IMA Journal of Management Mathematics (2022) **33**, 1–18
https://doi.org/10.1093/imaman/dpab029
Advance Access publication on 12 August 2021

Forty years of score-based soccer match outcome prediction: an experimental review

Ondřej Hubáček†, Gustav Šourek and Filip Železný
Faculty of Electrical Engineering, Department of Computer Science, Czech Technical University in Prague, Technická 2, 166 27 Prague 6
†Corresponding author: hubacon2@fel.cvut.cz

[Received on 28 November 2019; accepted on 09 July 2021]

We investigate the state-of-the-art in score-based soccer match outcome modelling to identify the top-performing methods across diverse classes of existing approaches to the problem. Namely, we bring together various statistical methods based on Poisson and Weibull distributions and several general ranking algorithms (Elo, Steph ratings, Gaussian-OD ratings) as well as domain-specific rating systems (Berrar ratings, pi-ratings). We review, reimplement and experimentally compare these diverse competitors altogether on the largest database of soccer results available to identify true leaders. Our results reveal that the individual predictions, as well as the overall performances, are very similar across the top models tested, likely suggesting the limits of this generic approach to score-based match outcome modelling. No study of a similar scale has previously been done.

Keywords: forecasting; soccer; score-based models; review.

1. Introduction

Soccer, being arguably the most popular sport in the world, continues to attract researchers and practitioners competing for the design of the most accurate game outcome forecasting models. Indeed, there has been a plethora of such models published in the past decades. However, due to a lack of a standardized dataset, it has been difficult to draw conclusive statements about relative performances of the diverse approaches. Creation of such a dataset has been, however, further complicated by the fact that the proposed models often utilize varying details of match and background information in order to gain more advantage over the competition.

While for some of the top leagues complete information about the game, including player-tracking data, can be obtained, such an approach does not generalize onto the vast amount of the lower leagues, where merely the results with basic metadata are all that is being stored for each match. Moreover, the fine-grained data are often proprietary and rather expensive, rendering them unsuitable for use in academic benchmarks.

For the purpose of a sound experimental comparison, we propose to target the broadest possible scope of the domain by considering solely the *score-based* models, i.e. the models that use the final scores, teams' names and dates as the only input covariates. Such an elementary modelling paradigm allows us to calculate predictions for virtually all existing matches and, consequently, unify the training and evaluation protocol across the diverse approaches.

Conveniently, a large dataset containing 218,916 match results from 52 leagues since the season 2000/01 was released recently by Dubitzky *et al.* (2019). The records in the dataset consist merely of the league names, dates, team names and the resulting scores. The availability of such a large dataset provides an ideal opportunity to finally shed some light on the relative performance of the respective

score-based state-of-the-art methods. For that purpose, we have reimplemented the most promising models from the literature to analyse their performance under a unified protocol.

The rest of the paper is organized as follows. In Section 2, we summarize the relevant research. Section 3 provides a brief description of the implemented models. Section 4 explains the protocol for fitting and evaluation of the models. Experimental results are compiled in Section 5, and we conclude the paper in Section 6.

2. Related Work

The body of related work on score-based predictive models can be generally divided into three categories: (i) *statistical models*, where the goals scored are assumed to follow a particular parametric probability distribution; (ii) *rating systems* that assign a real-valued rating(s) to each team to capture its strength; and (iii) *machine learning models* where various complex features are usually derived from the data and passed to an off-the-shelf learning algorithm.

2.1 *Statistical models*

The research in the domain of score-based soccer modelling has traditionally been dominated by statistical approaches. In his pioneering work, Maher (1982) came up with a double Poisson model and bivariate Poisson model, where the latter provided a better fit for the data. Maher also introduced the notion of teams' attacking and defensive strengths and how to use them for forecasting of the match results. This notion is still used in the current research nowadays.

Dixon & Coles (1997) extended Maher's ideas as they introduced a dependency between the home and away teams' goals scored for the double Poisson model, effectively increasing the probabilities of low-scoring draws. Also, while Maher considered the strength of the team to be time invariant, here the idea of weighting the likelihood during fitting of the model was introduced. Particularly, the authors used exponential time weighting to discount the effects of past results. The simplicity of exponential time weighting allows for its use with other models too (Ley *et al.*, 2019). A different approach to the time evolution of teams was used in Rue & Salvesen (2000), where the authors used a Brownian motion to tie together the teams' strength parameters in consecutive rounds. Crowder *et al.* (2002) used an autoregressive model for the evolution of teams' strengths, improving on results by Dixon & Coles (1997) and on the computational complexity of Rue & Salvesen (2000). A static hierarchical model based on the double Poisson distribution was introduced by Baio & Blangiardo (2010), claiming that performance is non-inferior to the bivariate Poisson model (Karlis & Ntzoufras, 2003). Owen (2011) used a random walk to model the teams strength evolution in the double Poisson model; however, a comparison against the established likelihood weighting approach was not done. Koopman & Lit (2015) introduced time dynamics into the bivariate Poisson model using a state-space model representation. The authors also pointed out that the dependency between scores had little effect on the out-of-sample forecasting performance of the model. This observation was latter supported by Ley *et al.* (2019). Angelini & De Angelis (2017) investigated another technique for implementing the time-dynamics with a PARX model (Agosto *et al.*, 2016). The PARX model outperformed Dixon & Coles (1997) in forecasting the number of scored goals. Koopman & Lit (2019) compared bivariate Poisson, Skellam and ordered probit models where the teams' strengths were updated according to a time series model. The bivariate Poisson model achieved the best results.

Karlis & Ntzoufras (2003) noticed that the bivariate Poisson models tend to underestimate the probabilities of draws and introduced a diagonal-inflated bivariate Poisson model. Karlis & Ntzoufras

(2008) then eliminated the need to explicitly model the scores dependency by using the Skellam distribution (Skellam, 1946), where the evolution of the teams' strengths was implemented using Bayesian updates. McHale & Scarf (2011) experimented with negative binomial and bivariate Poisson models where the dependence structure was implemented using copulas. The most recent novelty in statistical approaches is the use of bivariate Weibull count model (Boshnakov *et al.*, 2017). Unlike in the Poisson distribution, where the mean is equal to the variance, the Weibull count distribution is determined by two parameters, allowing for better handling of both under and over-dispersed data. The bivariate model is constructed using a copula function. The model provides a better fit for the data than the Poisson model at the expense of higher computational time, as the calculation of the probability density function of the Weibull count model is very demanding. A great review of the statistical approaches can be found in Ley *et al.* (2019).

