Serve-Based Player Metrics in Tennis

Aiwen Li, Amrita Balajee, Audrey Kuan, Jonathan Pipping

October 2025

Abstract

Traditionally, tennis ratings systems (e.g., ELO) are designed to capture overall player strength. However, they overlook the serve's independent impact on point outcomes. This study develops serve-specific player metrics using generalized linear mixed models (GLMMs) applied to point-by-point data from Wimbledon and the U.S. Open. For each tournament and gender (male/female), we use separate models for first and second serves to estimate a server quality score (SQS). These scores are based on magnitude of serve speed, speed variability, and serve placement features, with random intercepts capturing unobserved player effects. The resulting SQS isolates the value of the serve and is evaluated against weighted ELO as a benchmark. Across three of the four datasets, the SQS exhibits significantly stronger correlation and lower predictive error for serve efficiency, while weighted ELO remains better for predicting overall point win percentage. These findings highlight the complementary roles of serve-specific and holistic performance metrics, offering actionable insights for coaching, forecasting, and player evaluation that separates serving skill from broader match performance.

1 Introduction and Motivation

Player ratings such as ELO are widely used to quantify performance in tennis by estimating player strength from match-level wins and losses. While effective for forecasting, these models compress the complexity of play into a single number and do not isolate the value of specific skills. In particular, the serve, which is the only shot entirely under the player's control, is treated implicitly rather than modeled directly. This prevents a clear separation between serving, returning, and rally performance. The serve particularly plays a defining role in shaping point outcomes, especially on faster surfaces, which motivates the need for a framework that models serve performance explicitly at the point level.

This paper establishes a framework to estimate a new metric for server quality using generalized linear mixed models (GLMMs). We call this metric a player's *server quality score* (SQS). SQS captures both measured skill, which is derived from serve features such as average serve speed, variability of speeds, and serve location characteristics; as well as unmeasured server effects modeled through player-specific random intercepts. To distinguish between aggressive and defensive serving contexts, we fit separate models for first and second serves. We also create a combined metric using a weighted average of the first- and second-serve scores to obtain a holistic serve performance measure.

Using point-by-point data from Wimbledon and the U.S. Open (2018-2019 and 2021-2024) (Sackmann, 2025), we benchmark these serve metrics against weighted ELO (wELO) (Angelini, Candila, & De Angelis, 2022). The comparison reveals how serve-specific models complement the existing wELO ratings system, while also offering unique predictive power when predicting the outcomes of shorter points.

The next section reviews related work. Section 3 describes our methodology with data preparation, GLMM models, and construction of server metrics. Section 4 reports out-of-sample testing results from the server quality scores and compares them against wELO. We proceed with discussion of our results in Section 5, then conclude with limitations and future directions in Section 6.

2 Literature Review

2.1 Match-Level Ratings and Forecasting

ELO-style ratings, adapted from chess and widely used in tennis, estimate a player's strength from match results. ELO (and some of its variants) has repeatedly forecast match winners well on large historical datasets (Klaassen & Magnus, 2003) (S. A. Kovalchik, 2016). A recent extension of this metric is weighted ELO (wELO), which scales the ELO update based on the margin of victory. In out-of-sample tests on ATP/WTA matches, wELO outperformed standard ELO and other common baselines, and a simple value-betting exercise using best available odds produced a positive rate of return (Angelini et al., 2022). Beyond pre-match ratings, combining a pre-match prior with in-play updates improves point-by-point forecasting. For example, Kovalchik and Reid (2019) show improvements from dynamic empirical-Bayes updates during a match (S. Kovalchik & Reid, 2019).

Other approaches for forecasting treat tennis as a hierarchical Markov process. In these models, point-level win probabilities map to game, set, and match win probabilities through closed-form or recursive calculations. Early work tested whether points are independent and identically distributed (IID) and found that this assumption is not always true due to changes in momentum and pressure (Klaassen & Magnus, 2001). Building on this, O'Malley (2008) computes exact game, set, and match win probabilities under an IID baseline and study how these change with serve and return strength (O'Malley, 2008). Later papers simulate variations of these models that allow state-dependent dynamics to forecast live win probabilities based on context from each point (Klaassen & Magnus, 2003) (Newton & Aslam, 2009).

