



# Forecasting the FIFA World Cup – Combining Result- and Goal-Based Team Ability Parameters

Pieter Robberechts<sup>(✉)</sup>  and Jesse Davis 

Department of Computer Science, KU Leuven, Leuven, Belgium  
{pieter.robberchts,jesse.davis}@cs.kuleuven.be

**Abstract.** In this study we compare result-based Elo ratings and goal-based ODM (Offense Defense Model) ratings as covariates in an ordered logit regression and bivariate Poisson model to generate predictions for the outcome of the 2018 FIFA World Cup. To this end, we first estimate probabilities of match results between all competing nations. With an evaluation on the four previous World Cups between 2002 and 2014, we show that an ordered logit model with Elo ratings as a single covariate achieves the best performance. Secondly, via Monte Carlo simulations we compute each team’s probability of advancing past a given stage of the tournament. Additionally, we apply our models on the Open International Soccer Database and show that our approach leads to good predictions for domestic league football matches as well.

**Keywords:** Football match outcome prediction ·  
Tournament simulation

## 1 Introduction

Association football (hereafter referred to simply as “football”) is currently the most popular spectator sport in the world [25]. This popularity can be partly explained by its unpredictability [15]. Because football is such a low-scoring game, a single event can make the difference between a win, a draw or a loss. Especially on the top-level, many games are decided by an extraordinary action of a single player, a rare defensive slip, a refereeing error, or just luck. As a consequence, many football games are ultimately won by the proclaimed underdog.

Despite the fact that a large part of the outcome of soccer matches is governed by chance, every team has its strengths and weaknesses and most results reflect these qualities. Eventually, skill does prevail and the best teams typically distinguish themselves during the course of a season or a tournament. This indicates that statistical forecasting methods could be used to predict the outcome of football matches.

Making these predictions is also one of the favorite pastimes of many football fans. However, they typically base their predictions on subjective opinions. It is

a challenging task to quantify the strength of a team objectively. There has previously been a fair amount of research on this topic. The Elo rating system is one of the most adopted approaches. While its origins are in chess [10], Elo ratings are commonly used for other sports, including football. Hvattum and Arntzen [24] have shown for English league football that an ordered logit model with the relative difference between the Elo ratings of two competing teams as a single covariate is a highly significant predictor of match outcomes. While the Elo system is essentially a result-based rating (i.e., Elo ratings are computed from the win-draw-loss records of a team), other rating systems are goal-based (i.e., they are based on the number of goals a team scores). These approaches typically extend ratings to two parameters – an offensive and defensive rating.

In this paper we compare both rating systems (as well as a combination of both) in terms of their predictive performance on previous World Cups and we provide our predictions for the 2018 World Cup. Our approach consists of three steps: First, we estimate the strength of the participating teams using past results. Second, we estimate probabilities of match results between each competing nation based on the pairwise rating differences. Finally, the predictions are used to determine the most likely tournament outcome in a Monte Carlo simulation.

The remainder of this paper is structured as follows: In Sect. 2, we discuss various proposals that have been made for modeling the outcome of football matches, as well as for rating the strength of teams; Next, Sect. 3, describes our models, followed in Sect. 4 by a discussion of three different performance metrics to compare the predictive power of these models. Finally, Sects. 5 and 6 validate the predictive strength of our models on respectively international and domestic league football.

## 2 Related Work

The literature on modeling the outcomes of football games can be divided in two broad categories, namely goal-based and result-based models. The first category models the number of goals scored and conceded by both competing teams. Predictions of win-draw-loss outcomes can then be derived indirectly by aggregating the probabilities assigned to all possible scorelines. The second category models win-draw-loss outcomes directly.

The simplest goal-based models assume that the number of goals scored by both teams are independent and can be modeled with two separate models. Poisson regression models are used predominately, but the negative binomial distribution [2] and Gaussian distribution (by employing the least squares regression method) [40] appear as well in the literature. For example, Lee et al. [31] applied a Poisson regression model to data from the 95/96 Premier League season, using the offensive and defensive strength of both teams and the home advantage as parameters. These parameters are then estimated using maximum-likelihood estimation on historic data. Although these independent Poisson models give a reasonably accurate description of football scores, they tend to underestimate

the proportion of draws. Maher et al. [34] were the first to identify a slight correlation between scores and therefore proposed bivariate Poisson models as an alternative. They showed that such a bivariate Poisson distribution gives a better fit on differences in scores than an independent model. Yet, they did not use this insight to predict the results of future matches. Dixon and Coles [7] made the same observation, but instead of using a bivariate Poisson model, they extended the independent Poisson model by introducing an ad-hoc adjustment to the probabilities of low-scoring games. Karlis and Ntzoufras [26] further developed the idea of bivariate Poisson distributions for forecasting football games.

