Evaluating Real-Time Probabilistic Forecasts With Application to National Basketball Association Outcome Prediction

Chi-Kuang Yeh\textsuperscript{a}, Gregory Rice\textsuperscript{a}, and Joel A. Dubin\textsuperscript{a,b}

\textsuperscript{a}Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada; \textsuperscript{b}School of Public Health Sciences, University of Waterloo, Waterloo, ON, Canada

**ABSTRACT**

Motivated by the goal of evaluating real-time forecasts of home team win probabilities in the National Basketball Association, we develop new tools for measuring the quality of continuously updated probabilistic forecasts. This includes introducing calibration surface plots, and simple graphical summaries of them, to evaluate at a glance whether a given continuously updated probability forecasting method is well-calibrated, as well as developing statistical tests and graphical tools to evaluate the skill, or relative performance, of two competing continuously updated forecasting methods. These tools are demonstrated in an application to evaluate the continuously updated forecasts published by the United States-based multinational sports network ESPN on its principle webpage espn.com. This application lends statistical evidence that the forecasts published there are well-calibrated, and exhibit improved skill over several naïve models, but do not demonstrate significantly improved skill over simple logistic regression models based solely on a measurement of each team’s relative strength, and the evolving score difference throughout the game.

**1. Introduction**

Probabilistic predictions and forecasts are ubiquitous in modern society, and many individuals consider, and make decisions based on, such forecasts on a routine basis. For example, in the U.S. probability of precipitation forecasts became widely publicly available starting in the late 1960s, and are now a critical factor in countless people’s daily decisions (Murphy 1998; Council 2006). Over time the number and scope of probabilistic forecasts readily accessible to the public has increased at a steady pace, and now covers prediction of phenomena ranging from sports (Silver, Boice, and Paine 2019), to politics (Erikson and Wlezien 2012), to medicine (Spiegelhalter 1986), to geology (Gomberg et al. 2015), among many other, some more exotic (Rowe and Beard 2018), areas.

Many such forecasts are made initially well before the event in question occurs, and are then continuously updated as new information becomes available. The example that we focus on throughout this article is basketball game outcome prediction in the National Basketball Association (NBA). Websites like espn.com, the main web page of the United States-based multinational sports network, ESPN, publish and update, in real time, probabilistic forecasts of the home team winning for each NBA game played. Although the method by which ESPN produces these forecasts is largely proprietary, ostensibly initial probability forecasts of the home team winning are constructed based on information that is available before the game starts, for example, the usual home court advantage in the NBA, relative team strength, player injuries, etc., and then after the game commences and progresses these forecasts are updated with new information such as the score, game time remaining, ball possession, fouls, and in-game player injuries. The resulting probabilistic forecasts and their fluctuations may be viewed as a curve that is a function of the in-game time; see Figure 1 for an example. Such curves arise in any similar continuously updated probabilistic forecasting task, and are evidently not unique to basketball game outcome prediction; see, for example, Silver (2020).

Natural questions to ask when faced with any probabilistic forecast, including those that are continuously updated, are “are these forecasts accurate?” and “could these forecasts be improved upon?” Following the seminal work of Murphy and Winkler (1987, 1992) on evaluating the quality of probabilistic forecasts in meteorology, evaluating a method for producing probabilistic forecasts is often broken into the tasks of measuring its calibration and skill. A model is deemed well-calibrated if its forecasts are compatible with the observed outcomes. In other words, a model that predicts an outcome with a given probability is well-calibrated if the relative frequency that the outcome occurs matches the probability forecast in the long run. A model is deemed to have higher skill than a competing model if its predictions are “sharper” or “more concentrated” than its competitor. For example, a model that forecasts the probability of a rainy day in New York City on a given day with the long-run background rate of rainy days (which happens to be about 33.1% for New York City), will in the long run be well-calibrated, but has less skill than a model, perhaps based on

**CONTACT**  
Chi-Kuang Yeh \texttt{chi-kuang.yeh@uwaterloo.ca}  
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada.

**Supplementary materials** for this article are available online. Please go to www.tandfonline.com/r/TAS.

© 2021 American Statistical Association
more complete weather data, that correctly predicts rainy days with probabilistic forecasts of zero and one. Excellent reviews and more in-depth discussions of these concepts can be found in Gneiting, Balabdaoui, and Raftery (2007) and Gneiting and Katzfuss (2014).

The goal of this article is to develop simple, easily interpretable, tools for evaluating the calibration and skill of continuously updated probabilistic forecasts, and to apply these methods to evaluate the probabilistic forecasts pertaining to NBA basketball game outcome prediction published on ESPN. In terms of evaluating model calibration of probabilistic forecasts, standard tools are reliability diagrams and calibration plots, in which outcome frequencies are plotted against binned forecast probabilities; see Gneiting, Balabdaoui, and Raftery (2007). Below we show how such curves can be extended in the continuously updated case to calibration surfaces, and how such surfaces can be summarized to show at a glance whether a given method is well-calibrated. In order to evaluate the relative skill of one continuously updated forecasting model against another, we employ the method of Lai, Gross, and Shen (2011) to construct confidence intervals for the average loss difference measured by the Brier score (Brier 1950) between the two models at given time points throughout the updating process. Estimating these intervals pointwise can be used to construct a simple graphical summary of the relative skill of one model versus another. In order to measure the cumulative statistical significance of differences observed in such a plot, we develop a new significance test for the skill differences aggregated across time based on a novel large sample result for estimating continuous loss difference curves.

For the purpose of demonstrating these methods and evaluating ESPN’s forecasts, we introduce a number of “competing” continuously updated forecasting methods for basketball outcome prediction. Some are designed to be “straw men” for the purpose of demonstration, whereas others are based on logistic or probit generalized linear models making use of in-game information such as the score difference. We show using the proposed methods that ESPN’s model is generally well-calibrated, and exhibits significantly better skill than some naïve models, although it does not demonstrate superiority over relatively simple logistic regression models based on the score difference and relative team strength alone.

