Journal of Big Data

https://doi.org/10.1186/s40537-026-01486-6

Article in Press

Bayesian dynamic Bradley-Terry model with commensurate spike-and-slab priors

Received: 26 September 2025

Accepted: 31 May 2026

Published online: 27 June 2026

Cite this article as: Macrì-Demartino R., Egidi L. & Torelli N. Bayesian dynamic Bradley-Terry model with commensurate spike-and-slab priors. *J Big Data* (2026). https://doi.org/10.1186/s40537-026-01486-6

Roberto Macrì-Demartino, Leonardo Egidi & Nicola Torelli

Bayesian dynamic Bradley-Terry model with commensurate spike-and-slab priors

Roberto Macrì-Demartino[1*], Leonardo Egidi[1] and Nicola Torelli[1]

[1]Department of Economics, Business, Mathematics and Statistics "Bruno de Finetti", University of Trieste, Via Valerio 4/1, Trieste, 34127, Friuli-Venezia Giulia, Italy.

*Corresponding author(s). E-mail(s): roberto.macridemartino@deams.units.it;
Contributing authors: legidi@units.it; nicola.torelli@deams.units.it;

Abstract

We propose a Bayesian dynamic Bradley-Terry model that introduces team- and time-specific commensurate spike-and-slab priors on the innovation precisions governing the evolution of teams' strength parameter. This sparse state-space allows the model to adaptively borrow information from a team's past performance, strongly shrinking towards it when performance is stable or switching to a more diffuse prior to capture sudden changes. We apply our model to the last ten NBA seasons consisting of 12,841 matches played across the regular season, NBA Cup, play-in tournament, and playoffs. The results show that the model captures major changes in team performance that align with well-documented events such as roster changes and injuries. Furthermore, out-of-sample predictions exhibit better forecasting accuracy compared to two well-known dynamic models, particularly in later playoff stages, as measured by lower Brier scores. These findings suggest that the proposed model provides a valid dynamic extension of the Bradley-Terry model, incorporating adaptive temporal borrowing to improve both interpretability and predictive performance.

Keywords: Basketball, Borrowing, Pairwise comparison, Sport Analytics, State-space models

1 Introduction

The analysis of sports data has long attracted considerable interest. In many competitive contexts, outcomes are determined not by isolated individual performances but through direct head-to-head matches. Sports such as chess, football, tennis, basketball, and volleyball involve direct matches between individuals or teams. Consequently, modelling pairwise comparisons has become an important problem, with a well-established literature in areas such as preference elicitation and item ranking (David 1988; Cattelan 2012).

The foundation for modern paired comparison models was developed by Bradley and Terry (1952) and Luce (1959), based on earlier ideas from Thurstone (1927). The Bradley-Terry model (Bradley and Terry 1952) assigns a strength parameter to each competitor (e.g., each team or player), and the probability that one defeats another is determined by the ratio of their strengths. In its simplest form, these strength parameters are assumed to be constant, and the model's estimates provide a rating system for ranking competitors based on their match outcomes.

Over the years, several extensions and generalisations of the basic Bradley-Terry model have been proposed to address various limitations. An important extension deals with the possibility of tied outcomes in competitions. Rao and Kupper (1967) and Davidson (1970) expanded the model to scenarios where a draw (tie) is possible by introducing an additional parameter to capture the probability of a tie. Similarly, Agresti (2002) discussed paired comparison models with ties in the broader context of ordinal response models, using cumulative link frameworks. Other generalisations have focused on incorporating context-specific or team-specific effects. For instance, Springall (1973) proposed a Bradley-Terry generalisation with team-dependent linear predictors, allowing team-specific covariates or effects to influence match outcomes. The Bradley-Terry model has been further adapted to account for order effects, such as home-field or first-move advantage. Specifically, Beaver and Gokhale (1975) and Davidson and Beaver (1977) introduced additive and multiplicative order effects, respectively. Bayesian formulations have also been widely explored, beginning with conjugate priors for teams' log-strengths parameter (Davidson and Solomon 1973) and evolving into more flexible approaches using non-conjugate multivariate Gaussian priors (Leonard 1977). Since then, a wide range of Bayesian extensions have been proposed, introducing novel prior structures, advanced computational techniques, and several applications (Chen and Smith 1984; Caron and Doucet 2012; Whelan 2017; Osei and Davidov 2022; Wainer 2023; Macrì Demartino et al. 2024, among others).

Another crucial extension of the Bradley-Terry model is to allow team strengths to vary over time. In practice, paired comparison data often span multiple seasons or weeks, during which team performance levels may change. To capture such temporal dynamics, Fahrmeir and Tutz (1994) developed a state-space generalisation of the Bradley-Terry model for tournament data, treating team strength parameters as latent variables that evolve over time. Similarly, Cattelan et al. (2012) described the evolution of team strength through an exponentially weighted moving average process, and Bong et al. (2020) proposed a nonparametric time-varying generalisation of the Bradley-Terry model. In a Bayesian context, dynamic Bradley-Terry models have often been implemented using state-space approaches (Fahrmeir and Tutz 1994; Glickman

1999; Knorr-Held 2000; Glickman 2001; Koopmeiners 2012; Baker and McHale 2014; Lopez et al. 2018; Ingram 2021; Duffield et al. 2024; Glickman 2025, among others). Specifically, the Glicko (Glickman 1999) and Glicko-2 (Glickman 2001) rating systems implement such dynamic pairwise comparison models in practice. These systems have been widely adopted by national and international organisations, online platforms, and competitive gaming leagues as essential tools for assessing competitor strength (e.g. in chess), underscoring the practical value of dynamic rating models (Glickman and Jones 2024). Importantly, Glickman (2001) noted that assuming a fixed (unknown) innovation variance for the evolution of team strengths is often too restrictive, particularly when teams experience rapid changes in performance. To address this, he proposed a model in which the evolution variance varies stochastically between teams and over time, allowing for sudden shifts in team ability.

In this paper, we introduce a novel Bayesian dynamic Bradley-Terry model that offers flexibility in modelling time-varying abilities. We adopt a weighted (sparse) state-space approach with flexible innovation variances for each team at each time period, adapting according to the similarity (commensurability) between current and past performance. This formulation allows the model to adaptively borrow information from a team's past performance, but only to the extent supported by the data. Our approach is based on the Bayesian weighted framework of Macrì-Demartino et al. (2025), developed for dynamic goal-based football models, and implemented in the `footBayes` R package (Egidi et al. 2025). We extend this by making the innovation variance itself both time- and team-specific, providing a more flexible way to capture heterogeneous dynamic patterns across teams.

The paper is organised as follows. Section 2 presents the Bradley-Terry model and some of its extensions. Furthermore, Section 3 describes the commensurate prior framework and introduces our proposed state-space approach for the commensurate innovation variances of teams' strengths in the dynamic Bradley-Terry model. In Section 4, we apply our methodology to the National Basketball Association (NBA), using the matches played in the last ten seasons from 2015 to 2025. Finally, Section 5 provides concluding remarks, discussing limitations, advantages, and potential future research directions.

2 The Bradley-Terry Model

The Bradley-Terry (BT) model (Bradley and Terry 1952) is a widely used approach to model players or teams based on pairwise comparisons. The model assumes that each team T_k, with $k = 1, \ldots, N_T$, has an associated latent strength parameter $\alpha_k \in \mathbb{R}^+$. The outcome of a match between two distinct teams T_i and T_j is modelled as a Bernoulli trial, where the probability that T_i defeats T_j is given by the ratio of their strengths

$$p_{i,j} = \mathbb{P}(T_i \text{ defeats } T_j) = \frac{\alpha_i}{\alpha_i + \alpha_j}, \tag{1}$$

where α_i and α_j are the strength parameters of teams T_i and T_j, respectively. Note that these strength parameters are only defined up to an arbitrary multiplicative constant.

A common and useful reparameterization formulates the model in (1) on the log-scale. The win probability can then be written in logistic form as

$$p_{i,j} = \frac{\exp(\lambda_i)}{\exp(\lambda_i) + \exp(\lambda_j)} = \text{logit}^{-1}(\lambda_i - \lambda_j), \tag{2}$$

where $\lambda_i = \log(\alpha_i)$. The log-strength parameters λ_k are identifiable up to an additive constant, which is typically addressed by imposing a sum-to-zero constraint (Baio and Blangiardo 2010) as follows

$$\sum_{k=1}^{N_T} \lambda_k = 0.$$

This transformation offers several advantages. Notably, it enables the estimation of the log-strength parameter across an expanded parameter space $\lambda_k \in (-\infty, +\infty)$, facilitates the application of a broader class of priors in Bayesian inference, and enables parameter estimation using standard generalised linear models (GLMs) within the frequentist framework (Cattelan 2012).

The Bayesian formulation of the BT model was first introduced by Davidson and Solomon (1973), who used a family of conjugate prior distributions to derive the posterior analytically. Leonard (1977) later proposed the use of more flexible non-conjugate priors, specifically multivariate Gaussian distributions. This approach offers greater flexibility for model extensions, simplifies the interpretation and specification of prior beliefs, and naturally accommodates hierarchical structures (Issa Mattos and Martins Silva Ramos 2022). Furthermore, Whelan (2017) proposed a set of desirable properties for priors in pairwise comparison models to ensure fairness and avoid any advantages or disadvantages for a particular team. Both multivariate Gaussian distributions (Leonard 1977) and independent identical Gaussian priors satisfy these criteria. Based on this, let Y_{ij} be a binary indicator that team T_i wins against team T_j. Then the model can be specified as

$$\begin{aligned} Y_{i,j} \mid p_{i,j} &\sim \text{Bernoulli}(p_{i,j}), \\ \lambda_k \mid \mu_k, \sigma_k^2 &\sim \text{N}(\mu_k, \sigma_k^2), \end{aligned} \tag{3}$$

where μ_k is the mean for the team's log-strength and σ_k^2 denotes the corresponding variance. In practice, it is common to adopt non-informative priors by setting $\mu_k = 0$ and assuming a common variance $\sigma_k^2 = \sigma^2$ across all teams. This choice ensures that no team is assigned an a priori advantage over any other (Whelan 2017).

