Rillor

EXPLAINER

Brier score and log loss for probability forecasts

The Brier score and log loss both reward forecasters for stating the probabilities they believe, but they penalize confident misses differently, and the choice can reorder rankings.



Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

- The Brier score and log loss are both strictly proper: a forecaster gets the best expected score only by reporting the probability they believe [4, 5, 17]. The linear score and mean absolute error are not [5, 23, 26].
- They differ most on confident misses. The Brier penalty for one binary forecast is at most 1 on the 0 to 1 scale, while log loss rises without limit as the probability on the outcome falls toward zero [8, 12]. Among 10 forecasters, the Brier and log rankings had a Spearman correlation of 0.15, and one forecaster moved from 3rd to 10th [19].
- On rare events the Brier score gives little credit for a correct small probability. For an event with a frequency of 0.001, forecasting 0.001 improves on forecasting 0 by 0.1% in expected Brier score [9]. The Brier score prefers a forecast one thousand times too small to one four times too large, and the log score prefers the reverse [20].
- Murphy's 1973 decomposition splits the Brier score into reliability, resolution and uncertainty [2], and proper scores in general split the same way [10, 11]. A lower Brier score does not by itself show better calibration [13].
- Published Brier scores use a 0 to 2 scale (Brier's original, Good Judgment Open) or a 0 to 1 scale (ForecastBench, scikit-learn's binary default) [1, 14, 22, 24]. Skill scores read a score against a reference and give up strict propriety in small samples [3, 5, 12].

Background

Glenn Brier, at the U.S. Weather Bureau, set out the problem of scoring probability forecasts in 1950. Under arbitrary scores, he wrote, a forecaster may choose "to ignore the verification system or to let it do the forecasting for him by 'hedging' or 'playing the system.'" [1]. His score for probability statements was built so that the forecaster "is encouraged to state unbiased estimates of the probability of each event when he cannot forecast perfectly" [1].

Gneiting and Raftery trace the name: "The term proper was apparently coined by Winkler and Murphy (1968, p. 754), whereas the general idea dates back at least to Brier (1950) and Good (1952, p. 112)" [5]. Good's proposal is the logarithmic score [5]. Murphy later separated three kinds of goodness: consistency between a forecaster's judgment and the forecast, quality (the match with observations) and value (the benefit to the people who use the forecast) [4]. Strictly proper rules reward consistency, and accuracy scores such as the Brier score measure quality [4]. Value depends on the decisions the forecasts feed [4, 12].

Both scores now run well beyond weather: scikit-learn lists them as strictly consistent scoring functions for class probabilities [13], Good Judgment Open reports Brier scores [22], Metaculus builds on the log score [23] and ForecastBench scores LLM forecasts with the Brier score [24].

Method

What was examined. For each score: its definition and orientation, whether it is proper, how it treats confident misses and rare events, how it decomposes, the scales and averaging used where it is published, skill scores against a reference, and extensions to continuous and quantile forecasts.

Sources. The original papers by Brier and Murphy [1, 2, 3, 4], theory papers [5, 6, 7, 8, 9, 10, 11, 27], empirical studies [18, 19, 20, 21], documentation from the WWRP/WGNE Joint Working Group on Forecast Verification Research, scikit-learn 1.9.1 and scoringutils [12, 13, 14, 15, 16, 17], a textbook [28], and platforms and benchmarks [22, 23, 24, 25, 26].

How. Every formula comes from a cited source, with its scale, orientation and log base stated. Study figures are quoted as published. Values that appear in no source are arithmetic from the cited formulas and marked "computed". Three of Murphy's papers [2, 3, 4] were read from scanned PDFs and Brier's [1] from an imperfect text layer, so their formulas were checked against the papers' verbal descriptions. Where sources disagree, both are reported.

Findings

1. What makes a scoring rule proper

"A scoring rule is proper if the forecaster maximizes the expected score for an observation drawn from the distribution F if he or she issues the probabilistic forecast F, rather than G ≠F. It is strictly proper if the maximum is unique" [5]. Gneiting and Raftery write scores as rewards, higher being better. Most practical sources write them as penalties, lower being better [12, 13, 17, 22]. Propriety holds in either orientation.