The rest of the article is organized as follows. Section 2 introduces the details of the ESPN forecasting data that we consider, as well as some competing forecasting methods that we develop and use for the purpose of comparison. Section 3 discusses the construction of calibration surfaces for such forecasts, as well as simple graphical summaries of these surfaces. Section 4 explains the proposed methods to evaluate the relative skill of two sets of real-time probabilistic forecasts. A Monte Carlo simulation study of these methods is given in the supplementary material (Section S1). A detailed comparison of the ESPN forecasts as well as those of the proposed models is given in Section 5. Technical details are provided in the supplementary material (Section S2).

2. Motivating Data and Competing Models

The specific data that we consider are play-by-play records and real-time probabilistic forecasts of NBA regular season games downloaded from espn.com/nba (ESPN 2020). The NBA is a major professional basketball league, which is often referred to as one of the “Big Four” professional sports leagues in North America. Since 2004, except for the lockout season in 2011 and the COVID-19-influenced seasons in 2020 and 2021, the NBA is comprised of 30 teams, with each team playing a schedule of 82 games in the regular season.

Starting in the 2017–2018 NBA season, ESPN Analytics began providing real-time in-game probabilistic forecasts of the home team winning each NBA game played; an example of the forecasts from one game is shown in Figure 1.
data available from ESPN are quite rich, including real-time information about details such as substitutions, fouls, and ball possession. We consider here only a subset of these data that includes the real-time probabilistic forecasts provided by ESPN, as well as the evolution of the score throughout the game, for the 2017–2018 and 2018–2019 seasons. These data are updated each time there is an “event” in the game, which includes primarily score changes, fouls, and changes of possession. A typical game features between 460 and 480 events.

We excluded a small portion of these data from our analysis due to two issues. Quite often, multiple events will occur at the same instant in a game. One of the main examples that contributes to this is multiple players substituting at the same time. Although these events are all logged at the same time point, they occur in the dataset in an ordered sequence. The forecasted probabilities published by ESPN during such an event are typically contingent on this order. Therefore, we simply average the forecasts together in such a scenario to produce a probabilistic forecast at that instant. We also tried a number of other ways to handle this situation, such as using the first or last probabilistic forecast among the events recorded, and the difference in the results was negligible.

The second issue is due to games that go to overtime. If two teams’ scores are tied at the end of the 48-min regulation game time, the teams will play an extra 5-min overtime period. For such games, we remove the overtime period from the analysis, and only consider probabilistic forecasts up to the end of the game so that they are comparable to those that arise from games that did not go to overtime. Overtime games represent slightly less than 10% of the total games. Additionally, a small number of data points were discarded due to evident defects or excessive missing values.

The remaining data that we analyze are summarized in Table 1, and in each season there are more than 1100 games with a total of over 350,000 play-by-play records available. Below we use the data from the 2017 to 2018 season as training data for our own models, and then we produce and evaluate forecasts for the 2018–2019 season.

Letting $N$ denote the total number of games with forecasts that we wish to evaluate, so $N = 1213$ when we consider the 2018 and 2019 forecasts, the data may then be denoted as $\tilde{P}^{ESPN}_i$, $1 \leq i \leq N, t \in [0,1]$ representing the probabilistic forecasts of the home team winning in the $i$th game at intragame time $t$. We assume that the game time parameter $t$ is normalized to be between zero and one so that it represents the proportion of the game complete. Although these forecasts are only available when events occur, due to the fact that events are very dense throughout the game, we complete these forecasts as a piece-wise constant function set to the last probability forecast in-between events to produce full probability forecast curves over the interval $[0,1]$, which also makes them more comparable from one game to the next. This is illustrated in Figure 1.

We also consider the data $H_i(t)$ and $A_i(t)$, $1 \leq i \leq N, t \in [0,1]$, denoting the home team score and away team score in the $i$th game, respectively, at proportion $t$ of the game. In our analysis below, we frequently make use of the score difference $S_i(t) = H_i(t) - A_i(t), 1 \leq i \leq N, t \in [0,1]$.

The goal of the methods that we develop below is to evaluate the quality of the forecasts $\hat{P}^{ESPN}_i$, $1 \leq i \leq N, t \in [0,1]$. To do this we also develop a number of benchmark models that are used for the purpose of comparison.

### 2.1. Benchmark Models for Predicting NBA Game Outcomes

We use the following notation below. Let $Y_i$ denote the indicator random variable that the home team wins in the $i$th game, so that $Y_i = 1$ if the home team wins the $i$th game, and $Y_i = 0$ if the home team loses the $i$th game. We are interested in forecasting or estimating the probability $p_i(t)$ that the home team wins, given the information up to time $t$ in the game, so that

$$p_i(t) = P(Y_i = 1 | \text{all information up to time } t \text{ in game } i).$$

A more formal definition of $p_i(t)$ is given in Section S2 in the supplementary material. $\hat{P}^{ESPN}_i(t)$ is in principle an estimate (forecast) of $p_i(t)$. In order to evaluate the quality of these forecasts, we consider a number of competing benchmark models, progressing from naïve to more realistic. The benchmark models that we consider are for the most part generalized linear models (GLMs) for binary response data, which are often termed logistic regression models. Using $g(\cdot)$ to denote the GLM link function, which we take to be the logit link, see, for example, McCullagh and Nelder (1989, chap. 4), each of our models are of the form

$$g(p_i(t)) = \beta_0(t) + \sum_{j=1}^d \beta_j(t)X_j(t),$$

where the terms $X_j(t)$ denote covariates that are used to predict $p_i(t)$. All GLMs were fit pointwise over a discrete grid of 721 equally spaced time points $t$, which corresponds to a 4 sec in-game resolution, using the R programming language, specifically the glm function in the stats package, version 4.0.2. For each model we used the 2017 and 2018 season data as training