2.1 Order effect

In some paired-comparison settings, the order in which two opponents are presented can bias the outcome. Typical examples are home-field advantage in sports (home team vs. visiting team) and first-move advantage in chess. To account for such order effects, Davidson and Beaver (1977) proposed an extension of the BT model in (2)

considering a multiplicative model with an additional parameter $\phi \in \mathbb{R}$. This multiplicative parameter becomes on the log-scale an additive term, leading to the following Bayesian BT model

$$\begin{aligned} p_{i,j} &= \text{logit}^{-1}(\lambda_i - \lambda_j + \phi \times \mathbb{1}_{\{\text{i is home}\}}) \\ Y_{i,j} \mid p_{i,j} &\sim \text{Bernoulli}(p_{i,j}), \end{aligned} \tag{4}$$

where $\mathbb{1}_{\{\text{i is home}\}}$ is an indicator function equal to 1 when the team T_i plays at home and 0 otherwise. With prior distributions

$$\begin{aligned} \lambda_k \mid \mu_k, \sigma_k^2 &\sim \text{N}(\mu_k, \sigma_k^2) \\ \phi \mid \xi^2 &\sim \text{N}(0, \xi^2). \end{aligned}$$

A positive value ($\phi > 0$) indicates an advantage for the home team, increasing its probability of winning. Conversely, a negative value ($\phi < 0$) favours the visiting team. When $\phi = 0$, the model reduces to the standard BT form, where the outcome depends solely on the strengths of the teams.

2.2 Dynamic Bradley-Terry model

The basic BT model assumes that each team has a static strength parameter that remains constant over time. However, team performance is often dynamic – it fluctuates across seasons and even from week to week – since roster changes, fatigue, injuries to key players, or coaching changes can cause substantial shifts in a team's strength. To capture these dynamics, the basic BT in (2) is extended to allow time-varying team strengths. Let $\lambda_{k,t}$ denote the latent log-strength of team T_k at time t, where $t = 1, 2, \ldots, \mathcal{T}$ indexes time periods (e.g., seasons or weeks). Let $Y_{i,j,t}$ be a binary indicator that team T_i defeats team T_j at time t. The model is

$$\begin{aligned} p_{i,j,t} &= \text{logit}^{-1}(\lambda_{i,t} - \lambda_{j,t}), \\ Y_{i,j,t} \mid p_{i,j,t} &\sim \text{Bernoulli}(p_{i,j,t}). \end{aligned} \tag{5}$$

In a Bayesian framework, dynamic BT models are typically expressed as state-space models in which the latent log-strength of each team evolves as a Gaussian stochastic process (Glickman 1999; Knorr-Held 2000; Glickman 2001; Koopmeiners 2012; Baker and McHale 2014; Lopez et al. 2018; Ingram 2021; Glickman 2025, among others). In particular, Glickman (2001) treated each $\lambda_{k,t}$ as a latent state that follows a random walk with team- and time-specific innovation variances. This allows team log-strengths to evolve gradually over time, naturally reflecting changes in performance. For each team T_k, with $k = 1, \ldots, N_T$, and for each period $t = 2, \ldots, \mathcal{T}$ the following state transition model for the log-strength is assumed

$$\lambda_{k,t} \mid \lambda_{k,t-1}, \sigma_{k,t}^2 \sim \text{N}(\lambda_{k,t-1}, \sigma_{k,t}^2), \tag{6}$$

where the log-strength of team T_k at time t is normally distributed around its value in the previous period, $\lambda_{k,t-1}$. The innovation variance $\sigma_{k,t}^2$ controls how much the

log-strengths can change: larger values allow for greater uncertainty, while smaller values imply more stable performance and stronger shrinkage of $\lambda_{k,t}$ towards $\lambda_{k,t-1}$. Additionally, Glickman (2001) modelled the innovation variances as follows

$$\log(\sigma^2_{k,t}) \mid \sigma^2_{k,t-1}, \psi^2 \sim \mathrm{N}(\log(\sigma^2_{k,t-1}), \psi^2). \tag{7}$$

Furthermore, for the initial period $t = 1$ the prior distributions are initialised as

$$\lambda_{k,1} \mid \zeta^2 \sim \mathrm{N}(0, \zeta^2). \tag{8}$$

The formulation is completed by assuming vague inverse-Gamma prior distributions for both hyperparameters ψ^2 and ζ^2 in (7) and (8), respectively. Finally, to ensure identifiability, a zero-sum constraint (Baio and Blangiardo 2010) on the random effects within each period is required

$$\sum_{k=1}^{N_T} \lambda_{k,t} = 0, \quad t = 1, \ldots, \mathcal{T}. \tag{9}$$

The stochastic innovation variance (SIV) model proposed by Glickman (2001) is an extension of the model with constant innovation variance (CIV) (Fahrmeir and Tutz 1994; Glickman 1999; Knorr-Held 2000; Ingram 2021; Glickman 2025), where $\sigma^2_{k,t} = \sigma^2$ for $k = 1, \ldots, N_T$ and $t = 1, \ldots \mathcal{T}$. The state-space modelling strategy with a CIV can be too restrictive because it does not account for the possibility of sudden changes in the ability of a team. Conversely, the model proposed by Glickman (2001) allows the innovation variance $\sigma^2_{k,t}$ to vary stochastically over time for each team, as described above. This increased flexibility implies that the model can capture unexpected performance changes or periods of increased volatility for specific teams when needed (e.g., a generally strong team undergoing a rebuilding period might get a larger $\sigma^2_{k,t}$ for that period).

3 A commensurate spike-and-slab proposal

Although flexible, the state-space model of Glickman (2001) may still suffer from limitations. It may under-borrow information when a team's performance is stable across consecutive periods, or over-borrow information following a substantial change in performance.[1] Furthermore, the vague inverse-Gamma prior on the hyperparameter ϕ^2 in (7) is not well suited for shrinkage, since it is bounded away from zero and cannot induce strong shrinkage (Frühwirth-Schnatter and Wagner 2010). This can lead to insufficient borrowing of information from the previous innovation variance in (7), which consequently affects the degree of information borrowed from the previous log-strength in (6).

[1] Here, under-borrowing refers to the situation where a team's performance is stable across consecutive periods, but the model's prior does not shrink strongly enough toward the previous state, preventing effective pooling of past information. Conversely, over-borrowing occurs when, after a genuine structural change in performance, the model does not adapt quickly enough, causing excessive reliance on outdated information.

The literature on dynamic shrinkage priors has grown substantially to address such limitations. Spike-and-slab priors (Mitchell and Beauchamp 1988) are finite mixture distributions with two components, where one component induces strong global shrinkage (the "spike"), while the other allows greater flexibility (the "slab"). Originally developed for variable selection in regression models, these mixture shrinkage priors were later extended to state-space modelling by Frühwirth-Schnatter and Wagner (2010) to shrink time-varying state variables toward fixed components. For time-varying parameter models specifically, Bitto and Frühwirth-Schnatter (2019) showed how shrinkage priors can automatically reduce time-varying parameters to static ones when warranted. Kowal et al. (2019) proposed dynamic shrinkage processes that model dependence among local scale parameters, inheriting horseshoe shrinkage behaviour while providing localised adaptivity. Cadonna et al. (2020) unified several shrinkage priors under the triple gamma framework, which encompasses the horseshoe as a special case and enables both variance shrinkage and variance selection in state-space models. Compared to continuous shrinkage priors (e.g., the inverse-Gamma prior), spike-and-slab priors offer greater flexibility by explicitly classifying time-varying coefficients – such as innovation variances – into dynamic, constant, or zero ones (Frühwirth-Schnatter and Knaus 2021).

To address the limitations of the SIV model, we propose a Bayesian dynamic BT model formulated as a sparse state-space model (Frühwirth-Schnatter and Knaus 2021) that incorporates team- and time-specific spike-and-slab commensurate priors (Hobbs et al. 2012, 2013) on the innovation precision of each team's log-strength. This approach, subsequently adapted to dynamic offensive and defensive abilities in football goal-based models by Macrì-Demartino et al. (2025), provides a formal mechanism for adaptively borrowing information from a team's previous performance only to the extent that the new data support it. This adaptive shrinkage mitigates the under- or over-borrowing issues of the SIV model by allowing data-driven calibration of how much past information is carried forward.

3.1 Proposed method

Commensurate priors (Hobbs et al. 2011, 2012) are a Bayesian approach developed to adaptively borrow information from related sources (e.g., historical data) by explicitly modelling their commensurability. The main assumption is that the current parameter of interest θ comes from a normal distribution centred on the corresponding parameter in the historical data θ_0, which is given by

$$\theta \mid \theta_0, \tau^2 \sim \mathrm{N}(\theta_0, \tau^{-2}),$$

where τ^2 is the precision (commensurability) parameter. When current and historical data are consistent, the posterior for τ^2 favours large values (small variance), encouraging strong borrowing. Conversely, substantial discrepancies lead to a posterior favouring smaller τ^2 (large variance), down-weighting the historical data's influence by treating them as incommensurate.

In our dynamic BT model, we extend the commensurate prior idea to time-varying team strengths. The performance of each team in the previous period serves as historical information for the next period. Specifically, for each team T_k, with $k = 1, \ldots, N_T$, and each period $t = 2, \ldots, \mathcal{T}$, we treated each log-strength parameter $\lambda_{k,t}$ as a latent state that follows a random walk with team- and time-specific innovation precisions (commensurability parameters) $\tau^2_{k,t} = 1/\sigma^2_{k,t}$, assuming the following state transition model for the log-strengths

$$\lambda_{k,t} \mid \lambda_{k,t-1}, \tau^2_{k,t} \sim \mathrm{N}\left(\lambda_{k,t-1}, \tau^{-2}_{k,t}\right), \tag{10}$$

where the innovation precisions $\tau^2_{k,t}$ are modelled as a continuous two-component mixture of a concentrated "spike" and a diffuse "slab" component (Hong et al. 2018; Macrì-Demartino et al. 2025), such that

$$\tau^2_{k,t} \mid \mu_s, \mu_l, \psi^2_s, \psi^2_l, p_s \sim \mathrm{TN}(\mu_s, \psi^2_s) \times p_s + \mathrm{TN}(\mu_l, \psi^2_l) \times (1 - p_s), \tag{11}$$

where p_s is the prior probability of the "spike" component, and $\mathrm{TN}(\mu, \psi^2)$ denotes a normal distribution with mean μ and variance ψ^2 truncated from below at zero (e.g., the half-normal distribution). Here μ_s and μ_l are the means of the spike-and-slab components, respectively, and ψ^2_s and ψ^2_l are their variances, with $0 < \psi^2_s < \psi^2_l$.

For each $t \geq 2$, the proposed model adaptively determines $\tau^2_{k,t}$ based on the degree to which the new results align with the performance of the previous period. If a team's performance in period t is consistent with its performance in period $t-1$, the posterior for $\tau^2_{k,t}$ will favour the spike component, implying strong borrowing (i.e., shrinkage toward past performance). Conversely, if the performance of a team changes sharply from $t-1$ to t, the posterior will favour the slab component, allowing the current data to dominate and resulting in weak borrowing from the previous period. For the initial period $t = 1$, no past information is available, so we assign diffuse but proper prior distributions. To ensure identifiability, we impose a zero-sum constraint on the log-strengths in each period, as specified in (9).

