doi.org/10.26398/IJAS.0030-010

A COMPARISON OF RATING SYSTEMS FOR COMPETITIVE WOMEN'S BEACH VOLLEYBALL

Mark E. Glickman[1]

Department of Statistics, Harvard University, Cambridge, MA, USA

Jonathan Hennessy

Google, Mountain View, CA, USA

Alister Bent

Trillium Trading, New York, NY, USA

Abstract *Women's beach volleyball became an official Olympic sport in 1996 and continues to attract the participation of amateur and professional female athletes. The most well-known ranking system for women's beach volleyball is a non-probabilistic method used by the Fédération Internationale de Volleyball (FIVB) in which points are accumulated based on results in designated competitions. This system produces rankings which, in part, determine qualification to elite events including the Olympics. We investigated the application of several alternative probabilistic rating systems for head-to-head games as an approach to ranking women's beach volleyball teams. These include the Elo (1978) system, the Glicko (Glickman, 1999) and Glicko-2 (Glickman, 2001) systems, and the Stephenson (Stephenson and Sonas, 2016) system, all of which have close connections to the Bradley-Terry (Bradley and Terry, 1952) model for paired comparisons. Based on the full set of FIVB volleyball competition results over the years 2007-2014, we optimized the parameters for these rating systems based on a predictive validation approach. The probabilistic rating systems produce 2014 end-of-year rankings that lack consistency with the FIVB 2014 rankings. Based on the 2014 rankings for both probabilistic and FIVB systems, we found that match results in 2015 were less predictable using the FIVB system compared to any of the probabilistic systems. These results suggest that the use of probabilistic rating systems may provide greater assurance of generating rankings with better validity.*

Keywords: *Bradley-Terry, paired comparison, ranking, sports rating system, volleyball.*

1. INTRODUCTION

Beach volleyball, a sport that originated in the early 1900s, has been played by athletes on a professional basis for over 50 years. The rules of competitive beach volleyball are largely the same as indoor volleyball with several notable differences. Beach volleyball is played on a sand court with teams consisting of two players as

[1] Corresponding author: Mark E. Glickman, email: *glickman@fas.harvard.edu*

opposed to six in indoor volleyball. Matches are played as a best of 3 sets, in which each of the first two sets is played to 21 points, and the deciding set (if the first two sets split) is played to 15 points. The popularity of beach volleyball has led to regular organized international competition, with the sport making first appearance in the Olympic games in 1996.

The main international organization governing volleyball competition is the Fédération Internationale de Volleyball (FIVB). The FIVB originated in the 1940s, and is involved in planning elite international volleyball tournaments including the Olympic Games, the Men's and Women's World Championships, the World Tour, various elite youth events. In addition to being the main organizers of many professional beach volleyball tournaments organized worldwide, the FIVB coordinates events with national volleyball organizations and with other international athletic organizations such as the International Olympic Committee. The FIVB is also responsible for the standardization of the rules of volleyball for international competition.

One of the most important functions of the FIVB is the determination of how teams qualify for international events, which is largely based on the FIVB's ranking system. FIVB rankings determine how teams are seeded on the World Tour, thereby affecting their performance and tournament earnings, as well as determining which teams compete in the Olympic Games. Currently, the FIVB relies on an accumulation point system to rank its players. The system awards points based on teams' finishing place at FIVB tournaments, with the most points being awarded to the highest-placing teams. Furthermore, greater point totals are at stake at larger tournaments, such as World Championships or Grand Slam tournaments.

The current FIVB ranking system has several desirable qualities, including its simplicity and ease-of-implementation. Because the ranking system involves fairly basic computation, the system is transparent. The system also behaves predictably, so that teams with better finishes in tournaments typically move up in the FIVB rankings. The convenience of ranking teams according to such a system, however, is not without its shortcomings. For example, because the FIVB system awards points based solely on the final standings in a tournament, information from earlier match results in a tournament does not play a role in computing rankings. Many tournaments include only four to five rounds of bracket play, with most teams only making it through one or two matches in this stage. Only the teams who advance further receive FIVB points. Pool play, meanwhile, often represents the majority of the matches played by a team in a tournament, even for those who make it into the championship bracket (many teams play only 1-2 bracket matches after 4-5 pool play matches). The results of matches in pool play are not evaluated as part of the FIVB ranking calculation. Thus the FIVB system misses out on key information available in individual match data from the entire tournament.

In contrast to the FIVB ranking system, rating systems have been developed to measure the probability of one team defeating another with the goal of accurately predicting future match outcomes. Many of these approaches have arisen from applications to games like chess, whose Elo system (Elo, 1978) and variants thereof have been used in leagues for other games and sports such as Go, Scrabble, and table tennis. The main difference between such probabilistic systems and the point accumulation system of the FIVB is that all match results are incorporated in producing team ratings, with each head-to-head match result factoring into the computation. Furthermore, the probabilistic systems smoothly downweight the impact of less recent competition results relative to more current ones. In the FIVB system, tournaments older than one year do not play a role in the current rankings, whereas in most probabilistic systems older match results are part of the computation though they receive small weight. Reviews of different sports rating systems, both of point accumulation systems and probabilistic ones, can be found in Stefani (1997) and Stefani (2011).

In this paper, we compare the FIVB system to four probabilistic systems that have been in use in other sports/games contexts. We examine the comparison of these different rating systems applied to match data collected on women's beach volleyball. We describe in detail in Section 2 the FIVB system along with the four probabilistic rating systems. This is followed in Section 3 by a description of the women's beach volleyball data and the implementation of the probabilistic rating systems. In Section 4 we describe the results of our analyses. The paper concludes in Section 5 with a discussion about the results, and the appropriateness of using a probabilistic rating system for FIVB competition.

2. RATING VOLLEYBALL TEAMS

We describe in this section the point system used by the FIVB to rank players, and then review the four probabilistic rating systems considered in this paper.

2.1 FIVB TOURNAMENTS

Typical FIVB events are organized as a combination of a phase of Round Robin competition (pool play) followed by single elimination. For example, the Main Draw Tournament (separately by gender) for FIVB Beach Volleyball World Tour Grand Slam & Open is organized as 32 teams divided into eight pools of four teams. The four teams within each pool compete in a Round Robin, and the top 3 within each pool advance to a single elimination knockout phase, with the top eight seeded teams automatically advancing to a second round awaiting the winners of the 16-

team first round. The losers of the semi-finals compete to determine third and fourth place in the event.