2.2 *Rating systems*

Another technique to estimate the strength of an individual or a team are the so-called rating systems. Ratings try to capture the team's strength into one or two scalar values, providing relative ordering of the teams, but not necessarily a way to obtain the probabilistic forecasts. The world's best-known rating system is the Elo rating (Elo, 1978), originally used for assessing the strength of chess players. The player's performance is assumed to be drawn from a Gaussian distribution with a fixed variance. The mean of such distribution is then the player's rating (skill). An application of the Elo rating in the domain of soccer was shown in Hvattum & Arntzen (2010). Recent work by Robberechts & Davis (2018) demonstrated that the method yields competitive results.

TrueSkill (Graepel *et al.*, 2007) is another system that enhances the Elo rating by operating not only with the variance of the player's performance, but also with the variance of his skill (rating). This variance reflects the uncertainty about the player's skill in situations when we have not yet observed enough data (performances). The authors demonstrated faster convergence and better predictive performance in comparison with the Elo rating. One of the caveats of the TrueSkill is that it does not propagate the newly obtained information backward to correct the past ratings. In other words, the method does filtering without smoothing. The work by Dangauthier *et al.* (2008) aimed to fix this issue. Also, the plain version of TrueSkill does not account for the score difference, as it only considers the ternary win-draw-loss outcome of a match. Guo *et al.* (2012) proposed an extension to take into account the score differences and claimed superior performance to the vanilla TrueSkill, also on a soccer dataset. The current evolution of the TrueSkill rating system is TrueSkill2 (Minka *et al.*, 2018); however, most of the improvements are domain specific to matchmaking in online games, which is the primary focus of the system.

A soccer domain-specific rating system called pi-ratings was introduced in Constantinou & Fenton (2013). The team's strength is represented by its home and away ratings, which are updated after each match according to manually set learning rates. Another score-based rating system was developed by Berrar *et al.* (2019), where the rating system parameters were tuned using particle swarm optimization and fed to a standard off-the-shelf learner. A method for ranking teams after an incomplete season was proposed by Csató (2021).

2.3 *Machine learning approaches*

Machine learning models are not very common in score-based modelling as they usually leverage extra features besides the scores or ratings. Some recent exceptions include models from the 2017 Soccer Prediction Challenge (Dubitzky *et al.*, 2019), where the dataset contained merely the historical results

with basic metadata on the matches. The winning learner of the competition was a model by Hubáček *et al.* (2019), ensemble of several models combined with expert features derived from the plain scores as an input into a variant of the gradient-boosted trees algorithm. For the challenge, Constantinou (2019) extended his pi-ratings model with a Bayesian network to obtain the probability distribution over possible match outcomes from the rating difference. Tsokos *et al.* (2019) tested several variations of the Bradley–Terry model (Baker & Scarf, 2020; Bradley & Terry, 1952) and a hierarchical Poisson model. In the end, the hierarchical Poisson model outperformed all the Bradley–Terry models. The inferiority of the Bradley–Terry model to other methods was further confirmed by Ley *et al.* (2019).

A different, unorthodox approach to the problem is to view the match data as a relational structure (graph) between the teams. This was first pointed out by Van Haaren & Van den Broeck (2015) where the authors achieved some promising results. An advanced relational learner (Natarajan *et al.*, 2012) was also tested by Hubáček *et al.* (2019), however, with a little success. The same authors also proposed relational team embeddings (Hubáček *et al.*, 2018), implemented in a framework for combining relational and neural learning (Sourek *et al.*, 2018), with somewhat promising results. The graph representation of the data was also utilized by Govan *et al.* (2008), who used the PageRank algorithm (Page *et al.*, 1999) to estimate the teams' strengths. The same author later proposed a so-called offense–defense model (Govan *et al.*, 2009), which can be seen as an analogy to the HITS algorithm (Kleinberg, 1999).

3. Models

The ambition of this work is to provide an experimental comparison of a variety of different approaches towards the problem of soccer match outcome prediction. For that goal, we have reimplemented and tested the most prevalent models in score-based soccer forecasting, as well as some models that, despite their promising results, have not received as much attention as the former. We recall that in this review, we do not consider any models utilizing specialized features (Goes *et al.*, 2021) other than the match scores (Section 1). Also, we focus solely on assessing the predictive performance of the models, as opposed to, e.g. their potential profitability on a betting market (Uhrín *et al.*, 2021). The selected models, together with the underlying reasoning, are then as follows.

The double Poisson model with exponential time weighting (Dixon & Coles, 1997; Maher, 1982) is probably the most established model to date. Recent work by Ley *et al.* (2019) showed that the model (and its bivariate variant) is still relevant nowadays. The most notable improvement upon these models was claimed by Boshnakov *et al.* (2017) using their bivariate Weibull model.

From the perspective of the rating systems, the Elo (Elo, 1978; Hvattum & Arntzen, 2010) proved to be competitive by Robberechts & Davis (2018). Constantinou & Fenton (2013) claimed to outperform the Elo considerably. Another rating system that claimed improvement over the Elo, was TrueSkill (Graepel *et al.*, 2007). Its extension for score-based forecasting by Guo *et al.* (2012) demonstrated better results on a soccer dataset, therefore we chose the extension over the original model. Steph ratings (Stephenson & Sonas, 2019) not only did well in a Kaggle competition, but they are also an extension of another successful rating system—the Glicko (Glickman, 1999). Recently, the ratings by Berrar *et al.* (2019) showed the most promising results in another competition (Dubitzky *et al.*, 2019).

As the machine learning models often utilize an ensemble of the former models (Hubáček *et al.*, 2019), including them in the comparison would seem unfair. The unorthodox relational learning approaches, despite showing promising results, then do not scale to the size of the dataset used in this review (Hubáček *et al.*, 2018).

3.1 *Statistical models*

In this section we take a closer look at the statistical models included in this review. As all the statistical models provide a likelihood function for the scores, the probability of a team winning/drawing/losing can be easily computed by marginalizing the probability distribution over the results.

3.1.1 *Double Poisson model.* A representation of one of the earliest (Maher, 1982) and simplest models. However, as was shown in Ley *et al.* (2019), it still remains very competitive. The model assumes the goals scored by the competing teams in a match to be independent. Therefore, the probability of a home team scoring x goals with the away team scoring y goals is given by

$$P(G_H = x, G_A = y|\lambda_H, \lambda_A) = \frac{\lambda_H^x e^{-\lambda_H}}{x!} \cdot \frac{\lambda_A^y e^{-\lambda_A}}{y!},$$

where λ_H and λ_A are the scoring rates of the teams (the means of the underlying Poisson distributions). The scoring rates of the teams for a particular match can be expressed in terms of the Maher's specification as

$$\begin{aligned}\log(\lambda_H) &= Att_H - Def_A + H\\ \log(\lambda_A) &= Att_A - Def_H,\end{aligned}$$

where H represents a home advantage, and Att and Def are respectively the offensive and defensive strengths of the teams (the actual model parameters).

