Evaluating strange forecasts: the curious case of football match scorelines

Article

Accepted Version

Evaluating Strange Forecasts: The Curious Case of Football Match Scorelines

J. James Reade
University of Reading

Carl Singleton
University of Reading

Alasdair Brown*
University of East Anglia

August 2020
Accepted for publication at Scottish Journal of Political Economy

Abstract

This study analyses point forecasts of exact scoreline outcomes for football matches in the English Premier League. These forecasts were made for distinct competitions and originally judged differently. We compare these with implied probability forecasts using bookmaker odds and a crowd of tipsters, as well as point and probability forecasts generated from a statistical model. From evaluating these sources and types of forecast, using various methods, we argue that regression encompassing is the most appropriate way to compare point and probability forecasts, and find that both these types of forecasts for football match scorelines generally add information to one another.

Keywords: Forecasting; Statistical modelling; Regression models; Prediction markets
JEL Codes: C53; G14; G17; L83

*Corresponding author: j.j.reade@reading.ac.uk, Department of Economics, University of Reading, Whiteknights Campus, RG6 6UA, UK.
c.a.singleton@reading.ac.uk, Department of Economics, University of Reading, UK. alasdair.brown@uea.ac.uk, School of Economics, University of East Anglia, UK.

We are grateful for advice and comments from Shixuan Wang, Jean-Louis Foulley, Michael Clements, Robert Simmons, and from seminar participants at the Universities of Reading and Plymouth, as well as at the 12th International Conference on Computational and Financial Econometrics, the 39th International Symposium of Forecasters and the 1st Reading Football Economics Workshop.

This study is based on data analysed with the permission of Superbru, Sport Engage Ltd. Throughout the study, the anonymity of users of the Superbru prediction game was maintained. The use of these data does not imply the endorsement of the data owners in relation to the interpretation or analysis of the data.

Singleton thanks the Economic and Social Research Council (UK) for support under grant ES/J500136/1

1 Introduction

Forecasts form a central part of everyday life; they are statements regarding the probability of particular states of nature occurring. In general, economic agents have preferences over different states of nature, which can have real consequences in money or other terms. As such, the evaluation of forecasts is important and in principle ought to relate to agents' preferences (e.g. Granger and Pesaran, 2000). But for many variables and contexts, the inability to observe or understand these preferences, plus the difficulty of constructing agents' loss functions based on what actually occurs, and allied with the generally (quasi-)continuous nature of macroeconomic variables, has led to more statistical measures of forecast evaluation in most circumstances (e.g. Fawcett et al., 2015).

In this study, we evaluate the forecasts of association football (soccer) match outcomes. Ultimately, after all the punditry is said and done, there are two important aspects to the outcome of a football match: the *result* and the *scoreline*. The result is a win for either team, or a draw (tie). The scoreline gives the exact number of goals scored by each team. A football scoreline is thus a pair of non-negative integers, which are correlated due to the common conditions faced by both teams in a match and the fact that teams and their tactics will respond within matches to the goals scored (or not) by their opponents (e.g. Heuer and Rubner, 2012). The different states of nature dictated by football match outcomes matter significantly; teams may progress in competitions, their fans may gain bragging rights, and bettors may make returns (or losses). While the result generally determines the state of nature (e.g. winning a round-robin or knock-out competition), the scoreline is sometimes the first tie-breaker after the result. League positions and championships, where the teams are tied on cumulative points totals from results, are usually determined by some function of scorelines (e.g. the difference between goals scored and conceded or head-to-head records between teams over multiple matches). Some cup competitions (e.g. the UEFA Champions League) have scoreline-related tie-breaker rules, such as 'away goals'.[1] Even more fundamentally, the result is a function of the scoreline.

The majority of attention in the academic literature on forecasting football match outcomes has focused on the result rather than the scoreline (e.g. Angelini and De Angelis, 2019; Forrest and Simmons, 2000; Forrest et al., 2005; Goddard, 2005). But scorelines also matter. Many forecasts are made regarding them, both formal and informal. Based on our observations and estimation from the world's largest sports betting exchange in 2019, *Betfair Exchange*, the exact scoreline in a football match is a popular outcome to predict and bet on: focusing on the state of markets at the beginning of important matches (i.e. high liquidity markets with £1million or more of matched bets, e.g. the English Premier League or competitive internationals), for every £1.00 of bets matched on the result outcome of the match, approximately £0.20 is matched on the exact scoreline markets. This compares with

[1] If two teams are equally matched after playing each other twice, home and away, i.e. the cumulative scoreline is a draw, then the team which has scored more goals away from home is the winner.

£0.70 bet on the total number of goals scored in a match, £0.25 on the Asian Handicap markets, and £0.20 on the margin (goal difference) between the two teams at the end of a match. Further, these other mentioned outcomes of football matches and popular prediction markets are all functions of the final scoreline. As there are only three possible outcomes for the result, and many times more potential outcomes for the scoreline, it follows that forecasting the scoreline is more difficult. Historically, the most likely *result outcome* from a football match is a home win (occurring roughly 48% of the time), while the most likely *scoreline outcome* is a 1-1 draw (occurring roughly 11% of the time).[2]

The scoreline is a *strange* variable. A general definition of *strangeness* is difficult, and may in fact vary by context. But we think that if one was written down, then it would contain parallels to some of the following aspects of a football match scoreline forecast:

Non-standard: it is non-continuous, made up of two non-negative integers, and generates a range of important sub-outcomes (e.g. the result).

Residual outcome: the tie is a third outcome between either team winning. Despite 1-1 being the most common outcome, it is a residual outcome.

Uncertainty: a large number of potential event outcomes ensures that the most likely has only about a 10-15% likelihood of occurring.

Fragility: the median number of goals is three, with a variance near to three, and over 10% of all goals are scored in the final five minutes of matches.

Salience: the scoreline determines the result of a football match, and attracts attention from the forecaster.

In this context, scorelines are *strange*, in that their salience generates utility from making precise picks of a *non-standard variable*, a variable whose outcome is highly *uncertain*, *fragile* and affected by a *residual outcome* that neither team has as their preferred outcome.

The *non-standard* nature of these strange variables mitigates against many standard methods of forecast evaluation. For example, a 1-1 draw may be forecast, but if the actual outcome is 2-2, then the *result forecast* is correct, but the *scoreline forecast* is incorrect. If the home team wins 2-1, then both the result and scoreline forecasts are wrong, yet the forecast is only one goal away from being correct, as opposed to two goals away if the match finishes 2-2. This highlights another *non-standard* aspect of football match outcomes: the *total number of goals* sub-outcome. A scoreline forecast implies a total number of goals scored in a match, as well as a *margin of difference* between the two teams competing. The 1-1 forecast has a margin of zero, and hence a 2-2 outcome similarly has a margin of zero, but a 2-1 outcome has a margin of one. These two variables, the total number of goals and the

[2]Author calculations using the entire history of football matches listed on *Soccerbase.com*, i.e. from 511,759 recorded matches up to 8 January, 2019.

margin of difference, take integer values, and they can be used to evaluate scoreline forecasts. Nonetheless, they each represent reductions in the information content of a scoreline, since the same number of total goals (greater than one) can yield all possible results. While the margin is consistent with only one result, a 1-0 or a 4-3 scoreline represent two very different match outcomes, as well as different experiences for spectators (a.k.a. consumers).

Despite this strangeness, or perhaps because of it, popular public competitions exist for forecasting football match scorelines. Historically, the football pools were the most popular of these competitions, in which the traditional and most popular game involved players picking a number of matches within a round of fixtures to end in a score draw for a substantial jackpot prize, which was rarely won (e.g., see Forrest, 2011; Forrest and Pérez, 2011). Although the football pools continue to this day, popular but similar alternatives have evolved from them. *Sky Sports*, a UK-based broadcasting company, runs the weekly *Super Six* competition for forecasts of professional matches in England, with significant cash prizes. Similarly, the sports predictor games website *Superbru* has over 1.5 million global users.[3] In addition, pundits in the media make well-publicised scoreline predictions. For example, former professional and international footballer Mark 'Lawro' Lawrenson has predicted the outcome of football matches for *BBC Sport*, a UK-based broadcaster and media outlet, for over a decade, typically competing against a celebrity.[4] The competition scoring rules used to judge these forecasts are in all cases a function of whether the scoreline and the result forecasts were correct, and potentially also include a measure of *closeness* to, or distance from, the correct outcome.

In all these competitions, the kind of forecast made is a point forecast: a pick of a particular scoreline. Within economic forecasting in recent decades there has been a trend towards probability forecasts, or density forecasts: attaching probabilities to different possible outcomes. Bookmakers essentially produce density forecasts by offering odds on a range of different scorelines. Well-established statistical methods for predicting scorelines generate probability forecasts, and such density forecasts allow a more forgiving evaluation. Rather than being judged as either exactly correct, or completely wrong, density forecasts are judged more on their distance, on average, from what actually occurred, and hence every forecast can be thought of as partly right.

Despite this, all the scoreline forecast competitions are for point forecasts, rather than densities. The *salience* aspect of the strangeness of scorelines arguably contributes to this; the utility gained from making a distinct pick outweighs that gained from attaching small probabilities to a range of scorelines. Nonetheless, the *uncertainty* and *fragility* aspects, with subsequent harsh forecast evaluation, may not be sufficient to offset the utility gain from providing a point forecast. This could explain why the evaluation metrics associated with

[3] As of 21st January 2019, the website claimed 1,532,572 users. See www.superbru.com/. The *English Premier League Predictor Game*, where users pick scorelines, has around 90,000 global players, with around one-sixth of these based in the United Kingdom.

[4] For example: www.bbc.co.uk/sport/football/

the point forecast competitions mentioned above generally have built-in ways to compensate forecasters for near misses, still rewarding them if the scoreline was incorrect yet their result was nonetheless correct, or if the margin of difference was right, or if the total number of goals scored was within one or two of what occurred. It may be that these kinds of mechanisms are sufficient to attract users to make point forecasts.

To demonstrate these issues, we compare a range of scoreline point forecasts for the *English Premier League* (EPL), made by forecasters who knew *ex ante* that they were being evaluated according to different rules within competitions. We also consider the probability forecasts implied by exact scoreline odds from a set of online bookmakers, as well as probability and point forecasts generated by a standard statistical model of goal arrival in football matches. We evaluate all these forecasts according to a range of metrics, attempting to illustrate their *strange* nature. We also find evidence that all the sources of forecasts studied appear to be inefficient in terms of predicting scoreline outcomes, including the probability forecasts implied by bookmaker odds and a crowd of tipsters. There is suggestive evidence that the competition scoring rule, which was originally used to judge a point forecaster's match outcome predictions, could have affected how much attention was given to correct scoreline picks relative to correct result picks. An analysis of forecast encompassing (e.g. Chong and Hendry, 1986; Fair and Shiller, 1989), we argue, provides the fairest way to compare the different sources of point and probability forecasts for the same set of events. In this way, we test whether the different sets of scoreline forecasts can on average predict one another's forecast errors. We find that the probability forecasts, such as from the statistical model or the bookmaker odds, tend to contain more relevant information than point forecasts. Nonetheless, some combination of probability and point forecasts is likely to be the most effective in predicting the outcomes of football matches. The statistical model that we create encompasses the forecasts implied by bookmakers.

The rest of the paper is organised as follows: Section 2 discusses the related literature; Section 3 introduces the data and documents their various sources; Section 4 sets out the methodology we employ; Section 5 presents our results; and Section 6 concludes.

2 Related literature

This study sits ostensibly in the forecasting literature; we seek to evaluate strange forecasts. In what follows, we briefly describe some of the most relevant literature on sports forecasting, in particular focusing on football match forecasting.