Murphy's proof for the Brier score, as a penalty on the 0 to 1 scale: a forecaster who believes p and reports f receives $(f - 1)^2$ if the event occurs and f^2 if not. The expected score "can be rewritten as EBS = p(1 - p) + (f - p)^2", which "is minimized by choosing f = p" [4]. The extra cost of misreporting, $(f - p)^2$, is the same above and below the judgment. Log loss is also minimized in expectation only by the true distribution [18], with an extra cost that is not symmetric (Finding 4).

Rules that fail. The linear score, which pays the probability given to the outcome, "is not a proper scoring rule, despite its intuitive appeal" [5]. In a Metaculus example, if a six has a true probability of 17%, the average linear score is 0.8 for a forecast of 5%, 0.72 for 17% and 0.63 for 30%. The average log scores are -0.54, -0.45 and -0.51, best at the true 17% [23]. Mean absolute error "incentivizes reporting the median of the outcome distribution (i.e., an extreme 0 or 1 forecast), not the true probability p" [26].

Why it matters for models. Bröcker and Smith argue that "only proper scores provide internally consistent evaluations of probability forecasts, justifying the focus on proper scores independent of any attempt to influence the behavior of a forecaster" [7]. An improper score can rank a forecast above another even when the data were drawn from the other, and "The question of whether the employed score is proper or not can be answered before any data are considered" [7]. Serafini and colleagues prove that the parimutuel gambling score, used to rank earthquake forecasts, is improper when more than two forecasts are tested [20].

Rescaling and locality. Multiplying a proper score by a positive constant and adding a term that depends only on the outcome keeps it proper [5]. The Forecasting Research Institute's Brier Index, (1 - √Brier score) × 100% with the Brier score on the 0 to 1 scale, is a nonlinear transformation [26]. In a 5 March 2026 addendum the institute corrected its earlier claim that the index is proper, noting that "monotone nonlinear transformations need not preserve properness". It applies the index after averaging, so rankings match the Brier score's [26]. The log score is local, depending only on the probability given to the outcome [17], and Gneiting and Raftery report Bernardo's result that, under regularity conditions, every proper local rule is equivalent to it [5]. The Brier score also uses the probabilities of outcomes that did not happen, which Benedetti counts against it [9].

2. Brier score decomposition: reliability and resolution

Brier's score averages, over occasions, the sum over all categories of $(f - E)^2$, where E is 1 for the category that occurred and 0 otherwise [1]. It "has a minimum value of zero for perfect forecasting and a maximum value of 2 for the worst possible forecasting" [1]. His ten rain forecasts scored 0.19 on this 0 to 2 scale, against

0.42 for a constant forecast of the observed 0.3 and 0.44 for a climatological 0.2 [1]. Most current sources score a binary event as $(f - o)^2$ averaged over forecasts, on a 0 to 1 scale [12, 24], which is half of Brier's sum (computed). scikit-learn calls it "equivalent to the mean squared error" [13].

Murphy's 1973 partition splits the score into "a measure of the uncertainty inherent in the events", "a measure of the reliability of the forecasts" and "a new measure of the resolution of the forecasts" [2]. The score equals uncertainty plus reliability minus resolution. Reliability is the weighted squared gap between each stated probability and the frequency observed when it was stated. Resolution is the weighted squared gap between those observed frequencies and the overall frequency: "the smaller this term the smaller the resolution and the greater the PS" [2]. Uncertainty is the score of forecasting the sample frequency every time [2]. The verification FAQ writes the binary form, with uncertainty $\bar{y}(1 - \bar{y})$ for an observed frequency $\bar{y}$ [12]. Murphy's ten-forecast example gives 0.480 + 0.136 - 0.330 = 0.286 on the 0 to 2 scale [2], or 0.143 on the 0 to 1 scale (computed). On the 0 to 1 scale, uncertainty is 0.25 for a 50% event, 0.09 for a 10% event and 0.0099 for a 1% event (computed from [12]).

