Calibration and the Brier Score in Sports Prediction Markets: A Hierarchical Shrinkage Approach to NFL Player Performance Forecasting

Fourth & Value Research Group

fourthandvalue.com

August 29, 2026

ABSTRACT

We examine the use of proper scoring rules, specifically the Brier score, for evaluating probabilistic forecasts of NFL player performance against sportsbook betting lines. We formalize the Brier score's strict propriety, apply Murphy's (1973) three-term decomposition into reliability, resolution, and uncertainty, and compare two forecast estimators: a single-window recency estimator using each player's trailing four games, and a hierarchical shrinkage estimator that partially pools player-level estimates toward position-level priors using an empirical Bayes weighting scheme. Forecasts are evaluated out-of-sample against 12,900+ graded player-prop outcomes from the 2025 NFL season across eight statistical markets, with probability calibration fit via isotonic regression and validated under grouped cross-validation. The hierarchical estimator reduces the Brier score from 0.337 to 0.276 (reliability component: 0.086 to 0.025) relative to the single-window baseline, and reaches 0.252 after calibration — approaching the coin-flip floor of 0.25 implied by the near-even base rate of the underlying markets. We discuss the distinction between calibration and resolution, and note that a well-calibrated forecast is not equivalent to a forecast with demonstrated predictive skill.

Keywords: proper scoring rules, Brier score, forecast calibration, isotonic regression, hierarchical shrinkage, empirical Bayes, sports prediction markets

1. Introduction

Evaluating a probabilistic forecaster by whether its favored outcome occurred more often than not — a "hit rate" or win-rate metric — is a common but statistically incomplete practice in sports prediction. Hit-rate metrics are invariant to the forecaster's stated confidence: a model reporting 55% and a model reporting 95% are scored identically if both favor the side that wins, even though the second model is making a categorically stronger and more falsifiable claim. This creates an incentive structure under which reporting extreme probabilities, rather than accurate ones, can appear advantageous by a naive read of accuracy alone.

A *proper scoring rule* corrects this by constructing a loss function under which a forecaster's expected score is optimized only by reporting their true subjective probability — no relabeling, hedging, or overstatement of confidence can improve the expected score. The Brier score [1] is the most widely used proper scoring rule for binary probabilistic forecasts and is standard in meteorology, epidemiology, and machine learning classifier calibration.

This note has three aims: (i) formalize the Brier score and its decomposition, (ii) describe a hierarchical shrinkage estimator for player-level performance forecasting motivated by the small-sample instability of single-window estimators, and (iii) report an empirical evaluation of both estimators, and the sportsbook market itself, against real outcomes from the 2025 NFL season.

2. Proper Scoring Rules and the Brier Score

2.1 Definition

For a binary outcome $o \in \{0, 1\}$ and a forecast probability $f \in [0, 1]$ that $o = 1$, the Brier score over N forecast–outcome pairs is defined as:

$$BS = (1/N) \sum_{i=1}^{N} (f_i - o_i)^2 \tag{1}$$

The score ranges over $[0, 1]$, with 0 indicating perfect foresight. A forecaster who reports $f = 0.5$ unconditionally scores exactly 0.25 regardless of the outcome sequence — the

"coin-flip" reference score used throughout this note as a baseline for markets with a near-even base rate.

2.2 Strict Propriety

A scoring rule is *strictly proper* if, for a forecaster with true belief p, the expected score is uniquely minimized by reporting $f = p$. For the Brier score, the expected loss of reporting f when the true probability is p is:

$$E[BS] = p(1-f)^2 + (1-p)f^2 \tag{2}$$

Differentiating with respect to f and setting the result to zero yields $f^* = p$, with positive second derivative confirming a minimum. No reporting strategy other than the forecaster's true belief minimizes expected loss [2]. This property is what distinguishes the Brier score from accuracy-based metrics, which admit no such guarantee.

2.3 Murphy's Decomposition

Murphy (1973) [3] shows that, for forecasts grouped into K bins by predicted value, the Brier score admits an exact three-term decomposition:

$$BS = \underbrace{\frac{1}{N}\sum_k n_k(\bar{f}_k - \bar{o}_k)^2}_{\textit{Reliability}} - \underbrace{\frac{1}{N}\sum_k n_k(\bar{o}_k - \bar{o})^2}_{\textit{Resolution}} + \underbrace{\bar{o}(1-\bar{o})}_{\textit{Uncertainty}} \tag{3}$$

where n_k is the count of forecasts in bin k, $\bar{f}_k$ the mean forecast in that bin, $\bar{o}_k$ the observed outcome frequency in that bin, and $\bar{o}$ the overall base rate. **Reliability** penalizes the gap between stated confidence and observed frequency (lower is better; zero indicates perfect calibration). **Resolution** rewards the forecaster for sorting cases into bins with outcome frequencies that differ from the overall base rate (higher is better; it measures discrimination, independent of calibration). **Uncertainty** is a property of the outcome data alone — the irreducible variance of the event being forecast — and is identical for every forecaster evaluated on the same outcome sample.