Forrest et al. (2005) studied bookmakers in the 1990s and 2000s, finding that they became more accurate at forecasting outcomes throughout this time, reflecting growing commercial pressure. Štrumbelj and Šikonja (2010) later updated this finding, but highlighted one aspect of the strangeness of football scorelines: the draw. They found that bookmaker odds provide little predictive information on the relative frequency of draws, and noted that Pope and Peel (1989) and Dixon and Pope (2004) found something similar in earlier studies. Štrumbelj

and Šikonja (2010) suggested that this reflected the *residual* nature of the draw outcome; it is the remaining probability mass after the home and away teams' strengths have been taken into account. Angelini and De Angelis (2019) studied the odds of online bookmakers on football matches throughout the major professional leagues in Europe between 2006 and 2017. Using a forecast-based approach, they tested whether these markets are generally efficient, finding that they are in most countries, even if the best odds on match outcomes are selected from among bookmakers. This result was supported by Elaad et al. (2020), who found that after accounting for heterogeneity among online bookmakers, prices set on result outcomes in the EPL were generally unbiased as forecasts. Dixon and Pope (2004) is one of the few papers in the literature that has considered football scoreline outcomes rather than just results, finding that the markets for exact scoreline predictions were generally inefficient in the 1990s.

Another literature has looked at so-called tipsters as forecasters. Forrest and Simmons (2000) evaluated the predictions of British newspaper tipsters, i.e. journalists providing forecasts of forthcoming football match outcomes, finding that they did better than random forecasting methods would have performed. However, these tipsters did not build into their forecasts readily available information, and mostly appeared to rely on information contained in one another's forecasts. The tipsters studied by Forrest and Simmons (2000) picked match result outcomes, rather than scorelines. Spann and Skiera (2009) also looked at newspaper tipsters, evaluating them against bookmakers and prediction markets. They found that the tipsters were outperformed by both. This finding was corroborated by Reade (2014), looking at the users of a betting odds comparison website, *Oddsportal.com*. This latter study generalised the description of a tipster from somebody providing tips in a newspaper into the realm of social media, as *Oddsportal.com* operates a network where users share their predictions with one another online. However, Brown and Reade (2019) noted that these particular tipster picks do in fact provide relevant information not contained within bookmaker prices, embodied within the crowd of tipsters on the website rather than in any one individual. In all these cases, tipsters were predominantly picking match result outcomes rather than exact scorelines.[5] While it is not known how individual tipsters construct forecasters, it is perhaps reasonable to view them as producing judgemental forecasts (see Lawrence et al., 2006).

Scorelines are salient, and hence attract considerable attention, as exemplified by the community picks from *Oddsportal.com*. Considering the forecasting performance of such groups of tipsters, much has been published on the 'wisdom of crowds' idea of Surowiecki (2004). Peeters (2018) considered whether crowd valuations of football players can help in predicting football results, and O'Leary (2017) found that crowd-based predictions were comparable with bookmaker odds at the 2014 FIFA World Cup. Simmons et al. (2010) emphasised the role of knowledge in the wisdom of crowds, and we usually think of those with

[5]The *Oddsportal.com* sample studied in these latter cases contained a huge range of different events, a tiny fraction of which may have been football match scorelines.

more knowledge than others to be experts. As an example outside of the sport forecasting context, Genre et al. (2013) evaluated combinations of expert forecasters from the European Central Bank's Survey of Professional Forecasters, finding that it was not obvious that individuals or combinations of some experts would have provided more accurate forecasts than using a simple weighted average over the whole set of experts in the survey.

There is a substantial literature studying behavioural biases implied by sports forecasts. Perhaps most famously and extensively studied is the favourite-longshot bias, whereby the probability forecasts implied by prediction market prices typically suggest that favourites, i.e. those most likely to win, are underbet. Rational explanations of this bias focus on the potential for relative risk-love among gamblers (see Ottaviani and Sørensen, 2008 for a summary). In football, Cain et al. (2000) showed that this bias appears in UK football and Angelini and De Angelis (2019) find that this bias is generally present throughout European betting market odds for match results. Football match scorelines are uncertain and fragile. On such low-probability events, Snowberg and Wolfers (2010) and Vaughan Williams et al. (2018) found evidence from horse-racing and online poker markets, respectively, that supports an explanation for the favourite-longshot bias based on agent misperceptions; bettors cannot distinguish between events with different low probabilities of occurring. If this were true, we might expect tipsters to perform particularly poorly when it comes to forecasting exact scorelines when compared with statistical models. Away from the favourite-longshot bias, Ayton et al. (2011) have found experimental evidence of the role that very simple heuristics can play in successfully forecasting match result outcomes in English football, when individuals pick the team to win which they recognise or relate to the most.

There are many studies which have also attempted to statistically model the outcome of football matches and which have subsequently evaluated the forecasting performance of such models. Maher (1982) analysed both independent and bivariate Poisson processes for goal arrival and hence football scorelines, while Dixon and Coles (1997) augmented that model for low-scoring games, a common feature of English football in the early 1990s, the period they were studying. Dixon and Coles (1997) focused on inefficiencies in betting markets as the main purpose of their modelling, looking at the outcomes of betting on home or away wins based on their model. Karlis and Ntzoufras (2003, 2005) also developed a bivariate Poisson model for modelling football scorelines. Goddard (2005) has investigated whether this class of model, which reflects the process of goal arrival within a football match, is more effective at forecasting match result outcomes when compared with methods which are more directly aimed at estimating these outcomes. Boshnakov et al. (2017) introduced a bivariate Weibull count model of goals to this topic, which they documented as improving upon the model of Karlis and Ntzoufras (2005). As with Dixon and Coles (1997), they evaluated their estimates against traditional statistical measures, but also used it to inform a potentially successful betting strategy, looking at both result outcomes and whether more than 2.5 goals were scored in a match. Similarly, Buraimo et al. (2013) have shown that using straightforward

betting strategies for football match results, whenever positive returns were expected based on the University of Warwick's 'Fink Tank' statistical model's probability forecasts, which were published in a British newspaper, would have generated positive expected returns for each season of the English Premier League between 2006/07 and 2011/12.

Beyond football, there is a vast literature that looks to model the outcomes of other sports, makes forecasts and subsequently evaluates them. For example, there have recently been notable contributions in the challenging area of forecasting the outcome of cricket matches in-play (see Akhtar and Scarf, 2012 for test match cricket and Asif and McHale, 2016, 2019 for international limited overs cricket). Another sport that has attracted substantial interest and advances in forecasting theory and evaluation is tennis (e.g. Klaassen and Magnus, 2003; del Corral and Prieto-Rodríguez, 2010; McHale and Morton, 2011).

We believe that to date there has been only one other study which has asked how to evaluate the strange case of football match scoreline forecasts, in spite of the fact so many people make these forecasts. Foulley and Celeux (2018) have suggested a method of penalising scoreline predictions based on a novel metric of distance between forecasts and outcomes, attempting to reduce the two-dimensional scoreline into a single measure. We will describe and apply this metric, alongside others, in what follows. However, we go further in this study by comparing a wide range of forecasting methods, as well as different sources and types of forecasts.

3 Data

We gather data from several sources. Our attention is focused on the EPL for a couple of reasons: it is widely regarded as the foremost domestic club competition globally;[6] practically, the EPL is the league for which the widest range of forecasts is available. Across all the sources we consider, these data cover forecasts for the 380 matches played in each of the 2016/17 and 2017/18 EPL seasons. We focus on forecasts during these two recent seasons as our general tipster data, which we describe later, was only available for this period. We extract the data on the outcomes of the football matches from *Soccerbase.com*. Table 1 presents the distribution of scorelines across the two seasons. The left panel is the 2016/17 season and the right panel is the 2017/18 season. There were 33 unique scorelines in 2016/17 and 32 in 2017/18, of which around two thirds involved each team scoring at most two goals. Within each panel, each row represents the number of goals scored by the home team, and each column gives scorelines where the away team scored a particular number of goals. Hence, the top left entry in each panel is a 0-0 draw, and 7.1% of games in 2016/17, and 8.4% in 2017/18 had 0-0 scorelines. There were slightly more draws in

[6]It is a derivative of the Football League, the first football league competition founded in 1888. The total club revenues for the EPL at £5.3bn are almost equal to the sum of the next two leagues combined, Spain's La Liga (£2.9bn) and Germany's Bundesliga (£2.8bn) (see 2018 Deloitte Annual Review of Football Finance; www2.deloitte.com/uk/.

2017/18 than 2016/17, and fewer goals, but these differences between seasons are generally not statistically significant.

The right panel of Table 2 displays the distribution of results in 2016/17 and 2017/18, showing that there were more home wins in 2016/17, and fewer draws, though again differences were generally insignificant between seasons. As home wins happen almost half the time, this provides a naïve forecasting method. In fact, Forrest and Simmons (2000) document that newspaper tipsters tended to have a lower success rate than such a naïve forecasting method as this.

3.1 Bookmaker odds

While bookmakers exist to profit maximise, it nonetheless remains the case that to do this they must forecast future events sufficiently well. We consider the decimal odds d set by a bookmaker. Decimal odds are inclusive of the stake (the money amount bet), such that if the potential event outcome being bet on occurs, the bettor is paid dz, where z is the stake. If it does not occur, then the bettor loses their stake z. The implied outcome probability is $p = 1/d$.[7] In reality, there is an overround included in bookmaker prices; if the implied probabilities for all events in the event space are summed, they will add to more than one. Various methods have been suggested to correct for the overround when interpreting odds as implied probabilities (see the summary by Štrumbelj, 2014). For the analysis which follows, we use the most simple of these corrections, by dividing the quoted odds on each outcome through by the booksum, which is the sum of the odds offered for the various possible outcomes on some event (e.g. over all possible scorelines).[8]

We obtain bookmaker odds for all EPL match outcomes listed on *Oddsportal.com*. Within this we have information on 51 individual bookmakers, and also a betting exchange, *Matchbook*. Calculating implied probabilities from the posted decimal odds enables a comparison of the average probability of outcomes according to bookmakers, and this is presented in the left panel of Table 2 for results, without adjusting for the overround. Bookmakers were consistent over the two seasons we study, predicting home teams to win 46% of the time, away teams to win 32% of the time, and a draw to occur 25% of the time (implying an overround of about 3%). In the right panel of Table 2 we present the actual frequencies, suggesting that bookmakers over-estimated the likelihood of an away win. Online Appendix Table B1 presents the implied probability, or frequency, from the average bookmaker odds for each scoreline in each season.[9] The scoreline odds-implied probabilities indicate a sizeable overround of about 12%, with the majority of implied probabilities being

[7] Decimal odds relate to fractional odds, f, which is how bets are traditionally priced in some areas, by $d = f + 1$.

[8] The implied probability of match outcome i from the bookmaker odds is then given by: $p_i = (1/d_i)/\sum_i(1/d_i)$.

[9] In 2016/17, bookmakers offered odds on scorelines of 7-4, 7-5, 7-6 and 6-7 for the Premier League, but in 2017/18 such odds were not offered. In the entire history of the English Football League, of more than 220,000 matches, there have been 21 7-4 scorelines, 5 7-5 scorelines, and no 7-6 or 6-7 scorelines.

higher than the actual proportions from Table 1. Variation between the seasons is smaller in these implied probabilities than in the actual proportions of scoreline outcomes.

3.2 Tipsters

We were given data for 50 anonymous users, or tipsters, from the online *Superbru Premier League Predictor Game*, from the 2016/17 season, and 150 from the 2017/18 season. These data samples were selected in different ways by *Superbru* administrators from the populations of users, without any particular instruction from us in this regard.[10] The 50 tipsters from 2016/17 were sampled randomly from all game users. The 150 tipsters from 2017/18 were randomly sampled from users who 'completed' the game, i.e. they forecast the outcome of every match that season. On the representativeness of these samples, first, we would speculate that the typical user of Superbru is a keener and more knowledgeable fan of EPL football than the typical person, or even the typical football fan, given that they self-selected into playing the game. Second, users who completed the game are likely to be particularly devoted to following the EPL, so we could speculate that they have better than normal knowledge and expertise about the events being forecast. Third, to persist with the game, these players probably attain greater than normal utility from making forecasts and from getting them correct.