In summary, the full hierarchical specification of the proposed model is

$$\begin{aligned}
\text{Likelihood:} \quad & Y_{i,j,t} \mid p_{i,j,t} \sim \mathrm{Bernoulli}(p_{i,j,t}), \\
& p_{i,j,t} = \mathrm{logit}^{-1}(\lambda_{i,t} - \lambda_{j,t} + \phi), \\
\text{State evolution:} \quad & \lambda_{k,t} \mid \lambda_{k,t-1}, \tau^2_{k,t} \sim \mathrm{N}(\lambda_{k,t-1}, \tau^{-2}_{k,t}), \quad t = 2, \ldots, \mathcal{T}, \\
\text{Spike-and-slab prior:} \quad & \tau^2_{k,t} \mid \mu_s, \mu_l, \psi^2_s, \psi^2_l, p_s \sim \mathrm{TN}(\mu_s, \psi^2_s) \times p_s \\
& \quad + \mathrm{TN}(\mu_l, \psi^2_l) \times (1 - p_s), \\
\text{Home advantage prior:} \quad & \phi \mid \xi^2 \sim \mathrm{N}(0, \xi^2), \\
\text{Identifiability constraint:} \quad & \sum_{k=1}^{N_T} \lambda_{k,t} = 0, \quad t = 1, \ldots, \mathcal{T}.
\end{aligned}$$

4 Application on the NBA data

We evaluate the proposed model using data from the ten most recent NBA seasons (2015/2016 to 2024/2025). The dataset comprises 12,841 games across the regular season, NBA Cup, play-in tournament, and playoffs, with each season treated as a distinct time period.

The models are implemented in the probabilistic programming language Stan (Carpenter et al. 2017) using the `cmdstanr` R package (Gabry et al. 2024), which employs the Hamiltonian Monte Carlo (HMC) algorithm with the No-U-Turn sampler (Hoffman et al. 2014). For posterior sampling, we run four independent chains of 2000 iterations each, discarding the first 1000 iterations as burn-in. In the spike-and-slab specification given in (11), the spike component has mean $\mu_s = 30$ and standard deviation $\psi_s = 0.1$, while the slab component has mean $\mu_l = 0$ and standard deviation $\psi_l = 5$. That is

$$\tau^2_{k,t} \mid p_s \sim \mathrm{TN}(30, 0.1^2) \times p_s + \mathrm{TN}(0, 5^2) \times (1 - p_s),$$

where p_s is fixed at 0.05. The "spike" prior ensures strong pooling when past and current performances are consistent (Ouma et al. 2022; Zheng and Wason 2022; Chen et al. 2018; Hobbs et al. 2012). Conversely, the "slab" prior is approximately uniform in $[0, 3]$ and then decays, allowing minimal borrowing or even complete discounting of prior information when performance shifts occur (Alt et al. 2025). A comprehensive sensitivity analysis examining the robustness of our results to these hyperparameter choices is provided in Appendix A. Finally, we include an order effect for home advantage, modelled with a weakly informative prior (Gelman et al. 2008) as follows

$$\phi_{\text{home}} \sim \mathrm{N}(0, 10^2).$$

4.1 Parameter estimates

We now examine the posterior smoothed estimates of the model parameters, obtained by conditioning on the full dataset across all ten seasons, to assess whether the proposed approach effectively captures the dynamics of NBA team performance. Figure 1 shows the posterior mean log-strengths for each NBA team over the ten seasons, with 95% credible intervals. The model successfully identifies several shifts in team performance. For instance, the Golden State Warriors' period of dominance was followed by a pronounced decline in the 2019/2020 season. This drop, captured clearly by the model, coincides with the departure of key players (Kevin Durant, Andre Iguodala) and major injuries to Stephen Curry and Klay Thompson, which resulted in a 15th-place finish in the Western Conference. A similar trend is observed for the Cleveland Cavaliers, whose decline in 2018/2019 corresponds to LeBron James' move to the Los Angeles Lakers – whose subsequent peak in log-strength during the 2019/2020 season corresponds to their NBA championship win. This shows that the proposed dynamic BT effectively adapts to real shifts in ability and avoids simply remembering past strengths when teams have changed significantly.

Interestingly, many teams that experience sharp declines or rapid improvements in performance tend to revert toward their previous levels within one or two seasons. This behaviour is evident for the Golden State Warriors, who recovered substantially following their 2019/2020 collapse, as well as for the Cleveland Cavaliers and Los Angeles Lakers. This trend may partly reflect structural features of the NBA, most notably the draft system, which awards higher draft picks to weaker-performing teams and thus facilitates the acquisition of elite talent. In addition, roster adjustments, coaching changes, and player recovery from injury likely contribute to this trend.

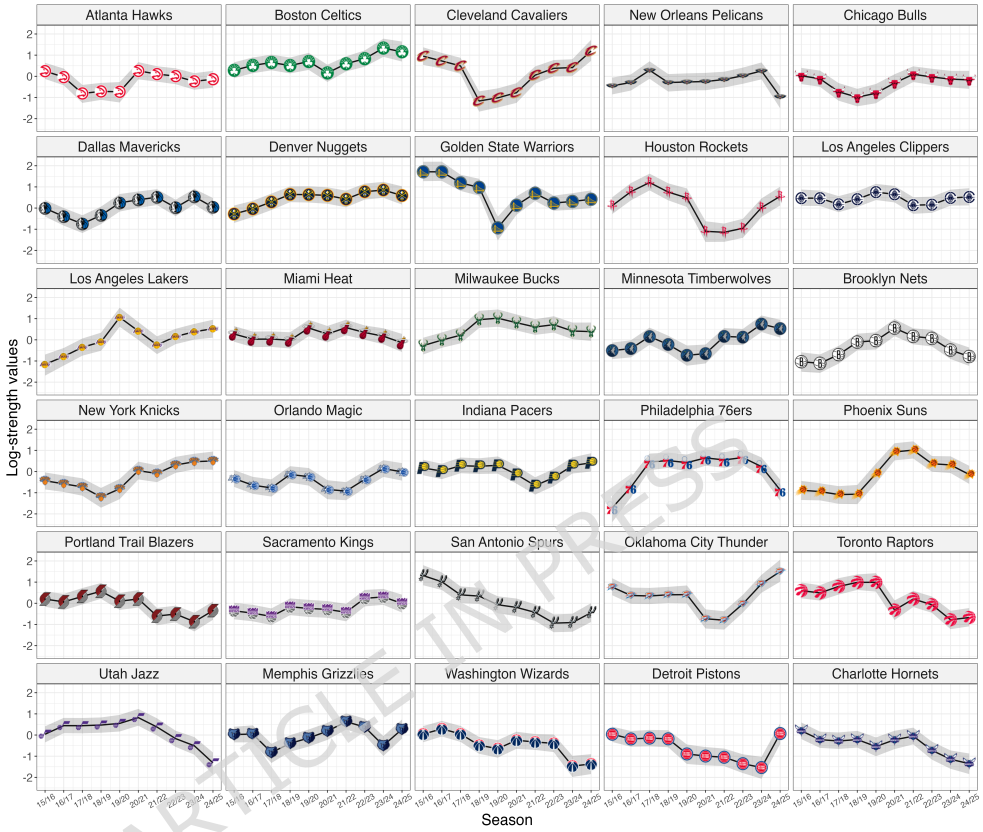

Fig. 1 Trajectories of posterior mean log-strengths (solid lines) with their 95% credible intervals (grey ribbons) over ten seasons for each NBA team.

Figure 2 shows the joint posterior distribution of the log-strength parameter $\lambda_{k,t}$ and the innovation standard deviation $\sigma_{k,t}$ for the Oklahoma City Thunder and the Philadelphia 76ers, comparing the SIV model with our proposal. During periods of stable performance – such as for the Oklahoma City Thunder from 2016/2017 to 2019/2020 or the Philadelphia 76ers from 2018/2019 to 2022/2023 – our model produces more sharply peaked and concentrated joint posterior distributions. Conversely, following a sudden performance shift – such that of the Oklahoma City Thunder in

2020/2021 or the Philadelphia 76ers in 2024/2025 – our model responds by inflating the innovation standard deviation. This yields a more diffuse joint posterior, reflecting greater uncertainty about the team's new strength level. This behaviour demonstrates the model's ability to dynamically down-weight historical information when new data indicate a sudden change in performances.

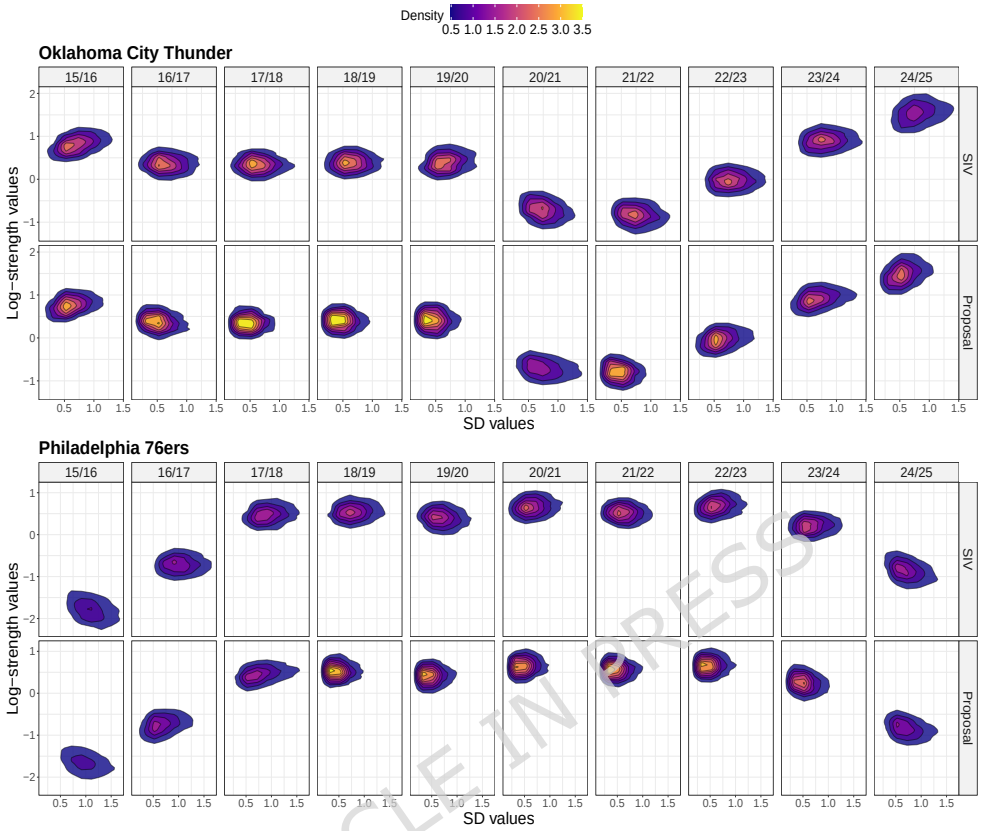

Fig. 2 Joint posterior of team log-strength $\lambda_{k,t}$ and innovation standard deviation (SD) $\sigma_{k,t}$ for the Oklahoma City Thunder and the Philadelphia 76ers across ten NBA seasons. Each panel compares the stochastic innovation variance (SIV) model with our commensurate spike-and-slab proposal.