The seeding of teams within events is computed based on information from FIVB points earned at recent events. In particular, a team's seeding is based on Athlete Entry Points, which are the sum of the FIVB points for the teammates earned from the best six of the last eight FIVB events within the year prior to 14 days before the tournament. In the case of ties, the ranking of teams based on the sum of FIVB points over the entire year (called the Technical Ranking) is used. Given that the top eight seedings among teams who qualify for the elimination phase of a tournament have a distinct advantage by not having to compete in a first round, the ranking computation is an important component of competition administration.

2.2 FIVB POINT SYSTEM

Beach volleyball players competing in FIVB-governed events earn FIVB ranking points based on their performance in an event and on the category of the event. The more prestigious the event, the greater the number of ranking points potentially awarded. Table 1 displays the ranking points awarded per player on a team based on their result in the event, and based on the event type.

Table 1 indicates that teammates who place first in the World Championships will each earn 500 points, whereas finishing in first place at a Continental Cup will earn only 80 points. Teams who finish tied in fifth through eighth place (losing in the quarter-final round) all receive the same ranking points as indicated by the 5th place row in the table. Because points earned in an event are based exclusively on the final place in the tournament, and do not account for the specific opponents during the event, FIVB points can be understood as measures of tournament achievement, and not as compellingly as measures of ability. Additionally, rankings, seeding and eligibility are computed based on the accumulation of points based on a hard threshold (e.g., only points accumulated in the last year) as opposed to a time-weighted accumulation of points. Thus, a team whose players had an outstanding tournament achievement exactly 365 days prior to an event would be high-ranked, but on the next day would lose the impact of the tournament from a year ago.

The event-based FIVB points are used for a variety of purposes. In addition to seeding teams, they are used for eligibility for international events. For example, one qualification of teams to participate in the 2016 Olympics in Rio de Janeiro involved determining an Olympic Ranking, which was the sum of teams' FIVB points over the 12 best performances from January 2015 through June 12, 2016. Other factors were involved with the selection process, but the use of FIVB points was an essential element.

Tab. 1: Point scores by event type and place achievement in FIVB competition.

Tournament Rank	Senior World Ch	Grand Slam	Open/Cont. Tour Final	Cont. Tour Master/ Challenger	Cont. Tour Zonal/FIVB Age World Ch	Cont. Cup	Cont. Age Group Champs	Homolgated Nat'l Tour
1st	50	400	250	160	140	80	40	8
2nd	450	360	225	144	126	72	36	6
3rd	400	320	200	128	112	64	32	4
4th	350	280	175	112	98	56	28	2
5th-8th	300	240	150	96	84	48	24	1
9th-16th	250	180	120	80	70	40	20	0
17th-24th	200	120	90	64	56	32	16	0
25th-32nd	-	80	60	48	42	24	12	0
33rd-36th	150	40	30	0	0	0	0	0
37th-40th	100	0	0	0	0	0	0	0
41st-	-	20	15	0	0	0	0	0

2.3. PROBABILISTIC APPROACH TO RANKING

A major alternative to point accumulation systems is rating systems based on probabilistic foundations. The most common foundation for probabilistic rating systems is the class of linear paired comparison models (David, 1988). Suppose team i and j are about to compete, and let $y_{ij} = 1$ if team i wins and $y_{ij} = 0$ if team j wins. If we assume parameters θ_i and θ_j indicating the strengths of each team, then a linear paired comparison model assumes that

$$\Pr(y_{ij} = 1 | \theta_i, \theta_j) = F(\theta_i - \theta_j) \tag{1}$$

where F is a continuous cumulative distribution function (cdf) with a domain over $\mathbb{R}$. Choices of F typically used in practice are a logistic cdf or a standard normal cdf. In the case of a logistic cdf, the model can be written as

$$\operatorname{logit} \Pr(y_{ij} = 1) = \theta_i - \theta_j \tag{2}$$

which is known as the Bradley-Terry model (Bradley and Terry, 1952). The model was first proposed in a paper on tournament ranking by Zermelo (1929), and was developed independently around the same time as Bradley and Terry by Good (1955). The alternative when a standard normal distribution is assumed for F can be expressed as

$$\Phi^{-1}(\Pr(y_{ij} = 1)) = \theta_i - \theta_j \tag{3}$$

which is known as the Thurstone-Mosteller model (Mosteller, 1951; Thurstone, 1927). Two general references for likelihood-based inference for the strength parameters for these models are David (1988) and Critchlow and Fligner (1991). In linear paired comparison models such as Bradley-Terry and Thurstone-Mosteller, a linear constraint is usually assumed on the strength parameters to ensure identifiability such as that the sum of the strength parameters is 0.

Linear paired comparison models can be extended to acknowledge that teams may change in strength over time. Glickman (1993) and Fahrmeir and Tutz (1994) present state-space models for the dynamic evolution of team strength. The state-space model framework assumes a linear probability model for the strength parameters at time t, but that the parameters follow a stochastic process that governs the evolution to time $t+1$. For example, an auto-regressive paired comparison model may be implemented in the following manner. If θ_{it} is the strength of team i at time t, then the outcome of a match between teams j and k at time t is given by

$$\Pr(y_{jk} = 1 | \theta_{jt}, \theta_{kt}) = F(\theta_{jt} - \theta_{kt}) \tag{4}$$

and that for all $i = 1, \ldots, n$ (for n teams),

$$\theta_{i,t+1} = \rho \theta_{it} + \varepsilon_{it} \tag{5}$$

where $\varepsilon_{it} \sim \mathrm{N}(0, \sigma^2)$ and $|\rho| < 1$. Bayesian inference via Markov chain Monte Carlo simulation from the posterior distribution may be implemented as described by Glickman (1993). Other approaches to team strength evolution can be developed on the θ_{it} following a flexible function, such as a non-parametric smoother. Baker and McHale (2015) used barycentric rational interpolation as an approach to model the evolution of team strength.