2.2 Estimating Serve Probabilities

A key step in point-based models is estimating each player's "serve probability", which is the chance that a player wins the next point when they are serving. Serve probability is often split between first and second serves, and is sometimes adjusted based on surface or match context. A framework developed by Barnett and Clarke (2005) estimates these chances by using each player's past serve and return results along with their opponents' averages, then combining them to produce head-to-head serve probabilities for the upcoming match (Barnett & Clarke, 2005). Later work adds surface adjustments, shrinkage to stabilize small samples, and common-opponent adjustments to compare the server and returner against the same set of opponents. However, when comparing these methods, approaches that tie these estimates to ELO-type ratings often forecast serve outcomes better with lower RMSE than those without any adjustments for ELO (Gollub, 2021).

A growing line of work calibrates point-level inputs using player ratings, which keeps the interpretability of Markov models while leveraging the predictive strength of match-level information. Kovalchik & Reid (2019) show that starting from a rating-based prior and updating during play improves win-probability forecasts (S. Kovalchik & Reid, 2019). Gollub (2021) finds that aligning serve parameters with ELO improves serve performance forecasts compared to point-only inputs (Gollub, 2021). Together, these results support incorporating ratings into point-level models for serve forecasting and in-play prediction.

Despite frequent use of serving variables in match-level models, there remains no widely adopted measure that cleanly isolates the serve as a component of player quality. Creating this separation helps in both interpretation and decision-making: it clarifies how much advantage comes from the serve itself and informs coaching plans that are specific to serve and return. In response, we introduce a framework for server metrics focused just on serving performance. The next section describes the methodology for constructing these metrics.

3 Methodology

3.1 Data Preparation

We use Jeff Sackmann's point-by-point data for singles matches at the Wimbledon and U.S. Open tournaments (Sackmann, 2025). We combine the 2021–2024 and 2018–2019 seasons into a single pool for each of four datasets: Wimbledon males, Wimbledon females, U.S. Open males, and U.S. Open females. All modeling, metric construction, and validation are performed separately for each tournament/gender dataset.

Within each dataset, we split the combined six years of points into disjoint training and testing sets by randomly choosing 80% of the points for training, and the rest of the points for testing. We then fit models on the training set to construct the server metrics and perform out-of-sample validation on the testing set.

From the raw data, we remove points with missing or empty serve location, restrict to valid first- or second-serve observations, and filter out serves that have a speed of 0. There are two variables for serve location, which are ServeWidth and ServeDepth. ServeWidth takes on the five values "Body," "Body/Center," "Body/Wide," "Center," and "Wide"; and ServeDepth takes on the two values "Close to Line" or "Not Close to Line." For every serve, we define a location bin:

$$\texttt{location\_bin} = \text{W}\langle\text{ServeWidth}\rangle\_\text{D}\langle\text{ServeDepth}\rangle.$$

For each server j and serve type $s \in \{1, 2\}$ (first or second), we aggregate the following server-level features on the training set:

- avg_speed_{js}: average serve speed (mph),
- sd_speed_{js}: standard deviation of serve speed (a proxy for variability/unpredictability in serve speed),
- $\text{location\_entropy}_{js} = -\sum_i p_{ijs} \log_2 p_{ijs}$ where p_i is the proportion of serves landing in location bin i,
- $\text{modal\_location}_{js}$: the modal location bin (one-hot encoded in the model),
- n_{js}: number of serves observed of type s.

We keep servers with $n_{js} > 20$ to remove servers with sparse data from both the training and testing sets. The continuous features are standardized (z-scored) within each dataset and serve type (first/second). Then, $\text{modal\_location}_{js}$ is treated as a set of categorical indicator variables.

3.2 Fitting GLMMs to Compute Metrics

We fit separate generalized linear mixed models (GLMMs) for first and second serves to produce distinct first- and second-serve metrics. GLMMs are fit to the binary point outcome of whether the server wins the point.