Players of the *Superbru* game are offered financial incentives to make correct forecasts; five or six correct scoreline picks in a round of ten matches wins an item of clothing, while seven or more correct scoreline picks earns a cash prize, up to £50,000 for picking all ten scorelines correctly in a round of EPL fixtures. In our dataset, which amounts to 7,526 tipster-round observations, the most correct scoreline picks in a round of fixtures is five, which happens on eight occasions. With the existence of mini-leagues between players, there are also non-financial incentives for tipsters, as well as an overall game leaderboard.

Online Appendix Table B2 provides the distribution of scoreline picks by the tipsters over the two seasons. Tipsters rarely predicted goalless draws, and predicted the vast majority of matches to involve each team scoring at most two goals, reflecting the empirical regularity displayed in Table 1. In fact, tipsters predicted about 75% of matches to lie within this range of outcomes, when in reality only about 66% of matches finished this way.[11]

3.3 'Experts': Lawrenson and Merson picks

BBC Sport publishes forecasts by Mark Lawrenson, a former professional and international footballer, around 24 hours before each round of matches in the EPL. These are typically published on a Friday for weekend fixtures, in advance of matches through Saturday

[10] Across all the games on the *Superbru* platform there are over 1.5 million registered users, who each play one or more of the games. As of September 2018, there were 89,000 players of *Premier League Predictor Game*. Despite focusing on English football, the players are global, with just over a sixth based in the United Kingdom.

[11] See Singleton et al. (2019) and Butler et al. (2020) for further analysis of these tipsters and their forecasting behaviour.

lunchtime to Monday evening. Similarly, *Sky Sports* publishes forecasts before fixtures by Paul Merson, another former professional and international footballer. These two 'experts' have been making forecasts in this way since at least 2003/04 and 2014/15, respectively.[12] In both cases, the forecasts are published on the websites of the respective broadcasters and on social media, in advance of being further publicised on television within a couple of days of the weekend's matches beginning. Both the forecasts of Merson and Lawresnson serve an entertainment purpose for their respective broadcasters. However, we know from speaking with Lawrenson that he at least takes the role seriously, and does produce the forecasts himself, rather than a BBC researcher doing so, for example.[13]

In the conventional literature, forecasts produced by individuals considered to have significant knowledge and experience of the events being forecast are referred to as *expert forecasts*. We treat the forecasts of Lawrenson at *BBC Sport* and Merson at *Sky Sports* as expert forecasts, since both individuals are hired by their respective broadcasters as pundits, and both are former professional footballers and team managers in the English Football League. Forecasts created by experts are by their nature not replicable; there is no clearly defined process by which these forecasts are created that could be applied in other situations.

Online Appendix Table B3 summarises the scoreline forecasts of Lawrenson and Merson, in the same format as Table 1. The two experts had a narrower range of scoreline predictions than the tipsters and actual outcomes, particularly in the 2016/17 season, when Lawrenson never picked a team to score more than three goals and Merson did so just five times. Both experts picked substantially more 2-0 scorelines for the home team than actually occurred, and Lawrenson heavily favours 1-1 draws.

4 Methodology

In addition to those described above, we generate a set of probability forecasts with a statistical model. Using these, we can then apply rules to create point forecasts, or picks, comparable with the tipsters and experts. The type of model we select for this purpose is well-known and could be considered the 'standard' statistical model for football match scorelines (e.g. Goddard, 2005). We briefly describe this model in Section 4.1. In Section 4.2, we discuss the various evaluation methods we will employ.

[12] Both sets of forecasts for recent seasons have been collected and made available in the online archive of quantitative football-themed blog *EightyFivePoints.com*.

[13] In April 2019, the present authors, under the guise of the University of Reading's 'supercomputer' called RED, competed against Lawrenson in picking a round of EPL scorelines; https://www.bbc.co.uk/sport/football/47886436. During filming for a TV slot, we asked Lawrenson how he goes about making predictions.

4.1 Generating scoreline forecasts from a statistical model

To generate forecasts, we first estimate the goal arrival process in football matches using a bivariate Poisson regression model, of the form proposed (and coded) by Karlis and Ntzoufras (2003, 2005). That is, the goals scored by each team in a football match are modelled as jointly Poisson distributed. The counts of goals scored in match i for the home and visiting teams can be thought of as functions of their own strengths X_{i1} and X_{i2}, respectively, and some third common factor X_{i3}, representing the match conditions (e.g. weather, time of the year). If the goals of the home team in match i are denoted by h_i, and those of the visiting team by a_i, then we can define three Poisson distributed random variables X_{i1}, X_{i2}, X_{i3}, such that $h_i = X_{i1} + X_{i3}$ and $a_i = X_{i2} + X_{i3}$, and we say that these are jointly distributed according to a bivariate Poisson distributed, with $BP(\lambda_{i1}, \lambda_{i2}, \lambda_{i3})$. The regression model is written as:

$$\begin{aligned} (h_i, a_i) &\sim BP(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}) \ , \\ \log(\lambda_{ik}) &= \mathbf{w}'_{ik}\boldsymbol{\beta}_k \ , \qquad k = 1, 2, 3 \ , \end{aligned} \tag{1}$$

where $\mathbf{w}_{ik}$ is a vector of explanatory variables and $\boldsymbol{\beta}_k$ is a vector of coefficients. Team fixed effects are added into the model for each k to allow for teams having particular goal scoring or defensive strengths irrespective of their opposition. The explanatory variables also include day of the week and month dummies for the modelling of λ_{i3}, to reflect the fact that midweek matches may have different properties to weekend ones, and matches in the middle of winter may be different to those in the autumn or spring. We also add an indicator for whether a match follows a break in the season for international matches. We include information in the model about the league positions and recent form of each team, following Goddard (2005), as well as our calculations of each team's measured Elo (1978) strengths as they varied throughout the season, based on the historical results for all relevant teams, including those not playing in the EPL in the period studied.[14] We add a variable for whether a team is still in the main domestic cup competition, the FA Cup, as Goddard (2005) found this to matter for goal arrival in league matches, and others have found this to matter for league attendance, and attendance to matter for home advantage. We also add variables for whether a team can still achieve a top-two position in the league, and a variable for whether a team is returning to domestic action having played in European competition in their previous match, since this may affect squad rotation and player tiredness.

The statistical model is estimated by maximum likelihood up to each round of matches in each season, using the past calendar year of matches, and the estimated parameters are subsequently used to make predictions. Values of λ_{ik} are estimated for the upcoming round of out-of-sample matches, and used to generate probabilities for a range of scorelines. Combinations of the λs give predictions of the mean (or expected) number of goals scored

[14] This is an increasingly common method used in both practical football applications (see, for example, https://www.eloratings.net/), but also in academic research (e.g. Hvattum and Arntzen, 2010).

within matches by teams. A scoreline point forecast comparable to that provided by the tipsters and experts can then be generated.

We generate scoreline point forecasts in three ways. First, we simply use whatever the statistical model outputs as the most likely scoreline as the pick, which we call *Unconditional* forecasts. Second, we condition the scoreline pick on the most likely forecast result outcome. In this case, if all the probabilities of home win scorelines sum to a larger number than all the probabilities of draw or away win scorelines, then we would choose the most likely home win scoreline as the pick. We call these *Conditional* forecasts; i.e. conditional on the most likely result outcome, what is the most likely scoreline? This tends to generate differences, as empirically the most common scoreline is a 1-1 draw, but the most likely result outcome is a home win. Third, because of the large number of possible scorelines, many more are for wins than are for draws. As such, it is infrequent that a draw is the most likely forecast outcome. To address this, we develop *Fuzzy Conditional* forecasts. These return a draw prediction if the three result outcome probabilities are sufficiently close to one another. For example, if in some match the statistical model outputs the probability of a home win as 35%, the probability of an away win as 33%, and the probability of a draw as 32%, then this relatively even match according to the model estimates would have a draw Fuzzy Conditional scoreline pick, rather than the prediction of a home win, which our Conditional forecast would return. To determine whether a draw pick should be returned in this way, we use the entropy measure of Shannon (1948), which is a measure of the 'decidedness' of a market. If the three forecast result outcome probabilities are at a third each, then the entropy measure is maximised, while if one of the three outcome probabilities is exactly one, then the entropy measure is minimised. If the entropy measure is above 1.09, then we return a draw prediction.[15] The choice of 1.09 is naturally arbitrary; it was chosen such that if the probability of a home and away win became arbitrarily close, then a draw was the outcome produced as the Fuzzy Conditional scoreline pick.

4.2 Forecast evaluation and comparison

The difficulty of forecasting football scorelines task is emphasised by considering the variation in goals scored by teams over matches. In our sample of 760 matches, the mean number of goals scored per game is 2.73 and the variance is 2.78. Conditional on a home win, the variance of home goals is 1.5, and the variance of total goals is 2.7, while conditional on an away win occurring, the variance of away goals is 1.3, and the variance of total goals is 2.3. A

Any match has a number of outcomes and sub-outcomes that can matter in terms of how scoreline forecasts are evaluated:

[15] The example described above has an entropy score of 1.098.

The scoreline: the actual goals scored by each side. The scoreline is a pair of numbers, $\mathbf{s}_i = (h_i, a_i)$, where the number of goals scored by the home team is always listed first. We denote the actual scoreline by $\mathbf{s}_i$ and any forecast of it by $\widehat{\mathbf{s}}_i$.

The result: whether either team wins, or the game is a draw. We denote the result of some match i as r_i. The result can be defined as a single variable taking three values, one each for a home win, an away win, and a draw. For example, we could define the following values:

$$r_i = r(\mathbf{s}_i) = \begin{cases} 0 & \text{if } h_i < a_i \\ 0.5 & \text{if } h_i = a_i \\ 1 & \text{if } h_i > a_i \ . \end{cases} \tag{2}$$

Note that the result r_i is a function of the scoreline, so $r_i = r(\mathbf{s}_i)$.

Closeness: a way of giving credit for 'close' picks. However, there is no unique metric for closeness. In the *Superbru EPL Predictor Game*, a tipster gets one point for a correct result, three points for a correct scoreline, and 1.5 points for a 'close' scoreline. Similarly, Foulley and Celeux (2018) suggest a method to penalise scoreline forecasts based on the distance from the correct outcomes. Both these closeness metrics are described in Online Appendix A. Closeness is a function of the forecast scoreline and the actual scoreline, hence $c_i = c(\mathbf{s}_i, \widehat{\mathbf{s}}_i)$. It can be the function of two further sub-outcomes, the *margin* and the *total goals scored*, which taken together define a match scoreline:

Margin: the difference between the goals scored by two teams in match i;
$m_i = m(\mathbf{s}_i) = h_i - a_i$.

Total goals scored: the total number of goals scored by both teams in match i;
$t_i = t(\mathbf{s}_i) = h_i + a_i$.

4.2.1 Scoring rules

Generally, a scoring rule x of match i can be written as a function of the outcome scoreline $\mathbf{s}_i$ and the forecast scoreline pick $\widehat{\mathbf{s}}_i$, usually linearly:

$$score_{ix} = f_{ix}(\mathbf{s}, \widehat{\mathbf{s}}; A, B, C) = A_x \mathbb{1}\{r(\mathbf{s}_i) = r(\widehat{\mathbf{s}}_i)\} + B_x \mathbb{1}\{\mathbf{s}_i = \widehat{\mathbf{s}}_i\} + C_x c(\mathbf{s}_i, \widehat{\mathbf{s}}_i) \ , \tag{3}$$

where $\{A, B, C\}_x$ are parameters determining the weight given to picking correct and close outcomes. There are many more possible scorelines than there are possible results, which makes picking the scoreline correctly more challenging. As such, any reasonable scoring rule would have $A_x < B_x$ to encourage effort at the more challenging task.[16]

[16] Note that a penalty mechanism, like that of Foulley and Celeux (2018), would have $\{A_x, B_x, C_x\} < 0$.