---

**Table 1.** Summary of the data obtained from ESPN (ESPN 2020) from the 2017–2018 to 2018–2019 NBA regular seasons.

| Season | Mode     | Games | Events | Max. events | Min. events | Avg. # of events |
|--------|----------|-------|--------|-------------|-------------|------------------|
| 17–18  | Raw      | 1158  | 530,032| 606         | 234         | 457.7133         |
|        | Selected | 1137  | 517,983| 572         | 240         | 455.5699         |
|        | Processed| 1137  | 354,749| 375         | 173         | 312.0343         |
| 18–19  | Raw      | 1229  | 583,443| 700         | 124         | 474.7299         |
|        | Selected | 1213  | 572,346| 598         | 366         | 472.0082         |
|        | Processed| 1213  | 396,991| 385         | 241         | 327.2803         |

**Note:** Raw counts represent the total number of games that ESPN provides probability forecasts for. Selected refers to those games that do not contain errors or missing values. Processed represents the data after averaging out multiple events recorded at the same game time.
data, and then produced rolling forecasts on the 2018 and 2019 season data to compare to the ESPN forecasts.

The most complicated covariates that we consider to build these benchmark models are the score difference ScDi(t), and some measure of the relative strength of the teams, which we term RSi. We also consider the “leading status” covariate, which encodes whether or not the home team is winning at intra-game time \( t \). We define this as the covariate LS(t), which takes the value 1 if \( ScDi(t) > 0 \), \(-1\) if \( ScDi(t) < 0 \), and \(0\) if \( ScDi(t) = 0 \).

Regarding defining \( RSi \), there are a number of ways of evaluating the relative strength of teams, including using the Elo rating system (Elo 1978), which has been extensively used to rate the strength of basketball teams (see Silver 2014; Silver and Fischer-Baum 2015; Silver, Boice, and Paine 2019) and determine odds in betting markets. We instead use as a proxy of the relative team \( RS_i \) = \( \hat{p}_{ESPN}(0) \), the pregame probability of the home team winning as forecast by ESPN. We considered a number of alternate metrics to define \( RS_i \), and found that generally the results and conclusions of the below analyses did not change significantly, and so we use this quantity to avoid the need to introduce new metrics and data.

Descriptions of the basic benchmark models that we considered to forecast \( p_i(t) \) are collected in Table 2, and are listed in order from naïve to more realistic. We note that the GLMs with intercept terms are able to implicitly model the home team advantage, which refers to the phenomenon that in the NBA the home team tends to win a higher percentage of games than the away team, and so a model based on the score difference but without an intercept term, like ScDiNoInt in Table 2, could be expected to be poorly calibrated, at least at the beginning of the game. As mentioned, each GLM model is fit pointwise over the game-time parameter \( t \). This allows that, for example, the effect, as determined by the models, of relative team strength, home team advantage, and score difference can evolve throughout the game. Diagnostic plots of the pseudo \( R^2 \) (McFadden 1973) and relative variable importance over the course of the game of the variables in our “least naïve” model, which uses both the score difference and relative team strength and is termed PgRSScD, are displayed in Figure 2. It is clear from this figure that the models’ predictions improve as the game progresses, evidently since ultimately the score difference covariate determines the winner, and also that the relative importance of the score difference versus team strength changes inversely as the game progresses; relative team strength is the most important predictor early in the game, but becomes less important later in the game as the score difference becomes more informative.

### 3. Evaluating Calibration of Continuous Forecasts

We now turn to the task of evaluating the calibration for a given set of continuously updated forecasts \( \hat{p}_i(t) \) with realized outcomes \( Y_i \), \( i = 1, \ldots, N \). As mentioned in the introduction, traditionally when evaluating such forecasts one often considers what are called calibration plots or reliability diagrams. A calibration plot is a plot of binned probabilistic forecasts against the conditional event frequency associated with forecasts in a given bin. Since well-calibrated forecasts should have that the event frequency match the forecast probability, calibration may then be measured by comparing these points against a 45-degree diagonal reference line. Large departures from this line thus indicate poor calibration (cf. Dawid 1986; Murphy and

| Model | Covariates | Description |
|-------|------------|-------------|
| CF    | None       | A constant forecast for all intra-game times of 1/2, that is, \( \hat{p}_i(t) = 0.5 \) for all \( t \in [0, 1] \). Abbreviation is for Coin Flip. |
| HomeWP| None       | A constant forecast for all intra-game times set to the observed home team win rate in the prior 10 NBA regular season, spanning 2008–2017. This amount to forecasting \( \hat{p}_i(t) = 0.593 \) for all \( t \in [0, 1] \). |
| PgRS  | RS_i       | GLM for the home team win probability in terms of \textbf{Pre}game \textbf{Relative} team \textbf{Strength}, as measured by \( \hat{p}_{ESPN}(0) \). |
| LS    | LS_i(t)    | GLM for the home team win probability in terms of the \textbf{Leading} \textbf{Status} for the home team. |
| ScDiNoInt| ScDi(t) | GLM for the home team win probability in terms of the \textbf{Score} \textbf{Difference} between the home and away teams. \textbf{No Inter}cept term is included in the model. |
| ScD   | ScDi(t)    | GLM for the home team win probability in terms of the \textbf{Score} \textbf{Difference} between the home and away teams. |
| PgRSLS | RS_i, LS_i(t) | GLM for the home team win probability in terms of \textbf{P}regame \textbf{Relative} team \textbf{Strength} as measured by \( \hat{p}_{ESPN}(0) \), and the \textbf{Leading} \textbf{Status}. |
| PgRSScD| RS_i, ScDi(t) | GLM for the home team win probability in terms of \textbf{P}regame \textbf{Relative} team \textbf{Strength} as measured by \( \hat{p}_{ESPN}(0) \), and the \textbf{Score} \textbf{Difference}. |

![Figure 2. Left: the pseudo R² of the logistic regression model PgRSScD, which uses the covariates pre-game relative strength and score difference, as a function of the game time. Right: the variable importance of each covariate as it contributes to the pseudo R².](image-url)
Winkler 1992; Ranjan and Gneiting 2010). We refer the reader to Gneiting and Katzfuss (2014) and the references therein for a more comprehensive discussion.