For serve type $s \in \{1, 2\}$, let Y_{ijs} be the binary indicator that server j wins point i. The model is

$$\begin{aligned}\text{logit Pr}(Y_{ijs} = 1) = \beta_{0s} &+ \beta_{1s}\,\text{avg\_speed}_{js}^{(z)} + \beta_{2s}\,\text{sd\_speed}_{js}^{(z)} \\ &+ \beta_{3s}\,\text{location\_entropy}_{js}^{(z)} + \boldsymbol{\beta}_{4s}^{\top}\,\text{OneHot}(\text{modal\_location}_{js}) + u_{js}\end{aligned}$$

where $u_{js} \sim \mathcal{N}(0, \sigma_s^2)$ is a random intercept for server j that captures unmeasured server-specific effects for serve type s. The fixed effects use the server-level features from the data in each point. In R, the models are estimated using `lme4::glmer` with a binomial link.

Then, for each serve type s, we convert the fitted model into a server-level *server quality score* (SQS) by combining (i) measured skill from fixed effects and (ii) unmeasured server effects from the random intercept:

$$\underbrace{\text{SQS}_{js}^{(\ell)}}_{\text{log-odds}} = \underbrace{\hat{\beta}_{0s} + \hat{\beta}_{1s}\,\text{avg\_spd}_{js}^{(z)} + \hat{\beta}_{2s}\,\text{sd\_spd}_{js}^{(z)} + \hat{\beta}_{3s}\,\text{entropy}_{js}^{(z)} + \hat{\boldsymbol{\beta}}_{4s}^{\top}\,\text{OneHot}(\text{modal\_location}_{js})}_{\text{measured skill}} + \underbrace{\hat{u}_{js}}_{\text{server effects}}$$

For interpretability, we convert the logit-scaled server quality scores into a probability scale after all modeling and computing are completed.

To make a metric that reflects overall server quality between both first and second serves, we take a weighted average of the SQS's of each server on the log-odds scale. The weights are determined by the number of serves of each type (n_1 or n_2) for each server j. If a server had ≤ 20 serves for one type, then we just take the SQS using the serve type that had > 20 serves.

$$\text{SQS}_{j,\text{comb}} = \frac{n_{j1}\,\text{SQS}_{j1} + n_{j2}\,\text{SQS}_{j2}}{n_{j1} + n_{j2}}$$

After computing the weighted average on the log-odds scale, we again convert the SQS into a probability scale for interpretability.

3.3 Out-of-Sample Evaluation and Baselines

All evaluations are conducted on the held-out testing set of the same tournament/gender dataset (Wimbledon/U.S. Open and Male/Female) at the server level. For each server j, we aggregate the testing data to compute two outcome variables, win percentage and serve efficiency, that are defined below. *Win percentage* is the proportion of serving points won by the server, while *serve efficiency* is the proportion of serving points won by the server with a rally count of ≤ 3 (meaning the server won the point with only one or two shots).

$$\widehat{\text{WinPct}}_j = \frac{\#\{\text{serve points won}\}}{\#\{\text{serve points}\}}, \qquad \widehat{\text{ServeEff}}_j = \frac{\#\{\text{serve points won with RallyCount} \leq 3\}}{\#\{\text{serve points}\}}.$$

We then compare each SQS (on the probability scale) against these testing outcomes by fitting simple linear regression models of the testing outcomes vs. the SQS. All variables for the out-of-sample testing regression models are standardized. We record the correlation coefficient r, root mean squared error (RMSE), and two-sided p-values from r. Again, only servers with at least 20 testing serves are included in the testing data.

For each metric (based on first and second serves separately, as well as the combined metric), we evaluate their performance by using the server's average weighted ELO (wELO) as a benchmark. wELO scores are scraped from the R package `welo` (Angelini et al., 2022).

4 Results

4.1 Wimbledon Males

Table 1 reports out-of-sample results for the generalized linear mixed model (GLMM) server metrics (SQS) compared to the baseline predictor (weighted ELO). We evaluate the metrics' performance separately for first serves, second serves, and the combined serves metric. Each version of the SQS is tested against both outcomes (serve efficiency and win percentage) on the held-out data.

