Saturday

TECHNICAL SPECIFICATION

# The Saturday Model

***A predictive rating and outcome-distribution system for college football***

Saturday prices a game as a full distribution over the final margin, not a single number. This document specifies the estimators, the constants, the validation record, and the walk-forward discipline behind every figure the product displays.

| | |
|---|---|
| REVISION | 16 August 2026 · reflects the shipped production pipeline |
| SCOPE | FBS college football, 2021–2026. Market-comparison validation covers 2022–2025, the window every current validation script runs over. |
| PUBLISHER | Saturday Analytics |

SUMMARY

# The model in one page

Saturday prices a college football game as a probability distribution over the final margin rather than as a single number. Win probability, the projected score, season win totals, and the projected playoff field are all read off that distribution.

The model has two layers, split on how fast the thing being measured changes.

**Team strength** is estimated from the current season only. In the transfer-portal and NIL era a roster can be unrecognizable one year later, so Saturday never assumes this year's team resembles last year's. Each team is seeded before week one by a prior built entirely from offseason information (last season's rating, recruiting talent, coaching changes, the transfer portal, returning production, and preseason market expectations), and that seed is then progressively overwritten by a ridge-regularized rating fit as real games are played.

**Structure** is estimated over five to six seasons, because it does not churn season to season: how wide the range of game outcomes is, what a turnover is worth, and how much home field is worth at each particular stadium.

Everything is produced **walk-forward**. A game is priced using only games that finished before it; the model never sees the result of the game it is pricing. The outside check on that discipline is the closing betting line, which Saturday does not use as an input. Pooled over 2022–2025, Saturday's margin error is 16.04 RMSE against the market's 15.22, and its win-probability calibration error is 2.2%.

The sections below give the exact estimators, hyperparameters, and constants. The document is written to be sufficient to rebuild the model.

| 16.04 | +0.84 | +0.32 | 2.2% |
|---|---|---|---|
| MARGIN RMSE 2022–2025 | GAP TO CLOSING LINE ALL WEEKS | GAP TO CLOSING LINE WEEK 5 ONWARD | WIN-PROBABILITY CALIBRATION ERROR |

CONTENTS



**How to read the numbers.** Margins are stated home-minus-away (the exported data files use an away-perspective convention). RMSE is root-mean-square error in points of margin: lower is better, and the closing line is the benchmark, not a target.

# 1 Design philosophy

Two commitments shape every part of the model.

## 1.1 Single-season strength, multi-year structure

A college roster can be almost unrecognizable from one year to the next: a quarterback transfers in, a coordinator leaves, a recruiting class arrives. Saturday therefore estimates *how good a team is* from the current season's games only.

What it does borrow across years are the quantities that genuinely do not change season to season — the width of the outcome distribution, the value of a turnover, and the home-field edge at a given stadium. Those are measured over five to six seasons. Team strength is kept current.

## 1.2 Forward prediction, not curve-fitting to the outcome

Saturday fits one set of team ratings that best explain the margins of games *already played*, as classical rating systems (Massey, Colley, the simple rating system) do. That is an explanatory fit over past results used to make a forward prediction, not a fit of a game to its own final score. The difference is enforced mechanically: pricing a week-$w$ game uses a rating fit trained only on weeks strictly before $w$ (§6), so the result being predicted is never in the training set.

Before any games exist the fit reduces to the offseason prior of §4, built from recruiting, roster and coaching information rather than any current-season scoreboard. The whole approach is then scored out of sample against an outside scorekeeper: the closing betting line, which Saturday does not use as an input (§11).

> **THE ONE-SENTENCE VERSION**
>
> Saturday starts every season from a projection built out of recruiting, the portal, returning production, and coaching changes; it updates each team's strength only from games that have already happened; and it is graded on games it had not yet seen.

# 2 Data

All inputs come from the CollegeFootballData API (CFBD), pulled per season and cached locally, with one addition: preseason market expectations inform the preseason prior of §4. No point spreads or game betting lines are used anywhere in the model.

| ENDPOINT | WHAT IT SUPPLIES |
|---|---|
| `/games` | Final scores, week, venue, neutral-site flag, team classification — regular season and postseason. |
| `/teams, /venues` | Conference, colors, abbreviations; venue id, capacity, elevation, attendance. |
| `/talent` | Composite team recruiting-talent score. |

| ENDPOINT | WHAT IT SUPPLIES |
|---|---|
| `/player/portal` | Transfers: player, origin, destination, rating (0–1), eligibility. |
| `/coaches` | Per-coach season history, including SP+ overall. |
| `/player/returning` | Percent of last year's production returning, overall and passing. |
| `/ppa/*, /games/teams` | Per-play value (quarterback and team efficiency); team turnover counts. |
| `/rankings` | AP and CFP committee polls — display and playoff seeding context only, never strength. |
| `/lines` | Closing spread and total. Used *only* as an out-of-sample benchmark and to fit the totals model's shrinkage. Never a strength input. |

Games with missing scores are dropped. Opponents below FBS are handled in two tiers: FCS teams with an estimable prior-season rating get their own rating, and every other non-FBS opponent is pooled into a single `NON_FBS` anchor, so buy games still inform the FBS ratings without adding 130 noisy free parameters.

