A Goal-Arrival Generator for World Cup Forecasting

Johannes Bleher*

University of Hohenheim, Econometrics and Empirical Economics & Computational Science Hub
Schloss Hohenheim 1 C, 70599 Stuttgart, Germany

johannes.bleher@uni-hohenheim.de

*Corresponding author. Tel.: +49 711 459 23054

June 8, 2026

Abstract

This paper develops and evaluates a goal-arrival generator for World Cup forecasting. The statistical object is a predictive distribution over final scores. Win–draw–loss probabilities, exact-score decisions, tournament simulations, and total-goals forecasts are treated as functionals of that distribution rather than as separate forecasting problems. The generator is estimated as a state-dependent counting process and is compared with categorical econometric models, calibrated machine-learning benchmarks, and market-implied probabilities under fixed forecast vintages. Historical World Cup backtests show that the event-time generator improves the win–draw–loss proper scores relative to the final-score likelihood, while the final-score generator remains competitive for exact-score likelihood. Calibrated random forests are mean-best in the sparse strict pre-tournament categorical benchmark, but generator and logit forecasts remain in the predictive-ability set under World-Cup-clustered uncertainty. Market prices dominate on overlapping matchday or closing odds, but those prices define a reference frontier rather than a strict pre-tournament comparison. The paper therefore argues for forecast-object discipline: conclusions about predictive ability depend on the target, the information set, and the scoring rule.

Keywords: football forecasting; World Cup; point processes; Doi–Peliti generators; exact scores; proper scoring rules.

JEL codes: C11; C25; C52; C53.

1 Introduction

Forecasting a football tournament is not one prediction problem. A forecaster may be asked for the probability that a team wins a match, the exact final score, the number of goals in the tournament, or the submitted bracket implied by many conditional match outcomes. These targets are related, but they are not interchangeable. A model that gives a well-calibrated win–draw–loss probability need not give sensible exact-score probabilities. A model that is useful for an exact-score contest may not be best for ranking teams by match-winning probability. A tournament total-goals forecast requires a full distribution over all fixture scorelines and the tournament path, not only a sequence of modal scores. The first contribution of this paper is therefore conceptual: it treats the final-score distribution as the primitive forecast object and derives the other objects from it.

This distinction matters for World Cup forecasting because the tournament combines sparse target data with rich but uneven information. Historical World Cups provide only a small number of held-out tournament vintages. International friendly and competitive matches provide many observations, but the information set changes over time and differs sharply across teams. Betting markets provide a powerful benchmark, but market prices are horizon-specific: closing or match-day odds are not the same object as a strict pre-tournament forecast. Player availability, injuries, tactical news, and lineup information are plausible sources of incremental predictive information, yet they are difficult to use without timestamped provenance. A scientific forecast must therefore solve two problems at once. It must estimate a distributional object that can answer the relevant questions, and it must discipline the information set used to evaluate competing methods.

The paper is positioned between two established forecasting traditions. The first models match outcomes or scores directly, using ordered response models, Poisson-type score models, hierarchical score models, ratings, and machine-learning methods. The second treats the tournament as a simulation problem, often combining team-strength estimates, bookmaker information, and repeated tournament draws to obtain progression and title probabilities. These approaches have generated a mature literature on football forecasting and on World Cup prediction. The opening for the present paper is more specific. We ask whether an event-time generator can be used as the common probabilistic object from which the match result, exact score, tournament path, and total-goals targets are all derived under the same information-set discipline.

This paper asks whether a state-dependent goal-arrival generator provides a useful and backtestable

score distribution for this setting. The generator views goals as jumps in a continuous-time match state. Its transition law at full time yields a distribution over final scores. Win–draw–loss probabilities, exact-score decisions under a custom payoff, and total tournament goals are then functionals of that distribution. This is a modest but important shift in emphasis. The goal is not to claim that a particular point-process specification is universally optimal. The goal is to evaluate whether event-time structure contains predictive information that is not captured by final-score calibration or by direct categorical models.

The empirical exercise is deliberately designed as a method-performance study. Historical World Cups are predicted from fixed forecast vintages. Competing forecasts are scored on common matches and common forecast objects. Proper scores evaluate probabilistic targets. Contest points are reported separately because they depend on a decision rule rather than on a proper scoring rule. The main non-market competitors are ordered logit, multinomial logit, calibrated random forest, and calibrated gradient boosting. Market-implied probabilities are included where matched odds exist, but they are treated as a market-reference frontier because the available historical odds are matchday or closing prices rather than strict pre-tournament prices.

The evidence is informative but not final. In the current historical backtests, the event-time goal-arrival likelihood improves the win–draw–loss proper scores relative to the final-score likelihood. The final-score generator remains best on exact-score log score. Calibrated random forests are mean-best in the sparse strict pre-tournament categorical benchmark, but generator and logit forecasts remain in the predictive-ability set under World-Cup-clustered uncertainty. Gradient boosting performs poorly in this design. On market-overlap fixtures, no-vig market probabilities dominate the statistical models, as expected, but that result should be read as a frontier comparison rather than a strict pre-tournament horse race.

The application is the 2026 World Cup in North America. The current implementation produces score decisions under the workbook payoff: group-stage submissions receive two points for the correct win–draw–loss class, while knockout submissions receive five points for an exact score and two points for the correct result class otherwise. It also produces a total-goals forecast for a side target that excludes shootout goals and in-match penalty goals. These outputs are useful as freeze candidates, but they are not yet a final forecast vintage. Final paper-grade use still requires source-stamped 2026 market odds, verified post-playoff fixtures, real pre-playoff slot probabilities, licensed player and lineup covariates, and a frozen forecast vintage.

The paper proceeds as follows. Section 2 places the paper in the forecasting, econometric, and sports-modelling literature. Section 3 defines the goal-arrival generator. Section 4 formalizes the forecast targets and decisions. Section 5 describes the data and information sets. Section 6 sets out the evaluation design. Section 7 reports historical backtests. Section 8 reports the current 2026 forecast vintage and remaining gates. Section 9 concludes.

2 Related Literature and Contribution

The paper connects six literatures: categorical econometric forecasting, football score modelling, dynamic team ratings, tournament simulation, betting-market benchmarks, and predictive-ability testing. The connection is not simply that these literatures all study prediction. They imply different forecast objects. Categorical econometric models target a probability vector over result classes. Score models target a joint distribution over goals. Rating models target latent team strength. Tournament models propagate match-level probabilities through a bracket. Betting-market benchmarks measure the information content of prices at a particular horizon. Predictive-ability tests discipline comparisons across these objects.

Categorical response models provide the most direct econometric foundation for win–draw–loss prediction. Conditional logit and related qualitative response models are designed for discrete outcomes and are natural baselines when the target is a categorical probability vector (McFadden, 1974; Amemiya, 1981). Their advantage is clarity. The outcome is observed once per match, the likelihood is transparent, and proper scores can be applied directly to the predicted probability vector. Their limitation is equally clear: a categorical model does not produce a final-score distribution without an additional layer of assumptions.

Football score models address that limitation by modelling goals. The tradition starts with Poisson score models that represent attacking and defensive strength (Maher, 1982). The Dixon–Coles framework adds low-score dependence and time weighting, and shows how scoreline models can be used for football forecasts and betting-market analysis (Dixon and Coles, 1997). Subsequent work compares goal-based and result-based regression forecasts (Goddard, 2005), introduces bivariate Poisson specifications (Karlis and Ntzoufras, 2003), and develops Bayesian or dynamic score models in which team strengths evolve over time (Rue and Salvesen, 2000; Baio and Blangiardo, 2010; Koopman and Lit, 2015). These approaches are close to the forecast object in this paper because

an exact-score prediction is a cell of the modelled distribution. Their common simplification is that the match is usually observed through the full-time score. That compression can be efficient, but it uses goal timing only indirectly.