For first serves, the GLMM-based server quality score performs best for predicting serve efficiency, with the highest correlation ($r = 0.49$, $p < 10^{-7}$) and lowest RMSE (1.00). This indicates that the model successfully captures how first-serve attributes (speed, variability, and location features) translate into short-point success. Although weighted ELO (wELO) performs slightly better for overall point win percentage ($r = 0.31$, $p = 0.0015$), its correlation with serve efficiency is not significant ($r = 0.14$, $p = 0.16$), which shows that it primarily reflects all-court ability rather than serve-specific strength.

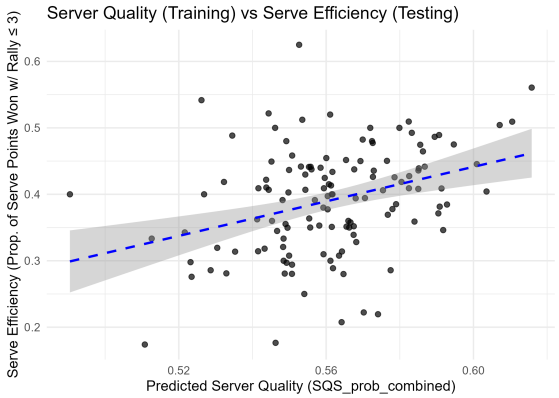
For second serves, the SQS also performs better for predicting serve efficiency, but the relationship is weaker ($r = 0.15$, $p = 0.23$). For win percentage, both the SQS and wELO perform comparably ($r \approx 0.30$–0.32).

When combining both serve types into a weighted overall metric, the SQS again performs best for predicting serve efficiency ($r = 0.34$, $p < 10^{-4}$) while maintaining similar RMSE to wELO. This reinforces the idea that the GLMM-based server metric captures the serve's direct contribution to short-point success more precisely than general-purpose ratings.

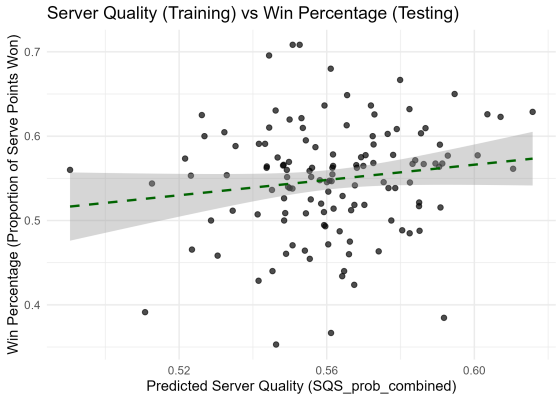
Figure 1a plots the combined SQS versus observed serve efficiency, and Figure 1b plots combined SQS versus overall win percentage for Wimbledon males. The first scatterplot exhibits a moderate positive relationship, validating the model's explanatory power for serve-specific outcomes. The second shows a weaker but still positive trend, consistent with the notion that serve quality is only one component of broader performance.

Table 1: Out-of-sample GLMM server metric performance for Wimbledon males.

Predictor	**Outcome**	**n**	**RMSE**	**Correlation (r)**	**p-value**	**Serve Type**
SQS	Serve efficiency	104	1.00	0.49	9.5×10^{-8}	First
wELO	Serve efficiency	104	1.31	0.14	0.16	First
SQS	Win percentage	104	1.20	0.28	0.004	First
wELO	Win percentage	104	1.17	0.31	0.0015	First
SQS	Serve efficiency	66	1.29	0.15	0.23	Second
wELO	Serve efficiency	66	1.34	0.09	0.49	Second
SQS	Win percentage	66	1.17	0.30	0.013	Second
wELO	Win percentage	66	1.15	0.32	0.008	Second
SQS	Serve efficiency	129	1.14	0.34	7.9×10^{-5}	Combined
wELO	Serve efficiency	129	1.35	0.08	0.35	Combined
SQS	Win percentage	129	1.30	0.14	0.10	Combined
wELO	Win percentage	129	1.29	0.16	0.064	Combined



(a) SQS (combined) vs. serve efficiency.



