International Journal of Forecasting 35 (2019) 573–579

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Interpreting the skill score form of forecast performance metrics

Edward Wheatcroft

Centre for the Analysis of Time Series, London School of Economics, London WC2A 2AE, United Kingdom

ARTICLE INFO

Keywords:
Evaluating forecasts
Forecasting practice
Nonlinear time series
Probability forecasting
Time series
Weather forecasting

ABSTRACT

Performance measures of point forecasts are expressed commonly as skill scores, in which the performance gain from using one forecasting system over another is expressed as a proportion of the gain achieved by forecasting that outcome perfectly. Increasingly, it is common to express scores of probabilistic forecasts in this form; however, this paper presents three criticisms of this approach. Firstly, initial condition uncertainty (which is outside the forecaster's control) limits the capacity to improve a probabilistic forecast, and thus a 'perfect' score is often unattainable. Secondly, the skill score forms of the ignorance and Brier scores are biased. Finally, it is argued that the skill score form of scoring rules destroys the useful interpretation in terms of the relative skill levels of two forecasting systems. Indeed, it is often misleading, and useful information is lost when the skill score form is used in place of the original score.

1. Introduction

Forecasting is a common endeavour in a wide range of disciplines, and as a result, the question of how forecasts can best be evaluated is of fundamental importance to much of the scientific community and beyond. One of the most common fields in which forecasting is deployed is weather forecasting, in which deterministic models of the atmosphere are used to simulate the future. Similar approaches are used in ecology (Hastings, Hom, Ellner, Turchin, & Godfray, 1993), hydrology (Smith & Beven, 2014) and biology (Strogatz, 2018), among other fields. In other areas, such as tourism (Smith, 1993), economics (Katz & Lazo, 2011) and agriculture (Hansen, Mason, Sun, & Tall, 2011), more statistical approaches tend to be taken, in which the key driving variables of some particular dependent variable are sought and used to make out-of-sample predictions. However, the issue of forecast evaluation is a more general one. Originally suggested as a means of comparing point forecasts, the skill score form of a forecast evaluation metric is an approach that expresses the relative skill levels of two competing forecasting systems (Murphy & Daan, 1985). This paper identifies a number of weaknesses of this approach and suggests an alternative approach.

In weather forecasting, as well as in the forecasting of other physical systems, deterministic models are used to simulate the underlying system. The dynamics of systems such as the atmosphere are often highly nonlinear (Lorenz, 1963), and thus, such physical models generally also have nonlinear, or even chaotic, dynamics. Combined with the fact that observations of physical variables are usually both incomplete and obscured by measurement error, a single model trajectory launched from a noisy observation would diverge from the truth even if the model reproduced the underlying dynamics perfectly. Thus, in general, a noisy observation of the initial condition can yield, at best, a *set* of model trajectories, called an ensemble, that are all consistent with that observation. Whilst accounting for observations stretching into the past can discount some of these trajectories, in a chaotic system it is never possible to narrow this set down to just the true initial condition (Smith & Judd, 2001, 2004), and thus, the best possible forecast of a nonlinear system is probabilistic, at best, even if the

E-mail address: e.d.wheatcroft@lse.ac.uk.

https://doi.org/10.1016/j.ijforecast.2018.11.010

underlying model/system dynamics themselves are deterministic. Thus, it is common to use ensembles to construct forecast probabilities (for discrete events) or probabilistic forecast densities (for continuous events); see Bröcker and Smith (2008).

Probabilistic forecasting is also used widely in applications in which purely statistical models, such as linear regressions, are utilised. For example, in sales forecasting, it is common to use regression models to identify key driving factors for sales and to use these to make predictions regarding future sales patterns (Böse, Flunkert, Gasthaus, Januschowski, Lange, Salinas, et al., 2017). In sports forecasting, there is typically some rating applied to each team, after which a statistical approach is used to relate those ratings to forecast probabilities (Constantinou, Fenton, & Neil, 2012). Statistical approaches are also often used in energy price forecasting (Ziel & Steinert, 2018) and population forecasting (Alkema, Gerland, Raftery, & Wilmoth, 2015), among many other fields.