Table 1 further quantifies the adaptive behaviour of the proposed model relative to the SIV specification. During periods of stable performance, the proposed approach produces substantially smaller innovation standard deviations (negative $\Delta\sigma$), indicating a stronger borrowing from past strength estimates. The largest reductions occur during post-transition stabilisation phases. For instance, the Oklahoma City Thunder's 2021/22 season exhibits $\Delta\sigma = -30.9\%$, with the innovation standard deviation decreasing from 0.826 to 0.571, while the Philadelphia 76ers' stable 2018/19 season shows $\Delta\sigma = -31.9\%$, corresponding to a reduction in σ from 0.861 to 0.586. Conversely, following major performance disruptions, the model appropriately increases

the innovation standard deviation (positive $\Delta\sigma$), thus discounting historical information. The Oklahoma City Thunder's 2020/21 collapse – during which log-strength decreased from 0.411 to −0.733 – is associated with $\Delta\sigma = +23.0\%$, as the innovation standard deviation increases from 0.813 to 1.001. Similarly, the Philadelphia 76ers' 2024/25 decline, where λ falls from 0.233 to −0.870, results in $\Delta\sigma = +16.8\%$, with the innovation standard deviation increasing from 0.827 to 0.966. These results demonstrate that the spike-and-slab mechanism effectively discriminates between stable and transitional periods, yielding tighter posterior inference when appropriate while remaining responsive to genuine structural changes in team performance.

4.2 Model checking

Once the model has been estimated, the next step is to evaluate its goodness of fit. Posterior predictive checks (PPCs) (Rubin 1984; Gelman et al. 1996, 2013) are the main tool used to assess whether a Bayesian model can generate replications that resemble the observed data. In PPCs, the observed data distribution is compared with the posterior predictive distribution. If these distributions differ substantially, the model is likely misspecified and should be refined. Specifically, hypothetical replications y^{rep} are generated from the posterior predictive distribution

$$f(y_{\text{rep}} \mid y) = \int f(y_{\text{rep}} \mid \theta)\pi(\theta \mid y)d\theta, \tag{12}$$

where $\pi(\theta \mid y)$ is the posterior distribution and $f(y_{\text{rep}} \mid \theta)$ is the likelihood function for hypothetical replicated values. Since this distribution is rarely analytically tractable, prediction typically requires a two-step simulation at each MCMC iteration: first sample $\theta^{(i)}$ from $\pi(\theta \mid y)$, and then generate $y_{\text{rep}}^{(i)}$ from $f(y_{\text{rep}} \mid \theta^{(i)})$. Figure 3 presents posterior predictive checks of home and away win proportions for each NBA team. The observed proportions, shown as horizontal lines, are compared with posterior predictive medians and 95% credible intervals, displayed as dots and whiskers. There is clear agreement between the replicated and observed distributions. For every team, the observed proportions of home and away wins fall within the 95% credible intervals of the corresponding posterior predictive distribution. This concordance indicates no evidence of model misspecification.

When the model fits well, summary statistics (e.g., the mean or standard deviation) calculated from the replicated data should align with those from the observed dataset. This agreement can be formally evaluated using posterior predictive p-values (PPPs) (Guttman 1967; Rubin 1984; Meng 1994; Gelman et al. 1996), defined as the probability that the test statistic from replicated data is greater than or equal to that of the observed data

$$\text{PPP}(y, T) = \mathbb{P}\left\{T(y_{\text{rep}}) \geq T(y) \mid y\right\},$$

where $T(\cdot)$ is the chosen test statistic. Extreme PPP values close to 0 or 1 suggest that the observed statistic lies in the tail of the predictive distribution, denoting poor model fit. However, PPPs are not classically calibrated and generally do not follow a

Table 1 Posterior means and 95% credible intervals (CI) for log-strength λ and innovation standard deviation σ comparing the stochastic innovation variance (SIV) model and the proposed commensurate spike-and-slab model for Oklahoma City Thunder (OKC) and Philadelphia 76ers (PHI). $\Delta\sigma$ is the percent change in mean.

Team	Season	Model	λ Mean [95% CI]	σ Mean [95% CI]	$\Delta\sigma$ (%)
OKC	15/16	SIV	0.817 [0.406, 1.240]	0.783 [0.333, 1.609]	−4.2
		Proposal	0.771 [0.379, 1.182]	0.750 [0.329, 1.915]	
	16/17	SIV	0.357 [−0.057, 0.759]	0.734 [0.307, 1.498]	−15.3
		Proposal	0.369 [−0.023, 0.743]	0.622 [0.302, 1.642]	
	17/18	SIV	0.339 [−0.066, 0.730]	0.698 [0.289, 1.405]	−20.4
		Proposal	0.347 [−0.035, 0.726]	0.556 [0.292, 1.307]	
	18/19	SIV	0.389 [−0.007, 0.803]	0.699 [0.291, 1.408]	−16.1
		Proposal	0.399 [0.018, 0.775]	0.587 [0.281, 1.503]	
	19/20	SIV	0.396 [−0.028, 0.831]	0.737 [0.330, 1.449]	−23.6
		Proposal	0.411 [−0.005, 0.842]	0.563 [0.292, 1.337]	
	20/21	SIV	−0.722 [−1.244, −0.241]	0.813 [0.388, 1.548]	+23.0
		Proposal	−0.733 [−1.191, −0.288]	1.001 [0.382, 2.687]	
	21/22	SIV	−0.844 [−1.322, −0.405]	0.826 [0.393, 1.583]	−30.9
		Proposal	−0.799 [−1.242, −0.375]	0.571 [0.290, 1.422]	
	22/23	SIV	−0.023 [−0.418, 0.391]	0.875 [0.416, 1.708]	−12.8
		Proposal	−0.029 [−0.439, 0.397]	0.763 [0.329, 2.062]	
	23/24	SIV	0.902 [0.490, 1.322]	0.916 [0.430, 1.845]	−4.4
		Proposal	0.920 [0.500, 1.343]	0.876 [0.346, 2.517]	
	24/25	SIV	1.558 [1.042, 2.144]	0.942 [0.394, 2.060]	−25.2
		Proposal	1.509 [1.021, 2.060]	0.705 [0.310, 1.923]	
PHI	15/16	SIV	−1.826 [−2.484, −1.231]	1.247 [0.637, 2.362]	+10.6
		Proposal	−1.724 [−2.389, −1.146]	1.379 [0.528, 3.583]	
	16/17	SIV	−0.719 [−1.161, −0.293]	1.134 [0.576, 2.126]	−19.8
		Proposal	−0.738 [−1.220, −0.282]	0.910 [0.337, 2.674]	
	17/18	SIV	0.474 [0.070, 0.893]	1.012 [0.508, 1.918]	+1.00
		Proposal	0.457 [0.070, 0.847]	1.022 [0.396, 2.787]	
	18/19	SIV	0.539 [0.147, 0.941]	0.861 [0.382, 1.693]	−31.9
		Proposal	0.523 [0.172, 0.893]	0.586 [0.292, 1.505]	
	19/20	SIV	0.396 [−0.028, 0.803]	0.767 [0.321, 1.562]	−25.0
		Proposal	0.414 [0.013, 0.804]	0.575 [0.287, 1.407]	
	20/21	SIV	0.660 [0.276, 1.065]	0.717 [0.288, 1.485]	−18.4
		Proposal	0.649 [0.274, 1.049]	0.585 [0.297, 1.441]	
	21/22	SIV	0.518 [0.149, 0.893]	0.701 [0.274, 1.492]	−19.5
		Proposal	0.552 [0.179, 0.924]	0.564 [0.287, 1.303]	
	22/23	SIV	0.689 [0.295, 1.107]	0.712 [0.288, 1.525]	−18.7
		Proposal	0.665 [0.305, 1.048]	0.579 [0.288, 1.485]	
	23/24	SIV	0.215 [−0.189, 0.621]	0.762 [0.315, 1.673]	−16.4
		Proposal	0.233 [−0.193, 0.666]	0.637 [0.300, 1.668]	
	24/25	SIV	−0.865 [−1.368, −0.377]	0.827 [0.338, 1.835]	+16.8
		Proposal	−0.870 [−1.380, −0.388]	0.966 [0.366, 2.699]	

uniform distribution, even when the model is correctly specified (Bayarri and Berger 2000). They are therefore used as diagnostic summaries rather than formal hypothesis tests (Gelman et al. 2013). Figure 4 presents the posterior predictive distribution of the mean – that is the home win probability. The observed mean (blue line) lies well

Fig. 3 Posterior predictive check of home and away win proportions by team. For each team, horizontal lines show the observed proportions (y), while dots and whiskers represent posterior predictive medians and 95% credible intervals (y_{rep}).

within the predictive distribution, and the PPP value of 0.501 further supports the adequacy of the model.

4.3 In-sample predictions

To evaluate the in-sample predictive accuracy of our model, we reconstruct the league standings in terms of won matches for each season using the posterior predictive distribution of the model, as in (12). Specifically, for each iteration of the MCMC sampling, we draw values from the model's sampling likelihood based on the parameter values generated at that iteration. This process yields a posterior predictive distribution of match outcomes. Figure 5 displays the 95% predictive intervals (grey ribbon) for the number of wins by each NBA team for the 2024/2025 season, along with the expected wins from the reconstruction of the sample (red dots) and the observed wins (black diamonds). The results show a strong agreement between the predicted and expected wins, with all observed values falling within the 95% predictive intervals.

