XC-SPI Data Dictionary & Variable Guide

Definitions, units, allowed values, provenance fields, and interpretation rules

Colin C. Caso
Oklahoma State University

August 28, 2026

Abstract

This document defines the variables used in the XC-SPI preseason ledger and planned in-season model. It is intended for readers, auditors, and future collaborators who need to know exactly what a field means, what units it uses, where it comes from, and whether it is observed, derived, declared, or estimated.

1. Variable Classes

Every XC-SPI field belongs to one of five classes: **observed** (copied from a source), **derived** (calculated deterministically), **declared** (a project rule or fixed launch parameter), **estimated** (learned statistically from historical data), or **simulated** (generated from a fitted/provisional probability model). Keeping these classes separate prevents a modeling choice from being mistaken for a fact.

2. Identifiers and Provenance

Table 1: Core identity and provenance fields.

Field	Class	Definition
athlete_id	declared	Persistent internal athlete identifier. A transfer changes team affiliation, not athlete identity.
athlete_name	observed	Publicly reported athlete name used for display and source matching.
team_id	declared	Stable team identifier.
team_name	observed	Current public school/team name.
season	observed	Competition year, eg, 2026.
source_url	observed	Direct URL used to support the record.
source_grade	declared	Evidence grade A, B, or C under the Data & Evidence Standards protocol.
accessed_at	observed	Timestamp when the source was checked.
cutoff_at	declared	Latest evidence timestamp allowed into a published release.
model_version	declared	Version string for the code/specification that generated the output.
random_seed	declared	Seed used to make a simulation reproducible.

3. Race and Result Fields

Table 2: Race/result variables.

Field	Class	Definition
race_id	declared	Stable identifier for a race instance.
race_date	observed	Calendar date of the race.
start_time	observed	Local race start time when available; important for weather matching.
distance_m	observed	Race distance in meters.
raw_time_s	observed	Athlete's official time in seconds.
overall_place P_i	observed	Athlete's raw finish position in the full race field.
scoring_place Q_i	derived	NCAA team-scoring-sequence position after applying the governing scoring filter. Five athletes score; sixth and seventh may displace.[1]
team_score	derived/observed	Sum of five NCAA scoring places; verified against official/TFRRS result where available.
field_size	observed	Number of finishers in the raw field.
complete_team	derived	Whether a school has enough eligible finishers to form a team score.

4. Week 0 Preseason Score Fields

Table 3: XC-SPI Preseason Score variables.

Field	Class	Definition
X_i / runner_baseline	derived	**Runner Baseline**: comparable prior NCAA cross-country scoring value assigned to athlete i.
c_i / confidence	declared	**Confidence**: rule-based Week 0 shrinkage weight: 0.99 verified, 0.85 probable, 0.70 materially unresolved, 0 verified unavailable. This is not an availability probability.
R_0 / replacement_score	declared	**Replacement Score**: conservative launch fallback used when historical athlete evidence is unavailable or discounted; Week 0 value 200.
newcomer_score	declared	**Newcomer Score**: common conservative prior for a newcomer without directly comparable NCAA XC evidence; launch value 150 where applicable.
C_i / adjusted_runner_score	derived	**Adjusted Runner Score**: $C_i = c_i X_i + (1 - c_i)R_0$. An uncertainty-adjusted shrinkage-index contribution, not an expectation or predicted finish.
$C_{j(1)} \dots C_{j(7)}$	derived	Team candidate contributions sorted from lowest/best to highest/worst.
λ	declared	Depth Risk coefficient; launch value 0.05.

Table 3: Preseason Score variables (continued).

Field	Class	Definition
F_j / depth_risk	derived	**Depth Risk**: term comparing sixth/seventh candidate values with scorer five; not actual NCAA displacement points.
PS_j	derived	Sum of five best adjusted contributions plus Depth Risk. Lower is better; not expected NCAA points.