Later, Ley *et al.* (2019) demonstrated that the number of the model's parameters can be effectively halved by considering only a single strength parameter for each team without any loss of predictive performance, i.e. reducing to

$$\begin{aligned}\log(\lambda_H) &= Str_H - Str_A + H\\ \log(\lambda_A) &= Str_A - Str_H.\end{aligned} \tag{1}$$

3.1.2 *Bivariate Poisson.* Karlis & Ntzoufras (2003) extended the double Poisson model by introducing dependence between the scored and conceded goals. The dependence is given by another Poisson distribution. The scored goals are modelled as $G_H = X_1 + X_3$, $G_A = X_2 + X_3$ where $X_1 \sim Pois(\lambda_H)$, $X_2 \sim Pois(\lambda_A)$ and $X_3 \sim Pois(\lambda_C)$. The probability function for the bivariate distribution is then given by

$$P(G_H = x, G_A = y|\lambda_H, \lambda_A, \lambda_C) = e^{-(\lambda_H+\lambda_A+\lambda_C)} \frac{\lambda_H^x \lambda_A^y}{x!\,y!} \sum_{k=0}^{min(x,y)} \binom{x}{k}\binom{y}{k} k! \left(\frac{\lambda_C}{\lambda_H\lambda_A}\right)^k, \tag{2}$$

where the scoring rates λ_H and λ_A are computed in the same fashion as for the double Poisson model, and $log\lambda_C$ is fitted together with the Str and H parameters.

3.1.3 *Double Weibull count model.* One of the pitfalls of the Poisson-based models is that the Poisson distribution does not consider under or over dispersion in the data since the variance of the distribution is strictly equal to the mean. Weibull-based models (Boshnakov *et al.*, 2017) aim to tackle this issue. The

Weibull count model was derived from the continuous Weibull distribution by McShane *et al.* (2008). The probability density function of the univariate Weibull count model is given by

$$P(G = x|\lambda, c) = \sum_{j=x}^{\infty} \frac{(-1)^{x+j}\lambda\alpha_j^x}{\Gamma(cj+1)}, \tag{3}$$

where c is the shape parameter of the distribution and α_j^x is defined recursively for $x = 0, 1, \ldots$ and $j = x+1, x+2, \ldots$ as

$$\alpha_j^0 = \frac{\Gamma(cj+1)}{\Gamma(c+1)}$$

$$\alpha_j^{x+1} = \sum_{m=x}^{j-1} \alpha_m^x \frac{\Gamma(cj-cm+1)}{\Gamma(c-j+1)}.$$

The probability of observing a score in a soccer match is then obtained by multiplying the two probability distributions for each of the opposing teams, analogically to Eq. (2). For this reason, the double Poisson and Weibull models are also referred to as 'independent'. Calculation of the scoring rates λ then also follows Eq. (1).

3.1.4 *Bivariate Weibull count model.* The bivariate version of the Weibull count model was introduced by Boshnakov *et al.* (2017). The Weibull marginals were tied together with Frank copula to form the bivariate model. The joint probability function is given by

$$\begin{aligned} P(G_H = x, G_A = y|\lambda_H, \lambda_A, c_H, c_A) = {} & C(F(x|\lambda_H, c_H), F(y|\lambda_A, c_A)) \\ & - C(F(x-1|\lambda_H, c_H), F(y|\lambda_A, c_A)) \\ & - C(F(x|\lambda_H, c_H), F(y-1|\lambda_A, c_A)) \\ & + C(F(x-1|\lambda_H, c_H), F(y-1|\lambda_A, c_A)), \end{aligned}$$

where F is a cumulative distribution function that can be computed using the probability density function from Eq. (3) and c_H, c_A are the shape parameters of the distribution. The C is Frank copula function given by

$$C(u, v) = -\frac{1}{\kappa}\ln\left(1 + \frac{(e^{-\kappa u}-1)(e^{-\kappa v}-1)}{e^{-\kappa}-1}\right),$$

where κ is the dependence parameter. Calculation of the scoring rates λ again follows Eq. (1).

3.2 *Rating systems*

One of the main differences between statistical models and rating systems is that rating systems were designed mainly to rank the competing teams in a league and not necessarily to produce a probability

distribution over the possible outcomes. However, this can be effectively solved by employing a subsequent regression model that transforms the ratings into the desired probability distribution, as was demonstrated in Hvattum & Arntzen (2010). The details on how the ratings and the subsequent regression are trained can be found in Section 4.1.

3.2.1 *Elo ratings.* This model by Elo (1978) is a general rating system, the modification of which is still used for evaluation of the strength of chess players. Hvattum & Arntzen (2010) proposed its modification for soccer and, consequently, Robberechts & Davis (2018) demonstrated the effectiveness of the method. The modification involves the use of an ordered logit model (McCullagh, 1980) to obtain the probability distribution over the possible match outcomes. At the core, each team's performance is assumed to be normally distributed around its true strength. The expected outcomes for both teams are then calculated as follows:

$$E_H = \frac{1}{1 + c^{(R_A - R_H)/d}}$$
$$E_A = 1 - E_H,$$

where R_H and R_A are the ratings of the home and away teams, and c and d are metaparameters of the method. The actual ternary outcome of the match is then encoded numerically as

$$S_H = \begin{cases} 1 & \text{if the home team won} \\ 0.5 & \text{if the match was drawn} \\ 0 & \text{if the home team lost} \end{cases}$$

Finally, after the match, the ratings of the teams are updated using

$$R_H^{t+1} = R_H^t + k(1+\delta)^\gamma \cdot (S_H - E_H)$$
$$R_A^{t+1} = R_A^t - k(1+\delta)^\gamma \cdot (S_H - E_H),$$

where δ is an absolute value of goal difference, k represent a learning rate and γ is a metaparameter scaling the influence of the goal difference on the rating change.