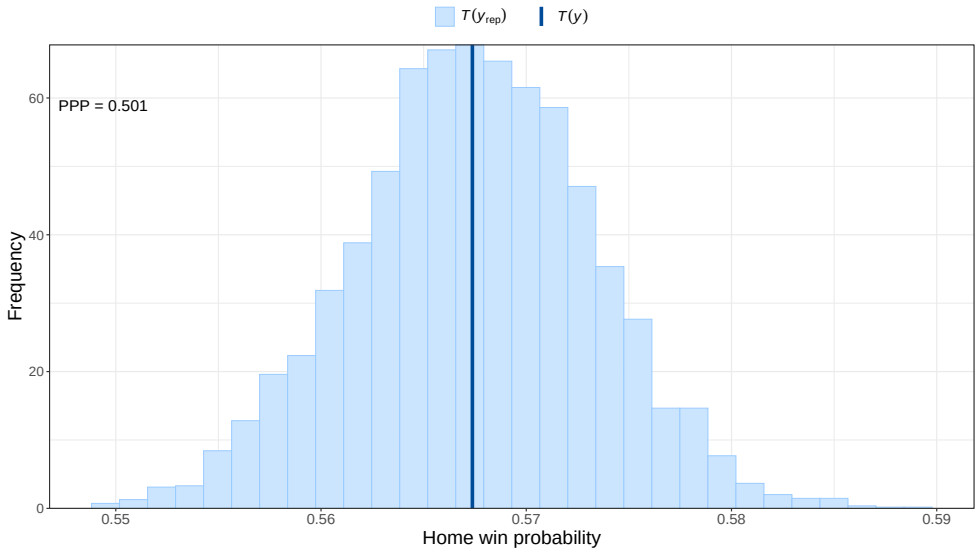

Fig. 4 Posterior predictive check for the global mean (home–win probability). The histogram shows the distribution of the mean computed from replicated datasets. The vertical line denotes the observed mean and the posterior predictive p-value (PPP) is reported in the top-left corner.

Similar results are obtained for all previous NBA seasons. Further details are provided in Appendix B.

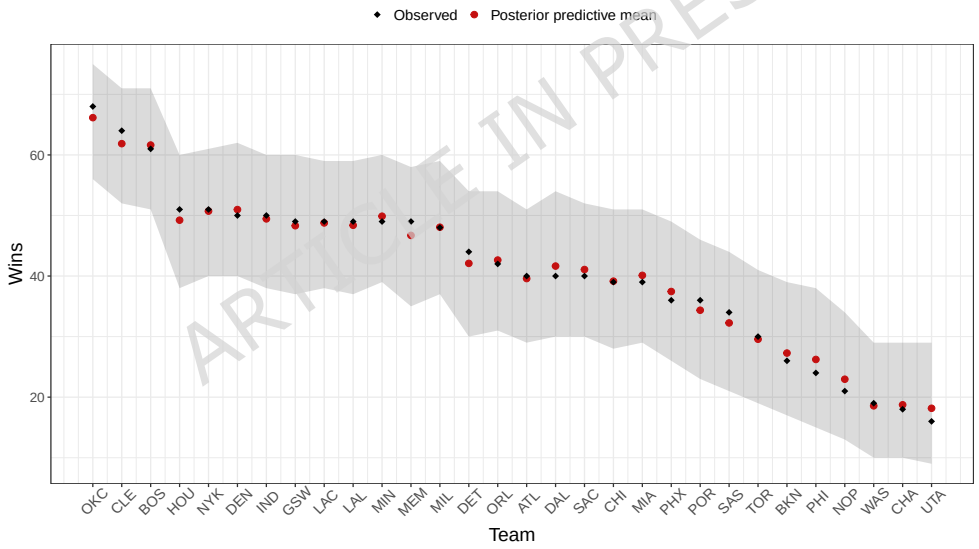

Fig. 5 In-sample predictions of NBA team wins for the 2024/2025 season. Grey ribbons represent the 95% predictive intervals, red dots denote expected wins from the in-sample reconstruction, and black diamonds indicate the observed wins.

4.4 Predictive performance comparison

This section evaluates the predictive performance of the proposed model compared to the CIV model (Fahrmeir and Tutz 1994; Glickman 1999; Knorr-Held 2000; Ingram 2021; Glickman 2025) and the SIV model proposed by Glickman (2001). We assess predictive accuracy using two approaches: cross-validated predictive fit on the training data and genuine out-of-sample predictions on held-out playoff games.

4.4.1 Cross-validated predictive fit

While Section 4.3 assessed the model's ability to reconstruct observed match outcomes, we now evaluate predictive accuracy using leave-one-out cross-validation (LOO) (Vehtari et al. 2017). LOO approximates the predictive performance that would be obtained by refitting the model with each observation held out in turn, providing a measure of internal predictive accuracy that accounts for model complexity. It provides a practical and stable alternative to exact cross-validation and is particularly well-suited for Bayesian models fitted via MCMC. The Bayesian LOO estimate, computed using the log-likelihood evaluated at the posterior simulations of the parameter values, is summarised by the LOO information criterion (LOOIC)

$$\text{LOOIC} = -2 \sum_{n=1}^{N} \log f(y_n \mid y_{-n}), \tag{13}$$

where $f(y_n \mid y_{-n})$ is the leave-one-out predictive density obtained by fitting the model without the n-th observation. Smaller LOOIC values – which are equivalent to larger sums of log predictive densities – indicate better predictive performance, analogous to information criteria such as AIC, DIC, and WAIC. In practice, we compute approximate LOO using the `loo` package (Vehtari et al. 2024), which implements Pareto smoothed importance sampling (PSIS) (Vehtari et al. 2024).

Table 2 presents the LOOIC values for all models across the different stages of the 2024/2025 NBA season. The proposed model yields consistently smaller LOOIC values compared to both the SIV and CIV models, indicating superior cross-validated predictive fit. As expected, LOOIC values increase as the season progresses, reflecting the growing number of games incorporated into the evaluation at each stage.

4.4.2 Out-of-sample predictions

While LOOIC provides a useful measure of cross-validated model fit, it is computed using the same data employed for model estimation. To evaluate genuine out-of-sample predictive performance, we therefore assessed the models on NBA games that were not included in the fitting process. The comparison focuses on five scenarios: the second half of the regular season, the first round, the conference semifinals, the conference finals, and the NBA Finals of the playoffs. After each stage, the observed outcomes are incorporated as new data before generating predictions for the subsequent scenario.

Bayesian models provide a natural framework for obtaining posterior probabilities of future outcomes through MCMC simulations from the posterior predictive distribution. Let $\tilde{y}$ denote a future observable value and θ the parameter of interest. Similar to

Table 2 Leave-one-out cross-validation information criterion (LOOIC) values under the proposed commensurate spike-and-slab model, the stochastic innovation variance (SIV) model, and the constant innovation variance (CIV) model.

Pred. Scenario	Proposal	SIV	CIV
Second Half	14949	14958	14956
First Round	15689	15698	15692
Conf. Semifinals	15740	15748	15744
Conf. Finals	15776	15784	15778
NBA Finals	15790	15798	15792

(12), the posterior predictive distribution is given by $f(\tilde{y} \mid y) = \int f(\tilde{y} \mid \theta)\pi(\theta \mid y)d\theta$, where $f(\tilde{y} \mid \theta)$ is the likelihood of a future observation and $\pi(\theta \mid y)$ is the posterior distribution of θ. Table 3 presents, for each match in the conference semifinals, conference finals, and NBA Finals, the posterior probability assigned to the actual outcome (home or away win) by our proposed BT model, the SIV model, and the CIV model. Across stages, our proposed model more frequently assigns higher probabilities to the true outcomes in the conference semifinals compared to the two alternative models. In the conference finals, our model assigns the highest probability to the realized outcome in most matches, while the NBA finals show mixed results across models.

To provide a more systematic comparison, we evaluate out-of-sample predictive accuracy using the Brier score (Brier 1950), a proper scoring rule recommended by Spiegelhalter and Ng (2009). The Brier score is defined as the mean squared error of the predicted probabilities:

$$\text{Brier Score} = \frac{1}{M}\sum_{m=1}^{M}\sum_{r=1}^{2}(p_{r,m} - \delta_{r,m})^2,$$

where $p_{r,m}$ denotes the predicted probability of outcome $r \in \{\text{home win, away win}\}$ for the m-th match to predict, and $\delta_{r,m}$ is the Kronecker delta, equal to 1 if outcome r occurs in the m-th match. The Brier score is bounded between 0 and 2, where 0 represents a perfect forecast and 2 indicates the worst possible forecast. In this formulation, a score of 0.5 corresponds to an uninformative forecast that assigns equal probability to both outcomes, providing a natural benchmark to assess model predictive performance. Table 4 reports the Brier scores for the three models across the various predictive scenarios considered. In all scenarios, the proposed dynamic Bradley–Terry (BT) model consistently achieves lower Brier scores than both the SIV and CIV specifications, indicating modest yet systematic improvements in predictive accuracy. In the second half and first-round scenarios of the 2024/2025 NBA playoffs, the proposed model shows modest improvements, with Brier scores of 0.423 and 0.387, respectively. More pronounced performance gains are observed in the conference semifinals and finals, where the proposed model achieves Brier scores of 0.603 and 0.420, compared to 0.637 and 0.437 for the SIV model and 0.609 and 0.426 for the CIV model. In the

Table 3 Posterior probabilities assigned to the realized match outcomes (home win, HW, or away win, AW) in the conference semifinals, conference finals, and NBA finals under the proposed commensurate spike-and-slab model, the stochastic innovation variance (SIV) model, and the constant innovation variance (CIV) model.

