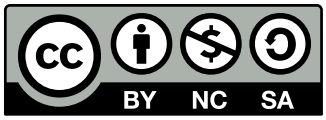
Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Mathematics and Operations Research

Probabilistic Forecasting of PGA Tour Events

Matias Linnankoski

Master's Thesis
2026



Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Matias Linnankoski

Title Probabilistic Forecasting of PGA Tour Events

Degree programme Mathematics and Operations Research

Major Systems and Operations Research

Thesis supervisor PhD Jukka Kohonen

Thesis advisor PhD Jukka Kohonen

Date 30 July 2026 **Number of pages** 67 **Language** English

Abstract

This thesis studies the probabilistic modelling of PGA Tour stroke-play tournament outcomes using historical round-level score data. Player abilities are modelled by weighting historical performances. These ability estimates are combined with a round-level residual score distribution to simulate rounds and tournament outcomes and estimate probabilities for making the cut and finishing in the top 20, top 10, and top 5.

The model is validated by comparing predicted outcome probabilities with realised tournament outcomes. Extensions such as an alternative weighting scheme, player-specific variance, weather adjustments, and player-course interactions are also examined. The results suggest that most of the predictive signal is captured by field-strength-adjusted historical performances weighted toward recent rounds, while additional extensions provide only marginal improvements.

Keywords golf, PGA Tour, probabilistic forecasting, Monte Carlo simulation, player skill estimation

Aalto-yliopisto
Aalto-universitetet
Aalto University

Tekijä Matias Linnankoski

Työn nimi PGA Tourin lyöntipeliturnausten todennäköisyysmallinnus

Koulutusohjelma Mathematics and Operations Research

Pääaine Systems and Operations Research

Valvoja FT Jukka Kohonen

Ohjaaja FT Jukka Kohonen

Päivämäärä 30.07.2026 **Sivumäärä** 67 **Kieli** englanti

Tiivistelmä
Tässä diplomityössä tutkitaan PGA Tourin lyöntipeliturnausten lopputulosten todennäköisyysmallintamista historiallisten kierrostulosten avulla. Pelaajien taitotasoa mallinnetaan painottamalla historiallisia suorituksia. Nämä taitotasoestimaatit yhdistetään kierroskohtaisen satunnaisvaihtelun jakaumaan, minkä avulla simuloidaan kierroksia ja turnausten lopputuloksia sekä estimoidaan todennäköisyydet cutin läpäisylle ja sijoittumiselle 20, 10 ja 5 parhaan joukkoon.

Malli vahvistetaan vertaamalla todennäköisyysennusteita toteutuneisiin lopputuloksiin. Lisäksi työssä tarkastellaan laajennuksia, kuten vaihtoehtoista painotusmenetelmää, pelaajakohtaista varianssia, sään vaikutuksia sekä pelaajan ja kentän yhteisvaikutuksia. Tulokset viittaavat siihen, että suurin osa ennustevoimasta saavutetaan, kun historialliset kierrostulokset normalisoidaan kilpailijoiden tason avulla ja lähihistorian tuloksia painotetaan enemmän. Laajennuksilla saavutetaan vain vähäisiä parannuksia.

Avainsanat golf, PGA Tour, todennäköisyysennustaminen, Monte Carlo -simulointi, pelaajan taitotason estimointi

Contents

Symbols and abbreviations

Symbols

Y	Round-level performance (player's score − field average score)
Y^{adj}	Round-level performance adjusted with the field skill (ability) level
μ	Player ability compared to an average player
σ^2	Player round-level variance

Abbreviations

EWMA	exponentially weighted moving average
NLL	negative log-likelihood
MAE	mean absolute error
RMSE	root mean squared error
NNLS	non-negative least squares
field	set of players participating in a round

1 Introduction

Golf is inherently stochastic. Each shot is played with an intended outcome, but the realised result is uncertain and can be viewed as a draw from a distribution around the intended outcome. Over the course of a round, this shot-level variability accumulates, resulting in variation in total round-level scores. It is a sport which is well suited for statistical modelling, as it has many repetitions and minimal opponent interaction compared to most other sports. Most modern tours collect extensive data for statistical and entertainment purposes, even at a shot-by-shot level. This combined with a long history of data gives a good foundation for statistical modelling on the subject.

One of the earliest statistical models in golf was the Hardy distribution, developed by G. H. Hardy in 1945. Its aim was to predict the score for a hole through a closed form function, using probability of good stroke and probability of bad stroke as inputs [1]. Later on, Larkey (1994) attempted to adjust tournament results for course difficulty and field quality, as discussed in [2]. This was followed by Berry (2001) modelling round scores as normally distributed around a player skill parameter, and Connolly & Rendleman (2008) extending this further by accounting for course difficulty and random variation through a generalised additive model [3, 4]. Currently, there are also commercial statistics and prediction providers, of which Data Golf is arguably the best known [5].

Golf is also a popular sport in betting markets, with larger tournaments attracting substantial wagering activity. Betting market prices are partly driven by supply and demand, and the markets can therefore produce very tightly quoted spreads. For instance, Rory McIlroy was quoted at decimal odds of 2.78–2.80 before the final round at the Masters 2026. Such a narrow spread, at least in the author's view, is primarily a reflection of liquidity, rather than evidence that the underlying win probability is known with the same degree of precision. Given the inherent randomness of golf, it is unlikely that a statistical model could justify such a narrow probability interval for the event itself.

While betting markets provide an interesting practical context, they are not the primary focus of this thesis. The aim is to explore how round-level golf scores can be modelled statistically, and how such models can be used to forecast tournament outcomes. Earlier studies have mostly aimed to explain past results rather than to forecast new ones, and commercial providers describe their methods only in outline and comment little on their results.

2 Problem Statement

The aim of this study is to create a *probabilistic forecasting* framework for PGA Tour stroke play events. Relevant terms, such as stroke and round, and the structure of golf tournaments, are explained in Section 3. Probabilistic forecasting requires an estimate (predicted mean) of how a player is expected to perform in a given event, together with the uncertainty (estimated distribution) around the expected performance. This study models the distribution of round-level scores for the players using historical performance data and relevant contextual factors, allowing these round-level distributions to be aggregated into tournament-level outcome distributions. Round-level scores contain sufficient information about player ability while avoiding unnecessary complexity, and align naturally with the structure of stroke play tournaments, which are played and scored sequentially over rounds. Hole-level modelling will not be considered, as individual hole outcomes are highly dependent on hole-specific characteristics, and small modelling errors at hole level can accumulate and amplify when summed up to round level.

The objective is to estimate the distribution of a player's performance compared to an average "field" (set of players participating in a round) on the tour at a given point in time. To formalise this objective, we define a player's round-level performance relative to the field during a round

$$Y_{p,i} = S_{p,i} - \frac{1}{|P_i|} \sum_{q \in P_i} S_{q,i}, \tag{1}$$

where $S_{p,i}$ denotes the total score of player p during round i, and P_i is the set of players participating in the round.

In this work, a player's round-level score (e.g., 70 strokes) is assumed to consist of four components: a latent skill term, contextual adjustments, course difficulty and an idiosyncratic error component, as follows

$$S_{p,i} = \mu_{p,i} + \Delta C_{p,i} + \delta_i + \epsilon_{p,i} \ . \tag{2}$$

This decomposition follows the established approach of separating skill from course difficulty [3, 4]. The time-varying latent skill term $\mu_{p,i}$ describes how well a player is expected to play against an average field on the Tour. The contextual adjustments $\Delta C_{p,i}$ consists of any adjustments made for player performance, such as weather or course suitability. The course difficulty term δ_i describes how easy or hard the course was to play on a specific round. However, since the objective is to model performance relative to the field (and ultimately, to compare players with each other), rather than absolute scoring levels, this term is irrelevant in the scope of this study. The course difficulty is assumed to affect all players equally within a round, and any player-specific deviations from common conditions can be captured through $\Delta C_{p,i}$. By inserting (2) into Equation (1), the summation simplifies to

$$\frac{1}{|P_i|}\sum_{p\in P_i} S_{p,i} = \bar{\mu}_i + \delta_i + \bar{\epsilon}_i, \tag{3}$$

where

$$\bar{\mu}_i = \frac{1}{|P_i|}\sum_{p\in P_i} \mu_{p,i}, \qquad \bar{\epsilon}_i = \frac{1}{|P_i|}\sum_{p\in P_i} \epsilon_{p,i}\ .$$

This cancels the δ_i term from (1), and simplifies it to

$$Y_{p,i} = (\mu_{p,i} - \bar{\mu}_i) + \Delta C_{p,i} + (\epsilon_{p,i} - \bar{\epsilon}_i)\ . \tag{4}$$

Note that the contextual adjustments here are defined to average to zero across the field. The main goal will be to find an appropriate estimate for the player's time-dependent ability μ, as well as the residual distribution ϵ, as these components are expected to have the most explanatory power in forecasting tournament outcomes. For simplicity, we will neglect the $\bar{\epsilon}_i$ term. If residuals are assumed to have constant variance (σ^2) across players and to be independent within a round, then the standard deviation of $\bar{\epsilon}_i$ is $\sigma/\sqrt{n}$ (where $n = |P_i|$). This term (or its variation) is insignificant compared to other sources of variation in round-level scores.

The estimate of player ability for each round is constructed from the player's prior results, which provide the available information about performance level. Here, the essential part is to find a natural balance in weighting between most recent rounds and the ones further in history. Once we have a sufficiently good estimate of player skill, we are able to examine the deviations of the expected and observed performances. By construction, the residuals are expected to exhibit slight right-tail skewness, as exceptionally good scores are naturally bounded, whereas poor scores can be substantially larger. For example, players with consistently higher driving accuracy could be expected to have less deviation in round scores than players with lower driving accuracy.

Although player ability and the players' variance around their ability are expected to provide most of the explanatory power, additional factors may also influence outcomes. This includes weather conditions, which affect the performance of the players. In general, players are expected to perform worse relative to their ability in windy conditions than in calm conditions [6]. Since players complete their rounds at different times of the day, they may experience different weather conditions, which can lead to differences in scoring that are unrelated to their average ability. Furthermore, players may also differ in how sensitive they are to differences in weather conditions.

Certain courses may suit some players better than others. For example, players with lower driving accuracy may be expected to perform worse on a narrow course surrounded by trees, where inaccurate tee shots are penalised more. Alternatively, courses where the majority of holes curve towards to the left or right may benefit the player who dominantly curves the ball in the same direction. Therefore, in theory,

we should not (heavily) decrease a player's underlying ability if they perform badly on a course which does not suit them. However, unlike weather variables, which tend to affect player performance in a similar direction, the course setup affects players differently and is harder to quantify. In addition, the same courses recur from season to season, which means that during one season the average adjustment should approximately average to zero (players cannot always blame the course).

Psychological pressure may influence player performance, particularly in situations where players compete for top finishing positions. Hickman and Metz provide empirical evidence that putting performance is affected by the monetary stake involved [7]. In the current study, we do not model pressure effects, but instead mitigate them by focusing on the top-n finishing positions rather than the winner alone. If pressure effects were to be modelled, they would need to be player-specific. A pressure effect common to all players would not affect relative performance, as the eventual winner must necessarily overcome the pressure in order to win.

3 Data

3.1 Fundamentals of golf

A **round** of golf is played on a golf course consisting of 18 different holes, and is considered as the primary unit of play. A player's **round score** is the total number of strokes taken across those 18 holes. A golf **hole** generally consists of a teeing area, fairway, rough, green and a cup (hole), where the aim is to use as few strokes as possible to get the ball from the teeing area into the cup. A **stroke** refers to any attempt to move the ball using a golf club. In addition, certain rule infractions or relief situations add a **penalty stroke**, which increases the player's score for the hole without an additional physical swing. A player's **hole score** is the total number of strokes taken to complete the hole, including any penalty strokes. Each hole has a certain par score, which describes the reference number of strokes required by a proficient golfer to complete the hole, and varies from 3 to 5 depending on hole length and difficulty. Teeing area is a predefined spot where a player starts the hole by playing their first stroke (called teeing off). The fairway is a short-cut grass area between the teeing area and the green, typically present on par 4 and par 5 holes, and is the preferred landing area for most shots when the green is unreachable. A green is a very fine-cut grass area where the ball is rolled towards the cup using a putter.

The teeing area, fairway, and green make up the designed playing route for the hole, which is often surrounded by rough and various hazards. Rough is longer-cut grass which makes the impact of the golf club and ball more unpredictable, thus making the shot more difficult for the player. Hazards are areas designed to make the hole more difficult, such as bunkers, water hazards, and out-of-bounds, and often force the player to take a penalty stroke or play from a worse position. Figure 1 illustrates the 18th hole of Pebble Beach, an iconic course that has hosted several significant tournaments.

Figure 1: Image of Pebble Beach's 18th hole (par 5). Labels: 1 teeing area, 2 fairway, 3 green (where the cup is located), 4 rough, 5 bunker (hazard), 6 water (hazard), and 7 out-of-bounds area (hazard).

3.2 Golf tournaments

A golf tour is essentially a league that organises and runs a season of professional golf tournaments under the same rules. There exist multiple golf tours across the world,

targeting different geographical areas and skill levels of players, with the PGA Tour widely regarded as the highest level of professional men's golf in terms of playing standard.

Each year the PGA Tour organises a season, which consists of multiple tournaments played in different locations of the United States. While players accumulate points across the season based on their tournament performances, competitive outcomes and recognition are primarily associated with individual tournaments rather than season-long standings.

A golf tournament is an event in which a "field" (set) of players competes over multiple rounds under a defined competition format. In stroke-play tournaments, the most common form of competitive golf, the winner is determined by the lowest cumulative score across the rounds played. The size of the field (number of players) at PGA Tour events is generally around 140 players, although this varies between tournaments depending on event type. The number of rounds played in a tournament is four, with some exceptions occurring due to weather interruptions or other unexpected circumstances. Traditionally, all four rounds of a tournament are played on the same course, with slightly different setups by varying teeing area and cup locations. However, in some tournaments play is distributed across multiple courses (up to three), with each player completing one round per day on different courses during the initial days of the tournament. Most of the tournaments exhibit a cut, where part of the field is eliminated after the first two rounds. The cut line is predefined as top n players and ties, who continue to the final rounds. In tournaments played across multiple courses, the cut rules may differ from standard events, and also no-cut tournaments exist where all players play all four rounds.

The play is completed in groups of two and three players, depending on tournament and schedule of play. The rounds before the cut (generally first two rounds), are played in groups predetermined by the tour. These rounds are scheduled so that each player competes once in an early tee-time and once in a late tee-time. The final rounds' tee-times are ordered by leaderboard position, such that the tournament leader starts last.

3.3 Scoring data

The main dataset used in this study is collected from the PGA Tour's website, and contains over 500 tournaments played from season 2015 to 2025. The dataset includes all stroke-play tournaments held during the period, with the exception of certain team-based events and tournaments which are co-sanctioned with another tour.

The raw dataset contains hole-level scoring data, meaning that for every tournament during the seasons, we have the observations of each player's score on every hole of every round, totalling over 3.5 million observations. While hole-level outcomes are inherently noisier than round-level aggregates, they provide additional information that may help identify player-course specific effects. Table 1 shows an example of what the hole level data looks like. The data concerns Davis Love III's first round at

the 2015 Frys.com Open. Here, roundId identifies the tournament in which the hole is played.

roundId	playerId	roundNo	holeNo	score	par	yardage
R2015464	01706	1	1	4	4	433
R2015464	01706	1	2	3	3	198
R2015464	01706	1	3	5	4	411
R2015464	01706	1	4	3	4	408
R2015464	01706	1	5	5	5	537

Table 1: Example of hole-level data.

Since the aim is to model round-level scores, it is beneficial to aggregate the hole-level scores into round-level scores to enhance structural and computational efficiency. Table 2 shows data from the same tournament as the hole-level example, but aggregated at round level. The table contains first-round scores for Davis Love III, Scott Verplank, Vijay Singh, Jerry Kelly, and Tom Gillis. The column totScore shows the total score (number of strokes) for the round, while totPar is the par for the course.

playerId	roundId	roundNo	courseId	date	totScore	totPar
01706	R2015464	1	552	2014-10-09	77	72
02239	R2015464	1	552	2014-10-09	73	72
06567	R2015464	1	552	2014-10-09	72	72
08075	R2015464	1	552	2014-10-09	69	72
08725	R2015464	1	552	2014-10-09	70	72

Table 2: Example of round-level data. Note that the 2015 season started during late 2014, and that the dates and season years might not match.

3.4 Strokes gained and statistics data

Strokes gained is a measure used to quantify a player's performance relative to the field average. The framework was originally introduced by Broadie (2012) [2] and has become a standard tool in golf analytics, allowing performance to be decomposed into specific components, such as off-the-tee, approach, around-the-green and putting, reflecting different aspects of player skill. Strokes gained for a single stroke is defined as

$$\mathrm{SG}_i = J(d_i, c_i) - J(d_{i+1}, c_{i+1}) - 1$$

where d_i denotes the location and c_i the condition (such as fairway, rough, or bunker) of shot i, $J(d_i, c_i)$ the expected number of strokes required to hole out from location d_i

and condition c_i, and the -1 term represents the stroke used in the process. In practice, the location parameter d describes the distance to the pin, although theoretically it could be coordinates as well (assuming more complete data). The strokes gained for a category C is calculated by summing the strokes gained of the shots within that category

$$\mathrm{SG}_C = \sum_{i \in C} \mathrm{SG}_i \ .$$

The PGA Tour divides the strokes into four different categories: off the tee, approach shots, around the green and putts, and publishes the statistics aggregated on a round level where the underlying data is available. The category does not strictly relate to the condition c_i and location d_i, as it describes the type of shot rather than the location or condition from which it is played. To illustrate this, let's consider an example where a player hits a tee shot from 400 yards, and the shot ends up on the fairway 80 yards from the pin. The expected strokes needed to hole the ball from the positions equal $J(400,$ Teeing Area) and $J(80,$ Fairway). If we suppose that $J(400, \text{Teeing Area}) = 3.9$ and $J(80, \text{Fairway}) = 2.75$, the strokes gained for the tee shot is then

$$\mathrm{SG}_i = 3.9 - 2.75 - 1 = 0.15 \ ,$$

meaning that the player gained 0.15 strokes from the shot. If the ball had instead ended in the rough at the same distance from the pin, and $J(80, \text{Rough}) = 2.95$, the strokes gained for the shot would have been -0.05, meaning that the player would have lost 0.05 strokes relative to the field by hitting it into the rough.

The PGA Tour estimates the expected strokes function J using empirical data including millions of strokes recorded on the tour. Although it follows the framework proposed by Broadie, the exact underlying methodology is not publicly disclosed. The Tour also normalises the strokes gained statistics for each round, such that the average strokes gained of the field is zero. This normalisation is important for the categorical metrics, as it improves comparability. Without normalisation, rounds played on courses with wide fairways would inflate off-the-tee performance metrics.

