ARTICLE IN PRESS

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

A calibration method with dynamic updates for within-match forecasting of wins in tennis

Stephanie Kovalchik [a,b,*], Machar Reid [b]

[a] Institute of Sport, Exercise and Active Living, Victoria University, PO Box 14428, Melbourne 8001, VIC, Australia
[b] Sport Science and Medicine Unit, Tennis Australia, PO Box 6060, Richmond South 3121, VIC, Australia

ARTICLE INFO

Keywords:
Calibration
Probability forecasting
Regression
Sports forecasting
Turning points

ABSTRACT

In-match predictions of player win probabilities for professional tennis matches have a wide range of potential applications, including betting, fan engagement, and performance evaluation. The ideal properties of an in-play prediction method include the ability to incorporate both useful pre-match information and relevant in-match information as the match progresses, in order to update the pre-match expectations. This paper presents an in-play forecasting method that achieves both of these goals by combining a pre-match calibration method with a dynamic empirical Bayes updating rule. We present an optimisation rule for guiding the specifications of the dynamic updates using a large sample of professional tennis matches. We apply the results to data from the 2017 season and show that the dynamic model provides a 28% reduction in the error of in-match serve predictions and improves the win prediction accuracy by four percentage points relative to a constant ability model. The method is applied to two Australian Open men's matches, and we derive several corollary statistics to highlight key dynamics in the win probabilities during a match.

1. Introduction

Many of the statistical and mathematical models that have been developed in sport have focused on forecasting the outcomes of competitive events (Stekler, Sendor, & Verlander, 2010). Beyond their use for the betting market, sports forecasting models have potential use to both players and coaches in setting expectations and putting performance outcomes in the context of those expectations (Jordan, Melouk, & Perry, 2009). Performance forecasts have also been used to improve ratings systems and track athletes' abilities over time (Constantinou & Fenton, 2013; Irons, Buckley, & Paulden, 2014). In addition, advanced performance statistics, like win expectations, are increasingly of interest to broadcasters as part of their efforts to enrich their commentary and increase fan engagement (Barnett, O'Shaughnessy, & Bedford, 2011).

For the sport of tennis, a number of forecasting models have been proposed for predicting match outcomes. Regression models have been a common approach, and authors have considered a range of predictive factors, including player rankings (Klaassen & Magnus, 2003), player seedings (Boulier & Stekler, 1999), player demographics (Del Corral & Prieto-Rodriguez, 2010), and prize money earnings (Gilsdorf & Sukhatme, 2008). Point-based models have been another popular method. The distinctive thing about these models is that they begin with a forecast of winning a point on serve and use this estimate to derive the conditional probabilities of other events based on the assumption that point outcomes on serve are independent and identically distributed (Barnett & Clarke, 2005; Knottenbelt, Spanias, & Madurska, 2012; Newton & Keller, 2005; Spanias & Knottenbelt, 2012). Other match

* Correspondence to: Institute of Sport, Exercise and Active Living, Victoria University, PO Box 14428, Melbourne 8001, VIC, Australia.
E-mail addresses: skovalchik@tennis.com.au (S. Kovalchik), mreid@tennis.com.au (M. Reid).

https://doi.org/10.1016/j.ijforecast.2017.11.008
0169-2070/

forecasting approaches have considered the use of player Elo ratings (Kovalchik, 2016) or bookmaker odds (Leitner, Zeileis, & Hornik, 2010).

A recent study compared the predictive performances of ten tennis match forecasting approaches and found a notable degree of variation among the methods (Kovalchik, 2016), with point-level models having the lowest accuracy for predicting match wins. When considering the accuracy over a representative set of professional matches, forecasts based on player Elo ratings came closest to the performance of the betting market.

One major drawback of most tennis forecasting models is that they do not update as a match progresses. Only point-level models have the ability to modify the win prediction conditional on the scoreline. The limitation of the proposed point-level models is that they do not consider in-match information when updating their expectations of player serve performances (Barnett & Clarke, 2005; Knottenbelt et al., 2012; Newton & Keller, 2005; Spanias & Knottenbelt, 2012). Furthermore, point-level models have been shown to provide less accurate forecasts of match outcomes than either bookmakers or Elo ratings (Kovalchik, 2016). This suggests that prior methods could also improve the specification of the pre-match serve parameters.

Ideally, an in-match forecasting approach would make use of the most accurate performance information that is available both before and during the match. In their point-level model, Klaassen and Magnus (2003) developed an inversion method for obtaining estimates of the pre-match serve probability using match win formulae based on independent and identically distributed (IID) model assumptions. One implementation of the model that used pre-match betting odds to set the win expectations found that the Klaassen and Magnus (2003) model was in strong agreement with the implied probabilities of in-play betting markets (Easton & Uylangco, 2010).

One shortcoming of the Klaassen and Magnus (2003) model is that it ignores in-match information. Barnett et al. (2011), on the other hand, describe an approach for updating the serve probability inputs to IID formulae dynamically using a weighted estimator of the pre-match serve expectations and the observed serve performance up to the current point in the match. However, these authors use an *ad hoc* approach for specifying the weights of this average and do not make use of established match forecasting methods for specifying the pre-match parameters. Furthermore, they do not evaluate any dynamic updating approach using empirical match data or provide any comparison to the performances of fixed serve models.

To address these limitations of prior in-play forecasting approaches, the present paper develops a dynamic in-play prediction method and contrasts its performance to the constant serve model of Klaassen and Magnus (2003) using a large set of professional match data. The main advance of our method is its use of an empirical Bayes estimator to update the serve expectations with in-match information, and we use empirical studies to determine the optimal weighting of the pre-match and in-match serve information. In the implementation that we present, pre-match parameters are derived from player Elo ratings, which makes our method fully reproducible.

2. Methods

2.1. Notation

The symbol π_{ij} will be used to denote the probability that the ith player will win a tennis match when facing the jth opponent. For consistency, we will let the ith player be the player who is favoured to win—here, the player with the higher pre-match Elo rating.

Analytical formulae are available for predicting the outcome of any event in a tennis match from any scoreline, under the assumption that each point on serve is IID (Newton & Keller, 2005; O'Malley, 2008). The only inputs to these functions are the player's and the opponent's probabilities of winning a point on serve, which are equivalent to specifying the player's probabilities of winning a point on serve and return, denoted by p and q, respectively.