One clear method then to check for calibration of continuously updated forecasts is to produce a calibration plot for each \( t \in [0, 1] \) based on the pairs \((\hat{p}_t(t), Y_t)\). While this is in essence what we propose, there are two main challenges in doing so. (i) Traditionally when producing calibration plots, the forecasted probabilities are binned into fixed bins, commonly by deciles. For example, often the event frequency corresponding to all forecasted probabilities between \([0, 0.1)\) are compared to 0.05, similarly for \([0.1, 0.2)\) to 0.15, and so on. With continuously updated forecasts, and as with the ESPN forecasts, it is typical that the forecasts fluctuate so that for certain time points \( t \) the forecasts cluster around some fixed values, and are hence far from being uniformly distributed in such fixed bins.

In the case of the ESPN forecasts, near the end of the game the majority of forecasts are clustered around 0 and 1. Fixed bins will often have the problem that the forecasts within them are not uniformly distributed within the bin. (ii) Having constructed calibration plots for each \( t \), one must examine a large number of such plots to pinpoint if a method appears to be well-calibrated, or to diagnose if there are a subset of times \( t \) at which the method is more or less calibrated than others. A simple summary of the many calibration plots produced would be useful.

In order to address (i), we propose to use adaptive bins in constructing the calibration plots. Specifically, for each \( t \), suppose we wish to construct \( M \) bins for the forecasts \( \hat{p}_t(t) \). By calculating the ranked forecasts \( \hat{p}_{(i)}(t), i = 1, \ldots, N \), we may group them into \( M \) bins so that \( \hat{p}_{(i)}(t) \) is in bin \( j \) if \((\lfloor N/M \rfloor (j - 1) + 1) \leq i < \lfloor N/M \rfloor j\). We denote the collection of \( \hat{p}_t(t) \)'s in the \( j \)th bin as \( \mbox{Bin}_j \). In simpler terms, the forecasts at a given time \( t \) are grouped into \( M \) bins based on their rank. As a reference point or summary of the forecasts in the \( j \)th bin, we use \( \hat{p}_t(t) = \text{Median}(\hat{p}_{(i)}(t)) \), \((\lfloor N/M \rfloor (j - 1) + 1) \leq i < \lfloor N/M \rfloor j\). A calibration plot at time \( t \) is constructed then by comparing \( \hat{p}_t(t) \) to \( \bar{Y}_j(t) = \text{Average}(Y_t \mid \hat{p}_t(t) \in \mbox{Bin}_j) \). Letting \( n_j \) denote the number of forecasts in Bin\(_j\), a 95% confidence interval for the mean of the events in Bin\(_j\) is constructed as

\[
\frac{n_j \bar{Y}_j(t) + \kappa^2/2}{n_j + \kappa^2} \pm \frac{\kappa n_j^{1/2}}{n_j + \kappa^2} \left( \bar{Y}_j(t)(1 - \bar{Y}_j(t)) + \kappa^2/(4n_j) \right)^{1/2},
\]

where \( \kappa = z_{\alpha/(2M)} \), and \( z_\beta \) denotes the \( \beta \) quantile of the standard normal distribution. Here, we have applied a Bonferroni correction to the significance level according to the number of bins used \( M \). \( \alpha \) is typically taken to be 5% so that 95% confidence intervals are computed. This interval is often referred to as the Wilson interval; see Brown, Cai, and DasGupta (2001), and Wilson (1927). We have noticed significant improvements by using the Wilson interval in this setting due to event frequencies that approach zero and one near the end of the game. Additionally, so that no bins contain predominantly zero or one probability forecasts, we discard all forecasts to produce a calibration plot at a given \( t \) and corresponding events, if \( \hat{p}_t(t) < 0.005 \) or \( \hat{p}_t(t) > 0.995 \), and we analyze those forecasts for calibration separately; see Table S2 in the supplementary material.

A calibration plot may then be made by plotting the bin references \( \hat{p}_t(t) \) against the above confidence intervals, and comparing these intervals to the reference line \( y = x \). The method is deemed well-calibrated at the given significance level if the reference line generally goes through each interval.

Before proceeding, we note that an interesting alternative that we implemented, with results presented in the supplementary material, is to choose the bins, both their locations and number, adaptively based on an isotonic regression as discussed in Dimitriadis, Gneiting, and Jordan (2020). In our application to the ESPN forecasts, this lead to similar results, and so we omit them here.

Although such surface plots are informative, it can be challenging to infer quickly based on these plots whether a given method appears to be calibrated. In order to produce a more
Figure 4. Plots of \( U_{0.95}^{\min}(t) \) and \( L_{0.95}^{\max}(t) \) against \( t \) based on \( M = 10 \) bins: Methods (a) historic home team win probability (HomeWP); (b) GLM using score difference without intercept (ScDnoInt); (c) ESPN forecasts; (d) GLM using pre-game relative strength and score difference (PgRSScD).