(b) SQS (combined) vs. win percentage.

Figure 1: Out-of-sample relationships between combined SQS and serve outcomes for Wimbledon males.

4.2 Wimbledon Females

Table 2 reports out-of-sample evaluation results for the Wimbledon females dataset. As with the male data, separate GLMM-based server quality scores (SQS) are estimated for first and second serves, and an overall combined metric is produced by weighting by serve frequency.

For first serves, SQS performs better with a higher correlation and lower RMSE than wELO for predicting serve efficiency ($r = 0.33$, RMSE $= 1.15$). For win percentage, all predictors show weaker associations

($r \approx 0.17$–0.20), indicating that rally-based play contributes more strongly to overall point outcomes in this group.

For second serves, the relationships are even weaker overall. Weighted ELO (wELO) performs best for both outcomes, achieving $r = 0.36$ ($p = 0.006$) for serve efficiency and $r = 0.27$ ($p = 0.04$) for win percentage. This suggests that for women's second serves, broader skill factors captured by ELO may better explain point-level success than serve-specific features alone.

Finally, the combined serve metric performs comparably to wELO. Both have moderate positive correlations with serve efficiency ($r = 0.33$–0.35), though wELO performs slightly better for win percentage ($r = 0.43$, $p < 10^{-5}$). These results indicate that the GLMM-based SQS provides an interpretable, serve-focused complement to general rating systems like ELO, particularly for evaluating serve efficiency.

Table 2: Out-of-sample GLMM server metric performance for Wimbledon females.

Predictor	**Outcome**	**n**	**RMSE**	**Correlation (r)**	**p-value**	**Serve Type**
SQS	Serve efficiency	79	1.15	0.33	0.0026	First
wELO	Serve efficiency	79	1.31	0.13	0.26	First
SQS	Win percentage	79	1.27	0.18	0.11	First
wELO	Win percentage	79	1.26	0.20	0.08	First
SQS	Serve efficiency	57	1.31	0.13	0.34	Second
wELO	Serve efficiency	57	1.12	0.36	0.0056	Second
SQS	Win percentage	57	1.37	0.05	0.72	Second
wELO	Win percentage	57	1.20	0.27	0.042	Second
SQS	Serve efficiency	99	1.15	0.34	6.9×10^{-4}	Combined
wELO	Serve efficiency	99	1.13	0.35	3.2×10^{-4}	Combined
SQS	Win percentage	99	1.22	0.24	0.015	Combined
wELO	Win percentage	99	1.06	0.43	8.1×10^{-6}	Combined

4.3 U.S. Open Males

Table 3 summarizes the out-of-sample testing results for the U.S. Open males dataset.

For first serves, the SQS again outperforms wELO when predicting serve efficiency ($r = 0.29, \text{RMSE} = 1.19$). For win percentage, correlations across all predictors are modest ($r = 0.15$–0.26), though wELO performs best in that category.

For second serves, none of the serve-specific predictors correlate significantly with serve efficiency. In fact, wELO exhibits a small but significant negative relationship ($r = -0.29$, $p = 0.0028$). This shows that servers who succeed through second-serve defense often have broader tactical or return strengths not reflected in serve-only data. Interestingly, the GLMM-based SQS slightly outperforms others in predicting win percentage ($r = 0.23$, $p = 0.018$). Nonetheless, none of the metrics perform strongly for predicting server-level outcomes.

The combined SQS has slight explanatory power, with correlations of $r = 0.14$ for serve efficiency ($p = 0.08$) and $r = 0.16$ for win percentage ($p = 0.034$). While wELO continues to yield stronger results for overall win percentage ($r = 0.28$, $p < 0.001$), its poor performance for serve efficiency illustrates the distinction between holistic player ratings and serve-specific skill metrics. Overall, these findings suggest that while SQS captures a consistent signal from serve features, its explanatory strength is reduced on hard courts where non-serve factors (e.g., rally length, movement, and return quality) play a larger role.