Table 3 shows an example of the strokes-gained data, which is available for only a part of the rounds in the dataset. The variables sgOTT, sgAPP, sgARG, and sgPUTT denote strokes gained off the tee, on approach shots, around-the-green, and putting, respectively, and their sum equals sgTOT, which equals the difference between the player's score and the average score during the day (same as Equation (1)).

playerId	roundId	roundNo	sgOTT	sgAPP	sgARG	sgPUTT	sgTOT
25493	R2015002	1	1.04	1.89	-0.35	1.50	4.08
25686	R2015002	3	0.24	-1.22	0.18	2.02	1.22
35541	R2015002	2	1.02	-0.16	-1.10	1.05	0.81
35541	R2015002	4	-1.85	2.16	0.44	-0.57	0.18
29535	R2015002	2	0.47	-0.42	1.45	0.30	1.81

Table 3: Strokes-gained metrics.

Figure 2 shows the 50 round rolling average of the categorical values for the current (as of March 2026) world number 1 and 2 players Scottie Scheffler and Rory McIlroy. Scottie Scheffler's main strengths are his superior approach play and off-the-tee performance, while putting has been his relative weakness in the past. Rory McIlroy, on the other hand, is particularly strong off the tee, while also having a more all-around profile with no clearly observable weakness.

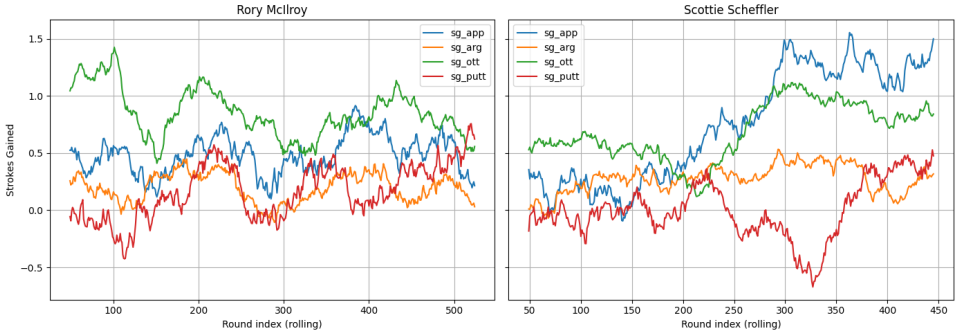


Figure 2: Fifty round moving average for the strokes gained category values for Rory McIlroy and Scottie Scheffler.

While the strokes-gained data is available for only a part of the tournaments, we have traditional statistics for all rounds, including driving accuracy, average driving distance, green in regulation (GIR) percent, scrambling percent and putts per GIR. Green in regulation means reaching the green in two strokes fewer than par, leaving two putts for par. Driving accuracy is measured by the percentage of tee shots finishing on the fairway. An example of this data is shown in Table 4. The expected use-case for this is to identify player driving distance and accuracy, which can be used to estimate how well a certain course suits a player.

playerId	roundId	roundNo	drvAcc	drvDist	GIR	scrambling	puttsGIR
25493	R2015002	1	0.77	291.0	0.72	0.80	1.54
25686	R2015002	3	0.69	276.5	0.61	0.71	1.64
35541	R2015002	2	0.85	287.5	0.61	0.57	1.82
35541	R2015002	4	0.54	284.0	0.72	0.80	1.62
29535	R2015002	2	0.46	275.5	0.72	0.80	1.77

Table 4: Round level statistics data.

3.5 Weather data

To adjust player performances for weather conditions, hourly meteorological data were retrieved from the Visual Crossing Weather API [8]. To construct the weather dataset, all tournament courses were manually mapped to their geographic coordinates, as the course locations are not provided in the main dataset. In addition, official tee-times

were collected for every round and converted into Coordinated Universal Time (UTC), to ensure consistency across timezones.

By estimating the daily time window of play through the tee-times and combining it with the mapped course coordinates, weather data were retrieved separately for each tournament. The weather data includes temperature (°C), wind speed (kph), wind gust (kph), wind direction (deg), precipitation (mm), pressure (hPa) and humidity (%). Although the majority of variables are nearly complete, the wind gust variable contains missing values in approximately 40% of observations. These missing entries are due to unavailable measurements or the absence of noticeable gusts during the respective hour (according to Visual Crossing's documentation).

Table 5 shows an example of what the weather data looks like for an individual round. This is an example and does not show all columns of the full dataset where course coordinates and precise timestamps are also included.

	datetime	temp	windspeed	windgust	winddir	precip	pressure	humidity
14	14:00	15.00	3.80	N/A	279.00	0.00	1012.60	81.43
15	15:00	15.90	8.50	N/A	200.00	0.00	1012.60	77.22
16	16:00	17.30	9.40	N/A	345.00	0.00	1012.90	72.62
17	17:00	17.70	10.20	N/A	335.00	0.00	1013.20	74.66
18	18:00	19.10	10.40	N/A	67.00	0.00	1013.50	68.67

Table 5: Example data of hourly weather observations.

4 Simulation and validation methods

4.1 Simulating tournament outcomes

In this work, we model the player's expected relative score in round i of tournament t as

$$\hat{Y}_{p,t,i} = \hat{\mu}_{p,t} + \Delta C_{p,t,i} + \varepsilon_{p,t,i},$$

where $\hat{\mu}_{p,t}$ is player p's estimated skill prior to tournament t, $\Delta C_{p,t,i}$ is the contextual adjustment term, and $\varepsilon_{p,t,i}$ is the residual term modelling the round-level variation in the player's performance. The distribution of the residual term is assumed to remain the same between rounds, but a new realisation is drawn for each round. The shape of the residual distribution will be addressed in Section 5.2, where both normal and skewed distributions are considered. We assume no autocorrelation between the round-level scores, as it simplifies the simulation, and also because no strong evidence of autocorrelation was found in the study conducted by Connolly and Rendleman (2008) [4].

All simulated tournaments consist of four rounds played over four consecutive days. Thus, each tournament has been assigned a cut rule defined by three variables: whether the event has a cut, the round number after which the cut is applied, and the cut threshold defined as the top n players and ties. These variables were inferred from the dataset together with available knowledge of tournament formats, and filled in manually where necessary.

Let CR_t be the cut round and CN_t be the cut threshold for tournament t. For tournaments without a cut, we set $\mathrm{CR}_t = 4$, meaning that all players advance to the final round. Using these variables, we can simulate the pre-cut and post-cut round scores (against the field average) for each player as

$$S_{p,t,\mathrm{pre}} = \sum_{i=1}^{\mathrm{CR}_t} (\hat{\mu}_{p,t} + \Delta C_{p,t,i} + \varepsilon_{p,t,i})$$

$$S_{p,t,\mathrm{post}} = \sum_{i=\mathrm{CR}_t+1}^{4} (\hat{\mu}_{p,t} + \Delta C_{p,t,i} + \varepsilon_{p,t,i}) \, .$$

Our model treats round-level scores as continuous. Since actual scores are whole numbers, and we need to account for ties (after cut round and at end of tournament), we round the simulated scores to the nearest integer. This introduces small rounding errors, and to minimise the possible distortion, we round the scores only twice (or only once if there is no cut) as follows

$$\tilde{S}_{p,t,\mathrm{pre}} = \mathrm{round}\big(S_{p,t,\mathrm{pre}}\big) \, ,$$
$$\tilde{S}_{p,t,\mathrm{post}} = \mathrm{round}\big(S_{p,t,\mathrm{post}}\big) \, .$$

After this, we calculate the ranks for all players pre-cut, and calculate the scores for the whole tournament

$$\tilde{S}_{p,t} = \begin{cases} \infty, & \text{if } \operatorname{rank}\big(\tilde{S}_{p,t,\text{pre}}\big) > \text{CN}_t, \\ \tilde{S}_{p,t,\text{pre}} + \tilde{S}_{p,t,\text{post}}, & \text{if } \operatorname{rank}\big(\tilde{S}_{p,t,\text{pre}}\big) \leq \text{CN}_t. \end{cases}$$

We repeat this simulation multiple times, such that the probabilities for the outcomes converge. After the simulations, we calculate the outcomes for each simulation j and calculate the number of players who have outscored the player $B_{p,t,j}$ and the number of players who are tied with the player $T_{p,t,j}$

$$B_{p,t,j} = |\{q : \tilde{S}_{q,t,j} < \tilde{S}_{p,t,j}, q \in P_t\}|,$$
$$T_{p,t,j} = |\{q : \tilde{S}_{q,t,j} = \tilde{S}_{p,t,j}, q \in P_t\}|,$$

which are crucial to decide how much we award players based on their performance. The outcome for player finishing among the top n_{top} in simulation j will be evaluated using

$$I_{p,t,j}(n_{\text{top}}) = \begin{cases} 1, & B_{p,t,j} + T_{p,t,j} \leq n_{\text{top}} \\ \dfrac{n_{\text{top}} - B_{p,t,j}}{T_{p,t,j}}, & \text{if } B_{p,t,j} < n_{\text{top}} \text{ and } B_{p,t,j} + T_{p,t,j} > n_{\text{top}} \\ 0, & \text{else.} \end{cases} \tag{5}$$

Using this metric, a player receives a value of 1 if the player and all tied players are ranked within the top n_{top} positions. If only part of the tied group falls within the top n_{top}, the value is split equally among the tied players. Otherwise the metric will be zero. This ensures that the sum of the probabilities of players finishing top n_{top} equals n_{top}.

Now, by running N_{sims} simulations, we can estimate the probability for player p finishing top n_{top} in tournament t as

$$\text{Pr}_{p,t}^{(n_{\text{top}})} = \frac{1}{N_{\text{sims}}} \sum_{j=1}^{N_{\text{sims}}} I_{p,t,j}(n_{\text{top}}) \ .$$

4.2 Validation metrics

The simulated probabilities will be evaluated by comparing them to the observed outcomes. To construct the observed outcomes, we will first reconstruct the leaderboards for each tournament. From the leaderboards, we can use the same metric as introduced in (5) to indicate whether the player finished in a top n_{top} place in the tournament. Let $O_{p,t}(n_{\text{top}}) \in [0, 1]$ be the observed outcome for player p finishing top n_{top} in tournament t. Respectively, for each outcome, we have the simulated probability

$\Pr_{p,t}^{(n_{\text{top}})}$. The main measure for the loss for a probability estimate given an observed outcome will be measured using negative log-likelihood (NLL)

$$\text{NLL}_{p,t}^{(n_{\text{top}})} = -[O_{p,t}(n_{\text{top}}) \ln(\Pr_{p,t}^{(n_{\text{top}})}) + (1 - O_{p,t}(n_{\text{top}})) \ln(1 - \Pr_{p,t}^{(n_{\text{top}})})]$$

Note that NLL is equivalent to log loss, and since it is positive, we are aiming to minimise it. In practice, the estimated probabilities are clipped to $[\epsilon, 1 - \epsilon]$, with $\epsilon = 10^{-6}$, to avoid infinite loss from outcomes with estimated probability zero or one.

In most tournaments, we have a share of players whose estimates are constructed using the estimate for players with insufficient data, explained in section 5.1.3. These players will be included in the simulations as dummy players, but will not be used in the evaluation of the model, as we are only interested in modelling the players who have sufficient amount of data. In addition, tournaments where the share of these players exceeds 25% will be excluded from the results. This mainly excludes qualifying events for the tour and some individual tournaments that take place concurrently with larger tournaments, resulting in weaker fields. Players who have withdrawn or have been disqualified during the tournament will not be included in the simulations.

We focus on measuring the NLL for top-5, top-10, and top-20 finishes, as well as making the cut. These events occur substantially more often than tournament wins, resulting in a larger number of observed outcomes and therefore a more stable evaluation of the predicted probabilities.

For an individual tournament, the average NLL for the top n_{top} estimates is calculated as

$$\text{NLL}_{t}^{(n_{\text{top}})} = \frac{1}{|P_t|} \sum_{p \in P_t} \text{NLL}_{p,t}^{(n_{\text{top}})} ,$$

where P_t is the set of players included in the evaluation for tournament t.

The total average NLL for the top n estimates for a set of tournaments T is calculated as

$$\text{NLL}^{(n_{\text{top}})} = \frac{1}{\sum_{t \in T} |P_t|} \sum_{t \in T} |P_t| \text{NLL}_{t}^{(n_{\text{top}})} .$$

5 Basic Model

The basic model used for explaining the problem will be a simple model, where we try to model the players round performances using only their previous scores as an input. To start solving the problem, we calculate the strokes gained $Y_{p,i}$ for each player and round, according to the Equation (1). By strokes gained versus field we mean how many strokes the player's score deviates from the field average on the round. Recall that by field, we mean the set of players participating in a round of play. Negative values indicate good performance while positive values indicate below average performance. The data was split into three periods: a burn-in period (2015–2017) used to initialise the skill and variance estimates, a fit period (2018–2022) used for calibrating the model, and a test period (2023–2025) for out-of-sample evaluation.

Figure 3 shows the distributions of Y for selected players during 2021. The titles of the subplots include the mean, standard deviation and p-value for normality calculated using the Shapiro-Wilk test. The selected players are well-known high-performing players on the tour, which is reflected in their performances, as all of the players have played on average better than the field. In addition, the plots suggest that the distributions are approximately normal, although the sample sizes are relatively small.

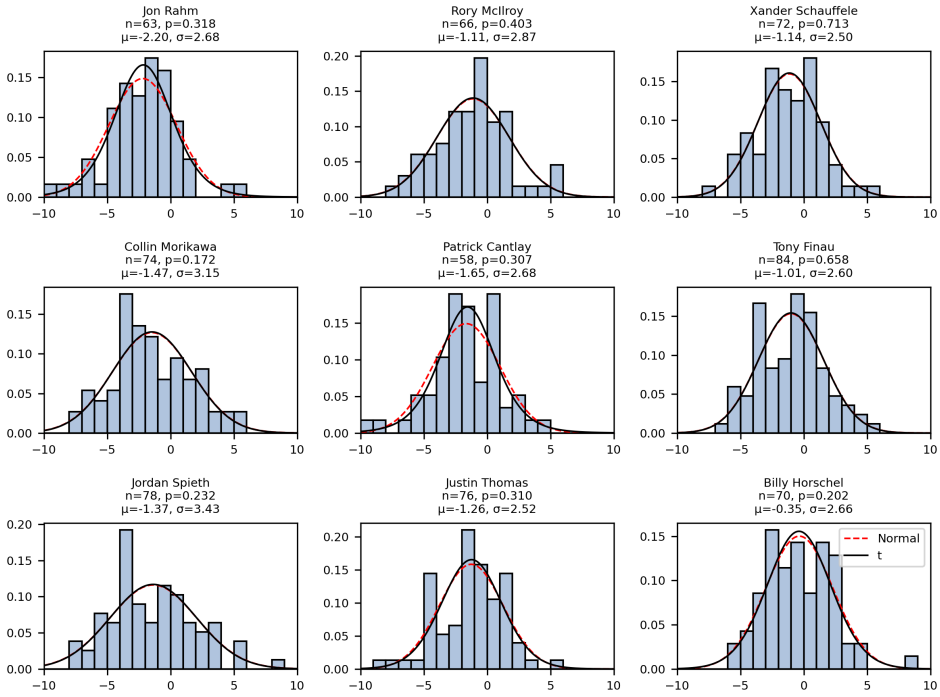


Figure 3: Distribution of Y for selected players during the year 2021.

5.1 Estimating location

Figure 4 shows Rory McIlroy's strokes gained versus field during seasons 2018 to 2022, along with the distribution of the respective metric grouped by year. The

time-ordered round-level observations in the left subplot represent the individual data points from which player ability is estimated.

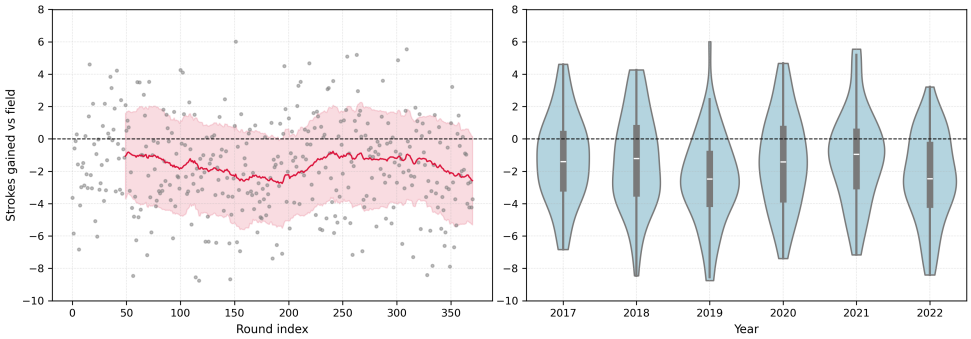


Figure 4: Rory McIlroy's scores versus field average during seasons 2018-2022. The dots represent individual rounds, the red line the 50-round moving average, and the shaded area the bounds of one standard deviation.

By inspecting the plot we see that the round-level scores exhibit substantial variability. Thus, creating a precise estimate of a player's single round performance is practically impossible. However, over a longer sample the underlying performance level is evident. In McIlroy's case, the distribution of observations is clearly centred towards negative values, implying that he outperforms the field more often than not.

5.1.1 Illustrative example of importance of ability estimate

It is more important to get an accurate estimate of player's ability $\mu_{p,t}$ than their volatility $\sigma_{p,t}$. To illustrate this, we define players A and B such that player A is one stroke better per round, while both players share a standard deviation of 2.5 strokes per round, reflecting a typical level of round-to-round variation in performance. Their performance distributions are assumed to be normal.

$$Y_{A,i} \sim N(\mu_{A,i} = -1, \sigma_{A,i} = 2.5)$$
$$Y_{B,i} \sim N(\mu_{B,i} = 0, \sigma_{B,i} = 2.5)$$

Figure 5 shows the sensitivity of the probability of player A beating player B over four rounds to changes in player B's parameters. The round level scores are in this example treated as continuous for simplicity and avoiding the handling of ties. The local sensitivity of the parameters is approximately 19 percentage points per stroke per round for μ and 3.8 percentage points per stroke per round for σ, making the sensitivity to μ five times greater than that to σ in this example.