Approaches that model the match outcome directly are a more recent development. Apart from their computational simplicity, these models have the advantage that they avoid the problem of having to model the interdependence between the scores of both teams [17]. Most studies in this category use discrete choice regression models, such as the ordered probit model and the ordered logit model [18, 28]. Goddard [17] compared a bivariate Poisson regression models with an ordered probit regression model and found that a hybrid model in which goal-based team performance covariates are used to forecast win-draw-loss match results yielded the best performance.

Although regression models are most common in the literature, any machine learning model could be used. For example, Groll et al. [21] found that a random forests model generally outperforms the conventional regression methods. Another popular class of models are Bayesian networks. Rue and Salvesen [38] proposed a Dynamic Bayesian Network (DBN) in order to take the time-dependent strength parameters of all teams in a league simultaneously into account. Baio and Blangiardo [1] extended this to a hierarchical goal-based Bayesian model. As such, they avoid the use of a more complex bivariate structure for the number of goals scored by assuming a common distribution at a higher level.

While the above studies focus on the actual prediction models, other studies have investigated the feasibility of possible covariates. Bookmaker odds are a first popular covariate. They reflect the (expert) predictions of bookmakers [37], who have strong economic incentives to make accurate predictions. Several studies have found that they are an efficient forecasting instrument [4, 14, 39].

A second popular covariate are ratings or rankings. The main idea is to estimate adequate ability parameters that reflect a team's current strength, based on a set of recent matches. A widely accepted approach in sports forecasting is the Elo rating system [10]. It has several generalisations, including the Glickman [16] and TrueSkill [22] rating systems. Besides Elo, there are numerous other approaches that fit into many categories [29]. Closely related to the regression-based forecasting models are the regression-based ranking methods. These models use maximum likelihood estimation to find adequate strength parameters for each team that can explain the number of goals scored or the win-tie-loss outcome in past games [33]. Other approaches are the Markov Chain based models such as Keener et al. [27] and the Power Rank rating system [45]; or the network based rating systems, such as the one by Park and Newman [36]. Yet another one

is the Sinkhorn-Knopp based ranking models such as the Offense Defense Model (ODM) [19]. Lasek et al. [30] found that these Elo ratings outperform several of these other ranking methods when predicting the outcome of individual games. Additionally, Van Haaren and Davis [43] found that Elo ratings perform well when predicting the final league tables of domestic football.

### 3 The Models

In this section, we describe the different components of our approach for predicting the outcome of football games. We computed both result-based Elo ratings and goal-based ODM ratings for each team based on past results. These ratings are then combined in an ordered logit regression or bivariate Poisson regression model in order to make predictions for future games. In the next sections, we compare various combinations of these rating systems and regression models.

#### 3.1 The Elo Rating System

We will briefly introduce both the basic Elo rating system and two football-specific modifications. An Elo rating system assigns a single number to each team that corresponds to a team's current strength. These numbers increase or decrease depending on the outcome of games and the ratings of the opponents in these games. Therefore, the Elo system defines an expected score for each team in a game, based on the rating difference with the opponent. Let  $R^H$  be the current rating of the home team and  $R^A$  the current rating of the away team, the exact formulae for the expected score  $E^H$  and actual score  $S^H$  of the home team are given by:

$$E^H = \frac{1}{1 + c^{R^H - R^A}/d} \quad \text{and} \quad S^H = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the match ended in a draw} \\ 0 & \text{otherwise} \end{cases}$$

The expected score and actual score for the away team are then respectively  $E^A = 1 - E^H$  and  $S^A = 1 - S^H$ .

When a team's actual score exceeds its expected score, this is seen as evidence that a team's current rating is too low and needs to be adjusted upward. Similarly, when a team's actual score is below its expected score, that team's rating is adjusted downward. Elo's original suggestion, which is still widely used, was a simple linear adjustment proportional to the amount by which a team overperformed or underperformed their expected score. The formula for updating the rating is

$$R'^H = R^H + k(S^H - E^H).$$

How much a team's rating increases or decreases is determined by both its expected score and the  $k$ -factor. The rating of a team that was expected to win by a large margin will therefore decrease with an accordingly large amount

if it actually loses. The  $k$ -factor is often called the recentness factor, because it determines how much weight is given to the results of recent matches. We added two additional factors: the competitiveness factor and the margin of victory. First, one of the difficulties in evaluating international football is that not all games are handled with the same seriousness. Friendlies, for example, are often used to experiment with new line-ups and players tend not to go to any extreme. Therefore, when computing the Elo ratings, we weight games differently depending on the importance of the competition. Second, because the best performing international teams play most of their matches against weak opponents (especially in European qualifiers) and record very few losses, we take the margin of victory (i.e., the absolute goal difference) into account [24]. So, we replace  $k$  by the expression

$$k = k_0 w_i (1 + \delta)^\gamma \quad (1)$$

with  $\delta$  the absolute goal difference,  $w_i > 0$  a weight factor corresponding to the competitiveness of the competition,  $k_0$  the recentness factor and  $\gamma > 0$  a fixed parameter determining the impact of the margin of victory on the update rule.