Brier saw the two parts in 1950: a good match between forecast and observed probabilities "is not sufficient to indicate how useful the forecasts are, for it is also necessary to know the frequency with which forecasts are made in the various categories" [1]. Bröcker extends the decomposition to every strictly proper score for finite-valued targets, the ignorance (log) score included [10]. Dimitriadis, Gneiting and Jordan note that "Murphy's decomposition is exact in the discrete case, but fails to be exact under continuous forecasts". Their CORP method recalibrates forecasts with the pool-adjacent-violators algorithm and yields miscalibration (MCB), discrimination (DSC) and uncertainty (UNC) terms that are "guaranteed to be non-negative" [11]. Their worked example scores 24-hour precipitation forecasts for Niamey, Niger, from July to September 2016.

Forecast	Mean Brier score (0 to 1)	MCB	DSC	UNC
ENS: 52-member ensemble	0.266	0.066	0.044	0.244
EPC: extended probabilistic climatology (reference)	0.234	0.022	0.032	0.244
EMOS: statistically postprocessed ENS	0.232	0.018	0.030	0.244
Logistic: data-driven statistical forecast	0.206	0.017	0.056	0.244

Source: [11].

The ensemble scored worse than the climatology-based reference because its miscalibration term was three times as large. Postprocessing cut that term to 0.018, and the logistic forecast combined the smallest miscalibration with the most discrimination [11]. scikit-learn warns that "A lower Brier loss, for instance, does

not necessarily mean a better calibrated model, it could also mean a worse calibrated model with much more discriminatory power" [13].

3. The log score and confident misses

Gneiting and Raftery write the logarithmic score as the log of the probability given to the outcome, higher being better [5]. scikit-learn's log loss is its negative, averaged: for a binary outcome, -(y log p + (1 - y) log(1 - p)). It is "also called logistic regression loss or cross-entropy loss", and "The logarithm used is the natural logarithm (base-e)" [13, 15]. Guo and colleagues note that negative log likelihood "is also referred to as the cross entropy loss in the context of deep learning" [18]. Roulston and Smith's ignorance score uses base 2, so it counts bits, and "Each bit of ignorance represents a factor-of-2 increase in uncertainty" [8]. A 50% forecast scores 0.693 nats or exactly 1 bit (computed).

The score is unbounded. "A forecast might receive a score of -∞ if an event claimed to be impossible is realized" [5]. Roulston and Smith advise replacing zero probabilities with small ones, since "Not to do so means reporting the improbable as the impossible" [8]. Merkle and Steyvers: "if one incorrectly makes a forecast of 0 or 1, then one's average score can never recover" [19]. Metaculus calls its log score "much more punitive of extreme wrong predictions than it is rewarding of extreme right predictions": moving from 99% to 99.9% gains 0.009 if right and costs 2.3 if wrong [23].

Critics point to the same feature. Gneiting and Raftery report that the log score "has been criticized for its unboundedness" [5], and Gneiting, Balabdaoui and Raftery say it "lacks robustness" [6]. Gneiting and Raftery add that it "involves a harsh penalty for low probability events and thus is highly sensitive to extreme cases" [5]. scoringutils notes that log scores "can easily be dominated by a few extreme values" [17]. Benedetti argues the other side, deriving the log score as the single measure that is additive, local and strictly proper [9].

Practice varies at the extremes. scikit-learn clips probabilities at machine precision, so its log loss is finite [15]. Metaculus "restricts binary predictions to be between 0.1% and 99.9%", which bounds its worst binary Baseline score at -897 [23]. The FluSight Challenge truncated the log score at -10, and Bracher and colleagues note that "when this truncation is performed, the score is no longer proper" [27].

In machine learning, Guo and colleagues found that "modern neural networks, unlike those from a decade ago, are poorly calibrated", and that networks "can overfit to NLL without overfitting to the 0/1 loss". On CIFAR-100, "test error drops from 29% to 27% in the region where NLL overfits" [18]. Temperature scaling, a one-parameter recalibration, was "surprisingly effective" [18].

4. Where the two disagree: overconfidence and rare events

The first table works single binary forecasts through both scores.