Pred. Scenario	Home Team	Away Team	Outcome	Proposal	SIV	CIV
Conf. Semifinals	CLE	IND	AW	0.294	0.236	0.277
	BOS	NYK	AW	0.295	0.297	0.296
	OKC	DEN	AW	0.206	0.171	0.221
	CLE	IND	AW	0.275	0.234	0.273
	MIN	GSW	AW	0.349	0.385	0.360
	BOS	NYK	AW	0.288	0.313	0.292
	OKC	DEN	HW	0.784	0.825	0.779
	MIN	GSW	HW	0.664	0.618	0.625
	IND	CLE	AW	0.592	0.668	0.600
	DEN	OKC	HW	0.309	0.278	0.332
	NYK	BOS	AW	0.563	0.556	0.574
	GSW	MIN	AW	0.512	0.484	0.491
	DEN	OKC	AW	0.684	0.707	0.682
	IND	CLE	HW	0.402	0.344	0.414
	NYK	BOS	HW	0.436	0.429	0.424
	GSW	MIN	AW	0.493	0.487	0.484
	CLE	IND	AW	0.283	0.228	0.267
	OKC	DEN	HW	0.791	0.817	0.761
	BOS	NYK	HW	0.702	0.706	0.699
	MIN	GSW	HW	0.658	0.616	0.621
	DEN	OKC	HW	0.316	0.276	0.332
	NYK	BOS	HW	0.434	0.428	0.422
	OKC	DEN	HW	0.799	0.816	0.781
Conf. Finals	OKC	MIN	HW	0.627	0.608	0.625
	NYK	IND	AW	0.533	0.502	0.499
	OKC	MIN	HW	0.624	0.601	0.631
	NYK	IND	AW	0.544	0.496	0.490
	MIN	OKC	HW	0.496	0.545	0.493
	IND	NYK	AW	0.341	0.366	0.378
	MIN	OKC	AW	0.487	0.477	0.495
	IND	NYK	HW	0.661	0.619	0.626
	OKC	MIN	HW	0.627	0.594	0.643
	NYK	IND	HW	0.473	0.501	0.500
	IND	NYK	HW	0.650	0.629	0.625
NBA Finals	OKC	IND	AW	0.405	0.345	0.350
	OKC	IND	HW	0.615	0.668	0.659
	IND	OKC	HW	0.510	0.452	0.481
	IND	OKC	AW	0.482	0.543	0.514
	OKC	IND	HW	0.610	0.662	0.641
	IND	OKC	HW	0.517	0.449	0.474
	OKC	IND	HW	0.614	0.666	0.632

NBA finals scenario, the proposed model achieves a Brier score of 0.441, outperforming the SIV and CIV models, which achieve scores of 0.451 and 0.452, respectively. An additional comparison with a time-weighted BT model is reported in Appendix C.

Table 4 Brier scores for successive predictive scenarios under the proposed commensurate spike-and-slab model, the stochastic innovation variance (SIV) model, and the constant innovation variance (CIV) model.

Pred. Scenario	Proposal	SIV	CIV
Second Half	0.423	0.427	0.425
First Round	0.387	0.393	0.389
Conf. Semifinals	0.603	0.637	0.609
Conf. Finals	0.420	0.437	0.426
NBA Finals	0.441	0.451	0.452

5 Discussion

In this paper we introduce a novel Bayesian dynamic BT model in which the latent log-strength of each team evolves according to a team- and time-specific innovation precision. This precision is modelled using a spike-and-slab commensurate prior, allowing the model to adaptively borrow information from a team's past performance only when justified by the data. In practice, when a team's recent performances are consistent with its historical results, the model applies strong shrinkage toward the previous strength level (the "spike"). Conversely, when a team's performance changes suddenly, the model shifts to the more diffuse "slab," focussing on the information from the most recent observed data.

Posterior predictive checks confirm that the model provides a good fit. The results of the replicated matches closely align with the observed home and away wins frequencies for each team, and a key summary statistic – the overall probability of home win – falls well within the posterior predictive distribution. Furthermore, out-of-sample predictions underscore the practical effectiveness of our approach. In the NBA playoff scenarios examined, our model consistently assigned higher probabilities to actual winners compared to both the SIV and CIV models, particularly in the later stages of the playoffs. This improvement is quantitatively reflected in consistently lower Brier scores and LOOIC values across all predictive scenarios. Thus, the commensurate spike-and-slab formulation yields substantial predictive benefits without compromising the model's ability to accurately reconstruct league standings. Additionally, our method demonstrates a computational advantage over the SIV model proposed by Glickman (2001).

Several limitations and potential extensions remain. In this paper, we fixed the hyperparameters of the spike-and-slab prior (e.g., the spike probability p_s). In practice, hierarchical extensions could be introduced to allow these hyperparameters to adapt more flexibly. Furthermore, the current model assumes binary outcomes and does not accommodate ties. Extending the framework to sports where draws are common – such as football or handball – using a model like the Bradley-Terry–Davidson (Davidson and Beaver 1977) is a promising direction for future work. Finally, incorporating additional covariates, such as team market values, injury reports, or in-game statistics, could further improve predictive performance.

Additionally, our model assumes a single global home advantage parameter, consistent with the approach of Glickman (2001). Similarly, Fahrmeir and Tutz (1994) and Harville (2003) accounted for home advantage by including a single common parameter for all teams in their BT models. Furthermore, Knorr-Held (2000) did not find substantial evidence of heterogeneity among Bundesliga teams, and Harville and Smith (1994) reported similar findings for college basketball. However, the empirical evidence on home advantage heterogeneity yields contrasting conclusions. An alternative specification would allow for team-specific home effects (i.e., ϕ_k for each team T_k). Notably, the results of Clarke and Norman (1995), Kuk (1995), and Glickman and Stern (1998) support team-specific home effects. Given our focus on the spike-and-slab mechanism for dynamic team strengths, we opted for parsimony in the home effect specification to avoid overparameterization. Extending the model to accommodate team-specific home advantages remains a natural direction for future work.

A further consideration concerns the choice of modelling binary win/loss outcomes rather than richer score-based data. In basketball, point differentials and margins of win are typically available and can provide additional information about the strength of the team beyond the binary result. Several modelling alternatives could incorporate this information, including models for point spreads (Stern 1991) or extensions of goal-based approaches such as the bivariate Poisson models commonly applied in football (Karlis and Ntzoufras 2003). Our choice of the BT framework was motivated by some main aspects. First, for playoff prediction, the primary quantity of interest is often which team advances rather than the margin of win. Second, the additional complexity introduced by team- and time-specific precision parameters benefits from a parsimonious likelihood to maintain computational tractability. However, extending the proposed adaptive borrowing framework to score-based models represents a promising direction for future research and may yield further predictive improvements.

Furthermore, our model operates at the season level, treating each NBA season as a discrete time period. As a result, all state transitions occur between consecutive seasons, representing approximately similar time gaps. Within this framework, the commensurate spike-and-slab prior on the innovation precision adaptively captures the degree of change occurring during each off-season: when a team's performance remains stable, the spike component induces strong shrinkage toward the previous season's strength, whereas substantial off-season changes (e.g., major roster moves or key injuries) lead the model to favour the slab component, allowing greater uncertainty. A finer-grained model operating at week level would enable analysis of within-season reactiveness – for instance, quantifying how many games are required for the model to detect a genuine performance shift following a mid-season injury. Such an extension would be particularly relevant for capturing within-season dynamics, where the time between consecutive games varies from days to weeks, and represents a promising direction for future research.

Software and Data Availability

All analyses were conducted in the R programming language version 4.4.3 (R Core Team 2025). The data and code to reproduce this manuscript is openly available at https://github.com/RoMaD-96/BayesDBTCSV.

Funding

This work has been supported by the project "SMARTsports: "Statistical Models and AlgoRiThms in sports. Applications in professional and amateur contexts, with able-bodied and disabled athletes", funded by the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2022 - grant n. 2022R74PLE (CUP J53D23003860006).

Appendix A Sensitivity analysis for spike-and-slab hyperparameters

We conducted a comprehensive sensitivity analysis by systematically varying the key hyperparameters of the spike-and-slab formulation in (11) over a predefined grid. The analysis considered the spike mean with values $\mu_s \in \{30, 50, 100, 150\}$, the slab standard deviation with values $\psi_l \in \{2, 3, 4, 5\}$, and the spike probability with values $p_s \in \{0.05, 0.10, 0.20, 0.30\}$. Taken together, these choices resulted in 64 distinct hyperparameter configurations. For each configuration, we re-estimated the full model using the same NBA dataset, which comprises 12,841 matches spanning ten seasons.

Figure A1 presents the posterior mean log-strength trajectories for all 30 NBA teams under each hyperparameter configuration. The results indicate a high degree of robustness, as the estimated team strength trajectories are nearly indistinguishable across all specifications. Additionally, Figure A2 provides an enlarged view for three selected teams whose trajectories reflect well-documented events discussed in the main text: the decline of the Golden State Warriors during the 2019–2020 season, the drop in performance of the Cleveland Cavaliers following LeBron James' departure, and the subsequent rise of the Los Angeles Lakers. In all three cases, the salient patterns identified in the main analysis are consistently recovered across all 64 configurations.

We also evaluated the sensitivity of the out-of-sample predictive performance. For each of the 64 hyperparameter configurations, we computed the Brier score on held-out games across all five predictive scenarios described in Section 4.4.2. Table A1 summarises the results showing that Brier scores are stable across configurations. Notably, in the second half scenario, the scores vary by at most 0.002 across all 64 configurations, while the first round and conference semifinals exhibit a range of 0.009 and 0.010, respectively.

These findings provide strong evidence that the main conclusions of our analysis are robust to reasonable variations in the spike-and-slab prior specification. The adaptive borrowing mechanism performs as intended across a wide range of hyperparameter values, appropriately distinguishing between periods of stable and changing team performance.

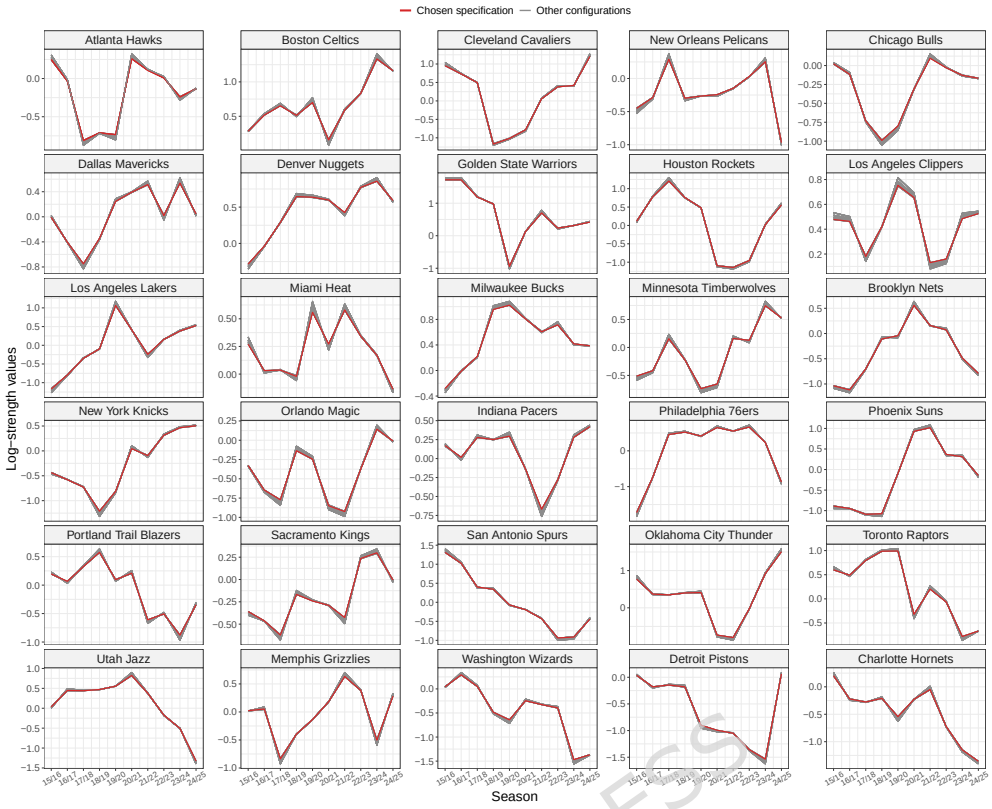

Fig. A1 Sensitivity of posterior mean log-strength trajectories to spike-and-slab hyperparameters. Each panel displays one NBA team's trajectory across ten seasons. Grey lines correspond to the 63 alternative hyperparameter configurations. The red line denotes the chosen specification ($\mu_s = 30$, $\psi_l = 5$, $p_s = 0.05$).