5. Roster Status Fields

Table 4: Roster and availability fields.

Field	Class	Definition
roster_status	observed/declared	Verified return, verified transfer in/out, professional departure, probable return, unresolved, newcomer, or unavailable.
former_team	observed	Previous school for transfer evidence.
current_team	observed	Destination/current school supported by the strongest public source.
eligibility_note	observed	Public class/eligibility information retained as evidence but not automatically treated as proof of availability.
status_reason	declared	Short audit note explaining why a weight/status was assigned.

6. Postmodern Era Historical Success Fields

Table 5: Historical Success Model fields.

Field	Class	Definition
postmodern_era	declared	XC-SPI project period beginning with the 2020 season; also described as Super-Shoe Era onward. The 2020 COVID-disrupted season is flagged.
hist_team_score	observed/derived	Prior NCAA five-runner team score, standardized within source season and sex.
hist_front_runner	observed/derived	First scorer's NCAA scoring place.
hist_fifth_runner	observed/derived	Fifth scorer's NCAA scoring place.
hist_depth_gap	derived	Seventh-runner minus fifth-runner scoring position; a conservative fill is used when a seventh finisher is unavailable.
historical_title_rate	modeled	Historical Success Model probability of next-season NCAA champion for a team with that profile.
historical_podium_rate	modeled	Probability of next-season finish 1st–4th. Used as the primary Historical Success ranking target because it is less sparse than title.
historical_top10_rate	modeled	Probability of next-season top-10 finish.

Table 5: Historical Success Model fields (continued).

Field	Class	Definition
historical_top20_rate	modeled	Probability of next-season top-20 finish.
historical_success_rank	derived	Rank ordered by historical podium rate, then title/top-10 as tie context.
historical_title_share	derived	Raw historical title rate normalized within the displayed Top 25 so the title shares sum to 100%.
historical_title_odds	derived	Fair American price from historical title share; Top-25 conditional, not full current-model XC-SPI odds.

7. Planned In-Season Athlete Model

Table 6: Planned model-state variables.

Symbol	Class	Definition
$\alpha_{i,t}$	estimated	Latent athlete ability at time t.
$V_{i,t}$	estimated	Uncertainty/variance attached to the athlete state.
σ_η^2	estimated	State-evolution variance per unit elapsed time. Innovation variance scales with Δt. State-space modeling provides the framework.[9]
$\delta(d)$	estimated	Smooth distance effect, planned as a penalized spline rather than a fixed pace conversion.[8]
γ_c	estimated	Persistent course effect, partially pooled across courses.
$h(w)$	estimated	Weather effect function.
$g(q)$	estimated	Ground/footing effect function.
u_r	estimated	Common race-wide residual effect; mainly a diagnostic/common-context term.
v_{jr}	estimated	Team-by-race shared shock creating modest correlation among teammates.
$\sigma_{\epsilon,r}$	estimated	Residual scale, optionally modeled as a function of difficult conditions so weather/ground can change outcome variance.
ϵ_{ir}	estimated	Athlete/race observation noise.
p_i^{avail}	estimated/optional	Calibrated probability the athlete is available, only if historical data support it.
p_i^{select}	estimated/rule-based	Championship-group selection conditional on availability; separate from availability.

8. Weather, Course, and Ground Fields

Table 7: Context variables.

Field	Class	Definition
course_id	declared	Persistent venue/course identifier.
elevation_m	observed	Course elevation/altitude metric when available.
elevation_gain_m	observed/derived	Total climb estimate where a defensible course trace exists.
surface	observed	Grass, dirt, gravel, mixed, etc, where documented.
temperature_c	observed	Start-time or nearest defensible temperature observation.
dewpoint_c	observed	Moisture/thermal-stress input.
wind_ms	observed	Wind speed tied to race time/location.
race_precip_mm	observed	Precipitation during the race window.
antecedent_precip_mm	observed/derived	Rainfall before the race window used as a footing proxy; window ends before race start to avoid double counting.
ground_grade	observed/declared	Direct footing evidence when available; otherwise missing rather than invented.
station_distance_km	derived	Distance from the weather source to the course.
weather_time_offset_min	derived	Time difference between the observation and scheduled race start.