Forecast for "yes"	Outcome	Brier, 0 to 1	Brier, 0 to 2	Log loss, nats	Log loss, bits
0.50	Either	0.250	0.500	0.693	1.000
0.70	Yes	0.090	0.180	0.357	0.515
0.70	No	0.490	0.980	1.204	1.737
0.90	No	0.810	1.620	2.303	3.322
0.99	No	0.980	1.960	4.605	6.644
0.999	No	0.998	1.996	6.908	9.966

Computed from the formulas in [8, 12, 13, 15]. The values match Good Judgment Open's 0.18 example [22], ForecastBench's 0.25 for a 50% forecast [24] and Metaculus's -4.6 and -6.9 for confident misses [23].

From 0.70 to 0.99 on a miss, the Brier penalty doubles and the log loss nearly quadruples. At 1.0 the Brier penalty stops at 1 on the 0 to 1 scale and the log loss is infinite [12].

The second table shows expected scores when the true probability is 0.70 (computed from [4] and [13]).

Reported	Expected Brier, 0 to 1	Excess	Expected log loss, nats	Excess
0.55	0.2325	0.0225	0.658	0.047
0.70	0.2100	0	0.611	0
0.85	0.2325	0.0225	0.683	0.072
0.99	0.2941	0.0841	1.389	0.778

Both scores are lowest at the truth. The Brier excess is symmetric, and the log excess is larger on the overconfident side. scoringutils states the pattern and reaches similar figures by simulation: log-score differences of 0.0717 for 0.85 and 0.0474 for 0.55 [17].

Rankings. Merkle and Steyvers compared "10 forecasters on 21 binary problems" from a public forecasting site. The Spearman correlation between Brier and log rankings was 0.15. One forecaster ranked 3rd under the Brier score and 10th under the log score because that forecaster "tended to report extreme forecasts,

and the incorrect forecasts were heavily penalized under the logarithmic score". Four others moved four places [19]. Sources differ on how often this matters. Researchers "often find that one's choice of strictly proper scoring rule has minimal impact", Merkle and Steyvers write, but they conclude that "conclusions vary greatly across different scoring rules", noting that "only 10 forecasters were involved" [19].

Rare events. Serafini and colleagues studied binary earthquake forecasts with a true probability of 0.001, "comparable to the estimated probability of having an event with magnitude greater than 5.5 calculated the days before the L'Aquila earthquake" [20]. The log score weighs relative differences and "heavily penalizes overconfidence". The Brier score weighs absolute differences, so a forecast of 0 is preferred to any forecast above twice the true probability [20]. In one comparison the Brier score preferred a forecast one thousand times smaller than the truth over one four times larger, and the log score preferred the reverse [20]. Benedetti, on the 0 to 2 scale: for frequency 0.001, a forecast of 0 scores 0.002 in expectation and the correct 0.001 scores 0.002 × 0.999, "a gain of only 0.1% in score". He calls the Brier score "very unfair for evaluating rare (or common) events forecasts" [9]. The verification FAQ agrees that "the more rare an event, the easier it is to get a good BS without having any real skill" [12]. Always stating the base rate of a 1% event scores 0.0099 on the 0 to 1 scale (computed).

Serafini and colleagues tie the choice to the use. For an alarm system, "being overconfident may put lives at risk and perhaps the log score would be the right choice", while "the Brier score may be suitable when being overconfident is less dangerous" [20].

5. Reading published Brier scores

A published Brier score needs its scale, its orientation, its averaging and its reference before it can be read.

Source	Brier scale	Better is	How it is built
Brier 1950 [1], **Murphy 1973** [2]	0 to 2	Lower	Sum over all categories
Gneiting and Raftery 2007 [5]	-2 to 0 (computed)	Higher	Negative of the sum
Verification FAQ [12]	0 to 1	Lower	Binary $(f - o)^2$
scikit-learn [13, 14]	0 to 1 binary by default, 0 to 2 multiclass	Lower	scale_by_half option, added in version 1.7
Good Judgment Open [22]	0 to 2	Lower	Sum over options, daily average
ForecastBench paper [24]	0 to 1	Lower	Binary $(f - o)^2$

Source	Brier scale	Better is	How it is built
ForecastBench leaderboard [25, 26]	Brier Index, 0% to 100%	Higher	Difficulty-adjusted, then transformed

scikit-learn's default "auto" setting rescales to 0 to 1 "only for binary classification (as customary)". The same four binary forecasts score 0.0375 by default and 0.075 with scale_by_half set to False [14].