International-tournament forecasting adds another layer. World Cup and European Championship papers typically estimate team strength or match-level goal distributions and then simulate the tournament. Early World Cup simulations used ratings-based Poisson models (Dyte and Clarke, 2000). Later work combines bookmaker-implied probabilities, ratings, and simulation for tournament forecasts (Leitner et al., 2010). Regularized Poisson models have been used for major international soccer tournaments and for the 2014 FIFA World Cup (Groll et al., 2015). Random forests and hybrid approaches combine covariates with ability estimates and have been applied to international tournaments and World Cup forecasts (Schauberger and Groll, 2018; Groll et al., 2019). This literature is the closest empirical benchmark for the present paper. The distinction is that the current paper treats the final-score distribution as the primitive object and asks whether an event-time generator improves the distribution from which match, tournament, and total-goals targets are derived.

Dynamic ratings and paired-comparison models provide a complementary view of team strength. Elo-type and Glicko-type systems summarize past performance in a low-dimensional state and are attractive because they update naturally over time (Glickman, 1999; Hvattum and Arntzen, 2010). Soccer-specific strength rankings have been estimated by weighted likelihood using Thurstone–Mosteller, Bradley–Terry, independent Poisson, and bivariate Poisson models (Ley et al., 2019). These models are useful sources of covariates and benchmark probabilities. In this paper, ratings and rolling strength measures enter as information about team quality rather than as the final forecast object.

The goal-arrival generator is related to counting-process and hazard-model thinking. In a counting process, events arrive with state-dependent intensities. The event of interest here is a goal, and the state includes the current score, time, team strength, and opponent composition. This formulation is conceptually close to survival and event-history models (Cox, 1972), but the output needed for the football application is not a single duration. It is the transition distribution over the two-dimensional score state at full time. The generator language is useful because it makes that transition explicit and keeps the event-time law, final-score law, and decision rule separate.

Machine-learning benchmarks enter for a different reason. Random forests and gradient boosting can approximate nonlinear functions and interactions (Breiman, 2001; Friedman, 2001). Soccer forecasting challenges have shown that outcome models, hierarchical Poisson models, ratings, and machine-learning methods can be compared under temporal validation designs (Tsokos et al., 2019). If calibrated machine-learning probabilities dominate econometric models under common forecast vintages, that is evidence that the simple structural specification is missing predictive information. If they do not dominate, the added flexibility may not be useful in the available sample. The paper therefore treats machine-learning rows as challengers in a predictive-ability exercise, not as black-box replacements for the forecast object.

Betting markets are a demanding benchmark because prices aggregate information from many participants. The football forecasting literature has long used odds both as predictors and as objects in market-efficiency tests. Bookmaker odds have been shown to contain substantial forecasting information in English football (Forrest et al., 2005), and tournament-forecasting papers often use bookmaker consensus measures as ability or probability inputs (Leitner et al., 2010; Groll et al., 2019). In this paper, market probabilities are not treated as just another model row. They define a horizon-specific information set. Closing odds can contain lineup, injury, weather, and tactical information unavailable to a pre-tournament model. A fair pre-tournament comparison therefore requires opening or pre-tournament odds observed at the same forecast date. The current historical market comparison is informative, but it is a reference frontier rather than a strict pre-tournament benchmark.

The evaluation draws on the literature on proper scoring rules and predictive ability. Proper scores align estimation and evaluation with the target probability distribution (Gneiting and Raftery, 2007). Ranked probability scores are useful for ordered outcomes (Epstein, 1969). Predictive-ability testing and model confidence sets provide a disciplined way to compare forecast losses when several models are considered (Diebold and Mariano, 1995; Giacomini and White, 2006; Hansen et al., 2011). The current application is deliberately modest because only a small number of World Cup vintages are available. The goal is to report paired evidence and uncertainty, not to overstate small score differences.

The contribution is thus threefold. First, the paper defines a coherent forecast object for a setting with multiple downstream targets: match result, exact score, tournament path, and total goals. Second, it evaluates whether event-time goal-arrival structure adds predictive information beyond

final-score and categorical calibration. Third, it embeds the 2026 forecasts in a reproducible vintage discipline that distinguishes implemented evidence from pending data blocks. The paper is intentionally conservative about claims: player, lineup, and news covariates are discussed as a schema and a future information set, not as active predictors.

3 A Goal-Arrival Generator

The statistical target is the conditional distribution of the final score. This target follows the football-score literature, where Poisson, bivariate Poisson, Bayesian hierarchical, and dynamic specifications model the scoreline directly rather than only the match result (Maher, 1982; Dixon and Coles, 1997; Karlis and Ntzoufras, 2003; Rue and Salvesen, 2000; Baio and Blangiardo, 2010; Koopman and Lit, 2015). The generator formulation keeps that final-score target but estimates it through a state-dependent event-time law. Let

$$S_t = \{G_A(t), G_B(t), D(t), X(t)\} \tag{1}$$

denote the state of a match at time t, where $G_A(t)$ and $G_B(t)$ are the current goals of teams A and B, $D(t) = G_A(t) - G_B(t)$ is the score differential, and $X(t)$ contains pre-match and match-state covariates. The generator acts on a test function f as

$$\mathcal{L}_t f(S_t) = \lambda_A(S_t, t; \theta)\{f(S_t + e_A) - f(S_t)\} + \lambda_B(S_t, t; \theta)\{f(S_t + e_B) - f(S_t)\}, \tag{2}$$

where e_A and e_B increment the score of the corresponding team by one goal. This is a birth process on the two-dimensional score state. The intensities λ_A and λ_B are allowed to depend on team strength, opponent strength, time, score state, venue, and other covariates that are known at the forecast time.

The final-score distribution is the transition law induced by this generator:

$$p_\theta(g_A, g_B \mid X) = \Pr_\theta\{G_A(90) = g_A, G_B(90) = g_B \mid X\}. \tag{3}$$

This object is the bridge between event-time modelling and the scoreline forecasts needed by the

contest. Once p_θ is available, the win probability is

$$\Pr(A \text{ wins} \mid X) = \sum_{g_A > g_B} p_\theta(g_A, g_B \mid X), \tag{4}$$

and analogous sums give draw and loss probabilities. Exact-score probabilities are individual cells of the same distribution. Expected goals are moments of the same distribution. Tournament total-goal distributions are convolutions or simulations over fixture-level score distributions. The generator therefore supplies one coherent probabilistic object rather than separate forecasts for each downstream target.

The implemented final-score generator uses a parsimonious rate specification. Team strength enters through rolling pre-match ratings, following the broad rating and current-strength tradition in football forecasting (Glickman, 1999; Hvattum and Arntzen, 2010; Ley et al., 2019). Score state and a late-match component allow intensities to change with tactical incentives and time. The current model is intentionally simple: it is strong enough to define the forecast object and to support historical backtests, but it is not yet a player-aware or lineup-aware process. This distinction is important for claim discipline. The paper can report evidence about the information in goal-arrival timing and team-level strength. It cannot yet claim that player form, injury news, or tactical lineups improve the forecast, because these covariates are not yet populated with licensed, source-stamped rows.

The event-time extension estimates a piecewise counting-process likelihood. The match is partitioned into time bins, and each team-bin contributes a goal count with exposure equal to time at risk. Team attack and defense effects are regularised using prior-match strength. In this specification, the generator for team i against team j is not obtained merely by relabelling a generic process. Team i's attack is composed with team j's defense, the match context, and score-state effects. Thus Germany–Brazil and Germany–Japan imply different generators because the opponent component changes the conditional arrival rates.

For the live forecast, estimation proceeds in four steps. First, the historical international match file is cut at the forecast date. From this sample we construct pre-match team-strength and recent-form covariates using only matches dated before the fixture being predicted. The resulting feature table supplies the sparse information set common to the generator and the categorical benchmarks. Player, lineup, injury, and news fields are carried as schema-checked missingness indicators, but

they are not active predictors in the current generator because source-stamped rows are not yet available.