On the *BBC Sport* website, Lawrenson and his celebrity guest competitor each week get 10 points for a correct result forecast and 40 points for a correct scoreline. The forecasts of Merson are released on *Sky Sports*, and are closely associated to a competition: the *Super Six* requires entrants to pick six scorelines within a given round of fixtures. This competition is run by *Sky Sports* and sponsored by the associated company *Sky Bet*. The betting odds for each of Merson's scoreline picks are provided alongside on the *Sky Sports* website. As such, it seems reasonable to associate the *Super Six* scoring rules to Merson's forecasts. In this competition, a correct result forecast gets 2 points and a correct scoreline gets 5 points. Therefore, the *BBC Sport* scoring rule more strongly rewards a correct scoreline pick relative to only the correct result ($40/10 = 4$), and the *Sky Sports* rule less so ($5/2 = 2.5$). The *Superbru* scoring rule is complicated slightly by the additional return from the closeness metric, but the reward from forecasting a correct scoreline relative to only a correct result is ($3/1 = 3$), and hence is roughly in between the other two rules.

Alternative scoring rules might evaluate solely based on results ($B_x = C_x = 0$) or scorelines ($A_x = C_x = 0$). Scoring rules could be further augmented to reward particularly 'good calls', by using some measure of the uncertainty associated with a given outcome. That is, the scoring rule could reward more generously a forecaster that is able to pick not just the more likely scorelines or results, but also the less likely ones. We can think of this as making 'better', or more 'bold' picks. In this case, A_x and B_x can be allowed to vary with i, for example according to *ex ante* bookmaker prices or *ex post* tipster crowd forecast performance measures for each match.

4.2.2 Return on investment

Evaluating scoreline forecasts according to betting prices is arguably the most natural evaluation method, since it reflects the potential payoffs from making decisions based on those forecasts. Therefore, we add to our scoring rules for evaluating scoreline picks with the returns from betting on the results, scorelines, total goals and the winning margin consistent with those picks, otherwise referred to as a return on investment (ROI). If d_i are the decimal odds in match i for the scoreline consistent with the forecast $\widehat{s}_i$, then the ROI from a one unit bet on that event outcome would be:

$$ROI_i = d_i \mathbb{1}\{\mathbf{s}_i = \widehat{\mathbf{s}}_i\} - 1 \ . \tag{4}$$

Throughout our analysis, in the case of scorelines, we use the mean of the bookmaker odds collected, and in the case of results, we take the best available bookmaker odds, all as posted right before matches began.

An alternative to the scoring rule in Equation (3) for a scoreline point forecast would be to place bets together on results (r_i), scorelines (s_i), total goals (t_i), and winning margins (m_i). Such a strategy may allow a financial compensation scheme to mimic those of scoring rules, allowing returns to still be made if the exact scoreline is not achieved. A one unit bet

on match i might be placed to maximise in expectation:

$$GROI_i = z_{i1}d_{i1}\mathbb{1}\{r_i = r(\widehat{\mathbf{s}}_i)\} + z_{i2}d_{i2}\mathbb{1}\{\mathbf{s}_i = \widehat{\mathbf{s}}_i\} + z_{i3}d_{i3}\mathbb{1}\{t_i = t(\widehat{\mathbf{s}}_i)\} + z_{i4}d_{i4}\mathbb{1}\{m_i = m(\widehat{\mathbf{s}}_i)\} - 1 \ , \tag{5}$$

where $\sum_1^4 z_{ij} = 1$. The stakes bet on each outcome type within a match may differ, with z_{ij} acting as weights. In principle, some set of optimal weights or stakes exists which maximises expected returns, given the available odds and beliefs about the outcomes, or which replicates a particular scoring rule of the form described by Equation (3).

4.2.3 Brier score

A more traditional (statistical) scoring rule for forecasts, particularly probability ones, is the Brier (1950) score, based on the mean squared forecast error (MSFE) for a generic forecast $\widehat{y}_i$ of event y_i, for some set of N events:

$$MSFE = \sum_{i=1}^{N}(y_i - \widehat{y}_i)^2/N \ . \tag{6}$$

In our case, $y_i \in \{\mathbf{s}_i, r_i, m_i, t_i\}$. Evaluation using Brier scores like (6) is on the basis of probability forecasts, such as those produced by our statistical model, or by bookmakers. Neither Lawrenson, Merson nor the individual *Superbru* game players produce probability forecasts; they pick a scoreline for each match. If we consider the unit of observation to be match scorelines, then the range of scorelines we consider will affect the number of observations being summed over and may affect conclusions. Such an approach also assumes that Lawrenson and Merson place a probability of one on each scoreline they pick and zero on all other scorelines, which is surely unfair. While picks by both experts go back a number of seasons before the period we study, it is nonetheless unclear how a probability distribution could be formed based on their past predictions.

4.2.4 Regression-based methods and forecast encompassing

We apply forecast tests in the form proposed by Mincer and Zarnowitz (1969), by regressing outcomes on forecast probabilities. In the case of scorelines, the outcome variable is binary and the observational unit is the match scoreline, while in the case of results, the outcome variable is categorical. If we denote $\widehat{y}_{ij}$ as our probability forecast of match i for event outcome j and y_{ij} as the relevant specific outcome (e.g. a scoreline), then the regression model is:

$$y_{ij} = \alpha + \beta\widehat{y}_{ij} + \epsilon_{ij} \ , \tag{7}$$

where α and β are the intercept and slope coefficients, respectively, and ϵ_{ij} is the the error term. The weak efficiency of a forecast depends on the restriction $\alpha = 1 - \beta = 0$ holding. A stronger test of efficiency includes other information available at the forecast origin, and

can be tested using the regression model:

$$y_{ij} = \alpha + \beta \widehat{y}_{ij} + \mathbf{z}_i' \boldsymbol{\gamma} + \nu_{ij} \ , \tag{8}$$

where $\mathbf{z}_i$ is a vector of potentially other important variables for explaining the outcome y_{ij} and ν_{ij} is the error term. Strong efficiency further requires that $\boldsymbol{\gamma} = \mathbf{0}$ holds in addition. If $\boldsymbol{\gamma} \neq \mathbf{0}$, then other known information at the forecast origin is relevant and the forecast is not efficient.

Taking expectations of (7) yields that for unbiasedness we require $E(\widehat{y}_{ij}) = \alpha/(1-\beta)$. To test for this, we could estimate the regression:

$$\widehat{e}_{ij} = \theta + \nu_{ij} \ , \tag{9}$$

where $\widehat{e}_{ij} = y_{ij} - \widehat{y}_{ij}$ is the forecast error and ν_{ij} is the error term, and then test the null hypothesis that $\theta = 0$. Strictly speaking, in addition to the hypothesised restrictions holding, we require that the residuals from each regression estimation are approximately normally distributed, and free from any autocorrelation or heteroskedasticity. In their application, Forrest and Simmons (2000) add a range of variables that are public information into $\mathbf{z}_i$, including the recent results of each team and league-standing-related information. We do similarly by using our derived dynamic Elo (1978) ratings of teams.[17]

Other forecasts could be added to the regression analysis. In doing so, we could test whether any of the various forecasts are *encompassing*. A forecast a is said to encompass forecast b if it can explain variation in the forecast errors from forecast b, and forecast b cannot explain any of the variation in the forecast errors from forecast a:

$$\widehat{e}_{ija} = \theta_a + \phi_a \widehat{y}_{ijb} + \nu_{ija} \ , \tag{10}$$

$$\widehat{e}_{ijb} = \theta_b + \phi_b \widehat{y}_{ija} + \nu_{ijb} \ , \tag{11}$$

and $\mathbf{H_0}$: $\phi_a = 0, \phi_b \neq 0$, i.e. can one forecast explain what another forecast cannot? Chong and Hendry (1986) and Fair and Shiller (1989) both consider the possibility of encompassing in this manner. If $\phi_a \neq 0$ *and* $\phi_b \neq 0$, then a linear combination of such forecasts would be more effective than taking any single forecast in isolation. For example, focusing on the case of the bookmaker implied probabilities, we can test whether our generated statistical model probabilities, or the forecasts from the experts and tipsters, add any information when trying to determine the accurate probability of a future event taking place. In such an analysis, the implied probabilities from bookmakers and the predicted probabilities from the statistical model would be real numbers on the unit interval, whereas picks by the experts would be binary variables, taking one if that particular outcome for match i is

[17] We prefer Elo ratings as they are not only time varying within a football season, capturing the natural and systematic variation in the strengths and form of teams, but also more agnostic about how to compare teams than alternative metrics, e.g., based on the squad payrolls of teams or player valuations from transfermarkt.de - Elo ratings are only a function of who beats whom and when.

picked and zero otherwise. We include the *Superbru* user forecasts in this analysis by taking the proportion of tipsters picking a particular outcome as $\widehat{y}_{ijb}$, i.e. treating the tipsters as a crowd.

4.2.5 Scoreline forecast evaluation in summary

We consider a range of methods for evaluating football match forecasts, ranging from common statistical methods, through to more sophisticated statistical measures tailored to scoreline pick evaluation, as well as returns from betting strategies. The important distinction is between point forecasts, or picks, and probability, or density, forecasts. This affects the appropriateness of a particular forecast evaluation metric, and arguably motivates the construction of scoring rules to evaluate pick forecasts. It may then be that schemes which compensate for 'close' picks influence forecasting behaviour, diverting attention away from picking the correct scoreline toward lesser and easier outcomes, such as the result or the total number of goals scored.

5 Results

5.1 Evaluation by scoring rule and return on investment

Considering the range of evaluation metrics discussed in Section 4.2, namely Brier scores, the scoring rules employed by *BBC sport*, *Sky Sports*, *Superbru* and Foulley and Celeux (2018), as well as betting returns, we present the three types of point forecast generated from our statistical model: Unconditional, Conditional forecasts and Fuzzy Conditional. We consider these against each tipster and the two experts, Lawrenson and Merson, for both the 2016/17 and 2017/18 EPL seasons.

Tables 3 presents the output from applying the various forecast evaluation metrics described above to each EPL season, with twelve different metrics displayed. The top set of rows presents summary results for the sample of 50 Superbru tipsters in 2016/17, with the statistics for the other five sources of forecasts beneath (two experts and three model generated). The panel under this presents the equivalent set of results for 2017/18. We prefer to look at the two seasons separately because of how the samples of Superbru tipsters were generated: random sampling from all game players in 2016/17 and random sampling from the restricted set who made forecasts for all 380 matches in 2017/18; these tipsters are different individuals.

Between the two seasons there is variation in the frequencies of scorelines that occur (see Table 1), and almost certainly in the relative predictability of the individual matches. The experts and tipsters, by virtue of their methods not being known or replicable, do not in themselves provide any information on the relative predictability of the two seasons. The model generated picks must do, since the method is identical in both seasons. We consider the Unconditional model picks as the best measure of this, since this is the simplest set of forecasts based on the most likely scorelines, as estimated in advance of each match.

Column (1) of Table 3 demonstrates the inapplicability of Brier scores here. For the Superbru tipsters, this metric is essentially the proportion of picks that are wrong out of all the games they pick scorelines for. On the contrary, for the experts, because we consider their forecasts alongside the probability forecasts from the model and implied by bookmaker odds, it gives the proportion of all of the possible scorelines considered for all matches. There is no variation between the different forecasters because in the latter case the number of incorrect picks dwarfs the number of correct picks.

