Calibrating KickCast: Per-Class Isotonic Recalibration of World-Cup Match Probabilities

Karim Semaan · Northeastern University · June 2026

Companion study to the KickCast project (github.com/karimsemaan/kickcast-worldcup)

Abstract

KickCast predicts international football matches as a 3-class (home / draw / away) problem and feeds those probabilities into a 10,000-iteration Monte-Carlo simulation of the 2026 World Cup. Because the downstream simulation consumes *probabilities*, not labels, the project's stated lever is calibration rather than top-1 accuracy. This study completes that work: the saved tuned, class-balanced XGBoost artifact is recalibrated with per-class isotonic regression and evaluated on the untouched 64-match 2022 World Cup holdout. Recalibration cuts holdout log-loss from **1.347 to 1.093**, 10-bin expected calibration error (ECE) from **0.157 to 0.120**, and multiclass Brier score from **0.739 to 0.646**. As a robustness check, the same calibrator applied to the full 3,552-match test split cuts log-loss from 1.031 to 0.923 and pooled ECE from 0.098 to 0.017, confirming the correction generalizes beyond the 64-match story.

1 · Motivation

KickCast's honest self-review noted that its 45.3% top-1 accuracy (29/64) on the 2022 World Cup holdout sits at the always-home baseline (~44%) and below a naive Elo-favorite rule (~52%) — precisely because the model predicts draws that favorite-picking baselines never do. For a system whose product is a full probability distribution over tournament outcomes, the correct optimization target is a proper scoring rule. This study measures how miscalibrated the deployed model actually is, applies a standard post-hoc correction, and reports the change under log-loss, ECE, and Brier score.

2 · Setup

Model under test. The saved tuned, class-balanced XGBoost artifact from the KickCast pipeline. Before any calibration work, the artifact was loaded and verified to reproduce the project's published test metrics exactly (log-loss 1.0308, accuracy 56.25% on the 3,552-match test split) — establishing that the model being studied is the model that shipped.

Data discipline. The KickCast pipeline uses leakage-safe chronological splits over a 21,371-match × 38-feature matrix. Three slices matter here:

- **Validation split** — 2,324 matches (2020 through Nov 2022). The *only* data the calibrator is fit on.
- **Test split** — 3,552 matches, used as the published-metric fidelity check and the robustness check.
- **2022 World Cup holdout** — 64 matches never touched during training, tuning, or calibrator fitting.

Measurement. Reliability is measured as pooled one-vs-rest over 192 match-class pairs (64 matches × 3 classes) in 10 uniform bins, alongside log-loss and multiclass Brier score.

3 · Diagnosis: how miscalibrated was the model?

On the 64-match holdout, the uncalibrated model shows **ECE 0.157, log-loss 1.347, multiclass Brier 0.739**. The reliability diagram shows a consistent pattern: under-confidence at low predicted probabilities and over-confidence above 0.6 — the model's strong opinions were too strong, and its weak opinions too weak.