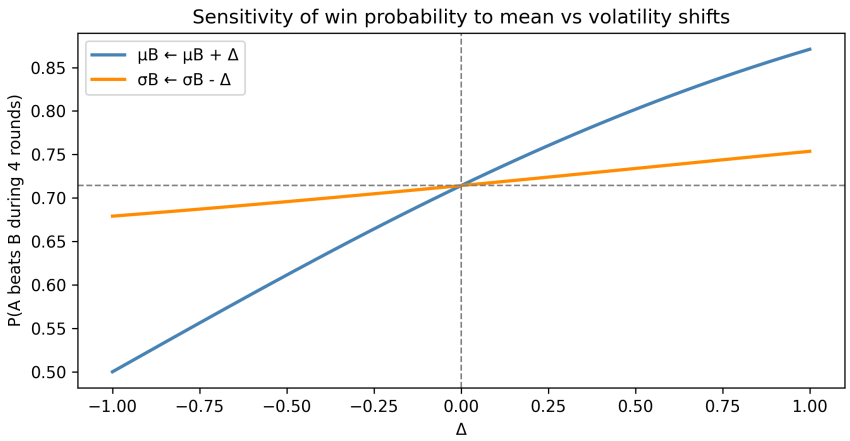


Figure 5: Sensitivity of winning probability to changes in distribution parameters.

5.1.2 Field strength adjustment

The problem with looking at raw strokes gained against field is that they are not normalised against field strength. By field strength, we mean the average skill of the players participating in the round. Field strength varies across rounds and tournaments, depending on which players are competing. The tournament schedule is structured into events of varying competitive tiers, each requiring a different qualification category or world ranking status. Consequently, lower-ranked players often do not qualify for higher-tier tournaments, which attract top-ranked players due to larger prize purses and higher ranking points. This tiered access structure results in systematic variation in field strength across events. Furthermore, most tournaments include a cut after two rounds, where only players within a predefined position limit, such as the top 65 and ties, continue to the final two rounds. As a result, roughly half of all rounds are played by a subset of entrants who have already outperformed the remainder of the field during first rounds. The data support this pattern, showing that the scoring average during rounds 3—4 is on average 0.47 strokes lower than during the first two rounds.

To adjust for the differing strengths of field, we normalise the score versus field as

$$Y_{p,i}^{\text{adj}} = Y_{p,i} + \frac{1}{|P_i|} \sum_{q \in P_i} \mu_{q,i} \ , \tag{6}$$

where P_i is the set of players participating in round i. This way Y^{adj} will describe how well the player performed on an individual round against an average field on the tour. However, this introduces a circular dependency, as the field-strength-adjustment requires estimates of player ability, while these ability estimates are themselves constructed from the adjusted observations.

To solve this issue, the player-level skill estimates μ are first constructed using the unadjusted Y values. After the initial estimation, the procedure can be iteratively refined by recomputing Y^{adj} and updating the corresponding μ values until convergence.

However, this iterative requirement does not suit all modelling approaches. To mitigate this issue, we can precompute the field strengths using a simplified approximation of player ability, for example by taking the mean performance from a set of recent rounds. We expect this approach to provide a sufficiently accurate field adjustment while reducing computational complexity in case it is required. The method for adjusting the scores for field strength is explained in Algorithm 1.

Algorithm 1 Description of how the field strength adjustments are calculated

Inputs

Strokes gained versus field for each player $\mathbf{Y} = (Y_{p1}, \ldots, Y_{pn})$ such that $Y_p = (y_{p,1}, ..., y_{p,T})$ are chronologically ordered.

Initial window size T_0 (e.g. end of first season), such that we have reasonable amount of data to construct initial estimates of player skills.

1: **for** $t = T_0 + 1$ **to** T **do**

2: **for** the player p in observation t **do**

3: Calculate the skill estimates for players:

$$\mu_{p,t} = \mu\big(\{\, y_{p,\tau} : y_{p,\tau} \in Y_p,\ \tau < t \,\}\big)\,,$$

4: **end for**

5: Calculate

$$\bar{\mu}_t = \frac{1}{|\text{field}(t)|} \sum_{p \in \text{field}(t)} \mu_{p,t}$$

6: **for** the player p in observation t **do**

7: Adjust the player's observation with the estimated field strength, such that $Y_p = (y_{p,1}, ...y_{p,T_0}, y^{\text{adj}}_{p,T_0+1}, ..., y^{\text{adj}}_{p,t})$ until t.

$$y^{\text{adj}}_{p,t} = y_{p,t} + \bar{\mu}_t$$
$$Y_p(t) \ \leftarrow \ y^{\text{adj}}_{p,t}$$

8: **end for**

9: **end for**

Where

$\mu(\cdot)$ is the player skill estimator (linear combination of historical performances).

field(t) denotes the set of players who appear in the same round or event at time t.

$\bar{\mu}_t$ represents the estimated field strength at time t, calculated as the mean of the current skill estimates of all players in the field.

$y_{p,t}$ denotes the observation for player p at time t before adjustment, and $y^{\text{adj}}_{p,t}$ denotes the adjusted value.

5.1.3 Estimate for players with insufficient data

In many events, sponsors are allowed to invite a small number of players to participate, and some tournaments also have open qualifying. In addition, at the start of each season, new debutants enter the tour. This is visible in the data, shown in Figure 6, where the average share of players with missing data across tournaments is 17 %.

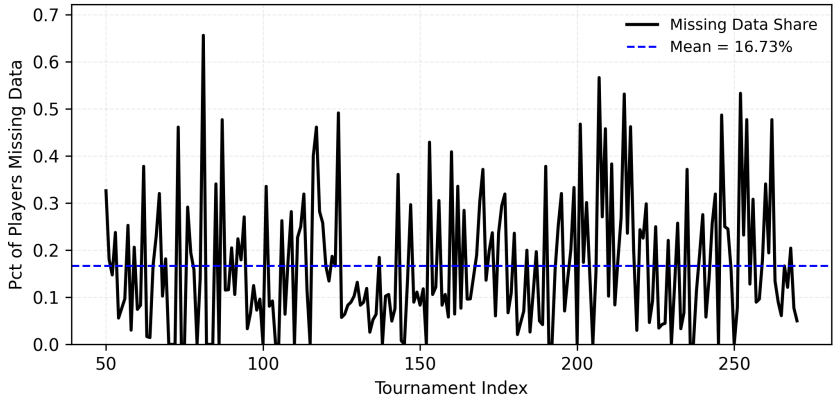


Figure 6: Share of players with missing data for different tournaments.

Thus, our dataset contains players with insufficient amount of data to create an estimate of their skill. To cope with this, we introduce an initial estimate for skill as a function of how many rounds they have played on the tour historically. While we are not interested in predicting the performance of these players, we still need to include them in the simulations, and also account for their contribution to the field strength.

We assume that a player making their debut will perform worse than a player who has completed 20 rounds at the tour. This assumption partly reflects the survivorship effect, as players who continue to compete on the tour have typically performed reasonably well. For players with insufficient round history, we define an exponential function $\mu_p^{\text{base}}(n_p)$ which describes the expected score of a player as a function of the number of historical observations

$$\mu_p^{\text{base}}(n_p) = a + b\, e^{-cn_p} \ . \tag{7}$$

The parameters of this function are estimated by regressing the empirical mean performance on the number of historical rounds per player, as shown in figure 7. The fitted parameters are $a = 0.34$, $b = 1.77$, and $c = 0.17$, implying that a debut-level player has an expected round-level performance of approximately 2.1, while for a player who has 20 observations the same estimate is 0.40.

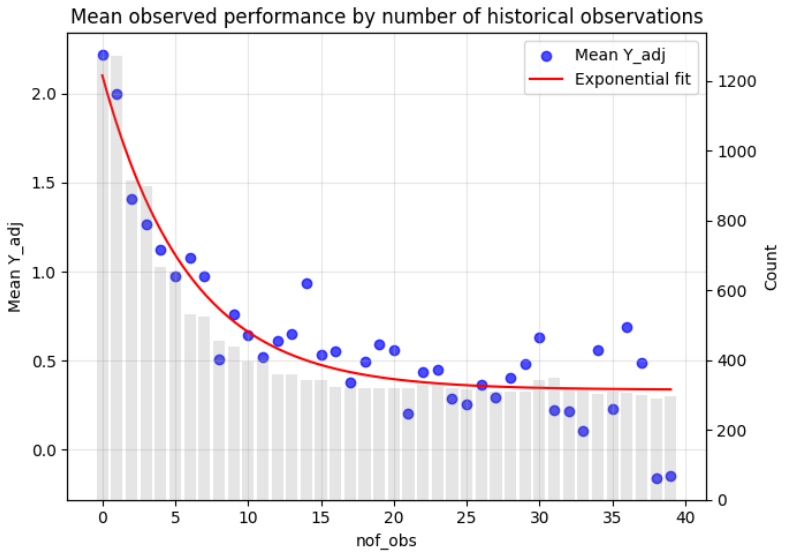


Figure 7: Fit for player ability as a function of number of observations in dataset.

Since most of the players which we label as having insufficient amount of data, actually have some data available, we combine the initial estimate with the existing player data.

The amount of data is limited for these players and therefore we use a traditional average of the previous scores instead of giving more weight to the most recent observations. After this, the two estimates can be combined using empirical Bayes shrinkage

$$\mu_p(n_p) = w(n_p)\,\bar{\mu}_p \;+\; \big(1 - w(n_p)\big)\,\mu_p^{\text{base}}(n_p)$$
$$w(n_p) = \frac{n_p}{n_p + \sigma^2/\tau^2}\,,$$

where $\bar{\mu}_p$ is the mean of player's observed performances (Y_{adj}), and σ^2 the within-player variance and τ^2 the between-player variance estimated from full dataset. Figure 8 shows the mean absolute errors for observation number bins, and shows that the empirical Bayes shrinkage is more efficient than using the baseline estimate, the player's sample mean or a simple n-weighted combination of these two, especially for the players with smaller sample size.

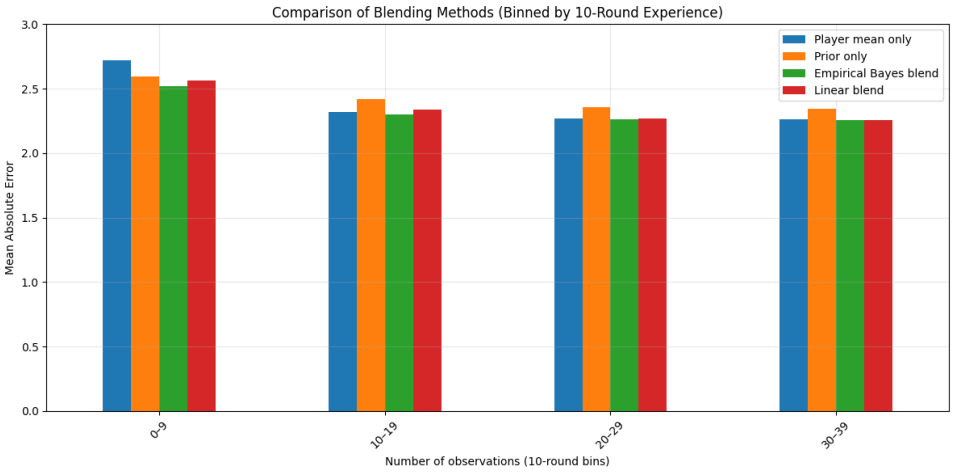


Figure 8: Mean absolute errors for players with insufficient data with different combinations of the base estimate and player mean performance.

Other variables which could enhance the initial estimate of player ability would include data from other tours which they have played, their world ranking, age and other accomplishments. However, modelling these players in detail is not the main focus of this work. We only require estimates that are sufficiently accurate to incorporate these players as dummy players in the dataset and simulations.

5.1.4 Weighting of past performance

We model a player's underlying skill as a weighted average of their previous performances. It is reasonable to assume that recently completed rounds are at least as informative as rounds played further in the past. This introduces a monotonicity constraint on the weighting scheme: the weights must decrease as a function of the time elapsed since the round was played.

To illustrate the weighting of historical rounds, we consider an illustrative regression problem. For each player p, the adjusted score $Y^{\text{adj}}_{p,t}$ is modelled as a linear combination of their last 150 adjusted scores $\{Y^{\text{adj}}_{p,\tau} : t - 150 \leq \tau < t\}$. By adjusted score, we mean strokes gained versus field, adjusted with the field strength, as explained in section 5.1.2. In this example, the skill estimator $\mu(\cdot)$, required for constructing the field strength adjustment, is defined as the arithmetic mean of the observations within the preceding two-year window.

The weights $w \in \mathbb{R}^{150}$ are obtained by solving the following optimisation problem:

$$\begin{aligned} \min_{w} \quad & \sum_{p \in P} \sum_{t=151}^{T_p} \left(Y^{\text{adj}}_{p,t} - \sum_{i=1}^{150} w_i \, Y^{\text{adj}}_{p,t-i} \right)^2 \\ \text{s.t.} \quad & w_i \geq 0, \quad \forall \, i = 1, \ldots, 150 \, . \end{aligned} \tag{8}$$

Note that for each player p, t indexes the t-th observed round in chronological order, rather than an absolute time point. Figure 9 shows the estimated weight vector w obtained using non-negative least squares (NNLS) and isotonic regression. Isotonic regression estimates the weights under the additional constraint that the weights are monotonous (non-increasing in this case) as a function of lag [9]. The implementation used Python's scikit-learn package [10]. We see that the most recent rounds contain the most information about the subsequent round.

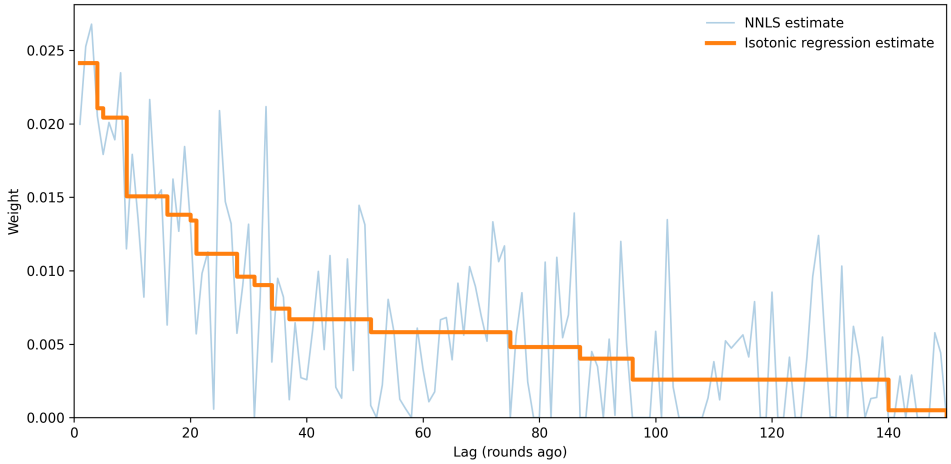


Figure 9: Optimal weights for the 150 most recent rounds, minimising next-round squared prediction error.

5.1.5 Player skill estimate

A natural choice for such a decaying weighting scheme is the exponentially weighted moving average (EWMA). This approach produces a smooth weighting function and provides a balance between recent and long-term information. The rate of decay is controlled by a single parameter, which will be fitted to the data.

In time-series applications, exponential moving averages are typically implemented recursively ($z_i = \lambda x_i + (1 - \lambda) z_{i-1}$) due to computational efficiency and minimal memory requirements. However, in this setting we normalise the weights so that they sum to one. Without normalisation, players with fewer rounds in their history would inherently receive smaller total weight than players with long histories. Any shrinkage of the estimated skill toward a prior mean is applied separately after constructing the EWMA estimate.

The exponential moving average to describe skill level is defined as

$$\hat{\mu}_{p,t} = \sum_{i=1}^{t-1} \alpha \, \lambda^{t-i} \, Y_{p,i}^{\text{adj}}, \quad \text{with } \alpha \text{ chosen such that } \sum_{i} \alpha \, \lambda^{t-i} = 1 \, ,$$

where $\lambda \in]0, 1]$ is the decay parameter and α a scaling constant such that the sum of the weights equal to 1.

To select an appropriate half-life parameter (half-life $= -\ln(2)/\ln(\lambda)$) for the model, we compute the mean absolute prediction error (MAE) for a range of half-life values and choose the parameter that minimises this error. We prefer MAE over the root mean squared error (RMSE) because the dataset contains some outliers (large deviations), and the squared error loss would place excessive weight on these observations. The field strength estimates are here computed using the corresponding skill estimates.

Figure 10 shows the mean absolute error (MAE) and root mean squared error (RMSE) of the estimates with different half-life parameters. We see that the error fairly quickly decreases and starts to stabilise at around 30 round half-life. In addition, the error surface becomes fairly flat once the parameter exceeds a certain threshold, indicating limited sensitivity to further increases. The results show that the MAE is minimised at a half-life of 38 and the RMSE at 36, which are consistent with each other, and we choose the model's half-life parameter to be equal to 38.

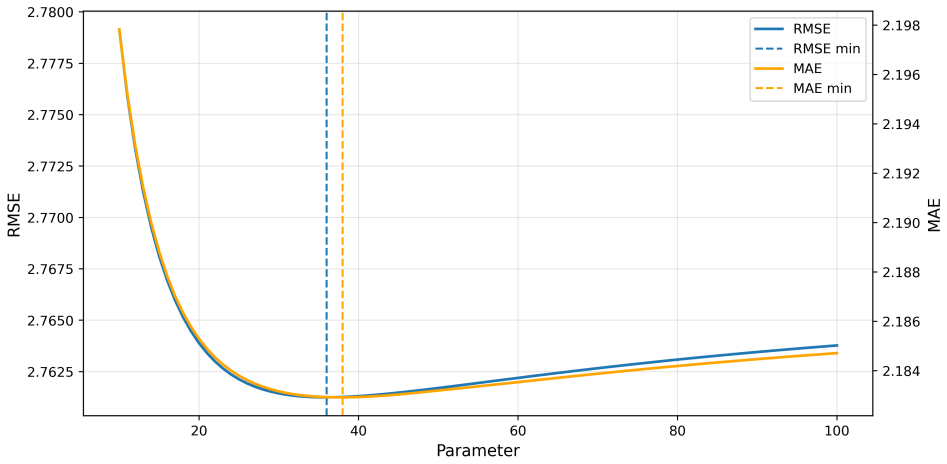


Figure 10: Errors for different half-life parameter of the EWMA estimate.

Next, we investigate whether the estimates should be shrunk towards the mean depending on the available sample size. This is particularly relevant when recent observations receive larger weights, and the weights are scaled such that they sum to one. The shrinkage target is set to zero, which corresponds to the average skill of a PGA Tour player at the beginning of the data series. A single population mean skill (a value for $\bar{\mu}$) is difficult to define, since the set of players in the data changes over time and stronger players appear more often. Thus, the estimated skills are shrunk (towards 0) and redefined as

$$\hat{\mu} \leftarrow \frac{n}{n+m}\hat{\mu}, \tag{9}$$

where n is the number of observations used to construct $\hat{\mu}$ and m is the shrinkage parameter. The parameter m is calibrated for the selected EWMA half-life and is not jointly optimised along with the half-life parameter. The choice of m has only a

minor effect on the results, and jointly selecting it together with the EWMA half-life is unnecessary in this context.

Figure 11 shows the mean absolute error (MAE) and root mean squared error (RMSE) evaluated across different values of m. The MAE is minimised between 2 and 3, while RMSE is minimised at $m = 4$. Based on the results, we choose $m = 3$ as the shrinkage parameter, as it is in between the two minimums.

