XC-SPI Historical Validation & Backtesting Plan

Version 2.7: separate validation for race prediction, team ranking, and championship probabilities

Colin C. Caso
Oklahoma State University

August 28, 2026

Abstract

This plan specifies how XC-SPI will be evaluated before fitted probabilities are promoted to public calibrated forecasts. Version 2.7 separates the abundant-data tasks (athlete/race prediction and weekly team ranking) from the rare-event task (national-title probability calibration). The latter cannot be established from one held-out NCAA final.

1. No-Leakage Principle

A historical Week t prediction may use only information genuinely available by Week t in that season. Later race results, final rosters, later injury knowledge, and the NCAA outcome cannot leak backward.

2. Tier 0: Postmodern Era Historical Success Benchmark

Before the full athlete simulator is promoted, XC-SPI maintains a simpler historical benchmark using the strict **Postmodern Era (Super-Shoe Era onward), 2020 forward**.[4] Source profiles from 2020–2024 predict 2021–2025 outcomes. The 2020 source championship is marked as COVID-disrupted and excluded in a sensitivity rerun. Candidate profile models are compared by held-out-season cross-validation. This is an era-analogy benchmark, not a substitute for the no-leakage rolling-origin validation required for the athlete model.

The benchmark reports title, podium, top-10, and top-20 rates; Brier/calibration summaries; comparison with simpler score-only specifications; and sensitivity to removing 2020. Historical Title Odds are Top-25 conditional. The small number of title events is reported explicitly.

3. Tier 1: Athlete and Race Prediction

Use recent seasons with the richest course/weather/roster reconstruction (initially 2021–2025, expanding when feasible). Evaluate rolling-origin next-race predictions, athlete intervals, residual structure, and context adjustments. Report MAE, RMSE, interval coverage, residual diagnostics, and performance by course/source-quality strata.

4. Tier 2: Weekly Team Ranking and Score

At each historical weekly cutoff, simulate the national field using only contemporaneous evidence. Report Spearman rank correlation, team-place MAE, NCAA-score MAE/RMSE, and comparisons with simpler baselines: prior NCAA order, raw current-season order, roster-only index, and simplified athlete model.

5. Tier 3: Championship Probability Calibration

Full current-athlete national-title probabilities require repeated championships. The Postmodern Era benchmark is the first modern calibration lens, but five outcome seasons are too small to establish fine-grained sportsbook-quality title calibration. For the athlete simulator, use as many reconstructable seasons as feasible and report clearly when older data are used only for calibration rather than for the Postmodern historical ranking. Do not claim full current-model probability calibration from a single men's and women's held-out championship.

Evaluation includes Brier score, log score, reliability/calibration plots, probability-bin counts, and calibration uncertainty. Sparse extreme bins must be reported with their sample sizes. [1;2]

6. Expected Score Versus Title Probability

Evaluate expected NCAA score and title probability separately. A model may order teams well by expected score while being miscalibrated in the tails. Likewise, the highest title probability need not always belong to the team with the lowest expected score if volatility differs.

7. Ablation Tests

Remove one family at a time: course, weather/ground mean terms, condition-dependent residual variance, track prior, dynamic state, team-race shock, and any future athlete-condition interaction. A component survives only when it improves held-out predictive performance, calibration, or uncertainty quality enough to justify complexity.

8. Observation-Uncertainty Audit

Test whether races with poor course/weather/ground measurement receive appropriately wider uncertainty and smaller athlete updates. If the model treats low-quality context as equally precise, the measurement-variance design has failed.

9. Irregular-Time State Audit

Compare elapsed-time-scaled state evolution with a fixed-per-race variance model. The scaled model should better handle long gaps without making athletes artificially volatile after tightly spaced races.

10. Team-Correlation Audit

Compare championship score distributions with and without the team-by-race shared shock. The term should remain only if it improves held-out team-score distribution/calibration rather than simply widening intervals.

11. Condition-Variance Audit

Test whether heat/mud/wind/technical conditions improve held-out residual-scale prediction. Athlete-specific condition slopes are excluded unless repeated-athlete evidence supports them and they improve out-of-sample performance.

12. Monte Carlo Convergence

Record draw count, seed, batch sequence, absolute MCSE, and relative precision for rare events. Stop only when preregistered tolerances are satisfied. A fixed 50,000-draw minimum may be used as a starting batch, not as a universal proof of adequate precision.[3]

13. Probability Promotion Gate

Public calibrated title probabilities require: no leakage; competitive proper scores versus baselines; acceptable calibration/reliability over the multi-season archive; stable behavior under reasonable sensitivity analyses; understandable major misses; and a versioned frozen data/code snapshot.

14. Versioned Results Report

The eventual report will include sample definition, historical seasons, model variants, baselines, metrics, calibration figures, ablation tables, sensitivity analyses, examples of failures, and a plain-language conclusion. Re-runs after bugs create new versions rather than overwriting old results.

References

[1] Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950;78(1):1–3.
[2] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc.* 2007;102(477):359–378.
[3] Metropolis N, Ulam S. The Monte Carlo method. *J Am Stat Assoc.* 1949;44(247):335–341.
[4] Caso CC. *XC-SPI Postmodern Era Historical Success Model: Super-Shoe Era onward, 2020–2025 team-profile persistence, 2026 success rates, and historical title odds.* Oklahoma State University; 2026.