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

February 2019
Link to latest version

Abstract

This study analyses point forecasts for a common set of events. These forecasts were made for distinct competitions and originally judged differently. The event outcomes were low-probability but had more predictable sub-outcomes upon which they were also judged. Hence, the forecasts were multi-dimensional, complicating any evaluation. The events were football matches in the English Premier League. The forecasts were of exact scoreline outcomes. We compare these with implied probability forecasts using bookmaker odds and a crowd of tipsters, as well as point and probabilistic forecasts generated from a statistical model suited to predicting football match scorelines. By evaluating these sources and types of forecast using various methods, we conclude that forecasts of this type are strange. We argue that regression encompassing is the most appropriate way to compare point and probabilistic forecasts, and find that both types of forecasts for football match scorelines generally add information to one another.

JEL Classification: C53, L83, G14, G17.
Keywords: Forecasting, statistical modelling, regression models, prediction markets

*j.j.reade@reading.ac.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 helpful comments and advice from Mike Clements, and from seminar participants at the Universities of Reading and Plymouth, as well as at the 12th International Conference on Computational and Financial Econometrics in Pisa in December 2018.

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, and those states of nature can have different consequences in money or other terms. As such, the evaluation of forecasts is important. In principle, forecast evaluation ought to relate to those preferences. Martinez (2018) has achieved precisely that, evaluating the 12-hour ahead forecasts of hurricane paths using the recorded damages they subsequently inflicted upon landfall. But for many variables and contexts the inability to construct loss functions of this kind, allied with the generally (quasi-)continuous nature of macroeconomic variables, has led to the use of more statistical measures in forecast evaluation (e.g. Fawcett et al., 2015).

In this study we evaluate forecasts of football match outcomes. Ultimately, after all the punditry is said and done, there are just two important aspects of the outcome of a football match: the *result* and the *scoreline*. The result is a win for either team, or a tie (draw). The scoreline gives the exact number of goals scored by each team. A football scoreline is thus a pair of correlated non-negative integers. 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, perhaps due to the more complicated nature of the scoreline. But scorelines also matter. Many forecasts are made regarding them, both formal and informal. As there are only three possible outcomes for the result, and a countably infinite number of 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 43% of the time), while the most likely *scoreline outcome* is a 1-1 draw (occurring roughly 11% of the time).[2]

[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.

[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.

The scoreline is a *strange* variable though. 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 aspect of football match outcomes: the *total number of goals*. 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 margin of difference, take integer values, and it may be that they are used to evaluate scoreline forecasts. However, 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 the spectators (a.k.a. consumers).

Despite these complexities, or perhaps because of them, popular public competitions exist for forecasting football match scorelines. *Sky Sports*, a 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.[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. Such closeness measures are typically a function of both the total number of goals and the margin of difference.

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 probabilistic 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 probabilistic 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 from what actually occurred, and hence every forecast can be thought of as partly right. The evaluation metrics associated with the scoreline forecast

[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, former professional and international footballer Mark 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, e.g. www.bbc.co.uk/sport/football/...

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.

We compare a range of scoreline 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 evaluate these forecasts according to a range of metrics, attempting to capture their *strange* nature. We find evidence that the tipsters within an online sports prediction game and two football experts made biased forecasts for football match results. However, this bias was not necessarily present when the same forecasters predicted scoreline outcomes. There is suggestive evidence that the competition scoring rule, which was originally used to judge a forecaster's match outcome predictions, 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 different sources of point and probability forecasts for the same set of events. We find that the probability-based forecasts tend to contain more 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 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 Literature

This study sits ostensibly in the forecasting literature; scorelines are low-probability events that are being forecast, and we look to evaluate these forecasts. Scorelines are multi-faceted events, with the result, the margin and the total number of goals all being functions of a scoreline, and each having a differing level of importance. Reflecting this, while the most utility is perhaps derived from a correct scoreline forecast, these three sub-outcomes matter and can be weighted accordingly in any evaluation. Forecasts are statements regarding different states of nature, and those different states imply different pay-offs for contestants, particularly in the EPL, where the 2018 champions Manchester City were awarded £149.4m.[5] As such, we could attempt to evaluate forecasts in terms of the monetary gains and losses implied. In a similar manner, Martinez (2018) evaluated forecasts of hurricane landfalls using estimates of the damages they caused, while Clements (2004) evaluated Bank of England inflation forecasts based on the consequences of missed targets by its Monetary Policy Committee.