An easily interpretable summary of such calibration surface plots, we instead consider a plot for each \( t \) of the minimum distance over \( p \) between the reference plane \( f(t, p) = p \), and the upper and lower calibration surfaces. Specifically, we consider plots of the functions \( U_{1-\alpha}(t) = \min_{1 \leq j \leq M} U_{1-\alpha}(t, \hat{p}_j(t)) - \hat{p}_j(t) \) and \( L_{1-\alpha}(t) = \max_{1 \leq j \leq M} L_{1-\alpha}(t, \hat{p}_j(t)) - \hat{p}_j(t) \) against \( t \). If the upper and lower confidence surfaces contain the reference plane \( f(t, p) = p \), then \( U_{1-\alpha}(t) \) should always lie above zero, and \( L_{1-\alpha}(t) \) should always lie below zero. Points with respect to \( t \) at which this does not hold can be used to identify times for which a given method does not appear well-calibrated. We note here that in this case, due to the high degree of fluctuations in continuously updated probabilistic forecasts for basketball prediction, we have found it also useful to aid in the interpretation of these plots to smooth them with respect to \( t \) using a simple moving average smoother over 5% of the gametimes.

These summary plots calculated based on the ESPN forecasts as well as from the benchmark models HomeWP, ScDnoInt, and PgRSScD are shown in Figure 4; see Table 2 for a description of these models. From these we see that the naïve model HomeWP, which predicts that the home team will win each game at all times \( t \) with the prior 10-year historic win rate of home teams in the NBA, is well-calibrated, as expected. Similar plots (not shown) for the method CF, which simply predicts that the home team will win with probability 50%, show that this method is not well-calibrated. Considering this plot for the method ScDnoInt (Panel (b) in Figure 4), we see that the forecasts are poorly calibrated at the beginning of the game, but calibration improves toward the end of the game. This is expected since the corresponding logit model is taken to be free of an intercept term, and so the model is unable to capture the home team advantage which should force the forecast probabilities to favor the home team at the beginning of the game. Both the ESPN forecasts and those from the model PgRSScD, which incorporates team strength as well as the score difference, demonstrated generally good calibration for all game times.

A summary of the outcomes corresponding to games in which a probability forecast at some intra-game time exceeded 0.995 or was less than 0.005 is given in Table S2 in the online supplementary for the ESPN forecasts as well as for the forecasts based on the models PgRSScD and ScDnoInt. The empirical home team win rate closely matched these thresholds in each case suggesting that the forecasts for each of these methods are reasonably well-calibrated at these extremal probability forecast levels.

4. Evaluating Skill of Continuously Updated Forecasts

In this section, we consider methods to produce pointwise confidence intervals for the difference in skill, measured by the Brier score, between two competing methods as a function of the intra-game parameter \( t \), which can be used to identify intra-game times where one method appears to perform significantly better than a benchmark. We also introduce a method to evaluate the statistical significance of differences in skill aggregated across all intra-game times.
4.1. Pointwise Confidence Intervals Measuring the Skill Difference Between Competing Methods

As described in the introduction, the skill of a probabilistic forecasting method generally refers to its acuity relative to a competing or benchmark method. Formally this can be measured by defining a loss function, or scoring rule, used to measure the accuracy of a given probabilistic forecast based on the realized events. Perhaps the most frequently used loss function is $L(a,b) = (a - b)^2$ and defines the Brier score (Brier 1950). We use this loss function below, but the following results generalize to any loss function that has a linear equivalent, which means that (1) $L'(x,b)$ is linear in $x$, and (2) $L'(x,b) - L(x,a)$ does not depend on $x$. This includes Kullback-Leibler Divergence (Kullback and Leibler 1951), and Good’s log-score, among others; see Bickel (2007) and Lai, Gross, and Shen (2011).

Following the work of Lai, Gross, and Shen (2011), ideally any method to forecast $p_i(t)$ would satisfy that $L(p_i(t), \hat{p}_i(t))$ is small, and further when averaged over all forecasts would minimize

$$L_N(t) = \frac{1}{N} \sum_{i=1}^{N} L(p_i(t), \hat{p}_i(t)).$$

(1)

Since the underlying true probabilities $p_i(t)$ are unobservable, a sensible estimate of $L_N(t)$ is obtained by replacing these probabilities with their point estimates based on the realizations $Y_i$ to produce

$$\hat{L}_N(t) = \frac{1}{N} \sum_{i=1}^{N} L(Y_i, \hat{p}_i(t)).$$

(2)

The quantity $\hat{L}_N(t)$ captures the skill of the forecasting method at a given time point $t$ as described above, since a given forecasting method is ascribed generally lower losses, or higher scores, if $\hat{p}_i(t)$ is closer $Y_i$, the latter of which takes the values 0 and 1.

Suppose we wish to compare two methods, call them method A and method B, for producing continuously updated probabilistic forecasts. We denote such forecasts by $\hat{p}_i^A(t)$ and $\hat{p}_i^B(t)$, and we compare them based on the corresponding realized events $Y_i$, $1 \leq i \leq N$. This can be done by comparing their average losses defined in Equation (2). Specifically, we consider the function of $t$

$$\hat{\Delta}_N(t) = \frac{1}{N} \sum_{i=1}^{N} \left[ L(Y_i, \hat{p}_i^A(t)) - L(Y_i, \hat{p}_i^B(t)) \right],$$

(3)

which can be viewed as an estimate of the true loss difference

$$\Delta_N(t) = \frac{1}{N} \sum_{i=1}^{N} \left[ L(p_i(t), \hat{p}_i^A(t)) - L(p_i(t), \hat{p}_i^B(t)) \right].$$

(4)

Larger than zero values of $\hat{\Delta}_N(t)$ favor method B at the given $t$, whereas negative values show favor for method A. In order to measure the statistical significance of any deviations of $\hat{\Delta}_N(t)$ from zero, we may view $\hat{\Delta}_N(t)$ as an estimator for $\Delta_N(t)$. By constructing suitable confidence intervals for $\Delta_N(t)$ based on $\hat{\Delta}_N(t)$, we may evaluate whether observed deviations suggest the superiority of one model over another, and further construct simple graphical summaries that illustrate the skill of one model compared to another as a function of $t$. To construct such confidence intervals, we first define the variance of $\hat{\Delta}_N(t)$ as

$$\hat{s}_N^2(t) = \frac{1}{N} \sum_{i=1}^{N} \delta_i^2(t) p_i(t)(1 - p_i(t)),$$

(5)

where $\delta_i(t) = \left[ L(1, \hat{p}_i^A(t)) - L(0, \hat{p}_i^A(t)) \right] - \left[ L(1, \hat{p}_i^B(t)) - L(0, \hat{p}_i^B(t)) \right]$. The following result is proven in Lai, Gross, and Shen (2011) and stated for each $t \in [0,1]$.