This decomposition is the reason a single Brier score number is an incomplete summary on its own: two forecasters can share an identical Brier score while one is well-calibrated with

low resolution (an honest "I don't know") and the other is poorly calibrated with high resolution (a confident, systematically wrong forecaster). Both terms are reported separately in Section 6.

3. Calibration Methodology

We calibrate raw forecast probabilities using isotonic regression [4], a non-parametric method that fits a monotonic step function mapping raw forecast value to empirical outcome frequency, subject to the constraint that the mapping is non-decreasing. Isotonic regression makes no assumption about the functional form of the miscalibration (unlike Platt scaling, which assumes a logistic form [5]), which is appropriate here given no prior reason to expect a particular parametric miscalibration shape.

To avoid overfitting the calibration map to the same data used to evaluate it, we use grouped k-fold cross-validation ($k=5$), with folds constructed by week number so that no week's outcomes inform the calibration map applied to that same week. All out-of-sample results in Section 6 use this protocol; the production calibration map is subsequently refit on the full sample for deployment.

4. Forecast Estimator: Hierarchical Shrinkage

We compare two estimators for the latent parameters (μ, σ) of a per-player, per-market performance distribution, from which forecast probabilities against a sportsbook line are derived via the corresponding CDF.

4.1 Estimator A: Single-Window Recency Estimate

Volume and efficiency statistics are estimated from each player's trailing $k = 4$ games within the current season, using exponential weights $w_t \propto \alpha^{(n-t)}$ with decay $\alpha = 0.4$. This estimator is standard in short-window recency-weighted forecasting but has a structural limitation: a standard deviation computed from four observations is itself a high-variance estimate of the true dispersion, and provides no mechanism for players with fewer than four current-season observations.

4.2 Estimator B: Hierarchical Shrinkage Estimate

Estimator B introduces two additional pooling levels, in the spirit of empirical Bayes / Stein-type shrinkage estimation [6] and hierarchical regression modeling [7]:

1. **Position-level prior.** For each statistical category, a prior mean and variance are computed from all player-games at the same position (QB/RB/WR/TE) across a multi-season reference sample (2019–2025, $n \approx 130{,}000$ player-games).
2. **Player career estimate.** Each player's own multi-season history is aggregated with season-level exponential decay ($\rho = 0.75$ per season), then shrunk toward the position-level prior with weight $w = n_{career} / (n_{career} + K_{pos})$, $K_{pos} = 40$, an approximation to the empirical Bayes shrinkage factor $\tau^2 / (\tau^2 + \sigma^2/n)$ with K_{pos} standing in for the prior-to-sample variance ratio.
3. **Current-season blend.** The resulting career estimate is blended with the current-season Estimator-A value with weight $w = n_{recent} / (n_{recent} + K_{recent})$, using separate constants for the mean ($K_{\mu} = 4$) and standard deviation ($K_{\sigma} = 8$) — reflecting that a variance estimate requires more observations than a mean estimate to be trusted at equal weight.

This construction guarantees a well-defined estimate for any player, including those with zero current-season observations, while allowing the estimate to converge toward the single-window value as within-season sample size grows.

5. Data and Experimental Design

Player performance outcomes are sourced from nflverse play-by-play and box-score aggregations for the 2025 NFL regular season. Sportsbook lines and prices are sourced from four national sportsbooks via a real-time odds aggregation API. The evaluation sample covers eight statistical markets — rushing yards, rushing attempts, receiving yards, receptions, passing yards, passing attempts, passing completions — across nine weeks of the season (weeks 4–14, non-contiguous due to data availability), yielding $N = 12{,}900$ graded (player, market, line, side) observations for Estimator A and $N = 12{,}904$ for Estimator B.

For each observation, the realized outcome is coded as $o = 1$ if the named side (over/under) of the sportsbook line was correct given the actual statistical outcome, and $o = 0$ oth-

erwise; pushes (exact ties to the line) are excluded. To prevent information leakage, current-season statistics available to each estimator at forecast time are truncated to weeks strictly prior to the forecast week; career-level statistics are restricted to seasons strictly prior to the current season.

A market-implied probability, computed from the de-vigged consensus price across all available books for a given line, is reported alongside both estimators as an approximate reference for market efficiency, consistent with prior findings that closing sportsbook lines are close to informationally efficient [8].