> RANKING HIERARCHY
>
> Saturday uses the **AP Poll** from the preseason until the College Football Playoff committee releases its first poll of the season. From that release onward, **CFP rankings are the source of truth** and remain so for the rest of the season. The **Coaches Poll is never used.**
>
> Rankings drive display badges and supply seeding context to the playoff projection (§12.2). No poll is ever an input to team strength.

## 3 Notation and the rating equation

Let $\boldsymbol{R_t}$ be team $\boldsymbol{t}$'s strength rating, in points. The atomic modeling assumption is that a game's margin is the rating difference plus a home-field term plus noise:

$$\text{margin(home - away)} \approx R_{\text{home}} - R_{\text{away}} + \text{HFA}_{\text{venue}} + \varepsilon$$

Two robustness rules apply to every rating fit. Margins are capped at `MOV_CAP = ±38` points before use, so a 63-point blowout cannot dominate the least-squares objective through garbage time. Ratings are identified only up to an additive constant, so after each fit the FBS ratings are re-centered to mean zero, with `NON_FBS` floating below them at typically around −19. The projected neutral-field edge for a hypothetical $\boldsymbol{A}$ vs $\boldsymbol{B}$ game is $\boldsymbol{R_A}$ − $\boldsymbol{R_B}$; adding the venue's home-field points gives the projected margin.

## 4 The preseason prior

Before any games are played, each FBS team is seeded with a prior assembled entirely from information available in the offseason. It is the sum of five additive terms:

$$\text{prior}_t = \gamma \cdot R_t^{(\text{prev})} + W_{\text{tal}} \cdot z^{\text{tal}}_t + W_{\text{coach}} \cdot \Delta\text{coach}_t + W_{\text{portal}} \cdot \text{portal}_t + \text{retLoss}_t$$

| TERM | WEIGHT | DEFINITION AND RATIONALE |
|---|---|---|
| **Carryover** | $\gamma = 0.95$ | Last season's final full-season rating (a plain ridge margin fit, §5, light penalty $\alpha = 1$), shrunk toward the mean. Teams promoted from FCS with no carryover start at the `NON_FBS` floor rather than at dead average. |
| **Recruiting talent** | 4.0 | The talent composite, standardized across FBS: $z^{\text{tal}}_t = (\text{talent}_t - \text{mean}) / \text{sd}$. Carryover alone badly compresses elite-vs-Group-of-5 mismatches, because a strong mid-major's record inflates its carry rating above its true level. |
| **Coaching change** | 0.3 | Points per point of coach-quality difference when the head coach changes. Catches, for example, a blueblood that carryover would otherwise keep elite after a marquee departure. |
| **Transfer portal** | 5.0 | Net portal flow, weighted by player rating and immediate eligibility. Weighting by rating captures coach-led mass migration automatically; an earlier net-*star-count* version was swamped by G5 volume and tuned to zero. |
| **Returning-production loss** | 3.0 | A downside-only penalty for a team gutted by departures that carryover and the portal would otherwise keep over-rated. |

The three terms with internal structure:

$$\text{coachQ}(c) = \left[ \sum_{y<Y} w_y \cdot \text{SP}_{c,y} \Big/ \sum w_y \right] \cdot n_{\text{eff}} / (n_{\text{eff}} + K)$$

$$w_y = 0.85^{(Y-1-y)}, \quad K = 2.0, \quad n_{\text{eff}} = \sum w_y$$

$$\Delta\text{coach}_t = \text{coachQ(new)} - \text{coachQ(old)} \quad \text{[only if both coaches have prior history]}$$

$$f_p = \max(0, \text{rating}_p - 0.80) \quad \text{[blue-chip weight; walk-on churn} \to 0\text{]}$$

$$\text{portal}_t = \sum_{\text{in}} f_p \cdot (1 \text{ if immediate else } 0.5) - \sum_{\text{out}} f_p \cdot 0.7$$

$$\text{retLoss}_t = 3.0 \cdot \min(0, (\text{ret}_t - \text{mean ret}) / \text{sd(ret)})$$

Coach quality is an empirical-Bayes, recency-weighted average of the coach's SP+ over seasons strictly before the year in question. Losing a portal star is penalized slightly less than gaining one helps (factor 0.7). Only teams *below* average returning production are docked.

One adjustment is applied after the five terms above. Preseason market expectations (§2) carry information about roster turnover that the data-driven terms miss, so the market's *ordering* of teams — re-expressed on this prior's own distribution, with its scale discarded — is blended in at weight 0.75. It has no decay term of its own: the walk-forward fit of §6 washes the whole prior out as real games arrive. Teams with no posted expectation are left exactly as the five terms leave them.

**A term that was tested and dropped.** A symmetric additive returning-production term carries weight zero. It is anti-correlated with talent — bluebloods churn rosters, small schools retain — so it inflated high-continuity G5 teams and docked powerhouses while adding nothing to the backtest. Returning production survives only as the downside penalty above and as an uncertainty multiplier (§7.3).

