Dynamic attack and defense strength in international rugby union

Christopher Claassen*

University of Glasgow

August 2026

*christopher.claassen@glasgow.ac.uk

Dynamic team attack and defense strength in international rugby union

Abstract

This article develops and compares dynamic Bayesian paired-score models for men's international rugby union. The models decompose team quality into latent attack and defense strengths and embed these in a paired negative-binomial score model. We compare three alternative specifications for the evolution of latent strengths over time: a weekly random walk, a continuous-time random walk, and a two-speed event-gap process that combines background elapsed-time diffusion with additional active-appearance variation. The models are evaluated using out-of-sample forecasts for 2025 and benchmarked against both a simple exponential error-correction smoother and the official World Rugby rankings algorithm. All three Bayesian models outperform these simpler benchmarks, although there is little to separate the three on predictive accuracy alone. The continuous-time and two-speed specifications are especially attractive, however, because they retain similar predictive performance while better reflecting the irregular timing of international rugby. Then, examining the evolution of latent attack and defense strengths among tier-one rugby teams from 2020 to 2025, we discuss how the models reveal meaningful differences in teams' attacking and defensive profiles, illustrating how this type of analysis can contribute to wider sports discourse and public discussion.

Keywords: rugby union; sports analytics; Bayesian modeling; forecasting; latent strength

Words: 6,518

1. Introduction

Rugby union is an attractive but still comparatively under-studied setting for dynamic scoreline analysis. In association football, there is now an extensive literature on paired-score models that distinguish between attacking and defensive quality and allow those latent strengths to evolve over time (Maher 1982; Dixon and Coles 1997; Karlis and Ntzoufras 2003; Crowder et al. 2002; Rue and Salvesen 2000; Koopman and Lit 2019). In American football, dynamic score-based models have also been used to study team strength and forecasting in a higher-scoring setting (Glickman and Stern 1998; Lopez et al. 2018). Like these sports, rugby produces paired scores that can plausibly be decomposed into attacking and defensive components, and like American football, it is a relatively high-scoring sport in which margins and scorelines are substantively informative. Yet rugby has attracted much less work of this kind, particularly work that combines dynamic latent strengths, paired-score modeling, and an explicit attack-defense decomposition.

This article focuses specifically on senior men's national-team 15-a-side rugby union. This is distinct from 13-a-side rugby league and rugby sevens, which have different formats and rules, as well as from women's senior national-team rugby, men's under-20 national-team rugby, and professional club and franchise rugby. International rugby union also presents a distinctive dynamic problem. Much of the existing literature on dynamic team-strength modeling focuses on domestic professional leagues, where teams play on relatively regular schedules and under broadly balanced itineraries. International rugby is different. Its calendar is irregular and clustered: matches come in bursts – the Six Nations, the Rugby Championship, autumn test windows, summer tours, and World Cups – separated by long periods in which national teams do not play. The schedule is also unbalanced, especially outside the top tier, with teams playing markedly different numbers of matches in a given year. A credible dynamic model of international rugby strength therefore needs to take seriously not only that team quality changes, but that it changes in a setting where periods of activity and inactivity alternate sharply and observational opportunities are uneven.

While research on latent team strengths in sport is often motivated by gambling applications

(e.g., Dixon and Coles 1997; Glickman and Stern 1998), there is another reason to care about estimates of this kind. Team "power" rankings are a familiar part of sports journalism and fan discourse across many sports and leagues. They are used to summarize team quality, frame major fixtures, interpret form, and debate whether observed results reflect durable strength or short-term noise. In international rugby union, the most prominent of these are the team rankings published by the sport's governing body, World Rugby. While overall team strength estimates are interesting, it is even more informative to move beyond a single scalar rating and distinguish between teams' attack and defense strengths. A team can be elite because it scores heavily, because it defends exceptionally well, or because it combines both. We therefore view attack–defense decomposition not only as a statistical convenience, but also as a way to produce estimates that are substantively interpretable and of interest to analysts, journalists, and fans.

This article develops and compares dynamic Bayesian paired-score models for men's international rugby union that decompose team quality into latent attack and defense strengths. The observation model uses conditionally independent negative-binomial score distributions for home and away points, linked through team-specific attack and defense strengths and a common home-advantage term. The state dynamics vary across three specifications: a weekly random walk, a continuous-time random walk that allows uncertainty in latent strengths to accumulate with the elapsed time between team appearances, and a two-speed event-gap process that combines this background elapsed-time diffusion with an additional active-appearance component. Thus, while all three models estimate teams' attack and defense strengths, they also allow us to evaluate alternative ways of representing the evolution of those strengths in an irregularly scheduled international sport.

We evaluate the three Bayesian models through an out-of-sample forecasting exercise and a benchmark comparison with two simpler alternatives: an exponential error-correction smoother and the official World Rugby rankings algorithm. The results show that all three Bayesian models perform similarly well and improve meaningfully on both benchmark algorithms. Although there is relatively little to separate the three Bayesian models on predictive accuracy alone, the

continuous-time and two-speed specifications are computationally simpler, and therefore more attractive, than the weekly random walk.

We then examine the estimated attack and defense strengths of tier-one men's teams over the 2020–2025 period. We show that our estimates provide more detailed insight into trends in international rugby, notably the different profiles of the two leading teams in Europe, Ireland and France. At the same time, the estimates depict familiar patterns that feature in broader narrative discussions, such as the rise of Argentina and the decline of Wales. We also show that, unlike the random-walk model, the continuous-time and two-speed models produce latent attack and defense estimates that remain largely stable while teams are idle, but can change, sometimes rapidly, once teams return to play. This is an attractive feature because it seems counterintuitive for a team's estimated strength to move substantially during periods in which we do not observe its performance.