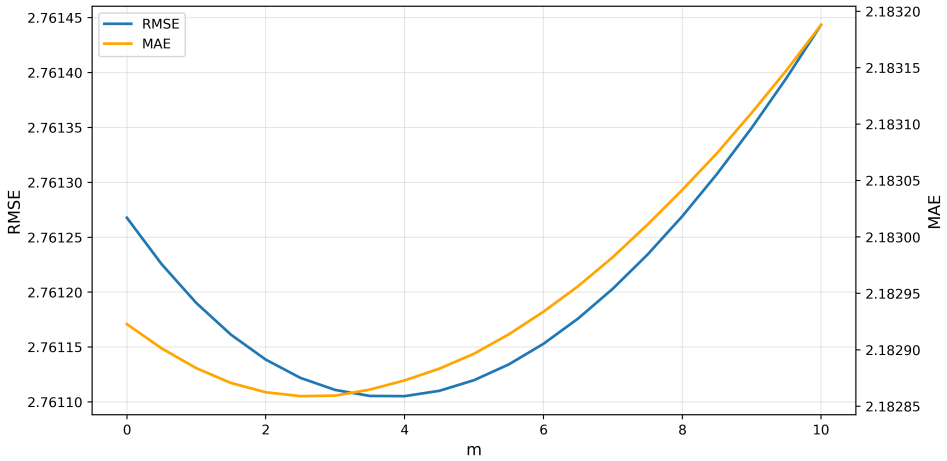


Figure 11: Errors for different values of the shrinkage parameter.

While shrinkage improves the stability of the estimates, accounting for field strength is particularly important for avoiding systematic prediction biases across different skill levels. To illustrate this, we compare the residuals of the model with and without field strength adjustment across skill levels. The unadjusted residuals are constructed using strokes gained relative to the field, without accounting for differences in field strength across tournaments and rounds in either the estimate or the observed outcome. In the adjusted version, both components are corrected for field strength using the estimates constructed in this section.

Figure 12 illustrates the average residuals as a function of estimated ability, grouped into one-stroke intervals. The model based on unadjusted scores systematically overestimates the performance of strong players and underestimates the performance of players whose ability is estimated to be below average. After adjusting for field strength, this bias decreases substantially, and the average residual is closer to zero across the different bins. The only exceptions occur in the lowest ability bins, where the number of observations is small.

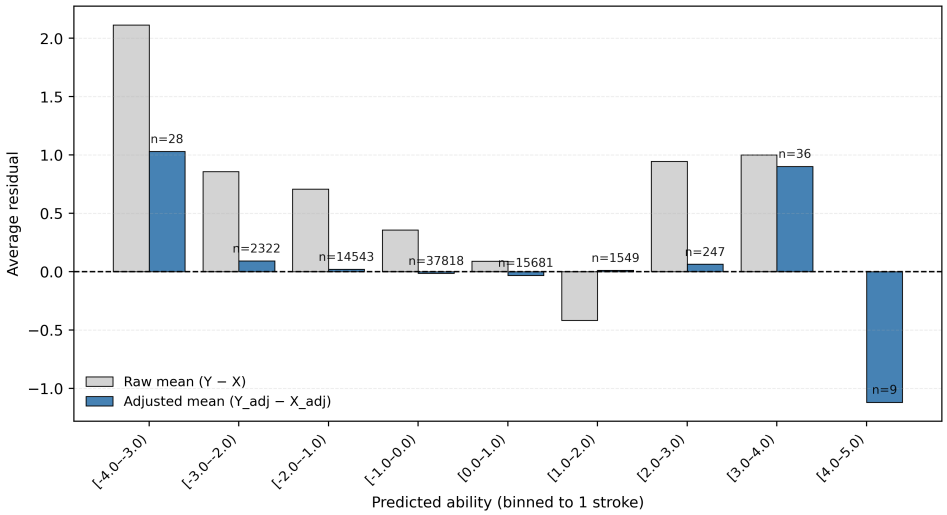


Figure 12: Systematic biases of predictions based on predicted skill ability.

Figure 13 shows the estimated field strengths for different rounds in the fit dataset. The average field strength is slightly below zero (negative field strength corresponds to a stronger field). This is mainly driven by the smaller-field events, as these typically consist of stronger players and therefore skew the average field strength downward. The downward spikes towards end of years are explained by the tournament schedule. During August–September, the PGA Tour holds its playoffs, where the field size decreases from tournament to tournament, such that only the top 30 ranked players qualify for the final event. In addition, the last stroke-play tournament of the year is the Hero World Challenge, consisting of 20 highly ranked players, and the following season begins with the Sentry Tournament of Champions, in which the previous year's winners are eligible to play.

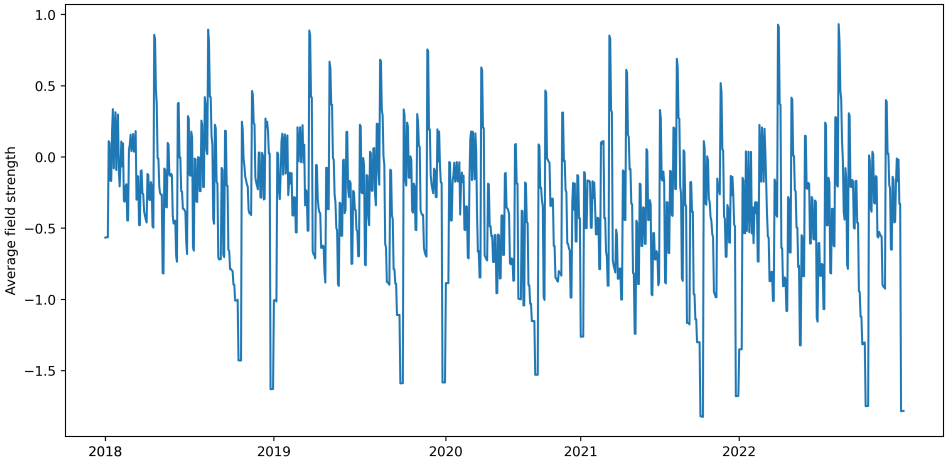


Figure 13: Estimated field strength for the rounds in the fit period.

5.2 Residual distribution

5.2.1 Empirical distribution of residuals

We define the residual for player p for round i as

$$\varepsilon_{p,i} = Y^{\text{adj}}_{p,i} - \hat{\mu}_{p,i} \tag{10}$$

where $Y^{\text{adj}}_{p,i}$ denotes the field strength adjusted observed score and $\hat{\mu}_{p,i}$ the estimated skill. The estimated skill is constructed using the exponentially weighted moving average with the half-life parameter 38, as explained in section 5.1.

The residuals are centred close to zero, indicating that the prediction model is approximately unbiased. The distribution exhibits positive skewness meaning a longer right tail, and a slight excess kurtosis, indicating slightly heavier tails than a normal distribution. These results, especially the skewness, are supported by the hypothesis that good rounds are physically bounded while bad rounds can be larger by deviation. Table 6 shows the descriptive statistics of the residuals.

Table 6: Descriptive statistics of residual errors.

Location and dispersion		Quantiles	
Observations	72,233	5%	-4.35
Mean	-0.006	25%	-1.89
Std. dev.	2.761	Median	-0.121
Minimum	-9.98	75%	1.76
Maximum	17.54	95%	4.70
Distribution shape		**Tail asymmetry**	
Skewness	0.271	$P(\varepsilon > 7.5)$	0.007
Excess kurtosis	0.317	$P(\varepsilon < -7.5)$	0.0018
		Right/Left ratio	4.01

The empirical probability of a player performing 7.5 strokes worse than the estimate $P(\varepsilon > 7.5)$ is four times as large as performing the same number of strokes better than the estimate. Therefore, using a symmetric distribution to model the residuals would overestimate the frequency of strong rounds compared to poor rounds. To mitigate this issue, we introduce a skew-normal distribution to model the residuals. This choice preserves model simplicity while allowing for asymmetry in the residuals. For comparison, we also use the symmetric normal distribution.

The skew-normal distribution is a modification of the normal distribution which allows skewness. Its probability density function is defined as

$$f(x;\, \xi, \omega, \alpha) = \frac{2}{\omega}\, \phi\left(\frac{x-\xi}{\omega}\right)\, \Phi\left(\alpha\, \frac{x-\xi}{\omega}\right),$$

where α is the shape parameter controlling skewness, ξ is the location parameter determining the centre of the distribution, and $\omega > 0$ is the scale parameter determining the width of the distribution. The function $\phi(\cdot)$ denotes the probability density function of the standard normal distribution and $\Phi(\cdot)$ its cumulative distribution function. [11]

The mean and variance of the distribution are given by

$$\mathbb{E}[X] = \xi + \omega\delta\sqrt{\frac{2}{\pi}}, \qquad \mathrm{Var}(X) = \omega^2\left(1 - \frac{2\delta^2}{\pi}\right) \tag{11}$$

where X denotes a residual and $X \sim SN(\xi, \omega, \alpha)$ with $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$.

Figure 14 shows the residuals fitted with a skew-normal and a normal distribution. By inspecting the plot, we see that the skew-normal distribution seems to model the data well and provides a visually better fit than the traditional normal distribution, as it matches the slightly negative mode better, and also models the asymmetrical tails more accurately. With large sample sizes, goodness-of-fit tests often become very sensitive, rejecting deviations that are statistically detectable but practically unimportant (at least in our use-case).

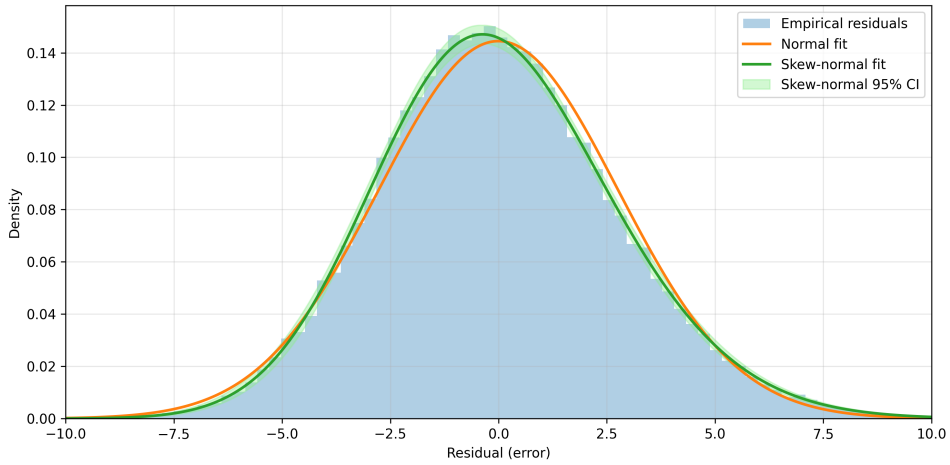


Figure 14: Empirical residuals with a normal fit and skew-normal fit.

Table 7 shows the estimated parameters for the skew-normal distribution and their confidence intervals. Skew-normal distribution does not have closed form solution for calculating the confidence intervals, so they were calculated by bootstrapping the samples 10 000 times. This distribution will be used as the global distribution for player residuals.

Table 7: Bootstrapped parameter estimates for the skew-normal distribution fitted to the residuals

Parameter	Estimate	2.5% CI	97.5% CI
Shape (α_g)	1.348	1.295	1.402
Location (ξ_g)	-2.309	-2.369	-2.248
Scale (ω_g)	3.595	3.551	3.639

5.2.2 Player specific variance and residual estimate

The variability of round-to-round performance is not constant across players. Figure 15 shows the residuals for high and low variance players in the dataset. The plots on the first row show residuals of the 20 players with the highest and lowest variances during each season grouped together by category. The plots on the second row shows the residuals of the same players during the subsequent season. The results imply that historical variance predicts future variance, but with only partial persistence, as the difference of the standard deviation between the groups shrinks from 0.93 to 0.19 between the in-sample and out-of-sample periods. While using a single global residual variance for all players may be preferable compared to relying solely on the noisy player-specific variance estimates, the results indicate that player-level variability is somewhat persistent.

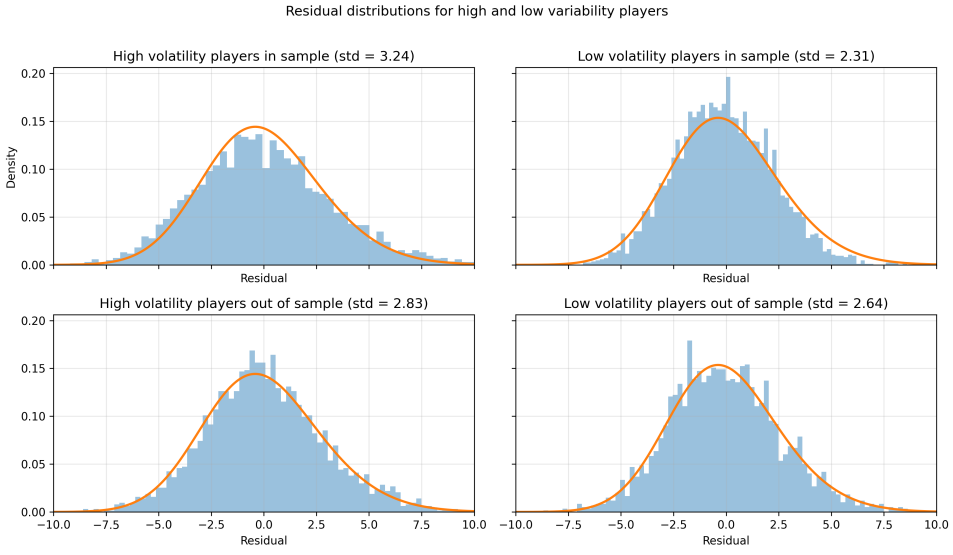


Figure 15: Residuals of players which are labelled as high variance and low variance.

To construct estimates for player specific variance, we combine the information about the global distribution and the player's historical residuals. Because large differences from the global average variance tend to revert toward the mean, the observed player

variance is shrunk toward the global variance. This prevents extreme estimates based on small samples while still allowing players with consistently high or low variability to differ from the field. The following linear shrinkage will be used

$$\hat{\sigma}^2_{p,i} = \frac{n\sigma^2_{p,i} + m\sigma^2_g}{m+n}\,, \tag{12}$$

where $\sigma^2_{p,i}$ is the player's empirical variance, n the sample size of player's empirical variance, m the weighting parameter, and σ^2_g the global variance.

The player-specific residual variance is estimated from the player's most recent rounds, using at most the latest 150 observations. Since player attributes and performance can change over time, older rounds may be less informative of the player's current level of variability. A window of 150 rounds corresponds to roughly two seasons for players who compete regularly, providing a balance between using enough data for a stable estimate while still focusing on recent history. Based on the data, this window length appears to be a reasonable choice and the results are not highly sensitive to the exact value, provided it is not too small or too large.

We try to use the skew-normal distribution to model the player-specific residual distributions. The shape parameter α is fixed at the value estimated from the global residual distribution, while player-level variance is modelled through the location and scale parameters.

Given a fixed skewness parameter α_g, the scale and location parameters ω and ξ can be selected to match a desired mean and variance, defined in Equation (11). Since the residuals are assumed to have mean zero and target variance $\hat{\sigma}^2_{p,i}$, we can set

$$\omega_{p,i} = \frac{\hat{\sigma}_{p,i}}{\sqrt{1-\delta^2\frac{2}{\pi}}}, \qquad \xi_{p,i} = -\,\omega_{p,i}\delta\sqrt{\frac{2}{\pi}}\,, \tag{13}$$

where $\delta = \frac{\alpha_g}{\sqrt{1+\alpha_g^2}}$.

Thus, the estimated residual distribution for player p at round i

$$\varepsilon_{p,i} \sim SN(\xi_{p,i}, \omega_{p,i}, \alpha_g)$$

will have a mean of zero and a variance of $\hat{\sigma}^2_{p,i}$, while following the skewness of the global residual distribution.

5.2.3 Validation

We have now constructed a framework for estimating player-specific residual distributions. The remaining task is to determine the shrinkage parameter m in (12), which controls the degree of shrinkage toward the global variance. Small values of m place

more weight on the player's observed variance, whereas large values assign more weight to the global estimate.

To select an appropriate value, the variance estimates are computed for a range of candidate values of m. For each candidate value, a skew-normal and a zero mean normal residual distributions are constructed using the corresponding variance estimates, and the performance of the resulting model is measured using negative log-likelihood (NLL) between the estimated distributions and observed deviations.

Figure 16 shows the NLLs for the estimated residual normal and skew-normal distributions. The solid lines use player specific variances which are estimated using the shrinkage, while the dashed lines illustrate the NLL using the global variance for residuals ($m \to \infty$).

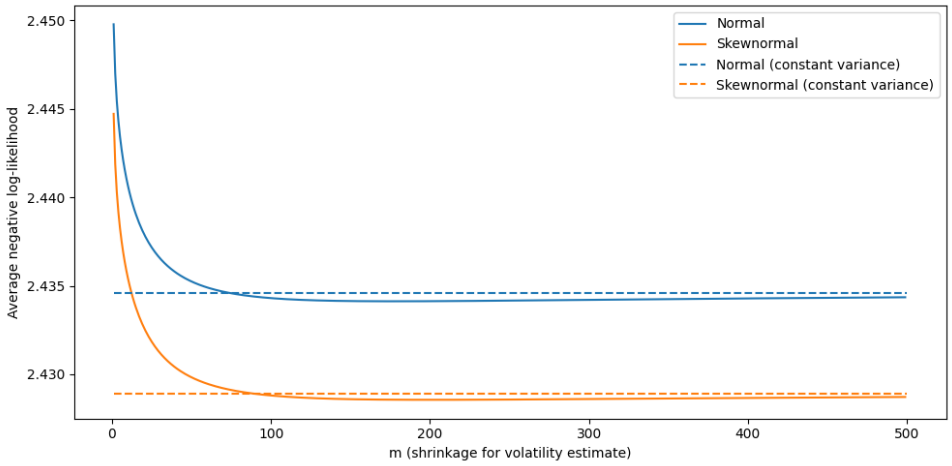


Figure 16: Negative log-likelihood for residuals as a function of the shrinkage weight.

The results further support the choice of the skew-normal distribution due to consistently smaller NLLs. The optimal m for both residual models (178 for normal and 201 for skew-normal) seems to be around 200, which gives a 43 % weight for a player's observed variance, assuming that it is calculated using the full window of 150 rounds. As a conclusion, we set $m = 200$ to construct the variance estimates. If a player does not have sufficient amount of data to construct the estimates, their residuals will be assumed to follow the global distribution.

5.3 Results

This section presents the results for the simulated probabilities. Unless otherwise specified, all of the simulations use the EWMA estimate with half-life parameter 38 and shrinkage parameter $m = 3$, as chosen in Section 5.1. The simulations are generated using four different residual distribution specifications: skew-normal and normal distributions, each estimated either with player-specific parameters or with non–player-specific parameters (global).