[5] See www.sportingintelligence.com/....

Many previous attempts have been made to evaluate forecast competitions (e.g. Makridakis and Hibon, 2000), in particular in the context of sport. Those competitions usually involved some mixture of statistical models developed for the purpose of forecasting, bookmaker prices, which are a form of outcome forecasts, and expert tipsters, who generated forecasts without necessarily using any kind of statistical modelling.

Forrest et al. (2005) considered the role of bookmakers as forecasters, noting that over their sample period in the 1990s and 2000s, bookmakers became more accurate at forecasting outcomes, reflecting growing commercial pressures during that time. Forrest and Simmons (2000) evaluated the predictions of British newspaper tipsters, journalists providing forecasts of forthcoming 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 in Forrest and Simmons (2000) only 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 these particular tipster picks do provide some information not contained within bookmaker prices, finding it is embodied within the crowd of tipsters on the website rather than in any one individual. In all these cases, tipsters were picking match result outcomes rather than exact scorelines.[6] While it is not known how individual tipsters construct forecasts, it is reasonable to view them as judgemental forecasts (see Lawrence et al., 2006). Further, Ayton et al. (2011) have found evidence of the role very simple heuristics can play in successfully forecasting match result outcomes in English football.

Considering the forecasting performance of 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 scorelines, 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 more knowledge than others to be experts. Genre et al. (2013) evaluated such expert forecasts, studying combinations of expert forecasters from the European Central Bank's Survey of Professional Forecasters. Clements

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

(2006, 2009) also evaluated forecasts from the experts in the US Survey of Professional Forecasters.

On low-probability events, Snowberg and Wolfers (2010) suggested that one explanation for the commonly observed favourite-longshot bias in betting prices is agent misperceptions, i.e. bettors cannot distinguish between events with different low probabilities of occurring. If this is true, we might expect tipsters to perform particularly poorly when it comes to forecasting exact scorelines when compared with statistical models.

Goddard (2005) has investigated whether statistical models of goal arrival, or more direct methods of estimating match outcomes, are more effective when forecasting. 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. They also 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. Boshnakov et al. (2017) introduced a bivariate Weibull count model to this topic, which they documented improves upon the bivariate Poisson 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 betting strategy, looking at both result outcomes and whether more than 2.5 goals were scored in a match.

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.

3 Data

We extract data from several sources. Our attention is focused on the EPL for a number of reasons. It is widely regarded as the foremost domestic club competition globally.[7] Practically, the EPL is the league for which the widest range of forecasts is available. Across all the sources we consider here, these data cover forecasts for all 380 matches played in each of the 2016/17 and 2017/18 EPL seasons. We extract the data on outcomes of these football matches from *Soccerbase.com*. Table 1 presents the distribution of scorelines across the two

[7]It is a derivative of the Football League, the first football league competition founded in 1888, and 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 (£2.9nb) and Germany's Bundesliga (£2.8nb) (see 2018 Deloitte Annual Review of Football Finance; www2.deloitte.com/uk/....

seasons. The left panel is the 2016/17 season and the right panel is the 2017/18 season. There were 33 different 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 2017/18 than 2016/17, and fewer goals, but these differences between seasons are generally not statistically significant.

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
	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
Home goals	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*

The right panel of Table 2 displays the distribution of results in 2016/17 and 2017/18, showing 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

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 using 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. 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

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*

25% of the time (implying an overround, i.e. the expected profit margin of bookmakers is about 3%, which we address later). 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.[8] The scoreline implied probabilities indicate a sizeable overround, with the majority of implied probabilities being comfortably 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 have 50 anonymous users, or tipsters, from the online *Superbru Premier League Predictor Game*, selected from the 2016/17 season, and 150 from the 2017/18 season. These samples were selected in different ways. 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.[9]

Online Appendix Table B2 provides the distribution of scoreline picks by the tipsters over the two seasons. Tipsters under-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. Tipsters were over-conservative in this sense, however, as they predicted about 75% of matches to lie within this range of outcomes, when in reality only about 66% of matches finished this way.

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 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.[10] 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.

[8] 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.