2. Existing research

Statistical models for sports results are often grouped into three broad families: paired-score models, which model both teams' scores jointly; margin models, which model score difference or some transformation of it; and outcome models, which focus only on win, draw, and loss. In the (association) football forecasting literature, these approaches are frequently compared as alternative predictive devices rather than merely as different descriptive summaries of the same process (Koopman and Lit 2019). For the present paper, paired-score models are an attractive starting point as these contain more information than binary outcomes. That additional information is useful because team performance is plausibly decomposable into offensive and defensive components.

The canonical reference point is Maher (1982), who models home and away goals in association football as functions of team-specific attack and defense strengths plus home advantage. Much of the subsequent methodological literature can be understood as relaxing one or more of Maher's basic assumptions. Dixon and Coles (1997) modify the independent Poisson framework to improve fit in football and make recency weighting central to estimation. Karlis and Ntzoufras (2003) move to a fully bivariate Poisson specification, introducing dependence directly into the

paired-score distribution. Crowder et al. (2002) and Rue and Salvesen (2000) add dynamic latent strengths, allowing attack and defense parameters to evolve over time rather than remain fixed. In American football, Glickman and Stern (1998) provide a classic state-space treatment of team strength and score differences, showing how dynamic latent-quality models can be adapted to a higher-scoring sport.

The extension of these methods to international rugby union matches raises two questions: first, what count distribution is appropriate for rugby scores; and second, how should latent team strengths be modeled over time in a sport with highly irregular fixture timing?

For the first question, the Poisson distribution is a natural starting point because match scores are non-negative counts (Karlis and Ntzoufras 2003). The key restriction of the Poisson model, however, is that the conditional variance equals the conditional mean, which is likely to be too limiting for rugby union. Like American football, rugby scores are generated through several scoring routes – tries, worth five points, conversions, worth an additional two, and penalty kicks and drop goals, worth three points each – and match outcomes are shaped by factors such as tempo, tactics, refereeing, and weather. These sources of heterogeneity likely produce greater score variation than the Poisson model permits, or overdispersion. One approach, following Pledger and Morton (2011), would be to model rugby union matches by decomposing final scores into tries, penalties, etc., with each modeled as a Poisson count. In the present application, we instead focus on total home and away points directly. A negative-binomial likelihood is attractive for this task because it retains the count-data interpretation of the Poisson while adding a dispersion parameter that allows the variance to exceed the mean (see also Higgs and Stavness 2021).

The second question concerns dynamics. Once team strengths are allowed to vary over time, the literature offers several main approaches. One is a random walk, in which latent attack and defense drift stochastically from one period to the next (Rue and Salvesen 2000). Another is an autoregressive process, in which shocks decay over time and strengths revert toward a long-run level (Crowder et al. 2002; Glickman and Stern 1998). A third uses recency weighting or discounting in the style of Dixon and Coles (1997). These choices are substantively and practically

important as different dynamic structures can alter both computational burden and forecast quality (Koopman and Lit 2019).

For international rugby union matches, the key distinction is that unlike sports leagues like NFL and English Premier League football, which have attracted much of the research attention, international rugby teams do not play weekly schedules. Instead, they play in bursts – the autumn tests, Six Nations, Rugby Championship, and (every four years) the World Cup. These bursts are interspersed with long idle spells that differ sharply across teams and seasons. Both the weekly random walk and the mean-reverting AR specification evolve on a regular calendar grid, treating each week as an equivalent time step regardless of whether a team actually plays. Both can be useful approximations, but neither maps especially naturally onto the international rugby union calendar.

The paper therefore compares several dynamic formulations designed to handle that problem in different ways, and evaluates whether accounting more explicitly for elapsed time improves either predictive performance or the substantive interpretation of team trajectories. We describe these models in the next section.

3. A dynamic model of rugby attack and defense strengths

3.1. Observation model

Let $y_{g,h}$ and $y_{g,a}$ denote the home and away scores in match $g = 1, \ldots, N$. For a match between home team $i(g)$ and away team $j(g)$ at time $t(g)$, expected home and away scoring are denoted by $\mu_{g,h}$ and $\mu_{g,a}$. We model the paired scoreline using a paired negative-binomial specification with team-specific latent attack and defense strengths and a common home-advantage term. The linear predictors are

$$\log(\mu_{g,h}) = \kappa_h + \eta H_g + \alpha_{i(g),t(g)} - \delta_{j(g),t(g)},$$
$$\log(\mu_{g,a}) = \kappa_a + \alpha_{j(g),t(g)} - \delta_{i(g),t(g)}.$$

Here $\alpha_{k,t}$ is team k's latent attack strength at time t, $\delta_{k,t}$ is its latent defense strength, and κ_h, κ_a are intercepts governing average home and away scoring rates. H_g is an indicator for non-neutral fixtures, with 1 indicating matches in which the designated home team has home-field advantage, and 0 otherwise, while η is the home-advantage parameter. The latent strengths are therefore interpreted on the log scale: higher attack raises expected scoring, while higher defense lowers opponent scoring.

Conditional on these linear predictors, home and away scores follow negative-binomial distributions:

$$y_{g,h} \sim \text{NB2}(\mu_{g,h}, \phi_h),$$
$$y_{g,a} \sim \text{NB2}(\mu_{g,a}, \phi_a).$$

Here, $\mu_{g,h}$ and $\mu_{g,a}$ are the expected home and away scores in match g, while $\phi_h, \phi_a > 0$ are home- and away-score dispersion parameters. We use the quadratic variance parameterization of the negative binomial distribution, in which $\text{Var}(y) = \mu + \mu^2/\phi$, so smaller values of ϕ imply greater overdispersion (Cameron and Trivedi 2013). Note that this is also referred to as the NB2 parameterization and, in the Stan modeling language, as the "alternative parameterization" (Stan Development Team 2025). Smaller values of ϕ imply greater overdispersion relative to a Poisson model, while the Poisson model is recovered in the limit as $\phi \to \infty$. The model therefore relaxes the equidispersion restriction of a Poisson score model and allows the variance of each team's score to exceed its mean.