**Validation window.** All five terms are knowable before week 1 and were validated walk-forward on 2022–2025, the post-COVID NIL/portal regime; pre-2021 transfer behavior is a different world and is deliberately excluded. Weights were locked on a 2022–23 tune and held out on 2024–25. The shipped carryover of γ = 0.95 is raised from the γ = 0.8 the distribution was validated at, to de-compress preseason favorites: it cut week-1 RMSE by roughly 0.56 out of sample with no effect on mature weeks. FCS teams are merged in afterward with their own prior-season ratings.

> PROVISIONAL PRIORS
>
> Early in an offseason some feeds (talent, coaches, returning production) may not yet be published. The pipeline records which inputs were available; if the NIL-era set is incomplete, the season is flagged `preliminary` and the prior falls back to carryover plus portal only. The app surfaces this as a "model still filling in" banner and per-game markers.

## 5 In-season power ratings

As games are played, the prior is updated to fit the season's actual margins. For a vantage at week $w$, using only games with week < $w$, Saturday solves a ridge regression that pulls each rating toward its prior:

$$R_w = \operatorname{argmin}_R \sum_{g:\ \text{week}<w} \left[ (m_g - h_g) - (R_{\text{home}(g)} - R_{\text{away}(g)}) \right]^2 + \lambda \sum_t (R_t - \text{prior}_t)^2$$

$$\text{solved in closed form:} \quad (X^\top X + \lambda I)\, R = X^\top y + \lambda \cdot \text{prior}, \qquad \lambda = 4.0$$

Here $m_g$ is the capped margin and $h_g$ is the venue's home-field points (zero at neutral sites); $X$ carries one +1/−1 column pattern per game for the two teams involved.

The ridge penalty toward the prior does the early-season work. With no games it returns the prior exactly; with a few games it barely moves; by midseason the data dominates and the prior fades. It also keeps thinly-played teams from acquiring extreme ratings and lets pooled `NON_FBS` buy games contribute without distorting the scale. λ was tuned on 2022–23 and held out on 2024–25. As an independent check, the full-season version of this fit correlates **0.97** with SP+ — a rating Saturday never sees — and recovers a sensible three-point home field on its own.

## 6 The walk-forward protocol

This is the discipline that makes every displayed number honest. The app never serves a full-season (hindsight) rating for an in-progress season. Ratings, margins, totals and win probabilities are computed *as of* the vantage week, fitting only on games that had already happened. The exporter fits one rating snapshot per week and prices each game off its own week's snapshot; the win-probability timeline on a matchup page is the sequence of these as-of-week reads.

Because "now" is defined as the earliest week with an unplayed game, re-pulling results and re-running the pipeline rolls the entire app forward automatically — new power ratings, a new playoff field, updated schedules — with no code change and no leakage.

**Concrete illustration.** At its real week 8 the model had Texas over Georgia by about 11 points. The full-season hindsight rating, after Texas later regressed, was only about 6.7. Saturday shows the former, because that is what it actually knew at the time.

### 6.1 Early-season efficiency blend

Raw final margin is a noisy read on true strength in the opening weeks, when only a handful of games exist. Opponent-adjusted per-play efficiency is a cleaner early signal. For each completed game Saturday forms a net-efficiency differential from play-by-play PPA ( `offense.overall - defense.overall` ), fits the *same* ridge-toward-prior estimator on it, and converts the rating difference to points with a fixed scale $B_1 \approx 36$. The projected neutral edge is a week-decaying blend of the two signals:

```
edge = (1 - w)·margin_edge + w·(B1·efficiency_edge),   w(week) = 0.7·exp(-0.4·(week-1))
```

Efficiency therefore carries 70% of the week-1 edge, fades to roughly 14% by week 5, and is negligible thereafter. Home field is added separately per venue; games where either side lacks a PPA rating fall back to pure margin. Only the projected margin blends — the power ratings stay pure-margin.

The weight schedule was tuned on 2021–22, locked, then tested untouched on 2023–25: week-1 margin RMSE falls from **19.01 to 18.43**, the gap to the closing line narrows through week 4, and mature weeks are unaffected ($\Delta \approx -0.01$). Efficiency helps only early, which is why the weight is engineered to vanish by week 5 — and the opening weeks are exactly where Saturday's gap to the market is largest (§11).

### 6.2 Automatic preseason-to-live handoff

The offseason prior is not meant to persist once real games exist, and the switch off it is automatic rather than a flag someone has to remember to flip. Before exporting a season the pipeline checks the pulled schedule for any completed FBS game. If none exist, the season exports on the **preseason board**, every game priced purely from §4's prior. The moment the first FBS game finals, the next export switches that season onto the **as-of-week board** for the rest of the season.

A manual override ( `SATURDAY_PRESEASON` ) exists for testing, but the automated refresh (§13.1) never sets it. Used carelessly after a season has started it forces the board back onto the offseason prior and discards every completed result for that run, so the exporter prints a loud warning on that combination — a warning rather than a hard failure, since deliberately forcing the preseason vantage is occasionally legitimate during QA.