A *scoring rule* is a function of a probabilistic forecast, and its corresponding outcome is intended for measuring predictive performances. Due to the probabilistic nature of the forecasts, though, the scores are only meaningful when multiple forecasts and outcomes are considered. Thus, it is common for the mean or median score to be given and used for comparison purposes.

A skill score is defined as the gain in forecast accuracy, given some measure, as a proportion of the total gain in accuracy that would be possible were a perfect point forecast to be issued; i.e., were the forecast able to predict the outcome perfectly (Murphy & Daan, 1985). The aim of a skill score is to give some context to the gain in skill that is achieved by using a given forecasting system over some other reference one. Whilst the skill score form of a scoring rule is intended to yield an intuitive measure of the relative skill levels of two forecasting systems, this paper argues that a number of shortcomings of this approach tend to outweigh the benefits of taking it.

In the weather forecasting literature, the scores of probabilistic forecasts are often converted into skill score form (Christensen, Moroz, & Palmer, 2015; Siegert, Bröcker, & Kantz, 2011; Weigel, Liniger, & Appenzeller, 2007; Wilks, 2001) before they are presented. This approach is also used commonly in operational weather forecasting. For example, skill scores are used as headline evaluation tools at both the European Centre for Medium Range Weather Forecasting (ECMWF)[1] and the UK Met-Office.[2] Although skill scores are used most commonly in the forecasting of physical systems such as the weather, they have also been used in a wide range of fields such as macroeconomic forecasting (Bluedorn, Decressin, & Terrones, 2016; Lahiri & Wang, 2013), the forecasting of baseball (Richards, 2014) and association football (Haave & Høiland, 2017), and medicine (Karoly, Ung, Grayden, Kuhlmann, Leyde, Cook, & Freestone, 2017).

This paper starts by arguing that, since the presence of observational uncertainty in an initial condition makes a perfect point forecast impossible in the context of simulation models, the skill score form of a scoring rule represents the gain in skill as a proportion of the total possible gain were a perfect forecasting system available *and* no observational uncertainty present in the initial condition from which the forecasts were launched. It is therefore argued that this is not a useful measure, as the observational noise is usually outside the control of the forecaster. Next, using a number of examples, we show that the skill score forms of a number of scoring rules are biased when a finite number of forecasts and outcomes are evaluated. This particular criticism is common to both point and probabilistic forecasts, and is demonstrated in both cases. Finally, we question whether the proportion of possible skill gained has a useful interpretation regarding the relative value of two given forecasting systems.

This paper is organised as follows. Section 2 presents the background methodology describing the scoring rules and skill scores. Section 3 discusses the relevance of the 'optimal' score that is necessary for calculating the skill score form and argues that the optimal score renders the skill score form unusable in some cases and unachievable in many cases unless the accuracy of the observations can be improved (which is not usually an option for a forecaster). Section 4 demonstrates analytically that the skill score form of the mean squared error (for point forecasts) can be biased. Using an empirical example, it then shows that the skill score forms of the ignorance and Brier scores can also give biased results. Section 5 discusses whether the skill score forms of scoring rules have useful interpretations, and Section 6 provides discussion and conclusions.

2. Background definitions

2.1. Evaluating point forecasts

Although this paper is concerned mostly with probabilistic forecasts, some of the issues raised also apply to point forecasts, and these are demonstrated using the two measures of point forecast accuracy given below. The two measures considered are the mean squared error, defined as

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(f_i - Y_i)^2, \tag{1}$$

and the mean absolute error, defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|f_i - Y_i|, \tag{2}$$

where f_i and Y_i represent the point forecast and the outcome for the ith period, respectively.

2.2. Scoring rules

A *scoring rule* is a function of a probabilistic forecast and its outcome that evaluates forecast performances. Since scoring rules consider only probabilistic forecasts, this means that measures of the performances of point forecasts, such as the mean squared error, do not fall under the definition of a scoring rule. By convention, scoring rules

[1] See https://www.ecmwf.int/en/forecasts/quality-our-forecasts.