Second, the final-score component estimates a Poisson exposure model for team goals. For team i against team j, the match-level goal rate is written as

$$\log \mu_{ij} = \alpha + \beta R_{ij} + a_i - d_j, \tag{5}$$

where R_{ij} is the pre-match strength differential, a_i is a team attack effect, and d_j is the opponent defense effect. The global coefficients are estimated from the pre-vintage match sample. Team effects are then empirical-Bayes shrinkage estimates. If n_i is the number of observed team matches, $\bar{g}_i^+$ and $\bar{g}_i^-$ are goals for and against per match, and $\bar{g}$ is the global goal mean, the attack effect has the form

$$a_i = \frac{n_i}{n_i + \kappa} \log\left(\frac{n_i \bar{g}_i^+ + \kappa \bar{g}}{(n_i + \kappa)\bar{g}}\right), \tag{6}$$

with an analogous expression for defense. The tuning constant κ is the prior match count. This construction makes the generator team-specific while preventing teams with little information from receiving unstable attack or defense effects.

Third, the event-time component replaces the full-time score likelihood by a piecewise counting-process likelihood on World Cup goal-event data. Let Y_{mikr} be the number of goals scored in fixture m by side i against side j in time bin r, and let E_{mikr} be the corresponding minutes at risk. The event-time generator uses the bin-level intensity

$$\log \lambda_{mikr} = \eta_r + \gamma Z_{mikr} + \beta R_{mij} + a_i - d_j, \tag{7}$$

where η_r is the baseline time-bin component, Z_{mikr} collects score-state and late-match indicators, R_{mij} is the pre-match strength differential from the perspective of side i, and $a_i - d_j$ is the team-specific attack-defense composition from Equation (5). The event-time log likelihood is

$$\ell_{\mathrm{ET}}(\theta) = \sum_{m,i,r} \left[Y_{mikr} \log\{E_{mikr}\lambda_{mikr}\} - E_{mikr}\lambda_{mikr} - \log(Y_{mikr}!) \right]. \tag{8}$$

The fitted intensity therefore contains a side term, the same strength differential, a late-match time component, and the shrinkage team effects. The prior match count used in these team effects is

selected from the historical sensitivity exercise under the held-out outcome log score. The 2026 event-time generator therefore uses a documented validation rule rather than an arbitrary prior strength.

The benchmark set also includes an empirical-characteristic-function GMM variant. This is not presented as the main estimator. Its role is to ask whether a moment-based distributional calibration can add information beyond the likelihood-based generator. Let $\mathrm{i} = \sqrt{-1}$, and let $\varphi_m(u;\theta) = \sum_{g_A,g_B} \exp\{\mathrm{i}(u_A g_A + u_B g_B)\} p_\theta(g_A, g_B \mid X_m)$ denote the generator-implied characteristic function for fixture m, evaluated on a fixed grid of frequencies u. The ECF-GMM benchmark starts from the team-specific generator and estimates a small calibration vector that rescales the overall goal rate, the strength slope, and the shrinkage team effects. The criterion minimizes the squared distance between the average model-implied characteristic function and the empirical characteristic function of training scorelines,

$$\widehat{\theta}_{\mathrm{ECF}} = \arg\min_{\theta} \sum_{u \in \mathcal{U}} \left| \frac{1}{n} \sum_{m=1}^{n} \varphi_m(u;\theta) - \frac{1}{n} \sum_{m=1}^{n} \exp\{\mathrm{i}(u_A G_{A,m} + u_B G_{B,m})\} \right|^2 . \tag{9}$$

The implemented version regularizes the calibration vector toward the likelihood-based generator and selects the penalty on a time-ordered split of the calibration window. This prevents the characteristic-function match from collapsing team-strength effects when the calibration window is short. The benchmark is attractive because it estimates a distributional object directly, while remaining agnostic about individual scoreline likelihood contributions. In the current implementation it is used as a historical comparator and table column, not as the primary 2026 forecast engine.

The historical tables use these equations as model definitions. The final-score generator is the global version of Equation (5) with $a_i = d_j = 0$. The team-specific generator adds the shrinkage effects in Equation (6). The regularized ECF-GMM generator applies the characteristic-function calibration in Equation (9) to the team-specific generator. The event-time generator is defined by the intensity and likelihood in Equations (7)–(8), and its implied score distribution is again the transition law in Equation (3).

Fourth, each 2026 fixture is forecast by composing the fitted team-specific generator with the two teams in the fixture and propagating the birth process over regular time on a bounded score grid.

This produces

$$\widehat{p}(g_A, g_B \mid X_m), \qquad 0 \le g_A, g_B \le G_{\max}, \tag{10}$$

for fixture m. Win, draw, and loss probabilities are obtained by summing this distribution over the corresponding score regions. The primary printed forecast uses this event-time score distribution as its scoreline shape and then recalibrates the three result-class masses by a convex combination of the generator and the calibrated random forest. The combination weight is selected on historical held-out World Cup outcome log score. Within each result class, relative exact-score probabilities remain those of the generator; only the aggregate win–draw–loss mass is rescaled. This keeps the forecast a coherent score distribution while allowing the best categorical benchmark to contribute information about result-class calibration.

The generator language is useful because it separates three layers that are often conflated. The first layer is the event-time law: how goals arrive conditional on the current state. The second layer is the final-score law: the transition distribution at full time. The third layer is the decision layer: how a forecaster maps the distribution into a reported score, a W/D/L probability, or a total-goals point forecast. A direct categorical model skips the first two layers and estimates a W/D/L probability vector. That can be statistically efficient for the categorical target. It cannot, by itself, answer exact-score or total-goals questions without an additional distributional model.

The model also clarifies why simple Poisson score models are not the end of the problem. Independent Poisson rates produce a score distribution, but they do not automatically encode score-state dependence, time-varying incentives, or opponent-specific interactions. Dixon–Coles-type corrections can adjust low-score dependence, but the generator formulation provides a more general state-space object. The point is not that every extension is automatically useful. The point is that extensions can be evaluated as additions to a common final-score distribution and scored on held-out forecast objects.

Estimation is performed only on information available before the forecast vintage. For a historical World Cup c, all training matches must predate the tournament cutoff. For event-time likelihoods, goal-event records must be internally consistent with final scores and available before the held-out tournament. This timing restriction is what makes the backtest interpretable. Without it, the richer generator would simply become another channel for leakage.

4 Forecast Targets and Decisions

The paper distinguishes forecast objects from decisions. A forecast object is a probability distribution or probability vector. A decision is an action chosen under a loss or payoff function. The final-score distribution $p_\theta(g_A, g_B \mid X)$ is the primitive forecast object in the generator approach. A reported exact score under a customized payoff is a decision. The distinction matters because the score that maximizes expected payoff need not be the modal scoreline.

Let $Y = (G_A(90), G_B(90))$ denote the realized final score, and let $w(Y) \in \{A, D, B\}$ denote the realized result class. The contest workbook uses stage-specific scoring. In the group stage, the submitted score receives two points if it has the correct result class and zero points otherwise; there is no exact-score bonus. For a reported group-stage score $s = (g_A, g_B)$, expected payoff is

$$\mathbb{E}\{\text{points}(s, Y) \mid X\} = 2 \Pr\{w(s) = w(Y) \mid X\}.$$

In the knockout phase, an exact score receives five points; if the exact score is missed but the result class is correct, the score receives two points. The knockout expected payoff is therefore

$$\mathbb{E}\{\text{points}_{\text{KO}}(s, Y) \mid X\} = 5 \Pr(Y = s \mid X) + 2 \Pr\{w(s) = w(Y), Y \neq s \mid X\}.$$

The optimal score decision is the score s that maximizes the relevant stage-specific expression over an admissible score grid. If p_θ is diffuse, a slightly less likely exact score can dominate the modal score because it has a more valuable result class. This is why the paper reports exact-score likelihood and payoff points separately. A model can be better probabilistically but worse for a particular payoff, and vice versa.