Good Judgment Open sums squared differences over every answer option, so a 70% forecast that resolves yes scores $(1 - 0.7)^2 + (0 - 0.3)^2 = 0.18$ on its 0 to 2 scale [22]. It scores each day with an active forecast, carries a forecast forward until it is updated and averages the daily scores. Its Relative Brier Score subtracts the crowd's average daily median score and multiplies by the user's participation rate, so negative values beat the crowd [22].

Metaculus time-averages log-based scores. When a question resolves early, the remaining time counts with a score of 0, because "To preserve properness, it is imperative that each day is weighted the same in the final average (or at least that the weights be decided in advance)" [23].

The ForecastBench paper uses $(f - o)^2$, noting that "a score of 0.25 is associated with the uninformed forecast of 0.5" [24]. On 200 standard questions at 7, 30, 90 and 180-day horizons, the superforecasters' median forecast scored 0.096 (95% interval 0.076 to 0.116), the public's median forecast 0.121 and the top model 0.122, on the 0 to 1 scale [24]. "The overall score is calculated as the average of the mean dataset Brier score and the mean market Brier score", unresolved market questions were scored against community aggregates, and the top models had the crowd forecast on market questions as an input [24]. The leaderboard now reports a "difficulty-adjusted Brier score" converted to a Brier Index, adjusting for question difficulty with a two-way fixed-effects model because the human groups answered different questions [25]. The institute's March 2026 post gives 0.086 for superforecasters (Brier Index 70.6%) and 0.103 for the best models (67.9%) [26]. These and the paper's 0.096 rest on different questions and methods [24, 26].

6. Skill scores against a reference forecast

The generic skill score is (score of the forecast - score of the reference) / (score of a perfect forecast - score of the reference). It is 1 for an optimal forecast and 0 for the reference, and negative values mean worse than the reference [5]. "The reference forecast is typically a climatological forecast", and climatological forecasts "are calibrated by construction, but often lack sharpness" [5]. The verification FAQ gives BSS = 1 - BS / BS of the reference, ranging from minus infinity to 1, with persistence or climatology as the usual reference [12].

Skill makes events of different difficulty comparable. A Brier score of 0.081 on the 0 to 1 scale is a skill of 0.10 against climatology for a 10% event (reference 0.09) and 0.68 for a 50% event (reference 0.25) (computed from [12]).

Skill scores give up strict propriety. The verification FAQ calls the BSS "Not strictly proper. Unstable when applied to small data sets" [12]. Murphy showed in 1973 that two skill versions of the Brier score encourage hedging, one of them then used by the National Weather Service, but that the version defined on collections of forecasts "is approximately strictly proper for large collections of forecasts ($K \geq 100$)" [3]. Gneiting and Raftery call skill scores "generally improper, even if the underlying scoring rule S is proper", note that "Little is known about the propriety of skill scores", and report Briggs and Ruppert's view that the departures "do little harm" [5]. Scores compare directly only "if they refer to exactly the same set of forecast situations" [5].

On the log scale, skill is a difference: the skill of a forecast relative to another can be found "by simply calculating the difference of the ignorance of the forecasts" [8]. Metaculus's Baseline score rescales the log score so that equal probabilities on all outcomes score 0 and a perfect binary forecast scores +100. An 80% forecast scores +68 on yes and -132 on no [23]. No source examined gives a level of Brier score or log loss that counts as good in general. Each reads a score against a reference [5, 12, 22, 23, 24].

7. Calibration and sharpness, read together

"Calibration refers to the statistical consistency between the distributional forecasts and the observations ... Sharpness refers to the concentration of the predictive distributions and is a property of the forecasts only" [5]. The goal is "maximizing the sharpness of the predictive distributions subject to calibration" [5, 6]. A climatology forecast "has no sharpness" [12]. In a simulation by Gneiting, Balabdaoui and Raftery, calibration histograms for an ideal forecaster and three flawed competitors were "essentially uniform", so the check "cannot distinguish between the ideal forecaster and her competitors". Both proper scores they report ranked the ideal forecaster highest [6].