[2] See the MOSAC-21 Annex II: forecast accuracy, https://www.metoffice.gov.uk/binaries/content/assets/mohippo/pdf/library/mosac/2016/mosac21_annex_ii_forecast_accuracy.pdf.

are defined to be oriented negatively; that is, lower scores imply better forecast accuracies. Many different scoring rules have been proposed over the years, and the decision as to which to use to evaluate a set of forecasts is of great importance.

A scoring rule is *proper* if it is optimised in expectation by a perfect probabilistic forecast; that is, the true distribution from which the outcome was drawn. To be useful, a scoring rule should always be proper, otherwise there would be no incentive to choose a perfect forecasting system if one was available. In addition, under a perfect model with some well-defined but unknown parameters, optimising those parameters with respect to an improper scoring rule would result in convergence to the wrong values. For these reasons, this paper considers only proper scoring rules. One particular score that fits this requirement is the ignorance score, introduced by I.J Good in 1951 (Good, 1952; Roulston & Smith, 2002) and defined for discrete forecasts as

$$\mathrm{IGN} = -\log_2(p(Y)), \tag{3}$$

where $p(Y)$ represents the probability placed on the outcome by the forecast. In the continuous case, the probability is replaced by the probability density, meaning that ignorance is defined as

$$\mathrm{IGN} = -\log_2(f(Y)), \tag{4}$$

where $f(Y)$ is the forecast density placed on the outcome Y.

Another proper scoring rule is the Brier score (Brier, 1950), which is defined so as to evaluate the performance of binary probabilistic forecasts. It is given[3] by

$$BS = (p(Y) - Y)^2, \tag{5}$$

where $p(Y)$ represents the forecast probability and Y is one (zero) if the event occurred (did not occur). The Brier score is bounded between zero and one, with a score of zero corresponding to the case in which a probability of one is placed on the eventual outcome and a score of one if the probability placed on the outcome is zero.

2.3. Skill scores

It is commonly argued that measures of the forecast accuracy should be expressed in the form of a skill score (Christensen et al., 2015; Murphy & Epstein, 1989; Siegert et al., 2011; Tödter & Ahrens, 2012). A *skill score* is defined as

$$SS = \frac{A_f - A_r}{A_p - A_r}, \tag{6}$$

where A_f and A_r represent the 'accuracy', according to some given measure, of the forecasting system of interest and some reference forecasting system, respectively. The quantity A_p represents the optimal value of the measure; that is, the value of the metric if the outcome were known perfectly. The value of SS can be interpreted as the increase in accuracy achieved by using some forecasting system of interest, as a proportion of the total possible increase in accuracy. The reference forecasting system could be either a competing forecasting system over which improvement is sought or some benchmark forecasting system such as a climatology (a forecast based purely on past states).

[3] Note that the Brier score is often defined as an average over N forecasts and outcomes. Here it is defined for one particular forecast and outcome, for the sake of consistency with the definition of the ignorance score.

Table 1
Values of A_p for discrete and continuous (where applicable) forecasts for the scoring rules considered in this paper.

Scoring rule	Discrete forecast A_p	Continuous forecast A_p
Ignorance	0	∞
Brier score	0	NA

3. Defining a 'perfect score'

The skill score representation of a measure of the forecast accuracy, as defined in Eq. (6), can be interpreted as the improvement in accuracy, according to the measure, as a proportion of the total possible improvement if the true outcome were known perfectly. The value A_p does not depend on the forecast, but is in fact a property of the accuracy measure itself. The values of A_p for discrete and continuous (where applicable) forecasts for each scoring rule considered in this paper are shown in Table 1. Note that, for the ignorance score, A_p is infinite in the continuous case, and thus the skill score representation is not informative in this case. The Brier score is defined only for binary categorical forecasts.