There are five parameters in this rating system. The parameters  $c$  and  $d$  determine the scale of the ratings. We set them respectively to 10 and 400. Other values lead to the same rating system, but one has to determine matching weight parameters  $k_0$ ,  $w$  and  $\gamma$ . The optimal values for  $k_0$  and  $\gamma$  are determined from historical data. We explain this procedure in Sect. 3.3. The values for  $w$  are application dependent and based on expert knowledge.

### 3.2 Offense-Defense Ratings

An important factor in football games is the playing style of both teams and the balance between offense and defense. We argue that, besides the relative strength of both teams, this difference in playing style might be an important factor in deciding the final outcome of a game. For example, a game between two teams that are known to rely on a very strong defense might have a higher probability to end up in a draw than a game between two teams that are known to play very offensively.

A rating system that can capture these differences in offensive and defensive strengths of a team is the Offense Defense Model (ODM) [19]. As opposed to the Elo system discussed above, it captures the offensive and defensive strength of a team as two separate parameters. Therefore, it uses goals scored as a measure of offensive strength and goals conceded as a measure of defensive strength. Whether a game eventually results in a win, a tie or a loss does not affect the ratings. We will again briefly discuss the basic ODM rating system, followed by a couple of modifications for its application to international football.

Define  $A_{ij}$  as the goals scored by team  $j$  against team  $i$ . The offensive and defensive ratings of team  $j$  are

$$o_j = \sum_{i=1}^n \frac{A_{ij}}{d_i} \quad \text{and} \quad d_j = \sum_{i=1}^n \frac{A_{ji}}{o_i}.$$

Since the offensive and defensive ratings are interdependent, they must be approximated by an iterative refinement procedure. We refer to the original paper for details.

This approach works for domestic league football where each team plays the same number of games against every other team. In international football, however, there can be large disparities between the number of games played and the strength of the opponents. A team that plays few games against strong opponents will likely score fewer goals and concede more goals, which leads to weaker attack and weaker defence ratings. To address these problems, we update ratings sequentially. In each game, a team has two sets of ratings: pre-game ratings and post-game ratings. The pre-game ratings are a weighted sum of a team's post-game ratings in previous games. Similar to what we did for the Elo rating system, these weights are determined by the recentness of a game and its competitiveness. To compute the post-game ratings, we apply the iterative procedure from the original ODM as if the competition has only two teams and using the pre-game ratings as initial estimates for the offensive and defensive ratings. Algorithm 1 below defines the exact procedure.

---

**Algorithm 1.** Computing post-game ratings from pre-game ratings

---

```

1: procedure SCALE( $A, x$ )
2:    $y \leftarrow A \frac{1}{x}$ 
3:    $x_{\text{post}} \leftarrow A^T \frac{1}{y}$ 
4:    $y_{\text{post}} \leftarrow A \frac{1}{x}$ 
5:   return  $x_{\text{post}}, y_{\text{post}}$ 
6: end procedure

7:  $A_{ij} \leftarrow$  score team  $j$  generated against team  $i$ 
8:  $o_{\text{post}} \leftarrow [o_1, o_2]$  ▷ The pre-game offensive ratings of both teams
9:  $d_{\text{post}} \leftarrow [d_1, d_2]$  ▷ The pre-game defensive ratings of both teams
10: for  $k=0, \dots, \text{nb.iter}$  do
11:    $o_{\text{rel}_o}, d_{\text{rel}_o} \leftarrow \text{scale}(A, o_{\text{post}})$ 
12:    $d_{\text{rel}_d}, o_{\text{rel}_d} \leftarrow \text{scale}(A^T, d_{\text{post}})$ 
13:    $o_{\text{post}} \leftarrow o_{\text{rel}_o} + o_{\text{rel}_d}$ 
14:    $d_{\text{post}} \leftarrow d_{\text{rel}_o} + d_{\text{rel}_d}$ 
15: end for

```

---

### 3.3 Match Result Predictions

The rating systems defined above can be combined with a regression model to obtain predictions for future matches. Therefore, we consider the rating differences  $R^H - R^A$ ,  $o^H - d^A$  and  $d^H - o^A$  as covariates. Additionally, a fourth covariate indicates whether a home advantage applies to the home team. In Sect. 5, we compare the predictive power of an ordered logit regression model [35] and a bivariate Poisson regression model [26], as well as various combinations of these covariates in terms of their predictive power on previous World Cups.

We use the L-BFGS-B algorithm [5] with the Ranked Probability Score (RPS) [11] as a loss function to determine the optimal set of parameters for these models. This approach allows us to jointly optimize the parameters for both the rating systems and regression models. Therefore, we order the games in our dataset chronologically and define two subsets: a validation set and a test set. The test set contains the matches that we would like to predict and ideally the validation set contains matches from previous editions of the same tournament or league. Then we repeatedly evaluate each match in the complete dataset sequentially, updating the ratings based on the actual outcome and making a prediction for the matches in the validation set. Once all matches are evaluated, we compute the RPS on the validation set and update the parameters for both the rating system and regression model in order to minimize the RPS.