Win–draw–loss evaluation uses the probability vector

$$q_\theta(X) = \left(\Pr(A \text{ wins} \mid X), \Pr(D \mid X), \Pr(B \text{ wins} \mid X)\right).$$

Direct categorical models estimate $q(X)$ without modelling the scoreline cells. Ordered logit imposes an ordinal structure on the result classes, multinomial logit estimates a direct categorical response, and calibrated machine-learning models provide flexible alternatives using the same pre-match information set. These models are therefore valid W/D/L benchmarks, but they are not complete substitutes for a final-score distribution.

Tournament total goals are another functional of the match-level score distributions. Let M denote the tournament fixture set and $T = \sum_{m \in M}(G_{A_m} + G_{B_m})$ denote total goals, excluding shootouts. If the contest target excludes in-match penalty goals, the all-goals distribution must be transformed to a non-penalty target. The current pipeline uses historical goal-event coverage to estimate a penalty-goal share from prior information and maps the all-goals distribution into the contest-counted distribution. Under absolute-error loss, the optimal integer forecast is the predictive median. Under squared-error loss, it is the rounded mean. The contest wording therefore determines the point forecast.

Forecast vintages are treated as part of the estimand. A pre-playoff forecast must integrate over unresolved playoff slots using source-stamped candidate probabilities. A post-playoff forecast may condition on the resolved field, but only after the playoff outcomes are known and recorded. A pre-tournament forecast may use the final field, final ranking snapshot, and source-stamped market or player information available at the freeze date. These vintages are not interchangeable. A method that uses final playoff results cannot be evaluated as a pre-playoff forecast.

The 2026 application uses this discipline operationally. The current group-stage table is a freeze candidate based on the loaded field and ranking snapshot. The post-playoff update path is implemented, but it still requires source verification against final playoff results. Optional player and lineup schemas are present, but they do not enter active covariates because the source-stamped rows are not yet populated. This conservative treatment prevents the paper from claiming predictive gains from data blocks that are not yet empirically active.

5 Data and Information Sets

The empirical design depends on what is known at each forecast date. The project separates raw source inputs, processed feature tables, metadata manifests, and report artifacts. This organization is not only a software choice. It is part of the econometric design. A forecast is meaningful only relative to an information set, and each information block must have timing and provenance.

The historical backbone is an international match-results database. These rows supply final scores, dates, teams, venues where available, and the match outcomes used for training and backtesting. Rolling team-strength features are constructed so that a match forecast uses only information available before that match. These features are intentionally sparse. They provide a clean baseline

information set for comparing the generator, logit models, and machine-learning benchmarks.

Goal-event data supply timing information. The event-time likelihood uses goal counts over minute bins, with exposure offsets and team attack/defense effects. Goal-event rows are used only when their final-score totals are internally consistent with the historical results. This consistency check matters because an event-time model can otherwise become less reliable than the simpler final-score likelihood it is meant to enrich. The timing data are also used in the 2026 forecast as a late-match adjustment.

Market odds are available for a historical overlap sample. The implemented market benchmark converts decimal 1X2 odds into no-vig implied probabilities. These probabilities are valuable because they summarize a high-information market reference. They are also limited because the current historical odds are matchday or closing prices. The paper therefore uses them to measure distance from a market frontier on matched fixtures, not to claim that the statistical models were beaten in a strict pre-tournament setting.

Player, lineup, injury, and news covariates are not active predictors in the current estimates. The pipeline contains schema and missingness audits for these data, but the source-stamped rows have not yet been populated. This is the correct status for a scientific paper. The model can discuss how these covariates would enter the generator, and the pipeline can validate their shape, but the paper cannot claim that they improve predictive ability until the rows exist, the licenses are reviewed, and the incremental value is backtested.

The 2026 forecast layer contains group-stage fixtures, team-strength snapshots, forecast manifests, total-goals simulation output, and post-playoff update templates. The current loaded field supports a group-stage score table and a 104-match total-goals simulation. The pre-playoff workflow is implemented as an algorithm, but real slot probabilities are still required. The post-playoff workflow is also implemented, but the resolved field must be verified against a source-stamped file before it becomes a final contest vintage.

Table 1 summarizes the information-set status. Historical results and team strength are fully active. Goal-event timing is active with high minute coverage. Market odds are active only as a historical reference frontier. Player, lineup, and news rows are schema-only. Forecast vintages exist as workflow artifacts but still require final source locks before a final forecast freeze. This table is deliberately included in the main paper because it prevents overclaiming. It tells the reader which

Table 1: Information-set availability by empirical layer.

The table separates implemented information blocks from partial or unavailable covariate blocks. It is a leakage and claim-control audit: rows identify the timing control, active model use, and limitation for each empirical layer.

Block	Rows	Coverage	Active use	Limitation
Historical results and strength	49306	strength 100.0%; as-of 100.0%	generator, categorical, and ML benchmarks	sparse team-strength feature set
Goal-event timing	47601	minute 99.5%; as-of 100.0%	timing adjustment and event-time likelihood	team-shrinkage effects before richer player and match-context effects
Market odds	3234	overlap available; matched odds 2316; matches 1001	no-vig 1X2 benchmark	not a strict pre-tournament information set
Player, lineup, and news	0	status contract only; historical pair 0.0%; 2026 pair 0.0%	excluded from active covariates until source-stamped rows exist	no licensed/source-stamped player rows currently available
World Cup 2026 fixture strength	72	strength 100.0%; snapshot 100.0%	2026 exact-score and custom score decisions	needs final source freeze before final forecast use
Forecast vintages	3	pre-playoff 3/3; post-playoff 6/6; pre-tournament 9/9	forecast-vintage workflow and post-playoff update path	real pre/post-playoff vintages still need final inputs

empirical claims can be supported now and which require additional data.

The data design also clarifies the missingness problem. More data do not automatically improve estimates if their missingness is endogenous or if their timestamps are weak. Player participation, for example, is valuable only if absence, uncertainty, and late lineup revelation are recorded in a way that corresponds to the forecast horizon. Otherwise, the model may learn from post-kickoff information. The current treatment is therefore conservative: missing player information is not imputed into active forecasts; it is represented as an absent information block until source-stamped records exist.

This discipline is central to the paper's claim. The current backtests are fair for the sparse information set they use. They are not a final ranking of all conceivable sports-forecasting methods. A richer paper-grade vintage should add timestamped odds, squad snapshots, player-form aggregates, injury flags, travel and rest covariates, and venue conditions. Each addition should be evaluated incrementally under the same fixed-vintage design.

6 Econometric Evaluation Design

The evaluation design follows the forecast object. Exact-score distributions are evaluated with exact-score log score. Win–draw–loss probability vectors are evaluated with outcome log score, Brier score, and ranked probability score. Custom score decisions are evaluated with expected and

realized payoff points. Tournament total-goals distributions are evaluated with absolute error, log score, and distributional scores where available. These scores answer different questions, so the paper does not collapse them into a single ranking.

Proper scoring rules are central because they reward honest probability forecasts. The outcome log score penalizes assigning low probability to the realized class. The Brier score measures squared probability error across classes. The ranked probability score is useful for ordered categories because it compares cumulative probabilities. Exact-score log score evaluates the full scoreline distribution at the realized score. These scores are not cosmetic alternatives. They define the statistical target that a model is trying to forecast.

Historical backtests use fixed forecast vintages. For each held-out World Cup, all training data must predate the cutoff. Forecasts are evaluated on common matches. The common-match restriction is especially important when comparing models with different output objects or data requirements. A model should not appear better merely because it is evaluated on an easier subset. When a market benchmark is included, the comparison is restricted to fixtures with matched odds. Because the available historical odds are matchday or closing prices, those rows are interpreted as a market-reference frontier rather than as a strict pre-tournament comparison.