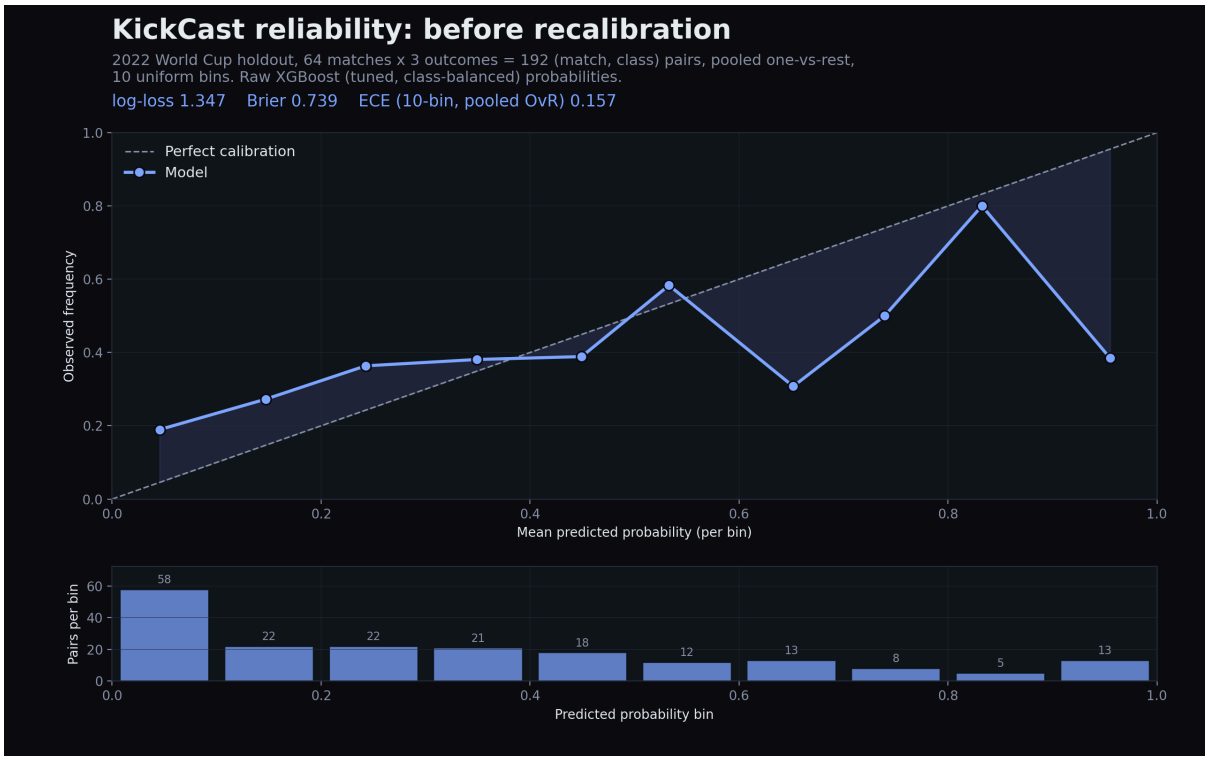

Figure 1 — Reliability diagram before recalibration (64-match 2022 WC holdout, pooled one-vs-rest, 10 uniform bins).

4 · Correction: per-class isotonic regression

A per-class isotonic regression (one monotone map per outcome class, followed by row renormalization so each match's probabilities still sum to 1) was fit *only* on the chronologically earlier 2,324-match validation split. The holdout stayed unseen during fitting — the same discipline the original pipeline applies to model training.

After recalibration, the same 64 matches score:

Metric (64-match WC holdout)	Before	After	Δ
Log-loss	1.347	**1.093**	−0.254
ECE (10-bin, pooled OvR)	0.157	**0.120**	−0.037
Multiclass Brier	0.739	**0.646**	−0.093
Top-label ECE	0.247	**0.148**	−0.099
Top-1 accuracy	45.3%	**50.0%**	+4.7 pp