When forecasting is performed using deterministic simulation models, the existence of observational uncertainty in an initial condition prevents the point forecasts from being perfect (i.e., predicting the exact outcome consistently), regardless of the accuracy of the forecasting system. In addition, all real-world models contain some degree of structural error. Inasmuch as observational uncertainty can be considered an unavoidable feature of the real world, it can be argued that any limitations to predictability that result from this factor should be differentiated from those that stem from the forecasting system itself. If this is not done, the impression that the forecasts could be improved upon further might be given even if the forecasting system is already as informative as it can possibly be, given the information available. In the presence of observational noise, the best possible forecast will be a probability distribution, henceforth referred to as a perfect probabilistic forecast. Since the only uncertainty is in the initial condition, a perfect probabilistic forecast can be considered to be the distribution from which the eventual outcome is drawn, given only the initial condition uncertainty. This is achieved by evolving forward the distribution of possible initial conditions that are consistent with both the observation of the initial condition and the system dynamics. Thus, some uncertainty regarding the outcome will remain, and as a result, the best possible score given the observations available will come from a perfect *probabilistic* forecast rather than a perfect point forecast, meaning that the optimal score A_p will be unattainable. A more useful quantity, if available, would be

$$\mathrm{SSprob} = \frac{A_f - A_r}{A_{pp} - A_r}, \tag{7}$$

where A_{pp} represents the score achieved with a perfect probabilistic forecast. This quantity represents the skill gained over the reference forecast as a proportion of the total possible gain in skill, given the observation and the distribution of the observational uncertainty. In practice, though, SSprob is never available, since the skill of a perfect probabilistic forecast is never expected to be known. Thus, the proportion of potential skill gained is not expected ever to be available for probabilistic forecasts. This calls into question the value of the skill score representation of scoring rules.

Some studies and applications calculate the ensemble mean and treat it as a point forecast, rather than using ensembles to construct probabilistic forecasts. It could be argued that such an idea is ill-advised, since the approach discards important information regarding the shape of the distribution. Worse still, the mean of the forecast distribution is often an unlikely quantity when the dynamics of the model are nonlinear. Consider a forecast distribution of the waiting time between eruptions of the Old Faithful geyser in Yellowstone National Park in the USA, which famously has a bimodal distribution (Rinehart, 1969). A probabilistic forecast distribution may suggest that either a relatively long or a relatively short waiting time is likely. A point forecast based on the mean, on the other hand, would make a prediction somewhere between the two, which is a relatively unlikely outcome. Nonetheless, it should be pointed out that, since the outcomes in such situations will be drawn from some underlying distribution and the forecast will, at best, represent the mean of that distribution, the initial condition uncertainty that differentiates the ensemble members means that perfect point forecasts (i.e., forecasts that always coincide with the outcome) are not attainable, and the arguments presented above still apply.

The arguments above consider the case in which forecasts are generated using deterministic simulation models. In many applications, though, it is actually statistical models that are applied; for example, with linear regression, the resulting forecast is a single point estimate that forms the mean of some Gaussian forecast distribution. However, the simulation model approach taken in numerical weather prediction could theoretically be applied to such cases, though in practice statistical models may be deemed more effective in relating important variables to the predictand (economic data to sales volumes, for example). Whilst one can never expect statistical models to yield perfect point forecasts, in theory the deterministic modelling approach could be taken (though building such a model may be highly impractical). Consider, for example, making a forecast of the outcome of a football match. In general, forecasting of this kind is done using statistical approaches, but if one knew the dynamics of the world perfectly and had perfect observations of its exact state at a given time, theoretically it would be possible to make a perfect point forecast of the state of the world (Laplace, 2012), and therefore of the outcome of that match. However, once the assumption that perfect observations are available is removed, the very best forecast of that match would be a probability distribution again, even with a perfect model (Frigg, Bradley, Du, & Smith, 2014). Of course, obtaining perfect observations of the world is out of the control of the forecaster (and impossible in practice), and thus, as discussed above, skill scores do not represent the proportion of the possible skill achieved by the forecaster.

4. Sampling distributions

So far, we have considered only the general properties of scoring rules and their skill score form, but the question of how each behaves in the context of a finite sample is also of importance. In practice, any measure of the forecast accuracy is calculated over a finite sample of forecasts and outcomes. The skill score form of a scoring rule aims to provide a direct comparison of the skill levels of two forecasting systems. However, this section shows that the skill score form of a measure of forecast accuracy can be biased for finite samples. We start by demonstrating this analytically using a point forecasting example in which the mean squared error is used as the measure of accuracy. We then demonstrate it using the skill score form of scoring rules in the context of probabilistic forecasts. This is compared with an alternative approach to expressing the relative skill levels of two forecasting systems that is shown to be unbiased.