Given a best-of-n set match, let $M_n(p, q; \mathcal{S})$ be the IID probability that the favoured player will win the match, conditional on the match score $\mathcal{S}$. The score $\mathcal{S}$ contains information about the points, games, and sets won by each player up to the current point in the match. Under the IID assumption, it can be shown that at any score, the conditional probability of winning the match is a polynomial function of the probabilities p and q (Barnett & Clarke, 2005).

2.2. Pre-match win prediction

Although the calibration method developed in this paper could be applied to any pre-match prediction, we will use an approach based on the Elo ratings system to obtain the pre-match win prediction, $\tilde{\pi}_{ij}$. Elo ratings were first developed for rating player strength in chess (1978). Since its introduction, versions of Elo have been adopted for the measurement of player/team strength in many different sports (Stefani, 2011), including tennis.

For tennis, the Elo rating system developed by the data journalists at *FiveThirtyEight* has been shown to be one of the most accurate of the published methods for predicting match outcomes (Kovalchik, 2016). Their version uses the following recursive formula for updating the ith player's rating:

$$E_i(t+1) = E_i(t) + K_i * (\tilde{\pi}_{ij}(t) - W_i(t)), \tag{1}$$

where t refers to the tth match, $W_i(t)$ is an indicator of whether or not the player won, $E_i(t)$ is the player's Elo rating at the start of the match, and $\tilde{\pi}_{ij}(t)$ is the Elo-based prediction for a match win against the jth opponent. The parameter K_i is a function of the ith player's number of career matches (including all tour-level main draw matches) played up to time t. If we denote this match number by $m_i(t)$, then $K_i = 250/(m_i(t)+5)^{0.4}$. For Grand Slam tournaments, K_i is multiplied by 1.1 to give outcomes at the majors a 10% greater weight than all other tournaments.

The Elo prediction that player i will win the tth match, $\tilde{\pi}_{ij}(t)$, is equal to

$$\tilde{\pi}_{ij}(t) = \left(1 + 10^{\frac{(E_j(t)-E_i(t))}{400}}\right)^{-1}, \tag{2}$$

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

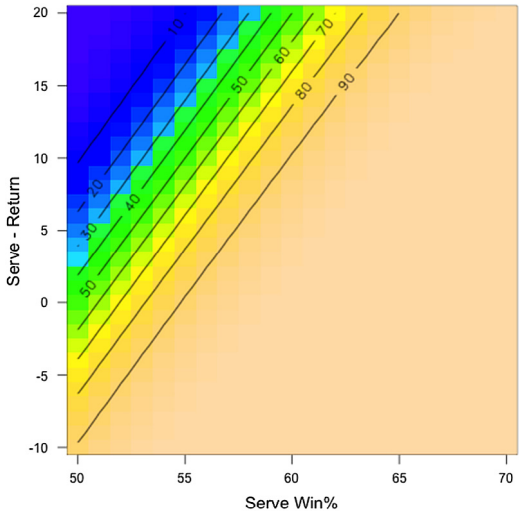

Fig. 1. Contour plot of a player's match win probability over a grid of server win probabilities and differences between serve and return win probabilities. The gradient denotes the match win probability, with the shading going from blue to yellow for low to high win probabilities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which approaches one as the rating difference, $E_j(t) - E_i(t)$, becomes more negative. For their first tour match, players begin with a rating of $E(1) = 1500$.

Like other direct models of match wins, the inputs into the standard Elo model are fixed at the start of the match, and therefore the forecast does not change as the match progresses.

2.3. Calibration of pre-match serve parameters

Suppose that we have a pre-match prediction of player i's chance of winning a match against opponent j of $\tilde{\pi}_{ij}$. Our goal is to develop an in-play model for tennis that is consistent with the pre-match prediction, $\tilde{\pi}_{ij}$. To do this, we use the IID formulae to 'back-calculate' the serve and return probabilities that give the equivalent prediction at the first point of the match. Mathematically, given the score at the start of the match before any points have been played, $\mathcal{S}_1$, the problem is to solve for p and q,

$$M_n(p, q; \mathcal{S}_1) = \tilde{\pi}_{ij}. \tag{3}$$

In this form, $M_n(p, q; \mathcal{S}_1)$ is a two-parameter non-linear system, which cannot guarantee unique solutions for p and q. This can be seen from the contour plot in Fig. 1, where the same match win probability can be achieved by multiple combinations of p and q.

The non-uniqueness problem can be dealt with by making the solution of Eq. (3) a one-parameter problem by introducing the difference $\delta = p - q$. If this difference were known, the match win probability could be written as a one-parameter function $M_n(p) = M_n(p, p - \delta)$. This corresponds to a two-stage process by which we first specify δ then solve for p, using $M_n(p) = \tilde{\pi}_{ij}$.

Since δ is not known, our analysis includes an investigation of different approaches for specifying δ.

2.4. In-match parameter updating

When the calibration to the pre-match prediction $\tilde{\pi}_{ij}$ is complete, it yields a fixed estimate for p. Then, one approach is to regard the serve probabilities as constants throughout the match, which is equivalent to assuming that the players are playing to expectation, what we will call a *constant ability* assumption. Under this assumption, changes in match predictions are due, not to changes in player performance, but only to the changing scoreline. Note that this is the model that was proposed originally by Klaassen and Magnus (2003) and evaluated by Easton and Uylangco (2010).

We also consider a *dynamic ability model* that relaxes the constant probability assumption and allows the in-match performance to influence the estimates of each player's expected performance on their serve. To achieve this, we need a point-by-point updating rule for the player serve probabilities. Accordingly, we propose the empirical Bayes estimator

$$\hat{p}(\mathcal{S}_m) = \theta(\mathcal{S}_m)p_0 + (1 - \theta(\mathcal{S}_m))\bar{p}(\mathcal{S}_m), \tag{4}$$

where p_0 is the calibrated pre-match serve probability, $\bar{p}(\mathcal{S}_m)$ is the observed in-match average number of points won on serve up to the nth point of the match with scoreline $\mathcal{S}_m$, and θ is a proportional weight that always takes a value between zero and one, but is allowed to change as the match progresses. When $\theta = 1$, the updates are equivalent to the constant probability approach.