3.2.2 *Steph ratings.* This model by Stephenson & Sonas (2019) is an evolution of another rating system known as Glicko (Glickman, 1999) developed for a chess rating competition at Kaggle. However, the ratings can be easily adapted to other sports. The strength of a team is represented with a tuple (r, v) to capture the team's rating and its variance. Unlike in Elo, the variance of the rating is not constant. Before each match, the variance is increased based on the time passed (Δt) since the last match of the team and a scaling factor c.

$$v^t += c\Delta t$$

The expected outcome (e) is then computed, accounting for the rating difference (Δr) between the competing teams (i and j), and a home advantage (γ).

$$w = \begin{cases} -1 & \text{for home team} \\ 1 & \text{for away team} \end{cases}$$

$$q = \frac{\ln 10}{400}$$

$$\Delta r = w \cdot (r_i^t - r_j^t + \gamma)$$

$$k = \frac{1}{1 + 3q^2 v_j^t / \pi^2}$$

$$e = \frac{1}{10^{-k\Delta r/400}}$$

Finally, the rating and its variance is updated as follows:

$$s = \begin{cases} 1 & \text{if team won} \\ 0.5 & \text{in case of draw} \\ 0 & \text{if team lost} \end{cases}$$

$$d = q^2 k^2 e(1-e)$$

$$v_i^{t+1} = \left(\frac{1}{v_i^t + h} + d \right)^{-1}$$

$$r_i^{t+1} = r_i^t + q v^{t+1} k(s - e + b) + \lambda(r_j^t - r_i^t),$$

where λ is scaling factor for rating difference and h controls the increase in rating's variance over time. The b serves as a bonus to players/teams that play more often. When $h = b = \gamma = 0$, the computations reproduce the Glicko ratings. The learning rate k depends on the rating's variance, allowing for faster changes when the rating is not yet well supported by evidence.

3.2.3 *Pi-ratings.* This model by Constantinou & Fenton (2013) represents a domain-specific rating system. Each team is assigned two ratings, representing its strength when playing home (R^α) and when playing away (R^β). For each match, the expected goal difference is calculated based on home team's home rating (R_H^α) and away team's away rating (R_A^β).

$$\widehat{G}_H = 10^{\frac{|R_H^\alpha|}{c}} - 1$$

$$\widehat{G}_A = 10^{\frac{|R_A^\beta|}{c}} - 1$$

$$R_H^\alpha < 0 \implies \widehat{G}_H := -\widehat{G}_H$$

$$R_A^\beta < 0 \implies \widehat{G}_A := -\widehat{G}_A$$

$$\widehat{\Delta G} = \widehat{G}_H - \widehat{G}_A,$$

where c is a metaparameter of the ratings. After a match is played, the expected score is compared to the actual outcome and both R^{α} and R^{β} get updated, with each of the updates having a separate learning rate.

$$
\begin{aligned}
e &= \Delta\widehat{G} - \Delta G \\
\psi(e) &= c\log_{10}(1+|e|) \\
R_H^{\alpha} &+= \lambda\psi(e)\cdot sign(e) \\
R_H^{\beta} &+= \gamma\psi(e)\cdot sign(e) \\
R_A^{\beta} &+= \lambda\psi(e)\cdot sign(-e) \\
R_A^{\alpha} &+= \gamma\psi(e)\cdot sign(-e),
\end{aligned}
$$

where λ and γ are the ratings' learning rates.

3.2.4 *Gaussian-OD ratings.* Gaussian-OD ratings are an extension of the TrueSkill rating system (Graepel *et al.*, 2007). The TrueSkill system was originally designed for ranking players in a computer game called 'Halo'. The motivation was to match equally skilled players against each other to maximize the overall enjoyment of the game. This further illustrates the usefulness of models presented in this paper beyond the sole purpose of predicting future outcomes. In TrueSkill, each team is assigned a Gaussian distribution representing the user's prior about the team's skill. Unlike in Elo, the variance of each team rating is a parameter that changes value over time. In Guo *et al.* (2012), the authors promoted a version of the TrueSkill, where each team is assigned a separate Gaussian distribution for its offensive ($p(o) := \mathcal{N}(\mu_o, \sigma_o^2)$) and defensive ($p(d) := \mathcal{N}(\mu_d, \sigma_d^2)$ skill. TrueSkill generally assumes that even if we knew the exact value of the team's skill (the variance of the Gaussian would be equal to 0), its performance would still be stochastic, as the teams do not perform the same each day. The defensive ($p_d := \mathcal{N}(d, \beta^2)$) and offensive ($p_o := \mathcal{N}(o, \beta^2)$) performances are thus affected by the performance variance β^2. The home goals scored generation process is then assumed to be $G_H \sim \mathcal{N}(p_{o_H} - p_{d_A}, \gamma^2)$. β^2 and γ are metaparameters for performance and score variance. Finally, the prior distributions are updated after each match according to the following equations:

$$
\begin{aligned}
\pi_{o_H} &= \frac{1}{\sigma_{o_H}^2} + \frac{1}{2\beta^2+\gamma^2+\sigma_{d_A}^2} \\
\pi_{d_A} &= \frac{1}{\sigma_{d_A}^2} + \frac{1}{2\beta^2+\gamma^2+\sigma_{o_H}^2} \\
\tau_{o_H} &= \frac{\mu_{o_H}}{\sigma_{o_H}^2} + \frac{G_H+\mu_{d_A}}{2\beta^2+\gamma^2+\sigma_{d_A}^2} \\
\tau_{d_A} &= \frac{\mu_{d_A}}{\sigma_{d_A}^2} + \frac{\mu_{o_H}-G_H}{2\beta^2+\gamma^2+\sigma_{o_H}^2} \\
\sigma_{o_H}^2 &:= \pi_{o_H}^{-1} \\
\mu_{o_H} &:= \sigma_{o_H}^2\tau_{o_H}
\end{aligned}
$$

The equations for updating the remaining skills are analogous.

3.2.5 *Berrar Ratings.* Berrar ratings were introduced in Berrar *et al.* (2019) as input features to a more complex model. The idea behind these ratings it to use a logistic function to predict the number of goals scored using once again offensive and defensive strengths of the teams. The formulas for estimating the expected goals scored are as follows:

$$\widehat{G}_H = \frac{\alpha}{1 + \exp(-\beta_H(o_H - d_A) - \gamma_H)}$$
$$\widehat{G}_A = \frac{\alpha}{1 + \exp(-\beta_A(o_A - d_H) - \gamma_A)},$$

where o and d are the offensive and defensive ratings of the competing teams and α stands for the maximum possible number of goals scored predicted. The authors set the $\alpha = 5$ as more than five goals are scored very rarely. β then determines the steepness of the logistic function, while γ defines the threshold (also known as bias). The updates to the ratings are then done in the following fashion:

$$o_H \mathrel{+}= \omega_{o_H}(G_H - \widehat{G}_H)$$
$$d_H \mathrel{+}= \omega_{d_H}(G_A - \widehat{G}_A),$$

where ω stands for the learning rate for the particular rating. Updates of the away team are done analogously.

4. Validation Framework

All the data used in this review came from the Open International Soccer Database v2 (Dubitzky *et al.*, 2019). We divided the data into two sets. Matches before 07/2010 formed a validation set, used for hyperparameter tuning, and matches after 07/2010 formed a test set, used solely for evaluation. The first five rounds of each season were used as a burn-in period, omitted from the evaluation. This left us with 91,155 matches in the test set. The validation set was used to validate our implementations, training the parameters of subsequent regression models for rating systems, tuning hyperparameters of the models and trying out several optimization algorithms. All the presented results are computed on the test set.