The final model selection will be done by inspecting the in-sample simulation results. The in-sample simulations contain tournaments (total of 124) played between the beginning of 2019 and end of 2022, excluding the first year of the fit period (2018). This is done to allow the variance estimates to have a burn-in period, as these estimates can only be constructed after a sufficient amount of historical data has been accumulated.

Figure 17 shows the negative log-likelihoods for the predictions using the in-sample. The in-sample results indicate that normally distributed residuals with player-specific variance achieve the best performance under the NLL validation metric.

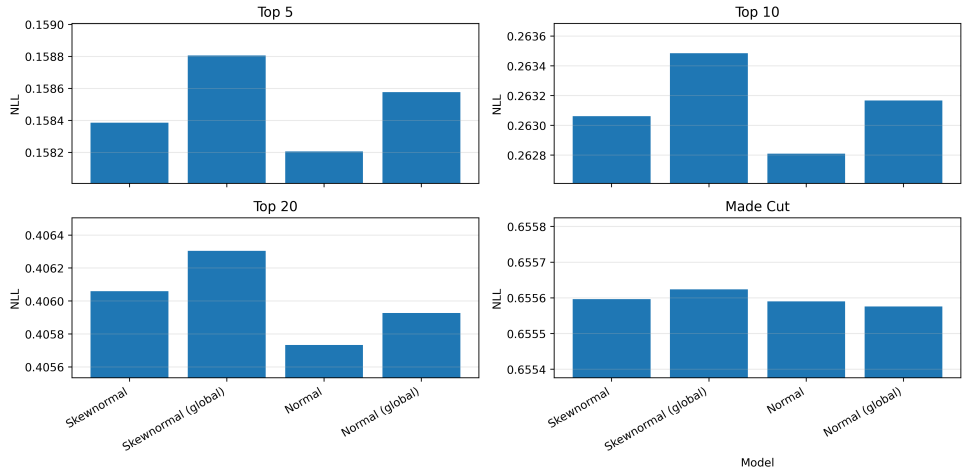


Figure 17: NLLs of simulated probabilities and observed outcomes in the in-sample dataset.

It is intuitive that allowing the variance to vary across players improves the results compared to assuming a constant variance for all players. However, we observe that the skew-normal residuals do not perform as well as the normal ones, despite earlier findings suggesting otherwise. This is because the skew-normal model consistently assigns lower probabilities to lower-skilled and higher variance players, which results in a higher loss. The loss measure is particularly sensitive to events with predicted probabilities close to 0 or 1, due to incorrect predictions in these cases causing large penalties, since $\log(x) \to -\infty$ as $x \to 0$. Regardless of whether the differences arise from random variation or model bias, we select the normal residual model with player specific variance, as it performs best under our evaluation metric.

Table 8 provides an illustrative example of what the simulated probabilities look like. The table shows the simulated and market-implied probabilities for the 2025 Tour Championship Event. The market-implied probabilities are calculated from the last traded price at Betfair marketplace before the start of the event. As the table shows, the simulated probabilities follow the market fairly closely, indicating that they are sensible and not arbitrarily generated.

Player	Win (mkt)	Win (sim)	T5 mkt	T5 (sim)	T10 (mkt)	T10 (sim)
Scheffler	0.357	0.308	0.676	0.698	0.855	0.868
McIlroy	0.100	0.079	0.385	0.335	0.602	0.560
Fleetwood	0.069	0.061	0.317	0.305	0.543	0.539
Åberg	0.048	0.033	0.278	0.190	0.485	0.376
Henley	0.045	0.044	0.286	0.240	0.503	0.454
Hovland	0.038	0.027	0.200	0.170	0.403	0.351
Burns	0.033	0.029	0.196	0.171	0.388	0.352
Morikawa	0.028	0.025	0.208	0.156	0.362	0.331
Thomas	0.025	0.034	0.167	0.188	0.379	0.374
Young	0.025	0.020	0.175	0.130	0.360	0.287

Table 8: Comparison of market-implied probabilities and simulation-based probabilities for win, top 5, and top 10 outcomes. Includes the 10 players (out of 30) with largest market-implied probability for win.

The out-of-sample period contains tournaments (total of 93) between start of 2023 and end of 2025. These will be used to evaluate the model, and find out whether it works out-of-sample or not. Unless otherwise specified, the results presented in this section refer to the out-of-sample period.

Figure 18 shows the NLLs for the different residual distributions for the out-of-sample tournaments. The NLL for the in-sample and out-of-sample periods are not directly comparable, as differences in player abilities affect the skewness of the outcome probabilities. Furthermore, the tournament structures differ across the two periods, as the out-of-sample period includes the newly introduced signature events, which have smaller fields than traditional PGA Tour events and therefore increase the proportion of players finishing in the top n or making the cut. Nevertheless, the same overall pattern can be observed in both periods.

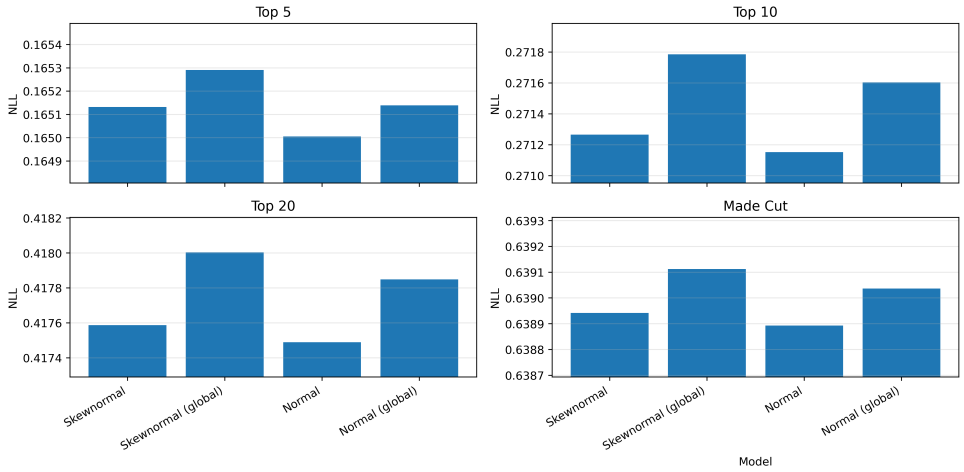


Figure 18: NLLs of simulated probabilities and observed outcomes in the out-of-sample dataset.

To inspect the result from a different perspective, we compare the results of the following specifications.

1. Assuming all modelled players to have same ability estimates ($\hat{\mu}$) and variance, resulting in equal probabilities across players.
2. The model with $N(0, 0.2)$ noise added to the round-level ability estimates. This is a moderate noise, as the average standard deviation for player abilities within a tournament is 0.70.
3. The model with normally distributed residuals and player-specific variance.
4. The expected NLL under the assumption that the outcomes follow the probabilities provided by the model.

Figure 19 presents the corresponding NLLs across both in and out-of-sample tournaments. The value for the expected NLL (Normal EV) should not be interpreted as the theoretical lower bound for the loss, as the realised outcomes are not generated from this distribution. However, the fact that our results are within the same ballpark is encouraging.

Adding $N(0, 0.2)$ noise to the ability estimates increases the loss marginally. Even with this additional noise, the losses remain far below those obtained when assigning equal probabilities to all players, which suggests that the estimated abilities capture meaningful differences in player skill.

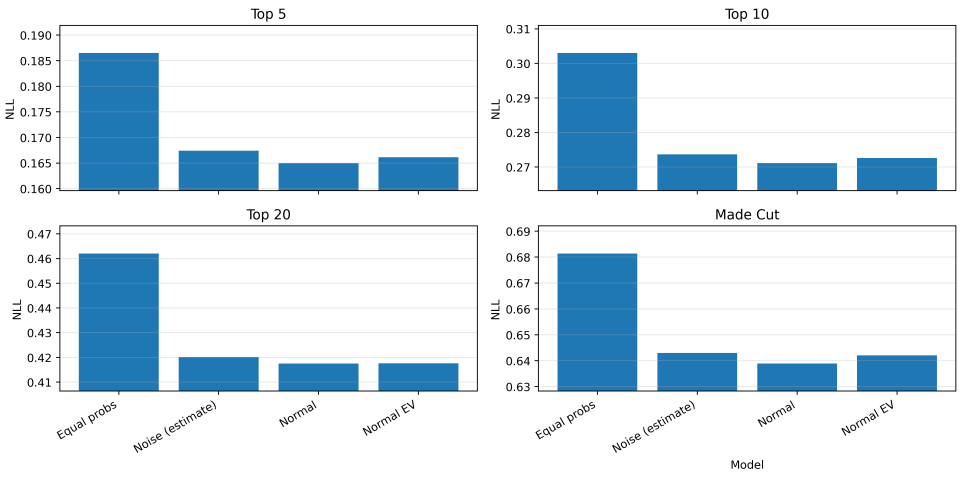


Figure 19: Mean NLLs of the model and the listed specifications.

To further assess the performance of the model, a permutation test was conducted. The tournaments were simulated 1000 times while randomly reassigning the estimated player parameters across players within each tournament. Figure 20 shows a histogram of the NLL values obtained from the 1000 permutation simulations. The permutation simulations also perform worse than the equal-probability benchmark. This is because the permutations produce skewed probability estimates, but assign them to the wrong

players, whereas the equal-probability benchmark assigns flat probabilities across players. This supports the importance of correctly identifying differences in player ability for predicting tournament outcomes.

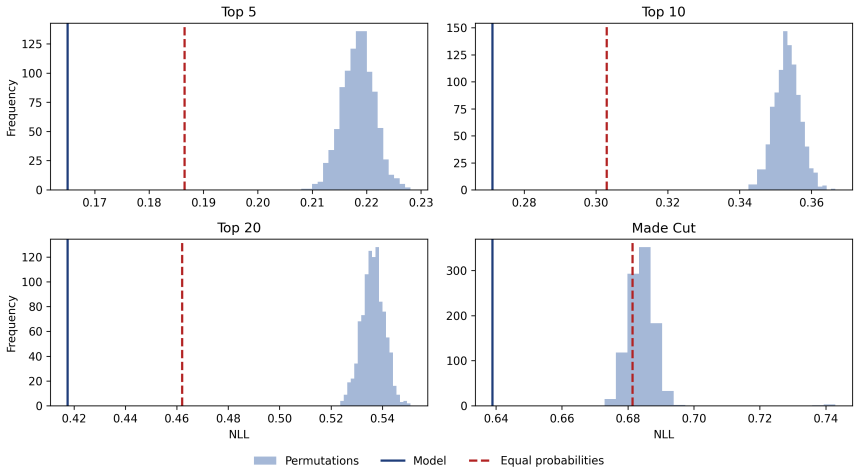


Figure 20: Histogram of the mean NLL for the permutations.

The final assessment of the results is to compare how well the model predicts outcomes across different skill levels. Figure 21 shows the realised outcomes relative to the expected probabilities for different deciles of player ability. For each tournament, the players are divided into ten deciles based on their estimated ability, and the results are then aggregated across these groups.

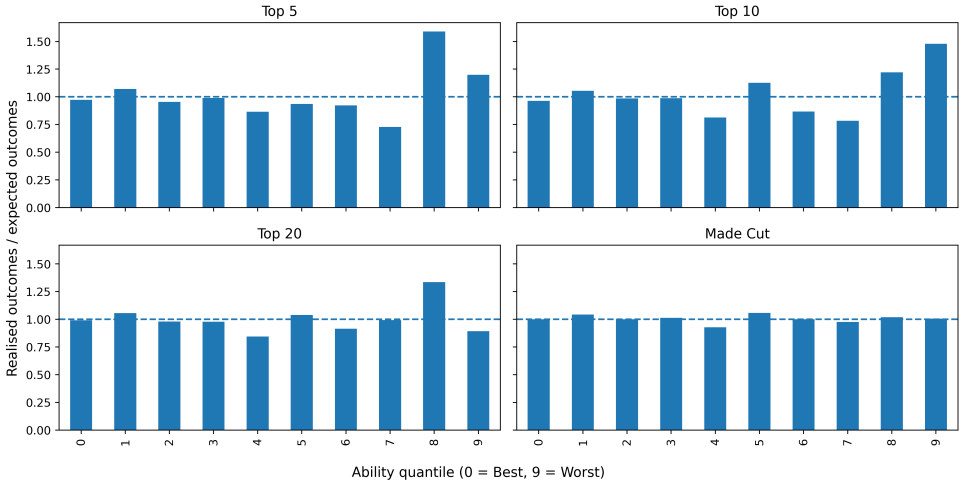


Figure 21: Ratio between realised and expected outcomes.

For the top 20 and made cut events, the ratios are fairly well balanced across the ability deciles. As the event becomes more selective, the deviations increase, with the ratio for the lowest ability decile exceeding two in the Top 5 market. This is partially explained

by the fact that the median probability of the lowest-decile players finishing in the Top 5 is only 0.36%, making the estimates sensitive to small deviations in realised outcomes. However, we cannot rule out the possibility of the model systematically underestimating the probabilities for these players. Overall, we are fairly satisfied with the results, as especially the best players, who are the most interesting ones, have a ratio which is fairly close to one.

6 Further enhancements and studies

6.1 Weather effects

To examine the effects of weather conditions on scoring, we gathered a dataset containing hourly weather observations from the tournament locations during each round of play. The weather variables considered in the analysis, and their expected effects, are as follows:

- **Wind speed (kph):** Likely to be the most important weather feature as it increases unpredictability of ball flight.
- **Wind gust (kph):** Unpredictable wind makes it more difficult to adjust for the wind. However, wind gust is unknown for many observations.
- **Temperature (Celsius):** Warmer temperatures let the player move more freely.
- **Precipitation (mm):** Influences course softness, making bounces more predictable. However, makes playing more uncomfortable due to hands getting wet.
- **Air density ($\mathbf{kg/m^3}$):** Lower air density makes ball travel further, making play easier. High correlation with temperature.

Since we are modelling relative scores, i.e. how players play compared to each other, we are interested in the conditions each player faced during their rounds of play. This can be calculated by combining information about their tee-time and course location. We have the data for the players' tee-times, which is generally accurate, but may have inconsistencies during extreme weather due to delays. Play is often suspended if the weather conditions are dangerous due to lightning or wind being too strong making the conditions unplayable. In these cases play is delayed or moved to next day, making our data overstate the conditions. These records will be excluded where obvious, but some incorrect observations will persist due to limitations in identifying them.

Note that the results in this section are created using observed weather conditions. While this approach is appropriate for examining the relationship between weather and scoring, the results cannot be directly used for beforehand predictions, as the weather realisations are not known in advance. For predictive purposes, we would need to rely on historical weather forecasts, while also accounting for the uncertainty around the forecast. Thus, the results of this section should be interpreted accordingly.

To approximate the time intervals over which players completed their rounds, the tee-times are combined with the latest PGA Tour average round-durations. In mid 2025, the PGA Tour published that their year to date median duration for a round was 4 hours and 46 minutes [12]. While 2-player pairings generally play faster than this, and 3-player groupings slower, we use a constant of 5 hour duration for a round.

Tee-times are specified with minute-level precision, whereas the weather data consists of hourly observations. Thus, if we define the round i tee-time for player p as $TT_{p,i}$, we

can compute the average weather conditions faced by the player for weather variable W as

$$W_{p,i} = \frac{\sum_h W_h f_h}{\sum_h f_h},$$

where W_h denotes the hourly observation of weather variable W, and $f_h \in [0, 1]$ is the fraction of hour h included in the interval $[TT_{p,i}, TT_{p,i} + 5]$.

Using these values, we compute the deviation between the weather conditions experienced by player p and the average conditions in round i as

$$\Delta W_{p,i} = W_{p,i} - \overline{W}_i \,,$$

where $\overline{W}_i$ is the average condition experienced by players participating in round i.

Figure 22 shows the correlation of the observed differences in the weather variables.

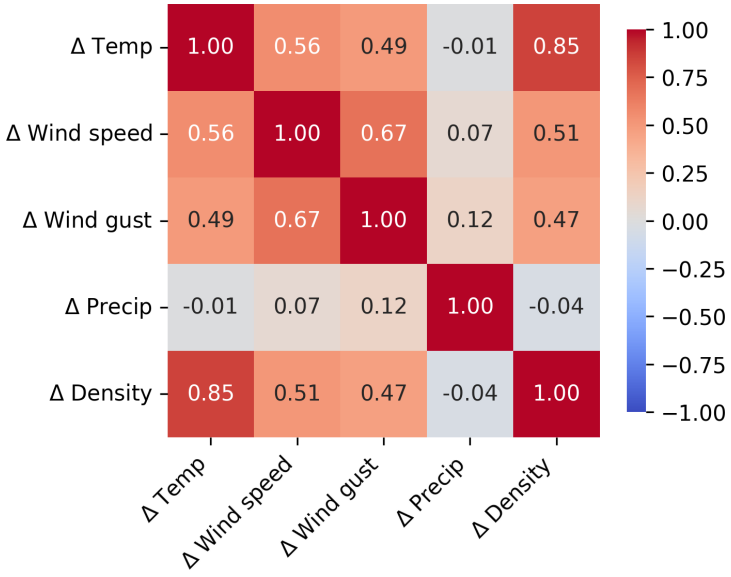


Figure 22: Correlation between the weather condition deviations.

We observe that wind speed and wind gust are strongly correlated, as are air density and temperature. Since wind gust is available only for part of the dataset, and air density is derived from other weather variables while temperature is directly measured, we retain wind speed and temperature in the analysis. Precipitation is also included, as it shows little correlation with the other variables.

The next step is to find whether there exists a relationship between the weather variables and the scoring. To examine this, we try to explain the round-level residuals, i.e. the difference between observed and expected outcome provided by the basic model, using the weather variables.

Figure 23 shows scatter plots of the residuals against the differences between the observed weather conditions and the corresponding field-average weather conditions.

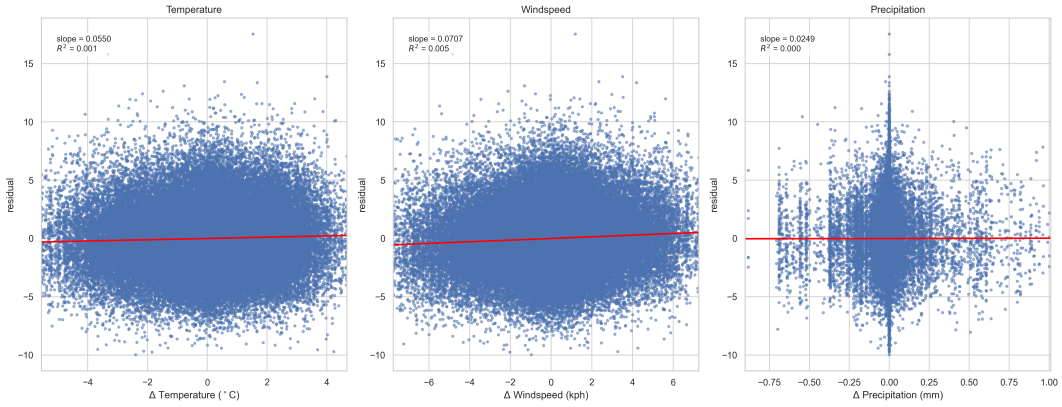


Figure 23: Residuals as a function of the difference between a player's observed weather conditions and the corresponding field-average weather.