[9] Brown and Reade (2019) consider tipsters from *Oddsportal.com* when looking at the wisdom of crowds. As here we are interested in scorelines, and these attract relatively few tipsters, we do not consider this information, despite collecting bookmaker data from *Oddsportal.com*.

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

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 over-picked 1-1 draws.

4 Methodology

In addition to those described above, we generate a set of probabilistic 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, also giving some further details on the range of other forecasts we will evaluate. In Sections 4.2–4.4 we discuss the various evaluation methods we will employ.

4.1 Candidate forecasts

4.1.1 Statistical model

To generate forecasts, we first estimate goal arrival 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. 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. 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). 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 strengths as they varied throughout the season (Elo, 1978), based on the historical results for all teams.[11] 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 up to each round of matches in each season, and the estimated parameters 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 within matches by teams. A scoreline point forecast comparable to that provided by the tipsters and experts can then be generated.

We generate scoreline forecasts in a number of ways. First, we simply use what the statistical model outputs as the most likely scoreline as the pick, what we call *Unconditional* forecasts. Alternatively we condition the scoreline pick on the most likely forecast result outcome. In the latter case, if all the probabilities of home win scorelines sum to a larger number than all the probabilities of draw or away win scorelines, 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 (occurring in about 11% of all match outcomes (both in our sample and more generally), and yet the most likely outcome is a home win (around 45% of all match outcomes).

Because of the large number of possible scorelines, many more are for wins rather than draws, and as such, it is very infrequent that a draw is the most likely forecast outcome. In a way to address this, we also develop *Fuzzy Conditional* forecasts. These return a draw prediction if the three result outcome probabilities are sufficiently close to one another. Hence, 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

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

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, the entropy measure achieves its highest value, 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, we return a draw prediction. 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.

4.1.2 Bookmakers

While bookmakers exist to profit maximise, it nonetheless remains the case that to do this they must forecast future events sufficiently well. We thus include a range of bookmaker forecasts in our comparison. Bookmakers set the prices of bets, and bettors can choose whether to accept the available price to back an event. We consider the decimal odds set by a bookmaker, d. 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. We also include information from betting exchanges, or prediction markets, where prices respond to individuals seeking to place bets on outcomes, and bettors can both *back* events to occur, as with bookmakers, and *lay* them, essentially acting as the bookmaker themselves. The implied probability of a given decimal odd set is $p = 1/d$.[12] In reality however, 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 overrorund when interpreting odds as implied probabilities (see for a smummary Štrumbelj, 2014). For the analysis which follows, we use the most simple of these corrections, normalising by dividing the quoted odds on each outcome through by the booksum, the sum of the odds offered for the various possible outcomes on some event, for example over all possible scorelines.[13]

4.1.3 Tipsters

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.

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

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

4.1.4 Experts

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 football team managers in the English Football Leauge). 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.

4.2 Forecast evaluation and comparison

Forecasting scorelines is interesting along a number of dimensions. First, as already discussed, these are low probability events. The difficulty of the 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 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. Second, any match has a number of outcomes and sub-outcomes that can matter:

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. Denote for some match i the result 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 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 is commonly 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;
$t_i = t(\mathbf{s}_i) = h_i + a_i$.

Generally, the 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:

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

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 much more challenging. As such, any reasonable scoring rule would have $A_x < B_x$ in order to encourage effort at the more challenging task.[14]

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 for a given weekend, drawn from any division, potentially, but typically the EPL and one or two from the second tier of English professional football. 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 forecasts 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 ($A_x = C_x = 0$). Simple result and scoreline percentages can be constructed, whereby a forecaster gets a point for either a correct result or scoreline. Scoring rules could

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

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.

Evaluating scoreline forecasts according to betting prices is arguably the most natural method for evaluating forecasts. In this case, actions are based on forecasts, actions that have different payoffs in different states of nature. Such payoffs can be found using betting prices. 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. If d_i are the decimal odds in match i for the scoreline consistent with the forecast $\widehat{s}_i$, then the return from a £1 bet on that event outcome would be:

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

Throughout our analysis, in the case of scorelines, we use the mean of the bookmaker odds we 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 point forecast would be to place bets 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 total £1 of bets on match i might be placed to maximise in expectation:

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

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. The total bet on a match may be larger than £1, but 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).

