Metric	Explanation
CRPS (Continuous) ranked probability score	The crps is a proper scoring rule that generalises the absolute error to probabilistic forecasts. It measures the 'distance' of the predictive distribution to the observed data-generating distribution. The CRPS is given as $$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - 1(x \geq y))^2 \, dx,$$ where y is the true observed value and F the CDF of 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 and caveats** Smaller values are better. The crps is a good choice for most practical purposes that involve decision making, as it takes the entire predictive distribution into account. If two forecasters assign the same probability to the true event y, then the forecaster who assigned high probability to events far away from y will still get a worse score. The crps (in contrast to the log score) can at times be quite lenient towards extreme mispredictions. Also, due to it's similarity to the absolute error, the level of scores depend a lot on the absolute value of what is predicted, which makes it hard to compare scores of forecasts for quantities that are orders of magnitude apart.

(continued)

Metric	Explanation
Log score	The Log score is a proper scoring rule that is simply compuated as 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. For integer-valued forecasts, the log score can be computed as $$\text{log score} = \log p_y,$$ where p_y is the probability assigned to outcome p by the forecast F. **Usage and caveats**: Larger values are better, but sometimes the sign is reversed. The log score is ensitive to outliers, as individual negative log score contributions quickly can become very large if the event falls in the tails of the predictive distribution, where $f(y)$ (or p_y) is close to zero. Whether or not that is desirable depends ont the application. In scoringutils, the log score cannot be used for integer-valued forecasts, as the implementation requires a predictive density. In contrast to the crps, the log score is a local scoring rule: it's value only depends only on the probability that was assigned to the actual outcome. This property is desirable for inferential purposes, for example in a Bayesian context (@winklerScoringRulesEvaluation1996). In settings where forecasts inform decision making, it may be more appropriate to score forecasts based on the entire predictive distribution.

(continued)

Metric	Explanation
WIS (Weighted) interval score	The (weighted) interval score is a proper scoring rule for quantile forecasts that converges to the crps for an increasing number of intervals. The score can be decomposed into a sharpness (uncertainty) component 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 $1 - \frac{\alpha}{2}$ quantiles of the predictive distribution 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, $$WIS = \frac{1}{K + 0.5} \cdot (w_0 \cdot |y - m| + \sum_{k=1}^{K} w_k \cdot IS_\alpha(F, y)),$$ where w_k is a weight for every interval. Usually, $w_k = \frac{\alpha_k}{2}$ and $w_0 = 0.5$. **Usage and caveats**: Smaller scores are better. Applicable to all quantile forecasts, takes the entire predictive distribution into account. Just as the crps, 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. The decomposition into sharpness, over- and underprediction make it easy to interpret scores and use them for model improvement.

(continued)

Metric	Explanation
DSS Dawid-Sebastiani score	The Dawid-Sebastiani-Score is a proper scoring rule proposed by Gneiting and Raftery in [@gneitingStrictlyProperScoring2007] that only relies on the first moments of the predictive distribution and is therefore easy to compute. It is given as $$\text{dss}(F, y) = \left(\frac{y - \mu}{\sigma}\right)^2 + 2 \cdot \log \sigma,$$ where F is the predictive distribution with mean μ and standard deviation σ and y is the true observed value. textbfUsage and caveats The dss is applicable to continuous and integer forecasts and easy to compute. Apart from the ease of computation we see little advantage in using it. MAYBE SCRATCH IT ALTOGETHER FROM THE PAPER?
Brier score	Proper scoring rule for binary forecasts. The Brier score is computed as $$\text{Brier Score} = \frac{1}{N}\sum_{n=1}^{N}(f_n - y_n),$$ where f_n, with $n = 1, \ldots, N$ are the predicted probablities that the corresponding events, $y_n \in (0, 1)$ will be equal to one.) **Usage**: Applicable to all binary forecasts.

(continued)

Metric	Explanation
Interval coverage	Interval coverage measures the proportion of observed values that fall in a given prediction interval range. Interval coverage for a single prediction interval range can be calculated as $$IC_\alpha = \text{nominal coverage} - \text{empirical coverage},$$ where nominal coverage is $1 - \alpha$ and empirical coverage is the percentage of true values actually covered by all $1 - \alpha$ prediction intervals. To summarise interval coverage over different over multiple interval ranges, we can compute coverage deviation defined as the mean interval coverage over all K interval ranges α_k with $k = 1, \ldots, K$: $$\text{Coverage deviation} = \frac{1}{K} \sum_{k=1}^{K} \text{IC}_{\alpha_k}$$ **Usage**: Interval coverage for a set of chosen intervals, (e.g. 50% and 90%) gives a good indication of marginal calibration and is easy to interpret. Reporting coverage deviation has the advantage of summarising calibration in a single number, but loses some of the nuance.
Quantile coverage	Quantile coverage for a given quantile level is the percentage of true values smaller than the predictions corresponding to that quantile level. **Usage**: Quantile coverage is similar to interval coverage, but conveys more information. For example, it allows us to look at the 5% and 95% quantile separately, instead of jointly at the 90% prediction interval). This helps to diagnose whether it is the upper or lower end of a prediction interval that is causing problems. Plots of quantile coverage are conceptually very similar to PIT histograms.