The scatter plots show that wind and temperature seem to have an effect on scoring, while precipitation seems to have little to no effect. As an initial analysis, a simple linear regression estimate indicates that a one degree Celsius increase in temperature increases score approximately 0.06 strokes, while a one kilometre per hour increase in wind speed corresponds to roughly 0.07 strokes. The counter-intuitive positive temperature coefficient reflects the correlation between the wind and temperature variables, rather than a real relationship between temperature and score, as the single-variable estimate absorbs part of the wind effect. The explanatory power of these variables is small, and due to the collinearity between temperature and wind speed, the individual coefficients should be interpreted with caution. However, this does not imply that weather has no effect on scoring. Rather, it indicates that the effect is small relative to the overall noise in round scores.

Next, we inspect whether the marginal effect varies across different levels. To do this, we model the effect of the weather deltas (ΔW) as a function of the field-average weather level ($\overline{W}$).

Figure 24 shows the estimated slope of how much a one-unit difference in the weather variable relative to the field average affects scoring at different field-average weather levels. The effect of temperature differences on scoring appears small at low temperatures but increases as temperature rises. For the wind, we see a clearly positive slope (β) already at lower wind speeds, which increases as the wind level rises.

The finding that a one metre per second wind difference has a larger effect at higher wind speeds (e.g., 30 vs. 10 kph) is intuitive, as aerodynamic drag is proportional to the square of velocity. This, however, holds only to some extent, as playing in extremely strong winds would make scoring difficult for all players. For temperature, we do not identify a clear rationale for the observed pattern.

These results are presented as an empirical observation, and the weather adjustment in the model will not be conditioned on the level of the weather variable.

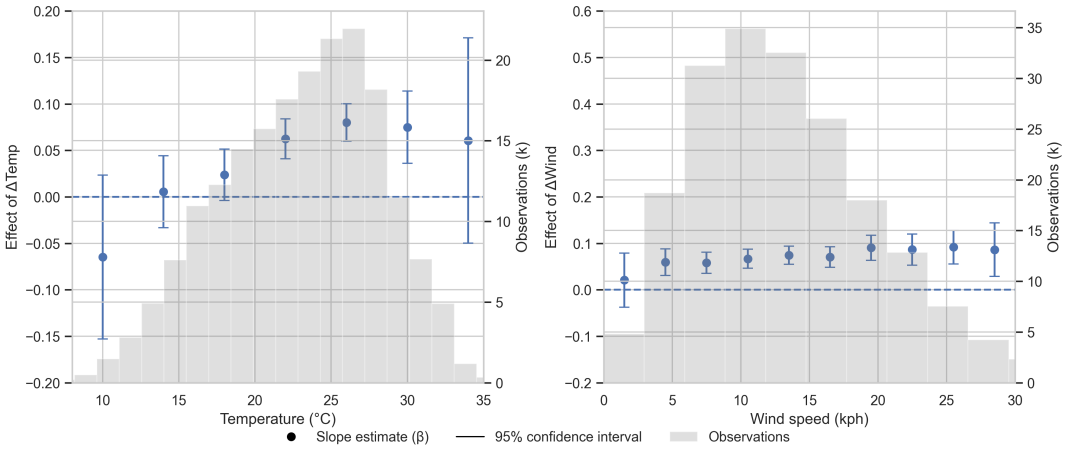


Figure 24: Slope coefficient as a function of average weather level during round.

While weather variables capture differences in playing conditions, they may not fully explain systematic differences related to the timing of play. Earlier studies have suggested that tee-times affect the scoring of players [13], [14]. The exact reason for this is not explained, but they suggest that the cause would be both wind which tends to increase during afternoon, and that putting becomes more difficult due to footprints on green, making the scoring average increase for the players who have late tee-times.

Figure 25 shows how the tee-times are distributed between rounds 1–2 and 3–4. First two rounds often include more players than the subsequent rounds, creating an increased need to spread out the play across the day. This is apparent also from the histogram, as we see an increased frequency of tee-times during the first couple of hours and 6 hours after the first tee-time.

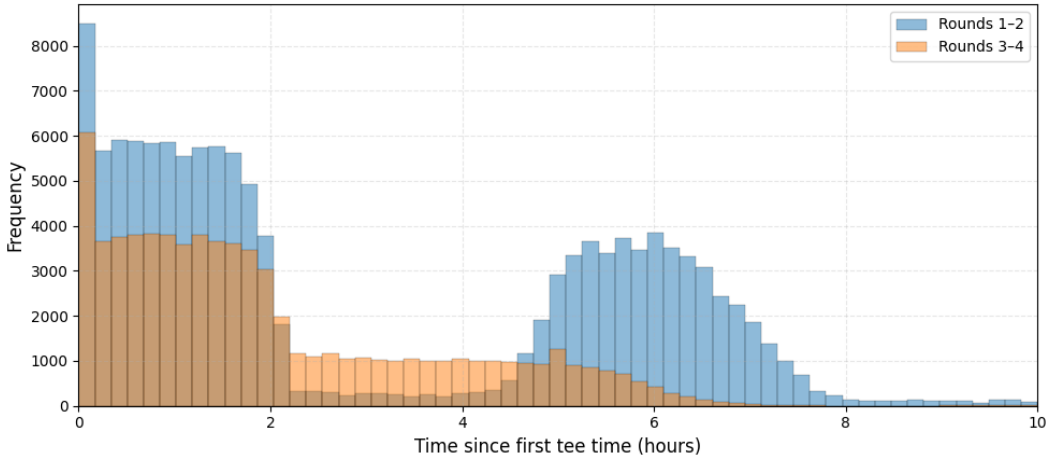


Figure 25: Tee-time distribution for rounds.

To examine whether differences in scoring between morning and afternoon rounds are driven by factors beyond the observed weather variables, we categorise tee-times within each round. This is done using the k-means clustering algorithm, resulting in the following categories: morning and afternoon tee-times. To ensure that the clustering creates two meaningful groups, we introduce two restrictions: the distance between the cluster centres must exceed 3 hours, and the total span of tee-times within the round must exceed 2 hours. The tee-times of rounds which do not satisfy these criteria are classified as neutral.

Figure 26 shows the differences in player performance between morning and afternoon tee-times. The scoring difference represents the difference in average scores between players in the morning and afternoon groups, while the residual difference measures the difference in average residuals, i.e. performance relative to the basic model's expectations for each round. The results show that players with a morning tee-time perform 0.24 strokes lower than those playing in the afternoon.

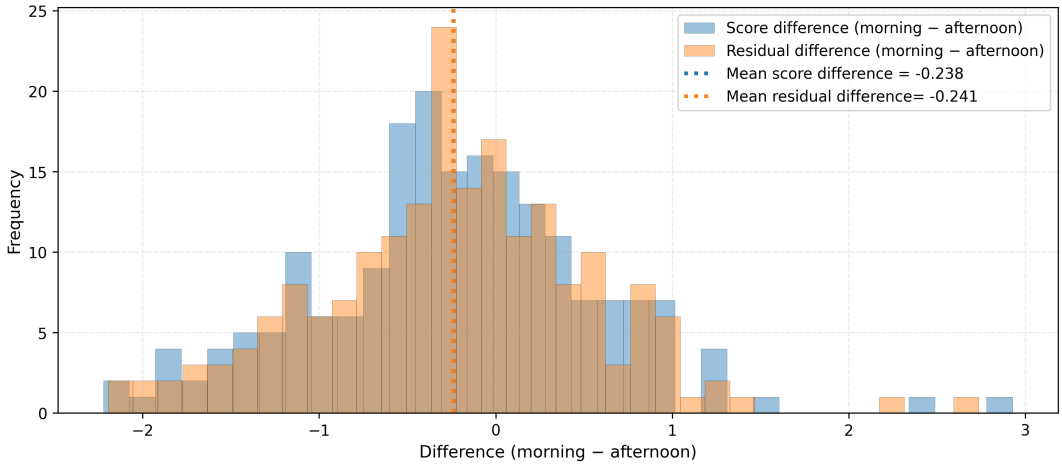


Figure 26: Residual distribution between the classified waves.

To adjust for observed weather conditions and other possible external differences between morning and afternoon tee-times, we include a tee-time indicator in the weather adjustment regression:

$$A_{p,i} = \beta_w \Delta\text{Wind}_{p,i} + \beta_t \Delta\text{Temp}_{p,i} + \beta_D D_{p,i},$$

where $D_{p,i} \in \{-1, 0, 1\}$ equals -1 for morning tee-times, 1 for afternoon tee-times, and 0 otherwise.

Variable	Coef.	Std. Err.	t	P> |t|	[0.025, 0.975]
Temp	-0.0004	0.008	-0.053	0.958	[-0.016, 0.015]
Wind	0.074	0.005	15.403	0.000	[0.065, 0.084]
Morning / Afternoon	-0.0386	0.023	-1.666	0.096	[-0.084, 0.007]
R-squared					0.005
Wind	0.0707	0.004	18.308	0.000	[0.063, 0.078]
R-squared					0.005
Observations					71,677

Table 9: OLS regression results for weather variables predicting the residual error.

Table 9 shows the multivariate linear regression (without intercept) results for all three variables, as well as the estimates obtained when fitting the model using only wind. The estimated slope parameters are not aligned with prior expectations. In particular, the coefficient for temperature is statistically insignificant when wind is included in the model. In addition, the morning–afternoon indicator coefficient has an opposite sign than expected. We suspect this is driven by correlation between the variables. Wind, whose coefficient was aligned with prior expectations, appears to explain most of the variation, while both temperature and wind increase over the course of the day (as does the morning–afternoon indicator by construction), as shown in Figure 27. This may bias the results, and therefore we use only wind speed to adjust the scores for weather.

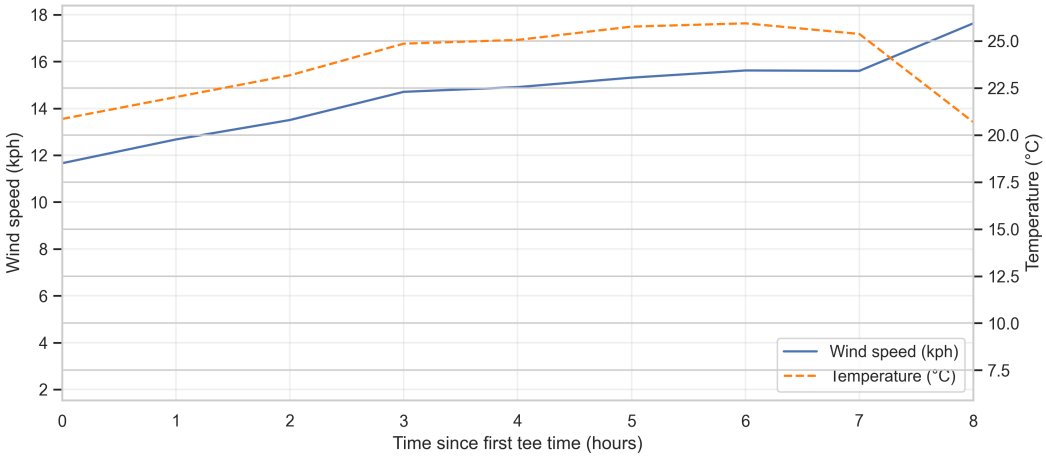


Figure 27: Average conditions faced as a function of time since first tee-time of the day.

When adjusting the scores for weather, we assume that a one kph average wind advantage corresponds to 0.07 fewer strokes. This value is taken from Table 9, using the regression estimate where wind is the only explanatory variable. Note that the value is rounded to two decimals, as the exact estimate is not essential for the analysis.

Next, we extend the basic model to account for the effect of wind on scoring:

$$\hat{Y}_{p,i} = \hat{\mu}_{p,i} + 0.07\Delta\text{Wind}_{p,i} + \varepsilon_{p,i} \ .$$

Here, the previously defined variables are used, and the coefficient 0.07 reflects the estimated coefficient for the wind differences.

Figure 28 shows the errors of the estimates, when the basic model is adjusted for the weather (wind) conditions. The adjustment is applied both to the historical rounds used to estimate player ability and to the round being predicted. The results are shown separately for the fit period, test period and for rounds where observed wind difference exceeds 3.6 kph (1 m/s). We observe a noticeable but moderate improvement in round-level prediction errors. On average, the weather-adjusted model reduces the error by approximately 0.2%–0.3% compared to the unadjusted model. When conditioning on rounds where the experienced wind deviates by more than 3.6 kph from the round (or field) average, the improvement exceeds 1%.

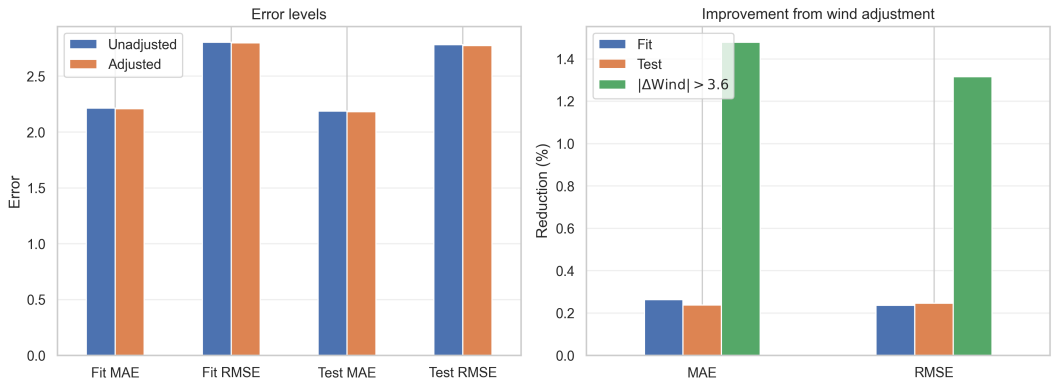


Figure 28: Improvement in residual errors when adjusting for wind conditions.

In order to examine how the weather adjustments affect the probability estimates, we consider the following specifications.

1. Basic model, where the historical observations are adjusted for weather
2. Specification 1 along with adjusting the first two rounds according to the observed weather ($\beta = 0.07$).

The first specification has no forward looking bias, as they rely only on information which is available prior to the tournament. Regarding the second specification, first two round tee-times are usually disclosed prior to the tournament. Weather adjustments for these rounds can be made using forecasts available prior to the tournament. Assuming that forecasts issued immediately prior to the tournament reasonably correspond to the realised conditions, they can serve as proxy inputs for the forecast. However, it is important not to interpret the results as strictly out-of-sample predictive performance. Simulating all four rounds conditional on observed weather is not possible, as only

players who make the cut have weather observations for the final two rounds, while in the simulations the players may complete all four rounds.

Figure 29 shows the improvements in NLL compared to the basic model when adjusting for weather. The improvements are defined as

$$\text{Improvement} = 1 - \frac{\text{Specification NLL}}{\text{Basic Model NLL}}. \tag{14}$$

We observe marginal but consistent improvements across the results. Adjusting historical scores for wind alone yields slightly lower errors, suggesting that the adjustment improves the quality of the player ability estimates by reducing weather-related noise in past round scores. The improvements are larger in magnitude during the test period than in the fit period. The reason for this difference is not directly identified, but it may be related to more extreme wind conditions in the later period, differences in the tournament sample, or changes in the distribution of predicted probabilities across the two periods. Incorporating the observed wind conditions for the first two rounds further improves the results, with the exception of the Top 20 category in the fit period. The largest improvements are observed for the made-cut market, which is intuitive since the cut is determined after the first two rounds, for which the weather conditions are known.

Overall, the results support using the 0.07 strokes per 1 kph wind adjustment rather than ignoring wind effects. The gains are small, but consistent, suggesting that the adjustment provides a modest improvement in predictive accuracy.

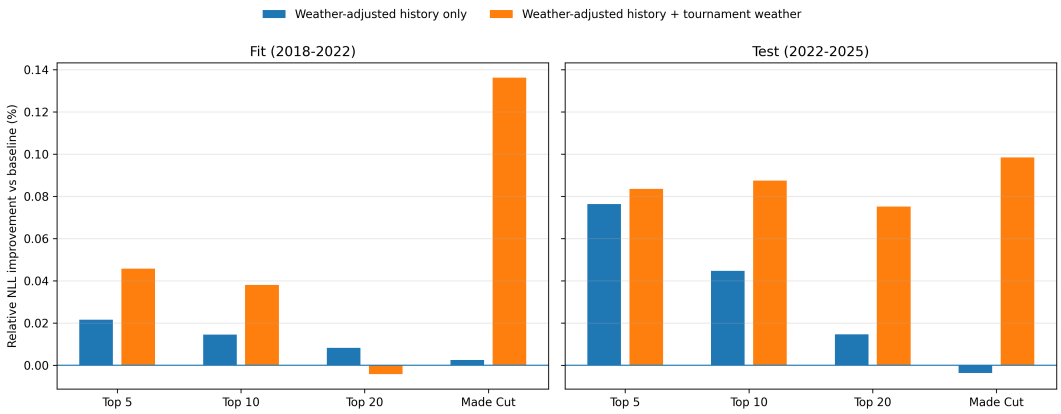


Figure 29: Improvements in NLL from adjusting for wind conditions relative to the basic model.

6.2 Alternative weighting scheme

In the basic model, we used an exponentially weighted moving average to construct our player ability estimate. While the EWMA is a simple and robust way of weighting the historical rounds, the findings in Section 5.1.4 imply that the decay might not be constant across lags. Thus, we introduce a more flexible weighting scheme

$$\hat{\mu}^{\text{alt}}_{p,t} = \sum_{i=1}^{t-1} w_i Y^{\text{adj}}_{p,i},$$
$$w_i \propto \alpha_1 e^{-(t-i)/\tau_1} + \alpha_2 e^{-(t-i)/\tau_2},$$
$$\alpha_1, \alpha_2 > 0, \quad \tau_1, \tau_2 > 0, \quad \tau_1 < \tau_2, \quad \alpha_1 + \alpha_2 = 1\ , \tag{15}$$

which is a mixture of two exponential decays. Here Y^{adj} denotes the field-strength-adjusted performance from the basic model. Recall Equation (8), where we model the weights of historical rounds as a function of lag (minimising round-level prediction least squares errors using the prior 150 rounds as input). In order to estimate the parameters of the weighting scheme, we solve the optimisation problem in Equation (8), subject to the constraint that the weights satisfy Equation (15). Using exactly 150 rounds does not make a large difference here, as it is large enough history for fitting the shape, and we can scale the weights (to sum to one) for shorter and longer histories in later stages. The fitting is done numerically using the Nelder-Mead method, and as a result, we obtain the following values for the parameters

$$\alpha_1 = 0.60,\ \alpha_2 = 0.40,\ \tau_1 = 13.10,\ \tau_2 = 90.39\ .$$

Figure 30 shows the more flexible scheme along with the EWMA (half-life parameter of 38), used in the basic model. Note that the NNLS estimates might slightly differ from the results shown in Figure 9, as here we use different field strength adjustments (provided by the basic model). Visually, the alternative weighting scheme seems to fit the unconstrained solved weights slightly better. However, the fits cannot yet be evaluated numerically. The EWMA is a special case of the alternative weighting scheme, making the alternative scheme always fit at least as well.