The choice of θ determines the balance between the influences of the pre-match information (the 'prior') and the in-match serve information. The larger θ is, the more weight is given to the pre-match serve expectation. We will determine a reasonable balance between these two sources of information by setting θ equal to the dynamic weight

$$\theta(\mathcal{S}_m) = n_0/(n_0 + m), \tag{5}$$

where m is the number of points played on serve by the current player and n_0 is a parameter to be specified. In this specification, θ increases proportionally with n_0. Thus, n_0 can be thought of as the pre-match confidence in the pre-match expectations about a player's server performance; or, to put it another way, the amount of in-match evidence that is required to overturn the pre-match expectations.

We describe the above estimator as an empirical Bayes estimator because it is related closely to a special case of the beta-binomial conjugate model in Bayesian inference. The update rule for the serve probability in Eq. (4) is equivalent to the posterior estimate of the mean of a binomial random variable after m trials that has a beta prior distribution $Beta(\alpha, \beta)$, with $n_0 = \alpha + \beta$ (Casella, 1985). Thus, n_0 can be interpreted as the effective sample size placed on the prior evidence about a player's serve.

The proposed update rule requires a value for n_0. Rather than choosing an arbitrary value, we investigate an *optimal* choice of n_0 that reduces the inconsistency between the dynamic and pre-match serve expectations, while also reducing the error between the observed and dynamic serve

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

expectations. We achieve this goal by seeking to minimize the following loss function with respect to n_0:

$$Loss(n_0) = -m^{-1} \sum_i \lambda_i \{[y_i log(\hat{p}(\mathcal{S}_i)) + (1 - y_i) \times log(1 - \hat{p}(\mathcal{S}_i))] + (\hat{p}(\mathcal{S}_i) - p_0)^2\}, \quad (6)$$

where y_i indicates the number of points won on serve by a given player, out of m service points. This loss function is the sum of two loss measures. The first is the log-loss of the observed outcomes on serve and the dynamic estimate of the expected serve probability up to the mth point using the update rule in Eq. (4), which captures the error from the observed outcomes (Yuan et al., 2015). The second component is the squared error of the dynamic serve probability and the pre-match serve expectation. This component encourages point-to-point stability in the estimator, as it penalizes choices of n_0 that would result in large swings from one point to the next. Because maximising the accuracy of predictions is more crucial later in the match than earlier in the match, the loss components are weighted by $\lambda_i = i/\sum_{i=1}^{m} i$, an arithmetic decay from the first point up to the current point.

2.5. Validation data

Two datasets of professional tennis matches were gathered for evaluating the validity of the pre-match and within-match performances of the proposed models through out-of-sample testing.

2.5.1. Pre-match validation data

Match data were gathered from www.tennis-abstract.com for 24,884 Association of Tennis Professionals (ATP) matches and 20,716 Women's Tennis Association (WTA) matches. All matches were singles matches on the men's and women's world tours — the highest level of professional competition — played between 2011 and 2016. The dataset had one row per match, and included match-level summary statistics of the points won on serve for both the winner and the loser of the match. One year of burn-in data (2010) was used in order to give the 52-week moving averages of points won on serve adequate time to stabilise. The observed proportions of service points won were calculated for each player and match; these summaries formed the ground truth against which the different approaches for setting the pre-match serve expectations were compared.

2.5.2. In-match validation data

We evaluated the current performance of the in-play estimates of the constant ability and dynamic ability models by compiling point-level data for 688 men's matches from the 2017 season from the betting site www.flashscore.com. Of these matches, 331 (48%) were at Grand Slam events, and a total of 122,896 points were played. For each match, this dataset includes the score at the start of the point, an indicator for the serving player, and the winner of the point. The dataset does not include point-by-point betting odds, which would have to be collected from live scoring rather than from completed matches.

2.6. Statistical analysis

2.6.1. Pre-match validation

The first set of analyses focused on the initialisation of the pre-match parameters of the model. Several approaches for estimating the pre-match expected serve probability, $\hat{p}_i$, were considered. These included a constant δ method, a 52-week moving average of the expected serve frequency won, the opponent-adjusted estimate of the serve frequency won (proposed by Barnett & Clarke, 2005), and a linear regression approach of the serve frequency won, based on player Elo ratings, which were derived by the authors using the *FiveThirtyEight* model for all tour matches in the Open Era. Each method results in an estimate of δ, from which we can solve for Eq. (3) using the constant δ assumption described above. All methods used information up to but not including the current match, thus allowing out-of-sample validation. The comparative performances of the various approaches were based on their accuracies at predicting each player's in-match frequency of points won on serve in the validation dataset. The root mean squared error (RMSE) was used to assess this accuracy, and we selected the approach with the lowest RMSE.

Because there were several possible regression models for estimating the pre-match serve parameters, we evaluated the performances of a number of alternatives prior to conducting a comparison with the constant δ, moving average, and opponent-adjusted methods. For the regression approach, eight different model specifications were evaluated. The models examined were guided by a graphical analysis of the relationship between p and the pre-match Elo ratings in the 2011–2016 match data. One of the exploratory plots is shown in Fig. 2. It suggests a moderate linear relationship between the serve probability p and the log ratio of the player and opponent Elo ratings ($r = 0.35$), which turns out to be one of the stronger correlations considered. Models were trained with a ten-fold cross-validation of 70% of the validation dataset. The predictive performances of the different methods were evaluated by the cross-validation error in the training data and the RMSE in the 30% test sample.

Three of the approaches considered provide estimates of δ through estimates of the player and opponent serves. One method uses a constant δ and sets this constant equal to the average difference between player serve and return performances. Our exploratory work has found that the variation in the difference between a player's serve and return performances, δ, is generally small. Thus, we use constants that are equal to the averages observed in the men's and women's professional tours, namely $\delta = 0.25$ for the men and $\delta = 0.15$ for the women. We note that this yields initial conditions for the differences between serve and return performances that are in close agreement with those reported by Easton and Uylangco (2010).