The main statistical uncertainty is tournament-level dependence. Matches within a World Cup share shocks: tournament conditions, qualification composition, tactical trends, travel environments, and team-selection information. The paper therefore reports uncertainty clustered at the World Cup level where paired comparisons are made. This is a conservative design with only six held-out World Cup vintages. It is not intended to deliver overconfident dominance claims. It is intended to prevent small average score differences from being interpreted as decisive evidence.

Predictive-ability sets are used in the spirit of model confidence set comparisons (Hansen et al., 2011). For a given score, the mean-best model defines the reference loss. Other models enter the predictive-ability set when the data do not detect inferior predictive ability at the chosen level after clustering by World Cup. This framework is more informative than a leaderboard alone. It separates point-estimate rankings from uncertainty about those rankings. In the current sparse-covariate design, that distinction matters: the calibrated random forest is mean-best for W/D/L proper scores, but the generator and logit models are not detectably worse.

The paper also reports bootstrap rank distributions. Tournament-cluster bootstrap draws resam-

ple held-out World Cups and recompute mean losses. The resulting rank-one probabilities show whether a model's apparent advantage is stable across vintages or concentrated in a small number of tournaments. This diagnostic does not replace proper score comparisons. It adds a stability statement that is useful when the sample of tournaments is necessarily small.

Payoff-point evaluation is deliberately kept separate from probabilistic evaluation. The custom score rule is a decision problem. Optimizing reported scores for expected points can improve payoff value even if the underlying probability model is unchanged. Conversely, a model could improve proper scores without changing the reported exact score. The correct comparison is therefore two-layered: first evaluate probability forecasts, then evaluate the decision rule applied to those forecasts.

Runtime is not a ranking criterion. Computational feasibility matters for reproducibility and for producing forecasts before a contest deadline, but the scientific question is predictive information. A method matters if it improves a prespecified forecast object under the same information set, or if it reveals structure that the generator does not yet use. The evaluation therefore treats random forests, gradient boosting, logit models, generator likelihoods, and market probabilities as sources of predictive information, not as competitors in a speed contest.

7 Historical Backtest Results

The historical evidence is best read as a sequence of forecast-object comparisons. The first question is whether additional generator structure adds predictive information beyond a final-score likelihood. The second is whether categorical econometric and machine-learning methods dominate when the target is only win–draw–loss. The third is how far the statistical models are from market-implied probabilities where market odds are observed.

Table 2 reports the main ablation. The generator ladder separates the final-score generator in Equation (5), the team-specific extension in Equation (6), the regularized ECF-GMM calibration in Equation (9), event-time arrival information in Equations (7)–(8), and rolling forecast-combination weights. The categorical and market rows are kept in separate panels because they answer different questions. Categorical models produce W/D/L probabilities directly, while market prices are observed at a later information horizon on the overlap sample. This table is therefore the most direct empirical answer to the question of which information block improves which score.

Table 2: Incremental forecast information by model family.

The table separates the generator ablation from categorical and market reference panels. Lower values are better for log, Brier, RPS, and exact-score log losses; higher values are better for contest points. Bold cells mark the best value within each panel and scoring rule. Missing cells indicate that a model does not produce the required forecast object.

Model	Increment	Out. log	Brier	RPS	Exact log	Exp. pts	Real. pts	N
Generator ladder								
Final-score generator	Final-score likelihood	1.0008	0.5978	0.2106	**2.9072**	1.1349	1.0208	384
Team-specific generator	Adds shrunken team attack and defence effects	1.0405	0.6193	0.2201	2.9689	**1.1772**	0.9740	384
Regularized ECF-GMM generator	Adds regularized ECF-GMM calibration	1.0310	0.6154	0.2184	2.9489	1.1255	0.9896	384
Event-time generator	Adds goal timing and score-state arrival information	0.9832	0.5789	0.2017	2.9081	1.1744	**1.0885**	384
Rolling log-score stack	Adds rolling validation weights	**0.9732**	**0.5749**	**0.2006**	–	–	–	320
Categorical benchmarks								
Ordered logit	Ordered categorical benchmark	1.0030	0.5995	0.2113	–	–	–	384
Multinomial logit	Multinomial categorical benchmark	**1.0028**	**0.5994**	0.2114	–	–	–	384
Calibrated random forest	Flexible categorical benchmark	1.0045	0.6007	**0.2113**	–	–	–	384
Calibrated gradient boosting	Boosted categorical benchmark	1.0767	0.6525	0.2375	–	–	–	384
Market overlap								
Final-score generator	Final-score likelihood	1.0225	0.6155	0.2223	**2.9904**	**1.1137**	**1.0000**	192
No-vig market	Archived no-vig 1X2 probabilities	**0.9613**	**0.5667**	**0.2001**	–	–	–	192

The main score-by-method comparison appears in Table 3. The event-time generator is best on outcome log score, Brier score, and ranked probability score. This result suggests that goal timing and state-dependent arrival information contain useful signal for W/D/L probabilities. The final-score generator remains competitive for exact-score log score, which is consistent with its likelihood being fitted directly to full-time scorelines. The regularized ECF-GMM calibration improves some decision-oriented quantities, but it does not dominate the proper W/D/L losses. Realized contest points remain noisy in small tournament samples.

Table 3: Score-by-method leaderboard for historical World Cup backtests.

Rows are non-market forecasting methods and columns are score domains evaluated on held-out historical World Cup vintages where the score is available. Exact-score and payoff columns require a scoreline distribution and an exact-score decision; W/D/L-only models are evaluated on proper W/D/L scores and result-class points. Bold cells mark the best method within each score.

Model	Out. log	Brier	RPS	Exact log	Exp. pts	Real. pts	Result pts	N
Final-score generator	1.0008	0.5978	0.2106	**2.9072**	1.1349	1.0208	0.5104	384
Team-specific generator	1.0405	0.6193	0.2201	2.9689	**1.1772**	0.9740	0.4870	384
ECF-GMM generator	1.0310	0.6154	0.2184	2.9489	1.1255	0.9896	0.4948	384
Event-time generator	**0.9832**	**0.5789**	**0.2017**	2.9081	1.1744	**1.0885**	0.5443	384
Ord. logit	1.0030	0.5995	0.2113	–	–	–	0.5130	384
Mult. logit	1.0028	0.5994	0.2114	–	–	–	0.5130	384
RF	1.0045	0.6007	0.2113	–	–	–	**0.5443**	384
GBC	1.0767	0.6525	0.2375	–	–	–	0.4349	384

The predictive-ability set in Table 4 sharpens the same conclusion. Once the event-time generator is included in the strict pre-tournament W/D/L comparison, it is mean-best for all three proper

scores. Several simpler challengers are detectably worse under World-Cup-clustered uncertainty. Calibrated random forests remain a useful challenger, especially for outcome log score, but they do not dominate the structural event-time forecast in this design. The table therefore separates two statements that are easy to conflate: flexible categorical models are informative, but event-time arrival structure carries additional predictive signal for result probabilities.

Table 4: Predictive-ability set for strict pre-tournament W/D/L forecasts.

The table compares the strict pre-tournament W/D/L models on identical held-out World Cup matches. For each proper score, the mean-best model is shown in bold. Loss gap is the paired mean loss difference relative to the best model for that score; larger positive gaps are worse. The clustered one-sided p-value tests whether a model is detectably worse than the best model. In set means that inferior predictive ability is not detected at the 10 percent level with World-Cup-clustered uncertainty.