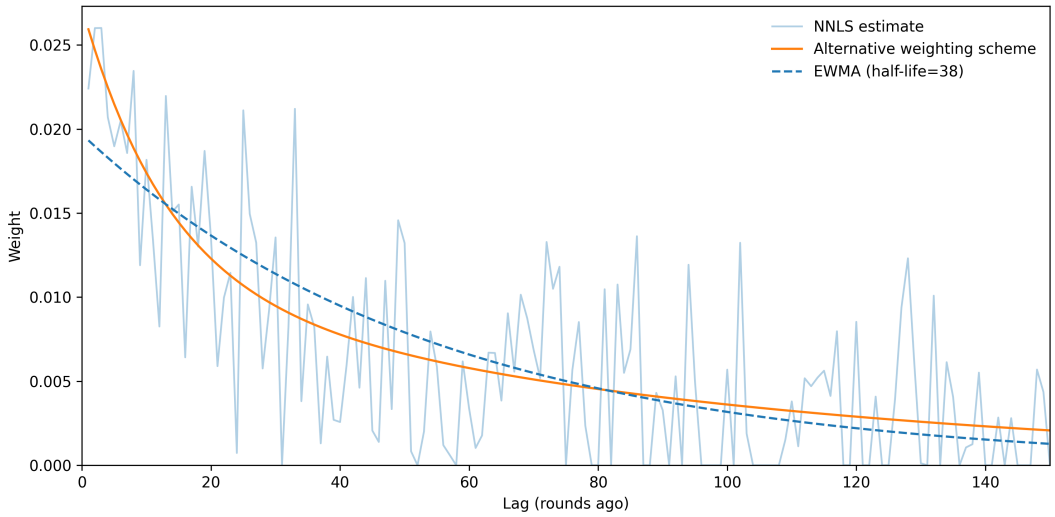


Figure 30: Fitted weighting scheme and the EWMA used in the basic model (for 150 rounds).

Next, we calculate the ability estimates for the fit period (2018-2022) using the alternative weighting scheme. Since it gives more weight to the most recent rounds, we expect to shrink it more heavily for players with shorter histories. Similarly as with the basic model, we shrink the estimates towards zero, as a function of the shrinkage parameter (m) and number of observations used for the estimate, shown in Equation (9). Figure 31 shows the RMSE and MAE for the different values of the shrinkage parameter. Contrary to what we suggested, the optimal shrinkage parameter does not seem to be significantly larger than for the basic model. We select $m = 4$ for shrinking the ability estimates provided by the enhanced weighting scheme.

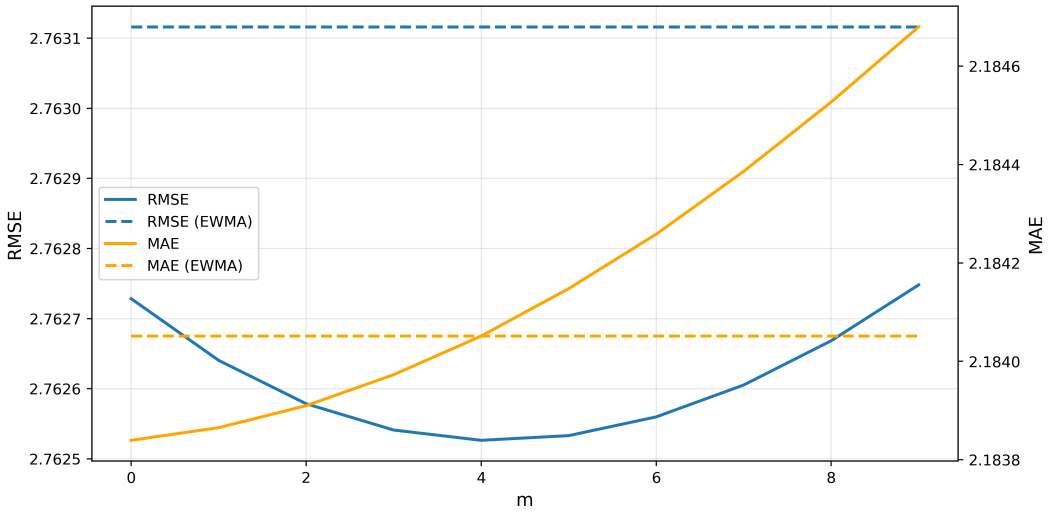


Figure 31: Optimal shrinkage parameter m when evaluating the residuals of the alternative weighting scheme under MAE and RMSE.

We then run the simulations using the ability estimates from the new weighting scheme, reusing the basic model's residual distributions for simplicity. Figure 32 shows the improvements (using Equation (14)) in the NLL versus the basic model for the fit and test (2023-2025) periods. The improvements are considerably larger for the fit period, whereas for the test period it is debatable whether the results improved or not.

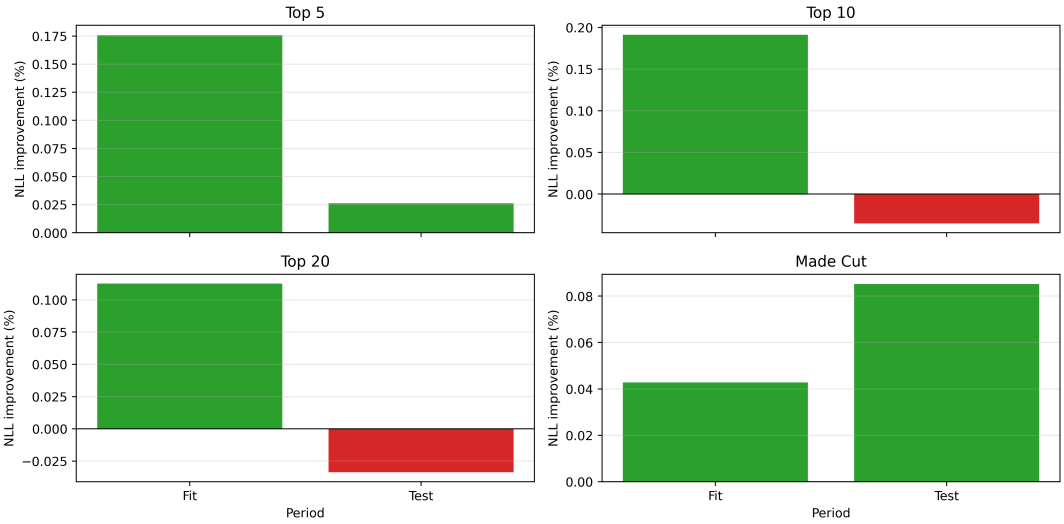


Figure 32: Improvements in NLL when comparing the alternative weighting scheme to the basic model across markets and periods.

This kind of behaviour is often caused by overfitting. However, in this case, the scheme was fitted using the errors between the estimates and the observed round-level performances. When inspecting the improvements in MAE and RMSE between the periods (Figure 33), we notice that the errors improve more during the test period.

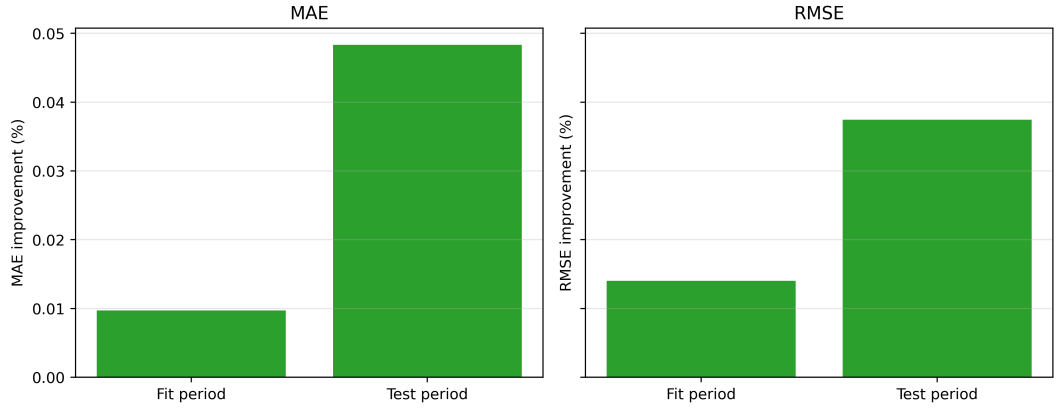


Figure 33: Improvements in residual errors compared to the basic model.

This implies that in general, the new weighting scheme does estimate round-level performances more accurately. One possible explanation for why the residual improvements do not translate into the simulation results is the shared conditions across most consecutive rounds. In the round-level data, the most recent prior round(s) are most often played at the same tournament and course as the round being predicted. When minimising round-level prediction errors, it seems natural to place more weight on a round played under the same conditions. This reduces the round-level residual error without noticeably improving the tournament-level probability estimates.

6.3 Player-course interactions

The aim of this section is to find out whether course data and player-specific data can be used to identify if a course suits a given player. We do not aim to identify the variables with the greatest explanatory power, but rather to conduct an exploratory analysis of whether such relationships can be detected at all.

Golf courses differ in their characteristics. Some courses are longer, while others may be narrow or have small greens. Players also differ in their strengths and tendencies, as some players might hit the ball farther, more accurately, or tend to miss in a certain direction. In this section, we study the relationship between the following:

- Course length and player's driving distance
- Importance of hitting fairway on course and player's driving accuracy
- Left-right balance of course hazards and player miss tendency
- Course historical GIR percent and player short game ability.

Since each course is played a maximum of four rounds per year, we will not split the dataset into fit and test datasets. We focus instead on finding out whether explanatory power exists rather than focusing on out-of-sample predictions. For driving distance, we are able to calculate the player-course interactions for each course, as yardage is known for the courses. For the other interactions, we have to empirically estimate the course coefficients, which requires sample size. For these, we consider courses which are played at least 20 times in our dataset (and have the required shot-by-shot data available).

The following subsections describe the construction of the variables, while the results for all variables are presented together at the end.

6.3.1 Driving distance

Based on our dataset, the average length of a course (at round-level) on the PGA Tour is 7192 yards with a 220 yard standard deviation. Based on PGA Tour 2025 statistics, the average approach distance, given that the player made par, was 165 yards [15]. Figure 34 shows the expected strokes needed to hole from different yardages, and the marginal benefit of being 10 yards closer at different yardages. Around 165 yards, the marginal benefit of gaining 10 yards is lower than the corresponding number at 180 yards, implying that longer courses benefit long hitters.

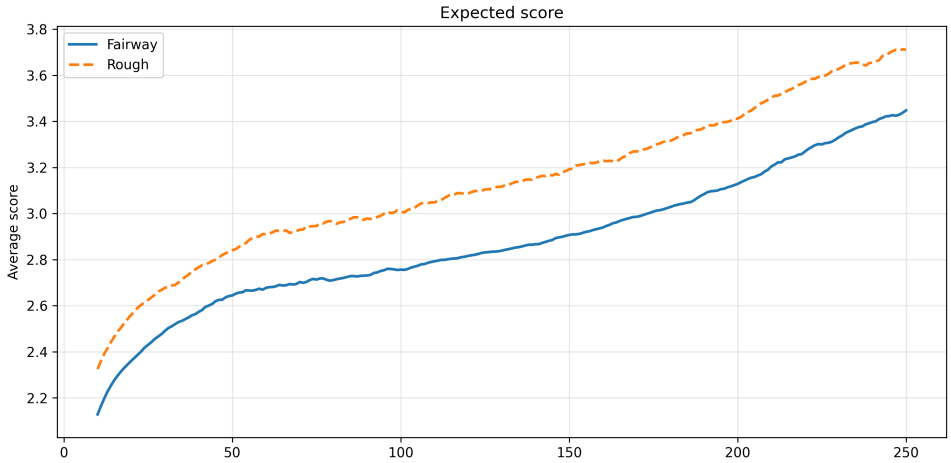


Figure 34: Expected amount of strokes needed from different yardages. Calculated based on shot-by-shot data consisting of over 30 million observations.

Using the statistics data, we estimate each player's driving distance as the average over their last 75 rounds (roughly corresponding to one season). To account for changes in course lengths over time, course length is normalised relative to the three-year average length on tour. We also normalise driving distance by measuring how much longer a player's driving distance is relative to the field average in each round, rather than using absolute values. These adjustments are necessary because driving distances (and course lengths as a result) have increased over the recent decades, to the extent that changes to equipment regulations, such as the introduction of a new golf ball, are being considered [16].

The player–course interaction term is then defined as the product of these two normalised quantities

$$I_{p,c}^{\mathrm{DD}} = (L_c - \overline{L}_{3Y}) \cdot (D_p - \overline{D}_{\mathrm{field}})$$

Figure 35 shows the average driving distance of players and the average length of course (per round) during 2024 and 2025. The figure indicates that there is meaningful variation in both player driving distance and course length, suggesting that driving distance may be more valuable on some courses than on others.

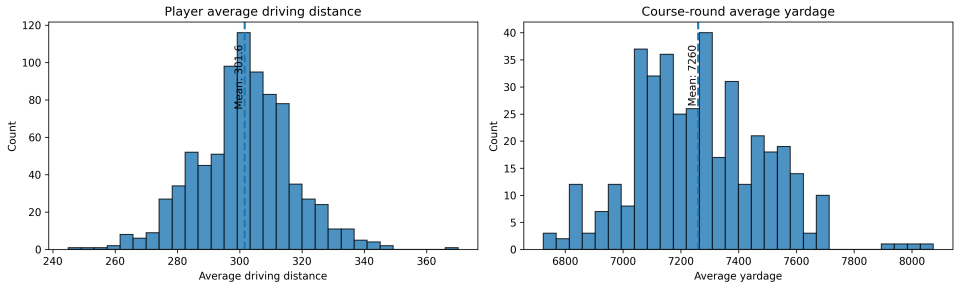


Figure 35: Average driving distance of players and course lengths (per round) during 2024 and 2025.

6.3.2 Driving accuracy

The impact of driving accuracy on scoring on different courses is less straightforward than for driving distance. On a very wide course, accuracy matters little because nearly everyone hits the fairway regardless of skill. Conversely, we argue that accuracy may be irrelevant on an extremely tight course. Thus, we measure the importance of accuracy on a course as

$$\text{AI} \propto \text{Var}(\text{fw hit}) \times \text{E}(\text{penalty} \mid \text{miss}) \ ,$$

where the variance term captures how much the course separates players in fairway outcomes, and the penalty term captures the expected cost of a miss. In practice, this importance will be calculated from hole level scores as

$$\text{AI}_c = \frac{1}{|H|} \sum_{h}^{H} p_{fw,h}(1 - p_{fw,h})(\text{E}[\text{net score} \mid \text{fw missed}, h] - \text{E}[\text{net score} \mid \text{fw hit}, h]) \ .$$

where H is the set of par 4 and 5 holes on the course c, $p_{fw,h}$ the empirical average fairway hit percentage. The driving accuracy estimate for players will be calculated similarly as for distance (average of past 75 rounds) such that the interaction term will become

$$I^{\text{A}}_{p,c} = (\text{AI}_c - \overline{\text{AI}}) \cdot (A_p - \overline{A}_{\text{field}})$$

Table 10 shows the 4 courses with least and most driving accuracy importance according to our metric.

Table 10: Courses where driving accuracy is least and most important, with larger values indicating more importance.

Course	Importance	N
Plantation Course at Kapalua	0.052	44
Memorial Park Golf Course	0.063	20
Sea Island Golf Club (Seaside Course)	0.065	44
Monterey Peninsula Country Club (Shore Course)	0.066	24
East Lake Golf Club	0.105	44
TPC Potomac at Avenel Farm	0.107	12
TPC Sawgrass (THE PLAYERS Stadium Course)	0.109	40
Muirfield Village Golf Club	0.113	48

6.3.3 Left-right miss bias

Many players tend to curve the ball in a particular direction, and courses where hazards are concentrated on one side may therefore be more suitable for some players than for others. In this context, a miss means the player failed to hit the fairway. Figure 36 shows the cumulative number of left and right misses (and the cumulative score following the misses) for Phil Mickelson, who was at times notorious for missing to the left.

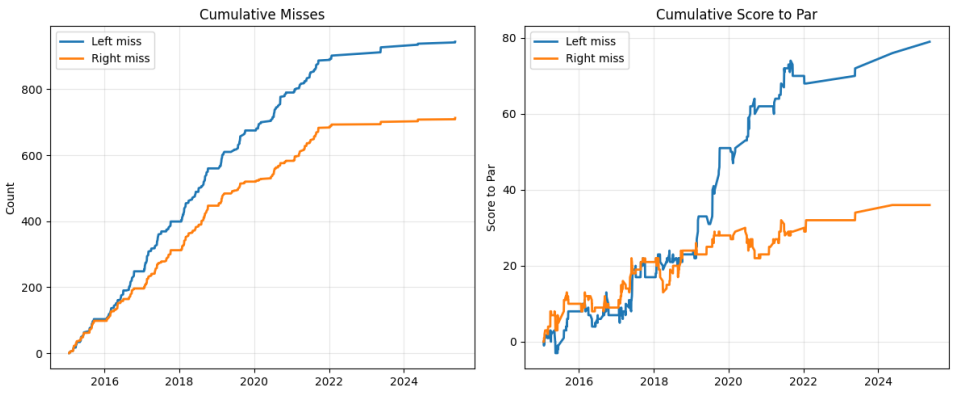


Figure 36: Phil Mickelson's cumulative left and right misses throughout the dataset.

We could argue that a course where all the water hazards are on the left side of the fairways would play more difficult for Phil Mickelson. Thus, we introduce a left–right miss bias index for each player as the (average) number of misses to the left minus the number of misses to the right over the last 75 rounds.

In addition, we calculate the left-right difficulty bias for each course c as

$$\mathrm{LRB}_c = \frac{1}{|H|} \sum_{h}^{H} (\mathrm{E}[\text{net score} \mid \text{left miss}, h] - \mathrm{E}[\text{net score} \mid \text{right miss}, h]) \ ,$$

which tells whether the course has more trouble left or right, which will give us a left-right miss bias interaction term

$$I_{p,c}^{\text{LRB}} = \text{LRB}_c \cdot \text{LRB}_p$$

Table 11 shows the courses where left-right driving bias is most important. Positive-valued courses are expected to favour players who predominantly miss right, and negative-valued courses those who miss left.

