SURVEYING THE FRONTIER FOR FOOTBALL BETTING MODELS: A BREADTH-FIRST THEN DEPTH-FIRST LITERATURE REVIEW WITH RANKINGS AND EVALUATION CRITERIA

RICHARD OLDHAM

ABSTRACT. We construct a rigorous, two-pass literature review (breadth-first, then depth-first) to map the state of the art for modeling American (NFL, NCAA) football outcomes and for pricing derivative bets (alt lines, teasers, pleasers, SGPs). The review targets methodological families directly useful for our discrete-margin/key-number engine: (i) dynamic state-space and hierarchical Bayes for team strengths, (ii) discrete scoring models (Poisson/Skellam and dynamic bivariate variants), (iii) spread→win-probability calibrations, (iv) machine-learning pipelines for win probability (WP) and exact-score forecasting, and (v) market microstructure/efficiency papers that explain sportsbook pricing, shading, and bettor bias. We formalize evaluation criteria (methodological novelty, empirical rigor, calibration quality, reproducibility, domain-transfer to NFL/NCAA, and practical value for pricing teasers/alt-lines), score and rank candidate works, and present a Top 10 with annotated strengths/limitations. We conclude with a research agenda aligned to the discrete-margin engine: NFL-vs-NCAA key-number conditioning, dynamic reweighting around $\{3, 6, 7, 10\}$, posterior EV distributions for teaser portfolios, and correlation-aware SGP pricing.

1. SCOPE AND GUIDING QUESTION

Our goal is to identify academic methods that most effectively power a *pricing-first* toolkit for football betting: produce calibrated integer-margin pmfs $P(M = k)$ and downstream prices for spreads, alt lines, teasers, pleasers, and SGP correlations. We emphasize papers whose models (a) natively handle discrete scoring and/or dynamic team strength, (b) yield well-calibrated probabilities, and (c) *transfer* to NFL/NCAA (where key numbers dominate decision value).[1]

2. SEARCH PROTOCOL: BREADTH FIRST, THEN DEPTH FIRST

Breadth-first pass (discovery). We cast a wide net across: *Journal of the American Statistical Association* (JASA), *The American Statistician* (TAS), *JRSS* (Series A/C), *Journal of Applied Statistics*, *International Journal of Forecasting*, *Journal of Quantitative Analysis in Sports* (JQAS), *Journal of Sports Economics*, and major preprint repositories (arXiv: stat.AP/stat.ML; SSRN).[2] Candidate clusters emerged:

(1) **Dynamic team-strength models:** state-space and paired-comparison (Elo-like) approaches in NFL; hierarchical pooling for variance control.
(2) **Discrete scoring models:** independent Poisson (Maher-style), Dixon–Coles dependence corrections, bivariate Poisson and dynamic bivariate Poisson; Skellam and generalizations.
(3) **Spread calibration and exact-score forecasting:** mapping point spreads to win probabilities; NFL exact-score models.
(4) **Machine learning for WP:** in-play/drive-level RF/GBM pipelines and calibration diagnostics.
(5) **Market efficiency/microstructure:** bookmaker behavior, bias/longshot effects, and implications for pricing menus (teasers/alt-lines).

---

[1]Throughout, we adopt the discrete-margin framework developed earlier: fair point-buys and teaser values equal weighted sums of integer-atom probabilities across crossed thresholds.

[2]Key phrases included: "state-space NFL team strength", "bivariate Poisson football", "Skellam difference model", "exact score NFL forecasting", "spread to win probability", "sportsbook market efficiency", and "random forest win probability NFL".

Depth-first pass (selection and appraisal). Within each cluster we prioritized (i) seminal origins, (ii) well-cited refinements with better calibration, and (iii) works demonstrably portable to NFL/NCAA score mechanics. For each candidate we extracted: data scope, likelihood or algorithmic core, treatment of discreteness/dependence, validation (Brier/log-loss, reliability curves), and code/data availability. We then scored against the criteria in §3.