2.6.2. Choosing n_0

After identifying the best-performing method for the pre-match serve parameters, we used this method to investigate the optimal prior confidence for the empirical Bayes dynamic probability model. The minimisation of the

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

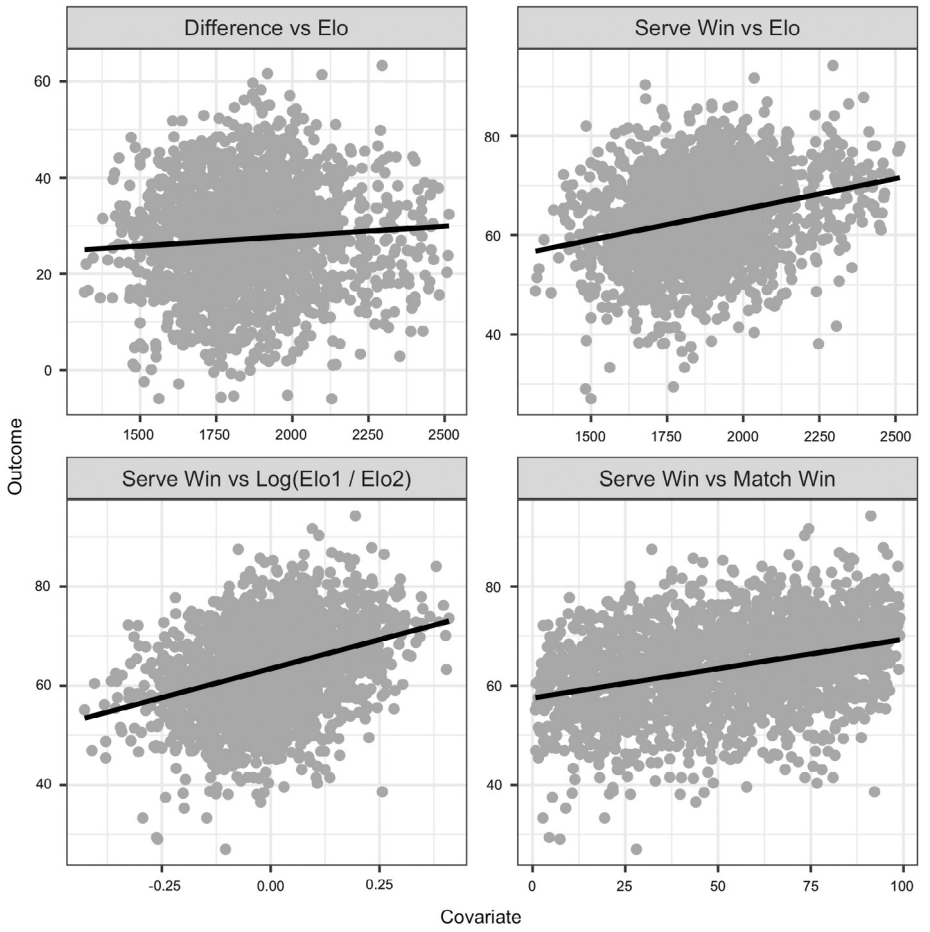

Fig. 2. Linear correlation plots exploring the relationships between (a) the difference in serve and return percentages won and a player's Elo rating ($r = 0.08$) (top left), (b) the serve percentage won against a player's Elo rating ($r = 0.27$) (top right), (c) the serve percentage won and the log ratio of player and opponent Elo ratings ($r = 0.35$) (bottom left), and (d) the serve percentage won and a player's Elo forecast for winning the match ($r = 0.34$) (bottom right).

loss function described in Eq. (6) was performed separately for each match, and the search interval for n_0 was 10 to 1000, which are equivalent to prior confidence of 10 and 1000 service points played, respectively. We summarise the distribution of the optimal prior level of confidence across matches and inspect examples of matches that have been assigned low and high prior confidence.

2.6.3. In-match validation

The validity of the in-play model was evaluated on the out-of-sample 2017 validation dataset. The validation included comparisons of the expected and actual serve performances and the expected and actual match winners. The serve performances were evaluated by calculating the point-by-point RMSEs for the predicted serve probability. All performance measures were calculated for each point in the sequence in which the points occurred, which emulates a real-time forecasting application and ensures that the performance measures were out-of-sample measures. We then evaluated the win performance by calculating the point-by-point frequency that the correct winner was selected, based on the player who had a greater than 50% win chance. For the dynamic model, the value of n_0 was set to a constant of 50 throughout, based on the characteristics of the optimal n_0 from the training data analysis. Predictive performance summaries are reported for all matches together and for the subgroups of Grand Slam and lower-tier matches separately.

2.6.4. Illustration

We demonstrate the application of the in-match prediction method by illustrating the dynamic and fixed methodologies for a men's singles match and contrasting the predictions with in-play betting odds. Historical betting data are not open to the public, but the authors were able to obtain historical betting data for one men's singles match from Fracsoft Ltd (http://www.fracsoft.com) at no charge. These betting data are based on Betfair Australia betting exchange prices over the duration of the match. These data were merged with a point-by-point dataset for a Grand Slam match that includes the elapsed time at the start of each point, obtained from the repository https://github.com/JeffSackmann/tennis_slam_pointbypoint. Pre-point odds were determined by matching the time at the start of the point with the time in the betting data, and a probability was derived from the odds by using the Shin method to correct for overround (Shin, 1993).

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

Table 1
Ten-fold cross-validation (CV) and root mean squared prediction errors (RMSE) of pre-match serve prediction models.

Covariate	ATP		WTA	
	CV error	RMSE	CV error	RMSE
Elo_i	0.74	8.67	0.83	9.26
$log(Elo_i/Elo_j)$	0.68	8.35	0.77	8.85
$\tilde{\pi}(Elo_i, Elo_j)$	0.68	8.32	0.77	8.87
$\hat{q}_j$	0.74	8.72	0.85	9.41
$\hat{p}_{BC}$	0.61	7.91	0.76	8.83
$log(Elo_i/Elo_j) + \tilde{\pi}(Elo_i, Elo_j)$	0.68	8.32	0.77	8.82
$log(Elo_i/Elo_j) + \hat{q}_j$	0.66	8.23	0.76	8.80
$log(Elo_i/Elo_j) + \hat{p}_{BC}$	0.58	7.71	0.72	8.56

Notes: Elo_i is the Elo rating of the ith player, $\tilde{\pi}(Elo_i, Elo_j)$ is the Elo-based match win expectation, $\hat{q}_i$ is the 52-week average return proportion won by the ith player, and $\hat{p}_{BC}$ is the Barnett and Clarke serve expectation. The performance measures have been multiplied by 100.

Table 2
Root mean squared errors of the pre-match expectations for the proportion of points won on serve.