One difficulty with likelihood-based inference (including Bayesian inference) for time-varying linear paired comparison models is evident when the number of teams, n, involved in the analysis is large. In such instances, the number of model parameters can be unwieldy, and the computational requirements for model fitting are likely to be challenging. Instead, a class of approximating algorithms for time-varying paired comparisons have relied on filtering algorithms that update strength parameter estimates based on current match results. These algorithms typically do not make use of the full information contained in the likelihood, so inference from these approaches is only approximate. However, the computational ease is the major benefit for using these approaches, which have become popular in settings for league competition that involve hundreds or thousands of competitors. Below we present several rating algorithms that are in current use for estimating competitor ability.

2.4. ELO RATING SYSTEM

In the late 1950s, Arpad Elo (1903-1992), a professor of physics at Marquette University, developed a rating system for tournament chess players. His system was intended as an improvement over the rating system in use by the United States Chess Federation (USCF), though Elo's system would not be published until the late 1970s (Elo, 1978). It is unclear whether Elo was aware of the development of the Bradley-Terry model, which served as the basis for his rating approach.

Suppose time is discretized into periods indexed by $t = 1, \dots, T$. Let $\hat{\theta}_{it}$ be the (estimated) strength of team i at the start of time t. Suppose during period t team i competes against teams $j = 1, \dots, J$ with estimated strength parameters $\hat{\theta}_{jt}$. Elo's system linearly transforms the $\hat{\theta}_{it}$, which are on the logit scale, to be on a scale that typically ranges between 0 and 3000. We let

$$R_{it} = C + \left(\frac{400}{\log 10}\right)\hat{\theta}_{it}$$

to be the *rating* of team i at the start of time period t, where C is an arbitrarily chosen constant (in a chess context, 1500 is a conventional choice). Now define

$$\mathrm{We}(R_{it}, R_{jt}) = \frac{1}{1 + 10^{-(R_{it} - R_{jt})/400}} \tag{6}$$

to be the "winning expectancy" of a match. Equation (6) can be understood as an estimate of the expected outcome y_{ij} of a match between teams i and j at time t given their ratings.

The Elo rating system can be described as a recursive algorithm. To update the rating of team i based on competition results during period t, the Elo updating algorithm computes

$$R_{i,t+1} = R_{it} + K \sum_{j=1}^{J} \left(y_{ij} - \mathrm{We}(R_{it}, R_{jt})\right) \tag{7}$$

where the value of K may be chosen or optimized to reflect the likely change in team ability over time. Essentially (7) updates a team's rating by an amount that depends on the team's performance (the y_{ij}) relative to an estimate of the expected score (the $\mathrm{We}(R_{it}, R_{jt})$). The value K can be understood as the magnitude of the contribution of match results relative to the pre-event rating; large values of K correspond to greater weight placed on match results relative to the pre-event rating, and low values of K connote greater emphasis on the team's pre-event rating. In some implementations of the Elo system, the value K depends on the team's pre-event rating, with larger values of K set for weaker ratings. This application of large K for weaker teams generally assumes that weaker teams have less stable strength and are more likely to change in ability.

Initial ratings by first-time teams in the Elo system are typically set in one of two ways. One approach is to estimate the team's rating by choosing a default starting rating R_{i0}, and then updating a rating using a large value of K. This is the approach implemented in the PlayerRatings R library described by Stephenson and Sonas (2016) in its implementation of the Elo system. An alternative approach, sometimes used in organized chess, is to compute a rating as a maximum likelihood estimate (e.g., for a Bradley-Terry model) based on a pre-specified number of matches, but treating the opponents' pre-event ratings as known in advance. Once an initial rating is computed, then the ordinary Elo updating formula in (7) would apply thereafter.

2.5. GLICKO RATING SYSTEM

The Glicko rating system (Glickman, 1999) was to our knowledge the first rating system set in a Bayesian framework. Unlike Elo's system in which a summary of a team's current strength is a parameter estimate, the Glicko system summarizes each team's strength as a probability distribution. Before a rating period, each team has a normal prior distribution of their playing strength. Match outcomes are observed during the rating period, and an approximating normal distribution to the posterior distribution is determined. Between rating periods, unobserved innovations are assumed to each team's strength parameter. Such assumed innovations result in an increase in the variance of the posterior distribution to obtain the prior distribution for the next rating period. West et al. (1985), Glickman (1993) and Fahrmeir and Tutz (1994) describe Bayesian inference for models that are dynamic extensions of the Bradley-Terry and Thurstone-Mosteller models. The Glicko system was developed as an approximate Bayesian updating procedure that linearizes the full Bayesian inferential approach.

A summary of the Glicko system is as follows. At the start of rating period t, team i has prior distribution of strength parameter θ_{it}

$$\theta_{it} \sim \mathrm{N}(\mu_{it}, \sigma_{it}^2). \tag{8}$$

As before, assume team i plays against J opposing teams in the rating period, each indexed by $j = 1, \ldots, J$. The Glicko updating algorithm computes

$$\begin{aligned}
\mu_{i,t+1} &= \mu_{it} + \frac{q}{1/\sigma_{it}^2 + 1/d^2} \sum_{j=1}^{J} g(\sigma_{jt})(y_{ij} - E_{ij}) \\
\sigma_{i,t+1} &= \left(\frac{1}{\sigma_{it}^2} + \frac{1}{d^2} \right)^{-1} + \delta^2
\end{aligned} \tag{9}$$

where $q = \log(10)/400$, and

$$\begin{aligned}
g(\sigma) &= \frac{1}{\sqrt{1 + 3q^2\sigma^2/\pi^2}} \\
E_{ij} &= \frac{1}{1 + 10^{-g(\sigma_{jt})(\mu_{it} - \mu_{jt})/400}} \\
d^2 &= \left(q^2 \sum_{j=1}^{J} g(\sigma_{jt})^2 E_{ij}(1 - E_{ij}) \right)^{-1},
\end{aligned} \tag{10}$$

and where δ^2 (the innovation variance) is a constant that indicates the increase in the posterior variance at the end of the rating period to obtain the prior variance for the next rating period. The computations in Equation (9) are performed simultaneously for all teams during the rating period.