9. Simulation and Forecast Evaluation Fields

Monte Carlo simulation repeats generated championship outcomes to estimate quantities that are difficult to obtain analytically.[5] Public probabilities must later be evaluated using proper scoring and calibration rather than only rank accuracy.[3;4]

Table 8: Simulation/evaluation variables.

Field	Class	Definition
M / simulations	declared	Number of Monte Carlo championship simulations.
$\hat{p}_j$	simulated	Estimated title probability: fraction of simulations won by team j.
MCSE	derived	Monte Carlo standard error from finite simulation count.
Brier score	derived	Mean squared error of probabilistic forecasts against binary outcomes. Lower is better.[3]
log score	derived	Proper score that strongly penalizes extreme confidence in outcomes that do not occur.[4]
calibration_bin	derived	Probability grouping used to compare stated probability with observed frequency.
rank_rho	derived	Spearman rank correlation between model ordering and comparison/outcome ordering.
MAE / RMSE	derived	Absolute/squared prediction-error summaries for continuous or rank-like targets.

10. Betting-Translation Fields

The probability/fair-price framework translates a calibrated probability into no-vig prices. The former public Week 0 scenario bridge is deprecated. Fair-price translation is arithmetic; it does not make an uncertain probability estimate correct. Kelly staking is included only as a theoretical risk diagnostic, not as a recommendation.[6]

Table 9: Odds/market variables.

Field	Class	Definition
fair_decimal	derived	$1/p$.
fair_american	derived	American-odds representation of model probability.
market_implied_p	derived	Probability implied by a quoted market price before removing vig.
overround	derived	Sum of market implied probabilities minus 1.
no_vig_p	derived	Market implied probability normalized by the total implied probability.
model_edge	derived	XC-SPI probability minus no-vig market probability.
EV	derived	Expected return per dollar at a stated price: $pd - 1$.
kelly_fraction	derived	Theoretical log-growth-optimal fraction under assumed probability/odds; not a project staking recommendation.[6]

11. Naming and Missing-Value Rules

A missing field should be genuinely missing, not replaced with zero unless zero has a real meaning. Text fields use concise controlled vocabulary where possible. Historical records are not rewritten to reflect later knowledge; corrections are recorded through a new version/change-log entry. Units belong in field documentation rather than being inferred from the column name alone.

References

[1] National Collegiate Athletic Association. Cross Country and Track and Field Playing Rules. Accessed August 28, 2026. https://www.ncaa.org/championships/playing-rules/cross-country-track-and-field-playing-rules/

[2] TFRRS. NCAA Division I Cross Country Championships. November 22, 2025. Gans Creek Cross Country Course, Columbia, Missouri. Accessed August 28, 2026. https://www.tfrrs.org/results/xc/27301/NCAA_Division_I_Cross_Country_Championships

[3] Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev.* 1950;78(1):1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[4] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc.* 2007;102(477):359–378. doi:10.1198/016214506000001437.

[5] Metropolis N, Ulam S. The Monte Carlo method. *J Am Stat Assoc.* 1949;44(247):335–341. doi:10.1080/01621459.1949.10483310.

[6] Kelly JL Jr. A new interpretation of information rate. *Bell Syst Tech J.* 1956;35(4):917–926. doi:10.1002/j.1538-7305.1956.tb03809.x.

[7] Gelman A, Hill J. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge University Press; 2007.

[8] Wood SN. *Generalized Additive Models: An Introduction With R.* 2nd ed. Chapman & Hall/CRC; 2017.

[9] Durbin J, Koopman SJ. *Time Series Analysis by State Space Methods.* 2nd ed. Oxford University Press; 2012.