The goal of the evaluation was to answer the following research questions:

1. How do the models compare in terms of predictive performance?
2. Do mathematically similar models produce similar predictions?

4.1 *Model fitting*

The models' parameters and outputs are summarized in Table 1.

Statistical Models All the statistical models are fitted by maximizing their respective weighted likelihood functions on the set of historical matches M:

$$L = \prod_{i=1}^{|M|} P(G_H^i = x, G_A^i = y|\Theta) \cdot w_i,$$

where w_i represents the weight of each observation, G_H^i and G_A^i are the goals scored by home and away team in match i and P is the probability of the respective match result as parametrized by Θ. The parameters belonging to Θ are summarized in Table 1. During the evaluation on the test set the

TABLE 1 *Models' parameters and outputs overview. The τ marks parameters belonging to each team*

	Metaparameters	Parameters	Outputs
Double Poisson	α	Str_τ, H	P(HDA)
Bivariate Poisson	α	Str_τ, H, λ_c	P(HDA)
Double Weibull	α	Str_τ, H, c_H, c_A	P(HDA)
Bivariate Weibull	α	$Str_\tau, H, c_H, c_A, \kappa$	P(HDA)
Elo	k, γ, H		R_τ, E
Steph	c, h, b, γ, λ		r_τ, v_τ, e
pi-ratings	λ, γ, c		$R^\alpha_\tau, R^\beta_\tau, \widehat{\Delta G}$
Gaussian-OD	β, γ, σ_0		$\mu_{o_\tau}, \mu_{d_\tau}, \sigma_{o_\tau}, \sigma_{d_\tau}$
Berrar	$\beta, \gamma, \omega^\alpha, \omega^\beta$		$o_\tau, d_\tau, \widehat{G}_\tau$

parameters (Table 1) are refitted after each league's round to account for the newly obtained information. To reduce the computational time, we limit the set of historical matches M by removing matches older than 5 years in each iteration. The matches from the past 5 years can be viewed as the training set for the iteration.

Since the first successful application (Dixon & Coles, 1997), exponential time weighting is being commonly used as

$$w_i = e^{-\alpha t_\Delta},$$

where t_Δ is the number of days passed since the match was played and α is a metaparameter. We found $\alpha = 0.002$ to perform best on our validation set. The same value was found in Boshnakov *et al.* (2017) and in Ley *et al.* (2019) value of $\alpha = 0.0019$ was used.

Ratings As the outputs of the ratings are not directly the probability distribution over the match outcomes ($P(HDA)$), a subsequent model has to be applied (Hvattum & Arntzen, 2010). We use multinomial logistic regression for this purpose. The parameters of the regression are optimized inside the meta-optimization routine for finding the rating's metaparameters. First, the ratings' computations (given the current set of metaparameters) are carried out through the data. The pre-match ratings then serve as input to the regression model. The regression model then produces in-sample predictions based on the given features. The in-sample loss is then reported as the loss belonging to the current set of metaparameters. The meta-optimizer then selects another set of metaparameters to be evaluated, and the process is repeated. The inner routine is summarized in the following pseudo-code:

```
def optimize_rating(data, res, metaparams, loss_func)
    ratings = compute_ratings(data, res, metaparams)
    lr = LogisticRegression()
    lr = lr.fit(ratings, res)
    predictions = lr.predict_proba(ratings)
    loss = loss_func(predictions, res)
    return loss.mean()
```

TABLE 2 *Comparison of the tested models via the evaluation metrics*

	xEnt	RPS	ACC
Berrar	1.0246	0.2101	48.54
Bivariate Poisson	1.0251	0.2103	48.58
Double Poisson	1.0254	0.2103	48.57
Double Weibull	1.0255	0.2103	48.60
pi-ratings	1.0258	0.2103	48.56
Bivariate Weibull	1.0260	0.2105	48.60
Elo	1.0263	0.2105	48.49
Steph	1.0291	0.2114	48.26
Gaussian-OD	1.0347	0.2134	47.84

TABLE 3 *Average Jensen–Shannon divergence between the models' predictions (BP = Bivariate Poisson, BW = Bivariate Weibull, DP = Double Poisson, DW = Double Weibull)*

	Berrar	BP	BW	Elo	DP	DW	pi-ratings	Steph	Gauss.-OD
Berrar	0.000	0.051	0.054	0.027	0.051	0.051	0.030	0.040	0.063
BP	0.051	0.000	0.011	0.052	0.014	0.014	0.052	0.058	0.071
BW	0.054	0.011	0.000	0.054	0.023	0.017	0.054	0.061	0.075
Elo	0.027	0.052	0.054	0.000	0.051	0.051	0.024	0.032	0.062
DP	0.051	0.014	0.023	0.051	0.000	0.009	0.051	0.056	0.069
DW	0.051	0.014	0.017	0.051	0.009	0.000	0.050	0.057	0.070
pi-ratings	0.030	0.052	0.054	0.024	0.051	0.050	0.000	0.036	0.064
Steph	0.040	0.058	0.061	0.032	0.056	0.057	0.036	0.000	0.058
Gauss.-OD	0.063	0.071	0.075	0.062	0.069	0.070	0.064	0.058	0.000

We observed, that multiple runs of meta-optimization result in different metaparameters while achieving the same loss. We therefore do not state any concrete values of the metaparameters found. While the parameters of the regressors and the metaparameters (Table 1) of the rating systems are determined on the validation set, the ratings (denoted as 'Outputs' in Table 1) are updated after each match in the test set to serve as input features for the regressor in the next league round.

The optimization of the regression parameters has been done by the L-BFGS-B algorithm (Byrd *et al.*, 1995). Optimization of the ratings' hyperparameters has been carried out by the PSO (Kennedy & Eberhart, 1995), as was done in Berrar *et al.* (2019). We experimented with other meta-optimization techniques but they were inferior to the PSO in terms of predictive performace and (mainly) computational time.

During the validation, we noticed that the meta-optimization of the Berrar ratings fails to converge occasionally even when given more iterations. We have resolved the issue by halving the number of metaparameters. In the original model the metaparameters for updating the ratings are separate for home and away team. As stated in Table 1, we use the same metaparameters for the home and away team ratings updates.