Column (2) presents the percentage of correct scoreline forecasts from each source; each season there were 380 matches. With the exception of Merson in 2017/18, the experts and model generated forecasts picked more scorelines correctly than the average Superbru tipster, and in 2016/17 Lawrenson and two of the three model-derived forecasts picked almost as many or more than the *best* tipster in our sample. The Unconditional model had 43 correct scorelines in 2016/17 and 40 in 2017/18, suggesting that the 2017/18 season was less predictable in terms of scorelines than 2016/17. Merson made 11 fewer correct scoreline picks in 2017/18 compared with 2016/17, yet Lawrenson made four more. The Superbru tipsters performed similarly well as a crowd by all measures across the two seasons, though with less variance in 2016/17, which may in part reflect that tipsters in 2017/18 picked every outcome and all 'completed' the season.[18]

Column (3) gives the percentage of correct result picks. With the exception of the unconditional model picks, all forecasters picked more results than the average Superbru tipster, although none of the model generated or expert forecasters picked as many correct results as the best tipster in both seasons. The two types of Conditional model-based forecasts picked many more results correctly than the Unconditional forecasts, which is to be expected since the former factors into the forecast the most likely result outcome from the statistical model.

Column (4) looks at the percentages of close picks by each forecaster, as measured by the Superbru metric (see Appendix A). With the exception of Merson in 2017/18, all the expert and Model-derived forecasts yielded more close picks than the average tipster, although none achieved more than the best tipster. In both seasons Lawrenson got more close picks than Merson, and the Unconditional model generated more close picks than either conditional model.

Columns (5)-(8) relate explicitly to scoring rules: first, the *BBC Sport* rule, ($A_{BBC} = 1$, $B_{BBC} = 3$ and $C_{BBC} = 0$), second the *Sky Sports* rule ($A_{SS} = 1$, $B_{SS} = 1.5$ and $C_{SS} = 0$), third the *Superbru* rule ($A_{bru} = 1$, $B_{bru} = 2$ and $C_{bru} = 0.5$), and finally the Foulley and Celeux (2018) penalty score measure (see Appendix A). Considering the *BBC Sport* rule, again the expert and model picks were generally above the average tipster, and in 2016/17 Lawrenson and the Conditional model were better than the best tipster. With

[18]There were 1,151 matches for which tipsters did not make a pick in 2016/17. On average, each sampled user picked 356 games in 2016/17 but all 380 in 2017/18.

the *Sky Sports* rule, essentially identical patterns are observed as with the *BBC Sport* rule. With the *Superbru* rule, similar patterns are observed again, although the ordering of the expert and model-generated picks is different; Lawrenson performed best in both seasons. The distinction between the *Superbru* and the broadcaster rules is that the former factors in closeness. Similar to *Superbru*, the Foulley and Celeux (2018) penalty rule rewards closeness, although in a more continuous manner. The model-generated picks all incurred a larger cumulative penalty than the average tipster in 2016/17, although not in 2017/18.

The final four columns of Table 3 consider returns on investment from systematically betting the same amount on the outcomes implied by the forecasts for every match. In other words, these returns are derived by assuming that the forecaster used their scoreline point forecast, for each of the 380 matches in a season, to place a £x bet on each of the correct result, correct scoreline, the margin being equal or greater than that implied by the predicted scoreline, and the total goals being equal or greater than that implied by the predicted scoreline. In general, betting on results, according to these forecasters, can generate positive returns (assuming that the bettor makes use of the best available odds from the range of bookmakers available in the UK). It would have done so for the average tipster, and all expert and model-generated forecasts apart from the Unconditional picks in 2016/17 and the Conditional picks in 2017/18. Betting on scorelines according to almost every source of forecasts here would have generated a substantial negative return, and the same is true for betting on the total goals scored in matches. Betting on the margin of difference between teams in 2016/17 would have been more successful than in 2017/18 based on the forecasts.

When considering these scoring rules more broadly, the relative rankings of the different forecasts are particularly informative, presented for each season in Online Appendix Table B4. The individual tipster rankings are implicit in these tables. Presumably, the *BBC Sport* rule is what Lawrenson forecasts to, the *Sky Sports* rule is what Merson forecasts to, and the *Superbru* rule is what the tipsters play to. The former rule most heavily rewards scoreline picks and the latter rule rewards close picks, as does the Foulley and Celeux (2018) penalty rule. Of the experts and the model-generated picks, Lawrenson was second best in 2016/17 and was best in 2017/18 at the *BBC Sport* rule. In both seasons he was ranked better according to the *BBC Sport* rule than the *Sky Sports* rule. Merson was ranked better according to the *Sky Sports* rule than the *BBC Sport* rule in both seasons, although in neither season was he ahead of Lawrenson on either rule. One conclusion consistent with these findings is that the scoring rules may influence forecasting behaviour, but that Lawrenson is a superior forecaster to Merson. Focusing on the model-generated forecasts, the Conditional forecasts always ranked better than the Unconditional picks in both seasons according to the *BBC Sport*, *Sky Sports*, and *Superbru* rules, but not according to the penalty rule.

5.2 Forecast efficiency

In this section, we describe the results of Mincer and Zarnowitz (1969) regression tests to evaluate the various candidate forecast methods for scorelines. We pool the two seasons, so the number of matches studied in each of these regressions is 760. When we refer to "Model" forecasts, we are evaluating the probability forecasts produced using the bivariate Poisson model set out in Section 4.1. By "Bookmaker" forecasts we are referring to the implied probabilities of outcomes derived from odds, as described before. Finally, by "Tipster" forecasts we refer to the crowd of forecasts generated by the samples of *Superbru* users and the probabilities of outcomes which these imply.

Table 4 presents the outcomes from regressions evaluating the strong efficiency of scoreline forecasts as per Equation (8), with a column for each forecast type.[19] Variables are added to $\mathbf{z}_i$ for the number of league points the home team has, the difference between the home and away team league points, the form of the home team, measured by the number of league points gained in the their last six matches, and the difference in form between the two teams. We also add an Elo prediction for the match outcome and a variable representing the historical frequency of each scoreline. Across all forecast methods, these extra variables are insignificant, i.e. $\boldsymbol{\gamma}$ in Equation (8) is insignificant from $\mathbf{0}$. This is not unexpected. While these team-specific variables must matter for result outcomes, given the sheer number of possible scoreline outcomes they simply are not important. It might be anticipated that the historical frequency of each scoreline would be significant, but our findings suggest that this is factored into each forecast. The bottom row of Table 4 reports an F-test of strong efficiency, which here is the null hypothesis that $\alpha = 0$, $\beta = 1$, and $\boldsymbol{\gamma} = \mathbf{0}$. The null hypothesis is heavily rejected in each case at standard levels of significance. In other words, the forecasts are inefficient. The $\widehat{\beta}$ slope coefficients on the Model and Bookmaker forecasts are closest to one, and the coefficients on the two experts are smallest, implying that if an expert makes a pick, that scoreline is about 10 percentage points more likely to occur than otherwise. As already mentioned, rather than indicating that the Model or Bookmaker forecasts are any more efficient, this merely reflects that this particular test is not a fair or appropriate comparison between the two, since it implies that the experts placed zero weight on every possible scoreline other than the one they picked. It is also worth noting that the β coefficient on the Bookmaker regression is significantly greater than one at standard levels, which is indicative of the well-known favourite-longshot bias. Hence we can document the existence of this bias among football match scorelines odds, whereas it has typically only been described for result outcomes in the previous literature.

For completeness, we also briefly consider the (implied) probability forecasts of result outcomes. For the Model and Tipster forecasts, we sum up all the scoreline forecast probabilities corresponding to each result outcome. For the individual experts, we

[19] The equivalent results and tests of weak efficiency as per Equation (7) are presented in Online Appendix Table B5, with next to no quantitative or qualitative difference to the strong efficiency testing results.

consider a binary variable of whether or not either expert picked that result outcome. In Online Appendix Table B6 we present the strong efficiency regression results, estimating equivalent regression models as before with scorelines, i.e. Equation (8), including the Elo prediction as an explanatory variable.[20] For the draw outcome, we take the squared difference of the Elo prediction from 0.5, referring to this as a "Balance" measure.[21] The table of results has three panels: the top panel for the home win outcome, the middle panel for the draw, and the bottom panel for the away win. We also present the F-test of efficiency (null hypothesis of $\alpha = 0$, $\beta = 1$ and $\boldsymbol{\gamma} = \mathbf{0}$). Despite some individually significant coefficients for γs, the test nonetheless does not reject the null of strong efficiency for the Model and Bookmaker forecasts in all three outcome cases at standard levels, and for the Tipster forecasts in the case of the draw outcome. As with the scoreline picks, the α and β coefficients are closest to zero and one respectively for the Model and Bookmaker forecasts, and are some distance away for the two experts and Tipster forecasts, even being negative for the latter for home and away wins. The β coefficient on the Bookmaker regression is greater than one for the home and away win results, but only significantly so for the latter at standard levels, suggesting that the typical favourite-longshot bias in the EPL during this period and sample of bookmakers only shows up in the away win odds.

5.3 Forecast encompassing

We now consider the outcomes of encompassing regressions, described by Equations (10)-(11). We mix both the density and point forecasts for scorelines, arguing that encompassing is the fairest way to compare these probability and pick forecasts, since the method simply asks whether either can add more information to the other. As such, we ask if the picks by either of the experts, or the point forecasts we derive from our model, add more information to the Model, Bookmaker and Tipster (implied) probability forecasts discussed in the previous section. In total, we consider the bilateral regression encompassing tests for every combination of eight different sources of forecasts.

The forecast encompassing results are summarised in Table 5, using the t-statistics for the equivalent of the estimated ϕ_a and ϕ_b coefficients in each case. The results are presented such that the row is the particular forecast error in the regression equation (the dependent variable), and the column is the other forecast being added into the model (the explanatory variable). Hence for the Model Probabilities, the entry in the first row and column is blank, since we cannot enter the Model Probability forecast into the Model Probability forecast error regression model. The bold faced numbers in the table indicate t-statistics that are very significant, i.e. 3.8 or larger, based on the rule of thumb established in Campos et al. (2003) for adjusting t-statistics with large sample sizes. Using our notation and definition of

[20] The equivalent results and tests of weak efficiency, as per Equation (7), are presented in Online Appendix Table B7, with next to no quantitative or qualitative difference to the strong efficiency testing results.

[21] As the Elo prediction lies on the unit interval, where 0 implies a certain away win and 1 a certain home win, we can take 0.5 to imply a 'certain' draw.

encompassing from before, reading from right to left in the table for a particular source of forecast errors a (b), the t-statistics give the values of ϕ_b (ϕ_a) for each of the other considered sources of forecasts (column). For example, when asking if the Model Probabilities (a) encompass the Bookmakers (b), $\{\widehat{\phi_b} : t\text{-stat} = 1.80\}$ and $\{\widehat{\phi_a} : t\text{-stat} = \mathbf{8.77}\}$. To repeat, one forecast source is said to encompass another if $\mathbf{H_0}$: $\phi_a = 0, \phi_b \neq 0$, and vice versa if $\mathbf{H_0}$: $\phi_a \neq 0, \phi_b = 0$. If $\phi_a \neq 0$ *and* $\phi_b \neq 0$, then a linear combination of such forecasts would be more effective than taking any single forecast in isolation.

Focusing on the Model Probabilities (a) to begin with, adding the model unconditional scoreline picks (b) to the forecast error regression model, $\{\widehat{\phi_b} : t\text{-stat} = \mathbf{-9.7}\}$, suggests that the former adds information to the Model's own forecast probabilities of each scoreline. The corresponding opposite entry, is shown in the first column and second row, $\{\widehat{\phi_a} : t\text{-stat} = \mathbf{-51.0}\}$, and suggests that the Model probability forecasts are able to explain the variation in Model point forecast errors. Taken together, this suggests that the probability forecasts add more information to scoreline point forecasts than vice versa, as would be expected. This occurs for all three types of Model point forecast considered. However, the top right section of Table 5, where the four different types of forecasts generated by the statistical model are presented, overall suggests that some combination of the probability and point forecasts would be optimal, rather than taking any one in isolation, since we do not find that any two sets of these forecast sources encompass each other.