4.3 Forecast efficiency and encompassing

A more traditional (statistical) scoring rule for forecasts, particularly probabilistic 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 probabilistic forecasts, such as those produced by our statistical model, or by bookmakers. Neither Lawrenson, Merson nor the individual *Superbru* game players produce probabilistic forecasts; they pick a scoreline for each match. If we consider the unit of observation to be match scorelines, then given that these are countably infinite, 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 these past predictions.

Forecast tests in the form of Mincer and Zarnowitz (1969) have previously been applied in this context by Forrest and Simmons (2000) when looking at newspaper tipsters. However, Forrest and Simmons (2000) considered result outcomes, of which there are only three, and as such the method appears more appropriate for their focus rather than our generalisation to scorelines. Nonetheless we can regress outcomes, a binary or categorical variable in this case, 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, 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:

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

where $\mathbf{z}_i$ is a vector of potentially other important variables for explaining the outcome y_{ij}, and ε_{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 run 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 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 to be 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, deriving all information included from historical results between all teams in the league using Elo (1978) ratings.

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, in this way 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 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.

4.4 Scoreline forecast evaluation in summary

We look at forecasts from a range of methods. We present a statistical model that generates probabilistic forecasts, but from which point forecasts can be constructed. We also present the point forecasts from two experts and from random samples of tipsters. We consider a

range of methods for evaluating such forecasts, ranging from common statistical methods, through more sophisticated statistical measures tailored to scoreline evaluation, to returns from betting strategies. The important distinction is between point forecasts, or picks, and density, or probabilistic, 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 towards lesser and easier outcomes, such as the result or the total number of goals scored.

5 Results

5.1 Evaluating scoreline point forecasts

Considering the range of evaluation metrics discussed in Section 4.2, namely Brier scores, the scoring metrics 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 and 4 present the output from applying these various forecast evaluation metrics to each EPL season. The top panels present summary statistics for the Superbru tipsters (50 in 2016/17 and 150 in 2017/18), while the bottom panels present the actual statistics for the other five sources of forecasts (two experts and three model generated).

TABLE 3: Scoreline point forecast evaluation using different metrics for the 2016/17 EPL season

	Scoring Rule								Returns			
	Brier (1)	Scoreline (2)	Result (3)	Close (4)	BBC (5)	Sky (6)	Sbru (7)	Penalty (8)	Result (9)	Scoreline (10)	Margin (11)	Total (12)
Sbru: (N = 50)												
Mean	0.91	33.1	183.1	168.3	282.5	232.8	294.3	-306.1	9.4	-81.5	-5.7	-13.1
Median	0.91	34.0	185.0	171.5	287.0	234.2	295.2	-307.4	8.2	-74.0	-6.9	-12.0
Minimum	0.88	19.0	139.0	118.0	196.0	167.5	209.5	-363.1	-41.8	-175.0	-38.4	-48.2
Maximum	0.95	44.0	236.0	222.0	344.0	288.5	373.5	-259.8	100.4	-17.1	16.3	12.6
St. dev.	0.02	5.8	21.4	21.6	33.3	26.6	35.1	26.4	27.8	39.7	12.2	12.6
Merson	0.98	40	220	181	340	280	376	-272.0	62.5	-15.9	17.1	-1.1
Lawrenson	0.98	44	215	213	347	281	387	-273.3	74.8	-41.2	39.2	-9.6
Unconditional	0.98	43	172	195	301	236.5	292.0	-319.5	-13.0	-68.8	-14.8	-7.3
Conditional	0.98	48	216	186	360	288	375.5	-317.2	48.3	-30.7	11.6	-1.6
Fuzzy Cond.	0.98	45	197	183	332	264.5	326.5	-320.5	5.1	-51.9	-5.0	-1.6

Notes: From left to right, "Brier" gives the MSFE as per Equation (6). "Scoreline" gives the number of correct scoreline picks (out of 380). "Result" gives the number of correct result picks (out of 380). "Close" gives the number 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. "Penalty" applies the metric of Foulley and Celeux (2018) (see Appendix A). Columns (9)-(12) give implied betting returns in £ over the whole season from placing £1 on each match, i.e. a total investment of £380 for either the result, scoreline, margin of difference and total number of goals in a game, consistent with the scoreline point forecast.