## 4 Evaluation Procedures

In this section, we consider several evaluation measures to compare our models with each other and to the odds determined by bookmakers. These odds can serve as a natural benchmark for our models. In contrast to the methods above, bookmaker odds are not solely based on results in past games. They include expert judgments from the bookmakers, which have a strong economic motivation to rate the competitors accurately [32]. After removing the profit margin of the bookmaker, the inverted odds can be interpreted as outcome probabilities [20].

Both our models and the bookmaker odds have in common that they assign a probability to all three possible outcomes of a match. One can evaluate these probabilities in three ways: First, one can consider the outcome with the highest assigned probability as the predicted outcome. Second, one can look at the probability that was assigned to the true outcome. As a third evaluation, one can judge these three probabilities as a whole. Each of these evaluations leads to a different evaluation measure, which we define below.

In the following paragraphs, we use the ordered vector  $\hat{p} = (p_1, p_2, p_3)$  to denote a probability forecast of all possible match outcomes  $r = (\text{win}, \text{tie}, \text{loss})$ . Additionally,  $y = (y_1, y_2, y_3)$  denotes the true outcome of a match, with  $y_i$  a binary indicator of whether or not  $i$  is the true outcome.

**Accuracy.** This measure compares the outcome with the highest assigned probability to the true outcome. The accuracy for a single match is computed using the following formula:

$$\mathbb{1}[\arg \max_i y_i = \arg \max_i \hat{p}_i] \quad (2)$$

where  $\mathbb{1}[\cdot]$  is the indicator function that equals 1 if the statement between brackets holds, and 0 otherwise.

**Logarithmic loss.** This measure the uncertainty of the prediction based on how much it varies from the actual outcome. The logarithmic loss is computed as

$$-\sum_{i=1}^{|r|} y_i \log \hat{p}_i \quad (3)$$

with  $|r|$  the number of possible outcomes. A perfect classifier would have a logarithmic loss of precisely zero. Less ideal classifiers have progressively larger values.

**Ranked Probability Score (RPS).** The ranked probability score (RPS) was introduced in 1969 by Epstein [11] to evaluate probability forecasts of ranked categories. In contrast to the two previous measures it explicitly accounts for the ordinal structure of the predictions. This means that predicting a tie when the actual outcome is a loss is considered a better prediction than a win. For our purpose, it can be defined as

$$\frac{1}{|r| - 1} \sum_{k=1}^{|r|-1} \left( \sum_{l=1}^k (\hat{p}_l - y_l) \right)^2 \quad (4)$$

As the RPS is an error measure, a lower value corresponds to a better fit.

We will use the same metrics to evaluate the prediction of a tournament outcome. For that purpose, we define the set of possible outcomes as  $r = (\text{elimination in the group stage, elimination in the round of } 16, \dots, \text{win})$ .

## 5 Validation on Previous World Cups

This section evaluates the predictive performance of our models on the four FIFA World Cups between 2002 and 2014. Therefore, we adopt a leave-one-out procedure where we iteratively tune the parameters of our model on three out of the four world cups and evaluate the performance on the left out one. For example, while predicting the matches of the 2010 World Cup (i.e., the test set), we use all available international games played between the end of the second world war and the start of the 2018 World Cup to determine the ratings for each team; but only matches of the 2002, 2004 and 2014 World Cups to tune the parameters (i.e., the validation set). Although we use data from future World Cups to determine the parameter settings for a previous World Cup, note that we only include information from past games when rating teams. Therefore, the predictions do not depend on results in future matches.

Our dataset was scraped from <http://eloratings.net> and includes all international games played between the end of the second world war (January 1, 1946) and the start of the 2018 World Cup (June 13, 2018). For each of these games, we scraped the competing teams, the date of the game, the competition, the outcome after the official game time, the outcome after extensions or penalties, and whether a home advantage applies. Additionally, we scraped the average



assigned three-way odds by multiple bookmakers from <http://betexplorer.com> for all<sup>1</sup> World Cup matches between 2002 and 2014.

We classified all international competitions into three categories, corresponding with their competitiveness and relative importance. Each category is assigned a weight, which corresponds to  $w$  in Formula 1. We assigned a “very high” weight ( $w = 1$ ) to World Cup games; a “high” weight ( $w = 0.833$ ) to each of the six (AFC, CAF, CONCACAF, CONMEBOL, OFC and UEFA) continental championships; a “medium” weight ( $w = 0.66$ ) to their qualifiers as well as to the World Cup qualifiers; a “low” weight ( $w = 0.5$ ) to the less important tournaments, such as the African Games, Balkan Cup,...; and finally a “very low” weight ( $w = 0.33$ ) to Friendlies. These five categories originate from the FIFA/Coca-Cola World Ranking. In comparison, we added an additional category of very low importance for friendlies (the FIFA ranking gives friendlies the same weight as small tournament games) and used different weights.