Unlike many implementations of the Elo system, the Glicko system requires no special algorithm for initializing teams' ratings. A prior distribution is assumed for each team typically with a common mean for all teams first entering the system, and with a large variance (σ_{i1}^2) to account for the initial uncertainty in a team's strength. The updating formulas in Equation (9) then govern the change from the prior distribution to the approximate normal distribution.

By accounting for the uncertainty in team's strength through a prior distribution, the computation recognizes different levels of reliability of strength estimation. For example, suppose two teams compete that have the same mean strength, but one team has a small prior variance and the other has a large prior variance. Suppose further that the team with the large prior variance wins the match. Under the Elo system, the winning team would have a mean strength increase that equals the mean strength decrease by the losing team. Under the Glicko system, a different dynamic takes place. Because the winning team has a high prior variance, the result of the match outcome has a potentially great impact on the distribution of team strength resulting in a large mean increase. For the losing team with the low prior variance, the drop in mean strength is likely to be small because the team's ability is already reliably estimated and little information is gained from a loss to a team with a large prior variance. Thus, the winning team would likely have a mean strength increase that was large, while the losing team would have a mean strength decrease that was small. As of this writing, the Glicko system is used in a variety of online gaming leagues, including *chess.com*.

2.6. GLICKO-2 RATING SYSTEM

The Glicko system was developed under the assumption that strengths evolve over time through an auto-regressive normal process. In many situations, including games and sports involving young competitors, competitive ability may improve in sudden bursts. This has been studied in the context of creative productivity, for example, in Simonton (1997). These periods of improvement are quicker than can be captured by an auto-regressive process. The Glicko-2 system (Glickman, 2001) addresses this possibility by assuming that team strength follows a stochastic volatility model (Jacquier et al., 1994). In particular, Equation (5) changes by assuming $\varepsilon_{it} \sim \mathrm{N}(0, \delta_t^2)$, that is, the innovation variance δ_t^2 is time-dependent. The Glicko-2 system assumes

$$\log \delta_t^2 = \log \delta_{t-1}^2 + \nu_t \qquad (11)$$

where $\nu_t \sim \mathrm{N}(0, \tau^2)$ and where τ is the volatility parameter.

The updating process for the Glicko-2 system is similar to the Glicko system, but requires iterative computation rather than involving only direct calculations like the Glicko system. The details of the computation are described in Glickman (2001). The Glicko-2 system, like the Glicko system, has been in use for various online gaming leagues, as well as for over-the-board chess in the Australian Chess Federation.

2.7. STEPHENSON RATING SYSTEM

In 2012, the data prediction web site `kaggle.com` hosted the FIDE/Deloitte Chess Rating Challenge in which participants competed in creating a practical chess rating system for possible replacement of the current world chess federation system. The winner of the competition was Alec Stephenson, who subsequently implemented and described the details of his algorithm in Stephenson and Sonas (2016).

The Stephenson system is closely related to the Glicko system, but includes two main extra parameters. First, a parameter is included that accounts for the strengths of the opponents, regardless of the results against them. A rationale for the inclusion of the opponents' strengths is that in certain types of tournaments in which teams compete against those with similar cumulative scores, such as knockout or partial elimination tournaments, information about a team's ability can be inferred by the strength of the opponents. Second, the Stephenson system includes a "drift" parameter that increases a team's mean rating just from having competed in an event. The inclusion of a positive drift can be justified by the notion that teams who choose to compete are likely to be improving.

The mean update formula for the Stephenson system can be written as

$$\mu_{i,t+1} = \mu_{it} + \frac{q}{1/\sigma_{it}^2 + 1/d^2} \sum_{j=1}^{J} g(\sigma_{jt})(y_{ij} - E_{ij} + \beta) + \lambda(\bar{\mu}_t - \mu_{it}) \tag{12}$$

where $\bar{\mu}_t = J^{-1}\sum_{j=1}^{J} \mu_{jt}$, the average pre-event mean strength of the J opponents during period t, β is a drift parameter, and λ is a parameter which multiplies the difference in the average opponents' strength from the team's pre-period strength. An implementation of Stephenson's system can be found in Stephenson and Sonas (2016).

3. DATA AND RATINGS IMPLEMENTATION

Women's beach volleyball game data and end-of-year rankings were downloaded from `http://bvbinfo.com/`, an online database of international volleyball tournament results going back to 1970. All match results from FIVB-sanctioned tournaments from the years 2007-2015 were compiled, keeping record of the two teams involved in a match, the winner of the match, and the date of the match. We used match data from 2007-2014 to construct ratings from the four probabilistic rating systems, leaving match outcomes during 2015 for validation.

The data set consisted of 12,241 match game results. For the 2007-2014 period in which the rating systems were developed, a total of 10,814 matches were included, leaving 1427 match results in 2015 for validation. The matches were played by a total of 1087 unique teams. For our analyses, we considered a single athlete who partnered with two different players as two entirely different teams. This is a conservative assumption for our analyses because we treat the same player on two different teams as independent. However, this assumption can be justified by acknowledging that different levels of synergy may exist between player pairs.

During the 2007-2015 period, 72 teams played in at least 100 matches. The greatest number of matches any player pair competed in our data set was 550. At the other extreme, 243 teams competed exactly once in the study period.

The probabilistic rating systems described in Section 2 were implemented in the R programming language (R Core Team, 2016). The core functions to perform rating updates of the Elo, Glicko and Stephenson systems were implemented in the PlayerRatings library (Stephenson and Sonas, 2016). We implemented the Glicko-2 system manually in R.

We optimized the system parameters of the probabilistic rating systems in the following manner. Matches from 2007-2014 were grouped into rating periods of 3-month periods (January-March 2007, April-June 2007, ..., October-December 2014) for a total of 32 rating periods. The period lengths were chosen so that team strengths within rating periods were likely to remain relatively constant, but with the possibility of change in ability between periods. Given a set of candidate system parameters for a rating system, we ran the rating system for the full eight years of match results. While updating the ratings sequentially over the 32 periods, we computed a predictive discrepancy measure for each match starting with month 25, and averaged the discrepancy measure over all matches from month 25 through 32. That is, the first 75% of the rating periods served as a "burn-in" for the rating algorithms, and then the remaining 25% served as the test sample.