TABLE 4: Scoreline point forecast evaluation using different metrics for the 2017/18 EPL season

	Scoring Rule								Returns			
	Brier (1)	Scoreline (2)	Result (3)	Close (4)	BBC (5)	Sky (6)	Sbru (7)	Penalty (8)	Result (9)	Scoreline (10)	Margin (11)	Total (12)
Sbru: (N = 150)												
Mean	0.91	35.4	192.7	181.7	298.9	245.8	308.4	-339.2	9.5	-94.8	-14.2	-22.3
Median	0.91	35.0	193.0	182.5	296.0	244.8	306.5	-333.9	9.1	-92.6	-13.5	-21.7
Minimum	0.87	21.0	135.0	139.0	201.0	168.0	208.0	-539.7	-56.8	-194.7	-43.3	-63.7
Maximum	0.94	53.0	280.0	293.0	439.0	359.5	451.5	-299.1	68.1	3.90	14.2	11.9
Variance	0.02	6.2	13.6	18.7	27.1	19.4	24.1	29.5	23.8	41.2	12.5	11.8
Merson	0.98	29	197	172	284	240.5	316.5	-314.3	5.6	-126.6	-3.7	-18.7
Lawrenson	0.98	48	192	190	336	264	359	-307.8	16.8	-31.5	-9.5	-14.0
Unconditional	0.98	40	183	193	303	243	303	-324.3	18.2	-98.4	-25.6	-24.7
Conditional	0.98	40	198	189	318	258	341.5	-334.2	-0.8	-101.1	-24.7	-28.9
Fuzzy Cond.	0.98	44	194	188	326	260	325	-324.8	5.0	-69.9	-26.4	-28.1

Notes: see Table 3.

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, on the other hand, 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) in each table 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 total number 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 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 improved as a crowd by all measures, which may in part reflect that tipsters in 2017/18 picked every outcome.[15]

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

Column (3) gives the total number of correct result picks. With the exception of the unconditional model picks, all forecasters picked more results than the average tipster, although none of the model generated or expert forecasters picked as many correct results as the best tipster. 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 total number 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 Appenidix 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 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, whereas with regards the broadcaster rules the Conditional model was best. The distinction between the *Superbru* and the broadcaster rules is that the *Superbru* rule factors in closeness. Similar to *Superbru*, the Foulley and Celeux (2018) penalty Rule rewards closeness, although in a more continuous manner. Here, the model-generated picks all incurred a larger cumulative penalty than the average tipster in 2016/17, although not in 2017/18, again almost certainly reflecting the greater number of picks made by each tipster in 2017/18 relative to 2016/17.

The final four columns of Tables 3 and 4 consider cumulative absolute betting returns. These returns are derived by assuming the forecaster used the scoreline point forecast for each of the 380 matches in a season to place a £1 bet on the correct result, correct scoreline, and on the margin being equal or greater than that implied by the predicted scoreline, and on the total goals being equal or greater than that implied by the predicted scoreline. In general, betting on results can generate positive returns (assuming the forecaster 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 in almost every case would have generated a negative return, and the same is true

for betting on total goals scored in matches. In 2016/17 betting on the margin of difference between teams would have been more successful than in 2017/18.

When considering these scoring rules more broadly, the relative rankings of the different forecasts are particularly informative. In Tables 5 and 6 the rankings of the experts and model-generated forecasts are presented for each season. The individual tipster rankings are implicit in these tables.

TABLE 5: Selected ranks out of 55 of forecasters according to different scoring rules in the 2016/17 EPL season

	Scoring Rule								Returns			
	Brier (1)	Scoreline (2)	Result (3)	Close (4)	BBC (5)	Sky (6)	Superbru (7)	Penalty (8)	Result (9)	Scoreline (10)	Margin (11)	Total (12)
Merson	5.00	14.00	3.00	20.00	5.00	4.00	2.00	7.00	4.00	1.00	2.00	9.00
Lawro	3.00	5.00	7.00	2.00	2.00	3.00	1.00	8.00	2.00	11.00	1.00	23.00
Unconditional	4.00	8.50	52.00	9.00	17.50	29.00	33.50	39.00	44.00	27.00	48.00	20.00
Conditional	1.00	1.00	6.00	16.00	1.00	2.00	3.00	35.00	6.00	6.00	7.00	12.00
Fuzzy Cond.	2.00	3.00	25.00	19.00	8.00	11.00	12.00	40.00	30.00	18.00	27.00	11.00

Notes: see Table 3.