4.2 *Evaluation measures*

Ranked Probability Score The ranked probability score was proposed by Epstein (1969) for evaluating ordinal outcomes. For the ternary outcome game of soccer, the formula is as follows:

$$\mathrm{RPS}(p_1, p_2, p_3, y_1, y_2, y_3) = \frac{1}{2} \sum_{i=1}^{2} \left(\sum_{j=1}^{i} \left(p_j - y_j \right) \right)^2,$$

where p_j is the estimated probability of outcome j, and $y_j \in \{0, 1\}$, with $y_j = 1$ indicating that outcome j was realized. The suitability of using this metric for evaluating soccer outcome predictions was heavily proclaimed in Constantinou & Fenton (2012) and has been widely used ever since. However, Hubáček *et al.* (2019) pointed out that while the outcomes of a soccer match can be intuitively seen as ordinal, the underlying probability distribution does not have to be as such, due to the low prior probability of a draw. Simply put, a win of a home team does not imply that the draw was more likely than the visiting team winning. This effect is even more obvious in high scoring games, such as basketball. Nevertheless, we still report the measured RPS, mainly for compatibility with the previous research.

Crossentropy The crossetropy loss is one of the most established measures for classification tasks. It does not raise any assumptions about the ordinality of the outcomes and is simply calculated as

$$\mathrm{xEnt}(p_1, p_2, p_3, y_1, y_2, y_3) = -\sum_{i=1}^{3} y_i \log(p_i). \tag{4}$$

Also, since the statistical models are fitted by maximizing their respective likelihood functions, the crossentropy is a natural evaluation metric to be used.

Accuracy Accuracy serves as the most crude evaluation measure. It simply represents how many times on average, the outcome with the highest estimated probability was realized.

Similarity measures Besides the predictive performance of the models, we are also interested in analysing how much the predictions of the models differ from each other since there are many similarities both among the statistical models and among the ratings. For this purpose, we compute the average Jensen–Shannon divergence between the models' predictions. The Jensen–Shannon divergence between two probability distributions P and Q is given by

$$JSD(P \,||\, Q) = \frac{1}{2} D_{KL}(P \,||\, M) + \frac{1}{2} D_{KL}(Q \,||\, M)$$

$$M = \frac{1}{2}(P + Q)$$

$$D_{KL}(P \,||\, Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)}$$

5. Results

The results are summarized in Table 2. The first thing to notice is that the models' performances are very close to each other. The Berrar ratings achieved best results by both the RPS and ENT measures, while the Double Weibull and Bivariate Weibull models reached the highest accuracy score. The Double

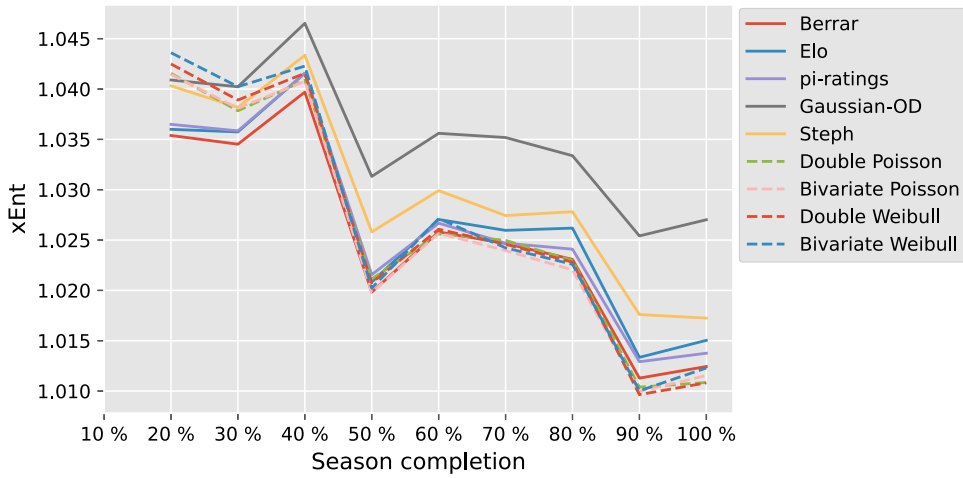

FIG. 1. Error of the models, as measured using the crossentropy (Eq. (4)), as a function of percentage of season completed.

Weibull model placing ahead of its Bivariate variant might look suspicious at first. However, during the validation, we noticed that while the Bivariate Weibull sometimes provided the best fit for the training data, the performance did not always translate into the test set. This suggests that finding the right dependence parameter κ is difficult. It is remarkable how well the general rating system Elo with only minor modifications works for soccer. This is not the case for the other two general rating systems—the Steph ratings and the Gaussian-OD ratings. Another result that catches the eye is the performance of the Double Poisson model. This only confirms its competitiveness, as suggested by Ley *et al.* (2019). The only model that significantly falls behind are the Gaussian-OD ratings.

5.1 *Predictions' similarity*

As the evaluation metrics of plurality of the models are very close to each other, it is also not surprising that the actual predictions exhibit many similarities, too, as shown in Table 3. Even without knowing that there are two classes of models, we would be able to distinguish the statistical models from the ratings based purely on their prediction similarities. We can observe that especially the predictions of the statistical models are very close to each other. This behavior was anticipated, as the very definitions of the models are very similar (with a certain parameter setup, they all reduce to the Double Poisson model). The similarity of the Gaussian-OD to other models is lower mostly due to its inferior performance.

5.2 *Model adaptability*

Another property of the models we were interested in is how quickly they adapt to new information. Between the seasons, the team's composition, and therefore also its strength, can change dramatically. We thus divided the matches into 10 groups based on which part of the season they occurred (Fig. 1).

The plot shows that the models' performance is generally higher in the second half of the season when more data are available. The statistical models trail behind the rating systems in the first third of the season, while providing a generally better fit in the second half of season. Berrar rating seems to outperform the competition in most parts of the season. The slight decrease in the models' performances, right after half of the season was played, might be due to breaks in the schedule that typically occur in

the middle of a season. Another explanation could be that in some leagues, there are transfer windows opened during this period of time, which could lead to changes in teams' compositions.

6. Conclusion

We reimplemented a wide selection of the top-performing score-based models for soccer outcome forecasting from the past decades and benchmarked them on the largest soccer dataset published to date. We asked two core research questions regarding the models' performances and similarities. We conclude from the experiments that the individual predictions, as well as the overall performances, were very similar across the top models tested, likely suggesting for the limits of this generic approach to score-based match outcome modelling. Additionally, we observed that the rating systems adapt faster to changes in teams' strengths and achieve better performance in the beginnings of the seasons, while the statistical models catch up and take a small lead at the ends.

Our results suggest that any dramatic improvement in the predictive performance of any rating or statistical method seems unlikely now. We therefore propose that further research should attempt to address the problem with a significantly different 'class' of models. These could possibly produce more diverse predictions, opening new possibilities for ensembling and other machine learning techniques.

Acknowledgements

We thank the anonymous reviewers for their constructive comments.