## 7 The outcome distribution

Saturday predicts the whole spread of outcomes, not just the center. The projected margin $\mu = R_{home} - R_{away} + HFA_{venue}$ is the mean of a distribution whose width and shape are measured, not assumed.

### 7.1 Shape and width

Pooling strict walk-forward residuals (predicted minus actual margin) across four seasons shows college-football margin errors are approximately **normal**: excess kurtosis is about zero. The visible mass of upsets and blowouts comes from how *wide* the curve is, not from fat tails — a 6.5-point favorite still loses about one time

in three straight out of a normal curve of the right width. The single-game distribution is therefore Normal(μ, $\sigma(w)^2$).

## 7.2 The week-aware width curve

A flat σ is overconfident in September, when only a prior is available. Bucketing pooled walk-forward residuals by week and fitting a smooth exponential decay gives

```
σ(w) = floor + span·exp(-k·(w-1))

2025 export (pools 2022–2024):  σ(w) = 15.63 + 3.00·exp(-0.426·(w-1))
2026 export (pools 2022–2025):  σ(w) = 15.55 + 2.98·exp(-0.411·(w-1))
```

The fit is weighted nonlinear least squares (weights proportional to √n per week; bounds floor ∈ [13,17], span ∈ [0,12], k ∈ [0.05, 3]) over residuals pooled only from seasons **strictly before** the season being priced, with a minimum of three seasons. No foresight applies to this structural fit either, which is why the curve differs by exported season. Both curves give σ ≈ 18.5 in week 1, decaying to 15.5–15.6 once ratings mature; σ is clamped at the week-13 value thereafter.

## 7.3 Decomposition: rating uncertainty vs outcome noise

That walk-forward σ bundles two very different things, and separating them is essential for season-level quantities:

- **Outcome noise** $\sigma_{outcome}$ — the irreducible randomness of one game. Even a correctly-rated 14-point favorite sometimes loses. Roughly flat all year, about 15.2.
- **Rating uncertainty** $\tau(w)$ — how unsure we are that the team truly *is* a 14-point favorite. Large preseason, shrinking as results arrive. Crucially it is *shared* across all of a team's games: if a team is better than its rating, it is better in every game, which correlates that team's outcomes.

Orienting each game's residual onto both teams and running a one-way random-effects (ANOVA) decomposition over (team, season) groups recovers $\sigma_{outcome}$ as the within-team scatter — about 15.2, matching σ's mature floor — and τ as the between-team scatter. $\sigma_{total}(w)$ is left **untouched**, so single-game win probabilities never change; the identity $\sigma_{total}^2 = \tau^2 + \sigma_{outcome}^2$ only *adds* the split that the season simulation needs.

THE SPLIT BY WEEK

| Week | $\sigma_{total}$ | $\tau(w)$ | $\sigma_{outcome}$ | $\rho = \tau^2 / (\tau^2 + \sigma_{outcome}^2)$ |
|---|---|---|---|---|
| 1 | 18.5 | 4.3 | 15.2 | 0.07 |
| 4 | 16.4 | 3.3 | 15.2 | 0.04 |
| 8 | 15.7 | 2.8 | 15.2 | 0.03 |
| 12 | 15.6 | 2.7 | 15.2 | 0.03 |

One subtlety is guarded here. The naive identity $\tau(w) = \sqrt{(\sigma_{total}^2 - \sigma_{outcome}^2)}$ attributes *all* early-season inflation to correlated rating error, giving τ(1) ≈ 10.5 and ρ(1) ≈ 0.33 — roughly four times the ANOVA evidence. Most

early inflation is instead larger *independent* noise and misspecification. Saturday therefore **caps** τ(w) at the directly measured ANOVA estimate: an exponential decay from the early-window τ (about 4.3, weeks 1–4) to the mature-window τ (about 2.7, weeks 6–15) at cap rate 0.35, refit to the same floor + span·exp family for export.

τ was also designed to be *team-specific*, since some teams are harder to rate. Returning production (overall and at quarterback) gives a hindsight-free multiplier meant to inflate the shared slice for churned or new-quarterback teams and shrink it for continuity-rich ones:

$$\text{mult}_t = \text{clamp}(\ 1 + 0.45\cdot(0.5 - \text{retPct}_t) + 0.55\cdot\max(0,\ 0.40 - \text{retPassPct}_t),\ \ 0.80,\ 1.30\ )$$

SHIPPED DISABLED

The τ multiplier exists in the pipeline but **ships off, pending validation.** It was hand-set with no validation artifact behind it, so every exported game uses $\text{mult}_t = 1.0$ and the reported `stdDev` is exactly $\sigma_{total}$(week). `SATURDAY_TAU_MULT=1` enables it for validation runs, where a game inflates only its τ slice by the geometric mean of the two teams' multipliers. Enabled, it does **not** preserve $\sigma_{total}$: scaling the τ slice changes the game's total σ and therefore its win probability. That is why it is held to the same walk-forward bar as every other shipped parameter.