The home and away scores are conditionally independent given the latent states and model parameters. Any dependence in observed scorelines is instead captured indirectly through the shared latent strengths and common match context.

3.2. Dynamic latent strengths

The main quantities of interest are the evolving latent attack $\alpha_{k,t}$ and defense $\delta_{k,t}$ states for each team. The models considered in the paper differ according to the processes used to model the evolution of these latent attack and defense strengths over time.

Random-walk dynamics: The first specification models attack and defense as weekly random walks:

$$\alpha_{k,t} \sim \mathcal{N}\Big(\alpha_{k,t-1},\, \sigma_\alpha^2\Big),$$
$$\delta_{k,t} \sim \mathcal{N}\Big(\delta_{k,t-1},\, \sigma_\delta^2\Big)$$

for $t \geq 2$, and where $\sigma_\alpha, \sigma_\delta > 0$. Initial states are drawn from

$$\alpha_{k,1} \sim \mathcal{N}(0, \sigma_{\alpha 0}^2), \qquad \delta_{k,1} \sim \mathcal{N}(0, \sigma_{\delta 0}^2),$$

where $\sigma_{\alpha 0}, \sigma_{\delta 0} > 0$. Unlike an autoregressive specification, the random walk model does not impose mean reversion of attack or defense strength towards a long-run level or population average. We also tested an autoregressive specification in which latent strengths follow AR(1) processes governed by parameters ρ_α and ρ_δ. These parameters were estimated to be very close to one, implying dynamics that were nearly indistinguishable from a random walk; we therefore report the AR(1) specification as a robustness check in the supplementary materials. In the random walk model, latent strengths instead evolve as a stochastic drift over calendar time. This is a natural benchmark for international rugby union, where team strengths are very persistent, with nations such as England, South Africa, and New Zealand remaining leading rugby powers over many decades. The random walk model, however, treats all weeks alike regardless of whether a team actually plays.

Continuous-time random-walk dynamics: The second specification relaxes the assumption that each team evolves by the same amount each calendar week. Let $g_{k,t}$ denote the elapsed time, in weeks, since team k last appeared before week t. Attack and defense then evolve according to

$$\alpha_{k,t} \sim \mathcal{N}\Big(\alpha_{k,t-1},\, g_{k,t}\,\sigma_\alpha^2\Big),$$
$$\delta_{k,t} \sim \mathcal{N}\Big(\delta_{k,t-1},\, g_{k,t}\,\sigma_\delta^2\Big)$$

where $\sigma_\alpha, \sigma_\delta > 0$. As in the discrete-time random walk, the model does not impose mean reversion. However, the innovation variance is proportional to elapsed time, so uncertainty grows more during long gaps between appearances than during short gaps. As such, the innovation variance grows linearly with the length of the gap, while the innovation standard deviation grows only with the square root of elapsed time. This allows uncertainty to accumulate during long gaps without implying implausibly explosive changes in latent strength after extended periods without international matches.

Two-speed event-gap dynamics: The third specification extends the continuous-time random-walk idea by allowing latent attack and defense to evolve at two distinct speeds. Let $g_{k,t}$ again denote the elapsed time, in weeks, since team k last appeared before week t. Attack and defense then evolve according to

$$\alpha_{k,t} \sim \mathcal{N}\Big(\alpha_{k,t-1},\, \tau_\alpha^2 + g_{k,t}\sigma_\alpha^2\Big),$$
$$\delta_{k,t} \sim \mathcal{N}\Big(\delta_{k,t-1},\, \tau_\delta^2 + g_{k,t}\sigma_\delta^2\Big)$$

where $\sigma_\alpha, \sigma_\delta, \tau_\alpha, \tau_\delta > 0$. This structure distinguishes between two sources of latent change. The first, governed by $\sigma_\cdot$, is a continuous time component that captures uncertainty accumulating in the background, whether a team plays or not, just as it does in the CTRW model above. The second source of change, governed by $\tau_\cdot$, is an active component that applies only in weeks when a team plays. The model is therefore closely related to the continuous-time random walk, but

disaggregates the variance associated with match participation and the variance associated with the passage of calendar time. This is designed for sports like international rugby union where teams play in bursts of several weeks separated by long gaps of inactivity.

3.3. Prior distributions

The priors are weakly informative and are intended to regularize the latent state process without imposing strong substantive restrictions. The scoring intercepts κ_h, κ_a use $\mathcal{N}(0, 1)$ on the log-score scale, which allows a wide range of plausible rugby scoring rates. The home-advantage parameter η uses $\mathcal{N}(0, 0.5^2)$, allowing moderate multiplicative home effects while discouraging implausibly large advantages. The negative-binomial dispersion parameters ϕ_h, ϕ_a use Gamma$(2, 0.5)$ priors, using the shape–rate parameterization. In the NB2(μ, ϕ) parameterization, this prior permits substantial overdispersion relative to a Poisson model while still regularizing very small dispersion values.

In all specifications, the attack and defense innovation scales $\sigma_\alpha, \sigma_\delta, \tau_\alpha$, and τ_δ are assigned $\mathcal{N}^+(0, 0.2^2)$, reflecting the expectation that team strengths usually change gradually between periods while allowing larger shifts where supported by the data. The initial-state scales $\sigma_{\alpha 0}, \sigma_{\delta 0}$ use $\mathcal{N}^+(0, 0.5^2)$, allowing meaningful initial differences between teams without permitting the initial rankings to dominate the likelihood. Taken together, these priors allow substantial movement in latent strength over time while discouraging implausibly large week-to-week shifts unless the data strongly support them (see the supplementary material for alternative prior specifications).