Funding

Czech Science Foundation project 20-29260S; computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the programme 'Projects of Large Research, Development, and Innovations Infrastructures'.

References

Agosto, A., Cavaliere, G., Kristensen, D. & Rahbek, A. (2016) Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *J. Empir. Finance*, **38**, 640–663.

Angelini, G. & De Angelis, L. (2017) PARX model for football match predictions. *J. Forecast.*, **36**, 795–807.

Baio, G. & Blangiardo, M. (2010) Bayesian hierarchical model for the prediction of football results. *J. Appl. Stat.*, **37**, 253–264.

Baker, R. & Scarf, P. (2020) Modifying Bradley–Terry and other ranking models to allow ties. *IMA J. Manag. Math*. https://academic.oup.com/imaman/advance-article-abstract/doi/10.1093/imaman/dpaa027/6029099?redirectedFrom=fulltext.

Berrar, D., Lopes, P. & Dubitzky, W. (2019) Incorporating domain knowledge in machine learning for soccer outcome prediction. *Mach. Learn.*, **108**, 97–126.

Boshnakov, G., Kharrat, T. & McHale, I. G. (2017) A bivariate Weibull count model for forecasting association football scores. *Int. J. Forecast.*, **33**, 458–466.

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

Byrd, R. H., Lu, P., Nocedal, J. & Zhu, C. (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

Constantinou, A. C. (2019) Dolores: a model that predicts football match outcomes from all over the world. *Mach. Learn.*, **108**, 49–75.

Constantinou, A. C. & Fenton, N. E. (2012) Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *J Quant. Anal. Sports*, **8**. https://www.degruyter.com/journal/key/JQAS/8/1/html.

Constantinou, A. C. & Fenton, N. E. (2013) Determining the level of ability of football teams by dynamic ratings based on the relative discrepancies in scores between adversaries. *J. Quant. Anal. Sports*, **9**, 37–50.

Crowder, M., Dixon, M., Ledford, A. & Robinson, M. (2002) Dynamic modelling and prediction of English football league matches for betting. *J. R. Stat. Soc. Ser. D*, **51**, 157–168.

Csató, L. (2021) Coronavirus and sports leagues: obtaining a fair ranking when the season cannot resume. *IMA J. Manag. Math*. https://academic.oup.com/imaman/advance-article/doi/10.1093/imaman/dpab020/6297161.

Dangauthier, P., Herbrich, R., Minka, T. & Graepel, T. (2008) Trueskill through time: revisiting the history of chess. *Advances in Neural Information Processing Systems*, Curran Associates, Inc., New York. pp. 337–344.

Dixon, M. J. & Coles, S. G. (1997) Modelling association football scores and inefficiencies in the football betting market. *J. R. Stat. Soc. Ser. C*, **46**, 265–280.

Dubitzky, W., Lopes, P., Davis, J. & Berrar, D. (2019) The open international soccer database for machine learning. *Mach. Learn.*, **108**, 9–28.

Elo, A. E. (1978) *The Rating of Chessplayers: Past and Present*. Arco Pub, New York.

Epstein, E. S. (1969) A scoring system for probability forecasts of ranked categories. *J. Appl. Meteorol.*, **8**, 985–987.

Glickman, M. E. (1999) Parameter estimation in large dynamic paired comparison experiments. *J. R. Stat. Soc. Ser. C*, **48**, 377–394.

Goes, F. R., Kempe, M., van Norel, J. & Lemmink, K. A. P. M. (2021) Modelling team performance in soccer using tactical features derived from position tracking data. *IMA J. Manag. Math*. https://academic.oup.com/imaman/advance-article-abstract/doi/10.1093/imaman/dpab006/6210047?redirectedFrom=fulltext.

Govan, A. Y., Langville, A. N. & Meyer, C. D. (2009) Offense-defense approach to ranking team sports. *J. Quant. Anal. Sports*, Degruyter, Berlin. **5**(1). https://www.degruyter.com/document/doi/10.2202/1559-0410.1151/html.

Govan, A. Y., Meyer, C. D. & Albright, R. (2008) Generalizing Google's PageRank to rank national football league teams. *Proceedings of the SAS Global Forum*, vol. **2008**. https://support.sas.com/resources/papers/proceedings/pdfs/sgf2008/TOC.html.

Graepel, T., Minka, T. & Herbrich, R. T. (2007) A Bayesian skill rating system. *Adv. Neural Inf. Process. Syst.*, **19**, 569–576.

Guo, S., Sanner, S., Graepel, T. & Buntine, W. (2012) Score-based Bayesian skill learning. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, pp. 106–121.

Hubáček, O., Šourek, G. & Železný, F. (2019) Learning to predict soccer results from relational data with gradient boosted trees. *Mach. Learn.*, **108**, 29–47.

Hubáček, O., Šourek, G. & Železný, F. (2018) Lifted relational team embeddings for predictive sport analytics. *Proceedings of the 28th International Conference on Inductive Logic Programming'*. CEUR-WS.org, pp. 84–91.

Hvattum, L. M. & Arntzen, H. (2010) Using ELO ratings for match result prediction in association football. *Int. J. Forecast.*, **26**, 460–470.

Karlis, D. & Ntzoufras, I. (2003) Analysis of sports data by using bivariate Poisson models. *J. R. Stat. Soc. Ser. D*, **52**, 381–393.

Karlis, D. & Ntzoufras, I. (2008) Bayesian modelling of football outcomes: using the Skellam's distribution for the goal difference. *IMA J. Manag. Math.*, **20**, 133–145.

Kennedy, J. & Eberhart, R. (1995) Particle swarm optimization (PSO). *Proc. IEEE International Conference on Neural Networks*. IEEE, New Jersey, Perth, Australia, pp. 1942–1948.

Kleinberg, J. M. (1999) Authoritative sources in a hyperlinked environment. *J. ACM*, **46**, 604–632.

Koopman, S. J. & Lit, R. (2015) A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League. *J. R. Stat. Soc. Ser. A*, **178**, 167–186.

Koopman, S. J. & Lit, R. (2019) Forecasting football match results in national league competitions using score-driven time series models. *Int. J. Forecast.*, **35**, 797–809.

Ley, C., Wiele, T. V. d. & Eetvelde, H. V. (2019) Ranking soccer teams on the basis of their current strength: a comparison of maximum likelihood approaches. *Stat. Model.*, **19**, 55–77.

Maher, M. J. (1982) Modelling association football scores. *Statistica Neerlandica*, **36**, 109–118.

McCullagh, P. (1980) Regression models for ordinal data. *J. R. Stat. Soc. Ser. B*, **42**, 109–127.

McHale, I. & Scarf, P. (2011) Modelling the dependence of goals scored by opposing teams in international soccer matches. *Stat. Model.*, **11**, 219–236.