### 7.4 Win probability with key numbers

Win probability is the mass of the margin distribution past zero. A featureless normal CDF misprices near-coin-flip games, because real football finals pile up on the key numbers 3 and 7 (and 10, 14) and almost never land on 2, 4, 5 or 9. Saturday convolves the normal margin density with an empirical per-integer key-number weight table, fit on five seasons of FBS finals, and reads off P(margin ≥ 1), split-crediting the impossible zero:

$$P(\text{win}) = \left[\ \sum_{m>0} \varphi(m;\ \mu,\sigma)\cdot\kappa(|m|)\ \ +\ \ \tfrac{1}{2}\cdot\varphi(0)\cdot\kappa(0)\ \right]\ \ \div\ \ \sum_{m} \varphi(m)\cdot\kappa(|m|)$$

One extreme-tail guard follows. A heavy favorite can round to 100%, which silently inflates season win totals over a slate of buy games, so Saturday softens probabilities above 97% toward an empirical upset floor of 1.5% — roughly the FCS-over-FBS and freak-result base rate — leaving ordinary games exactly as computed. The bell curve shown in the interface still uses the smooth density; only the numeric win percentage uses the key-number-aware version, so the two never visibly disagree near a spike.

## 8 Projected totals and scores

Each team carries an offense and a defense scoring rating. Every game is two observations, ridge-fit with penalty α = 8 on the offense/defense columns only:

$$\text{home\_pts} \approx \text{base} + O_{\text{home}} - D_{\text{away}} + \text{homeBump}$$
$$\text{away\_pts} \approx \text{base} + O_{\text{away}} - D_{\text{home}}$$
$$\text{projected total} = \text{home\_pts} + \text{away\_pts}$$

In season this is fit on completed games once at least 40 exist. Earlier, a preseason totals model fit on the prior two seasons is used, shrunk 50% toward its own mean and clamped to [35, 75]. The shrinkage is deliberate: the total signal is weak (correlation about 0.14 with actual totals), so overconfident totals would only add error. The offense/defense split shown on a game page is this model's split, rescaled to sum to the rating-model edge. A projected final score comes from snapping (total ± margin)/2 to realistic football scores — sums of field goals and touchdowns — so impossible finals never display, with the favorite nudged up if a snap would tie.

## 9 Season win totals

Summing a schedule's per-game win probabilities gives the right *mean* win total but a distribution that is too thin, because it ignores the shared rating error that correlates a team's games. Saturday instead runs a shared-latent Monte Carlo of 30,000 simulated seasons. The single-game $\sigma_g$ is *decomposed*, not inflated: the shared slice is $\tau_{team,g}^2 + \tau_{opp,g}^2$ and the independent slice is

$$\sigma_{ind,g}^2 = \max( \sigma_{outcome}^2,\ \sigma_g^2 - \tau_{team,g}^2 - \tau_{opp,g}^2 )$$

Each simulated season draws **one** standard-normal latent per team. Each game shifts its margin by $Z_{team} \cdot \tau_{team,g} - Z_{opp} \cdot \tau_{opp,g}$ and is resolved against its conditional win probability at the reduced $\sigma_{ind,g}$. Because the two variances sum back to $\sigma_g^2$, each game's *marginal* win probability reproduces the single-game number and the simulation mean tracks the sum of win probabilities — but the common per-team latent correlates that team's games and correctly fattens the win-total tails. $\tau$ is taken at each game's own week, so an early game moves more with the shared latent than a late one. The same engine, with selected games locked to a chosen result, powers the interactive choose-your-own-path win-total tool; P(wins > line) prices a season over/under.

## 10 The scenario engine

The toggles on each game page shift the projected margin and total by amounts measured from data. The big levers are real; the small ones are kept small on purpose, so they cannot quietly take over a forecast.

| SCENARIO | EFFECT | BASIS |
|---|---|---|
| **Turnover (net)** | `3.55 pts margin` | Regression of margin on turnover differential, controlling for rating strength (empirical ≈ 3.6). |
| **Starting QB out** | `1.5 – 8.5 pts` | Per-play PPA mapped linearly from replacement (12th percentile) to elite (90th). In-season deltas read the **prior season's** per-player PPA, matched to the current starter by name, so a transfer keeps their own history and no in-progress performance leaks in. Preseason, and for a true first-year starter with no prior-season record, a generic 4.5 (FBS) / 3.0 (FCS) applies. |
| **Missed field goal** | `∓2.3 margin, -3 total` | Expected value of a swung kick. |
| **Sack (net)** | `1.4 pts` | Backtested per net sack. |
| **Storm / heavy rain** | `-7 total` | Conditions heuristic (scoring suppression). |

| SCENARIO | EFFECT | BASIS |
|---|---|---|
| **Gusty wind (20+ mph)** | `-4 total` | Conditions heuristic. |
| **Cold / heat** | `-0.06 total per °F` | Deviation from 60°F. Open-Meteo, about 900 outdoor games; no detectable margin effect. |
| **Venue flip** | `2 × home-field pts` | Symmetric consequence of the venue model below. |

**Two factors retired from the shipped margin. Rest**, re-derived against Saturday's own walk-forward residual rather than a market residual, comes out at +0.032 ± 0.094 points per day (t = 0.35, n = 1,625). Saturday does not measurably misprice rest, so the coefficient is set to exactly zero — better none than a misspecified one. **Travel** is not applied at all. Both are excluded by design rather than estimated and hidden.