Appendix B In-sample predictions for previous seasons

Figure B3 displays the 95% predictive intervals (grey ribbons) for the number of wins by each NBA team across the 2015/2016 to 2023/2024 seasons with also the expected wins from the in-sample reconstruction (red dots) and the observed wins (black diamonds). Consistent with the findings of Section 4.3, the results indicate a strong agreement between predicted and expected wins, with all observed values lying within the 95% predictive intervals.

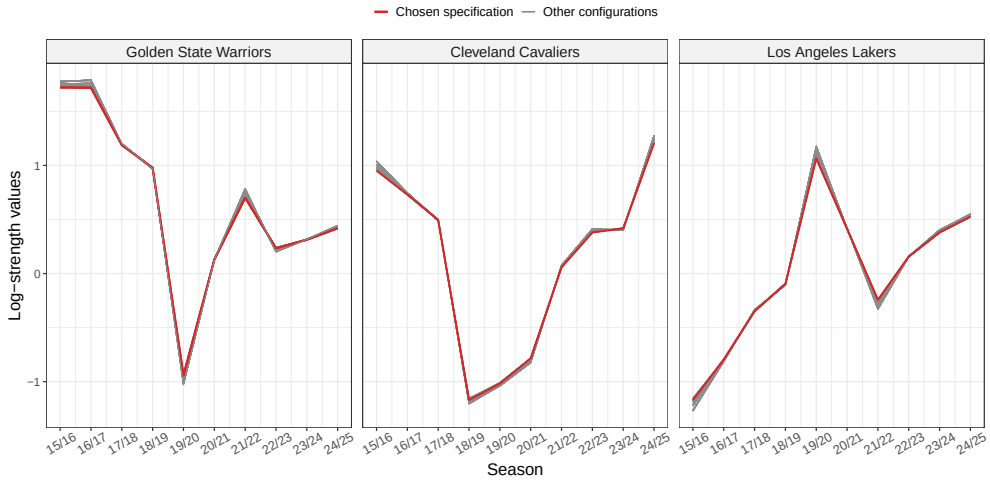

Fig. A2 Sensitivity of posterior mean log-strength trajectories for the Golden State Warriors, Cleveland Cavaliers, and Los Angeles Lakers. Grey lines correspond to the 63 alternative hyperparameter configurations. The red line denotes the chosen specification ($\mu_s = 30$, $\psi_l = 5$, $p_s = 0.05$).

Table A1 Summary of out-of-sample Brier scores for the proposed spike-and-slab model across all 64 hyperparameter configurations. For each predictive scenario, the table reports the minimum (Min), maximum (Max), mean, and standard deviation (SD) of the Brier scores.

Pred. Scenario	Min	Max	Mean	SD
Second Half	0.423	0.425	0.424	0.001
First Round	0.384	0.393	0.388	0.002
Conf. Semifinals	0.599	0.609	0.604	0.003
Conf. Finals	0.419	0.437	0.425	0.006
NBA Finals	0.432	0.458	0.448	0.007

Appendix C Comparison with time-weighted Bradley-Terry model

We compare our proposed model against a simpler time-weighted Bradley-Terry (TW-BT) variant. The TW-BT model provides a simpler baseline that does not require dynamic state-space inference. Unlike our proposed model where team strengths evolve according to (6), the TW-BT model estimates static team strengths but down-weights older observations exponentially. For predictions made at season t, each game g played in season t_g receives weight

$$w_g = \exp\{-\nu \times (t - t_g)\}, \tag{C1}$$

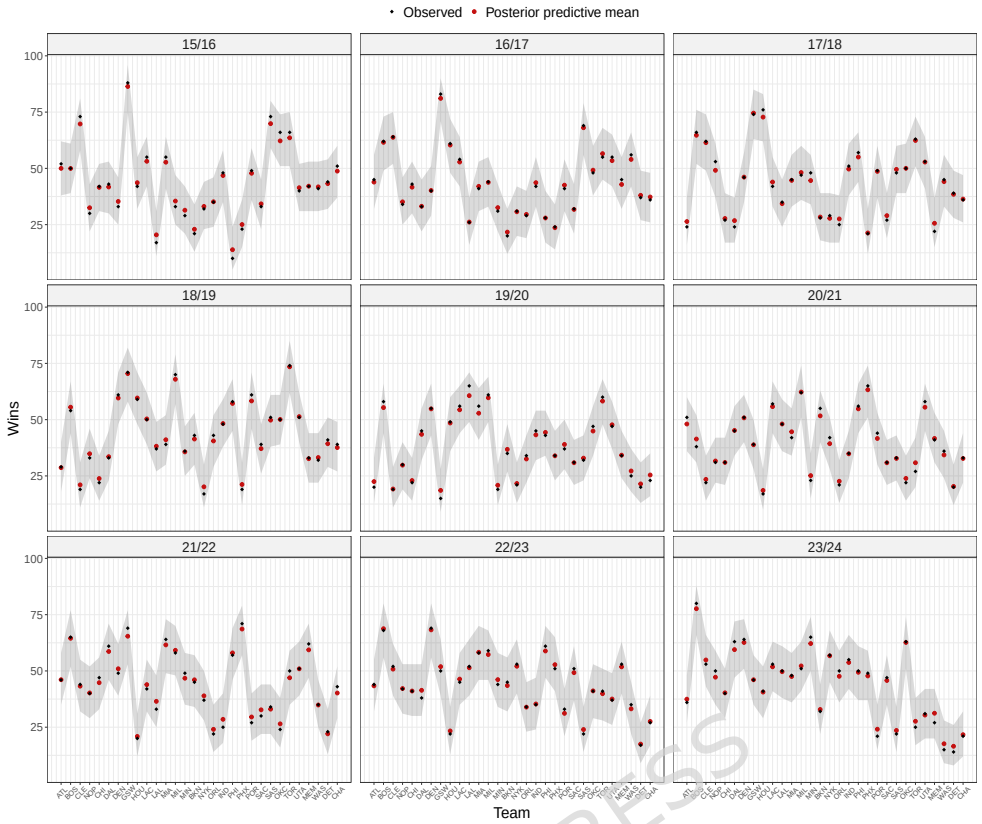

Fig. B3 In-sample predictions of NBA team wins for the 2015/2016–2023/2024 seasons. Grey ribbons represent the 95% predictive intervals, red dots denote expected wins from the in-sample reconstruction, and black diamonds indicate the observed wins.

where $\nu > 0$ controls the decay rate. Games from the reference season receive full weight ($w_g = 1$), while games from s seasons ago receive weight $\exp(-\nu \times s)$. An optional lookback parameter ρ excludes games more than ρ seasons old entirely.

Table C2 reports Brier scores for both the proposed model and the TW-BT baseline under the same predictive settings described in Section 4.4.2. The proposed model consistently yields lower Brier scores in all stages. The largest improvements occur in the conference finals, with scores of 0.420 versus 0.436, and in the NBA finals, with scores of 0.441 versus 0.450. These findings indicate that the adaptive borrowing mechanism induced by the spike and slab prior delivers meaningful predictive gains over simpler time weighting approaches

Table C2 Brier scores for successive predictive scenarios under the proposed commensurate spike-and-slab model and the time-weighted Bradley-Terry (TW-BT) model.

Pred. Scenario	Proposal	TW-BT
Second Half	0.423	0.428
First Round	0.387	0.392
Conf. Semifinals	0.603	0.604
Conf. Finals	0.420	0.436
NBA Finals	0.441	0.450

References

Alt, E.M., Chen, X., Carvalho, L.M., Ibrahim, J.G.: hdbayes: An R package for Bayesian analysis of generalized linear models using historical data. arXiv preprint arXiv:2506.20060 (2025)

Agresti, A.: Categorical Data Analysis, 2nd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, New Jersey (2002). https://doi.org/10.1002/0471249688

Bayarri, M., Berger, J.O.: P values for composite null models. Journal of the American Statistical Association **95**(452), 1127–1142 (2000)

Baio, G., Blangiardo, M.: Bayesian hierarchical model for the prediction of football results. Journal of Applied Statistics **37**(2), 253–264 (2010) https://doi.org/10.1080/02664760802684177

Bitto, A., Frühwirth-Schnatter, S.: Achieving shrinkage in a time-varying parameter model framework. Journal of Econometrics **210**(1), 75–97 (2019)

Beaver, R.J., Gokhale, D.V.: A model to incorporat within-pair order effects in paired comparisons. Communications in Statistics **4**(10), 923–939 (1975) https://doi.org/10.1080/03610927308827302

Bong, H., Li, W., Shrotriya, S., Rinaldo, A.: Nonparametric estimation in the dynamic bradley-terry model. In: International Conference on Artificial Intelligence and Statistics, pp. 3317–3326 (2020). PMLR

Baker, R.D., McHale, I.G.: Time varying ratings in association football: the all-time greatest team is.. Journal of the Royal Statistical Society Series A: Statistics in Society **178**(2), 481–492 (2014) https://doi.org/10.1111/rssa.12060 https://academic.oup.com/jrsssa/article-pdf/178/2/481/49338548/jrsssa_178_2_481.pdf

Brier, G.W.: Verification of forecasts expressed in terms of probability. Monthey Weather Review, **78**(1), 1–3 (1950)

Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)

Cattelan, M.: Models for Paired Comparison Data: A Review with Emphasis on Dependent Data. Statistical Science **27**(3), 412–433 (2012) https://doi.org/10.1214/12-STS396

Chen, N., Carlin, B.P., Hobbs, B.P.: Web-based statistical tools for the analysis and design of clinical trials that incorporate historical controls. Computational Statistics & Data Analysis **127**, 50–68 (2018)