Theorem 2, Lai, Gross, and Shen (2011): Suppose that for each $t \in [0,1]$ that $\hat{s}_N^2(t)$ converges in probability to a positive constant as $N \to \infty$, and that the variables $A_i(t) = L(Y_i, \hat{p}_i^A(t)) - L(p_i(t), \hat{p}_i^A(t))$ and $B_i(t) = L(Y_i, \hat{p}_i^B(t)) - L(p_i(t), \hat{p}_i^B(t))$ each form martingale difference sequences. Then, for each $t$,

$$\frac{\hat{\Delta}_N(t) - \Delta_N(t)}{\hat{s}_N(t)} \xrightarrow{D} N(0,1),$$

where $\xrightarrow{D}$ denotes convergence in distribution, and $N(0,1)$ denotes the standard normal distribution.

We note that the above result is similar to the main asymptotic result underpinning tests for equivalent predictive accuracy as discussed in Diebold and Mariano (2002), see also Gneiting and Katzfuss (2014). The two main conditions of the above theorem are that (1) $\hat{s}_N^2(t)$, the variance of $\hat{\Delta}_N(t)$, should for large $N$ behave like a positive constant, and (2) that the forecast loss differences should behave like martingale difference sequences. The first condition can be thought of as a nondegeneracy condition—this result only holds if the forecasts of the two methods to be compared do not coincide entirely. It is not valid for two methods that produce equivalent, or almost equivalent, forecasts. This is, in general, a reasonable assumption if the forecasts being compared are from entirely different models, or if one or both sets of forecasts to be compared come from unknown models, as with the ESPN forecast data, since it is unlikely in this case that they will produce forecasts that coincide. This assumption is in question when comparing forecasts of nested models, for example, comparing two GLM models whose only difference is that a covariate is included or excluded. A more in-depth discussion of comparing forecasts from nested models, and the problems that arise from it, may be found in Clark and McCracken (2015). Regarding the second condition, this is almost always satisfied when using a loss function with a linear equivalent and constructing genuine forecasting methods that must be based on available (past) information, rather than information from the unknown future, as discussed on page 2361 of Lai, Gross, and Shen (2011) (see also Seillier-Moiseiwitsch and Dawid 1993, p. 356).

The above results suggest constructing a 100(1-α)% confidence interval for $\Delta_N(t)$ as

$$\hat{\Delta}_N(t) \pm z_{1-\alpha/2} \frac{\hat{s}_N(t)}{\sqrt{N}}.$$
interval. Plots as a function of $t$ of $\hat{\Delta}_N(t)$ and the corresponding conservative confidence intervals can be used to evaluate the relative skill of one model compared to another. Points $t$ at which the associated confidence interval $1 - \alpha$ confidence interval for $\hat{\Delta}_N(t)$ do not contain zero indicate a significant improvement at the level $\alpha$ of the average loss of one method over another. Examples of plots of this form may be found in Figure 5, which we will discuss in more detail in Section 4.5. We note again that due to the high degree of fluctuations in continuously updated probabilistic forecasts for basketball prediction, we have found it also useful to aid in interpretation of these plots to smooth them with respect to $t$ using a simple moving average smoother over 5% of the gametimes.

4.2. Functional Tests to Measure Skill Aggregated Across $t$

Although the above confidence intervals can be used to evaluate whether two methods exhibit similar or significantly different skill at any given time point $t$, it is often also of interest to evaluate whether two continuously updated models have approximately equal predictive power when discrepancies between them are aggregated across $t \in [0,1]$. For example, it might be that one method exhibits similar but somewhat better skill at each game time that when viewed in aggregate suggest the superiority of one model over another. Conversely, it is also possible one method exhibits apparently improved performance at a single game time $t_0$, although when viewed in aggregate across $t \in [0,1]$ this improvement may appear rather insignificant.

To make this more precise, we formulate the null hypothesis of equal predictive power aggregated across $t \in [0,1]$ of two methods as

$$H_0 : \|\Delta_N\|^2 = 0,$$

where $\| \cdot \|^2$ is the standard squared $L^2$ norm of a function, so that $\|f\|^2 = \int_0^1 f^2(t)dt$. The hypothesis $H_0$ posits then that the two methods to be compared exhibit approximately on average (across $t$) equal skill. Let

$$Z_N(t) = \sqrt{N}\hat{\Delta}_N(t).$$

A measure of global discrepancy across time between the two forecasting methods may be obtained by considering $\|Z_N\|^2$. In order to determine the large-sample properties of $Z_N$ that would inform determining appropriate significance levels and estimating $p$-values for tests of $H_0$ based on $\|Z_N\|^2$, we make use of the following result, which we state rigorously and prove in the supplementary material (Section S2).