Subgroup	N	Constant δ	Moving average	Barnett and Clarke	Regression
ATP					
All	24,884	0.083	0.081	0.079	0.079
Top 30	13,618	0.082	0.078	0.075	0.077
Lower rank	11,266	0.084	0.083	0.083	0.082
WTA					
All	20,716	0.088	0.088	0.086	0.089
Top 30	11,270	0.087	0.088	0.084	0.089
Lower rank	9,428	0.090	0.088	0.087	0.089

The final analysis section presents several corollary statistics that can be derived from the point-by-point win predictions and that are of especial interest for highlighting key points of the match for either performance evaluation or broadcasting.

All analyses presented here were completed in the R statistical programming language. Code for implementing both the dynamic and constant ability models, along with an example dataset, are provided as supplementary material.

3. Results

3.1. Pre-match estimation

3.1.1. Regression model selection

Of the regression models considered in Table 1 for estimating the pre-match serve probability, the cross-validation error ranged from 0.58 to 0.74 for the ATP and 0.72 and 0.85 for the WTA. The RMSEs for the test data ranged from 7.71 to 8.61 for the ATP and 8.56 and 9.41 for the WTA. Although each method showed larger errors for the WTA serve estimates than for the ATP, the model that included the log ratio of player and opponent Elo ratings and the pre-match Barnett and Clarke expectation for the serve win percentage had the lowest cross-validation and prediction errors for both tours (0.58 and 7.71 for the ATP; 0.72 and 8.56 for the WTA), and was the model selected for comparison against the non-regression pre-match estimation methods.

3.1.2. Prediction performance

Table 2 shows the RMSEs for the pre-match serve expectations over all matches and in subgroups defined by the ranking of the highest-ranked player in the match. The calibration based on the opponent-adjusted inputs from the Barnett and Clarke method resulted in the smallest errors overall (0.079 for the ATP and 0.086 for the WTA). The constant δ method, which is the easiest of the four to implement, had higher errors than the other methods in general. Overall, the errors of the constant δ method were 5% larger than those of the best-performing Barnett and Clarke approach. The performances of the other approaches fell in between. Thus, the differences in errors among methods were relatively small.

More notable differences were found when we considered the performances in the rank subgroups. For matches involving a Top 30 player, the Barnett and Clarke method had an even lower error rate relative to the other approaches (0.075 for the ATP and 0.084 for the WTA; Table 2). All methods were less accurate in their serve predictions for lower-ranked players. We also observed higher error rates in general for the WTA comparisons. Both of these findings could be attributable to the smaller sample sizes of matches for these groups.

3.2. In-play estimation

3.2.1. Optimising dynamic probability updating

Using optimal priors for the parameter n_0 in the dynamic probability model was found to have a median of 23 and a mean of 252 in the ATP dataset, indicating a high degree of positive skew. Because we expected the optimal characteristics to be influenced by the discrepancy between the pre-match serve performance and the actual in-match serve performance, we evaluated the summary of the optimal distributions for the parameter n_0 against the magnitude of the discrepancy between the pre-match and

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

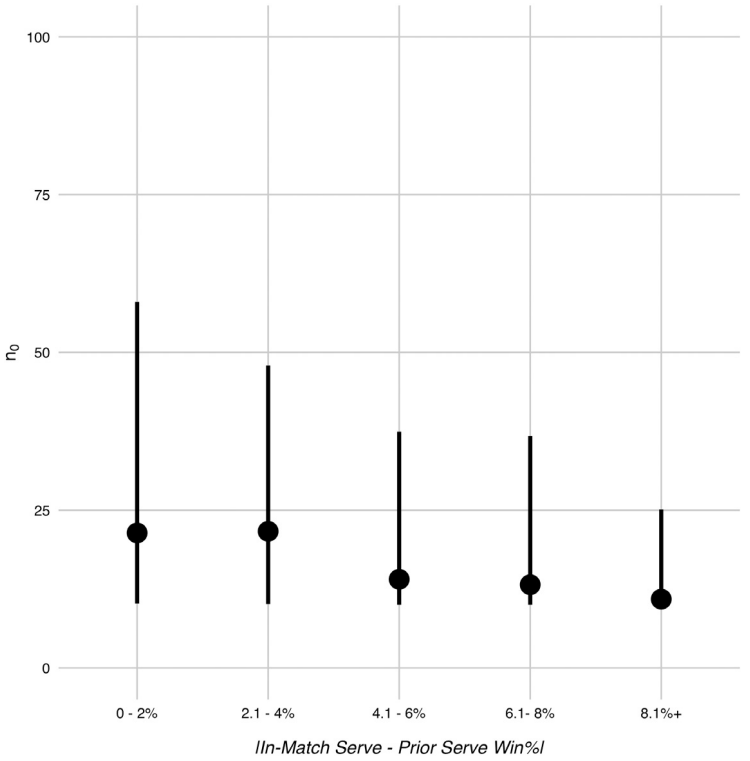

Fig. 3. Optimal n_0 for the pre-match serve expectation against the difference between the pre-match expected and observed in-match serve probabilities (as a percentage). The summaries are based on the ATP validation dataset. The points denote the mean and the lines the interquartile ranges.

Table 3
In-play predictive performances of the constant ability and dynamic ability models for the 2017 ATP validation dataset.

Matches	Serve RMSE		Win accuracy	
	Constant	Dynamic	Constant	Dynamic
Overall ($N = 688$)	0.144	0.104	69.8	74.0
Grand Slams ($N = 331$)	0.136	0.091	70.9	75.4
Other ($N = 357$)	0.155	0.120	68.2	70.8

Serve RMSE = point-by-point RMSE in the expected serve probability; win accuracy = point-by-point frequency that the correct winner was selected.

in-match performances. Fig. 3 shows that the median optimal prior tended to decrease as the discrepancy increased. Since a higher value of n_0 puts more weight on the pre-match serve performance, these results show that the loss function correctly placed less confidence on the prior serve performance when it was further from a player's true in-match performance.

Two specific matches illustrate this behaviour more clearly. The first match, shown in the top panel of Fig. 4, is an example where the server, Carreno-Busta, was highly consistent with the pre-match serve performance more than half-way through the match, yielding a minimal discrepancy between the pre-match and in-match serve performances. For this match, the optimal prior for Carreno-Busta was approximately 200 points, a high level of implied confidence, which allowed the pre-match serve expectation to dominate the point-to-point serve estimate throughout the match. In contrast to this low-discrepancy scenario, Delbonis's serve performance (in the bottom panel) was almost universally below the pre-match expectation. This resulted in a low value of 20 for the implied service points on the prior, allowing the in-match performance to dominate the point-to-point serve expectations as the match progressed.