All models use non-centered parameterizations (Betancourt and Girolami 2015) for the latent initial states and innovations, with the underlying raw error terms assigned standard normal priors. For identification, the latent attack and defense parameters are centered at each modeled time point after applying the relevant state evolution step. This centering fixes the location of the latent scale without changing the interpretation of the innovation scale parameters as governing movement in relative attack and defense strength.

3.4. Estimation

All models are estimated in Stan using Hamiltonian Monte Carlo with the No-U-Turn Sampler. For the main out-of-sample comparisons reported below, each weekly refit used three chains with 1,000 warmup iterations and 1,000 post-warmup iterations per chain, with `adapt_delta` = 0.95. For the final full-sample analyses used to interpret latent attack and defense trajectories, the same basic estimation strategy is retained, but four chains are used and the models are fit to the full sample rather than repeatedly re-estimated on rolling training windows.

The latent-state parameters are high dimensional, since each team has separate attack and defense strengths that evolve over many weeks. For that reason, model comparison focuses on a restricted set of scalar parameters and standard HMC diagnostics, including $\hat{R}$, effective sample size, and divergences. Posterior summaries of latent strengths are then obtained by extracting the weekly state matrices and transforming them to the index scale used in the substantive results.

4. Data and design

4.1. Data

The estimation sample comprises 584 men's international rugby union matches played between February 2020 and December 2025, and obtained from `ultimaterugby.com`. The dataset is not intended to be exhaustive for all international sides. Rather, it includes results from 32 national teams, with complete coverage for the ten most established teams, the so-called "tier-one" nations, and more sporadic coverage for a wider group of lower-tier national teams (see the supplementary materials for more information on the national teams included). This structure is appropriate for the paper's substantive aims, where the focus is on the evolution of tier-one attack and defense strengths, and including a broader set of opponents improves identification by locating those teams in a richer competitive network. At the same time, the dataset should not be interpreted as a comprehensive census of all men's international rugby union matches during the period.

The resulting match calendar is highly uneven. Fixtures occur in concentrated bursts asso-

ciated with recurring international competitions and tours. In a typical year, the men's international rugby calendar is structured around the Six Nations, contested by six tier-one European teams in February and March; the Rugby Championship, contested by four tier-one Southern Hemisphere teams in August and September; the mid-year matches in June and July; and the autumn matches in November, when Northern and Southern Hemisphere teams play each other. Every four years, these patterns are disrupted by the Rugby World Cup, which in our observation window took place in September and October 2023. Between these events lie long periods of relative inactivity. Weekly time provides a convenient indexing device for such a calendar, but international rugby is not played on a regular weekly schedule, and match spacing varies substantially across teams and seasons.

4.2. Research design

The empirical analysis has two stages. The first is an out-of-sample forecasting exercise designed to compare the three Bayesian models and benchmark them against two simpler algorithms. The second is a full-sample substantive analysis that examines our estimated latent attack and defense strengths among tier-one teams. The forecasting exercise therefore serves as a model-testing stage, while the final latent-strength analysis serves as an indication of the wider implications of such models.

The out-of-sample evaluation is based on event weeks in the 2025 calendar year. All matches from 2020 through 2024 are available for model training, and each week in 2025 containing at least one match is treated as a held-out evaluation week. For each evaluation week, the model is re-estimated using all matches strictly before that week as training observations. Predictive performance is then scored only on the matches played during the held-out week. Predictive performance is then scored only on the matches played during the held-out week. This rolling-origin design yields a fair comparison across models while preserving each model's own dynamic structure.

Predictive performance was assessed using several scoring metrics. For the paired scoreline

as a whole, we report the *mean joint log predictive density* (MJLPD), computed as the average log predictive density assigned to the observed home-away score pair across held-out matches. Formally, if $p(y_{g,h}, y_{g,a} \mid \mathcal{D}_{-g})$ denotes the model's posterior predictive density for the scoreline in held-out match g, then the joint log score is

$$\log p(y_{g,h}^{\text{obs}}, y_{g,a}^{\text{obs}} \mid \mathcal{D}_{-g}),$$

and MJLPD is the average of this quantity across matches. Higher values indicate better predictive performance, since the model assigns greater probability mass to the outcomes that actually occurred. MJLPD is a particularly demanding criterion because it evaluates the full joint scoreline distribution rather than only a reduced summary such as margin or match outcome.

For individual score components and point difference, we also report the continuous ranked probability score (CRPS). CRPS compares the full predictive distribution of a scalar outcome to the observed realization and may be interpreted as a distributional generalization of absolute error. Lower values indicate better predictive performance. In the present setting, CRPS is calculated separately for home score, away score, and points difference, allowing us to evaluate not only overall scoreline fit but also whether models differ in how well they predict each team's scoring and the match margin. We also include the Brier score for home-win probability and root mean squared error for point difference to facilitate comparisons with the two simple benchmarks, which produce scalar predictions. In addition to these predictive metrics, we record several standard MCMC diagnostics, including $\hat{R}$, effective sample size, and divergences, to ensure that predictive comparisons are not being driven by obvious sampling failures.

Alongside the Bayesian models, we include two simpler benchmarks. First, an exponential error-correction smoother (ECES) for point difference, which has been popularized in a rugby union context by statistician David Scott in the `www.statschat.org.nz` blog. In this benchmark, each team has a single scalar quality rating. Expected margin equals the home team's rating minus the away team's rating plus a fixed home-advantage term, and ratings are updated recursively by

a fraction of the forecast error. We tuned the smoothing and home-advantage parameters using results from calendar year 2024, which were then fixed for the 2025 out-of-sample evaluation.

Second, we also used the official World Rugby ranking algorithm (available at `https://www.world.rugby/rankings`). Although not a predictive model as such, the rankings are widely followed and discussed in rugby union circles, and provides a well known benchmark. For this comparison, pre-match ranking points differences, adjusted for home advantage, were treated as the benchmark prediction of match margin, and home-win probabilities were obtained via a normal approximation using the historical residual standard deviation. Because both these benchmarks focus on margins rather than pairs of scores, benchmark comparison focuses on margin RMSE and the Brier score for home-win probability.