The match-specific predictive discrepancy for a match played between teams i and j was

$$-\left(y_{ij}\log\hat{p}_{ij}+(1-y_{ij})\log(1-\hat{p}_{ij})\right) \tag{13}$$

where y_{ij} is the binary match outcome, and $\hat{p}_{ij}$ is the expected outcome of the match based on the pre-period ratings of teams i and j. This criterion is a constant factor of the binomial deviance contribution for the test sample. This particular choice has been used to assess predictive validity in Glickman (1999) and Glickman (2001). It is also a commonly used criterion for prediction accuracy (called "logarithmic loss," or just log loss) on prediction competition web sites such as `kaggle.com`.

For the Elo system, $\hat{p}_{ij}$ was the winning expectancy defined in (6). For the Glicko, Glicko-2 and Stephenson systems, the expected outcome calculation accounts for the uncertainty in the ratings. The expected outcome is therefore computed as an approximation to the posterior probability that team i defeats team j. Glickman (1999) demonstrated that a good approximation to the posterior probability is given by

$$\hat{p}_{ij}=\frac{1}{1+10^{-g(\sqrt{\sigma_i^2+\sigma_j^2})(\mu_i-\mu_j)/400}} \tag{14}$$

where the function g is defined as in (10).

The optimizing choice of the system parameters is the set that minimizes the average discrepancy over the test sample. We determine the optimal parameters through a the Nelder-Mead algorithm (Nelder and Mead, 1965), an iterative numerical derivative-free optimization procedure. The algorithm is implemented in the R function `optim`.

4. RESULTS

The probabilistic rating systems were optimized as described in Section 3. The following parameter values were determined to optimize the mean predictive discrepancy in (13):

Elo: $K = 19.823$

Glicko: $\sigma_1 = 200.074$ (common standard deviation at initial rating period), $c = 27.686$

Glicko-2: $\tau^2 = 0.000177$, $\sigma_1 = 216.379$, $c = 30.292$

Stephenson: $\sigma_1 = 281.763$, $c = 10.378$, $\beta = 3.970$, $\lambda = 2.185$

The resulting mean predictive discrepancy across the test sample of matches is reported in Table 2. In addition to the mean predictive discrepancy measure, we also calculated a misclassification rate of match results for the 25% test sample. For each match in the test sample, a result was considered misclassified if the expected score of the match was greater than 0.5 for the first team in the pair according to the pre-match ratings and the first team lost, or if the expected score was less than 0.5 and the first team won. Matches involving teams with equal ratings were ignored in this computation.

Tab. 2: Rating system summaries based on optimized parameter values. The first column reports 10, 000 × the mean log loss score from the 25% test sample. The second column reports the fraction of matches in which the result went the opposite of the favored team according to the pre-match ratings.

Rating System	10, 000 × mean log loss	Misclassification Rate
Elo	2652.55	0.318
Glicko	2623.03	0.319
Glicko-2	2622.08	0.319
Stephenson	2590.72	0.310

The table indicates that the Elo system had the worst predictive accuracy in terms of log loss, followed by the Glicko and Glicko-2 systems which had comparable predictive accuracy. The accuracy based on the misclassification rate were similar for Elo, Glicko and Glicko-2. The Stephenson system had the best predictive performance of the four systems with a lower mean log loss, and a slightly lower misclassification rate.

The rating systems were assessed for calibration accuracy as shown in Figure 1. For each rating system, we sorted the pre-match predicted probabilities for the 25% test sample relative to the higher-rated team (so that the winning probability was 0.5 or greater). These probabilities were divided into 10 consecutive groups. Within each group, we computed the average result for the higher rated team along with the endpoints of a 95% confidence interval. Each confidence interval along with the sample mean across the 10 groups was plotted as a vertical segment. If a rating system were well-calibrated, the pattern of confidence intervals would fall on the line $y = x$ (shown as diagonal lines on the figure).

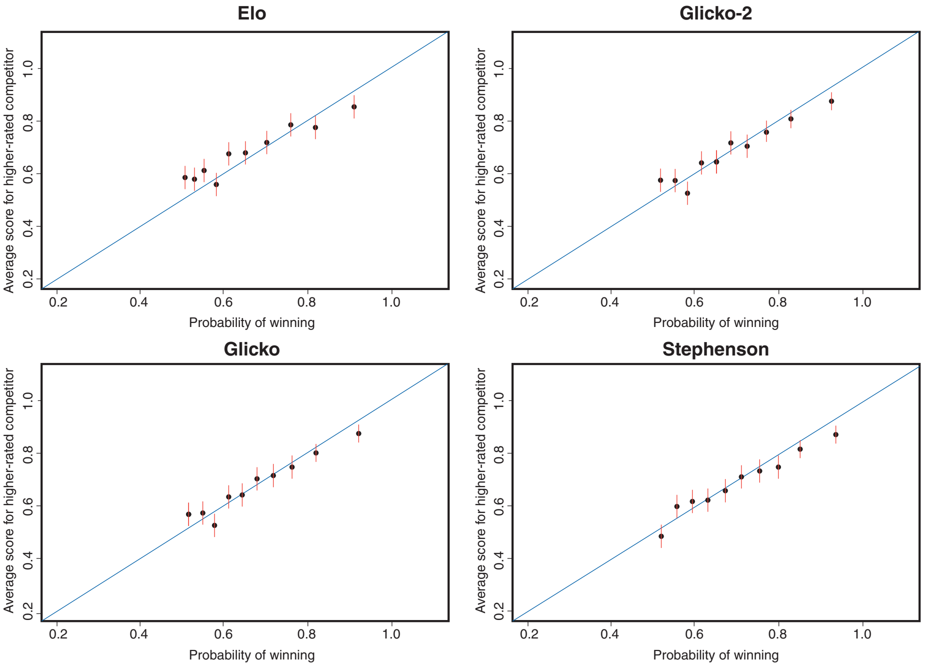

Fig. 1: Plots of average score and 95% confidence intervals computed from the 25% test sample for the favored team against the predicted proba- bility of winning for each of the four probabilistic rating systems.

Generally, the rating systems are all reasonably well-calibrated. In the case of Elo, Glicko and Glicko-2, small rating differences tend to underestimate the better team's performance, and in all cases large rating differences tend to over- estimate performances (indicated by the right-most confidence interval being en- tirely below the diagonal line). Elo has the least calibration consistency, with the fewest confidence intervals intersecting the diagonal line, and Glicko, Glicko-2 and Stephenson have the best calibration.