Table 3: Out-of-sample GLMM server metric performance for U.S. Open males.

Predictor	Outcome	n	RMSE	Correlation (r)	p-value	Serve Type
SQS	Serve efficiency	142	1.19	0.29	5.0×10^{-4}	First
wELO	Serve efficiency	142	1.40	0.01	0.91	First
SQS	Win percentage	142	1.30	0.15	0.068	First
wELO	Win percentage	142	1.21	0.26	0.0016	First
SQS	Serve efficiency	105	1.50	–0.13	0.18	Second
wELO	Serve efficiency	105	1.60	–0.29	0.0028	Second
SQS	Win percentage	105	1.24	0.23	0.018	Second
wELO	Win percentage	105	1.28	0.18	0.069	Second
SQS	Serve efficiency	168	1.31	0.14	0.077	Combined
wELO	Serve efficiency	168	1.49	–0.11	0.15	Combined
SQS	Win percentage	168	1.29	0.16	0.034	Combined
wELO	Win percentage	168	1.19	0.28	2.0×10^{-4}	Combined

4.4 U.S. Open Females

Table 4 presents the out-of-sample evaluation results for the U.S. Open females dataset. Relative to the other three groups, the women's U.S. Open sample shows the strongest overall relationships between serve metrics and outcomes.

For first serves, the SQS has a moderately strong and highly significant correlation with serve efficiency ($r = 0.54$, $p < 10^{-9}$), while weighted ELO (wELO) is weaker ($r = 0.27$, $p = 0.004$). The same pattern appears for win percentage: both predictors are significant, but the model-based SQS achieves the lowest RMSE (1.13) and slightly higher correlation ($r = 0.36$) than the baseline. These results indicate that on hard courts, variation in serve speed and precision is a strong driver of point-level success for top female players.

For second serves, none of the predictors display meaningful correlations with either outcome ($r < 0.20$, $p > 0.10$). This reflects the greater tactical variability in second serves, which are often shaped by risk avoidance and opponent quality rather than measurable speed or entropy effects.

For the combined serve metric, the SQS correlates significantly with both serve efficiency ($r = 0.43$, $p < 10^{-6}$) and win percentage ($r = 0.27$, $p = 0.001$), which outperforms wELO on RMSE and correlation for serve efficiency. This again highlights the GLMM-based metric's ability to capture serve-specific quality, while ELO remains more reflective of general match performance.

5 Discussion

The out-of-sample results across all four datasets reveal how serve-based metrics relate to point-level performance. Overall, the GLMM-derived server quality score (SQS) shows meaningful predictive power for serve efficiency, especially on first serves, while showing weaker relationships for second serves and total win percentage.

5.1 First Serves

The relationships between the server metrics and both outcomes are strongest when just looking at first serves. Across tournaments and genders, the SQS metric correlates strongly with serve efficiency and achieves the lowest or near-lowest RMSE among all predictors. This pattern is especially dominant for the U.S. Open females data, where the correlation reaches $r = 0.54$ ($p < 10^{-9}$), indicating that serve characteristics account for a substantial portion of success on first serves.

Table 4: Out-of-sample GLMM server metric performance for U.S. Open females.

Predictor	Outcome	n	RMSE	Correlation (r)	p-value	Serve Type
SQS	Serve efficiency	113	0.95	0.54	6.0×10^{-10}	First
wELO	Serve efficiency	113	1.20	0.27	0.004	First
SQS	Win percentage	113	1.13	0.36	8.4×10^{-5}	First
wELO	Win percentage	113	1.14	0.35	1.8×10^{-4}	First
SQS	Serve efficiency	74	1.33	0.10	0.37	Second
wELO	Serve efficiency	74	1.43	–0.04	0.75	Second
SQS	Win percentage	74	1.29	0.16	0.18	Second
wELO	Win percentage	74	1.26	0.19	0.10	Second
SQS	Serve efficiency	134	1.07	0.43	2.5×10^{-7}	Combined
wELO	Serve efficiency	134	1.28	0.18	0.040	Combined
SQS	Win percentage	134	1.20	0.27	0.0014	Combined
wELO	Win percentage	134	1.16	0.32	1.5×10^{-4}	Combined

For both Wimbledon samples, the SQS performs similarly well, capturing between one-quarter and one-half of the observed variance in efficiency. These results confirm that the GLMM framework effectively models how observable serving traits contribute to short-point success. While wELO alone remains a decent baseline predictor, SQS consistently provides additional explanatory value by integrating multiple features (magnitude of speed, variability of speed, and location entropy) into a single, cohesive model.