3. Evaluation criteria and scoring rubric

We evaluate each paper on a 0–5 scale per criterion (higher is better):

C1. Novelty/Influence (weight w_1=0.20): conceptual innovation and field impact.
C2. Empirical Rigor (weight w_2=0.20): sample size, holdout strategy, and statistical reporting.
C3. Calibration/Prob. Quality (weight w_3=0.20): reliability, proper scoring rules, discrete fit at key margins.
C4. Reproducibility (weight w_4=0.10): data/code and implementation clarity.
C5. Domain Transfer (NFL/NCAA) (weight w_5=0.15): applicability to American football scoring/idiosyncrasies.
C6. Practical Pricing Value (weight w_6=0.15): usefulness for alt-line/teaser/SGP pricing within the discrete-margin engine.

Total score $S = \sum_j w_j s_j$ with $s_j \in [0, 5]$. Ties are broken by calibration (C3) then transfer (C5).

4. Top-10 papers: rankings, synopses, and rationale

Table 1 summarizes our scores.[3]

Table 1. Top-10 works for football betting models (higher S is better).

#	Work (short citation)	C1	C2	C3	C4	C5	C6	S
1	Glickman–Stern (1998) [**?**]	5	5	4	3	5	5	4.55
2	Dixon–Coles (1997) [**?**]	5	4	4	3	4	5	4.30
3	Baio–Blangiardo (2010) [**?**]	4	4	4	3	4	5	4.05
4	Stern (1991) [**?**]	4	4	4	4	5	4	4.05
5	Koopman et al. (2015) [**?**]	4	5	4	3	4	4	4.00
6	Karlis–Ntzoufras (2003) [**?**]	4	4	4	3	3	4	3.80
7	Harville (1980) [**?**]	4	4	3	2	5	3	3.55
8	Lock–Nettleton (2014) [**?**]	3	4	4	3	4	3	3.55
9	Sauer (1998) [**?**]	4	4	3	3	4	3	3.50
10	Levitt (2004) [**?**]	4	4	3	2	4	3	3.35

1. Glickman & Stern (1998) [**?**] — **Dynamic state-space for NFL scores.** *Core:* Gaussian state-space linking latent team strengths to observed score *differences*; strengths evolve week-to-week.
Why it ranks #1: NFL-specific, time-adaptive, and directly outputs a calibrated distribution for margins.[4] *Toolkit role:* backbone for Stage-A ratings; plug into discrete reweighting for key numbers; yields posterior uncertainty for EV and Kelly sizing.

---

[3]Scores reflect combined judgments from model diagnostics reported in each paper and our assessment of transfer to discrete-margin pricing. For compactness, we omit raw rubric sheets; they can be provided as supplementary material.

[4]Score differences can be post-processed into a discrete $P(M = k)$ via continuity correction, then reweighted at key integers.

2. Dixon & Coles (1997) [?] — Poisson with short-term dependence. *Core:* Bivariate goal model with dependence correction for low scores; seminal for discrete scoring.
Transfer: Although developed for soccer, the bivariate framework and DC adjustment guide NFL/NCAA discrete modeling (swap goals for scoring events). *Toolkit role:* informs our categorical/Skellam layers and dependence handling; strong theoretical anchor for discrete atoms.

3. Baio & Blangiardo (2010) [?] — Hierarchical Bayes for team strengths. *Core:* Partial pooling of offense/defense effects with league-level priors; produces posterior predictive distributions for scores/results.
Transfer: Natural for NCAA (conference-level pooling) and early NFL weeks. *Toolkit role:* Stage-A Bayesian layer to stabilize thin data; integrates seamlessly with discrete-margin reweighting.

4. Stern (1991) [?] — Spread ⇒ win-probability mapping. *Core:* Normal-approximation map from spread to ML win probability; empirical calibration on NFL data.
Toolkit role: Converts market spreads to priors/targets; baseline for alt-line pricing; sanity check for model calibration.

5. Koopman, Lit, Lucas (2015) [?] — Dynamic bivariate Poisson. *Core:* Time-varying intensities in a bivariate Poisson for both teams' scoring; supports dynamic dependence.
Transfer: Provides a principled route to discrete $P(M = k)$ while adapting to form; more faithful than static Poisson. *Toolkit role:* Drop-in for Stage-B (continuous→discrete) with fewer ad hoc reweights.