**Theorem 1.** Under $H_0$ and conditions analogous to those of Lai, Gross, and Shen (2011, theor. 2), see S2 in the supplementary material for details, there exists an infinite sequence of constants $\{\lambda_i, i \geq 1\}$ that satisfy

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0, \text{ and } \sum_{i=1}^{\infty} \lambda_i < \infty,$$

so that

$$\|Z_N\|^2 \xrightarrow{D} \sum_{i=1}^{\infty} \lambda_i \chi_i^2(1),$$

where $\chi_i^2(1)$ are independent chi-squared random variables.
where $\chi^2(1)$, $i = 1, 2, \ldots$ are independent and identically distributed $\chi^2$ random variables with one degree of freedom. Moreover, the constants $\{\lambda_i, i \geq 1\}$ can be conservatively estimated by the eigenvalues of the function

$$\hat{C}_{\text{cons}}(t, z) = \frac{1}{N} \sum_{i=1}^{N} [\hat{p}^\lambda_i(t) - \hat{p}^\lambda_i(t)][\hat{p}^\lambda_i(s) - \hat{p}^\lambda_i(s)].$$

Namely with $\{\hat{\lambda}_i, i = 1, \ldots, N\}$ defined so that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_N \geq 0$ and satisfying that there exist functions $\lambda_i(t)$, $i = 1, \ldots, N$, $t \in [0, 1]$ with $\|\lambda_i\|^2 = 1$, such that

$$\hat{\lambda}_i \lambda_i(t) = \int_0^1 \hat{C}_{\text{cons}}(t, s) \lambda_i(s) ds,$$

then for any fixed $j \geq 1$, $P(\hat{\lambda}_j \geq \lambda_j) \to 1$ as $N \to \infty$.

This result suggests a simple way of conducting an approximate and conservative test of the hypothesis $H_0$.

Step 1: Evaluate $\|Z_N\|^2$.

Step 2: Estimate $\hat{C}_{\text{cons}}$ and the eigenvalues satisfying (7).

Step 3: Estimate the distribution of the random variable $Q_D = \sum_{i=1}^{D} \hat{\lambda}_i \chi_i^2(1)$ where $D$ is a large number (below we take $D = 10$ and have found this choice generally adequate). This can be done easily using Monte Carlo simulation, or using the numerical method of Imhof (1961).

Step 4: Calculate an approximate and conservative $p$-value of the test of $H_0$ as $p = P(Q_D \geq \|Z_N\|^2)$.

This $p$-value combined with the confidence intervals in Equation (6) allow for a detailed evaluation, both at particular game times $t$ and across all $t \in [0, 1]$, of the relative skill of competing continuously updated methods.

We studied this test as well as the pointwise confidence intervals introduced in Section 4.1 in an extensive simulation study of synthetic basketball game data, and found that both methods performed generally as expected, and exhibited strong power to differentiate models with differing skill levels. A description of these simulations and the corresponding results are available in the online supplement to this article.

5. Evaluating the Skill of ESPN Forecasts

In this section, we apply the methods described in Section 4.1 to evaluate the skill of ESPN’s continuously updated probabilistic forecasts.

Figure 5 shows plots $\hat{A}_{1213}(t)$ with conservative 95% confidence intervals, as well as approximate $p$-values of the test of $H_0$ for comparisons of the ESPN forecasts with the naïve models $\text{PgRS}$, $\text{ScD}$, $\text{LS}$, and $\text{PgRSLS}$. These plots suggest that, in aggregate, the ESPN forecasts significantly outperform these models. The specific points in the game at which the ESPN forecasts exhibit higher skill compared to these benchmarks is also clear in the plots. For the models that use relative team strength as encoded by ESPN’s initial home win probability as a covariate, the relative skill is similar to ESPN’s forecasts early in the game, and similarly those that make use of the score difference improve relative to the ESPN forecasts toward the end of the game. For example, $\text{ScDnoInt}$, the model based on the score difference alone, is strongly outperformed by the ESPN forecasts early in the game, but their forecasts have indistinguishable skill toward the end of the game.

We also compared the ESPN forecasts to those of the somewhat less naïve model $\text{PgRSscD}$ using a logit link. Figure 5 shows a plot $\hat{A}_{1213}(t)$ based on the 2018 and 2019 season with conservative 95% confidence intervals, as well as approximate $p$-values of the test of $H_0$ for this comparison. In absolute terms, the estimated skill as measured by the Brier score generally favored the simple logit model $\text{PgRSScD}$, with the exception of the last moments in the game. However, we do see from this analysis that the difference is apparently not statistically significant at the 5% level at any game time point based on the conservative confidence interval estimates, nor is it significant when the difference is aggregated across time points. We found it interesting that the ESPN’s sophisticated proprietary model, which ostensibly makes use of more nuanced information about the game status and more sophisticated models, did not significantly outperform a simple logistic regression model. One might draw the conclusion based on this that whatever additional information used by ESPN’s model in producing these forecasts is not clearly beneficial for the purpose of forecasting, except for perhaps in the final moments of the game.

6. Discussion

Motivated by evaluating forecasts of NBA basketball games, we have developed graphical tools and statistical tests for assessing the calibration and relative skill of continuously updated probabilistic forecasts. These were studied via a simulation study of synthetic “basketball games,” and applied to evaluating and comparing the forecasts published on ESPN and a number of competing models. In terms of calibration, the ESPN forecasts, as well as forecasts produced from simple logistic regression models using the in-game score difference and/or pregame relative strength of teams as covariates, appear reasonably well-calibrated. In terms of skill, the ESPN forecasts exhibited significantly higher skill over naïve models, but did not demonstrate superiority over simple logistic regression models based on the score difference and relative team strength.