"The reliability diagram plots the observed frequency against the forecast probability", with forecasts divided into bins. Points below the diagonal indicate "overforecasting (probabilities too high)", and a flatter curve means less resolution [12]. scikit-learn builds the same curve by binning [16]. Dimitriadis, Gneiting and Jordan show that "choices of m = 9, 10, or 11 equidistant bins yield drastically distinct reliability diagrams", and replace the bins with recalibration [11].

Other common measures check something else. The ROC curve "is not sensitive to bias in the forecast, so says nothing about reliability" [12]. Misclassification error is proper but not strictly proper, since only the side of one half matters [11]. A forecaster who says 51% for every event that happens and 49% for every event that does not is always on the right side but has a Brier Index of 51% [26].

8. Scores for continuous and quantile forecasts

The continuous ranked probability score (CRPS) "corresponds to the integral of the Brier scores for the associated binary probability forecasts at all real-valued thresholds" [5, 17]. It is reported in the units of the observations and reduces to the absolute error for a single-value forecast [5, 12]. Gneiting and Raftery found

it "less sensitive to extreme cases or outliers" than the log score [5]. Bracher and colleagues frame the trade-off as with Brier and log: the CRPS "does not diverge to ∞ even if a forecast assigns zero probability to the eventually observed outcome", and "in certain contexts the CRPS may seem lenient" [27].

The interval score adds the width of a central interval to a penalty, scaled by 2/α, for observations outside it [5]. Proper quantile scores exist at each level [5]. The quantile score is "sometimes called the 'pinball loss function'" and equals the absolute error at the median, and averaging quantile scores over all levels gives the CRPS [28]. The weighted interval score combines the median's absolute error with interval scores. It is "proper for any set of non-negative (un-normalized) weights", and with each interval weighted by half its alpha (and 1/2 on the median) over many equally spaced levels it approximates the CRPS [27].

Implications

For teams that score their own forecasts. Pick the score before seeing results, since propriety "can be answered before any data are considered" [7]. Match it to the cost of a confident miss [20]. Where the choice is open, report both and check whether rankings change [19]. State scale, orientation and log base, compare with a reference on the same questions [5, 12], and report decomposition terms with the total [2, 11, 13]. Put error bounds on scores and differences, "especially important (a) for rare events where the sample size is typically small" [12]. Serafini and colleagues warn that treating score differences as independent misleads when events cluster, and point to a Diebold-Mariano test [20]. Rillor's forecasting and market intelligence (https://rillor.com/markets/forecasting) work checks stated probabilities against observed frequencies and compares each forecast with simple baselines, such as no change or the historical rate.

For readers of published scores. For a binary question, a Brier score of 0.18 on Good Judgment Open's 0 to 2 scale is 0.09 on ForecastBench's 0 to 1 scale [22, 24]. Check the scale, question set, averaging, any adjustment or transformation, and the reference before comparing [22, 23, 24, 25, 26]. A low raw Brier score on rare events can come from forecasting the base rate [12].

For machine learning teams. Log loss, negative log likelihood and cross-entropy are one quantity [13, 18]. Accuracy can improve while log loss worsens [18]. scikit-learn recommends using the same strictly consistent function for training and evaluation [13].

For decision makers. A proper score measures quality, which differs from value [4, 12]. "When the cost of a missed event is high, the deliberate overforecasting of a rare event may be justified", and quadratic scores "will tend to penalise such forecasts harshly" [12]. Assel and colleagues simulated a clinical setting with 20% prevalence where overestimating risk was preferable. The Brier score (0 to 1 scale) still ranked the overestimating model below the underestimating one: 0.1708 against 0.1540, with 0.1386 for the well-calibrated model [21]. A decision-analytic measure, net benefit, "always favored the preferable test or model" in their scenarios [21].