Finally, after the forecasting comparison is complete, the three Bayesian models are estimated on the full sample. From these models we extract and discuss the latent attack and defense trajectories of tier-one teams, as well as the mean home-advantage parameter. The purpose of this final stage is descriptive and substantive: it allows us to examine how different teams' strength profiles evolved over time, whether the attack–defense decomposition reveals meaningful distinctions that are obscured by simpler ranking systems, and to what extent home advantage matters in international rugby once teams' attack and defense strengths are taken into account.

5. Results

5.1. Out-of-sample forecasting performance

Across the 116 held-out matches and 25 weeks of play in the 2025 season, the three Bayesian models perform similarly well, with no single specification clearly dominating across all metrics (see Table 1). The random-walk (RW), continuous-time random-walk (CTRW), and two-speed event-gap (TS) models have essentially identical mean joint log predictive densities. The CTRW model performs marginally best on CRPS for home score, CRPS for point difference, and RMSE for point difference, while the TS model performs marginally best on MAE for point difference.

The RW model has the lowest Brier score, but only by a very small margin. The main conclusion is therefore not that one Bayesian dynamic specification is decisively more accurate than the others, but that all three provide similar out-of-sample predictive performance.

Table 1. Out-of-sample predictive performance in the 2025 event-week evaluation

Metric	RW	CTRW	TS	ECES	WRRA
MJLPD	−7.74	−7.74	−7.74	–	–
CRPS_H	6.47	6.38	6.40	–	–
CRPS_A	6.69	6.73	6.74	–	–
CRPS_{PD}	10.17	10.15	10.19	–	–
RMSE_{PD}	19.52	19.45	19.51	20.55	22.47
MAE_{PD}	14.08	13.97	13.96	14.81	15.98
Brier	0.144	0.145	0.145	0.147	0.170
Time	11.8	5.6	8.2	< 0.1	< 0.1
Max $\hat{R}$	1.017	1.014	1.015	–	–
Min ESS	282	283	338	–	–

Notes: MJLPD = mean joint log predictive density (higher is better); CRPS_H = CRPS for home score (lower is better); CRPS_A = CRPS for away score; CRPS_{PD} = CRPS for point difference; RMSE_{PD} = Root mean square error for point difference (lower is better); MAE_{PD} = Mean absolute error for points difference (lower is better); Brier = Brier score for home-win probability (lower is better); Time = median computation time in minutes across all 25 iterations; Max $\hat{R}$ = the maximum potential scale reduction factor across all 25 iterations (closer to 1 is better); Min ESS = minimum bulk effective sample size (higher is better); CTRW = continuous time random walk; TS = two-speed model; ECES = error correction exponential smoother; WRRA = World Rugby rankings algorithm. All three Bayesian models had zero divergent transitions in the final run.

The stronger distinction across the Bayesian models is computational rather than predictive. The RW model is the slowest to estimate, with a median computation time of 11.8 minutes across evaluation weeks. The CTRW and TS models are faster, at 5.6 and 8.2 minutes respectively, while delivering predictive performance that is at least as good as the RW model on most margin-based measures. This suggests that, if one Bayesian specification must be selected for routine forecasting, the CTRW or TS models are more attractive than the weekly RW model: they retain the predictive advantages of the Bayesian state-space approach while better matching the irregular timing of international rugby fixtures.

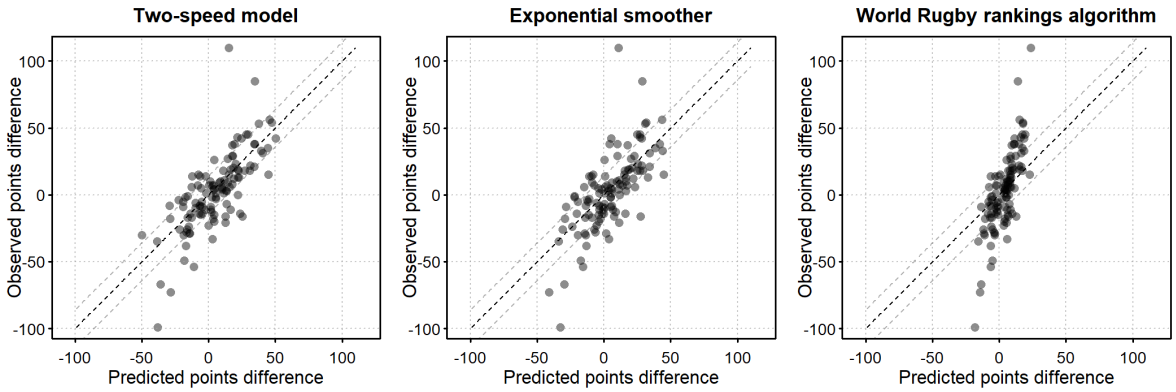
The benchmark algorithms provide a useful scale against which to interpret these results. The error-correction exponential smoother (ECES) performs respectably, confirming that a simple

rating model captures much of the structure in international rugby results. However, on the directly comparable point-difference measures, the Bayesian models are consistently stronger. Their RMSE values are around one point lower than ECES and around three points lower than the World Rugby rankings algorithm (WRRA), while their MAE values are also lower than both benchmarks. The gains in Brier score over ECES are modest, but all three Bayesian models are better calibrated for home-win prediction than WRRA.

To examine the predictions in more substantively meaningful terms, we show predicted against observed point differences for the 116 held-out matches in Figure 1, comparing the TS model with the two benchmark algorithms. The other Bayesian models produce very similar plots, so the TS figure is representative of the Bayesian specifications more generally. Many of the TS model predictions fall within 14 points of the observed point difference, which corresponds to two converted tries in rugby scoring. The mean absolute error of the OOS predictions is approximately 14 points, as shown in Table 1, although the median absolute errors are close to 10. As the figure shows, large margins remain difficult to predict for all methods, which is unsurprising in a high-scoring sport where occasional mismatches and late-game scoring can produce substantial variation in final margins. ECES also performs reasonably well by this standard, but WRRA is less well calibrated, especially in matches with large observed point differences.