### 10.1 Home-field advantage: the venue model

A single global home field ignores that some stadiums are genuinely tougher, but 130-plus free per-venue parameters are mostly noise, because team quality and venue are confounded. Saturday fits team-season ratings plus a global home field for each of six seasons (2019 and 2021–2025; COVID-empty 2020 is skipped), takes the per-home-game residual as each venue's raw home bonus net of team quality, then shrinks it toward the global mean:

$$\text{reliability} = \tau_V^2 / (\tau_V^2 + SE^2), \quad \tau_V^2 = 1.10, \quad SE = \text{sd(resid)}/\sqrt{n}$$
$$HFA_{venue} = \text{clamp}( 3.0 + \text{reliability} \cdot (\text{raw} - 3.0),\ 2.3,\ 4.3 )$$

$\tau_v$ — the true between-venue standard deviation, about 1.05 — is the noise-free split-half covariance of venue residuals. Capacity, fill rate and altitude did not replicate out of sample, so no mechanistic feature is used: each venue is trusted to its own multi-year home history, shrunk by how much of it there is. The global home field recovers to 2.97, ranging from about 2.3 at quiet stadiums to 4.3 at the loudest; venues with too few games fall back to the 3.0 default. As with every structural estimate here, the estimation window ends **before** the season being priced — the 2025 export uses 2019 and 2021–2024.

## 11 Validation

Because Saturday does not use point spreads or game betting lines as inputs, the closing line works as an outside scorekeeper. Each season is trained and tested on itself, walk-forward, and pooled here across the four seasons the validation pipeline runs over: 2022 through 2025. (There is no in-repo backtest against a market line before 2022.) Spread and total are scored the same way, against the median closing number.

WALK-FORWARD RMSE VS THE CLOSING MARKET, BY SEASON

| SEASON | SPREAD MODEL | SPREAD MARKET | GAP | TOTAL MODEL | TOTAL MARKET | GAP |
|---|---|---|---|---|---|---|
| 2022 | 16.58 | 15.37 | +1.21 | 16.29 | 15.35 | +0.94 |
| 2023 | 15.62 | 14.98 | +0.64 | 16.14 | 15.46 | +0.67 |
| 2024 | 16.26 | 15.45 | +0.81 | 16.37 | 16.28 | +0.09 |
| 2025 | 15.75 | 15.07 | +0.68 | 15.50 | 15.11 | +0.39 |
| **Pooled** | **16.04** | **15.22** | **+0.84** | **16.07** | **15.56** | **+0.51** |

The result is stable across all four independent seasons — the same shape, a small and consistent gap to the market on both spread and total — so it is not tuned to one lucky year.

Win-probability calibration is checked with a reliability curve (games called 70% should win about 70% of the time) and summarized by expected calibration error: the average gap between predicted and realized win rate across probability buckets, weighted by games per bucket. Scored under the exact rule a visitor sees — σ(week) from §7.2 plus the key-number convolution and extreme-probability soften of §7.4, not a plain normal curve — pooled calibration error across 2022–2025 is **2.2%**.

**One disclosed exception to walk-forward.** The σ(week) curve used only to score that calibration figure is fit on all four seasons being graded, rather than excluding the season under test the way every live export does — a small, known advantage this backtest gets that the live product does not. The spread and total RMSE figures above are unaffected; they are fully out of sample game by game.

HONESTY NOTE — THE HEADLINE NUMBERS ARE CARRIED BY WEEKS 5+

Split by week, the shipped pipeline's own walk-forward record, pooled 2022–2025, with the full production prior and no foresight:

| WINDOW | RMSE MODEL | RMSE SPREAD | GAP |
|---|---|---|---|
| Week 1 | 16.84 | 14.91 | +1.98 |
| Weeks 1–4 (early) | 16.91 | 15.20 | +1.73 |
| Weeks 5+ (mature) | 15.54 | 15.22 | +0.32 |
| **All weeks** | **16.04** | **15.22** | **+0.84** |

Early season is genuinely weaker. Week 1 sits about 2.0 RMSE points behind a market that has already priced rosters, portal acquisitions and pre-camp injuries; the gap closes to about 0.3 by week 5 as real games overwrite the preseason prior. σ is widened accordingly early (§7.2) rather than sitting flat, which is why early-season games carry a high-uncertainty flag.

**A harsher stress test.** Separately, a deliberately strict carry-only backtest — hyperparameters locked on 2022–23, tested untouched on 2024–25, seeding every team with *only* a last-season carryover prior — pushes week 1 to roughly σ 18.6 and +4.89 RMSE behind the closing line (weeks 1–4 +3.11, all weeks

> +1.45). That measures the bare carryover signal, not the shipped model, and tells the same story: early season is weaker, and the gap closes fast.