Tables 3 through 7 show the rankings at the end of 2014 of women's beach volleyball teams according to the different rating systems. Table 3 ranks teams according to total FIVB points (the sum over the two players in the team) while the ranks for the remaining tables are based on the order of the probabilistically-determined ratings.

Tab. 3: Top 15 teams at the end of 2014 according to FIVB points.

Rank	Team	Country	Points
1	Maria Antonelli/Juliana Felisberta	Brazil	6740
2	Agatha Bednarczuk/Barbara Seixas	Brazil	5660
3	April Ross/Kerri Walsh Jennings	United States	5420
4	Fan Wang/Yuan Yue	China	4950
5	Madelein Meppelink/Marleen Van Iersel	Netherlands	4640
6	Katrin Holtwick/Ilka Semmler	Germany	4610
7	Karla Borger/Britta Buthe	Germany	4580
8	Kristyna Kolocova/Marketa Slukova	Czech Republic	4420
9	Elsa Baquerizo/Liliana Fernandez	Spain	4360
10	Marta Menegatti/Viktoria Orsi Toth	Italy	4140
11	Ana Gallay/Georgina Klug	Argentina	3920
12	Talita Antunes/Larissa Franca	Brazil	3620
13	Carolina Salgado/Maria Clara Salgado	Brazil	3400
14	Maria Prokopeva/Evgeniya Ukolova	Russia	3220
15	Natalia Dubovcova/Dominika Nestarcova	Slovak Republic	3000

The probabilistic rating systems produce rank orders that have notable differences with the FIVB rank order. The team of Ross/Walsh Jennings is always either in first or second place on the probabilistic lists, but is third on the FIVB list. The top 10 teams on the FIVB list do appear on at least one probabilistic rating list, but it is worth noting that a non-trivial number of teams on the probabilistic rating lists do not appear on the FIVB top 15 list. For example, a team like Antunes/Franca are consistently in the top of the probabilistic rating systems, but is only ranked 30 in the FIVB rankings. This suggests that this team is having strong head-to-head results despite not achieving the tournament success of the top teams. The Elo top 15 list even includes a team ranked 83 on the FIVB list.

We compared the predictive accuracy of the four rating systems along with the FIVB system based on ratings/rankings at the end of 2014 applied to match results during 2015 in the following manner. A total of 1427 matches were recorded in 2015. Of the 1427 matches, 787 involved teams both having FIVB rankings in 2014 (only 183 teams appeared on the 2014 end-of-year FIVB list). We removed 4 of these games from our analyses as they involved teams with the same FIVB (tied) rank. We therefore restricted our predictive analyses to these $787 - 4 = 783$ matches. The result of each match played in 2015 was considered misclassified if the team with the higher rank from 2014 lost the match. Table 8 summarizes the misclassification rates for all five rating systems. The table indicates that the FIVB has the worst misclassification rate with greater than 35% of the matches incorrectly predicted. The Elo system is not much better, but Glicko, Glicko-2 and Stephenson have rates as low as 31-32%. McNemar's test (McNemar, 1947) for comparing the FIVB misclassification rate to the misclassification rates of the probabilistic systems was performed, with the p-values reported on Table 8. The difference in misclassification rates between the FIVB and Stephenson's system has a significantly low p-value (0.019), while the other differences are not significant at the 0.05 level.

Tab. 4: Top 15 teams at the end of 2014 according to Elo ratings.

Rating	Team	Country	FIVB Rank
1850	April Roùss/Kerri Walsh Jennings	United States	3
1839	Talita Antunes/Larissa Franca	Brazil	12
1819	Talita Antunes/Taiana Lima	Brazil	30
1775	Kristyna Kolocova/Marketa Slukova	Czech Republic	8
1773	Maria Antonelli/Juliana Felisberta	Brazil	1
1744	Laura Ludwig/Kira Walkenhorst	Germany	32
1727	Agatha Bednarczuk/Barbara Seixas	Brazil	2
1727	Katrin Holtwick/Ilka Semmler	Germany	6
1700	Carolina Salgado/Maria Clara Salgado	Brazil	13
1687	Madelein Meppelink/Marleen Van Iersel	Netherlands	5
1686	Fernanda Alves/Taiana Lima	Brazil	26
1674	Karla Borger/Britta Buthe	Germany	7
1672	Elsa Baquerizo/Liliana Fernandez	Spain	9
1665	Fan Wang/Yuan Yue	China	4
1662	Doris Schwaiger/Stefanie Schwaiger	Austria	83

Tab. 5: Top 15 teams at the end of 2014 according to Glicko ratings.

Rating	Team	Country	FIVB Rank
1918	April Ross/Kerri Walsh Jennings	United States	3
1903	Talita Antunes/Larissa Franca	Brazil	12
1847	Talita Antunes/Taiana Lima	Brazil	30
1763	Maria Antonelli/Juliana Felisberta	Brazil	1
1748	Laura Ludwig/Kira Walkenhorst	Germany	32
1747	Kristyna Kolocova/Marketa Slukova	Czech Republic	8
1730	Agatha Bednarczuk/Barbara Seixas	Brazil	2
1716	Madelein Meppelink/Marleen Van Iersel	Netherlands	5
1714	Carolina Salgado/Maria Clara Salgado	Brazil	13
1703	Fernanda Alves/Taiana Lima	Brazil	26
1691	Katrin Holtwick/Ilka Semmler	Germany	6
1684	Xinyi Xia/Chen Xue	China	27
1674	Elsa Baquerizo/Liliana Fernandez	Spain	9
1656	Karla Borger/Britta Buthe	Germany	7
1652	Laura Ludwig/Julia Sude	Germany	24

In addition to exploring the relationship between match outcomes in 2015 and a binary indicator of whether a team was more highly ranked in a given rating system, we investigated the relationship between match outcomes and the difference in rank on the 2014 lists. For this analysis, we included only matches involving teams that were in the top 200 in the end-of-2014 ranked lists from each rating system. This decision was to prevent the probabilistic rating systems incorporating matches involving teams that were far down the list and would result in a poor comparison to the analysis of matches involving FIVB-ranked teams. For each match, we computed the difference between the rank of the winner and loser. Boxplots of the match-specific rank differences appear in Figure 2. The figure shows that the four probabilistic rating system produce distributions of rank differences that are roughly comparable, with the Stephenson system having a slightly higher median rank difference for won matches than the other probabilistic systems. The FIVB system by comparison produces a substantially smaller median rank difference across the match winners. A 95% confidence interval for the mean rank difference based on FIVB 2014 rankings was (10.8, 15.5) whereas for the Stephenson 2014 rankings the 95% confidence interval was (18.3, 30.5). Based on simple two-sample t-tests, the mean rank differences between the FIVB and any of the probabilistic rating system ranks were significantly smaller at very low levels even conservatively accounting for test multiplicity.