On the women's side, the optimisation procedure placed slightly less confidence on the prior, having a median n_0 of 11 and mean of 49. We also observed the priors to be distributed more uniformly across groups, defined by the magnitude of the difference between the prior and the in-match serve performance (Fig. 5).

3.2.2. In-play validation

When we applied the dynamic and constant ability models to all points across 688 men's matches in the 2017 season, we found that the dynamic model improved the predictions of the in-match serve probability and win expectations (Table 3). The point-by-point RMSE for the dynamic model was 0.104 overall, which was 28% lower than that of the constant ability model. Furthermore, the dynamic model selected the correct winner more frequently than the constant ability model, being correct for 74% of

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

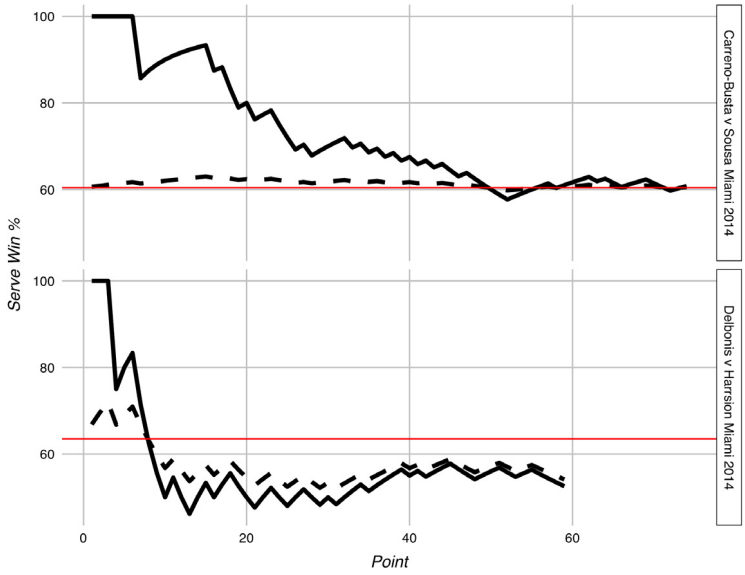

Fig. 4. Comparison between the in-match average points won on serve (solid black) and the dynamic serve prediction (dotted line) for a match, along with the pre-match serve expectation (red line). The top panel shows an example where the difference between the pre-match and observed match serve performances is small, while the bottom panel shows an example where this difference is large. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

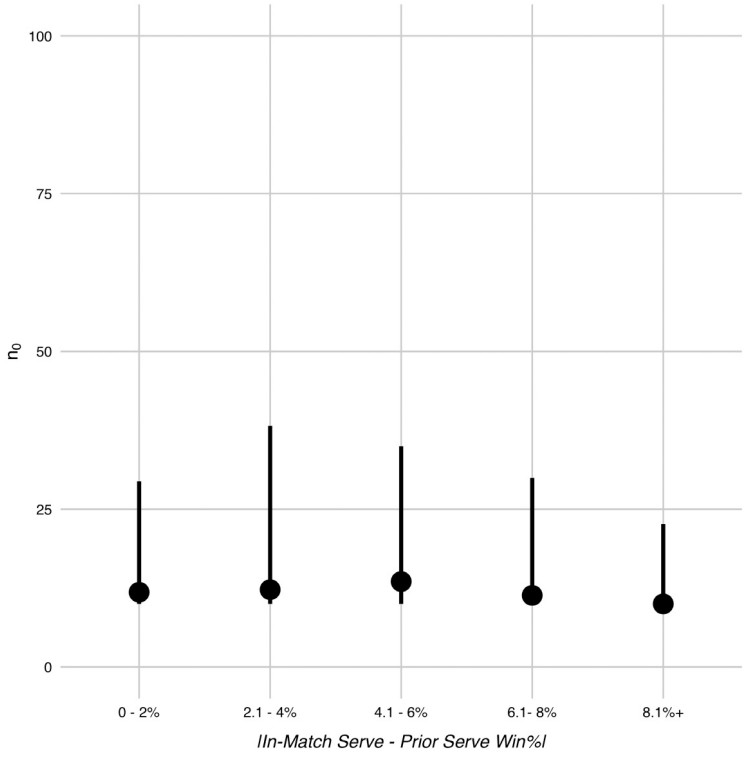

Fig. 5. Optimal n_0 for the pre-match serve expectation against the difference between the pre-match expected and observed in-match serve probabilities (as a percentage). The summaries are based on the WTA validation dataset. The points denote the mean and the lines the interquartile range.

points overall, while the static model was correct 70% of the time.

The performance differences were observed among both Grand Slam matches, where men play a best-of-five set format, and other events, where best-of-three is played (Table 3). This indicates that the prediction improvement with the dynamic model does not depend on either the match format or the tournament level.

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

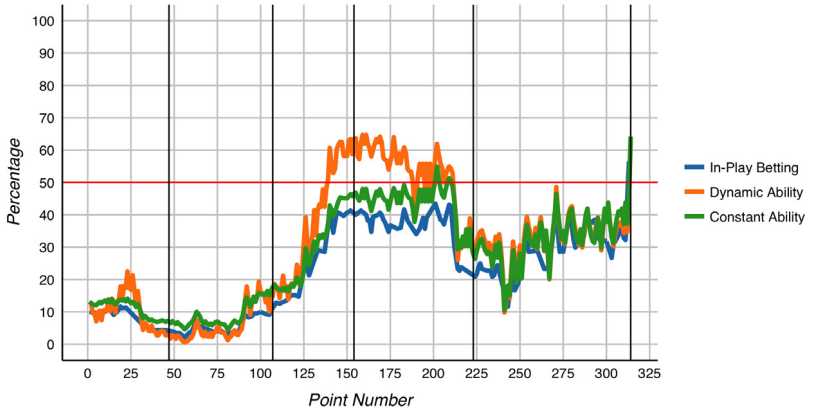

Fig. 6. In-match win predictions for Stan Wawrinka in the 2014 Australian Open quarter-final against Novak Djokovic. The win predictions shown are from in-play betting odds and the constant and dynamic probability methods of the current paper. The vertical lines denote the end of a set and the horizontal red line marks equal win chances for each competitor. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2.3. Illustration

This section provides a detailed illustration of the estimation of the dynamic and constant ability models. The match used for the illustration is the 2014 Australian Open quarter-final match between Stan Wawrinka and Novak Djokovic, which Wawrinka won in five sets, 2-6 6-4 6-2 3-6 9-7.