An alternative statistic to a skill score for comparing the performances of two forecasting systems is defined as

$$A_{rel} = A_f - A_r, \tag{8}$$

where A_{rel} will be referred to as the *relative skill*. The relative skill is related closely to a skill score, since

$$\mathrm{SS} = \frac{A_{rel}}{A_p - A_r}, \tag{9}$$

and, when $A_p = 0$,

$$\mathrm{SS} = -\frac{A_{rel}}{A_r}. \tag{10}$$

Thus, the skill score form of a measure of accuracy is a simple transformation of the relative skill, which is an unbiased estimator of the actual difference in skill. However, it is shown that the skill score form is *not* necessarily an unbiased estimator of the underlying skill score (that is, the skill score that would be obtained from an infinite number of forecasts and outcomes). We demonstrate this by presenting a simple example using point forecasts.

Consider a simple case in which, for each outcome Y_i, both the forecasts from the forecasting system of interest $u_{f,i}$ and the reference forecasting system $u_{r,i}$ are created by taking random draws from a Gaussian distribution $N(Y_i, 1)$ that is centred on the outcome Y_i. Thus, both forecasting systems are expected to have the same mean squared errors, on average. The mean squared error skill score in this case is given by

$$\mathrm{MSESS} = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(u_{f,i} - Y_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(u_{r,i} - Y_i)^2}. \tag{11}$$

Note that $\frac{\frac{1}{N}\sum_{i=1}^{N}(u_{f,i}-Y_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(u_{r,i}-Y_i)^2} \sim F(v_1, v_2)$, where $v_1 = v_2 = N$. Since the mean of an F distribution is $\frac{v_2}{v_2-2}$, the expected value of the skill score for this special case is $E(\mathrm{MSESS})$

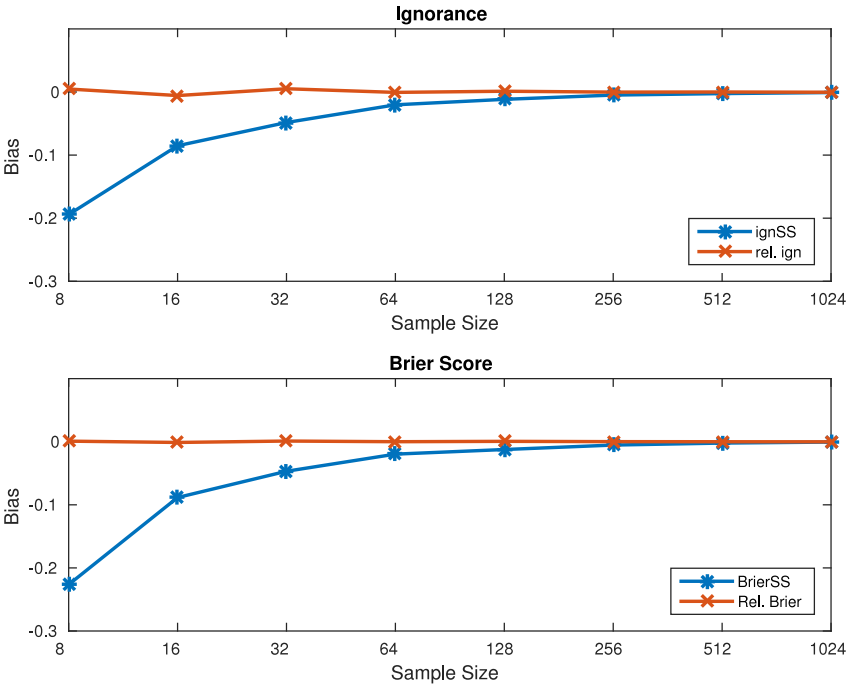

Fig. 1. The estimated bias of the relative skill (red) and the skill score forms (blue) of the ignorance (top panel) and Brier (lower panel) scores as a function of the sample size. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$= 1 - \frac{N}{N-2}$. Since the two forecasting systems are defined to have the same mean squared errors, on average, the mean squared error skill score is biased.