Table 1 shows the predictive performance of the models introduced in Sect. 3 on the individual games from these four World Cups. Additionally, we include the averaged bookmaker predictions and the predictions from the best-performing random forest model from Groll et al. [21] as a baseline. This random forest model includes amongst others the bookmaker odds and Elo ratings as covariates. We updated the ratings of each team after each game based on the true outcome. This approach is in line with how bookmaker odds are updated until close before the start of a game. It turns out that the simple ordered logit regression model with the Elo rating difference and home advantage as covariates outperforms the bookmaker predictions and all other models in terms of accuracy, logarithmic loss and RPS. Furthermore, we notice that the 2006 and 2014 are a lot easier to predict than the 2002 and 2010 World Cups. The ODM-based models perform slightly better on these hard to predict World Cups.

**Table 1.** Validation of different predictive models on individual games in the 2002, 2006, 2010 and 2014 World Cups. The models are ordered by increasing average RPS value. The last column shows the average RPS for each World Cup.

Team	Accuracy	Log loss	RPS	avg ● 2002 ■ 2006 ▲ 2010 ◆ 2014
ELO ordered logit	0.5938	0.9375	0.1860	
ELO bivariate Poisson	0.5898	1.0045	0.1866	
Random forest [21]	0.5560		0.1870	
Bookmakers	0.5240	0.9594	0.1877	
ELO+ODM ordered logit	0.5820	0.9437	0.1878	
ELO+ODM bivariate Poisson	0.5703	0.9637	0.1917	
ODM ordered logit	0.5586	0.9657	0.1949	
ODM bivariate Poisson	0.5664	0.9755	0.1968	

<sup>1</sup> The odds of six games were missing.

Next, we used these models to predict the tournament course for the World Cups between 2002 and 2014. Given match outcome probabilities for each possible match-up, we ran 20,000 Monte Carlo simulations for each World Cup. Occasionally two or more teams will finish the group phase on the same points tally. In that case, the FIFA defines a couple of tie-breakers to determine the ranking order, primarily based on the number of goals scored. However, since the ordered logit models can only sample win, draw and loss results, we resolve these ties randomly. Similarly, in the case of a draw in the knockout stage, we simulate the extra time by sampling a second result. The bivariate Poisson models make it possible to resolve these equal point tallies according to the official rules, but this does not lead to more accurate predictions.

Table 2 presents the performance of the Elo-based and Elo+ODM-based ordered logit models on the past World Cups. In contrast to the previous experiment, these are pre-tournament predictions, meaning that only games preceding the corresponding World Cup are considered when making these predictions. For comparison, we added the 2014 World Cup predictions by FiveThirtyEight [12]. The model using the Elo ratings as a single covariate is again the best performing one, but it turns out that these tournament forecasts are quite inaccurate. We could predict the actual round of elimination correctly for only about half of the participating teams.

**Table 2.** Performance of a Monte Carlo simulation on previous World Cups, using the Elo and Elo+ODM-based ordered logit models. Additionally, we provide the 2014 World Cup predictions of FiveThirtyEight [12] for comparison.






World Cup Model		Accuracy	Log loss	RPS	Correct by round <sup>a</sup>
2014	Elo	0.5000	0.2020	0.1364	10/5/3/1/0
	Elo+ODM	0.4375	0.2075	0.1381	10/4/3/1/0
	FiveThirtyEight	0.4063	0.1947	0.1334	8/6/3/1/0
2010	Elo	0.5625	0.1945	0.1330	12/4/2/1/1
	Elo+ODM	0.5313	0.1878	0.1324	11/5/2/1/1
2006	Elo	0.5313	0.1902	0.1317	12/6/2/0/0
	Elo+ODM	0.4063	0.2125	0.1422	11/5/1/1/0
2002	Elo	0.5000	0.1962	0.2322	11/4/0/0/0
	Elo+ODM	0.4688	0.2471	0.1456	10/4/0/0/0

<sup>a</sup> Number of teams correctly picked to advance to the next round as respectively *round of 16/quarter final/semi final/final/win*. For example, *11/4/3/1/0* means that 11 of the 16 teams advancing to the round of 16 were predicted.

Finally, we applied the ordered logit model with both Elo and ODM ratings as covariates to forecast the 2018 World Cup. According to our simulation, Brazil was the clear favorite with a win probability of 33% followed by Germany, Spain, France and Argentina. These results were in line with the bookmaker

odds, although the bookmakers were more conservative about the win probability of Brazil. Table 3 shows these probabilities for the five favourites. For a detailed overview, we refer to our interactive at <https://dtai.cs.kuleuven.be/sports/worldcup18/>.

**Table 3.** Estimated probabilities for reaching the different stages in the 2018 World Cup for the five most-likely winners based on 20,000 simulations.