5.2 Second Serves

The out-of-sample results with just second serves are more variable and generally weaker. Correlations between the SQS and serve efficiency are small or statistically insignificant across all four datasets, rarely exceeding $r = 0.15$. For both U.S. Open and Wimbledon samples, weighted ELO (wELO) occasionally outperforms SQS, especially for predicting overall win percentage. This likely reflects that points played on second serves depend more on consistency, rally tolerance, and contextual strategy. These factors are captured better by ELO, since it's a more general rating of match performance, rather than purely serving variables. Nonetheless, modest but significant correlations between SQS and win percentage in some cases (e.g., Wimbledon males: $r = 0.30$, $p = 0.013$) suggest that the model still identifies aspects of underlying serve quality, even if the explanatory power is limited.

5.3 Combined Serve Metric

When the metrics using first and second serves are combined via weighted averaging on the log-odds scale, the resulting SQS has moderate predictive strength for serve efficiency across all groups ($r = 0.33$–0.43). In three out of the four datasets, it outperforms weighted ELO for serve efficiency, which demonstrates its ability to isolate serve-specific performance effects. In the fourth dataset, the SQS performs comparably (similar correlation and RMSE) with wELO.

However, for win percentage, ELO remains the stronger predictor in three of the four datasets. This distinction reflects the conceptual boundary between the two measures: SQS captures the technical and tactical contribution of the serve itself, whereas ELO reflects composite performance across all phases of play. The results therefore highlight how SQS serves as a specialized, complementary performance indicator rather than a substitute for general player ratings.

5.4 Surface and Gender Effects

Surface and gender both influence the strength of these relationships. The Wimbledon datasets have relatively high SQS correlations, consistent with the greater impact of serving on faster grass courts. In contrast, the U.S. Open male sample exhibits the weakest relationships overall, suggesting that on hard courts, other factors such as return depth and rally length reduce the direct influence of serve characteristics on point outcomes.

Gender patterns follow a similar logic: correlations are typically higher in the women's datasets, especially for serve efficiency. This may reflect broader variation in serve ability among top female players and a greater payoff from effective serve placement and disguise from higher variability.

Taken together, the results show that the SQS reliably captures serve-specific skill in contexts where serve dominance plays a major role, such as with first serves and faster surfaces. The server metric maintains interpretability even when its predictive strength declines. The cross-tournament consistency in serve-efficiency correlations confirms the model's generalizability, while the divergence from ELO on win percentage shows that the metric is better for predicting serve-specific outcomes.

6 Conclusion

The results across tournaments and serve types demonstrate that the server quality score (SQS), estimated through a generalized linear mixed model, provides a robust and interpretable measure of serve-based performance. By modeling point outcomes with both fixed serve features and random player effects, the SQS isolates technical and tactical components of serving that traditional performance ratings, such as weighted ELO, cannot directly capture.

Across all four datasets, the SQS consistently predicts serve efficiency more accurately than existing baselines. These findings reinforce that a model-based approach can quantify serve skill in a way that is both generalizable and surface-sensitive. Nonetheless, serve-specific features alone cannot fully explain outcomes, which highlights the importance of contextual and behavioral factors in determining point success. Taken together, the SQS offers a scalable foundation for serve-based player evaluation. The metric is a complement, rather than a replacement, for holistic performance measures like ELO.