In the special case outlined above, the sampling distribution was derived analytically under some strong assumptions. In most cases, the sampling distribution will not be known; however, it can be estimated. This is now done for both the ignorance and Brier scores in a special case in which the two probabilistic forecasting systems are expected to have the same skill levels. Define $\boldsymbol{p}_f = p_{f,1}, \ldots, p_{f,N}$ and $\boldsymbol{p}_r = p_{r,1}, \ldots, p_{r,N}$ as two sets of *iid* random draws from a standard uniform distribution $U(0, 1)$. Let $p_{f,i}$ and $p_{r,i}$ represent two different probabilistic forecasts of the same binary outcome Y_i, such that each one represents a single forecast probability. The distribution of the outcome Y_i is then defined to be Bernoulli, with the parameter randomly chosen to be $p_{f,i}$ or $p_{r,i}$ with equal probability. The outcome Y_i is then a random draw from the randomly selected true distribution. This means that each of the forecast probabilities have equal chances of coinciding with the true probability. Given that it is not known with what probability the outcome was drawn, the result is that $\boldsymbol{p}_f$ and $\boldsymbol{p}_r$ represent equally useful probabilistic forecasts, on average. For an infinite number of forecasts and outcomes, the relative skill and the skill score form of any evaluation measure are both zero. We test whether there is any bias in either measure for finite samples, by randomly drawing sets of forecasts and outcomes of size N and calculating both the relative skill and the skill score. The mean of each is then calculated to give an estimate of the expected value, and thus of the bias. This is repeated for various values of N. The results of the experiment are shown in Fig. 1. The top panel shows the estimated bias of the relative skill and the skill score form of the ignorance score, whilst the lower panel shows these for the Brier score. It is clear here that the skill score forms of both scoring rules are biased, whilst such does not appear to be the case for the relative skill. The bias appears in the skill score because of the quotient that is required in its calculation.

The bias in the skill score also has an impact on tests of whether there is a significant difference between two forecasting systems. For the relative skill, bootstrap resampling can be applied to the differences to infer whether the mean difference in skill is significant. This is a reasonable thing to do because the relative skill is an unbiased estimator of the underlying difference in skill between the two forecasting systems. Whilst something similar could be applied to the skill score form, the bias means that there would be an overinflated probability of finding the reference forecasting system to be superior to the forecasting system of interest.

5. Interpreting and comparing skill scores

The skill score form of a measure of accuracy gives a scaling between 1 and $-\infty$, and measures the gain in skill, according to some measure, as a proportion of the total possible gain. For example, a skill score of 0.5 means that half of the total possible increase in the measure of accuracy has been achieved. In probabilistic forecasting, the measure of accuracy usually consists of a scoring rule. However, the skill score form of a measure of accuracy, as described in Eq. (6), was suggested initially for the evaluation of point forecasts (Murphy & Daan, 1985). Although this paper is concerned mostly with skill scores in the

context of probabilistic forecasts, it is useful to illustrate the intended interpretation of skill scores in the context of point forecasts, for comparison. Consider the mean squared error and mean absolute error, defined in Section 2.1. In both cases, the value A_p in Eq. (6) is zero, since the forecast and the outcome for an optimal point forecast would coincide. If the mean absolute errors achieved from the forecasting system of interest and the reference forecasting system are 3 and 4 respectively, the skill score form of the mean absolute error would be MAESS $= \frac{3-4}{0-4} = 0.25$, which can be interpreted as a reduction of 25% in the mean distance between the forecasts and the outcomes. This is an intuitive measure of the difference in accuracy between two forecasting systems. The value of using the skill score form of the mean squared error is less obvious. Consider for example a similar case in which the mean squared errors of the forecast system of interest and the reference forecasting system are again 3 and 4 respectively. The mean squared error skill score would then be 0.25 again. However, a 25% reduction in the mean squared error is harder to interpret than a 25% reduction in the mean absolute error. It could be argued that this is because the mean squared error has a less intuitive interpretation in the first place, so forcing it into skill score form adds little or nothing of value and potentially even makes it still less intuitive.