6. Karlis & Ntzoufras (2003) [?] — Bivariate Poisson for sports. *Core:* Foundational treatment of joint scoring processes; addresses covariance explicitly.
Toolkit role: Baseline parametric family for margin pmfs; extend with dynamic parameters or COM-Poisson for over/under-dispersion.

7. Harville (1980) [?] — Linear-model predictions for NFL. *Core:* Mixed linear models for NFL outcomes; early formalization of team-strength prediction.
Toolkit role: Simplicity and transparency; useful benchmark and feature generator for modern ML.

8. Lock & Nettleton (2014) [?] — Random forests for NFL win probability. *Core:* Drive/play-level RF for pre-play WP; strong calibration in later-game states.
Toolkit role: In-play module; powers live middles/derivative pricing and SGP correlations (side↔total).

9. Sauer (1998) [?] — Economics of wagering markets. *Core:* JEL survey of efficiency evidence and pricing dynamics; emphasizes mostly efficient but not perfect markets.
Toolkit role: Context for line shopping, vig structures, and why teaser/alt menus evolve; informs our execution layer.

10. Levitt (2004) [?] — Bookmaker microstructure. *Core:* Bookmakers exploit bettor preferences; prices need not be market-clearing.
Toolkit role: Theoretical rationale for shaded teaser/alt-line menus and for hunting outliers (our price-for-probability plays).

5. Synthesis: how these works power the discrete-margin engine

(1) **Stage-A team strengths.** Use a state-space (Glickman–Stern) or hierarchical Bayes (Baio–Blangiardo) backbone; in NCAA apply conference pooling. Outputs (μ, σ) or full posterior draws for the latent margin.

(2) **Stage-B discrete pmf.** Map to $P(M = k)$ via (i) Skellam/bivariate Poisson with dynamic intensities (Koopman et al., Karlis–Ntzoufras) or (ii) Normal-to-discrete with Dixon–Coles-style reweights at key integers $\{3, 6, 7, 10\}$ (NFL) and NCAA-specific sets. Calibrate with reliability curves and exact-margin fit around keys.

(3) **Pricing any derivative.** Stern's spread→probability map offers a baseline; our discrete pmf prices alt lines, teaser legs, and exact margins; ML (Lock–Nettleton) powers live/state-conditional pricing and SGP correlations.

(4) **Execution layer.** Sauer and Levitt explain why menus/payouts vary; their microstructure insights motivate shopping, combining legs across books, and preferring alt-parlays over teasers when menus are shaded.

6. Frontier directions and open problems

(1) **Key-number aware likelihoods.** Replace post-hoc reweighting with priors/likelihoods that *natively* elevate mass at $\{3, 6, 7, 10\}$ (NFL) and learned sets (NCAA), e.g., mixture-of-Skellams with atoms at key integers.

(2) **Posterior EV distributions.** Full Bayesian propagation to teaser/alt portfolios to size bets under parameter uncertainty (fractional Kelly on the EV distribution, not point estimates).

(3) **Correlation models for SGPs.** Joint side×total generative models with interpretable dependence (copula or shared-latent-drive processes) and game-state conditioning; validate with proper scoring rules.

(4) **Rule/trend drift.** Online change-point detection for PAT, 2-pt, and tempo changes; automatic retraining of discrete weights.

(5) **NCAA heterogeneity.** Multi-level priors (conference/program) with variance components that reflect scoring volatility; league-specific key-number maps.

Conclusion

The most valuable academic tools for a pricing-centric football betting stack are those that (i) infer *dynamic* team strengths, (ii) yield *discrete* and *calibrated* margin distributions, and (iii) connect cleanly to market mechanics. Our ranking elevates dynamic state-space and hierarchical Bayes (NFL-ready), discrete bivariate models (keys-aware), and calibration staples (Stern's mapping), augmented by ML for in-play and by market microstructure for execution. This synthesis directly empowers the discrete-margin/key-number engine to evaluate alt lines, teasers, and correlated parlays with PhD-level statistical rigor.

References