	Team	Sixteen	Quarter	Semi	Final	Win	Bookmakers
	Brazil	92%	73%	59%	46%	33%	15%
	Germany	81%	57%	41%	28%	15%	15%
	Spain	81%	70%	46%	26%	14%	12%
	France	79%	58%	41%	21%	11%	12%
	Argentina	77%	51%	32%	16%	7%	8%

The 2018 World Cup caught the attention of several other data scientist, trying to forecast the tournament outcome. Based on their SPI rating system, FiveThirtyEight [13] forecasted Brazil (19%) to win the World Cup, followed by Spain (17%) and Germany (13%). The same teams were determined as the major favorites by Zeileis, Leitner and Hornik [44]. By aggregating the winning odds of several bookmakers and transforming those into winning probabilities, they obtained a win probability of 16.6% for Brazil, 15.8% for Germany and 12.5% for Spain. The Swiss bank UBS [42] came up with the same three favorites, but in a different order. They obtained Germany as the main favorite (24.0%), followed by Brazil (19.8%) and Spain (12.5%). Also Groll et al. [21] came up with Spain (17.8%), Germany (17.1%) and Brazil (12.3%) as the main favorites. They combined a large set of 16 features with a random forest approach. As far as we know, only EA Sports [9] correctly predicted France as the World Cup winner. Yet, they did not publish any win probabilities. In Table 4 we compare these models with ours, looking both at predictive accuracy for individual games (if available) and the accuracy of the pre-tournament simulation. To allow comparison with FiveThirtyEight’s predictions, we convert the win-tie-loss probabilities for games in the knockout stage to win-loss probabilities. Therefore, we use the formula  $p'_{win} = p_{win} + p_{win}/(p_{win} + p_{loss}) * p_{tie}$  and analogous for  $p'_{loss}$ .

## 6 Validation on Domestic League Football

We also verified our models on The Open International Soccer Database that was provided as part of the 2017 Soccer Prediction Challenge [8]. The training set incorporates 216,743 match outcomes, with missing data (as part of the challenge), from 52 football leagues from all over the world in the seasons ranging from 2000–2001 until 2017–2018. The challenge involved using a single model to predict 206 future match outcomes from 26 different leagues. We used the 2010–2011 until 2017–2018 seasons of the training set as our test set to determine

**Table 4.** A comparison of our model’s predictions with others for the 2018 World Cup. Both the predictive accuracy of individual games (if available) and of the pre-tournament forecast are listed.

	Individual games			Tournament simulation			
	Accuracy	Log loss	RPS	Accuracy	Log loss	RPS	Correct by round <sup>a</sup>
Bookmakers	0.6094	0.8463	<b>0.1976</b>				
FiveThirtyEight [13]	<b>0.6250</b>	<b>0.8457</b>	0.1976	0.5313	<b>0.1822</b>	<b>0.1242</b>	14/4/1/0/0
Zeileis et al. [44]				0.5625	0.1845	0.1269	14/4/1/0/0
Groll et al. [21]				<b>0.5938</b>	0.1855	0.1261	<b>14/4/2/0/0</b>
ELO+ODM logit	0.6094	0.8865	0.2072	0.5625	0.2243	0.1315	14/4/1/0/0
UBS [42]				0.5000	0.2009	0.1923	13/4/1/0/0

<sup>a</sup> Number of teams correctly picked to advance to the next round as respectively *round of 16/quarter final/semi final/final/win*. For example, *11/4/3/1/0* means that 11 of the 16 teams advancing to the round of 16 where predicted.

the optimal parameters for our models. This corresponds to about half of the training data.

Table 5 compares the performance of our best-performing models to the four best results of the competition and the bookmaker odds. As for the World Cup predictions, these bookmaker odds are the average assigned three-way odds by multiple bookmakers scraped from <http://betexplorer.com>. Although we did not optimize our approach for domestic league football, it was found that our relatively simple model outperforms all other, more complex models in terms of RPS. Three models report a better accuracy. While we did not verify this, we think that the predictive accuracy could be improved by incorporating league-specific home advantages and (because of transfers) allowing faster rating updates after the summer and winter breaks.

**Table 5.** Our approach compared to the best performing models from the 2017 Soccer Prediction Challenge [8]. We omit the logarithmic loss, because we do not have it for the 2017 Soccer Prediction Challenge submissions.

Model	RPS	Acc
Bookmakers	0.2020 —	0.5194 —
ELO ordered logit	0.2035 —	0.5146 —
ELO+ODM ordered logit	0.2045 —	0.5146 —
Berrar et al. [3]	0.2054 —	0.5194 —
Hubáček et al. [23]	0.2063 —	0.5243 —
Constantinou [6]	0.2083 —	0.5146 —
Tsokos et al. [41]	0.2087 —	0.5388 —

## 7 Conclusion

In this work, we compared several models for match outcome prediction and tournament simulation in football. We considered all possible combinations of a

result- and goal-based regression model, with result-based Elo rating and goal-based ODM rating differences as covariates. In conclusion, we found that a very basic Elo-based ordered logit model outperforms all other models, including more complex models from the literature. The outcome of any match is unpredictable enough to confound these sophisticated computer models.