Taken together, the forecasting results show that the three Bayesian dynamic-strength models perform similarly well and improve meaningfully, though not dramatically, on simple benchmark rating algorithms. The choice among the Bayesian models should therefore be guided less by small differences in predictive accuracy than by computational cost and substantive fit to the structure of international rugby. On those grounds, the CTRW and TS specifications are the most attractive: both account for irregular gaps between team appearances, both are faster than the weekly RW model, and both retain near-best out-of-sample predictive performance.

Figure 1. Observed vs out-of-sample predicted points difference



Observed point differences for the 116 matches in the 2025 held-out sample (y-axes) plotted against predicted point differences from the two-speed model and the two benchmark algorithms: the exponential smoother and the World Rugby rankings algorithm. The diagonal lines indicate prediction errors of -14, 0, and $+14$ points, corresponding to two converted tries below, exactly equal to, or above the observed point difference.

5.2. Latent strength trajectories over time

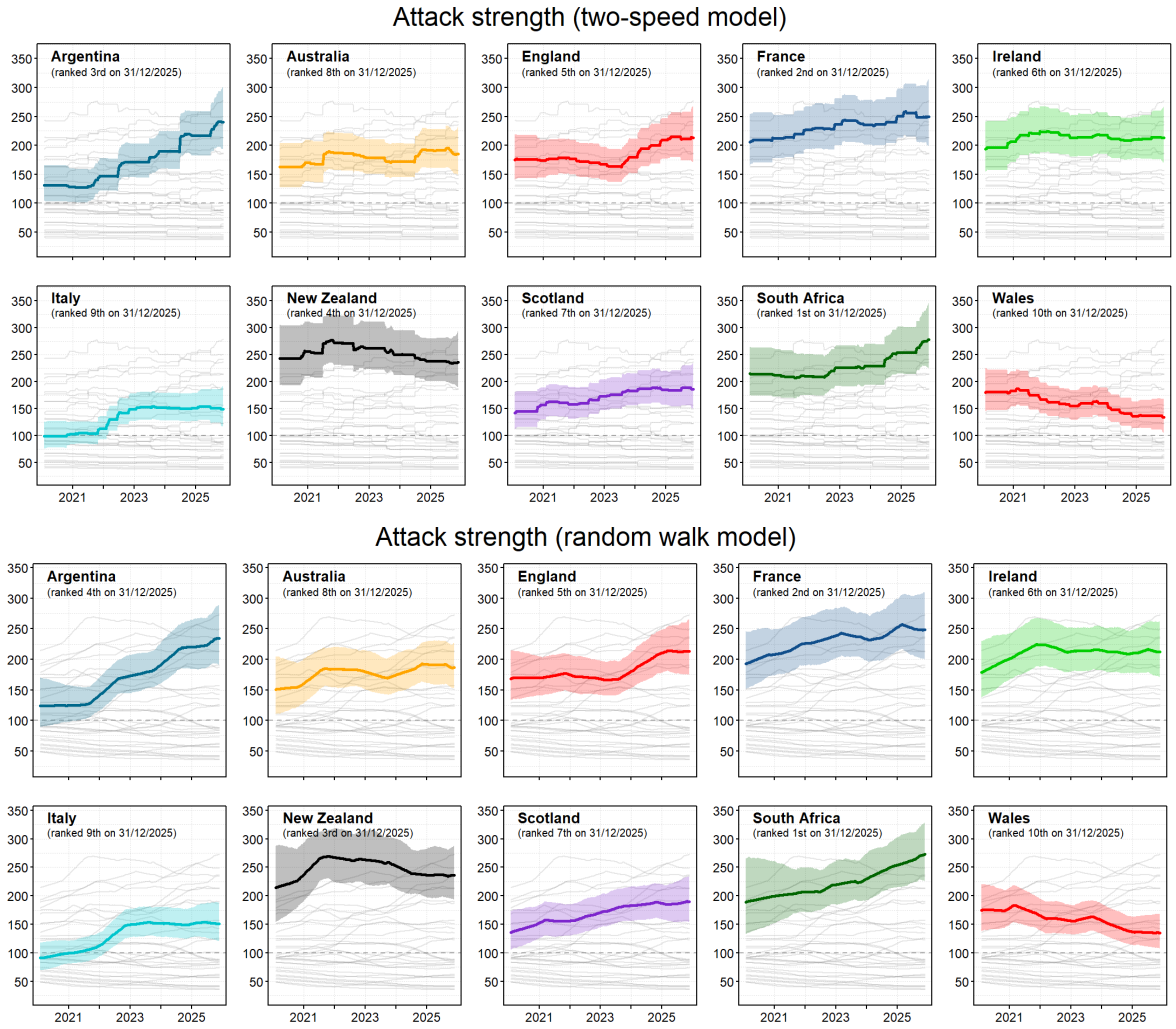
We now turn from predictive performance to substantive interpretation by examining the estimated latent strengths themselves, asking what the models imply about how the strength of leading international teams evolved over the 2020-2025 period.

The latent-strength estimates provide a richer picture of international rugby than existing rating systems can offer. Most obviously, the Bayesian models distinguish between attack and defense, allowing teams to differ not only in overall strength but also in the way that strength is produced. Elite international sides are not all strong in the same way: some are driven primarily by scoring power, others by defensive solidity, and others by a more balanced profile. The dynamic attack–defense framework therefore helps clarify whether changes in team performance reflect shifts in attacking potency, defensive resilience, or both.