Limits of this analysis

- This explainer reports no new forecasting data. Values marked "computed" are arithmetic from cited formulas.
- No source examined gives a general rule for when Brier and log rankings disagree, beyond extreme forecasts and rare events. Merkle and Steyvers' comparison covers 10 forecasters [19].
- No source examined gives a universal threshold for a good Brier score or log loss.
- Six papers were read in their arXiv versions [10, 11, 18, 20, 24, 27].
- Metaculus's public page did not load for automated requests. Its text was read from the page's source file in Metaculus's public repository, and the live page may differ [23].
- Leaderboard values change. ForecastBench figures are from the paper's version 5 [24] and the March 2026 post [26].
- Works cited inside the sources (Good, Selten, Bernardo, Briggs and Ruppert, Diebold and Mariano) were not opened and are reported as the citing sources describe them.
- This report describes scoring methods. It is not investment advice.

Sources

1. Verification of Forecasts Expressed in Terms of Probability. Glenn W. Brier, Monthly Weather Review 78(1), 1-3, DOI 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2. American Meteorological Society, January 1950. https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml

2. A New Vector Partition of the Probability Score. Allan H. Murphy, Journal of Applied Meteorology 12(4), 595-600, DOI 10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2. American Meteorological Society, June 1973. https://journals.ametsoc.org/view/journals/apme/12/4/1520-0450_1973_012_0595_anvpot_2_0_co_2.xml

3. Hedging and Skill Scores for Probability Forecasts. Allan H. Murphy, Journal of Applied Meteorology 12(1), 215-223, DOI 10.1175/1520-0450(1973)012<0215:HASSFP>2.0.CO;2. American Meteorological Society, February 1973. https://journals.ametsoc.org/view/journals/apme/12/1/1520-0450_1973_012_0215_hassfp_2_0_co_2.xml

4. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. Allan H. Murphy, Weather and Forecasting 8(2), 281-293, DOI 10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2. American Meteorological Society, June 1993. https://journals.ametsoc.org/view/journals/wefo/8/2/1520-0434_1993_008_0281_wiagfa_2_0_co_2.xml

5. Strictly Proper Scoring Rules, Prediction, and Estimation. Tilmann Gneiting and Adrian E. Raftery, Journal of the American Statistical Association 102(477), 359-378, DOI 10.1198/016214506000001437, author copy. University of Washington, March 2007. https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf

6. Probabilistic forecasts, calibration and sharpness. Tilmann Gneiting, Fadoua Balabdaoui and Adrian E. Raftery, Journal of the Royal Statistical Society Series B 69(2), 243-268, author copy. University of Washington, 2007. https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jrssb.pdf

7. Scoring Probabilistic Forecasts: The Importance of Being Proper. Jochen Bröcker and Leonard A. Smith, Weather and Forecasting 22(2), 382-388, DOI 10.1175/WAF966.1. American Meteorological Society, April 2007. https://journals.ametsoc.org/view/journals/wefo/22/2/waf966_1.xml

8. Evaluating Probabilistic Forecasts Using Information Theory. Mark S. Roulston and Leonard A. Smith, Monthly Weather Review 130(6), 1653-1660, DOI 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2. American Meteorological Society, June 2002. https://journals.ametsoc.org/view/journals/mwre/130/6/1520-0493_2002_130_1653_epfuit_2.0.co_2.xml

9. Scoring Rules for Forecast Verification. Riccardo Benedetti, Monthly Weather Review 138(1), 203-211, DOI 10.1175/2009MWR2945.1. American Meteorological Society, January 2010. https://journals.ametsoc.org/view/journals/mwre/138/1/2009mwr2945.1.xml

10. Reliability, Sufficiency, and the Decomposition of Proper Scores. Jochen Bröcker, published in the Quarterly Journal of the Royal Meteorological Society, DOI 10.1002/qj.456, arXiv:0806.0813, version 2, 1 December 2008. https://arxiv.org/abs/0806.0813

11. Evaluating probabilistic classifiers: Reliability diagrams and score decompositions revisited. Timo Dimitriadis, Tilmann Gneiting and Alexander I. Jordan, published as Stable reliability diagrams for probabilistic classifiers in the Proceedings of the National Academy of Sciences 118, e2016191118 (2021), arXiv:2008.03033, version 1, 7 August 2020. https://arxiv.org/abs/2008.03033

12. Forecast verification - methods, issues and FAQ. WWRP/WGNE Joint Working Group on Forecast Verification Research, viewed 24 October 2026. https://jwgfvr.github.io/forecastverification/