Model	Mean	Loss gap	Cl. p	Status
Brier score				
Event-time generator	0.5789	0.0000	–	Best
Final-score generator	0.5978	0.0189	0.010	Worse
Multinomial logit	0.5994	0.0205	0.010	Worse
Ordered logit	0.5995	0.0206	0.007	Worse
Calibrated random forest	0.6007	0.0218	0.052	Worse
ECF-GMM generator	0.6154	0.0365	0.014	Worse
Team-specific generator	0.6193	0.0404	0.002	Worse
Calibrated gradient boosting	0.6525	0.0736	0.005	Worse
Outcome log score				
Event-time generator	0.9832	0.0000	–	Best
Final-score generator	1.0008	0.0177	0.088	Worse
Multinomial logit	1.0028	0.0196	0.092	Worse
Ordered logit	1.0030	0.0198	0.078	Worse
Calibrated random forest	1.0045	0.0213	0.172	In set
ECF-GMM generator	1.0310	0.0478	0.048	Worse
Team-specific generator	1.0405	0.0574	0.005	Worse
Calibrated gradient boosting	1.0767	0.0935	0.017	Worse
Ranked probability score				
Event-time generator	0.2017	0.0000	–	Best
Final-score generator	0.2106	0.0089	0.007	Worse
Calibrated random forest	0.2113	0.0096	0.050	Worse
Ordered logit	0.2113	0.0096	0.007	Worse
Multinomial logit	0.2114	0.0097	0.007	Worse
ECF-GMM generator	0.2184	0.0167	0.021	Worse
Team-specific generator	0.2201	0.0184	0.005	Worse
Calibrated gradient boosting	0.2375	0.0358	0.005	Worse

The bootstrap rank distribution in Table 5 reinforces this conclusion. Event-time forecasts have the highest rank-one probability for the proper W/D/L scores. Random forests still receive nontrivial rank-one probability for outcome log score and frequently remain near the top two, but their advantage is not stable across the full set of proper scores. The evidence therefore points to a practical ranking, not a universal rule: flexible categorical models are useful challengers, while the event-time generator is the strongest current pre-tournament specification for W/D/L probabilities.

The event-time likelihood comparison in Table 6 isolates the structural question. Holding the his-

Table 5: Cluster-bootstrap rank distribution for W/D/L methods.

Each cell reports the tournament-cluster bootstrap probability that the method has rank one for the score. Bootstrap draws resample held-out World Cups with replacement and recompute mean losses on the tournament-level paired score matrix. Bold marks the largest probability within a score. The diagnostic turns point-estimate leaderboards into an uncertainty statement about method rankings.

Score	Final-score	Team-specific	ECF-GMM	Event-time	Ord.	Mult.	RF	GBC
Outcome log score	0.00	0.00	0.01	**0.86**	0.00	0.01	0.12	0.00
Brier score	0.00	0.00	0.00	**1.00**	0.00	0.00	0.00	0.00
Ranked probability score	0.00	0.00	0.00	**0.99**	0.00	0.00	0.00	0.00

torical World Cup vintages fixed, the event-time likelihood improves the W/D/L proper scores and expected contest points relative to the final-score likelihood. It does not improve exact-score log score. The interpretation is straightforward. Event-time structure improves result-class probabilities, perhaps because timing and score-state effects help allocate mass across win, draw, and loss regions. Exact-score likelihood remains more closely aligned with exact-score cells.

Table 6: Historical World Cup comparison of final-score and event-time likelihoods.

The event-time likelihood is a piecewise Poisson counting-process fit over team-by-minute-bin goal counts with exposure offsets and shrinkage team attack/defense effects. For each held-out World Cup, it is estimated only from prior World Cup matches whose goal-event records are internally consistent with final scores. The table compares this likelihood route with the final-score likelihood on the same fixed pre-tournament vintages. Lower values are better for loss scores; higher values are better for points.

Model	Out. log	Brier	RPS	Exact log	Exp. pts	Real. pts	N
Final-score likelihood	1.0008	0.5978	0.2106	**2.9072**	1.1349	1.0208	384
Event-time likelihood	**0.9832**	**0.5789**	**0.2017**	2.9081	**1.1744**	**1.0885**	384

Table 7 turns the same question into a rolling forecast-encompassing diagnostic. For each challenger, the table asks whether a two-model stack trained only on earlier held-out World Cups assigns positive validation weight to the challenger and improves later W/D/L scores relative to the final-score generator. This is a stricter interpretation than a leaderboard. A model can have a favourable mean score and still fail to add stable incremental information once the reference generator is already present.

Market overlap provides a demanding reference. Table 8 restricts all rows to fixtures with matched odds. No-vig market probabilities dominate W/D/L proper scores and realized result points. This is not surprising: matchday or closing prices aggregate information that pre-tournament statistical models do not have. The important point is horizon discipline. A closing-price comparison shows the distance from the market frontier. It does not show that a pre-tournament statistical model failed relative to a pre-tournament market forecast, because that market forecast is not yet observed

Table 7: Forecast-encompassing diagnostics by information horizon.

For each challenger, the table estimates a two-model rolling log-score stack using only earlier held-out World Cups, then scores the stacked forecast against the final-score generator on later World Cups. Negative score gaps favour the stack. The diagnostic tests whether a challenger contains incremental W/D/L information beyond the generator under the same forecast-vintage design.

Challenger	Weight	Log gap	Cl. p	Status	Interpretation
Market overlap					
No-vig market	0.6500	-0.0762	0.176	Suggestive	Improves mean log score; clustered evidence is weak.
Pre-tournament					
Event-time generator	0.9000	-0.0208	0.124	Suggestive	Improves mean log score; clustered evidence is weak.
Calibrated random forest	0.1900	-0.0080	0.156	Suggestive	Improves mean log score; clustered evidence is weak.
Multinomial logit	0.0000	-0.0000	–	Not rejected	No incremental log-score evidence.
Calibrated gradient boosting	0.0000	-0.0000	–	Not rejected	No incremental log-score evidence.
Ordered logit	0.0500	0.0004	0.374	Not rejected	No incremental log-score evidence.
ECF-GMM generator	0.0100	0.0004	0.374	Not rejected	No incremental log-score evidence.
Team-specific generator	0.0200	0.0009	0.374	Not rejected	No incremental log-score evidence.

in the same information set.

Table 8: Historical World Cup benchmark scores on market-overlap fixtures.

The table restricts all model rows to World Cup fixtures with matched Football-Data average 1X2 odds. Market prices are matchday or closing archived prices, so this is a market-reference comparison rather than a strict pre-tournament vintage. Lower values are better for log, Brier, and ranked-probability scores; higher values are better for contest points.

Model	Out. log	Brier	RPS	Exact log	Exp. pts	Real. pts	Result pts	N
Goal-arrival generator	1.0225	0.6155	0.2223	**2.9904**	1.1137	1.0000	–	192
Team-specific generator	1.0687	0.6438	0.2340	3.0585	1.1577	0.9479	–	192
ECF-GMM generator	1.0510	0.6356	0.2304	3.0428	1.1065	0.9583	–	192
Event-time generator	1.0249	0.6033	0.2159	3.0183	**1.1710**	**1.0833**	–	192
Ordered logit	1.0254	0.6167	0.2225	–	–	–	0.5104	192
Multinomial logit	1.0222	0.6151	0.2223	–	–	–	0.5104	192
Calibrated random forest	1.0135	0.6069	0.2183	–	–	–	0.5365	192
Calibrated gradient boosting	1.0772	0.6535	0.2423	–	–	–	0.4271	192
No-vig market	**0.9613**	**0.5667**	**0.2001**	–	–	–	**0.5573**	192

The custom score-decision backtest in Table 9 evaluates the decision layer. Optimizing the reported score for expected payoff changes many scorelines relative to modal exact-score choices. The expected-points gain is mechanically positive because the decision rule optimizes that objective under the model. The realized gain is positive in the historical sample but uncertain when clustered by World Cup. This is a useful distinction. The optimized rule is the correct decision under the model-implied distribution, but realized payoff gains remain noisy in small samples.

Total-goals evidence is more preliminary. Table 10 evaluates historical non-penalty tournament totals after transforming the all-goals distribution to the contest target. The rolling scaled generator improves over the raw generator and is competitive with the historical mean baseline, but the number of held-out tournaments is small. The 2026 side-question forecast should therefore be

Table 9: Historical custom score-decision backtest.