**Acknowledgements.** PR is supported by Interreg V A project NANO4Sports. JD is partially supported by KU Leuven Research Fund (C14/17/070 and C22/15/015), FWO-Vlaanderen (SBO-150033) and Interreg V A project NANO4Sports.

## References

1. Baio, G., Blangiardo, M.: Bayesian hierarchical model for the prediction of football results. *J. Appl. Stat.* **37**(2), 253–264 (2010). <https://doi.org/10.1080/02664760802684177>
2. Baxter, M., Stevenson, R.: Discriminating between the poisson and negative binomial distributions: an application to goal scoring in association football. *J. Appl. Stat.* **15**(3), 347–354 (1988). <https://doi.org/10.1080/02664768800000045>
3. Berrar, D., Dubitzky, W., Lopes, P.: Incorporating domain knowledge in machine learning for soccer outcome prediction. *Mach. Learn.* **108**(1), 97–126 (2019)
4. Boulier, B.L., Stekler, H.O.: Predicting the outcomes of National Football League games. *Int. J. Forecast.* **19**(2), 257–270 (2003). [https://doi.org/10.1016/S0169-2070\(01\)00144-3](https://doi.org/10.1016/S0169-2070(01)00144-3)
5. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. *SISC* **16**(5), 1190–1208 (1995). <https://doi.org/10.1137/0916069>
6. Constantinou, A.C.: Dolores: a model that predicts football match outcomes from all over the world. *Mach. Learn.* **108**(1), 49–75 (2018). <https://doi.org/10.1007/s10994-018-5703-7>
7. Dixon, M.J., Coles, S.G.: Modelling association football scores and inefficiencies in the football betting market. *J. Royal Stat. Soc. Ser. C (Appl. Stat.)* **46**(2), 265–280 (1997)
8. Dubitzky, W., Lopes, P., Davis, J., Berrar, D.: The open international soccer database for machine learning. *Mach. Learn.* **108**(1), 9–28 (2019)
9. EA Sports: EA Sport predicts France to win the FIFA World Cup, May 2018. <https://www.easports.com/fifa/news/2018/ea-sports-predicts-world-cup-fifa-18>
10. Elo, A.E.: *The Rating of Chess Players, Past and Present*. Arco Pub., New York (1978)
11. Epstein, E.S.: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.* **8**(6), 985–987 (1969). [https://doi.org/10.1175/1520-0450\(1969\)008<0985:ASSFPF>2.0.CO;2](https://doi.org/10.1175/1520-0450(1969)008<0985:ASSFPF>2.0.CO;2)
12. FiveThirtyEight: 2014 World Cup Predictions, June 2014. <https://fivethirtyeight.com/interactives/world-cup/>
13. FiveThirtyEight: 2018 World Cup Predictions, June 2018. <https://projects.fivethirtyeight.com/2018-world-cup-predictions/>
14. Forrest, D., Goddard, J., Simmons, R.: Odds-setters as forecasters: the case of English football. *Int. J. Forecast.* **21**(3), 551–564 (2005). <https://doi.org/10.1016/j.ijforecast.2005.03.003>