TABLE 6: Selected ranks out of 155 of forecasters according to different scoring rules in the 2017/18 EPL season

	Scoring Rule								Returns			
	Brier (1)	Scoreline (2)	Result (3)	Close (4)	BBC (5)	Sky (6)	Superbru (7)	Penalty (8)	Result (9)	Scoreline (10)	Margin (11)	Total (12)
Merson	5.00	130.50	22.00	98.00	114.00	100.00	45.00	22.00	85.00	123.00	35.00	60.00
Lawro	1.00	3.00	65.00	29.00	16.00	22.00	3.00	7.00	56.00	11.00	57.00	38.00
Unconditional	3.50	29.50	107.50	22.00	57.50	88.50	97.50	49.00	48.00	89.00	120.00	89.00
Conditional	3.50	29.50	16.50	38.00	31.50	29.50	15.00	81.00	104.00	93.00	118.00	111.00
Fuzzy Cond.	2.00	11.50	50.50	43.00	24.00	27.50	29.50	50.00	88.00	45.00	125.00	105.00

Notes: see Table 3.

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 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 conduct 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 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.

5.2.1 Scorelines

Table 7 presents the outcomes from the regressions investigating the weak efficiency of scoreline forecasts, with a column for each forecast type. From Equation (7), the null hypothesis of weak efficiency is that the intercept coefficient $\alpha = 0$ and the slope coefficient $\beta = 1$. We present an F-test of that hypothesis in the final row of the table, and in each case the null hypothesis of weak efficiency is heavily rejected at standard levels. The 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.[16] 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 the experts placed zero weight on every possible scoreline other than the one they picked.

TABLE 7: Weak efficiency tests for forecast scoreline outcomes

	Model	Bookmaker	Lawrenson	Merson	Tipster
	(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 (df = 61558)	0.107	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.

Table 8 presents the outcomes from regressions evaluating strong efficiency for scorelines as per Equation (8). Variables are added to $\mathbf{z}_i$ for the number of league points the home

[16] It is worth noting that the β coefficient on the Bookmaker regression is greater than one, which is indicative of the well-known favourite-longshot bias, hence documenting its existence among football match scorelines, as opposed to among results, where it is more typically described.

team has, the difference between the home and away team league points, for 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 estimated α and β coefficients are essentially identical between the two sets of results considered so far, as are the adjusted R^2 coefficients, suggesting that readily available (and obvious) information for football match outcomes has already been factored into scoreline forecasts. As before, the bottom row of Table 8 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.

Table 9 presents the outcomes from regressions investigating the bias in the scoreline forecasts as per Equation (9), once more presenting a column for each forecast source. The significance of the constant term implies significant bias in the forecasts. For all forecasts, the constant term is insignificant, suggesting unbiased forecasts for scorelines.

5.2.2 Result outcomes

We consider (implied) probability forecasts in each match of home and away wins and draws for the Model, Bookmakers and Tipster forecasts. For the Model and Tipster forecasts we sum up all the scoreline forecast probabilities corresponding to each result outcome. For the individual experts we consider a binary variable for whether or not either expert picked that result outcome. In Online Appendix Table B4, we present the weak efficiency output for these result picks from the five sources of forecasts. This table 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 find that the null hypothesis of weak efficiency for the Model and Bookmaker forecasts is never rejected, whereas it is always rejected for the experts, and in the case of Tipsters only for the draw is the null of weak efficiency not rejected.

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. These coefficients are not dissimilar to those found by Reade (2014) when considering tipster forecasts from *Oddsportal.com*.

In Online Appendix Table B5 we present the strong efficiency regression output for result picks, estimating equivalent regression models as before with the scoreline picks. The

TABLE 8: Strong efficiency tests for forecast scoreline outcomes

	Model	Bookmaker	Lawrenson	Merson	Tipster
	(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 (df = 61553)	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.