The pre-match Elo prediction gave Wawrinka a 13% probability of winning the match. The Barnett and Clarke serve calibration based on that pre-match win forecast put an initial expectation of 59% on Wawrinka winning a point on serve. Across the five sets played, Wawrinka's service percentage won ranged from a low of 52% to a high of 70%, with a considerable degree of set-to-set variation that demonstrates the rationale of dynamic in-match serve updating.

When we contrast the point-by-point forecasts of the constant and dynamic ability models against in-play betting odds for this match, we can see that the methods are generally in good agreement throughout the match (Fig. 6). After Wawrinka won the second set, though, the dynamic and constant ability models were more optimistic than the betting odds about Wawrinka's chances of winning. The greatest divergence among the methods happened between the middle of the third set and the middle of the fourth set. During this period, the dynamic model gave a win forecast for Wawrinka that was approximately 10 percentage points higher than that of the constant ability model and 20 percentage points higher than the in-play betting odds. This divergence can be explained by the fact that the second and third sets were the two in which Wawrinka had his highest serve percentage won and Djokovic had his lowest, in-match fluctuations that the dynamic model responds to the most.

3.2.4. Corollary statistics for broadcast

Using a second illustrative match, we derive several corollary statistics from the in-match forecasts that can help to identify key points in a match. The match that we consider in this section is the 2015 Australian Open men's final, in which Novak Djokovic defeated Andy Murray in four sets, 7-6(5) 6-7(4) 6-3 6-0.

Using the changes in win probabilities, we can define the biggest *turning points* in the match as the single points that had the largest changes in Djokovic's win expectations. The top five turning points shown in Table 4 reveal that the tiebreak points in the first and second sets were some of the most critical in the match. Murray missed two big opportunities to shift the odds in his favor in the first set tiebreak. Djokovic also suffered sizeable reductions in his win chances during the tiebreaks, with two of the biggest declines in his odds for the match coming in the first points of the first and second set tiebreakers.

We can also define the match *tipping point* as the first point at which a player takes a lead which is then maintained for consecutive points until the conclusion of the match. As the match progresses, only potential tipping points would be known, and, in a tightly contested match, we would expect the holder of the tipping point to reverse multiple times. In the Djokovic final against Murray, the definitive tipping point happened when Djokovic got to 15-0 in the eighth game of the third set on Murray's serve. At that point his win chance rose to 55%, and he increased that advantage over the remainder of the match.

We can also define the *peak win point* as the point at which a player achieved their highest win chance up to the current point in the match. For example, in the first set, Djokovic's peak win point occurred at the start of the sixth game when he was at 4-1, at which point his win chance increased to 97%, which was never exceeded for the remainder of the set. Peak points for the match or within a set can help to determine how definitive and dominant a player's win was.

4. Discussion

We have developed a novel approach for the dynamic in-match forecasting of winners of professional tennis matches. The features of our method build upon the inversion approach of Klaassen and Magnus (2003) and the dynamic updating suggested by Barnett et al. (2011). In

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.

Table 4
Top five turning points in the 2015 Australian Open men's final.

(Set, game, point) score	Player serving	Who won	Change in Djokovic win%
(1, 13, 5-5)	Murray	Djokovic	+16.2
(1, 13, 4-2)	Murray	Djokovic	+11.4
(1, 13, 0-0)	Djokovic	Murray	−11.1
(2, 10, 30-40)	Djokovic	Djokovic	+10.3
(2, 13, 0-0)	Murray	Murray	−8.9

combining the strengths of these two independent approaches, our results show that we improve on the accuracy of prior methods without adding greatly to their computational burden. These features should make our forecasting tool advantageous for practical applications.

One of the major contributions of the present paper is that it is the first to investigate an optimal specification for an empirical Bayes dynamic updating rule for serve probabilities. The minimisation function presented here promotes the selection of weights for the pre-match serve expectation that balance the in-match squared error and the point-to-point stability of estimates. Our assessment of the optimal n_0 suggested a range of 25–50 for the men's and women's prior confidence. The typical length of a men's Grand Slam match is 240 points, while that for women is 145 points (Kovalchik & Ingram, 2017). Thus, with $n_0 = 40$ for a men's match, we would expect the influence of the pre-match expectation to be 50% by the completion of one-third of a typical match at the Majors, while $n_0 = 20$ would give the equivalent influence in proportion to the expected match length for a women's match.

We have shown that the dynamic model with empirical Bayes updating improves the predictive performance of the constant ability model that was proposed by Klaassen and Magnus (2003). This improvement was found even once we included several simplifying assumptions in our updating rule. Namely, we use a constant prior confidence in the Bayes update across players and matches. One might prefer to allow the confidence to vary as a function of player attributes or match format, or to allow the uncertainty in the prior parameter to be incorporated formally through a fully Bayesian approach. We have opted for a simpler approach here because of its computational advantages, but recognise that these alternatives could offer performance gains that future research could explore.

A key component of the present paper's forecasting approach is its calibration of serve expectations at the start of each match. In this work, we have used the pre-match forecasts based on player Elo ratings to initialise the probabilities of winning on serve. However, our method is not limited to a specific type of pre-match forecast, but only requires a valid win expectation. Thus, practitioners could equally well apply our method with their own preferred pre-match win prediction, such as an expert forecast or the implied win probabilities from the betting market. With more accurate pre-match forecasts, we would expect a greater accuracy of in-match predictions. In this work, we have preferred to use the Elo model for our pre-match predictions because of both its demonstrated accuracy over published methods (Kovalchik, 2016) and its reproducibility.

In general, our analyses found a greater predictive error in the estimated serve probabilities for the WTA than for the ATP. Because of the large validation samples used for both, we believe that this reflects a greater difficulty in estimation for women's matches than for men's matches. Unlike male players, women do not play a best-of-five set format at Grand Slams. Also, the proportion of points won on serve is generally lower for the women's tour than for the men's tour. Additional research is needed to determine whether these or other factors can explain the gender differences in the performances of tennis forecasting methods.