The Model Probabilities (a) encompass the bookmaker and tipster-based probability forecasts, but not the picks made by Lawrenson. If Lawrenson picked a particular scoreline outcome, then the probability-based forecast error from the statistical model increases significantly, $\{\widehat{\phi_b} : t\text{-stat} = \mathbf{5.8}\}$; for Merson this similar effect is borderline significant $\{\widehat{\phi_b} : t\text{-stat} = 2.2\}$. This suggests that we cannot claim that the statistical model and its probabiltiy forecasts are *better* than the expert Lawrenson and his judgement-based point forecasts. Instead, we would conclude that these two sources of forecasts are complementary, despite being of a different type. However, we can conclude based on these results that the statistical model dominates the Bookmaker and Tipster implied probability forecasts, and that it is in a sense *better*. This is consistent with other attempts in the literature to compare statistical models and bookmakers as football match forecasters (e.g. Buraimo et al., 2013; Boshnakov et al., 2017), though in these previous cases the comparisons used betting strategies and returns on investment, and focused on match results rather than scorelines.

We find that the Bookmaker forecasts (a) encompass Merson's forecasts, $\{\widehat{\phi_a} : t\text{-stat} = \mathbf{-12.3}$; $\widehat{\phi_b} : t\text{-stat} = 2.7\}$, but not the Tipster crowd, $\{\widehat{\phi_a} : t\text{-stat} = \mathbf{-15.4}$; $\widehat{\phi_b} : t\text{-stat} = \mathbf{5.9}\}$ or Lawrenson's forecasts $\{\widehat{\phi_a} : t\text{-stat} = \mathbf{-24.0}$; $\widehat{\phi_b} : t\text{-stat} = \mathbf{7.1}\}$. The Tipster crowd implied probability forecasts do not encompass any of the other sources of forecasts studied here. The experts encompass some of the comparable point forecasts from the statistical model, suggesting that the Model would struggle to compete with them in any contest on like for like terms which focused on just picking correct scorelines. However, the Bookmaker

odds and Tipster crowd certainly add information to the experts and would improve their accuracy if that information was used. By and large, we find that the point forecasts add less information to other sources than the probability forecasts do, but this picture is not clear cut, and there is evidence that using combinations of these forecast sources would be optimal.

6 Summary and further discussion

We have studied forecasts of scorelines in association football matches. We described why we consider scorelines to be strange entities, and as such why it is difficult not only to forecast these outcomes but to subsequently evaluate those forecasts. To demonstrate this, we applied a range of scoring rules proposed both in the academic and non-academic realms, and also used conventional statistical methods for evaluating forecasts.

Scoring rules can sometimes compensate forecasters that make point forecasts for the difficulty of this endeavour, by awarding credit for sub-outcomes. Density or probability forecasts are unfairly favoured in conventional statistical approaches, such as regression-based efficiency testing, but probability forecasts cannot be used when applying scoring rules designed for point forecasts. There was vague evidence that the scoring rule by which a forecaster's scoreline predictions were being judged influenced their 'effort' devoted to making correct scoreline picks relative to result picks. Forecast encompassing, we believe, provides the fairest way to compare point and probability forecasts, and we found what might be anticipated: probability forecasts do tend to contain more information than point forecasts. Nonetheless, some combination of probability and point forecasts is likely to be most effective when attempting to predict the outcome of football matches. There was also evidence that the implied scoreline probabilities from bookmaker odds and a crowd of tipsters significantly explained the point forecast errors of football forecasting experts.

All of this begs an obvious question: is there a better way to evaluate point forecasts in this context? Or, is there a better way to compare them with probability forecasts? One potential answer, perhaps, is that we should focus on betting returns, aligning as far as possible with the objectives of the forecaster. If the forecaster has an objective or scoring rule that weights scorelines highly over results, then we should judge their forecast success using a betting portfolio which does so similarly. For probability forecasts, a way to judge performance is to consider whether the optimal distribution of a forecaster's budget over the available scoreline odds offered, for example, achieves a financial return. However, both these methods assume that the forecaster cares about making a return, which might not truly be their objective. Similarly, without information on the forecaster's risk preferences it is not clear that this is the best approach. Nevertheless, if point and probability forecasts of the same events were evaluated in this way, then it would be easy to compare the two different types based on the common metric of implied financial return on investment, notwithstanding reservations that this assumes identical risk preferences and motivations of

the individual forecasters. There is a further challenge of basing evaluation and comparison on implied market returns alone; it assumes that the bookmaker odds are exogenous to the forecasting behaviour of individuals (see Levitt, 2004 for evidence to the contrary).

This study focuses on a specific context. But this is a context where many people explicitly make forecasts and truly care about the outcomes of the events and the accuracy of those forecasts, far beyond any financial gain they might achieve from them. This justifies our attention and yet further research, and it would be interesting in the first instance to study scoreline forecasts from other sports, for example American and rugby football. Forecasts or picks of scorelines are particularly interesting because they contain sub-outcomes, which are as equally interesting or even more so than the scoreline outcome itself. Conditioning on these sub-outcomes might frequently suggest a different point forecast is made altogether. This is a feature especially caused by the draw being an acceptable, common, and final outcome of association football matches, which is unusual in other areas of society or even in other types of football, never mind more generally within professional sports. Although other sports feature draws, they are uncommon as an ultimate outcome of the contest, with one exception being first class cricket, though predicting a draw in that sport typically involves weather forecasting. On reflection, it is the prevalence of draws among the final outcomes of association football matches which makes forecasting these events most strange in our view.

References

Akhtar, S., and P. Scarf. 2012. "Forecasting test cricket match outcomes in play." *International Journal of Forecasting*, 28(3): 632–643.

Angelini, G., and L. De Angelis. 2019. "Efficiency of online football betting markets." *International Journal of Forecasting*, 35(2): 712–721.

Asif, M., and I. McHale. 2019. "A generalized non-linear forecasting model for limited overs international cricket." *International Journal of Forecasting*, 35(2): 634–640.

Asif, M., and I. McHale. 2016. "In-play forecasting of win probability in One-Day International cricket: A dynamic logistic regression model." *International Journal of Forecasting*, 32(1): 34–43.

Ayton, P., D. Önkal, and L. McReynolds. 2011. "Effects of ignorance and information on judgments and decisions." *Judgment and Decision Making*, 6(5): 381–391.

Boshnakov, G., T. Kharrat, and I. McHale. 2017. "A bivariate Weibull count model for forecasting association football scores." *International Journal of Forecasting*, 33(2): 458–466.

Brier, G. 1950. "Verification of forecasts expressed in terms of probability." *Monthly Weather Review*, 78(1): 1–3.

Brown, A., and J. J. Reade. 2019. "The wisdom of amateur crowds: Evidence from an online community of sports tipsters." *European Journal of Operational Research*, 272(3): 1073–1081.

Buraimo, B., D. Peel, and R. Simmons. 2013. "Systematic Positive Expected Returns in the UK Fixed Odds Betting Market: An Analysis of the Fink Tank Predictions." *International Journal of Financial Studies*, 1(4): 1–15.

Butler, D., R. Butler, and J. Eakins. 2020. "Expert performance and crowd wisdom: Evidence from English Premier League predictions." *European Journal of Operational Research*.

Cain, M., D. Law, and D. Peel. 2000. "The Favourite-Longshot Bias and Market Efficiency in UK Football Betting." *Scottish Journal of Political Economy*, 47(1): 25–36.

Campos, J., D. Hendry, and H.-M. Krolzig. 2003. "Consistent Model Selection by an Automatic Gets Approach." *Oxford Bulletin of Economics and Statistics*, 65(s1): 803–819.

Chong, Y., and D. Hendry. 1986. "Econometric evaluation of linear macro-economic models." *The Review of Economic Studies*, 53(4): 671–690.

del Corral, J., and J. Prieto-Rodríguez. 2010. "Are differences in ranks good predictors for grand slam tennis matches?" *International Journal of Forecasting*, 26(3): 551–563, Sports Forecasting.

Dixon, M., and S. Coles. 1997. "Modelling association football scores and inefficiencies in the football betting market." *Applied Statistics*, 47(3): 265—280.

Dixon, M., and P. Pope. 2004. "The value of statistical forecasts in the UK association football betting market." *International Journal of Forecasting*, 20(4): 697–711.

Elaad, G., J. J. Reade, and C. Singleton. 2020. "Information, prices and efficiency in an online betting market." *Finance Research Letters*, 35, p. 101291.

Elo, A. E. 1978. *The rating of chessplayers, past and present*. London Batsford.

Fair, R., and R. Shiller. 1989. "The Informational Context of Ex Ante Forecasts." *The Review of Economics and Statistics*, 71(2): 325–331.

Fawcett, N., L. Körber, R. Masolo, and M. Waldron. 2015. "Evaluating UK point and density forecasts from an estimated DSGE model: the role of off-model information over the financial crisis." Staff Working Paper 538, Bank of England.

Forrest, D. 2011. "The Past and Future of the British Football Pools." *Journal of Gambling Studies*, 15, p. 161–176.

Forrest, D., and L. Pérez. 2011. "Football pools and lotteries: substitute roads to riches?" *Applied Economics Letters*, 18(13): 1253–1257.

Forrest, D., J. Goddard, and R. Simmons. 2005. "Odds-Setters As Forecasters: The Case of English Football." *International Journal of Forecasting*, 21(3): 551–564.

Forrest, D., and R. Simmons. 2000. "Forecasting Sport: The Behaviour and Performance of Football Tipsters." *International Journal of Forecasting*, 16(3): 317–331.

Foulley, J.-L., and G. Celeux. 2018. "A penalty criterion for score forecasting in soccer." *arXiv preprint arXiv:1806.01595*.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann. 2013. "Combining expert forecasts: Can anything beat the simple average?" *International Journal of Forecasting*, 29(1): 108–121.

Goddard, J. 2005. "Regression Models for Forecasting Goals and Match Results in Association Football." *International Journal of Forecasting*, 21(2): 331–340.

Granger, C. W. J., and M. H. Pesaran. 2000. "Economic and statistical measures of forecast accuracy." *Journal of Forecasting*, 19(7): 537–560.

Heuer, A., and O. Rubner. 2012. "How Does the Past of a Soccer Match Influence Its Future? Concepts and Statistical Analysis." *PLOS One*, 7(11): 1–7.

Hvattum, L. M., and H. Arntzen. 2010. "Using elo ratings for match result prediction in association football." *International Journal of Forecasting*, 26(3): 460–470.

Karlis, D., and I. Ntzoufras. 2003. "Analysis of Sports Data Using Bivariate Poisson Models." *Journal of the Royal Statistical Society (Statistician)*, 52(3): 381–393.

Karlis, D., and I. Ntzoufras. 2005. "Bivariate Poisson and Diagonal Inflated Bivariate Poisson Regression Models in R." *Journal of Statistical Software*, 14(10): .

Klaassen, F., and J. Magnus. 2003. "Forecasting the winner of a tennis match." *European Journal of Operational Research*, 148(2): 257–267, Sport and Computers.

Lawrence, M., P. Goodwin, M. O'Connor, and D. Önkal. 2006. "Judgmental forecasting: A review of progress over the last 25 years." *International Journal of Forecasting*, 22(3): 493–518.

Levitt, S. 2004. "Why are gambling markets organised so differently from financial markets?" *The Economic Journal*, 114(495): 223–246.

Maher, M. 1982. "Modelling association football scores." *Statistica Neerlandica*, 36(3): 109–118.

McHale, I., and A. Morton. 2011. "A Bradley-Terry type model for forecasting tennis match results." *International Journal of Forecasting*, 27(2): 619–630.

Mincer, J., and V. Zarnowitz. 1969. "The evaluation of economic forecasts." In *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*. NBER, 1–46.