McShane, B., Adrian, M., Bradlow, E. T. & Fader, P. S. (2008) Count models based on Weibull interarrival times. *J. Bus. Econ. Stat.*, **26**, 369–378.

Minka, T., Cleven, R. & Zaykov, Y. (2018) TrueSkill 2: an improved Bayesian skill rating system. *Technical Report*.

Natarajan, S., Khot, T., Kersting, K., Gutmann, B. & Shavlik, J. (2012) Gradient-based boosting for statistical relational learning: the relational dependency network case. *Mach. Learn.*, **86**, 25–56.

Owen, A. (2011) Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter. *IMA J. Manag. Math.*, **22**, 99–113.

Page, L., Brin, S., Motwani, R. & Winograd, T. (1999) The PageRank citation ranking: bringing order to the web. *Technical report.* Stanford InfoLab.

Robberechts, P. & Davis, J. (2018) Forecasting the FIFA World Cup–Combining result-and goal-based team ability parameters. *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2018 workshop*, vol. **2284**. Springer, pp. 52–66.

Rue, H. & Salvesen, O. (2000) Prediction and retrospective analysis of soccer matches in a league. *J. R. Stat. Soc. Ser. D*, **49**, 399–418.

Skellam, J. G. (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. *J. R. Stat. Soc. Ser. A*, **109**, 296–296.

Sourek, G., Aschenbrenner, V., Zelezny, F., Schockaert, S. & Kuzelka, O. (2018) Lifted relational neural networks: Efficient learning of latent relational structures. *J. Artif. Intell. Res.*, **62**, 69–100.

Stephenson, A. & Sonas, J. (2019) PlayerRatings: dynamic updating methods for player ratings estimation. R package version 1.0-3.

Tsokos, A., Narayanan, S., Kosmidis, I., Baio, G., Cucuringu, M., Whitaker, G. & Király, F. (2019) Modeling outcomes of soccer matches. *Mach. Learn.*, **108**, 77–95.

Uhrín, M., Šourek, G., Hubáček, O. & Železný, F. (2021) Optimal sports betting strategies in practice: an experimental review. *IMA J. Manag. Math*. https://academic.oup.com/imaman/advance-article-abstract/doi/10.1093/imaman/dpaa029/6128334.

Van Haaren, J. & Van den Broeck, G. (2015) Relational learning for football-related predictions. *Latest Advances in Inductive Logic Programming*. World Scientific, Singapore, pp. 237–244.

A. Appendix - Calibration Curves

To analyze the models' performance in a greater detail we plot the calibration curves[1] in Fig. A2. A calibration curve shows how well the predicted probabilities correspond to the observed relative frequencies of an outcome conditioned on the predicted probability value. Firstly, it can be noticed (Fig. A2) that the calibration curves of the statistical models follow the desired identity ($x = y$) line more closely. The rating systems seem to be more conservative in the estimates and therefore less accurate in cases where there is a clear favourite to win. This is most apparent from the underestimated probabilities of the respective away teams' wins.

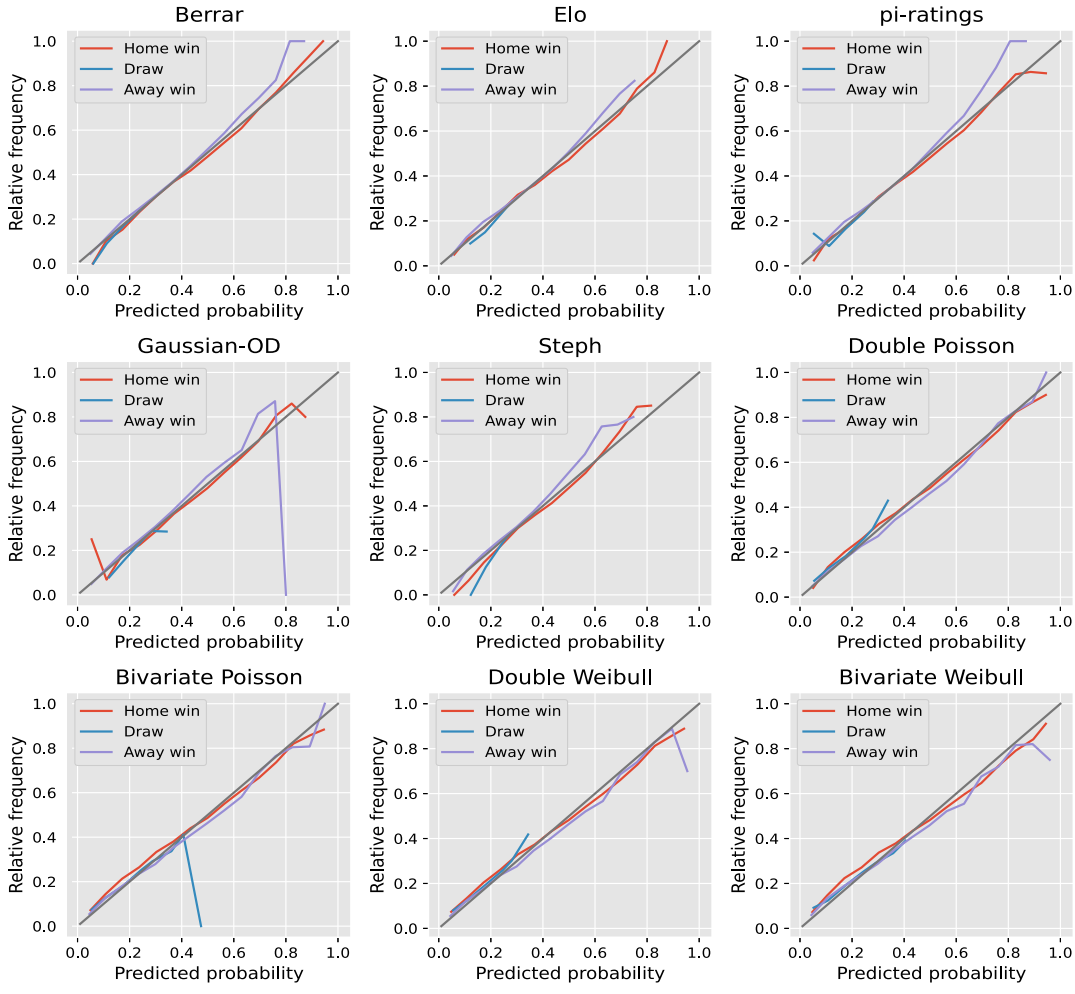

FIG. A2. Calibration curves of the reviewed models, visualizing the empirical relative frequencies of the match outcomes as a function of the probability values predicted by the respective models.

[1] Suggested by an anonymous reviewer.