Caron, F., Doucet, A.: Efficient Bayesian inference for generalized Bradley–Terry models. Journal of Computational and Graphical Statistics **21**(1), 174–196 (2012) https://doi.org/10.1080/10618600.2012.638220 https://doi.org/10.1080/10618600.2012.638220

Cadonna, A., Frühwirth-Schnatter, S., Knaus, P.: Triple the gamma—a unifying shrinkage prior for variance and variable selection in sparse state space and tvp models. Econometrics **8**(2), 20 (2020)

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of Statistical Software **76**(1), 1–32 (2017) https://doi.org/10.18637/jss.v076.i01

Clarke, S.R., Norman, J.M.: Home ground advantage of individual clubs in english soccer. Journal of the Royal Statistical Society: Series D (The Statistician) **44**(4), 509–521 (1995)

Chen, C., Smith, T.M.: A Bayes-type estimator for the Bradley-Terry model for paired comparison. Journal of Statistical Planning and Inference **10**(1), 9–14 (1984) https://doi.org/10.1016/0378-3758(84)90028-4

Cattelan, M., Varin, C., Firth, D.: Dynamic Bradley–Terry modelling of sports tournaments. Journal of the Royal Statistical Society Series C: Applied Statistics **62**(1), 135–150 (2012) https://doi.org/10.1111/j.1467-9876.2012.01046.x

Davidson, R.R.: On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association **65**(329), 317–328 (1970)

David, H.A.: The Method of Paired Comparisons. Griffin's statistical monograph. C. Griffin, London (1988). https://books.google.it/books?id=bB21VsB_GyYC

Davidson, R.R., Beaver, R.J.: On extending the Bradley-Terry model to incorporate within-pair order effects. Biometrics **33**(4), 693–702 (1977). Accessed 2024-04-20

Duffield, S., Power, S., Rimella, L.: A state-space perspective on modelling and inference for online skill rating. Journal of the Royal Statistical Society Series C: Applied Statistics **73**(5), 1262–1282 (2024)

Davidson, R.R., Solomon, D.L.: A Bayesian approach to paired comparison experimentation. Biometrika **60**(3), 477–487 (1973). Accessed 2024-04-20

Egidi, L., Macrì-Demartino, R., Palaskas., V.: footBayes: Fitting Bayesian and MLE Football Models. (2025). R package version 2.1.0. https://CRAN.R-project.org/package=footBayes

Frühwirth-Schnatter, S., Knaus, P.: Sparse Bayesian state-space and time-varying parameter models. In: Handbook of Bayesian Variable Selection, 1st edn., pp. 297–326. Chapman and Hall/CRC, New York (2021). https://doi.org/10.1201/9781003089018-13

Frühwirth-Schnatter, S., Wagner, H.: Stochastic model specification search for gaussian and partial non-gaussian state space models. Journal of Econometrics **154**(1), 85–100 (2010)

Fahrmeir, L., Tutz, G.: Dynamic stochastic models for time-dependent ordered paired comparison systems. Journal of the American Statistical Association **89**(428), 1438–1449 (1994). Accessed 2024-04-20

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. Chapman and Hall/CRC, New York (2013). https://doi.org/10.1201/b16018 . https://doi.org/10.1201/b16018

Glickman, M.E., Jones, A.C.: Models and rating systems for head-to-head competition. Annual Review of Statistics and Its Application **12** (2024)

Gelman, A., Jakulin, A., Pittau, M.G., Su, Y.-S.: A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics **2**(4), 1360–1383 (2008) https://doi.org/10.1214/08-AOAS191

Glickman, M.E.: Parameter estimation in large dynamic paired comparison experiments. Journal of the Royal Statistical Society. Series C (Applied Statistics) **48**(3), 377–394 (1999). Accessed 2024-04-20

Glickman, M.E.: Dynamic paired comparison models with stochastic variances. Journal of Applied Statistics **28**(6), 673–689 (2001) https://doi.org/10.1080/02664760120059219

Glickman, M.E.: Rating competitors in games with strength-dependent tie probabilities. arXiv preprint arXiv:2506.11354 (2025)

Gelman, A., Meng, X.-L., Stern, H.: Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, 733–760 (1996)

Glickman, M.E., Stern, H.S.: A state-space model for national football league scores. Journal of the American Statistical Association **93**(441), 25–35 (1998) https://doi.org/10.1080/01621459.1998.10474084 https://www.tandfonline.com/doi/pdf/10.1080/01621459.1998.10474084

Guttman, I.: The use of the concept of a future observation in goodness-of-fit problems. Journal of the Royal Statistical Society: Series B (Methodological) **29**(1), 83–100 (1967)

Gabry, J., Češnovar, R., Johnson, A., Bronder, S.: Cmdstanr: R Interface to 'CmdStan'. (2024). R package version 0.8.1.9000, commit 2d65dde3bc4a1a8c4d94e62a9185efaab473eda3. https://github.com/stan-dev/cmdstanr

Harville, D.A.: The selection or seeding of college basketball or football teams for postseason competition. Journal of the American Statistical Association **98**(461), 17–27 (2003) https://doi.org/10.1198/016214503388619058 https://doi.org/10.1198/016214503388619058

Hobbs, B.P., Carlin, B.P., Mandrekar, S.J., Sargent, D.J.: Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. Biometrics **67**(3), 1047–1056 (2011)

Hobbs, B.P., Carlin, B.P., Sargent, D.J.: Adaptive adjustment of the randomization ratio using historical control data. Clinical Trials **10**(3), 430–440 (2013)

Hong, H., Fu, H., Carlin, B.P.: Power and commensurate priors for synthesizing aggregate and individual patient level data in network meta-analysis. Journal of the Royal Statistical Society Series C: Applied Statistics **67**(4), 1047–1069 (2018)

Hoffman, M.D., Gelman, A., *et al.*: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res. **15**(1), 1593–1623 (2014)

Harville, D.A., Smith, M.H.: The home-court advantage: How large is it, and does it vary from team to team? The American Statistician **48**(1), 22–28 (1994)

Hobbs, B.P., Sargent, D.J., Carlin, B.P.: Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. Bayesian Analysis **7**(3), 639–674 (2012) https://doi.org/10.1214/12-BA722

Issa Mattos, D., Martins Silva Ramos, É.: Bayesian paired comparison with the bpcs package. Behavior Research Methods **54**(4), 2025–2045 (2022) https://doi.org/10.3758/s13428-021-01714-2

Ingram, M.: How to extend Elo: a Bayesian perspective. Journal of Quantitative Analysis in Sports **17**(3), 203–219 (2021)

Knorr-Held, L.: Dynamic rating of sports teams. Journal of the Royal Statistical Society. Series D (The Statistician) **49**(2), 261–276 (2000). Accessed 2025-09-02

Kowal, D.R., Matteson, D.S., Ruppert, D.: Dynamic shrinkage processes. Journal of the Royal Statistical Society Series B: Statistical Methodology **81**(4), 781–804 (2019)

Karlis, D., Ntzoufras, I.: Analysis of sports data by using bivariate Poisson models. Journal of the Royal Statistical Society: Series D (The Statistician) **52**(3), 381–393

(2003) https://doi.org/10.1111/1467-9884.00366

Koopmeiners, J.S.: A comparison of the autocorrelation and variance of NFL team strengths over time using a bayesian state-space model. Journal of Quantitative Analysis in Sports **8**(3) (2012)

Kuk, A.Y.: Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. Journal of the Royal Statistical Society Series D: The Statistician **44**(4), 523–528 (1995)

Leonard, T.: An alternative Bayesian approach to the Bradley-Terry model for paired comparisons. Biometrics **33**(1), 121–132 (1977)

Lopez, M.J., Matthews, G.J., Baumer, B.S.: 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 (2018). Accessed 2025-09-02

Luce, R.D.: Individual Choice Behavior: A Theoretical Analysis. Wiley, New York (1959). https://books.google.it/books?id=c519AAAAMAAJ

Mitchell, T.J., Beauchamp, J.J.: Bayesian variable selection in linear regression. Journal of the american statistical association **83**(404), 1023–1032 (1988)

Macrì Demartino, R., Egidi, L., Torelli, N.: Alternative ranking measures to predict international football results. Computational Statistics, 1–19 (2024)

Macrì-Demartino, R., Egidi, L., Torelli, N.: Bayesian weighted discrete-time dynamic models for association football prediction. arXiv preprint arXiv:2508.05891 (2025)

Meng, X.-L.: Posterior predictive p-values. The annals of statistics **22**(3), 1142–1160 (1994)

Osei, P.P., Davidov, O.: Bayesian linear models for cardinal paired comparison data. Computational Statistics & Data Analysis **172**, 107481 (2022) https://doi.org/10.1016/j.csda.2022.107481

Ouma, L.O., Grayling, M.J., Wason, J.M., Zheng, H.: Bayesian modelling strategies for borrowing of information in randomised basket trials. Journal of the Royal Statistical Society Series C: Applied Statistics **71**(5), 2014–2037 (2022)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2025). R Foundation for Statistical Computing. https://www.R-project.org/

Rao, P.V., Kupper, L.L.: Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. Journal of the American Statistical Association **62**(317), 194–204 (1967) https://doi.org/10.1080/01621459.1967.10482901 https://www.tandfonline.com/doi/pdf/10.1080/01621459.1967.10482901

Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics, 1151–1172 (1984)

Spiegelhalter, D., Ng, Y.-L.: One match to go! Significance **6**(4), 151–153 (2009)

Springall, A.: Response surface fitting using a generalization of the Bradley-Terry paired comparison model. Journal of the Royal Statistical Society Series C: Applied Statistics **22**(1), 59–68 (1973) https://doi.org/10.2307/2346303

Stern, H.S.: On the probability of winning a football game. The American Statistician **45**(3), 179–183 (1991) https://doi.org/10.1080/00031305.1991.10475798

Thurstone, L.L.: A law of comparative judgement. Psychological Review **34**, 278–286 (1927)

Vehtari, A., Gelman, A., Gabry, J.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Statistics and computing **27**(5), 1413–1432 (2017)

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., Gelman, A.: loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.8.0.9000 (2024). https://mc-stan.org/loo/

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., Gabry, J.: Pareto smoothed importance sampling. Journal of Machine Learning Research **25**(72), 1–58 (2024)

Wainer, J.: A Bayesian Bradley-Terry model to compare multiple ML algorithms on multiple data sets. Journal of Machine Learning Research **24**(341), 1–34 (2023)

Whelan, J.T.: Prior distributions for the Bradley-Terry model of paired comparisons. arXiv preprint arXiv:1712.05311 (2017)

Zheng, H., Wason, J.M.: Borrowing of information across patient subgroups in a basket trial based on distributional discrepancy. Biostatistics **23**(1), 120–135 (2022)