O'Leary, D. 2017. "Crowd performance in prediction of the World Cup 2014." *European Journal of Operational Research*, 260(2): 715–724.

Ottaviani, M., and P. N. Sørensen. 2008. "The Favorite-Longshot Bias: An Overview of the Main Explanations." In *Handbook of Sports and Lottery Markets*. Eds. by D. B. Hausch, and W. T. Ziemba, San Diego Elsevier, 83–101.

Peeters, T. 2018. "Testing the Wisdom of Crowds in the field: Transfermarkt valuations and international soccer results." *International Journal of Forecasting*, 34(1): 17–29.

Pope, P., and D. Peel. 1989. "Information, Prices and Efficiency in a Fixed-Odds Betting Market." *Economica*, 56(223): 323–341.

Reade, J. J. 2014. "Information and predictability: Bookmakers, prediction markets and tipsters as forecasters." *The Journal of Prediction Markets*, 8(1): 43–76.

Shannon, C. E. 1948. "A mathematical theory of communication." *Bell System Technical Journal*, 27(3): 379–423.

Simmons, J., L. Nelson, J. Galak, and S. Frederick. 2010. "Intuitive biases in choice versus estimation: Implications for the wisdom of crowds." *Journal of Consumer Research*, 38(1): 1–15.

Singleton, C., J. J. Reade, and A. Brown. 2019. "Going with your gut: The (In)accuracy of forecast revisions in a football score prediction game." *Journal of Behavioral and Experimental Economics*, p. 101502.

Snowberg, E., and J. Wolfers. 2010. "Explaining the Favorite-Longshot Bias: Is It Risk-Love or Misperceptions?" *Journal of Political Economy*, 118(4): 723–746.

Spann, M., and B. Skiera. 2009. "Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters." *Journal of Forecasting*, 28(1): 55–72.

Štrumbelj, E. 2014. "On determining probability forecasts from betting odds." *International Journal of Forecasting*, 30(4): 934–943.

Štrumbelj, E., and M. Šikonja. 2010. "Online bookmakers' odds as forecasts: The case of European soccer leagues." *International Journal of Forecasting*, 26(3): 482–488.

Surowiecki, J. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Brown Little.

Vaughan Williams, L., M.-C. Sung, P. Fraser-Mackenzie, J. Peirson, and J. Johnson. 2018. "Towards an Understanding of the Origins of the Favourite–Longshot Bias: Evidence from Online Poker Markets, a Real-money Natural Laboratory." *Economica*, 85(338): 360–382.

TABLE 1: Frequency of scoreline outcomes in the 2016–17 and 2017–18 EPL seasons (%).

		2016–17 Away goals								2017–18 Away goals						
		0	1	2	3	4	5	6	7	0	1	2	3	4	5	6
Home goals	0	7.1	5.5	4.5	2.6	1.8	0.3	0.0	0.0	8.4	6.1	3.9	3.2	1.8	0.0	0.3
	1	10.0	10.0	6.3	3.2	1.8	0.3	0.3	0.3	11.6	11.8	6.3	1.3	1.8	0.3	0.0
	2	8.7	7.9	4.5	0.8	0.5	0.0	0.0	0.0	7.1	8.4	5.0	2.9	0.3	0.3	0.0
	3	5.0	6.8	2.1	0.5	0.5	0.0	0.0	0.0	3.9	3.4	1.1	0.8	0.0	0.0	0.0
	4	2.9	1.3	1.6	0.5	0.0	0.0	0.0	0.0	2.4	2.9	0.3	0.5	0.0	0.0	0.0
	5	0.8	0.5	0.0	0.0	0.3	0.0	0.0	0.0	2.4	0.8	0.3	0.0	0.3	0.0	0.0
	6	0.0	0.5	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0

Source: *Soccerbase.com*

TABLE 2: Result outcomes in the 2016–17 and 2017–18 EPL seasons (%): comparison of actual outcomes with the average implied frequency from bookmaker prices

	Bookmakers			Actual		
Season	Home	Draw	Away	Home	Draw	Away
2016/17	46.1	25.3	32.3	49.2	22.1	28.7
2017/18	46.3	25.3	32.4	45.5	26.1	28.4

Source: author calculations using *Oddsportal.com* and *Soccerbase.com*

TABLE 3: Scoreline point forecast evaluation using different metrics for each of the 2016/17 and 2017/18 EPL seasons

	Scoring Rule								ROI (%)			
	Brier (1) ↓	Scores (%) (2) ↑	Results (%) (3) ↑	Close (%) (4) ↑	BBC (5) ↑	Sky (6) ↑	Sbru (7) ↑	Pen (8) ↓	Results (9) ↑	Scores (10) ↑	Margins (11) ↑	Total Gls (12) ↑
2016/17												
Sbru: (N = 50)												
Mean	0.91	9.2	50.8	46.6	282.5	232.8	294.3	-306.1	2.4	-21.4	-1.5	-3.4
Median	0.91	9.3	50.6	46.6	287	234.2	295.2	-307.4	2.2	-19.5	-1.8	-3.2
Minimun	0.88	5.4	39.8	35.5	196	167.5	209.5	-363.1	-11.0	-46.1	-10.1	-12.7
Maximum	0.95	11.9	60.1	56.5	344	288.5	373.5	-259.8	26.4	-4.5	4.23	3.3
St. dev.	0.0	1.5	4.2	4.1	33.3	26.6	35.1	26.3	7.3	10.4	3.2	3.3
Merson	0.98	10.5	57.9	47.9	340	280	376	-272.0	16.4	-4.2	4.5	-0.3
Lawrenson	0.98	11.6	56.6	56.2	347	281	387	-273.3	19.7	-10.9	10.3	-2.5
Unconditional	0.98	11.3	45.3	51.3	301	236.5	292	-319.5	-3.4	-10.8	-3.9	-1.9
Conditional	0.98	12.6	56.8	49.0	360	288	375.5	-317.2	12.7	-5.2	3.1	-1.6
Fuzzy Cond.	0.98	11.8	51.8	48.2	332	264.5	326.5	-320.5	1.3	-7.0	-1.3	-0.4
2017/18												
Sbru: (N = 150)												
Mean	0.91	9.1	49.4	46.5	298.9	245.8	308.4	-339.2	2.5	-24.9	-3.7	-5.9
Median	0.91	9.1	49.7	47.1	296	244.8	306.5	-333.9	2.4	-24.4	-3.5	-5.7
Minimum	0.87	5.5	35.4	35.6	201	168	208	-539.7	-15.0	-51.2	-11.4	-16.8
Maximum	0.94	13.4	54.7	53.9	439	359.5	451.5	-299.1	17.9	1.0	3.7	3.1
St. dev.	0.0	1.6	2.7	4.1	27.1	19.4	24.1	29.5	6.3	10.8	3.3	3.1
Merson	0.98	7.6	51.8	45.4	284	240.5	316.5	-314.3	1.5	-33.3	-1.0	-4.9
Lawrenson	0.98	12.6	50.5	50.4	336	264	359	-307.8	4.4	-8.3	-2.5	-3.7
Unconditional	0.98	10.5	48.2	50.8	303	243	303	-324.3	4.8	-25.8	-6.7	-6.5
Conditional	0.98	10.5	52.1	49.7	318	258	341.5	-334.2	-0.2	-26.6	-6.5	-7.6
Fuzzy Cond.	0.98	11.6	51.1	49.5	326	260	325	-324.8	1.3	-18.4	-7.0	-7.4

Notes: Percentages, competition scores and returns rounded to the nearest integer. ↑ indicates that better forecast performance is implied by an increased value of the metric, and vice versa for ↓. From left to right, "Brier" gives the MSFE as per Equation (6). "Score" gives the percentage of correct scoreline picks made. "Result" gives the percentage of correct result picks made. "Close" gives the percentage of close picks according to the Superbru metric (see Appendix A). "BBC" awards 4 points to a correct scoreline but 1 point to a correct result only. "Sky" awards 2.5 points to a correct scoreline but 1 point to a correct results only. "Sbru" awards 3 points to a correct scoreline but 1 point to a correct result only, with 0.5 points in addition for a close pick which is also a correct result but not a correct scoreline. "Pen" applies the metric of Foulley and Celeux (2018) (see Online Appendix A). Columns (9)-(12) give implied returns on investment from betting the same amount x over the whole season on each and every match, consistent with the scoreline point forecast made, i.e. a total investment by the forecaster/bettor over the season of $380x$ for either the result, scoreline, margin and total number of goals in a game.

TABLE 4: Strong efficiency tests for forecast scoreline outcomes

	Model	Bookmakers	Lawrenson	Merson	Tipsters
	(1)	(2)	(3)	(4)	(5)
Constant ($\widehat{\alpha}$)	0.002	−0.002	0.011***	0.011***	0.007***
	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Forecast ($\widehat{\beta}$)	0.839***	1.156***	0.111***	0.080***	0.458***
	(0.014)	(0.018)	(0.004)	(0.004)	(0.010)
Scoreline freq.	−0.00005	0.001	0.0002	−0.00004	0.0002
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Points (H)	0.00000	0.00001	0.00000	0.00000	−0.000
	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
Points diff.	−0.00000	−0.00001	−0.00000	−0.00000	0.000
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Form (H)	0.00000	−0.00003	0.00000	−0.00000	0.00001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Form diff.	0.00000	0.00001	−0.00000	0.00000	0.00000
	(0.0001)	(0.0001)	(0.0002)	(0.0002)	(0.0002)
Elo prediction	0.00001	−0.0001	0.00002	0.00001	−0.00001
	(0.004)	(0.004)	(0.004)	(0.005)	(0.004)
Observations	61,560	61,560	61,560	61,560	61,560
Adjusted R^2	0.052	0.063	0.012	0.006	0.036
Resid. std. error	0.108	0.107	0.110	0.110	0.108
F-test of efficiency	0.000***	0.000***	0.000***	0.000***	0.000***

Notes: *p<0.1; **p<0.05; ***p<0.01, two-tailed tests.

TABLE 5: Encompassing testing for scoreline forecasts

	Prob.	Uncond.	Cond.	Fuzzy	Book.	Tipster	Lawr.	Mers.
Model Prob.		**-9.70**	**-6.69**	**-6.92**	1.80	1.46	**5.75**	2.22
Model Uncond.	**-50.97**		**-113.05**	**-142.84**	**-26.64**	**-27.31**	**-14.83**	**-5.43**
Model Cond.	**-47.46**	**-115.20**		**-166.20**	**-22.99**	**-29.86**	-1.24	-3.22
Model Fuzzy	**-49.10**	**-145.99**	**-166.82**		**-24.82**	**-30.21**	**-7.94**	**-4.21**
Bookmaker	**8.77**	3.15	**5.97**	**5.77**		**5.91**	**7.11**	2.65
Tipster	**-11.47**	**-9.81**	**-9.49**	**-9.33**	**-15.35**		**-18.00**	**-18.09**
Lawrenson	**-17.23**	**-17.17**	-2.25	**-8.70**	**-23.96**	**-48.62**		**-29.05**
Merson	**-5.60**	-1.96	1.45	0.72	**-12.25**	**-33.73**	**-22.84**	

Note: bold-faced numbers indicate t-statistics larger than 3.8. The positive sign of the statistics implies that the column forecasts on average increase the errors of the row forecasts, and vice versa for a negative sign.

Evaluating Strange Forecasts: The Curious Case of Football Match Scorelines

Online Appendix

July 2020

Appendix A. Measures of 'closeness'

Superbru

The *Superbru* closeness metric is given by:

$$c_i = |\widehat{m}_i - m_i| + \left| \frac{\widehat{t}_i - t_i}{2} \right| \ . \tag{12}$$

Users get 1.5 points if $c_i \leq 1.5$ and the result is correct. In practice, this equates to the forecast having one goal more (less) for one or both teams than what actually occurred.