### 11.1 Accuracy by game type

The same pooled 2022–2025 predictions, split by matchup type instead of by week, show where the gap to the market actually lives. Conference games are the easiest read: every team plays a dense, well-connected schedule inside one league, giving the ridge fit plenty of shared opponents to triangulate from. Non-conference matchups between teams with largely disjoint schedules are the hardest.

| GAME TYPE | RMSE MODEL | RMSE MARKET | GAP | N |
|---|---|---|---|---|
| Conference | 15.58 | 15.32 | +0.26 | 2,104 |
| Neutral site | 15.08 | 14.78 | +0.31 | 99 |
| Non-conference FBS | 16.06 | 15.02 | +1.03 | 795 |
| FCS opponent | 16.04 | 15.15 | +0.89 | 463 |

Non-conference FBS is the largest remaining gap to the market in the whole model and the biggest lever on the roadmap. Teams that never share an opponent within a season are hard to compare directly, and today's ratings bridge that gap only through the ridge fit's connectivity across the rest of the schedule (plus, early on, the preseason prior) rather than through any dedicated cross-conference term. A validated bridge — for instance anchoring to an external preseason rating such as SP+ or FPI — is the leading candidate fix and has not shipped.

## 12 Derived products

### 12.1 Power rankings

Ratings map to a 0–99 power score by standardizing against the FBS distribution: $\text{power}_t = 70 + 14 \cdot z_t$, floored at 3. Above 97 the mapping **soft-compresses** toward a 99.9 asymptote rather than hard-clipping at 99, so the very best teams separate by fractional amounts near the ceiling instead of all reading an identical 99. FCS teams map onto the same scale, so a strong FCS team lands above a weak one instead of sharing a pooled value. The board is living: preseason it shows the projected order, then “through week N,” then the final end-of-season ratings fit on every completed game. Strength of schedule is average opponent power.

### 12.2 Playoff model

The 12-team College Football Playoff field is projected as of each week, with no foresight, using a transparent committee proxy that adds résumé terms to the rating in comparable ranking points:

$$\text{score}_t = R_t + 2.5 \cdot (\text{quality wins}) - 4.0 \cdot (\text{losses}) + 0.15 \cdot (\text{strength of schedule})$$

A quality win is a win over a top-25-by-rating team; strength of schedule is average opponent rating. The field follows the real 12-team rule: the **five highest-ranked conference champions** are guaranteed a slot (the as-of-week proxy treats each conference's current best team by committee score as its champion, with Independents

excluded from the guarantee), the remaining seven slots go to the highest-scoring at-large teams, and the field is straight-seeded 1–12 by score. Under the 2025 rules the top four seeds get first-round byes; the first round is 5v12, 6v11, 7v10, 8v9 at the higher seed, with later rounds at neutral sites.

Ranking context follows the hierarchy of §2: AP until the committee's first poll, CFP rankings thereafter, Coaches Poll never. Once the real playoff is played, the actual bracket is read from CFBD and shown alongside the week-by-week projected field.

## 13 Reproducibility

The pipeline is parameterized by two environment variables: `SATURDAY_YEAR` (the season) and `SATURDAY_PRESEASON` (prior-only versus as-of-week pricing — normally left to auto-detect per §6.2, not set by hand). The single-command rebuild is `python refresh.py --all`, which runs every step below for every configured season in order, warns on a stale games cache, runs smoke checks, and finishes with the frontend type-check. The steps, to rebuild a season end to end by hand:

1. **Priors and structure** are computed on demand from CFBD inside the exporter: the carryover fit (§5, α = 1 on the prior season), the enhanced prior (§4), the venue home-field map (§10.1), σ(w) (§7.2), and the τ decomposition (§7.3).
2. `export_season.py` — fits as-of-week ratings and totals once per week, prices every FBS game (margin, total, σ, win-probability history, quarterback deltas, records, ranks, venue), attaches results for completed games, and writes `season-data[-YEAR].json` with the σ and τ curves, $\sigma_{\text{outcome}}$, the end-of-season board, final records and a readiness manifest.
3. `export_power_rankings.py` — reads that file and writes the top-100 board (§12.1).
4. `playoff_model.py` — writes the projected field by week and, once available, the actual bracket (§12.2).
5. Auxiliary exporters build the accuracy-page and season-replay data.

The front end is a static consumer of these JSON files; the season win-total Monte Carlo (§9) and the key-number win probability (§7.4) run client-side from the exported curves. Advancing the season is simply re-pulling results and re-running `refresh.py --all --fresh`.

### 13.1 Running unattended: the refresh job and its safety gates

In production nobody runs those steps by hand. A scheduled job re-runs `refresh.py --all --fresh` every two hours during the season, re-pulling new final scores, repricing every game, and regenerating every JSON file this document describes; outside the season it idles to an occasional check.

Three gates stand between that automated run and the live site: a **schema gate** (every required field, type and range on every game and team record), an **aggregate smoke check** (plausible game counts, sane rating and score ranges, correct current-week math), and a full **frontend type-check**. No later step may proceed once one fails, so a bad API response, a malformed record, or a broken model run stops that cycle in place: the site keeps serving its last good data, and nothing partially built reaches production. A separate check (`scripts/check-parity.mjs`) confirms the iOS app's copy of the win-probability and scenario math has not drifted from the website's.