Because the final analysis is concerned with substantive interpretation rather than exhaustive model comparison, we focus on the ten tier-one teams. These are the most substantively important cases and are also the best observed throughout the sample. Figures 2 and 3 present estimated attack and defense strengths from the two-speed (TS) and random-walk (RW) models. We

show the TS model because it allows latent strength to evolve through both a background elapsed-time component and an additional active-appearance component. The RW model provides a useful comparison because it represents the simpler weekly dynamic specification. The continuous-time random-walk (CTRW) results are very similar to the TS estimates and are reported in the supplementary materials.

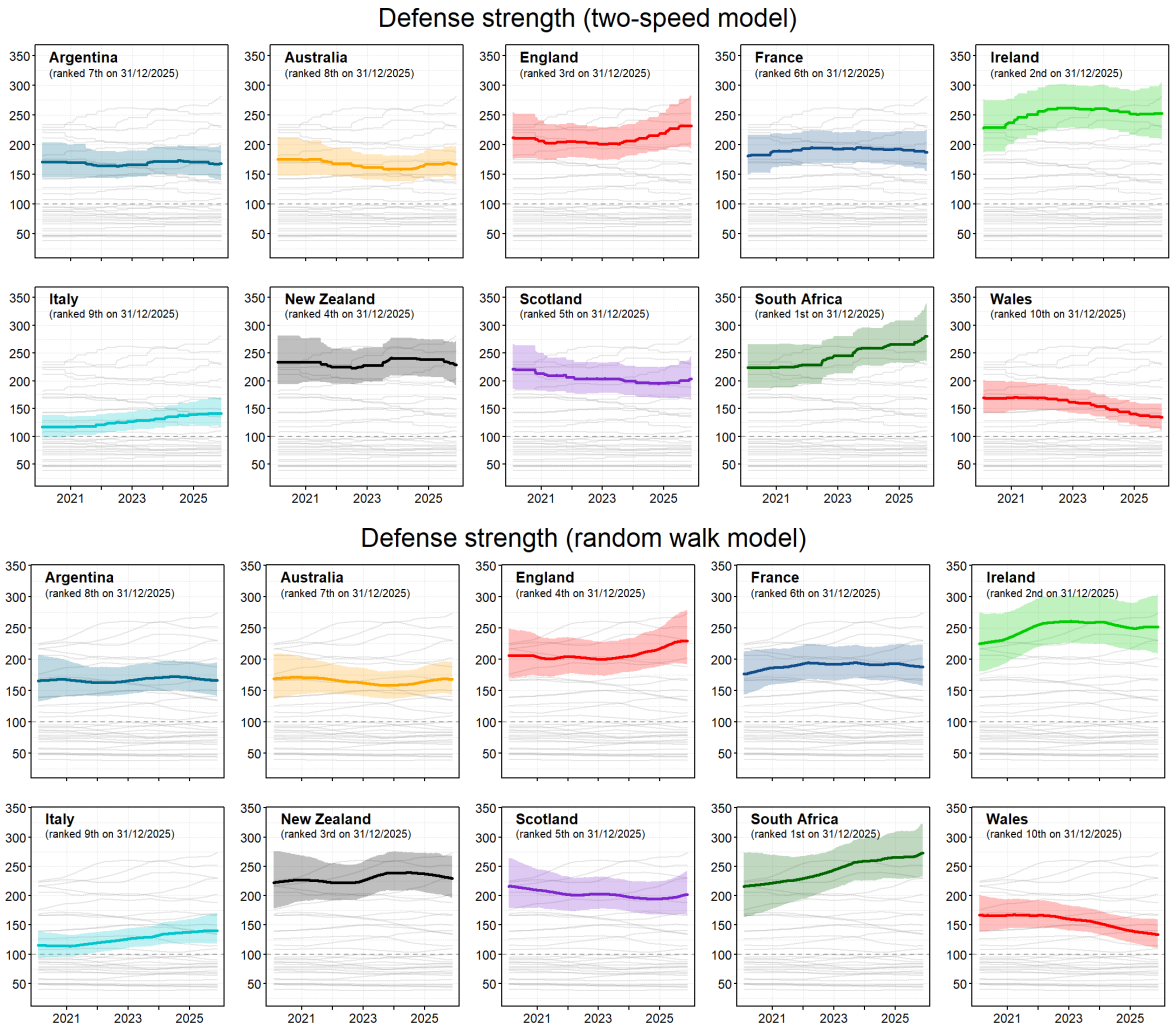
Figure 2. Estimated team attack strengths, TS and RW models



Estimated attack strengths for tier-one men's international rugby teams, 2020–2025. Shaded regions indicate 90% credible intervals. Estimates are shown on the index scale, $\exp(\alpha_{k,t}) \times 100$, so that 100 represents the contemporaneous average team in the model. The top panel shows the two-speed model; the bottom panel shows the random-walk model.

Several broad patterns emerge from the trajectories. First, the separation of attack and

Figure 3. Estimated team defense strengths, TS and RW models



Estimated defense strengths for tier-one men's international rugby teams, 2020–2025. Shaded regions indicate 90% credible intervals. Estimates are shown on the index scale, $\exp(\delta_{k,t}) \times 100$, so that 100 represents the contemporaneous average team in the model. Higher values indicate stronger defense. The top panel shows the two-speed model; the bottom panel shows the random-walk model.

defense strengths allows for more insight into the strategies and strengths of different teams. Some teams are stronger in attack than defense, while for others the reverse is true. For example, France is one of the strongest attacking sides in the system (ranked 2nd at the end of 2025) but is notably less dominant in defense (ranked 6th). Ireland shows the opposite profile: its defense is estimated to be very strong, while its attack, although still well above the sample average, is less distinctive

than France's.