Further directions for improving on the current work could include the investigation of alternative optimisation functions and the incorporation of covariate effects in the serve performance estimation. A growing body of research has demonstrated that players are affected by the scoreline, but that the magnitudes of these effects vary from player to player (González-Díaz, Gossner, & Rogers, 2012; Klaassen & Magnus, 2001; Kovalchik & Ingram, 2016). While such non-IID effects have been detected in large statistical association studies, it is yet to be determined whether these effects are large enough to have a meaningful impact on the point-to-point predictions of win outcomes at the elite level.

5. Conclusions

We have presented a novel approach for forecasting tennis matches in play that allows forecasters to predict win expectations at any point in a match. The method that we present allows the forecaster to set pre-match serve expectations that are consistent with the pre-match odds for a given matchup. Furthermore, we show how the most up-to-date information on each player's performance on serve can be used to update the pre-match expectations, thus allowing the method to adapt to the uncertainties of how players will actually perform on any given day. These improvements and the computational ease of the method make it a tool that is of practical use for tennis forecasting.

Acknowledgments

We would like to thank the staff of Fracsoft Ltd for their willingness to provide a historical dataset of in-play tennis betting odds for the purpose of this research study.

Funding

This research did not receive any specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supplementary model code and data

An R package with code to implement the in-match forecasting model and a data example has been made available as supplementary material and can be found online at https://doi.org/10.1016/j.ijforecast.2017.11.008.

References

Barnett, T., & Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, *16*(2), 113–120.

Barnett, T., O'Shaughnessy, D., & Bedford, A. (2011). Predicting a tennis match in progress for sports multimedia. *OR Insight*, *24*(3), 190–204.

Boulier, B. L., & Stekler, H. O. (1999). Are sports seedings good predictors?: An evaluation. *International Journal of Forecasting*, *15*(1), 83–91.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *The American Statistician*, *39*(2), 83–87.

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. *Journal of Quantitative Analysis in Sports*, *9*(1), 37–50.

Del Corral, J., & Prieto-Rodriguez, J. (2010). Are differences in ranks good predictors for Grand Slam tennis matches? *International Journal of Forecasting*, *26*(3), 551–563.

Easton, S., & Uylangco, K. (2010). Forecasting outcomes in tennis matches using within-match betting markets. *International Journal of Forecasting*, *26*(3), 564–575.

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

Gilsdorf, K. F., & Sukhatme, V. A. (2008). Testing rosen's sequential elimination tournament model incentives and player performance in professional tennis. *Journal of Sports Economics*, *9*(3), 287–303.

González-Díaz, J., Gossner, O., & Rogers, B. W. (2012). Performing best when it matters most: Evidence from professional tennis. *Journal of Economic Behavior & Organization*, *84*(3), 767–781.

Irons, D. J., Buckley, S., & Paulden, T. (2014). Developing an improved tennis ranking system. *Journal of Quantitative Analysis in Sports*, *10*(2), 109–118.

Jordan, J. D., Melouk, S. H., & Perry, M. B. (2009). Optimizing football game play calling. *Journal of Quantitative Analysis in Sports*, *5*(2), 1–34.

Klaassen, F. J., & Magnus, J. R. (2001). Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, *96*(454), 500–509.

Klaassen, F. J., & Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, *148*(2), 257–267.

Knottenbelt, W. J., Spanias, D., & Madurska, A. M. (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers and Mathematics with Applications*, *64*(12), 3820–3827.

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, *12*(3), 127–138.

Kovalchik, S., & Ingram, M. (2016). Hot heads, cool heads, and tacticians: Measuring the mental game in tennis (ID: 1464).

Kovalchik, S., & Ingram, M. (2017). Estimating the duration of professional tennis matches with varying formats. *Mathsport International*, *1*(1), 217–228.

Leitner, C., Zeileis, A., & Hornik, K. (2010). Forecasting sports tournaments by ratings of (prob) abilities: A comparison for the euro 2008. *International Journal of Forecasting*, *26*(3), 471–481.

Newton, P. K., & Keller, J. B. (2005). Probability of winning at tennis I. Theory and data. *Studies in Applied Mathematics*, *114*(3), 241–269.

O'Malley, A. J. (2008). Probability formulas and statistical analysis in tennis. *Journal of Quantitative Analysis in Sports*, *4*(2), 1–21.

Shin, H. S. (1993). Measuring the incidence of insider trading in a market for state-contingent claims. *The Economic Journal*, *103*(420), 1141–1153.

Spanias, D., & Knottenbelt, W. J. (2012). Predicting the outcomes of tennis matches using a low-level point model. *IMA Journal of Management Mathematics*, *24*(3), 311–320.

Stefani, R. (2011). The methodology of officially recognized international sports rating systems. *Journal of Quantitative Analysis in Sports*, *7*(4), 122.

Stekler, H. O., Sendor, D., & Verlander, R. (2010). Issues in sports forecasting. *International Journal of Forecasting*, *26*(3), 606–621.

Yuan, L.-H., Liu, A., Yeh, A., Kaufman, A., Reece, A., Bull, P., et al. (2015). A mixture-of-modelers approach to forecasting NCAA tournament outcomes. *Journal of Quantitative Analysis in Sports*, *11*(1), 13–27.

Stephanie Kovalchik is the lead data scientist in the Game Insight Group at Tennis Australia and is currently a Research Fellow at the Institute of Sport Exercise and Active Living at Victoria University in Melbourne, Australia. Dr. Kovalchik earned her Ph.D. in Biostatistics from UCLA. Her research focuses on the use of statistical methods to understand performance, game strategy, and mentality in high-performance tennis.

Machar Reid is one of international tennis's preeminent voices in sport and coaching science. He is currently the Innovation Catalyst at Tennis Australia; having previously served as its inaugural High Performance Manager (2011–14) and Sport Science & Medicine Manager (2008–2010). Dr. Reid has a Ph.D. in Biomechanics and has published over 100 peer-reviewed articles and books/book chapters in the field.

Please cite this article in press as: Kovalchik, S., Reid, M., A calibration method with dynamic updates for within-match forecasting of wins in tennis. International Journal of Forecasting (2018), https://doi.org/10.1016/j.ijforecast.2017.11.008.