6. Results

Table 1. Brier score and Murphy decomposition by estimator, pooled across all eight markets. Uncertainty is a property of the outcome sample and is therefore identical across rows; the base rate is 0.500 by construction of over/under lines.

Estimator	*N*	*Brier*	*Reliability↓*	*Resolution↑*	*Uncertainty*
A — single-window (k=4)	12,900	0.337	0.086	0.0004	0.250
B — hierarchical shrinkage	12,904	0.276	0.025	0.0002	0.250
Market (de-vigged consensus)	10,012	0.250	0.0001	0.0006	0.250

The reduction in Brier score from Estimator A to Estimator B is driven almost entirely by the reliability term (0.086 → 0.025); the resolution term is small and comparable across all three rows. This indicates the hierarchical estimator's improvement is a calibration effect — its stated confidence tracks reality substantially better — rather than a gain in the ability to discriminate winning from losing sides, which remains limited for all three, including the sportsbook market itself.

Table 2. Estimator B, Brier score and Pearson correlation between forecast probability and outcome, by market.

Market	*N*	*Brier (A)*	*Brier (B)*	*r (A)*	*r (B)*
Rushing yards	2,468	0.333	0.269	+0.014	+0.044
Receiving yards	4,960	0.338	0.280	-0.029	-0.010
Passing yards	1,558	0.333	0.279	-0.056	-0.078
Receptions	1,954	0.338	0.276	+0.044	+0.086
Rush attempts	880	0.335	0.275	+0.035	+0.050
Pass attempts	574	0.327	0.254	+0.073	+0.089
Pass completions	510	0.366	0.283	-0.042	-0.030

6.1 Out-of-Sample Calibration

Table 3 reports the grouped 5-fold cross-validated reliability diagram for Estimator B after isotonic calibration.

Table 3. Out-of-fold calibration table, Estimator B post-calibration. Confidence bins above 70% contain too few observations ($n \leq 17$) for the observed frequency to be a stable estimate.

Predicted bin	*N*	*Mean forecast*	*Observed freq.*
0.50–0.60	7,203	0.508	0.501
0.60–0.70	17	0.603	0.353
0.80–0.90	2	0.875	0.000
0.90–1.00	14	0.964	0.286

Out-of-fold Brier score: 0.2517, essentially at the uncertainty floor of 0.25. The dominant bin ($n = 7{,}203$ of 7,236 total, 99.5% of the sample) is well calibrated (0.508 predicted vs. 0.501 observed). The small number of observations in higher-confidence bins precludes a reliable calibration estimate there; we treat this as a genuine finding about the current model's resolution rather than a defect in the calibration procedure — the model rarely produces a confidently differentiated forecast, and when it does, the sample is too small to certify that confidence out-of-sample.

7. Discussion

The results in Section 6 illustrate the practical value of separating calibration from resolution. Estimator B is materially better calibrated than Estimator A and approaches the calibration quality of the sportsbook market itself. However, the resolution term for all three forecasters — including the market — is small relative to the uncertainty term, consistent with these markets being set close to a 50/50 proposition by design. A well-calibrated forecast under these conditions is a necessary but not sufficient condition for a forecaster to possess exploitable predictive skill; the correlation coefficients in Table 2, ranging from approximately −0.08 to +0.09 depending on market, quantify the (currently modest) extent of that skill directly.

This distinction matters for the interpretation of any single "edge" figure computed as the gap between a forecast probability and a market-implied probability: a large edge from a poorly-calibrated, low-resolution forecaster is not informative, whereas the same edge from a well-calibrated forecaster with demonstrated (if modest) resolution carries more evidentiary weight. We recommend reporting the reliability and resolution terms alongside any aggregate Brier score for this reason.

8. Limitations

- The evaluation covers a single NFL season (2025) and nine of eighteen weeks within it; results may not generalize across seasons with different scoring environments or rule changes.
- Entity resolution across data sources with inconsistent player-name conventions (e.g., generational suffixes) is a known source of silent join failure in multi-source sports data pipelines and was addressed via name normalization prior to this analysis; residual mismatches, if any, would bias estimates toward the position-level prior for affected players.
- The resolution term for both estimators is small in absolute magnitude; claims of predictive edge derived from this model should be sized accordingly.

9. Conclusion

We have presented a formal treatment of the Brier score as a strictly proper scoring rule, its decomposition into reliability and resolution components, and an empirical comparison of a single-window and a hierarchical shrinkage estimator for NFL player performance forecasting. The hierarchical estimator improves calibration substantially (reliability: 0.086 to 0.025) with a comparatively small change in resolution, and post-calibration performance approaches the theoretical floor implied by the near-even base rate of the underlying markets. We view continued, out-of-sample tracking of both calibration and resolution — rather than aggregate win rate — as the appropriate standard for evaluating forecasts of this kind going forward.