15. Forrest, D., Simmons, R.: Outcome uncertainty and attendance demand in sport: the case of English soccer. *J. Royal Stat. Soc.* **51**(2), 229–241 (2002). <https://doi.org/10.1111/1467-9884.00314>
16. Glickman, M.E.: Parameter estimation in large dynamic paired comparison experiments. *J. Royal Stat. Soc. Ser. C (Appl. Stat.)* **48**(3), 377–394 (2002). <https://doi.org/10.1111/1467-9876.00159>
17. Goddard, J.: Regression models for forecasting goals and match results in association football. *Int. J. Forecast.* **21**(2), 331–340 (2005). <https://doi.org/10.1016/j.ijforecast.2004.08.002>
18. Goddard, J., Asimakopoulou, I.: Forecasting football results and the efficiency of fixed-odds betting. *J. Forecast.* **23**(1), 51–66 (2004). <https://doi.org/10.1002/for.877>
19. Govan, A.Y., Langville, A.N., Meyer, C.D.: Offense-defense approach to ranking team sports. *J. Q. Anal. Sports* **5**(1) (2009). <https://doi.org/10.2202/1559-0410.1151>
20. Graham, I., Stott, H.: Predicting bookmaker odds and efficiency for UK football. *Appl. Econ.* **40**(1), 99–109 (2008). <https://doi.org/10.1080/00036840701728799>
21. Groll, A., Ley, C., Schaubberger, G., Van Eetvelde, H.: Prediction of the FIFA World Cup 2018 - a random forest approach with an emphasis on estimated team ability parameters. [arXiv:1806.03208](https://arxiv.org/abs/1806.03208) [stat], June 2018
22. Herbrich, R., Minka, T., Graepel, T.: TrueSkill<sup>TM</sup> : a Bayesian skill rating system. In: Schölkopf, B., Platt, J.C., Hoffman, T. (eds.) *Advances in Neural Information Processing Systems*, vol. 19, pp. 569–576. MIT Press (2007)
23. Hubáček, O., Sourek, G., Železný, F.: Learning to predict soccer results from relational data with gradient boosted trees. *Mach. Learn.* **108**(1), 29–47 (2019). <https://doi.org/10.1007/s10994-018-5704-6>
24. Hvattum, L.M., Arntzen, H.: Using ELO ratings for match result prediction in association football. *Int. J. Forecast.* **26**(3), 460–470 (2010). <https://doi.org/10.1016/j.ijforecast.2009.10.002>
25. Joy, B., Weil, E., Giulianotti, R.C., Alegi, P.C., Rollin, J.: Football. <https://www.britannica.com/sports/football-soccer>
26. Karlis, D., Ntzoufras, I.: Analysis of sports data by using bivariate Poisson models. *J. Royal Stat. Soc.* **52**(3), 381–393 (2003). <https://doi.org/10.1111/1467-9884.00366>
27. Keener, J.: The Perron–Frobenius theorem and the ranking of football teams. *SIAM Rev.* **35**(1), 80–93 (1993). <https://doi.org/10.1137/1035004>
28. Kuypers, T.: Information and efficiency: an empirical study of a fixed odds betting market. *Appl. Econ.* **32**(11), 1353–1363 (2000). <https://doi.org/10.1080/00036840050151449>
29. Langville, A.N., Meyer, C.D.: *Who’s #1?: The Science of Rating and Ranking*. Princeton University Press, Princeton (2012)
30. Lasek, J., Szlavik, Z., Bhulai, S.: The predictive power of ranking systems in association football. *Int. J. Appl. Pattern Recogn.* **1**(1), 27–46 (2013). <https://doi.org/10.1504/IJAPR.2013.052339>
31. Lee, A.J.: Modeling scores in the premier league: is Manchester united really the best? *Chance* **10**(1), 15–19 (1997). <https://doi.org/10.1080/09332480.1997.10554791>
32. Leitner, C., Zeileis, A., Hornik, K.: Forecasting sports tournaments by ratings of (prob)abilities: a comparison for the EURO 2008. *Int. J. Forecast.* **26**(3), 471–481 (2010). <https://doi.org/10.1016/j.ijforecast.2009.10.001>

33. Ley, C., Van de Wiele, T., Van Eetvelde, H.: Ranking soccer teams on basis of their current strength: a comparison of maximum likelihood approaches. eprint [arXiv:1705.09575](https://arxiv.org/abs/1705.09575), May 2017
34. Maher, M.J.: Modelling association football scores. *Stat. Neerl.* **36**(3), 109–118 (1982). <https://doi.org/10.1111/j.1467-9574.1982.tb00782.x>
35. McCullagh, P.: Regression models for ordinal data. *J. Royal Stat. Soc.* **42**(2), 109–142 (1980)
36. Park, J., Newman, M.E.J.: A network-based ranking system for American college football. *J. Stat. Mech. Theory Exp.* **2005**(10), P10014–P10014 (2005). <https://doi.org/10.1088/1742-5468/2005/10/P10014>
37. Pope, P.F., Peel, D.A.: Information, prices and efficiency in a fixed-odds betting market. *Economica* **56**(223), 323–341 (1989). <https://doi.org/10.2307/2554281>
38. Rue, H., Salvesen, O.: Prediction and retrospective analysis of soccer matches in a league. *J. Royal Stat. Soc. Ser. D (Stat.)* **49**(3), 399–418 (2000). <https://doi.org/10.1111/1467-9884.00243>
39. Spann, M., Skiera, B.: Sports forecasting: a comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *J. Forecast.* **28**(1), 55–72 (2008). <https://doi.org/10.1002/for.1091>
40. Stefani, R.T.: Improved least squares football, basketball, and soccer predictions. *IEEE Trans. Syst. Man Cybern.* **10**(2), 116–123 (1980). <https://doi.org/10.1109/TSMC.1980.4308442>
41. Tsokos, A., Narayanan, S., Kosmidis, G.I.B., Cucuringu, M., Whitaker, G., Kiraly, F.: Modeling outcomes of soccer matches. *Mach. Learn.* **108**(1), 77–95 (2019)
42. UBS AG: and the winner is.... investing in emerging markets (special edition, 2018 World Cup in Russia), May 2018
43. Van Haaren, J., Davis, J.: Predicting the final league tables of domestic football leagues. In: *Proceedings of the 5th International Conference on Mathematics in Sport*, pp. 202–207 (2015)
44. Zeileis, A., Leitner, C., Hornik, K.: Probabilistic forecasts for the 2018 FIFA World Cup based on the bookmaker consensus model, p. 19
45. Zyga, L.: New algorithm ranks sports teams like Google’s PageRank, December 2009. <https://phys.org/news/2009-12-algorithm-sports-teams-google-pagerank.html>