The nature of probabilistic forecasts means that the evaluation techniques described above cannot be applied without compromising the information content of the forecast, and thus, a scoring rule is required. While the skill score forms of some distance metrics like the mean absolute error can have simple and useful interpretations, as has been discussed, such is not necessarily the case for scoring rules. Consider for example the ignorance score. The relative skill of the ignorance score, as described in Section 4, can be interpreted as the mean bits of information gained from using one forecasting system over some reference forecasting system (say, the climatological distribution). This can then be converted back in order to infer how much more density or probability is placed on the outcome, on average. The skill score form can be interpreted as the number of bits of information gained over the total possible gain. However, the proportion of possible bits gained should not be considered a linear gain in value. In fact, the gain in probability or density placed on the outcome by using one forecasting system over another cannot be recovered using the skill score form alone. Thus, it does not seem to make sense to express the ignorance in the skill score form at all.

The Brier score can be interpreted as the mean squared distance between the probabilities and the outcome (either a one or a zero) in a binary probabilistic forecast. Similarly to the ignorance score, an increase or decrease in the Brier score does not correspond to a linear increase or decrease in the utility of the forecasts (this will depend on how the forecasts are to be used) in any conceivable way. It could be argued that one weakness of the Brier score over the ignorance score is that the former has a far less clear interpretation, and transforming the score into a skill score form will not create a useful interpretation if there is none in the first place.

Scoring rules such as the ignorance and Brier scores are clear, mathematically precise and informative scores. Forcing them into a skill score form destroys this utility, and there is no persuasive argument establishing any benefit from doing so anyway. It should also be clear from the discussion above that skill scores based on different measures of accuracy cannot be compared directly, even though they are required to be on the same scale.

6. Discussion

The skill score form of a forecast evaluation metric is designed to provide an intuitive measure of the gain in skill that can be achieved by using one forecasting system over another. However, while a skill score represents the mean gain in accuracy that can be achieved by using one forecasting system over another as a proportion of the total possible gain given a perfect point forecast, care needs to be taken to interpret this as an *achievable* gain in skill. When observational noise is present in the initial condition, even a perfect forecasting system cannot yield perfect point forecasts, and thus the skill score does not represent the proportion of the possible skill that could be obtained by improving the forecasting system. In addition, even if it were possible for the forecasting system to yield perfect point forecasts, it is not clear whether the proportion of the potential skill gained represents a useful indication of the proportion of the actual value that can be gained by using the forecasting system of interest. It has been shown that the skill score form of an unbiased measure of accuracy is not necessarily unbiased itself, due to the ratio that is introduced into the formula. The skill score form of a measure of accuracy is intended to give some context to the gain in skill obtained by using one forecasting system over another. In many cases, it could be argued that the intended interpretation is misleading as a measure of the proportion of the possible skill gained. Taken in combination with the fact that the skill score form can introduce a bias to the score, it should be treated with caution. Pressure to express the forecast system evaluation in terms of skill scores is found to be misplaced; in the absence of better motivation, the forecast evaluation can be more effective when considered in terms of raw scores with more meaningful units.

Acknowledgments

This work was supported by the Evaluating Probability Scores for the Insurance Sector (EPSIS) project funded by the LSE KEI fund, United Kingdom and the Lighthill Risk Network, United Kingdom. I am grateful to Leonard A. Smith for providing invaluable feedback on earlier versions of this paper.

References

Alkema, L., Gerland, P., Raftery, A., & Wilmoth, J. (2015). The United Nations probabilistic population projections: an introduction to demographic forecasting with uncertainty. *Foresight*, *2015*(37), 19.

Bluedorn, J. C., Decressin, J., & Terrones, M. E. (2016). Do asset price drops foreshadow recessions? *International Journal of Forecasting*, *32*(2), 518–526.