10. References

[1] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1–3.

[2] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

[3] Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600.

[4] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 694–699.

[5] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 61–74.

[6] Efron, B., & Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350), 311–319.

[7] Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

[8] Levitt, S. D. (2004). Why are gambling markets organised so differently from financial markets? *The Economic Journal*, 114(495), 223–246.

> Reproducibility: all figures reported in this note are computed from graded historical forecast–outcome pairs against real NFL box scores. Estimator and calibration source: `scripts/career_baseline.py`, `scripts/fit_calibration.py`. Evaluation source: `scripts/backtest_calibration.py`, `scripts/backtest_new_baseline.py`. Fitted calibration maps: `models/nfl_prop_calibration.json`. Source available at the repository backing this site.

Appendix A. Individual Forecast–Outcome Pairs

Aggregate statistics can obscure how a forecaster behaves on any single case. Table 4 reports individual (player, market, line) observations for Estimator B, sampled by a fixed, disclosed procedure rather than selected for effect: forecasts are grouped into five bins by raw forecast probability (`[0.50,0.60), [0.60,0.70), …, [0.90,1.00]`), restricted to the favored side (raw probability ≥ 0.5), and up to five observations are drawn from each bin by fixed-seed random sampling (`seed = 42`), for a maximum of 25 rows. Bins with fewer than five qualifying observations are shown in full. Both the raw Estimator B probability and the out-of-fold calibrated probability (Section 6.1) are reported alongside the realized statistical outcome.

Table 4. Stratified random sample of individual forecast–outcome pairs, Estimator B. "Result" indicates whether the named side (over/under) was correct given the realized statistic. Sorted by raw forecast probability, descending.

Player	*Market*	*Line*	*Side*	*P (raw)*	*P (calib.)*	*Actual*	*Result*	*Wk*
James Cook	Receptions	2.5	under	0.987	0.603	1	✓	9
Ladd McConkey	Receptions	5.5	under	0.983	0.980	5	✓	5
Jake Ferguson	Receptions	4.5	under	0.969	0.603	5	✗	9
Michael Pittman Jr.	Receptions	5.5	under	0.954	0.511	9	✗	14
Kirk Cousins	Pass Yds	208.5	under	0.907	0.573	199	✓	12
Malik Nabers	Receptions	5.5	under	0.864	0.519	2	✓	4
Jayden Daniels	Pass Comp	21.5	under	0.847	0.524	16	✓	9
Jameson Williams	Recv Yds	20.5	over	0.817	0.554	9	✗	5
Jared Goff	Rush Yds	0.5	over	0.817	0.511	-2	✗	14
Xavier Worthy	Rush Yds	3.5	over	0.805	0.554	9	✓	5
Saquon Barkley	Rush Yds	85.5	under	0.779	0.519	58	✓	6
Zach Charbonnet	Rush Att	6.5	over	0.744	0.509	9	✓	5
Rico Dowdle	Rush Yds	60.5	over	0.722	0.502	79	✓	7
David Montgomery	Recv Yds	8.5	over	0.712	0.502	13	✓	14
Jayden Daniels	Pass Att	31.5	under	0.704	0.524	22	✓	9
Daniel Jones	Pass Comp	20.5	over	0.696	0.537	19	✗	10
TreVeyon Henderson	Rush Yds	42.5	under	0.689	0.513	32	✓	4
TreVeyon Henderson	Rush Att	9.5	under	0.622	0.500	2	✓	7
Rico Dowdle	Recv Yds	14.5	under	0.612	0.500	11	✓	9
Josh Allen	Pass Yds	230.5	over	0.603	0.500	306	✓	10
Jordan Love	Pass Yds	234.5	over	0.599	0.500	176	✗	10

Player	*Market*	*Line*	*Side*	*P (raw)*	*P (calib.)*	*Actual*	*Result*	*Wk*
Jordan Love	Pass Att	29.5	under	0.593	0.500	43	✗	4
Cade Otton	Recv Yds	39.5	under	0.592	0.500	21	✓	12
Amon-Ra St. Brown	Recv Yds	80.5	under	0.543	0.500	58	✓	10
Cam Ward	Pass Comp	19.5	under	0.520	0.500	28	✗	12

Of the 25 sampled observations, 17 (68%) resolved in favor of the forecast — consistent with, though not a substitute for, the aggregate calibration reported in Section 6.1 given the small sample size here. Note the calibrated probability compresses toward 0.50 for most rows outside the top two bins, reflecting the limited resolution discussed in Section 7: the procedure is deliberately conservative about certifying confidence beyond what the out-of-fold evidence supports.