The estimates also differentiate the character of other leading teams. South Africa stands out as a side that is elite in both dimensions, with very high attack and defense throughout the later part of the sample. New Zealand also appears broadly balanced at a high level, though generally a little less dominant than South Africa at the end of the period. Argentina's recent improvement is especially visible in attack, while Wales' decline is apparent in both attack and defense. These patterns are substantively useful because they distinguish not just which teams are strong, but what kind of strong team each side is.

The comparison across specifications is also revealing. The RW model evolves latent strengths over weekly calendar time, regardless of whether a team plays in a given week. That produces somewhat implausible outcomes where international teams' strengths continue evolving even when those teams are dormant because players are playing in domestic leagues or it is the off-season. By contrast, the TS and CTRW models are more tightly tied to the event structure of the sport. When team do not play, their latent strength estimates are broadly flat, even while latent uncertainty can still grow.

As such, both the TS and CTRW models produce trajectories that align more naturally with the actual cadence of international rugby. For that reason, they are useful models not only for forecasting but also for contributing to understandings of patterns and trends in international rugby team strengths.

5.3. Home advantage

Finally, we briefly consider the home-advantage parameter estimates, η. Home advantage is of substantive interest to fans and analysts (e.g., Higgs and Stavness 2021; Pledger and Morton 2011) because it indicates whether teams systematically score more when playing at home, net of their own latent quality and that of their opponent. Across the Bayesian dynamic-strength models, the posterior estimates of home advantage are small and positive, ranging from 0.061 to 0.067. On the log-score scale, this implies approximately a 6–7% increase in expected home scoring for non-

neutral matches, holding latent attack and defense strengths fixed. In rugby terms, this corresponds to around 1.5 additional points for a team otherwise expected to score in the mid-20s. The 95% posterior intervals include zero, however, so the evidence for a distinct home-scoring advantage is surprisingly modest.

This estimate is smaller than that reported by Pledger and Morton (2011), who estimated home ground advantage in the 2004 Super Rugby (a professional rugby club competition) season to be 7.2 points. One reason for the more modest home advantage in our results is that "home" in international rugby is not equivalent to a fixed home ground in a professional league. National teams may play home fixtures across multiple stadia and cities, and some matches may therefore provide a stronger local advantage than others. A single average home-advantage parameter will tend to smooth over this variation. This suggests that home advantage in international rugby may be partly venue- or context-specific, an issue that could be explored in future research. It is also worth noting that the separately estimated exponential-smoothing benchmark selected a somewhat larger, though still modest, home-advantage adjustment of 2.75 points in its 2024 tuning exercise using RMSE as the selection criterion.

6. Conclusion

This article has developed and compared dynamic Bayesian paired-score models for men's international rugby union that decompose team power into latent attack and defense strengths. The first contribution is to bring this attack-defense framework to international rugby, showing that it yields substantively useful estimates of how teams differ not only in overall strength but also in the balance between scoring power and defensive resilience. The resulting trajectories provide a richer account of international team strength than scalar rating systems alone, and they help clarify the different ways in which leading teams sustain, gain, or lose competitive advantage over time. Our analyses show that the main story of recent tier one international rugby is not simply one of stable overall rankings, but of changing combinations of attack and defense: France and Argentina's ascents as attack-focused teams, Ireland's defensive consistency and South Africa's

increasing dominance on both fronts.

The second contribution is to compare alternative dynamic formulations in a sport where fixture timing is highly irregular. In out-of-sample prediction, all three Bayesian models improved on a simpler exponential error-correction benchmark, but there was relatively little to separate them on predictive accuracy alone. Yet examination of the strength estimates suggest that this issue is not merely technical. Models that differ only in how latent strengths evolve can produce meaningfully different forecasting profiles and different substantive trajectories.

More broadly, the results suggest that the most important modeling choice in this setting is not whether team quality changes over time, but how time itself should be represented. International rugby is not played on a smooth weekly clock. Models that allow latent uncertainty to evolve with elapsed time, and (in the two-speed case) also allow for the possibility of additional change in periods when teams are active, are more closely aligned with the structure of the sport. Future work could extend this framework by examining professional rugby leagues, applying our models to other international sports with irregular calendars, or by considering more complex modeling features such as interactions between attack and defense strengths. The central conclusion is that dynamic latent-strength models are both feasible and informative in international rugby, and that an attack-defense approach provides a useful foundation for both forecasting and substantive analysis.

References

Betancourt, M. and M. Girolami (2015). Hamiltonian monte carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, and A. Loganathan (Eds.), *Current Trends in Bayesian Methodology with Applications*, pp. 79–101. Boca Raton: CRC Press.

Cameron, A. C. and P. K. Trivedi (2013). *Regression Analysis of Count Data* (2 ed.). Cambridge: Cambridge University Press.

Crowder, M., M. Dixon, A. Ledford, and M. Robinson (2002). Dynamic modelling and prediction of english football league matches for betting. *Journal of the Royal Statistical Society: Series D (The Statistician) 51*(2), 157–168.

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

Glickman, M. E. and H. S. Stern (1998). A state-space model for national football league scores. *Journal of the American Statistical Association 93*(441), 25–35.

Higgs, N. and I. Stavness (2021). Bayesian analysis of home advantage in north american professional sports before and during covid-19. *Scientific Reports 11*(1), 14521.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *The Statistician 52*(3), 381–393.

Koopman, S. J. and R. Lit (2019). Forecasting football match results in national league competitions using score-driven time series models. *International Journal of Forecasting 35*(2), 797–809.

Lopez, M. J., G. J. Matthews, and B. S. Baumer (2018). How often does the best team win? a unified approach to understanding randomness in north american sport. *The Annals of Applied Statistics 12*(4), 2483–2516.

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

Pledger, M. J. and R. H. Morton (2011). Modelling the 2004 super 12 rugby union competition. *Australian & New Zealand Journal of Statistics 53*(1), 109–121.

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

Stan Development Team (2025). *Stan Reference Manual, Version 2.39*. Version 2.39.