relies on the first moments of the predictive distribution. If in doubt, we would recommend the crps, but the difference should not be large in practice.
Brier score	Proper scoring rule for binary forecasts. Brier Score = $\frac{1}{N} \sum_{n=1}^{N} (\text{prediction}_n - \text{outcome}_n)^2$, where prediction
interval coverage	(Interval) coverage for a single prediction interval can be calculated as IC_α = nominal coverage − actual empirical coverage, where nominal coverage is $1 - \alpha$ and empirical coverage is the percentage of true values actually covered by the $1 - \alpha$ prediction intervals. Interval coverage can then be aggregated over all interval levels: Coverage deviation = $\frac{1}{K} \sum_{k=1}^{K} \text{IC}_{\alpha_k}$

(continued)

Metric	Explanation
Quantile coverage	Quantile coverage for a given quantile level is the percentage of true values smaller than the predictions corresponding to that quantile level.
Bias	For continuous forecasts, bias is given as $B(F, y) = 1 - 2 \cdot (F(y))$. For integer-valued forecasts, bias can be calculated as$B(F, y) = 1 - (F(y) + P_t(x_t + 1))$ and for quantile forecasts as $B =$ the maximum percentile rank that satisfies prediction smaller than y, if the true value is smaller than the median of the predictive distribution. If the true value is above the median of the predictive distribution, then B_t is the minimum percentile rank for which the corresponding quantile is still larger than the true value. If the true value is exactly the median, both terms cancel out and B_t is zero. For a large enough number of quantiles, the percentile rank will equal the proportion of predictive samples below the observed true value, and this metric coincides with the one for continuous forecasts.
Mean score ratio	The mean score ratio is used to compare two models on the overlapping set of forecast targets for which both models have made a prediction. It is calculated as the mean score achieved by the first model over the mean score achieved by the second model.
Mean score ratio	Relative skill is used to create a ranking between models based on pairwise comparisons between all models. To compute the relative skill of model m, we take the geometric mean of all mean score ratios that involve model m.