Modeling Best Ball Fantasy Football Scores using Monte Carlo Simulation - Project Report

Harrison Stanton

February 2026

1 The problem and given data

In a typical fantasy football league, each manager sets their lineup prior to the individual players' games happening. However, in a best ball format, managers are not required to set a starting lineup; instead, your lineup is optimized dynamically during each game. While this removes start-sit decisions and scouring injury reports throughout the week to learn the availability of each player, it also fundamentally changes the optimal roster construction. Because any player on the roster can contribute in a given week, depth and week-to-week volatility become significantly more valuable than in traditional formats.

One challenge the format introduces is projecting team totals and win probability. In a standard fantasy league, a team projection is straightforward. Each player's weekly output is projected, and the projections of the players in the starting lineup are summed to give a team total. However, this method of projecting falls short for best ball. Since managers do not designate a starting lineup, bench players are not excluded from weekly outcomes. Ignoring bench depth leads to projections that systematically undervalue teams built around depth and volatility, making it difficult to evaluate roster strength or compare teams within the format.

Many best ball platforms, including Sleeper, still rely on traditional weekly player projections when projecting team scores. However, because only a subset of the highest-scoring players at each position contribute to the weekly total in best ball, team outcomes are driven not just by the expected value of the starting lineup, but by the distribution of possible outcomes across the entire roster. As a result, teams with similar projected point totals in a standard league projection can have meaningfully different probabilities of producing high weekly scores in best ball. In light of this, I sought to:

1. Develop a more accurate projection for team totals than Sleeper by accounting for the best ball scoring system and incorporating the strength of a team's bench.

2. Use said projections to develop a corresponding win probability model.
3. Have outputs from the model available to users weekly before Thursday Night Football.

Data used to create the best ball projections comes from Pro Football Reference and the play-by-play and NFL Draft datasets contained in the nflfastR package.

2 Approach

The projection methodology is built around Monte Carlo simulation, with 20,000 simulated outcomes generated for each team. Monte Carlo methods are well-suited for this problem because fantasy football outcomes are inherently uncertain, and best ball scoring rewards both ceiling performances and roster depth rather than a single expected outcome.

To apply Monte Carlo simulation, I first defined a range of outcomes for each player. This range was based on the player's most recent 16 games[1], using the 7.5th percentile as a conservative lower bound and the 90th percentile as an upper bound. These bounds were chosen to capture both downside risk and ceiling outcomes while reducing the influence of extreme outliers.

Using the range of outcomes, I simulated 20,000 potential fantasy scores for each player for the upcoming week. Each simulation represents one possible realization of how that week could unfold. For every simulated outcome, I then optimized the team's lineup according to best ball roster rules, producing 20,000 optimized lineups rather than a single static projection.

The final team projection is calculated as the average score across all optimized lineups. This approach accounts for positional depth, volatility, and best ball lineup selection, allowing the model to value high-upside bench players and depth in a way that traditional point projections cannot.

Win probabilities are derived from the optimized, simulated lineups. For each of the 20,000 lineups, Team A's optimized score is compared to Team B's. The win probability is then calculated as the proportion of simulations in which Team A outscores Team B.

Projections and win probabilities were calculated each week on Wednesday night and compiled into a spreadsheet that was sent out to users Thursday morning.

[1] See Appendix for how rookies were handled

3 Model Calibrations

There are cases in which a player's previous 16 games are not representative of their expected future performance, such as when a starting running back is injured or when a player changes teams in the offseason. In these situations, the distribution constructed from the player's past games fails to accurately capture their next week's range of outcomes.

When a starting running back is injured, the 16-game sample does not account for the anticipated increase in volume for the backup. An increase in volume raises the player's floor outcome, while the ceiling outcome increases by a larger margin. To account for this asymmetry, all simulated outcomes for the backup running back were multiplied by a constant between 1.35 and 1.75, depending on the player's prior level of involvement. This adjustment increased the projected mean while disproportionately expanding the upper tail of the distribution.

A similar approach was used for players whose roles were expected to decline. In these cases, simulated outcomes were multiplied by a constant between 0.35 and 0.85. This reduced the projected mean and compressed the distribution, with larger effects on the upper range of outcomes than the lower range.

There were also situations in which a player's distribution had an appropriate shape but required a shift upward or downward. This could occur if a player experienced unusually high touchdown variance, played through an injury, or faced an atypically strong stretch of defenses during the sample period. In these cases, rather than applying a multiplier that would substantially alter the distribution's shape, a constant value of 2 points was added to or subtracted from each simulated outcome. This adjustment shifted the mean while largely preserving the player's estimated floor and ceiling outcomes.

4 Results

During the 2025 NFL season, weekly projections were generated for each fantasy matchup using the model described above and evaluated after the season against observed fantasy scores. Model performance was compared to Sleeper's projections using three common forecasting error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Across the full season, the proposed model reduced forecasting error relative to Sleeper by 20.0% in MAE, 19.9% in RMSE, and 11.1% in MAPE, indicating consistently improved point predictions across multiple error measures.

To assess whether the observed reduction in absolute error was statistically significant, both a paired t-test and a Wilcoxon signed-rank test were conducted.

In both cases, the null hypothesis (H_0) stated that there was no difference in absolute projection error between the two models, while the alternative hypothesis (H_1) stated that the proposed model produced lower absolute error than Sleeper.

The paired t-test yielded a test statistic of $t = 3.60$ with a p-value of 0.000204, indicating strong evidence against the null hypothesis. The corresponding one-sided 95% confidence interval for the mean difference in absolute error was $[2.6, \infty)$, suggesting that the proposed model reduced absolute error by at least 2.6 points on average.

Results from the Wilcoxon signed-rank test were consistent (p = 0.000128), providing additional support that the improvement was not driven by distributional assumptions. Together, these findings provide strong statistical evidence that the proposed projection model significantly outperforms Sleeper's projections in terms of forecasting accuracy.

5 Appendix

Rookies, by definition, have no NFL experience. Because a player's last 16 games were used to create their range of outcomes, this presented a difficult situation. How do you project a rookie who has no data to go off of? The way this was handled was by grouping rookies based on their position and draft capital into cohorts. As an example, some of the cohorts included first round tight ends, top ten overall quarterbacks, etc. Instead of relying solely on each rookie's individual performances, the scores of other rookies from the same cohort in previous years were layered into each rookie's range of outcomes. This meant that in Week 1 of the NFL season, every player who was in the same cohort had the same range of outcomes and same projection. However, as rookies accumulated games played, their own scores were weighed heavily in the calculation of their range of outcomes by using a weighted quantile to calculate their 7.5th and 90th percentile outcomes. If a rookie had played between 1 game and 8 games, then their own scores accounted for 80% of the weight in the quantile, while the cohort represented the other 20%. When rookies had played between 9 and 12 games, this jumped to 90% for the individual rookie and 10% for the cohort. After the rookie had played in 12 games, their own fantasy point totals from the season accounted for 100% of their range of outcomes. This allowed for a reasonable projection for each rookie even with little data.