Table 11: Courses with the largest estimated left-right miss importance values.

Course	Importance	N
TPC Four Seasons Resort	-0.09	12
Colonial Country Club	-0.07	44
TPC Boston	-0.07	20
RTJ Trail (Grand National)	-0.06	12
Keene Trace Golf Club (Champion Trace)	0.04	20
TPC Potomac at Avenel Farm	0.05	12
TPC Sawgrass (THE PLAYERS Stadium Course)	0.05	40
Arnold Palmer's Bay Hill Club & Lodge	0.06	44

For example, the value for Arnold Palmer's Bay Hill Club & Lodge may be explained by holes 3, 6 and 11, where the fairway curves around water on the left, as shown in Figure 37.

Figure 37: Course map for Arnold Palmer's Bay Hill Club & Lodge.

6.3.4 Short game

We assume that short-game ability is more important on courses with lower green-in-regulation (GIR) percentages. By short-game ability, we mean a player's skill on shots played around the green (typically under 30 yards). Some courses are inherently more demanding, for example due to small greens or longer approach shots, which leads to lower historical GIR percentages and a higher number of short-game shots required. Players with better short-game ability are therefore expected to perform better on such courses.

The interaction variable for a player and a course is defined as

$$I_{p,c}^{\text{SHG}} = (\text{GIR-\%}_c - \overline{\text{GIR-\%}}) \cdot \text{SHG}_p \ ,$$

where $\text{GIR-\%}_c$ is the average green-in-regulation percentage for course c, $\overline{\text{GIR-\%}}$ is the average percentage across all courses, and SHG_p denotes the player's short-game ability measured as the mean strokes gained around-the-green over the past 75 rounds (minimum of 30 rounds required).

Figure 38 shows the distribution of the course-specific GIR percentages as well as the calculated players' short game abilities for 2024 and 2025. Table 12 shows the courses with empirically lowest and largest green-in-regulation percentages.

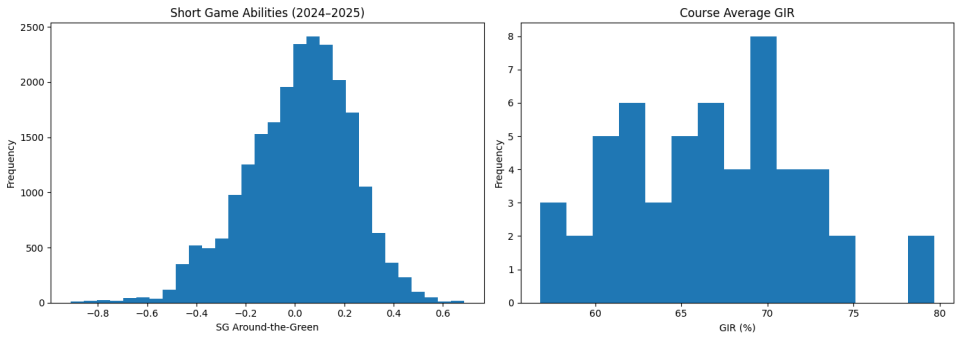


Figure 38: Distribution of the course-specific GIR percentages and the calculated players' short game abilities for 2024 and 2025.

Table 12: Courses with highest and lowest green-in-regulation percentages.

Course	GIR-%	N
El Cardonal at Diamante	80.0 %	12
Plantation Course at Kapalua	78.9 %	44
Monterey Peninsula Country Club (Shore Course)	73.9 %	24
Sea Island Golf Club (Seaside Course)	73.9 %	44
Arnold Palmer's Bay Hill Club & Lodge	58.8 %	44
Innisbrook Resort (Copperhead Course)	57.3 %	40
Firestone Country Club (South Course)	57.3 %	12
The Riviera Country Club	56.6 %	40

6.3.5 Results

In this section, we conduct linear regression to assess whether the interaction terms are relevant. We attempt to explain the residuals (Equation (10)) of the basic model using the interaction variables. Recall that the residuals are the differences between observed and expected outcomes, such that a positive residual means that player performed worse than expected. Since the different interaction variables are available for slightly different subsets of the dataset, we analyse each interaction term separately. This is also consistent with our objective, which is to evaluate whether the variables have explanatory power rather than to build a predictive model around them.

The interaction variables are constructed as products of two separate variables, of which one is player-specific and the other course-specific. The player-specific variables are estimated directly using player's historical data, and can be treated as best available estimates. The course-specific variables, on the other hand, are estimated with greater uncertainty. To assess the persistence of the course-specific variables, we conduct a split-half correlation test. This means, that for each course, we take the round-level observations and randomly split them into two equally sized subsets. We then compute

the course-level statistic separately for both subsets and calculate the correlation between the two resulting estimates across courses. This procedure is repeated 1000 times, producing a distribution of split-half correlations, which is visualised in Figure 39.

The Figure suggests that the accuracy importance metric and the average GIR-% are highly stable, whereas the left-right bias has somewhat weaker stability. While this does not directly evaluate the interaction terms, it provides support for using GIR-% and accuracy importance as course-specific variables.

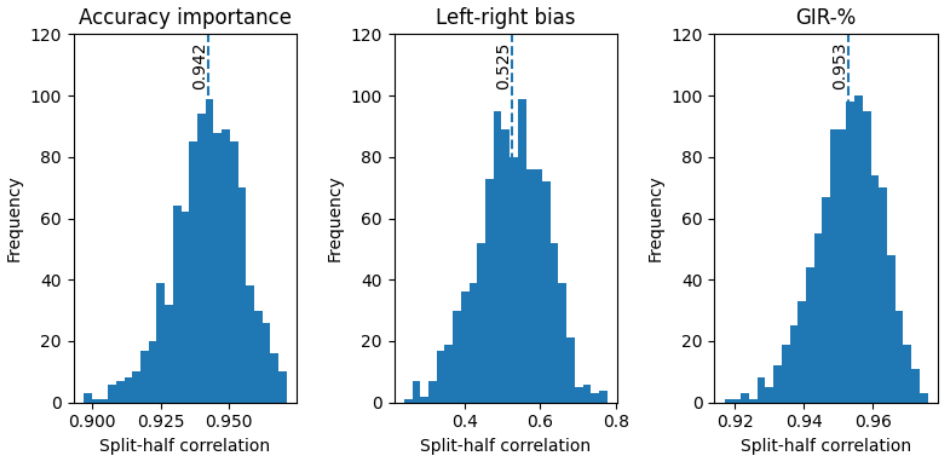


Figure 39: Correlation of random halves of round-level importance metrics.

Table 13 presents the regression results when explaining the residuals using the interaction variables without an intercept term.

Table 13: Univariate regression results for the variables. Sample sizes vary from 80 thousand to over 100 thousand, depending on variable.

Variable	Coef	P-value	R^2	CI low	CI high
Driving distance	$-2.2 \cdot 10^{-5}$	0.00	0.0002	$-3.1 \cdot 10^{-5}$	$-1.4 \cdot 10^{-5}$
Accuracy	-0.79	0.00	0.0003	-1.09	-0.50
LRB	-0.91	0.29	0.0000	-2.62	0.80
Short game	0.011	0.19	0.0000	-0.0057	0.0276

Both driving distance and accuracy yield marginal, but statistically significant coefficients. The negative coefficients imply that players with longer driving distance perform better on longer courses, and that more accurate players perform better on the courses where accuracy is more important. Although the explanatory power of the variables is limited, the results suggest that these interaction effects may contain some additional information beyond the basic model. For driving distance, the coefficient

implies that on a course one standard deviation longer than average (+220 yards), a player one standard deviation longer than average (+8.2 yards) has an advantage of 0.04 strokes per round. Similarly, for driving accuracy, on a course with an importance score one standard deviation above average (+0.013), a player one standard deviation more accurate than average (+5.1 pp) has an advantage of 0.05 strokes per round.

For the left-right miss bias (LRB) and the short-game interaction terms, the coefficients are not statistically significant. When further examining the course-specific part of the LRB variable, we also observe some inconsistencies. In the older portion of the data, hazards (areas which often cause penalty strokes), are not labelled left or right, whereas rough for example is. This causes the data to be biased, as for a hole where there is water immediately left of the fairway, the data shows no misses to left, and an increased amount of misses to the rough on the right (as players try to avoid the water). Thus, no conclusions can be drawn from the variable, although it was a bit exploratory by nature. For the short-game variable, the results are either caused by poor construction of the variable, or simply due to the lack of explanatory power.

6.4 Betting against the basic model

So far, we have evaluated the model under the NLL metric. While NLL measures how well the probabilities fit the realised outcomes, the absolute levels of the measure are unintuitive. To evaluate the results from a different perspective, we use the simulated probabilities to bet against historical traded odds on Betfair Exchange. Betfair Exchange is a marketplace where people can back (bet for) and lay (bet against) different outcomes. The market is organised as an order book with a back side and a lay side showing the available quotes. The best back odds are lower than the best lay odds, and the difference between them is the spread. For example, if the market consensus fair odds are around 2.0, the best back and lay quotes might be 1.95 and 2.05, with the gap between them representing the spread.

The odds are in decimal format, meaning that for a back bet with odds o and stake s, and fractional outcome $x \in [0, 1]$ (discussed in Section 4.1), the net payoff is

$$\text{payoff}_{\text{back}} = s(o-1)x - s(1-x) = s\big(ox - 1\big).$$

The corresponding net payoff for a lay bet is

$$\text{payoff}_{\text{lay}} = s(1-x) - s(o-1)x = s\big(1 - ox\big).$$

This means that if we back at odds 3 with stake 1, we win 2 if the event occurs fully ($x = 1$), and lose 1 if the event does not occur ($x = 0$). Similarly, for a lay bet at odds 3 with stake 1, we win 1 if the event does not occur, and lose 2 if it occurs fully.

The expected value (per unit of stake, $s = 1$) of a back bet with odds o is calculated as

$$EV(o|\text{Pr}) = \text{Pr}(o-1) - (1-\text{Pr}) = \text{Pr} \cdot o - 1.$$

This means that if we can back a fair coin flip at odds 2.2, the expected value for the bet is $0.5 \cdot 2.2 - 1 = 0.1$. While such a bet would not exist in an efficient market, in our setting the true probabilities are unknown, which may provide opportunities to place bets at favourable odds.

An important disclaimer is that we use last traded prices (odds) rather than the available back and lay offers. As a result, the odds reflect the most recent matched odd rather than the odds available in the order book, and may therefore differ from the true executable prices at the given moment. However, to mitigate the effect of this, we use the average last traded price (odds) over the 72 hours preceding the tournament start.

The objective is to identify bets with positive expected value. For each tournament, market (top 5, top 10, and top 20, where available), and player, we calculate the expected value for the bets using the average last traded odds and simulated probabilities. We place a back bet when the expected value exceeds a chosen threshold, and a lay bet when it falls below the negative of the threshold. Given a simulated probability $\widehat{\text{Pr}}$,

odd o and a threshold τ, our decision process is the following

$$\text{decision} = \begin{cases} \text{back with stake 1}, & \text{if } EV(o|\widehat{\Pr}) > \tau, \\ \text{lay with stake 1}, & \text{if } EV(o|\widehat{\Pr}) < -\tau, \\ \text{do nothing}, & \text{otherwise.} \end{cases}$$

We also cap the maximum odds at 20, 15, and 10, for top 5, 10, and 20 markets, respectively. This avoids placing bets on lower skilled players, whose estimates are generally more unstable, and avoids betting on low probability events, as they increase variance as we use the same stake for all bets.

We backtest the performance using tournaments from 2023–2025, with the following exceptions:

- We exclude tournaments where more than 25% of players do not have sufficient data to construct the estimates provided by our basic model.
- We do not bet on players who do not have sufficient data; they are only included as dummy players in the simulations.
- We exclude players whose names cannot be matched between the Betfair data and our dataset (e.g., due to diacritics).
- We exclude major championship tournaments, where the presence of additional players (e.g., from other tours) increases the quality of dummy players.

In addition, we must note that the simulations only include players who have completed all their rounds during the tournament, i.e., have not withdrawn. While this does not bias the results when evaluating the model under the NLL metric, it does introduce a survivorship bias when using the probabilities against the odds. As we do not have estimated probabilities for the withdrawn players, we do not place bets on them. This means we avoid losses from backing them, but also miss gains from laying them.

Table 14 shows the results for the 85 tournaments which were valid for betting. We used an EV threshold of 0.1.

Table 14: Betting results for top-5, top-10 and top-20 markets.

Market	Back bets	Back profit	Lay bets	Lay liab.	Lay profit
Top 5 Finish	572	133.2	1113	12047.2	136.3
Top 10 Finish	897	-66.3	1583	10987.7	130.3
Top 20 Finish	1211	52.9	1637	6801.7	106.8
Total	2680	119.8	4333	29836.6	373.4

The results suggest that the model can be used to identify bets that on average produce positive returns in both back and lay bets. The results are indicative, as we are relying on last traded prices, rather than executable prices, and whether the edge would survive this gap remains unknown. They should not be taken as evidence of profitability, as

a practical strategy would require bet sizing based on edge and risk, rather than the uniform stakes used in the example. The lay results may also be partly explained by market structure, where retail flow is concentrated on backing favourites, pulling their odds down towards tournament start. This is supported by the lay profits being larger when using a shorter time-window for calculating the average odds.

7 Conclusions

The results suggest that historical round-level performance contains substantial information about a player's future performance. In particular, the most recent rounds appear to carry the highest predictive value, while the rounds further in history contribute in stabilising the estimates. This supports the use of weighting schemes that place most of the weight on recent observations while still retaining some information from older rounds.

A key challenge in modelling golf scores is that raw round-level scores (or performances) are not directly comparable across tournaments, or even across rounds within the same tournament. The tournament schedule is structured in such way that only top-players qualify to the largest events, causing systematic differences in player levels across tournaments. Therefore, adjusting for field strength is essential when interpreting a player's performance relative to the competition faced.

However, adjusting for the differences in the field-strengths is not straightforward. In order to construct an estimate of a player's ability, we need the field strengths, and vice versa, creating a circular problem. In this thesis, this issue was addressed using an iterative procedure, where player estimates and field-strength-adjustments were updated jointly.

The player residual distributions were modelled separately for each player and assumed to be normally distributed. The round-level standard deviation was estimated using a blend of players' historical variance and the average variance across players. The results showed that the variance is only slightly persistent, and must be shrunk considerably towards population mean. The residual distributions showed skewness, which was evident in both histograms and descriptive statistics. However, when evaluating the results under the tournament-level negative log-likelihood metric, the results of the normal distribution were consistently better compared to the proposed skew-normal distribution.

The model produces point estimates of probabilities, but it does not provide confidence intervals around those probabilities. This is a relevant limitation if the model were to be used in practical decision-making such as betting. Not all probability estimates are equally reliable, as players with long and stable histories are easier to estimate than players with shorter or irregular histories. This was visible in the results, where the estimates were generally worse for weaker skilled players. This could be solved either by modelling these players separately, using more data from lower-tiered tours, or using a model which takes uncertainties better into account.

Overall, the main predictive signal is captured by a simple structure: field-strength-adjusted historical observations weighted toward recent rounds. Once this is implemented appropriately, further additions such as parameter fine-tuning, alternative weighting functions, weather adjustments, and course suitability variables appear to provide only marginal improvements. In this sense, the model seems to satisfy the Pareto-principle: most of the predictive power is obtained from a small number of

important steps, while the additional steps yield marginal benefits.

References

[1] Hardy, G. H., "A Mathematical Theorem about Golf," *The Mathematical Gazette*, vol. 29, pp. 226–227, Dec. 1945.

[2] Broadie, M., "Assessing Golfer Performance on the PGA Tour," *Interfaces*, vol. 42, pp. 146–165, 2012.

[3] Berry, S. M., "How Ferocious Is Tiger?" *Chance*, vol. 14, pp. 51–56, 2001.

[4] Connolly, R. A. and Rendleman, R. J., "Skill, Luck, and Streaky Play on the PGA Tour," *Journal of the American Statistical Association*, vol. 103, pp. 74–88, 2008.

[5] Data Golf, "Predictive Model Methodology," data-golf statistics and prediction service, Accessed 26.7.2026. https://datagolf.com/predictive-model-methodology

[6] Hunt, R., "How to Shoot Better Scores in Windy Conditions," *GolfWRX*, May 2015, blog post, Accessed 25.7.2026. https://www.golfwrx.com/297299/how-to-shoot-better-scores-in-windy-conditions/

[7] Hickman, D. C. and Metz, N. E., "The Impact of Pressure on Performance: Evidence from the PGA Tour," *Journal of Economic Behavior & Organization*, vol. 116, pp. 319–330, 2015.

[8] Visual Crossing Corporation, "Weather Data and Weather API," historical weather data service, Accessed 2.12.2025. https://www.visualcrossing.com/weather-api/

[9] Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D., *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. New York: Wiley, 1972.

[10] Scikit-learn, "IsotonicRegression," scikit-learn documentation, python package, Accessed 25.7.2026. https://scikit-learn.org/stable/modules/generated/sklearn.isotonic.IsotonicRegression.html

[11] Azzalini, A., "A Class of Distributions Which Includes the Normal Ones," *Scandinavian Journal of Statistics*, vol. 12, pp. 171–178, 1985.

[12] PGA Tour, "Multiple Factors Contribute to Improved Speed of Play," *PGATour.com*, Jun. 2025, news article, Accessed 25.7.2026. https://www.pgatour.com/article/news/latest/2025/06/03/multiple-factors-contribute-to-improved-speed-of-play-as-pga-tour-works-toward-further-advancements

[13] Hunt, R., "Want to Go Lower? The Stats Say You Need an Earlier Tee Time," *GolfWRX*, Aug. 2014, blog post, Accessed 25.7.2026. https://www.golfwrx.com/236637/want-to-go-lower-the-stats-say-you-need-an-earlier-tee-time/

[14] Rachesky, S., "USPGA Golf: The Impact of Tee Times," *Harvard Sports Analysis Collective*, Sep. 2014, blog post, Accessed 25.7.2026. https://harvardsportsanalysis.org/2014/09/uspga-golf-the-impact-of-tee-times/

[15] PGA Tour, "Average Approach Distance — Par," PGA Tour statistics site, Accessed 25.7.2026. https://www.pgatour.com/stats/detail/02339

[16] Sky Sports, "Golf Ball Rollback: Key Questions on What Is Happening, Why Change Is Coming and How It Will Affect You," *skysports.com*, Dec. 2023, news article, Accessed 25.7.2026. https://www.skysports.com/golf/news/12176/13024060/golf-ball-rollback-key-questions-on-what-is-happening-why-change-is-coming-and-how-it-will-affect-you