Tab. 6: Top 15 teams at the end of 2014 according to Glicko-2 ratings.

Rating	Team	Country	FIVB Rank
1927	April Ross/Kerri Walsh Jennings	United States	3
1914	Talita Antunes/Larissa Franca	Brazil	12
1850	Talita Antunes/Taiana Lima	Brazil	30
1766	Maria Antonelli/Juliana Felisberta	Brazil	1
1754	Kristyna Kolocova/Marketa Slukova	Czech Republic	8
1754	Laura Ludwig/Kira Walkenhorst	Germany	32
1734	Agatha Bednarczuk/Barbara Seixas	Brazil	2
1720	Madelein Meppelink/Marleen Van Iersel	Netherlands	5
1716	Carolina Salgado/Maria Clara Salgado	Brazil	13
1708	Fernanda Alves/Taiana Lima	Brazil	26
1693	Katrin Holtwick/Ilka Semmler	Germany	6
1684	Xinyi Xia/Chen Xue	China	27
1678	Elsa Baquerizo/Liliana Fernandez	Spain	9
1658	Karla Borger/Britta Buthe	Germany	7
1657	Laura Ludwig/Julia Sude	Germany	24

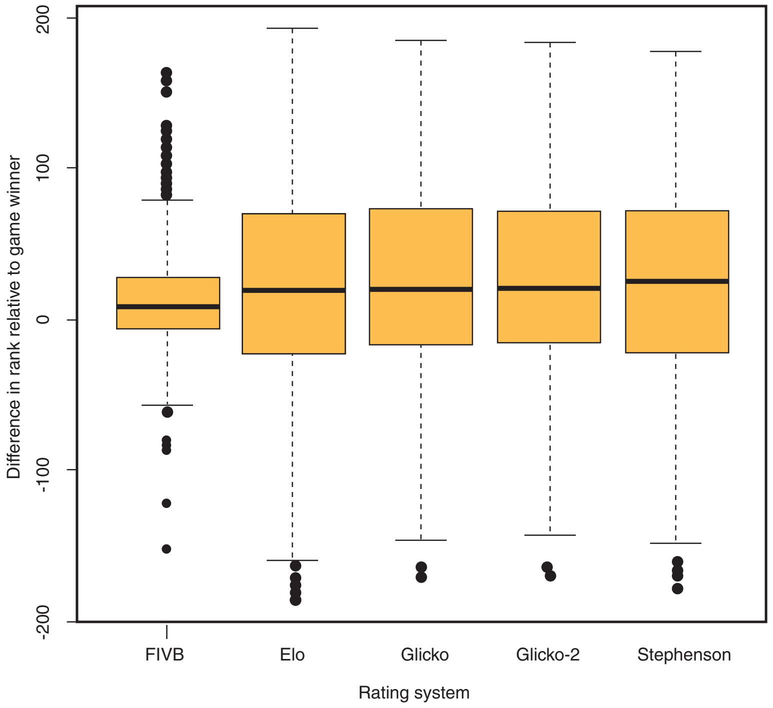

Fig. 2: Boxplots of the distribution of differences in 2014 rankings for each match played in 2015 relative to the winner of each match. A large rank difference indicates that the winner of a match had a much higher 2014 rank than the loser.

5. DISCUSSION AND CONCLUSION

The four probabilistic rating systems considered here appear to demonstrate solid performance in measuring women's beach volleyball team strength. The rating systems evidence roughly 31-32% misclassification rates for predicting future matches (the Elo system is slightly higher). By comparison, the FIVB point-based system has a greater than 35% misclassification rate. Given the fractional differences in misclassification rates among the probabilistic systems, the 4% misclassification difference is notable (and statistically significant comparing the FIVB and Stephenson systems). At a more fundamental level, the rating systems provide a means for estimating probabilities of match outcomes, a calculation not prescribed by the FIVB system. Because the focus of the probabilistic systems is in forecasting match outcomes, the ranked lists differ in substantive ways from the FIVB list. For example, the number 1 team on the 2014 FIVB list, Antonelli/Felisberta, is not only ranked lower on the probabilistic lists than the team Ross/Walsh-Jennings, but the estimated probability based on the probabilistic rating systems is that Ross/Walsh-Jennings would defeat Antonelli/Felisberta with a probability of between 0.71 and 0.75 for the Glicko, Glicko-2 and Stephenson systems.

Tab. 7: Top 15 teams at the end of 2014 according to Stephenson ratings.

Rating	Team	Country	FIVB Rank
2152	Talita Antunes/Larissa Franca	Brazil	12
2105	April Ross/Kerri Walsh Jennings	United States	3
2018	Talita Antunes/Taiana Lima	Brazil	30
1915	Maria Antonelli/Juliana Felisberta	Brazil	1
1900	Fernanda Alves/Taiana Lima	Brazil	26
1885	Laura Ludwig/Kira Walkenhorst	Germany	32
1879	Madelein Meppelink/Marleen Van Iersel	Netherlands	5
1859	Agatha Bednarczuk/Barbara Seixas	Brazil	2
1843	Kristyna Kolocova/Marketa Slukova	Czech Republic	8
1826	Laura Ludwig/Julia Sude	Germany	24
1823	Carolina Salgado/Maria Clara Salgado	Brazil	13
1818	Xinyi Xia/Chen Xue	China	27
1810	Katrin Holtwick/Ilka Semmler	Germany	6
1781	Elsa Baquerizo/Liliana Fernandez	Spain	9
1769	Marta Menegatti/Viktoria Orsi Toth	Italy	10

Among the four probabilistic rating systems, the Stephenson system appears to slightly outperform the other three. A curious feature of this system is that a team's rating increases due merely to competing regardless of the result. While this feature seems to be predictive of better performance, which may be an artifact that teams who are improving tend to compete more frequently, it may be an undesirable aspect of a system to be used on an ongoing basis to rate its teams. Teams could manipulate their ratings by choosing to compete frequently regardless of their readiness to compete. Nonetheless, for the purpose of predicting match outcomes, this system does the best out of the probabilistic methods we have considered.