13. 3.4 Metrics and scoring: quantifying the quality of predictions, scikit-learn 1.9.1 User Guide. scikit-learn developers, viewed 24 October 2026. https://scikit-learn.org/stable/modules/model_evaluation.html

14. sklearn.metrics.brier_score_loss, scikit-learn 1.9.1 API reference. scikit-learn developers, viewed 24 October 2026. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.brier_score_loss.html

15. sklearn.metrics.log_loss, scikit-learn 1.9.1 API reference. scikit-learn developers, viewed 24 October 2026. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html

16. 1.16 Probability calibration, scikit-learn 1.9.1 User Guide. scikit-learn developers, viewed 24 October 2026. https://scikit-learn.org/stable/modules/calibration.html

17. Scoring rules in scoringutils, scoringutils package vignette. Nikos Bosse, CRAN, R Project, 27 September 2026. https://cran.r-project.org/web/packages/scoringutils/vignettes/scoring-rules.html

18. On Calibration of Modern Neural Networks. Chuan Guo, Geoff Pleiss, Yu Sun and Kilian Q. Weinberger, Proceedings of the 34th International Conference on Machine Learning (ICML 2017), arXiv:1706.04599, version 2, 3 August 2017. https://arxiv.org/abs/1706.04599

19. Choosing a Strictly Proper Scoring Rule. Edgar C. Merkle and Mark Steyvers, Decision Analysis 10(4), 292-304, DOI 10.1287/deca.2013.0280, author copy. INFORMS, December 2013. https://mstrep.s3.amazonaws.com/publications/MerkleSteyvers.pdf

20. Ranking earthquake forecasts using proper scoring rules: binary events in a low probability environment. Francesco Serafini, Mark Naylor, Finn Lindgren, Maximilian J. Werner and Ian Main, published in Geophysical Journal International, DOI 10.1093/gji/ggac124, arXiv:2105.12065, version 2, 10 September 2021. https://arxiv.org/abs/2105.12065

21. The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. Melissa Assel, Daniel D. Sjoberg and Andrew J. Vickers, Diagnostic and Prognostic Research 1:19, DOI 10.1186/s41512-017-0020-3. PubMed Central, 2 December 2017. https://pmc.ncbi.nlm.nih.gov/articles/PMC6460786/

22. FAQ. Good Judgment Open, viewed 24 October 2026. https://www.gjopen.com/faq

23. Scores FAQ, source file of the help page in Metaculus's public repository on GitHub (main branch). Metaculus, viewed 24 October 2026. https://github.com/Metaculus/metaculus/blob/main/front_end/src/app/(main)/help/scores-faq/page.tsx

24. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. Ezra Karger and others, International Conference on Learning Representations (ICLR) 2025, arXiv:2409.19839, version 5, 28 February 2025. https://arxiv.org/abs/2409.19839

25. Leaderboards, ForecastBench. Forecasting Research Institute, viewed 24 October 2026. https://www.forecastbench.org/leaderboards/

26. Making Forecasting Scores Easier to Interpret: Introducing the Brier Index. Simas Kučinskas, Houtan Bastani and Matt Reynolds, with an addendum dated 5 March 2026. Forecasting Research Institute, 4 March 2026. https://forecastingresearch.substack.com/p/introducing-the-brier-index

27. Evaluating epidemic forecasts in an interval format. Johannes Bracher, Evan L. Ray, Tilmann Gneiting and Nicholas G. Reich, published in PLOS Computational Biology, DOI 10.1371/journal.pcbi.1008618, arXiv:2005.12881, version 3, 8 January 2021. https://arxiv.org/abs/2005.12881

28. Forecasting: Principles and Practice, third edition, section 5.9, Evaluating distributional forecast accuracy. Rob J Hyndman and George Athanasopoulos. OTexts, viewed 24 October 2026. https://otexts.com/fpp3/distaccuracy.html

About this report

Published by Rillor on 7 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Brier score and log loss for probability forecasts. 7 November 2026. https://rillor.com/insights/brier-score-and-log-loss

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send the question you want forecast, its horizon and the data you already hold. Contact us at rillor.com/contact?type=forecasting or contact@rillor.com. More on forecasting and market intelligence: rillor.com/markets/forecasting