6.1 Limitations and Future Work

While the mixed-effects framework successfully captures measurable components of serve performance, several limitations still remain. For instance, the current model does not incorporate match context such as score state, point importance, or player fatigue. These factors influence serve selection and execution and may explain why the SQS is weaker for second serves and overall win percentage. Also, although location entropy and modal serve location are included, these variables summarize complex spatial variables into discrete bins. A continuous spatial or probabilistic representation could capture finer distinctions in serve targeting.

Additionally, the quality and breadth of the data we use could be improved and expanded on. For example, serves with recorded speeds of zero corresponded to faults that were excluded during data cleaning. This may bias the analysis by filtering out risk-related variation that contributes meaningfully to overall serve performance. In addition, the model does not account for differences in return quality or the ability of players to adapt to different serves. In reality, the success of a serve depends on the interaction between server and returner skill. Finally, the datasets are restricted to major tournaments (Wimbledon and U.S. Open), which may limit the variability observed in serve performance compared to other tournaments, especially those played on different types of courts.

7 Reproducibility

The full code and results are in our Github repository for this project: https://github.com/wsabi-summer-lab/serve-profiles.

8 Acknowledgments

We thank our advisor, Abraham Wyner, and our mentor, Jonathan Pipping, for their invaluable feedback throughout the research process. We also thank the Wharton Sports Analytics and Business Initiative (WSABI) for supporting this work through their summer research program.

References

Angelini, G., Candila, V., & De Angelis, L. (2022). Weighted elo rating for tennis match predictions. , *297*(1), 120–132. Retrieved 2025-08-17, from https://www.sciencedirect.com/science/article/pii/S0377221721003234 doi: 10.1016/j.ejor.2021.04.011

Barnett, T., & Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches. , *16*(2), 113–120. Retrieved 2025-08-19, from https://doi.org/10.1093/imaman/dpi001 doi: 10.1093/imaman/dpi001

Gollub, J. (2021). Forecasting serve performance in professional tennis matches. , *7*(4), 223–233. Retrieved 2025-08-17, from https://doi.org/10.3233/JSA-200345 (Publisher: SAGE Publications) doi: 10.3233/JSA-200345

Klaassen, F. J., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model. , *96*(454), 500–509. Retrieved 2025-08-19, from https://doi.org/10.1198/016214501753168217 (Publisher: ASA Website _eprint: https://doi.org/10.1198/016214501753168217) doi: 10.1198/016214501753168217

Klaassen, F. J., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. , *148*(2), 257–267. Retrieved 2025-08-19, from https://linkinghub.elsevier.com/retrieve/pii/S0377221702006823 doi: 10.1016/S0377-2217(02)00682-3

Kovalchik, S., & Reid, M. (2019). A calibration method with dynamic updates for within-match forecasting of wins in tennis. , *35*(2), 756–766. Retrieved 2025-08-19, from https://www.sciencedirect.com/science/article/pii/S0169207017301395 doi: 10.1016/j.ijforecast.2017.11.008

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. , *12*(3). Retrieved 2025-08-19, from https://www.degruyter.com/document/doi/10.1515/jqas-2015-0059/html doi: 10.1515/jqas-2015-0059

Newton, P. K., & Aslam, K. (2009). Monte carlo tennis: A stochastic markov chain model. , *5*(3). Retrieved 2025-08-19, from https://www.degruyterbrill.com/document/doi/10.2202/1559-0410.1169/html?srsltid=AfmBOooFr93UcHyYxPWmYsaNauIFCgp3oT-cZRFnOVudDh5ldyieLfZE&utm_source=chatgpt.com (Publisher: De Gruyter) doi: 10.2202/1559-0410.1169

O'Malley, A. J. (2008). Probability formulas and statistical analysis in tennis. , *4*(2). Retrieved 2025-08-19, from https://www.degruyterbrill.com/document/doi/10.2202/1559-0410.1100/html (Publisher: De Gruyter) doi: 10.2202/1559-0410.1100

Sackmann, J. (2025). *JeffSackmann/tennis_slam_pointbypoint*. Retrieved 2025-07-28, from https://github.com/JeffSackmann/tennis_slam_pointbypoint (original-date: 2015-03-08T14:50:29Z)