TABLE 9: Bias regression tests for scoreline outcomes

	Model	Bookmaker	Lawrenson	Merson	Tipster
	(1)	(2)	(3)	(4)	(5)
Constant	0.00001	0.0002	0.0001	0.00005	0.0004
	(0.0004)	(0.0004)	(0.001)	(0.001)	(0.0004)
Observations	61,560	61,560	61,560	61,560	61,560
Adjusted R^2	0.000	0.000	0.000	0.000	0.000
Resid. Std. Error (df = 61559)	0.108	0.107	0.147	0.150	0.111

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

difference here compared with the weak efficiency tests is the addition of the Elo prediction as an explanatory variable in the regressions, representing the readily available information at the time of the match forecasts. For the draw outcome we take the squared difference of the Elo prediction from 0.5, referring to this as a "Balance" measure.[17] The Elo prediction is significant for all away win forecasts, and for all home win forecasts except for in the case of the Bookmaker forecasts. The Elo Balance measure is significant for the Model and both expert forecasts only. 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 cases at standard levels, and for the Tipster forecasts in the case of the draw outcome.

In Online Appendix Table B6 we present the forecasting bias regression outputs for the result outcomes. We again run regressions for each outcome (home win, draw, away win) and for each forecast method. We find slight though insignificant evidence for bookmakers under-predicting the likelihood of a home win. There is evidence that the two experts significantly over-predict home wins relative to their actual frequency, and the Tipster crowd under-predicts quite strongly the likelihood of a home win. With the exception of Merson, who under-predicts draws, there is no other significant bias found in the forecasts of draw outcomes. The two experts complement their over-prediction of home wins with an under-prediction of away wins, although Merson is insignificantly over-predicting. Similarly, Superbru tipsters complement their under-prediction of home wins with a strong over-prediction of away wins relative to how often they actually occur.

5.3 Forecast encompassing

Turning to the outcomes of encompassing regressions in Table 10, we present the t-statistics for the equivalent of the ϕ_a and ϕ_b coefficients 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 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-based forecasts discussed in the previous section.

The encompassing regression results in Table 10 are presented such that the row is the particular forecast error in the regression (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 can't

[17] 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.

TABLE 10: 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 4.

enter the Model probability forecast into the Model probability forecast error regression model. The bold faced numbers in the the table indicate t-statistics that are very significant, namely four or larger. In the notation and definition of encompassing described before by Equations (10)-(11), the t-statistics reading from right to left in the table for a particular source of forecasts (row) give the values of ϕ_a for all the other potential sources (column). Similarly, then reading down the column in the table corresponding to that same particular source one can look up the corresponding value of ϕ_b. 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 to begin with, the measured t-statistic of 9.7 ($\widehat{\phi_a}$) from adding the model unconditional scoreline picks to the forecast error regression model suggests that the former adds information to that provided by the Model's own forecast probabilities of each scoreline. The corresponding opposite entry, regressing the forecast error for Model point forecasts on the Model probability forecasts, is shown in the first column and second row. In this case the t-statistic is -51.0 ($\widehat{\phi_b}$), suggesting that the Model probability forecasts are able to explain the variation in Model point forecast errors. Taken together, this suggests that the probabilistic forecasts add much more information than do scoreline point forecasts, as would be expected. This occurs for all three types of Model point forecast considered. However, overall the top right section of Table 10, where the four different types of forecasts generated by the statistical model are presented, 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 they encompass one another.

The Model probabilities encompass all other sources of forecasts except those by Lawrenson. A pick by Lawrenson increases the forecast error significantly, although for Merson the effect is also borderline significant (t-statistic of 2.2). This suggests we cannot claim that the statistical model and it's more sophisticated probabiltiy-based forecasts are *better* than the expert Lawrenson and his point forecasts. Instead, we would conclude

that these two sources of forecasts are complementary, despite being of a different type. However, we can argue that the statistical model dominates the Bookmaker and Tipster implied probability forecasts, and is in a sense *better*.

We find that the Bookmaker forecasts encompass Merson forecasts, but not Tipster or Lawrenson forecasts. The crowd-based tipster 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 the forecasts of scorelines in 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 many standard conventional statistical approaches, such as efficiency and bias testing, but probability forecasts cannot be used when applying scoring rules designed for point forecasts. We found evidence that both tipsters and experts make biased forecasts in terms of football match result outcomes, although not necessarily in terms of scoreline outcomes, emphasising the complexity of evaluating the various sources of football match 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 the most effective when attempting to predict the outcome of football matches. There was also evidence from these encompassing regression results 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? And 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 his objective. Similarly, without information on the forecaster's risk preferences it is not clear this is the best approach. Nevertheless, once point and probability forecasts of the same events were evaluated in this way, it would be easy to compare the two different types based on the common metric of implied financial return, 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.