(continued)

Metric	Explanation
Probability integral transform (PIT)	The probability integral transform (PIT) @dawidPresentPositionPotential1984 represents a succinct way to visualise deviations between the predictive distribution F and the true data-generating distribution G. The idea is to transform the observed values such that agreement between forecasts and data can then be examined by observing whether or not the transformed values follow a uniform distribution. The PIT is given by $$u = F(y),$$ where u is the transformed variable and $F(y)$ is the predictive distribution F evaluated at the true observed value y. If $F = G$, then u follows a uniform distribution (for a proof see e.g. @angusProbabilityIntegralTransform1994). For integer outcomes, the PIT is no longer uniform even when forecasts are ideal. Instead, a randomised PIT (@funkAssessingPerformanceRealtime2019) can be used: $$u = P(y) + v \cdot (P(y) - P(y-1)),$$ where y is again the observed value $P()$ is the cumulative probability assigned to all values smaller or equal to y (where $P(-1) = 0$ by definition, and v is a standard uniform variable independent of y. If P is equal to the true data-generating distribution function, then u is standard uniform. also propose a non-randomised version of the PIT for count data that could be used alternatively. **Usage**: One can plot a histogram of u values to look for deviations from uniformity. U-shaped histograms often result from predictions that are too narrow, while hump-shaped histograms indicate that predictions may be too wide. Biased predictions will usually result in a triangle-shaped histogram. One can also test for deviations from normality, using for example an Anderson-Darling test. This, however, proves to be overly strict in practice and even slight deviations from perfect calibration are punished in a way that makes it very hard to compare models at all. I HAVE MADE A SIMULATION TO TEST THAT IN MY MASTER THESIS AND COULD ADD IT TO THE PAPER?

(continued)

Metric	Explanation
Sharpness	Sharpness is the ability to produce narrow forecasts and is a feature of the forecasts only and does not depend on the observations. Sharpness is therefore only of interest conditional on calibration: a very precise forecast is not useful if it is clearly wrong. As suggested by @funkAssessingPerformanceRealtime2019, we measure sharpness for continuous and integer forecasts represented by predictive samples as the normalised median absolute deviation about the median (MADN)), i.e. $$S(F) = \frac{1}{0.675} \cdot \text{median}(\lvert x - \text{median(x)}\rvert),$$ where x is the vector of all predictive samples and $\frac{1}{0.675}$ is a normalising constant. If the predictive distribution F is the CDF of a normal distribution, then sharpness will equal the standard deviation of F. For quantile forecasts we can directly use the sharpness component of the weighted interval score. Sharpness is then simply the weighted mean of the widths of the central prediction intervals.

(continued)

Metric	Explanation
Bias	Bias is a measure of the tendency of a forecaster to over- or underpredict. For continuous forecasts, bias is given as $$B(F, y) = 1 - 2 \cdot (F(y)),$$ where F is the CDF of the predictive distribution and y is the observed value. For integer-valued forecasts, bias can be calculated as $$B(P, y) = 1 - (P(y) + P(y + 1)),$$ where $P(y)$ is the cumulative probability assigned to all outcomes smaller or equal to y. For quantile forecasts, Bias can be calculated as the maximum percentile rank for which the prediction is 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 bias 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, bias 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. **Usage**: In contrast to the over- and underprediction penalties of the interval score it is bound between 0 and 1 and represents the tendency of forecasts to be biased rather than the absolute amount of over- and underprediction. It is therefore a more robust measurement, but harder to interpet. It largely depends on the application whether one is more interested in the tendency to be biased or in the absolute value of over- and underpredictions.

(continued)

Metric	Explanation
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. The mean score ratio is calculated as the mean score achieved by the first model over the mean score achieved by the second model. More precisely, for two models i, j, we determine the set of overlapping forecasts, denoted by $\mathcal{A}_{ij}$ and compute the mean score ratio θ_{ij} as $$\theta_{ij} = \frac{\text{mean score model } i \text{ on } \mathcal{A}_{ij}}{\text{mean score model } j \text{ on } \mathcal{A}_{ij}}.$$ The mean score ratio can in principle be computed for any arbitrary score. **Usage**: Mean scores ratios are usually calculated in the context of pairwise comparisons, where a set of models is compared by looking at mean score ratios of all possible parings. Whether smaller or larger values are better depends on the orientation of the original score used

(continued)

Metric	Explanation
Relative skill	Relative skill scores can be used to obtain a ranking of models based on pairwise comparisons between all models. To compute the relative skill θ_i of model i, we take the geometric mean of all [mean score ratios](mean $$\theta_i = \left(\prod_{m=1}^{M} \theta_{im} \right)^{1/M},$$ where M is the number of models. **Usage and caveats**: Relative skill is a helpful way to obtain a model ranking. Whether smaller or larger values are better depends on the orientation of the original score used. It is in principle relatively robust against biases that arise when models only forecast some of the available targets and is a reasonable way to handle missing forecasts. One possible precautionary measure to reduces issues with missing forecasts is to only compare models that have forecasted at least half of all possible targets (this ensures that there is always an overlap between models). If there is no overlap between models, the relative skill implicitly estimates how a model would have forecasted on those missing targets.