The table compares modal exact-score decisions with scores selected to maximize expected payoff points in historical pre-tournament World Cup vintages. Realized gains are clustered by World Cup.

Statistic	Value
Held-out matches	384
Different score decisions	263
Different result classes	236
Mean expected points, modal	0.8368
Mean expected points, optimized	1.1349
Mean expected-points gain	0.2981
Mean realized points, modal	0.8125
Mean realized points, optimized	1.0208
Mean realized-points gain	0.2083
Clustered p-value, realized gain	0.023
Exact-score hit rate, modal	0.1146
Exact-score hit rate, optimized	0.0964
Result-class hit rate, modal	0.4062
Result-class hit rate, optimized	0.5104

Expected-points score decisions improve realized payoff points in the historical backtest.

reported as a model-based freeze candidate rather than as a settled high-confidence estimate.

Table 10: Historical World Cup non-penalty total-goals backtest.

The table evaluates tournament total-goals forecasts after transforming the all-goals predictive distribution to the contest target that excludes in-match penalty goals. For each held-out World Cup, the penalty-goal share is estimated only from earlier World Cups with complete goal-event coverage, and the all-goals distribution is mapped to a non-penalty distribution through an exact binomial mixture. The rolling historical mean is a non-penalty goals-per-match point baseline. Lower absolute error, log score, and CRPS are better.

Model	Cups	Mean abs. err.	Median abs. err.	Log score	CRPS	80% hit
Goal-arrival generator	5	25.80	22.00	6.08	19.66	40.0%
Rolling historical mean	5	14.80	13.00	–	–	60.0%
Rolling scaled generator	5	**14.00**	14.00	4.41	**9.85**	60.0%

Taken together, the results support a disciplined conclusion. There is evidence that event-time structure adds predictive information for W/D/L outcomes. There is evidence that random forests are strong sparse categorical benchmarks. There is evidence that markets dominate when high-information prices are observed. There is not yet evidence that any one method dominates all targets, all information sets, and all forecast horizons.

8 The 2026 Forecast Vintage

The 2026 application translates the historical design into a forecast workflow. The current vintage produces group-stage exact-score forecasts, model-comparison artifacts, a deterministic knockout path, and a tournament total-goals forecast. The generator forecast is estimated as described in Section 3: historical matches are cut at the forecast date, team-strength and recent-form covariates are constructed without using future results, shrinkage attack and defense effects are estimated, and the event-time likelihood is fitted on source-consistent World Cup goal events. The primary group-stage score table uses the resulting event-time generator as the scoreline shape. A stacked win–draw–loss layer then recalibrates the three result-class masses using historical log-score weights for the event-time generator and the calibrated random forest. The purpose of this section is not to present the 2026 numbers as final truth. It is to show how the paper's information-set discipline governs a live forecasting problem.

The group-stage table printed below reports modal score forecasts from the stacked model. This is the scoreline with the largest estimated probability under the calibrated final-score distribution. It is therefore a forecast of the match score rather than a mechanical submission decision under the workbook's group-stage payoff rule. The replication files also store expected-points score decisions, which maximize the workbook rule that awards two points for the correct win–draw–loss class and no additional exact-score bonus. These two objects need not coincide. A draw can be the most likely exact score even when a narrow win maximizes expected contest points because the win-side result-class mass is larger than the draw mass. For the submitted tip files, the current vintage uses modal scorelines because they are the realistic score forecasts.

Table 11: Stacked modal group-stage score forecasts for the 2026 World Cup.

Date	Team A	Team B	Score	Date	Team A	Team B	Score	Date	Team A	Team B	Score
06-11	Mexico	S. Africa	1–0	06-11	S. Korea	Czechia	1–0	06-12	Canada	Bosnia-H.	1–1
06-12	USA	Paraguay	1–0	06-13	Qatar	Switzerland	1–1	06-13	Brazil	Morocco	1–1
06-13	Ivory C.	Ecuador	1–1	06-13	Haiti	Scotland	1–1	06-13	Sweden	Tunisia	1–1
06-14	Australia	Türkiye	1–1	06-14	Germany	Curaçao	3–0	06-14	Netherlands	Japan	1–1
06-15	Spain	C. Verde	3–0	06-15	Belgium	Egypt	1–0	06-15	Saudi	Uruguay	1–1
06-15	Iran	N. Zealand	2–0	06-15	Argentina	Algeria	2–0	06-16	Austria	Jordan	2–0
06-16	Portugal	DRC	2–0	06-16	France	Senegal	2–0	06-16	England	Croatia	1–0
06-16	Iraq	Norway	1–1	06-16	Ghana	Panama	1–1	06-16	Uzbekistan	Colombia	1–1
06-18	Czechia	S. Africa	1–1	06-18	Switzerland	Bosnia-H.	2–0	06-18	Canada	Qatar	1–1
06-18	Mexico	S. Korea	1–1	06-19	USA	Australia	1–1	06-19	Scotland	Morocco	1–1
06-19	Ecuador	Curaçao	2–0	06-19	Brazil	Haiti	3–0	06-19	Türkiye	Paraguay	0–1
06-20	Netherlands	Sweden	2–0	06-20	Germany	Ivory C.	2–0	06-20	Tunisia	Japan	0–1
06-20	N. Zealand	Egypt	0–2	06-21	Spain	Saudi	3–0	06-21	Belgium	Iran	1–1
06-21	Uruguay	C. Verde	2–0	06-22	Argentina	Austria	2–0	06-22	France	Iraq	3–0
06-22	Panama	Croatia	1–1	06-22	Norway	Senegal	1–1	06-22	Colombia	DRC	2–0
06-23	Portugal	Uzbekistan	2–0	06-23	England	Ghana	3–0	06-23	Jordan	Algeria	1–1
06-24	Switzerland	Canada	1–1	06-24	Bosnia-H.	Qatar	1–1	06-24	Scotland	Brazil	0–2
06-24	Morocco	Haiti	2–0	06-24	Japan	Sweden	1–0	06-24	Tunisia	Netherlands	0–2
06-24	Czechia	Mexico	1–1	06-24	S. Africa	S. Korea	1–1	06-24	Türkiye	USA	1–1
06-24	Paraguay	Australia	1–1	06-25	Ecuador	Germany	1–1	06-25	Curaçao	Ivory C.	1–1
06-25	C. Verde	Saudi	1–1	06-25	Uruguay	Spain	1–1	06-26	Egypt	Iran	1–1
06-26	N. Zealand	Belgium	0–2	06-26	Norway	France	0–2	06-26	Senegal	Iraq	2–0
06-26	Colombia	Portugal	1–1	06-26	DRC	Uzbekistan	1–1	06-26	Algeria	Austria	1–1

Date	Team A	Team B	Score	Date	Team A	Team B	Score	Date	Team A	Team B	Score
06-26	Jordan	Argentina	0–3	06-27	Panama	England	0–2	06-27	Croatia	Ghana	2–0

The current deterministic knockout path is obtained by applying the primary modal group-stage forecasts to the official qualification rules, resolving the best third-place teams, and then forecasting each subsequent knockout fixture conditional on the teams implied by the previous round. Knockout fixtures are also reported as modal score forecasts rather than expected-payoff decisions. Table 12 reports the resulting path through the final. If a knockout score is tied, the table advances the team with the higher model-implied win probability; the tied score itself is left unchanged because it is the model's score forecast.

Table 12: Primary deterministic knockout path for the 2026 World Cup.