Böse, J. -H., Flunkert, V., Gasthaus, J., Januschowski, T., Lange, D., Salinas, D., et al. (2017). Probabilistic demand forecasting at scale. *Proceedings of the VLDB Endowment*, *10*(12), 1694–1705.
Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1.
Bröcker, J., & Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A*, *60*(4), 663–678.
Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015). Evaluation of ensemble forecast uncertainty using a new proper score: application to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 538–549.
Constantinou, A. C., Fenton, N. E., & Neil, M. (2012). Pi-football: a Bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, *36*, 322–339.
Frigg, R., Bradley, S., Du, H., & Smith, L. A. (2014). Laplace's demon and the adventures of his apprentices. *Philosophy of Science*, *81*(1), 31–59.
Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B.*, *14*, 107–114.
Haave, H. S., & Høiland, H. (2017). *Evaluating association football player performances using markov models* (Master's thesis), NTNU.
Hansen, J. W., Mason, S. J., Sun, L., & Tall, A. (2011). Review of seasonal climate forecasting for agriculture in sub-Saharan Africa. *Experimental Agriculture*, *47*(2), 205–240.
Hastings, A., Hom, C. L., Ellner, S., Turchin, P., & Godfray, H. C. J. (1993). Chaos in ecology: is mother nature a strange attractor? *Annual Review of Ecology and Systematics*, *24*(1), 1–33.
Karoly, P. J., Ung, H., Grayden, D. B., Kuhlmann, L., Leyde, K., Cook, M. J., & Freestone, D. R. (2017). The circadian profile of epilepsy improves seizure forecasting. *Brain*, *140*(8), 2169–2182.
Katz, R. W., & Lazo, J. K. (2011). Economic value of weather and climate forecasts. In M. P. Clements, & D. F. Hendry (Eds.), *The Oxford handbook of economic forecasting, chap. 20* (pp. 559–584).
Lahiri, K., & Wang, J. G. (2013). Evaluating probability forecasts for GDP declines using alternative methodologies. *International Journal of Forecasting*, *29*(1), 175–190.
Laplace, P. -S. (2012). *Pierre-simon laplace philosophical essay on probabilities: translated from the fifth french edition of 1825 with notes by the translator, Vol. 13*. Springer Science & Business Media.
Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, *20*(2), 130–141.
Murphy, H., & Daan, H. (1985). Forecast evaluation. In A. Murphy, & R. Katz (Eds.), *Probability, Statistics and Decision Making in the Atmospheric Sciences* (pp. 379–437). Boulder, CO: Westview Press.
Murphy, A., & Epstein, E. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, *117*(3), 572–582.
Richards, J. A. (2014). Probabilities of victory in head-to-head team matchups. *Fall 2014 Baseball Research Journal*, *43*(2), 107–117.
Rinehart, J. (1969). Old faithful geyser performance 1870 through 1966. *Bulletin Volcanologique*, *33*(1), 153–163.
Roulston, M. S., & Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, *130*, 1653–1660.
Siegert, S., Bröcker, J., & Kantz, H. (2011). Predicting outliers in ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, *137*(660), 1887–1897.
Smith, L. K. (1993). The influence of weather and climate on recreation and tourism. *Weather*, *48*(12), 398–404.
Smith, P. J., & Beven, K. J. (2014). When to issue a flood warning: towards a risk-based approach based on real time probabilistic forecasts. In M. Beer, S. K. Au, & J. W. Hall (Eds.), *Vulnerability, uncertainty, and risk: quantification, mitigation, and management* (pp. 1395–1404).
Smith, L., & Judd, K. (2001). Indistinguishable states I. perfect model scenario. *Physica D*, *151*, 125–141.
Smith, L., & Judd, K. (2004). Indistinguishable states II. imperfect model scenario. *Physica D*, *196*, 224–242.
Strogatz, S. H. (2018). Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering. CRC Press.
Tödter, J., & Ahrens, B. (2012). Generalization of the ignorance score: continuous ranked version and its decomposition. *Monthly Weather Review*, *140*(6), 2005–2017.
Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2007). The discrete brier and ranked probability skill scores. *Monthly Weather Review*, *135*(1), 118–124.
Wilks, D. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, *8*(2), 209–219.
Ziel, F., & Steinert, R. (2018). Probabilistic mid- and long-term electricity price forecasting. *Renewable & Sustainable Energy Reviews*, *94*, 251–266.

Edward Wheatcroft is a research officer in the Centre for the Analysis of Time Series at the London School of Economics. He gained a BSc in Mathematics and Statistics and an MSc in Statistics from the University of Kent before studying for his Ph.D. at the London School of Economics in the area of forecasting nonlinear systems. His research interests include ensemble forecasting, probabilistic forecasting, forecast evaluation and data assimilation.