Foulley & Celeux

Foulley and Celeux (2018) propose a forecast penalty measure which is similar to *Superbru*'s measure, but which penalises the difference in result more and the distance from scoreline relatively less. The measure is summarised as:

$$FP(\mathbf{s}_i, \widehat{\mathbf{s}}_i) = C(\mathbf{s}_i, \widehat{\mathbf{s}}_i) + D(\mathbf{s}_i, \widehat{\mathbf{s}}_i) \ , \tag{13}$$

where:

$$C = \begin{cases} 0 & \text{if } r_i(\widehat{\mathbf{s}}_i) = r_i(\mathbf{s}_i) \\ c_0 & \text{if } |r_i(\widehat{\mathbf{s}}_i) - r_i(\mathbf{s}_i)| = 0.5 \\ 2c_0 & \text{if } |r_i(\widehat{\mathbf{s}}_i) - r_i(\mathbf{s}_i)| = 1 \ , \end{cases} \tag{14}$$

$$D(\mathbf{s}_i, \widehat{\mathbf{s}}_i) = \frac{\|\mathbf{s}_i - \widehat{\mathbf{s}}_i\|_2}{\|\mathbf{s}_i\|_2 + \|\widehat{\mathbf{s}}_i\|_2} \ , \tag{15}$$

where c_0 is some positive constant.

Appendix B. Additional tables

TABLE B1: Implied frequency (probability) from average bookmaker odds for scoreline outcomes in the 2016–17 and 2017–18 EPL seasons.

		2016–17 Away goals								2017–18 Away goals							
		0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
Home goals	0	8.8	7.6	4.0	1.7	0.9	0.7	0.6	0.3	8.5	7.3	3.6	1.5	0.9	0.6	0.6	0.4
	1	10.5	13.1	6.7	2.5	1.0	0.7	0.3	0.3	10.2	12.6	6.2	2.3	1.0	0.7	0.3	0.4
	2	6.8	9.1	5.9	2.3	0.9	0.4	0.3	0.2	6.5	8.7	5.6	2.1	0.9	0.4	0.3	0.3
	3	3.1	4.2	3.0	1.5	0.6	0.3	0.3	0.1	2.9	3.9	2.7	1.3	0.5	0.3	0.3	0.3
	4	1.4	1.7	1.3	0.7	0.4	0.3	0.2	0.1	1.4	1.6	1.2	0.6	0.4	0.3	0.3	
	5	0.9	0.9	0.6	0.3	0.3	0.2	0.2	0.1	0.9	0.9	0.5	0.4	0.3	0.3	0.3	
	6	0.6	0.4	0.3	0.3	0.2	0.2	0.2	0.1	0.7	0.3	0.4	0.3	0.3	0.3	0.3	
	7	0.3	0.3	0.2	0.1	0.1	0.1	0.1	0.1	0.7	0.6	0.4	0.3				

Source: author calculations using *Oddsportal.com* and *Soccerbase.com*

TABLE B2: Frequency of tips by Superbru tipsters for each scoreline outcome in the 2016–17 and 2017–18 EPL seasons (%).

		2016–17							2017–18						
		Away goals							Away goals						
		0	1	2	3	4	5	6	0	1	2	3	4	5	6
	0	2.1	5.8	6.6	1.8	0.2	0.0	0.0	1.5	5.1	5.9	2.1	0.5	0.1	0.0
	1	9.8	12.6	11.5	4.3	0.3	0.0	0.0	10.0	13.7	12.8	5.0	0.6	0.1	0.0
	2	10.1	14.3	4.6	0.7	0.1	0.0	0.0	10.6	15.2	3.7	0.7	0.1	0.0	0.0
Home goals	3	3.2	4.1	0.8	0.0	0.0	0.0	0.0	4.1	5.1	0.8	0.1	0.0	0.0	0.0
	4	0.4	0.3	0.1	0.0	0.0	0.0	0.0	1.2	0.6	0.2	0.0	0.0	0.0	0.0
	5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.0

Source: *Superbru*

TABLE B3: Frequency of tips by 'experts' for each scoreline outcome in the 2016–17 and 2017–18 EPL seasons (%).

			2016–17				2017–18					
			Away goals				Away goals					
			0	1	2	3	0	1	2	3	4	5
		0	0.3	0.5	16.4	0.5	0.0	0.3	15.9	0.8	0.0	0.0
		1	1.6	26.1	5.0	0.0	1.9	26.8	3.4	0.0	0.0	0.0
Lawrenson	Home goals	2	28.8	14.0	0.5	0.0	31.0	13.3	0.5	0.0	0.0	0.0
		3	6.1	0.3	0.0	0.0	5.3	0.3	0.0	0.0	0.0	0.0
		4	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0
		0	0.0	0.0	3.2	4.0	0.3	2.1	4.7	3.4	1.3	0.0
		1	2.9	14.8	9.0	7.4	8.2	9.8	9.0	7.4	0.0	0.3
Merson	Home goals	2	13.8	13.0	5.8	0.3	17.9	14.8	1.6	0.3	0.0	0.0
		3	11.9	9.8	2.6	0.3	9.5	4.2	0.8	0.0	0.0	0.0
		4	1.1	0.0	0.3	0.0	3.7	0.3	0.0	0.0	0.0	0.0
		5	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.0

Source: author calculations using *BBC Sport* & *Sky Sports*

TABLE B4: Selected ranks out of 55 (50 tipsters, 2 experts, 3 models) and 155 (150 tipsters, 2 experts, 3 models) according to different scoring rules in the 2016/17 and 2017/18 EPL seasons, respectively

	Scoring Rule								ROI			
	Brier (1)	Scores (%) (2)	Results (%) (3)	Close (4)	BBC (5)	Sky (6)	Sbru (7)	Pen (8)	Results (9)	Scores (10)	Margins (11)	Total Gls (12)
2016/17												
Merson	5	14	3	20	5	4	2	7	4	1	2	9
Lawrenson	3	5	7	2	2	3	1	8	2	11	1	23
Unconditional	4	8.5	52	9	17.5	29	33.5	39	44	27	48	20
Conditional	1	1	6	16	1	2	3	35	6	6	7	12
Fuzzy Cond.	2	3	25	19	8	11	12	40	30	18	27	11
2017/18												
Merson	5	130.5	22	98	114	100	45	22	85	123	35	60
Lawrenson	1	3	65	29	16	22	3	7	56	11	57	38
Unconditional	3.5	29.5	107.5	22	57.5	88.5	97.5	49	48	89	120	89
Conditional	3.5	29.5	16.5	38	31.5	29.5	15	81	104	93	118	111
Fuzzy Cond.	2	11.5	50.5	43	24	27.5	29.5	50	88	45	125	105

Notes: see Table 3 in main text.

TABLE B5: Weak efficiency tests for forecast scoreline outcomes

	Model	Bookmakers	Lawrenson	Merson	Tipsters
	(1)	(2)	(3)	(4)	(5)
Constant ($\widehat{\alpha}$)	0.002***	−0.002***	0.011***	0.011***	0.007***
	(0.0005)	(0.0005)	(0.0004)	(0.0004)	(0.0005)
Forecast/Pick ($\widehat{\beta}$)	0.839***	1.156***	0.111***	0.080***	0.458***
	(0.014)	(0.018)	(0.004)	(0.004)	(0.010)
Observations	61,560	61,560	61,560	61,560	61,560
Adjusted R^2	0.052	0.063	0.012	0.006	0.036
Resid. std. error	0.107	0.107	0.110	0.110	0.108
Ftest of efficiency	0.000***	0.000***	0.000***	0.000***	0.000***

Notes: *p<0.1; **p<0.05; ***p<0.01.

TABLE B6: Strong efficiency tests for forecast result outcomes (home win, draw, away win)

	Model	Bookmakers	Lawrenson	Merson	Tipsters
	(1)	(2)	(3)	(4)	(5)
Constant ($\widehat{\alpha}$)	0.005	0.071	0.045	0.056	0.277**
	(0.045)	(0.045)	(0.047)	(0.046)	(0.120)
Home-win forecast ($\widehat{\beta}$)	0.317**	1.158***	0.116***	0.163***	−0.255**
	(0.130)	(0.200)	(0.043)	(0.046)	(0.108)
Elo prediction	0.660***	−0.238	0.750***	0.657***	0.562***
	(0.138)	(0.215)	(0.109)	(0.115)	(0.176)
Adjusted R^2	0.142	0.176	0.145	0.151	0.142
F-test of efficiency	0.61	0.978	0.000	0.000	0.000
Constant ($\hat{\alpha}$)	0.195***	0.016	0.244***	0.271***	0.244***
	(0.065)	(0.102)	(0.025)	(0.023)	(0.051)
Draw forecast ($\widehat{\beta}$)	0.299	0.945***	0.102***	0.048	0.120
	(0.211)	(0.354)	(0.036)	(0.043)	(0.148)
Elo predict (balance)	−0.795**	−0.068	−0.833**	−0.964***	−0.757
	(0.393)	(0.508)	(0.364)	(0.364)	(0.493)
Adjusted R^2	0.011	0.020	0.020	0.010	0.009
F-test of efficiency	0.835	1.000	0.000	0.000	0.395
Constant ($\widehat{\alpha}$)	0.432***	−0.313**	0.546***	0.556***	0.620***
	(0.091)	(0.131)	(0.053)	(0.058)	(0.053)
Away-win forecast ($\widehat{\alpha}$)	0.442***	1.406***	0.229***	0.174***	−0.340***
	(0.124)	(0.169)	(0.044)	(0.044)	(0.111)
Elo prediction	−0.557***	0.343*	−0.624***	−0.640***	−0.362*
	(0.119)	(0.165)	(0.090)	(0.097)	(0.187)
Adjusted R^2	0.166	0.225	0.181	0.169	0.162
F-test of efficiency	0.67	0.916	0.000	0.000	0.000
Observations	760	759	756	757	760

Note: *p<0.1; **p<0.05; ***p<0.01, two-tailed tests.

TABLE B7: Weak efficiency tests for forecast result outcomes (home win, draw, away win)

	Model	Bookmakers	Lawrenson	Merson	Tipsters
	(1)	(2)	(3)	(4)	(5)
Constant ($\widehat{\alpha}$)	0.112^{***}	0.043	0.319^{***}	0.277^{***}	0.653^{***}
	(0.040)	(0.038)	(0.025)	(0.026)	(0.024)
Home-win forecast ($\widehat{\beta}$)	0.810^{***}	0.957^{***}	0.306^{***}	0.344^{***}	-0.558^{***}
	(0.080)	(0.076)	(0.035)	(0.035)	(0.052)
Adjusted R^2	0.117	0.173	0.092	0.115	0.131
F-test of efficiency	0.919	0.995	0.000	0.000	0.000
Constant ($\widehat{\alpha}$)	0.116^{**}	0.005	0.205^{***}	0.229^{***}	0.179^{***}
	(0.052)	(0.061)	(0.018)	(0.017)	(0.028)
Draw forecast ($\widehat{\beta}$)	0.482^{**}	0.979^{***}	0.122^{***}	0.072^{*}	0.278^{***}
	(0.191)	(0.246)	(0.035)	(0.042)	(0.107)
Adjusted R^2	0.007	0.019	0.015	0.003	0.008
F-test of efficiency	0.894	1.000	0.000	0.000	0.419
Constant ($\widehat{\alpha}$)	0.023	-0.047^{*}	0.200^{***}	0.190^{***}	0.531^{***}
	(0.028)	(0.027)	(0.017)	(0.018)	(0.025)
Away-win forecast ($\widehat{\beta}$)	0.892^{***}	1.090^{***}	0.398^{***}	0.361^{***}	-0.537^{***}
	(0.079)	(0.074)	(0.037)	(0.035)	(0.045)
Adjusted R^2	0.143	0.221	0.130	0.123	0.159
F-test of efficiency	0.973	0.979	0.000	0.000	0.000
Observations	760	759	756	757	760

Notes: *p<0.1; **p<0.05; ***p<0.01.