This study focuses on a very particular context. But the simple fact is that this is a context where many people explicitly make forecasts, and truly care about the outcomes of the events and those forecasts, far beyond any financial gain they might achieve from them. This justifies our attention and yet further research. The forecasts are also interesting in terms of containing sub-outcomes, which are as equally interesting or even more so than the main outcome being forecast by a scoreline pick. Conditioning on these sub-outcomes might frequently suggest a different point forecast altogether. This is a feature caused by the draw being an acceptable, common and final outcome of football matches, which is unusual in other areas of society or even generally within professional sports. Although other sports feature draws, they are uncommon as a final outcome, with one exception being first class cricket. On reflection, it is the prevalence of draws among the final outcomes of professional football matches which makes forecasting them strange.

References

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.

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

Clements, M. 2004. "Evaluating the Bank of England Density Forecasts of Inflation." *Economic Journal*, 114(498): 844–866.

Clements, M. 2006. "Evaluating the Survey of Professional Forecasters probability distributions of expected inflation based on derived event probability forecasts." *Empirical Economics*, 31(1): 49–64.

Clements, M. 2009. "Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters." In *The methodology and practice of econometrics. A festschrift in honour of David F. Hendry.* Eds. by J. Castle, and N. Shephard 206–226 Oxford University Press Oxford.

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

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

Fair, R. C., and R. J. Shiller. 1989. "The informational content of ex ante forecasts." *The Review of Economics and Statistics* 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., 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*, 16 317–331.

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.

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): .

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.

Maher, M. J. 1982. "Modelling association football scores." *Statist. Neerland.*, 36(3): 109–118.

Makridakis, S., and M. Hibon. 2000. "The M3-Competition: results, conclusions and implications." *International Journal of Forecasting*, 16(4): 451–476.

Martinez, A. 2018. "A False Sense of Security: The Impact of Forecast Uncertainty on Hurricane Damages." *University of Oxford, Department of Economics Discussion Paper*, 831.

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

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.

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.

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.

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.

Appendix A. Measures of 'closeness'

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 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.

Evaluating Strange Forecasts: The Curious Case of Football Match Scorelines

Online Appendix

J. James Reade Carl Singleton Alasdair Brown[†]

January 2019

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 Away goals							2017–18 Away goals						
		0	1	2	3	4	5	6	0	1	2	3	4	5	6
Home goals	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
	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*

[†]j.j.reade@reading.ac.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.

This study is based on data obtained from and analysed with the permission of Superbru, Sport Engage Ltd. Throughout the study, the anonymity of individual 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.

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
Lawrenson	Home goals	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
		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
Merson	Home goals	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
		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: Weak efficiency tests for forecast result outcomes (home win, draw, away win)

	Model	Bookmaker	Lawrenson	Merson	Tipster
	(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.

TABLE B5: Strong efficiency tests for forecast result outcomes (home win, draw, away win)

	Model	Bookmaker	Lawrenson	Merson	Tipster
	(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.

TABLE B6: Bias regression tests for result outcomes (home win, draw, away win)

	Model	Bookmaker	Lawrenson	Merson	Tipster
	(1)	(2)	(3)	(4)	(5)
Home win	0.027	0.024	-0.038^{*}	-0.100^{***}	0.152^{***}
	(0.017)	(0.016)	(0.021)	(0.021)	(0.025)
Residual std. error	0.471 (df = 759)	0.454 (df = 758)	0.589 (df = 755)	0.571 (df = 756)	0.689 (df = 759)
Draw	−0.019	0.0005	−0.033	0.078^{***}	0.019
	(0.016)	(0.015)	(0.021)	(0.020)	(0.016)
Residual std. error	0.428 (df = 759)	0.424 (df = 758)	0.576 (df = 755)	0.547 (df = 756)	0.439 (df = 759)
Away win	−0.008	−0.019	0.071^{***}	0.022	-0.171^{***}
	(0.015)	(0.015)	(0.018)	(0.018)	(0.024)
Resid. std. error	0.419 (df = 759)	0.400 (df = 758)	0.488 (df = 755)	0.507 (df = 756)	0.663 (df = 759)
Observations	760	759	756	757	760

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