Round	Match	Date	Team A	Team B	Score	Advances
Round of 32	73	06-28	S. Korea	Canada	1-1	S. Korea†
Round of 32	74	06-29	Germany	Australia	2-0	Germany
Round of 32	75	06-29	Netherlands	Morocco	1-0	Netherlands
Round of 32	76	06-29	Brazil	Japan	2-0	Brazil
Round of 32	77	06-30	France	Egypt	2-0	France
Round of 32	78	06-30	Ecuador	Senegal	1-1	Senegal†
Round of 32	79	06-30	Mexico	Ivory C.	1-1	Mexico†
Round of 32	80	07-01	England	Norway	2-0	England
Round of 32	81	07-01	USA	Qatar	1-0	USA
Round of 32	82	07-01	Belgium	Czechia	2-0	Belgium
Round of 32	83	07-02	Colombia	Croatia	1-1	Colombia†
Round of 32	84	07-02	Spain	Austria	2-0	Spain
Round of 32	85	07-02	Switzerland	Algeria	1-0	Switzerland
Round of 32	86	07-03	Argentina	Uruguay	2-0	Argentina
Round of 32	87	07-03	Portugal	Panama	2-0	Portugal
Round of 32	88	07-03	Paraguay	Iran	1-1	Iran†
Round of 16	89	07-04	Germany	France	1-1	France†
Round of 16	90	07-04	S. Korea	Netherlands	1-1	Netherlands†
Round of 16	91	07-05	Brazil	Senegal	2-0	Brazil
Round of 16	92	07-05	Mexico	England	1-1	England†
Round of 16	93	07-06	Colombia	Spain	1-1	Spain†
Round of 16	94	07-06	USA	Belgium	1-1	Belgium†
Round of 16	95	07-07	Argentina	Iran	2-0	Argentina
Round of 16	96	07-07	Switzerland	Portugal	1-1	Portugal†
Quarter-final	97	07-09	France	Netherlands	1-1	France†
Quarter-final	98	07-10	Spain	Belgium	1-0	Spain
Quarter-final	99	07-11	Brazil	England	1-1	Brazil†
Quarter-final	100	07-11	Argentina	Portugal	1-1	Argentina†
Semi-final	101	07-14	France	Spain	1-1	France†
Semi-final	102	07-15	Brazil	Argentina	1-1	Argentina†
Third-place play-off	103	07-18	Spain	Brazil	1-0	Spain

Round	Match	Date	Team A	Team B	Score	Advances
Final	104	07-19	France	Argentina	1-1	France[†]

[†] The predicted score is tied; the advancing team is selected by the higher model-implied win probability.

The total-goals side question asks for tournament goals excluding shootout goals and in-match penalty goals. The current simulation models 104 fixtures and produces a median of 285 goals before penalty adjustment. The expected in-match penalty adjustment is 25.8 goals, leaving a contest-counted median and point forecast of 259 goals. The current 10th–90th percentile range is 239–279. Table 13 reports the implemented forecast summary.

Table 13: Implemented forecast for total tournament goals.

Simulation-based tournament total. The contest count excludes shootouts and in-match penalty goals.

Quantity	Value
Modeled fixtures	104
Simulation paths	1000
Median total goals before penalty adjustment	285
Expected in-match penalty goals	25.8
Median contest-counted goals	259
Mean contest-counted goals	259.2
10th percentile	239
90th percentile	279
Contest point forecast	259

The forecast should still be read as a freeze candidate. Several gates remain. First, the post-playoff field must be verified against source-stamped playoff outcomes. Second, real pre-playoff slot probabilities are required if a pre-playoff vintage is reported. Third, forecast-horizon market odds must be populated if the final paper claims a horizon-matched market comparison. Fourth, player and lineup covariates must remain excluded until licensed timestamped rows are available. Finally, the selected total-goals point source and reported integer must be recorded before the forecast number is frozen. The replication package keeps a machine-readable claim audit, but the journal manuscript reports the implication in prose: the current evidence supports historical method comparisons and freeze-candidate 2026 outputs, not final player-aware performance claims or horizon-matched market parity.

9 Conclusion

This paper develops a generator-centered approach to World Cup forecasting and evaluates it as a method-performance problem. The key object is a conditional distribution over final scores. From that object one can derive win–draw–loss probabilities, exact-score decisions, tournament total-goal distributions, and customized payoff decisions. This structure is useful because it keeps the forecast object coherent while allowing different downstream decisions to be evaluated under their own loss or payoff functions.

The empirical evidence supports three conclusions. First, event-time goal-arrival structure contains useful information for win–draw–loss prediction in the historical World Cup backtests. The event-time generator improves the three W/D/L proper scores relative to the final-score likelihood. Second, the final-score generator remains competitive where the scoreline itself is the target: it is best on exact-score log score in the current backtest. Third, calibrated random forests are strong categorical benchmarks on sparse pre-tournament covariates, but their point-estimate lead does not imply statistical dominance over the generator and logit models under World-Cup-clustered predictive-ability comparisons.

The market comparison gives the expected discipline check. On fixtures with matched historical odds, no-vig market probabilities dominate the statistical models for W/D/L prediction. This does not invalidate the generator exercise. It shows that market prices remain a high-information frontier when they are observed near the match. The open empirical question is horizon-specific: whether source-stamped opening or pre-tournament market odds add information beyond the generator and categorical models at the same forecast date.

The 2026 application is best viewed as a freeze-candidate forecasting system. It produces group-stage exact-score decisions under the customized scoring rule and a total-goals forecast aligned with the side target. It also implements the post-playoff update path. However, the final forecast vintage still requires verified playoff outcomes, source-stamped market odds, final fixture and ranking inputs, and a documented data freeze. Player, lineup, injury, and news covariates remain schema-level inputs until licensed timestamped rows are populated.

The broader lesson is methodological. Predictive performance depends on the forecast object, the information set, and the scoring rule. The evidence is mixed across these dimensions. Event-time structure helps W/D/L probabilities. Final-score calibration matters for exact scores. Random

forests are strong in sparse categorical prediction. Markets dominate where high-information prices are available. The evidence therefore supports forecast-object discipline, leak-free information sets, and paired predictive-ability comparisons rather than a single-method dominance claim.

References

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19(4), 1483–1536.

Baio, G., and Blangiardo, M. (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, 37(2), 253–264.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.

Dixon, M. J., and Coles, S. G. (1997). Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C*, 46(2), 265–280.

Dyte, D., and Clarke, S. R. (2000). A ratings based Poisson model for World Cup soccer simulation. *Journal of the Operational Research Society*, 51(8), 993–998.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6), 985–987.

Forrest, D., Goddard, J., and Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21(3), 551–564.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232.

Giacomini, R., and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74(6), 1545–1578.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C*, 48(3), 377–394.

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

Goddard, J. (2005). Regression models for forecasting goals and match results in association football. *International Journal of Forecasting*, 21(2), 331–340.

Groll, A., Schauberger, G., and Tutz, G. (2015). Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: An application to the FIFA World Cup 2014. *Journal of Quantitative Analysis in Sports*, 11(2), 97–115.

Groll, A., Ley, C., Schauberger, G., and Van Eetvelde, H. (2019). A hybrid random forest to predict soccer

matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4), 271–287.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497.

Hvattum, L. M., and Arntzen, H. (2010). Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3), 460–470.

Karlis, D., and Ntzoufras, I. (2003). Analysis of sports data by using bivariate Poisson models. *Journal of the Royal Statistical Society: Series D*, 52(3), 381–393.

Koopman, S. J., and Lit, R. (2015). A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *Journal of the Royal Statistical Society: Series A*, 178(1), 167–186.

Leitner, C., Zeileis, A., and Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26(3), 471–481.

Ley, C., Van de Wiele, T., and Van Eetvelde, H. (2019). Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, 19(1), 55–77.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, 36(3), 109–118.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*. Academic Press.

Rue, H., and Salvesen, O. (2000). Prediction and retrospective analysis of soccer matches in a league. *Journal of the Royal Statistical Society: Series D*, 49(3), 399–418.

Schauberger, G., and Groll, A. (2018). Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5–6), 460–482.

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G., and Kiraly, F. (2019). Modeling outcomes of soccer matches. *Machine Learning*, 108(1), 77–95.

White, H. (2000). A reality check for data snooping. *Econometrica*, 68(5), 1097–1126.