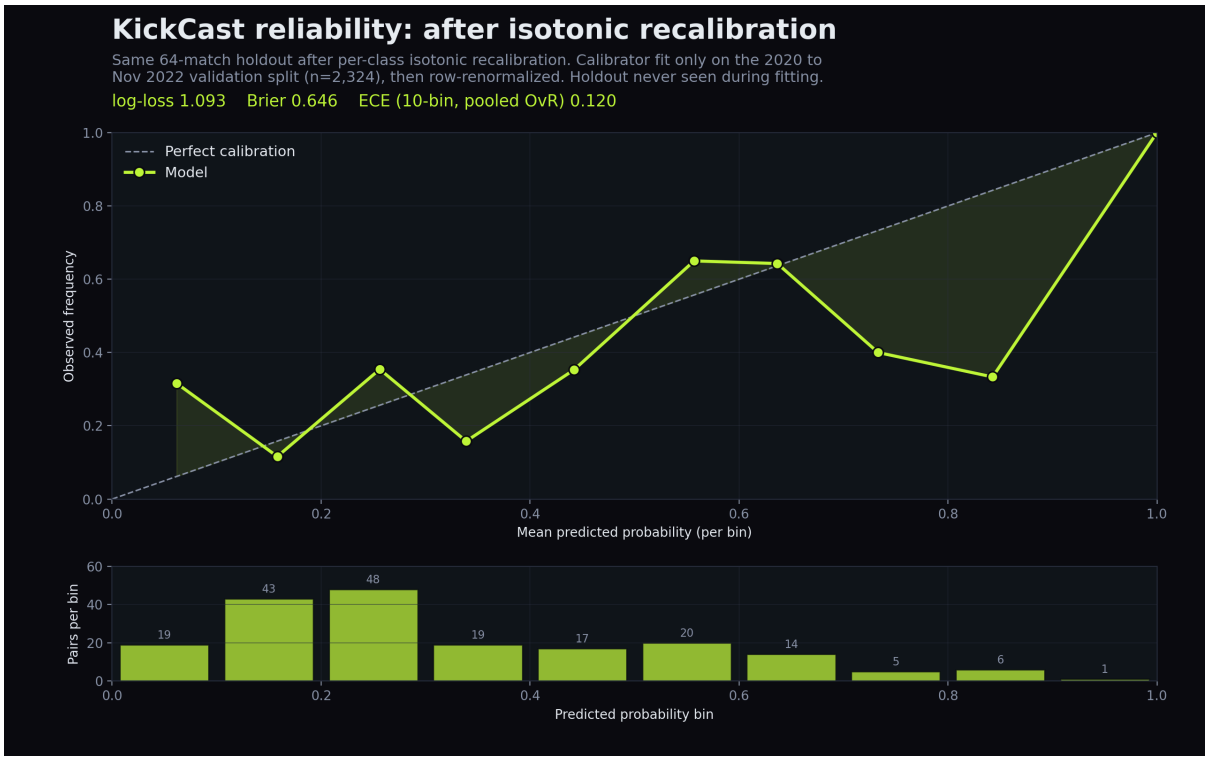

Figure 2 — Reliability diagram after per-class isotonic recalibration on the same holdout.

5 · Robustness check

n = 64 is a small sample: per-bin points are noisy and the deltas above carry wide uncertainty. To check that the correction is not an artifact of the small holdout, the same fitted calibrator was applied to the full 3,552-match test split:

Metric (3,552-match test split)	Before	After
Log-loss	1.031	**0.923**
ECE (10-bin, pooled OvR)	0.098	**0.017**

The near-elimination of pooled ECE at n = 3,552 is the strongest evidence that the isotonic maps capture a real, systematic miscalibration rather than holdout noise.

5.1 · Bootstrap confidence intervals on the holdout deltas

A percentile bootstrap (B = 10,000, calibrator fit once on the validation split and held fixed across resamples; resamples with a single outcome class skipped) puts error bars on the n = 64 deltas:

Delta (after – before, WC holdout)	Point	95% CI	Excludes 0?
Log-loss	−0.253	[−0.421, −0.101]	**yes**
Multiclass Brier	−0.093	[−0.151, −0.036]	**yes**
ECE (10-bin, pooled OvR)	−0.041	[−0.101, +0.026]	no

Read honestly: the proper-scoring-rule improvements (log-loss, Brier) are significant even at n = 64, but the holdout alone cannot certify the ECE gain — its interval crosses zero. The calibration claim therefore rests on the full-test robustness check above (n = 3,552, ECE 0.098 → 0.017), where the sample is large enough to resolve it.

6 · Honest limitations & open work

- **Small holdout.** Addressed with the bootstrap CIs in §5.1 — and they confirm the caveat was warranted: the ECE delta is not separable from zero at n = 64.
- **One method.** No head-to-head yet against temperature scaling or Dirichlet calibration; isotonic was chosen as the standard non-parametric default.
- **Downstream effect unmeasured.** The calibrated probabilities have not yet been fed back into the 10,000-run Monte-Carlo simulation to measure how tournament advancement odds shift.
- **Single tournament.** Backtesting calibration across the 2010, 2014, and 2018 cycles would establish whether the correction is stable across eras.

References & artifacts

- KickCast pipeline, model zoo, and result CSVs: *github.com/karimsemaan/kickcast-worldcup*
- Live Monte-Carlo dashboard: *kickcast-dashboard.vercel.app*
- Interactive case study with both reliability diagrams: *karimnsemaan.me/work/kickcast-calibration*
- Zadrozny & Elkan (2002), *Transforming classifier scores into accurate multiclass probability estimates* — the per-class isotonic + renormalization recipe used here.
- Guo et al. (2017), *On Calibration of Modern Neural Networks* — ECE measurement convention.