Two inputs are whole-season API responses with no per-week parameter to key a partial re-fetch on: the poll feed ( `/rankings` ) and per-play efficiency ( `/ppa/games` , behind §6.1's blend). The on-disk cache does not expire on its own, and re-pulling new final scores does not clear either one. Left unaddressed, both could silently freeze at whatever snapshot was first cached — stale rank badges cosmetically and, more seriously, the efficiency blend quietly falling back to pure-margin pricing for every week 1–5 game with no error shown anywhere. As of this revision, `--fresh` clears both alongside the games cache, and `export_season.py` hard-fails a live-season export if the efficiency blend returns zero coverage for weeks 1–5 while completed games already exist in that window.

## 14 Limitations

- **Early season.** Before roughly week 5, Saturday trails the market by several points of RMSE. The widened σ makes this honest but does not remove it.
- **Non-conference FBS games** are the largest remaining gap to the market by matchup type (+1.45 RMSE pooled, §11.1). No dedicated cross-conference bridge has shipped.
- **Totals are a weak signal.** The totals model barely beats a constant (correlation about 0.14). Projected scores are best read as plausible, shrunk estimates, not precise predictions.
- **Weather beyond temperature** is heuristic, not fit; the wind and precipitation deltas are conservative placeholders.
- **Committee behavior** in the playoff proxy approximates a human process and can diverge from the real committee on résumé edge cases. The résumé weights are inherited defaults, not backtested against actual committee history.
- **Provisional offseason priors** depend on feed availability; a season flagged `preliminary` is carryover plus portal only until the remaining feeds populate.
- **Rejected as of this revision.** Season-long efficiency ratings, turnover and luck-regression targets, situational margin terms (travel body-clock, altitude, bye week), and a wind term on totals were each backtested out of sample and did not improve on the existing estimator; the ridge-toward-prior fit already absorbs most of that signal. The one edge that survived, the early-season efficiency blend of §6.1, is shipped.

## A Appendix: constants and hyperparameters

| SYMBOL / NAME | VALUE | ROLE |
|---|---|---|
| MOV_CAP | ±38 | Margin cap before every rating fit |
| λ (LAM) | 4.0 | Ridge strength toward the prior, in season |
| γ carryover (shipped) | 0.95 | Last-season shrink in the prior |
| γ, λ (backtest-locked) | 0.8, 4.0 | Values σ and τ were validated at |
| $W_{tal}$ | 4.0 | Recruiting-talent z weight |

| SYMBOL / NAME | VALUE | ROLE |
|---|---|---|
| $W_{coach}$ | 0.3 | Coaching change; coachQ decay 0.85, K 2.0 |
| $W_{portal}$ | 5.0 | Portal quality; blue-chip cut 0.80, out-factor 0.7 |
| $W_{retloss}$ | 3.0 | Returning-production downside penalty |
| HFA center / clamp | 3.0 / [2.3, 4.3] | Venue home field; $\tau_v^2$ = 1.10, minimum 6 games |
| σ(w) | 2025: $15.63 + 3.00e^{-0.426(w-1)}$ 2026: $15.55 + 2.98e^{-0.411(w-1)}$ | Week-aware margin SD; refit per exported season on seasons strictly before it |
| $\sigma_{outcome}$ | ≈ 15.2 | Flat independent per-game noise |
| τ cap window | 4.3 → 2.7 | ANOVA early-to-mature ceiling, cap rate 0.35 |
| τ multiplier | [0.80, 1.30] | Returning production: 0.45 overall, 0.55 QB, QB pivot 0.40. **Disabled by default**; `SATURDAY_TAU_MULT=1` enables it for validation runs |
| Efficiency blend w(w), $B_1$ | $0.7e^{-0.4(w-1)}$ / 36 | Early-season weight decay and points scale |
| Totals ridge α | 8 | Offense/defense penalty; preseason shrink 0.5, clamp [35, 75] |
| Points per turnover | 3.55 | Shipped (empirical ≈ 3.6) |
| QB-out band | 1.5 – 8.5 | Per-play PPA, 12th to 90th percentile; in-season delta from prior-season PPA by starter name |
| Temperature | 0.06 / °F | Total points suppressed per °F away from 60 |
| Upset floor / soften from | 1.5% / 97% | Extreme win-probability guard |
| Win-total simulations | 30,000 | Shared-latent Monte Carlo; single-game simulations 10,000 |
| Playoff proxy | 2.5 / 4.0 / 0.15 | Quality-win, loss and SOS weights; five conference champions guaranteed, seven at large, straight-seeded |
| Power scale | 70 + 14z | 0–99 display score; soft-compresses above 97 toward a 99.9 asymptote |
| Poll source | AP → CFP | AP until the committee's first poll of the season, CFP thereafter; Coaches Poll never used |

**Saturday is an analytical instrument** for understanding the range of ways a game can unfold and how conditions change it. It is not market-aware, not an oracle, and not betting advice: it prices probabilities, not certainties, and a 75% favorite is meant to lose one time in four. All estimates derive from public CollegeFootballData inputs; the betting line is used only as an independent benchmark.