As mentioned previously, our approach to measuring women's beach volleyball team strength is conservative in the sense that we treat teams that share a player as entirely distinct. For example, the teams Antunes/Franca and Antunes/Lima who share Talita Antunes are both high on the probabilistic rating lists. In the probabilistic rating systems, we treated these two teams as separate competitors, and did not take advantage of Antunes being a member on both teams. Rating systems for beach volleyball could arguably be improved by accounting for the players involved in teams. Indeed, the FIVB system focuses on the players' FIVB points in determining a team's points, and this is an important difference in the way rankings were constructed. We argue, however, that it is not obvious how to account for individual player strength contribution in the construction of team abilities within a probabilistic system. One attempt might be to consider a team's ability to be the average of the two players' ratings of the team. This approach has been used, for example, in Herbrich et al. (2007). On the other hand, in a game like volleyball it may be that the team strength is more determined by the skill of the worse player given that the worse player is the source of vulnerability on the team. This is clearly an area for further exploration and is beyond the scope of this paper. However, even treating teams who share a player as entirely distinct still leads to the probabilistic rating systems outperforming the FIVB system in predicting future performance.

Tab. 8: Misclassification rates for 783 matches played in 2015 based on rank orders at the end of 2014, and McNemar's test p-values comparing misclassification rates of the probabilistic systems against the FIVB system.

Rating System	Misclassification Rate	p-value against FIVB
FIVB	0.3563	—
Elo	0.3448	0.550
Glicko	0.3282	0.128
Glicko-2	0.3244	0.074
Stephenson	0.3142	0.019

One weakness of the probabilistic systems in their most basic form is that they do not distinguish between elite events and events on national tours that are not as competitive. Teams competing in elite events may display performances that are more representative of their underlying abilities and preparation. These events could therefore be considered more relevant in measuring team strength than lower-prestige events. The FIVB system explicitly captures the difference in levels of tournament prestige. Various modifications of the probabilistic systems can account for different levels of prestige. The most direct change would involve having the sum of residuals (difference of observed and expected outcomes) inflated or deflated by a multiplicative constant that depends on the prestige of the event. Elite events would be associated with larger multiplicative factors, which would reflect the greater opportunity for teams' ratings to change as a result of their observed performance. Incorporation of these factors, or other related solutions, is an area for further exploration and beyond the scope of this paper.

Should the FIVB be considering a probabilistic system as a replacement to the existing point-accumulation system? An argument can be made that it should. The point-based systems were developed in a setting where it was important for the ranking system to require only simple arithmetic to perform the computation. With the stakes being so high for whether teams are invited to elite tournaments, it is arguably more important to rank teams based on systems with a probabilistic foundation than to keep the ranking computation simple. Such a move would involve a change in culture and a clarification of the goals of a ranking system, but our feeling is that a probabilistic system is more consistent with the goals set for identifying the best women's beach volleyball teams.

REFERENCES

Baker, R.D. and McHale, I.G. (2015). Deterministic evolution of strength in mul- tiple comparisons models: Who is the greatest golfer? In *Scandinavian Journal of Statistics*, 42 (1): 180–196.

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

Critchlow, D.E. and Fligner, M.A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementa- tion on glim. In *Psychometrika*, 56 (3): 517–533.

David, H.A. (1988). *The method of paired comparisons*. Oxford University Press, New York, 2nd edn.

Elo, A.E. (1978). *The rating of chessplayers, past and present*. Arco Pub., New York.

Fahrmeir, L. and Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. In *Journal of the American Statistical Association*, 89 (428): 1438–1449.

Glickman, M.E. (1993). *Paired comparison models with time-varying parameters*. Ph.D. thesis, Harvard University. Unpublished thesis.

Glickman, M.E. (1999). Parameter estimation in large dynamic paired comparison experiments. In *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48 (3): 377–394.

Glickman, M.E. (2001). Dynamic paired comparison models with stochastic vari- ances. In *Journal of Applied Statistics*, 28 (6): 673–689.

Good, I.J. (1955). On the marking of chess-players. In *The Mathematical Gazette*, 39 (330): 292–296.

Herbrich, R., Minka, T. and Graepel, T. (2007). TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, 569–576.

Jacquier, E., Polson, N.G. and Rossi, P.E. (1994). Bayesian analysis of stochastic volatility models. In *Journal of Business & Economic Statistics*, 12 (4): 371– 389.

McNemar, Q. (1947). Note on the sampling error of the difference between cor- related proportions or percentages. In *Psychometrika*, 12 (2): 153–157.

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. In *Psychometrika*, 16 (1): 3–9.

Nelder, J.A. and Mead, R. (1965). A simplex method for function minimization. In *The Computer Journal*, 7 (4): 308–313.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www. R-project.org.

Simonton, D.K. (1997). Creative productivity: A predictive and explanatory model of career trajectories and landmarks. In *Psychological Review*, 104 (1): 66.

Stefani, R.T. (1997). Survey of the major world sports rating systems. In *Journal of Applied Statistics*, 24 (6): 635–646.

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. In *Journal of Quantitative Analysis in Sports*, 7 (4).

Stephenson, A. and Sonas, J. (2016). *PlayerRatings: Dynamic Updating Meth- ods for Player Ratings Estimation*. URL https://CRAN.R-project.org/ package=PlayerRatings. R package version 1.0-1.

Thurstone, L.L. (1927). A law of comparative judgment. In *Psychological review*, 34 (4): 273.

West, M., Harrison, P.J. and Migon, H.S. (1985). Dynamic generalized linear models and Bayesian forecasting. In *Journal of the American Statistical Asso- ciation*, 80 (389): 73–83.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. In *Mathematische Zeitschrift*, 29: 436– 460.