We conclude with a few remarks about some ideas that we considered but chose not to include in the article, and avenues for future work. It is noteworthy that the confidence intervals defined in Equation (6) may be made narrower and less conservative by using auxiliary information to improve the approximation of replacing $p_i(t)(1 - p_i(t))$ with the upper bound 1/4. We considered a number of methods to achieve this, including employing auxiliary models and covariates to estimate $p_i(t)(1 - p_i(t))$ solely in the variance estimation step, but found both that this changed the intervals little, and lead to poor performance near the end of the game. For a discussion of similar methods, see Lai, Gross, and Shen (2011, sec. 3.4 and 3.5).
The basic reason why nested models often cannot be compared by means of the confidence intervals introduced and the test of $H_0$ derived using Theorem 1 is that nested models may lead to a covariance kernel $C_{\text{cons}}$ that is approximately (and asymptotically) degenerate. It might be interesting to adapt this result, and the corresponding intervals and test, to this case, for instance using methods similar to those described in Clark and McCracken (2015).

**Supplementary Materials**

The supplementary material document contains the proof of Theorem 1, rigorous definition of $p_i(t)$, a simulation study that examines the power of the proposed test, results for automatic bin selection, and a table that enumerates the discarded forecasts in Section 3.

**Acknowledgments**

We thank to two anonymous referees, an associate editor, and the editor Joshua M. Tebbs, for the helpful comments.

**Funding**

This work has been partially supported by two Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants (GR, JAD), and the Ontario Graduate Scholarship provided by the Province of Ontario (YCK).

**References**

Bickel, J. E. (2007), "Some Comparisons Among Quadratic, Spherical, and Logarithmic Scoring Rules," *Decision Analysis*, 4, 49–65. [220]

Brier, G. W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3. [215,220]

Brown, L. D., Cai, T. T., and DasGupta, A. (2001), "Interval Estimation for a Binomial Proportion," *Statistical Science*, 16, 101–117. [218]

Clark, T. E., and McCracken, M. W. (2015), "Nested Forecast Model Comparisons: A New Approach to Testing Equal Accuracy," *Journal of Econometrics*, 186, 160–177. [220,223]

Council, N. R. (2006), *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*, Washington, DC: The National Academies Press. [214]

Dawid, A. P. (1986), "Probability Forecasting" (Vol. 7), in *Encyclopedia of Statistical Sciences*, ed. S. Kotz, N. L. Johnson, and C. B. Read, New York: Wiley, pp. 210–218. [217]

Diebold, F. X., and Mariano, R. S. (2002), "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 20, 134–144. [220]

 Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2020), "Evaluating Probabilistic Classifiers: Reliability Diagrams and Score Decompositions Revisited," arXiv:2008.03033. [218]

Elo, A. E. (1978), *The Rating of Chessplayers, Past and Present*, New York: Arco Pub. [217]

Erikson, R. S., and Wlezien, C. (2012), "Markets vs. Polls as Election Predictors: An Historical Assessment," * Electoral Studies*, 31, 532–539. [214]

ESPN (2020), "ESPN, ESPN Internet Ventures, National Basketball Association Teams, Scores, Stats, News, Standings, Rumors," available at [www.espn.com/nba/]. [215,216]

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007), "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society*, Series B, 69, 243–268. [215]

Gneiting, T., and Katzfuss, M. (2014), "Probabilistic Forecasting," *The Annual Review of Statistics and Its Application*, 1, 125–151. [215,218,220]

Gomberg, J., Atwater, B., Beeler, N., Bodin, P., Davis, E., Frankel, A., Hayes, G., McConnell, V., Melbourne, T., Oppenheimer, D., Parrish, J., Roeloffs, E., Rogers, G., Sherrod, B., Vidale, J., Walsh, T., Weaver, C., and Whitmore, P. (2015), "Earthquake Forewarning in the Cascadia Region," U.S. Geological Survey Open-File Report 2015–1151. [214]

Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419–426. [222]

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86. [220]

Lai, T. L., Gross, S. T., and Shen, D. B. (2011), "Evaluating Probability Forecasts," *The Annals of Statistics*, 39, 2366–2382. [215,220,221,222]

McCullagh, P., and Nelder, J. (1989), *Generalized Linear Models*, 2nd ed. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. New York: Chapman & Hall. [216]

McFadden, D. (1973), "Conditional Logit Analysis of Qualitative Choice Behaviour," in *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press, pp. 105–142. [217]

Murphy, A. H. (1998), "The Early History of Probability Forecasts: Some Extensions and Clarifications," *Weather and Forecasting*, 13, 5–15. [214]

Murphy, A. H., and Winkler, R. L. (1987). "A General Framework for Forecast Verification," *Monthly Weather Review*, 115, 1330–1338. [214]

Ranjan, R., and Gneiting, T. (2010), "Combining Probability Forecasts," *Journal of the Royal Statistical Society*, Series B, 72, 71–91. [218]

Rowe, T., and Beard, S. (2018), "Probabilities, Methodologies and the Evidence Base in Existential Risk Assessments," Centre for the Study of Existential Risk, University of Cambridge. [214]

Seillier-Moiseiwitsch, F., and Dawid, A. P. (1993), "On Testing the Validity of Sequential Probability Forecasts," *Journal of the American Statistical Association*, 88, 355–359. [220]

Silver, N. (2014), "Introducing NFL ELO Ratings," available at [https://fivethirtyeight.com/features/introducing-nfl elo-ratings/]. [217]

Silver, N. (2020), "2020 Election Forecast," available at [https://projects.fivethirtyeight.com/2020-election-forecast/]. [214]

Silver, N., Boice, J., and Paine, N. (2019), "How our NBA Predictions Work," available at [https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/]. [214,217]

Silver, N., and Fischer-Baum, R. (2015), "How We Calculate NBA ELO Ratings," available at [https://fivethirtyeight.com/features/how-we-calculate-nba elo-ratings/]. [217]

Spiegelhalter, D. J. (1986), "Probabilistic Prediction in Patient Management and Clinical Trials," *Statistics in Medicine*, 5, 421–433. [214]

Wilson, E. B. (1927), "Probable Inference, the Law of Succession, and Statistical Inference," *Journal of the American Statistical Association*, 22, 209–212. [218]