Edward Wheatcroft

Interpreting the skill score form of forecast performance metrics

Article (Accepted version) (Refereed)

Interpreting the Skill Score Form of Forecast Performance Metrics

Edward Wheatcroft

Abstract

Performance measures of point forecasts are commonly expressed as skill scores in which the gain in performance from using some forecasting system over another is expressed as a proportion of the gain made by forecasting that outcome perfectly. It is increasingly common to express scores of probabilistic forecasts in this form. Three criticisms of this approach are presented. Firstly, initial condition uncertainty (out of the control of the forecaster) 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 the 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 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. The question of how best to evaluate forecasts is therefore 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 et al., 1993], hydrology [Smith and Beven, 2014] and biology [Strogatz, 2018], among other fields. In other areas, such as tourism [Smith, 1993], economics [Katz and Lazo, 2011] and agriculture [Hansen et al., 2011], more statistical approaches tend to be taken in which key driving variables of some dependent variable are sought and used to make out-of-sample predictions. The issue of forecast evaluation,

however, is a more general one. Originally suggested as a means with which to compare point forecasts, the skill score form of a forecast evaluation metric is an approach that expresses the relative skill of two competing forecasting systems [Murphy and Daan, 1985]. In this paper, a number of weaknesses of this approach are identified and an alternative approach is suggested.
In weather forecasting and 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 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 underlying dynamics were reproduced perfectly by the model. A noisy observation of the initial condition can thus usually, at best, yield a *set* of model trajectories, called an ensemble, all of which are 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 just to the true initial condition [Smith and Judd, 2001, Smith and Judd, 2004] and thus the best possible forecast of a nonlinear system is, at best, probabilistic, even if the underlying model/system dynamics are, themselves, deterministic. It is thus common to use ensembles to construct forecast probabilities (for discrete events) or probabilistic forecast densities (for continuous events) [Bröcker and Smith, 2008].
Probabilistic forecasting is also widely used in applications in which purely statistical models, such as linear regression, are utilised. For example, in sales forecasting, it is common to use regression models to identify key driving factors for sales and use these to make predictions of future sales patterns [Böse et al., 2017]. In sports forecasting, typically some rating is applied to each team and a statistical approach is used to relate those ratings to forecast probabilities [Constantinou et al., 2012]. Statistical approaches are also often used in energy price forecasting [Ziel and Steinert, 2018] and population forecasting [Alkema et al., 2015] among many other fields.
A *scoring rule* is a function of a probabilistic forecast and its corresponding outcome intended to measure predictive performance. Due to the probabilistic nature of the forecasts, scores are only meaningful when multiple forecasts and outcomes are considered and thus, commonly, the mean or median score is 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 possible gain in accuracy were a perfect point forecast to be issued, i.e. were the forecast able to predict the outcome perfectly [Murphy and Daan, 1985]. The aim of a skill score is to give some context to the gain in skill achieved from using one forecasting system over another reference one. Whilst the skill score form of a scoring rule is intended to yield an intuitive measure of the relative skill of two forecasting systems, it is argued in this paper that a number of shortcomings tend to outweigh the benefits of taking this approach.

In the weather forecasting literature, scores of probabilistic forecasts are often converted into skill score form [Siegert et al., 2011, Christensen et al., , Weigel et al., 2007, Wilks, 2001] before they are presented. This approach is also commonly used 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) [ECMWF,] and the UK Met-Office [Met Office,]. Although most commonly used in the forecasting of physical systems such as the weather, skill scores have also been used in a wide range of fields such as macroeconomic forecasting [Lahiri and Wang, 2013],[Bluedorn et al., 2016], forecasting of baseball [Richards, 2014] and association football [Haave and Høiland, 2017] and in medicine [Karoly et al., 2017].

In this paper, firstly, it is argued that, in the context of simulation models, since the presence of observational uncertainty in an initial condition makes a perfect point forecast impossible, 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 were present in the initial condition from which the forecast were launched. This is arguably not a useful measure, however, as the observational noise is usually out of the control of the forecaster. Secondly, using a number of examples, it is shown that the skill score form of a number of scoring rules is biased when a finite number of forecasts and outcomes are evaluated. This particular criticism is common both to point and probabilistic forecasts and is demonstrated in both cases. Finally, whether the proportion of possible skill gained has a useful interpretation regarding the relative value of two forecasting systems is called into question.

This paper is organised as follows. In section 2, background methodology describing scoring rules and skill scores is presented. In section 3, the relevance of the 'optimal' score necessary to calculate the skill score form is discussed and it is argued that, in some cases, the optimal score renders the skill score

form unusable and, in many cases, unachievable without improving the accuracy of the observations (which is not usually an option for a forecaster). In section 4, it is demonstrated analytically that the skill score form of the mean squared error (for point forecasts) can be biased. Using an empirical example, it is then shown that the skill score form of the Ignorance and Brier scores can also give biased results. Section 5 discusses whether the skill score form of scoring rules have a useful interpretation and section 6 is used for discussion and conclusions.

2 Background Definitions

2.1 Evaluating Point Forecasts

Although this paper is mostly concerned 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, described by

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

and the mean absolute error, described by

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

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

2.2 Scoring Rules

A *scoring rule* is a function of a probabilistic forecast and its outcome that evaluates forecast performance. Since scoring rules consider only probabilistic forecasts, this means that measures of the performance of point forecasts such as the mean squared error do not fall under the definition of a scoring rule. By convention, scoring rules are defined to be negatively oriented, that is lower scores imply better forecast accuracy. Many scoring rules have been proposed over the years and the choice of which to use to evaluate a set of forecasts is of great importance.

A scoring rule is *proper* if 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 be proper since, otherwise, there would be no incentive to choose a perfect forecasting system, were one 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, only proper scoring rules are considered in this paper. One particular score that fits this requirement is the ignorance score, introduced by I.J Good in 1951, [Good, 1952, Roulston and Smith, 2002] defined, for discrete forecasts, by

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

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 and the ignorance is thus defined by

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

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 to evaluate the performance of binary probabilistic forecasts. It is given[1] by

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

where $p(Y)$ represents the forecast probability and Y is 1 or 0 if the event did or didn't occur respectively. 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 for measures of forecast accuracy to be expressed in the form of a skill score [Murphy and Epstein, 1989, Siegert et al., 2011, Christensen et al., , Tödter and Ahrens, 2012]. A *skill score* is defined as

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

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

Tab. 1: Values of A_p for discrete and continuous (when 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

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 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 Defining a 'Perfect Score'

The skill score representation of a measure of forecast accuracy, as defined in equation 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 and is, in fact, a property of the measure of accuracy itself. Values of A_p for discrete and continuous (when applicable) forecasts for each scoring rule considered in this paper are shown in table 1. Note that, for the ignorance score, in the continuous case, A_p is infinite and thus the skill score representation is not informative in this case. The Brier score is only defined for binary categorical forecasts.

When forecasting is performed using deterministic simulation models, the existence of observational uncertainty in an initial condition prevents point forecasts from being perfect (i.e. consistently predicting the exact outcome), regardless of the accuracy of the forecasting system. In addition, all real world models will contain some degree of structural error. Inasmuch as observational uncertainty can be considered an unavoidable feature of the

real world, limitations to predictability resulting from this factor should, arguably, be differentiated from limitations stemming from the forecasting system itself. Were this not the case, the impression that the forecasts could be further improved upon could be given even if the forecasting system were already as informative as it could possibly be, given the information available. In the presence of observational noise, the best possible forecast would 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 would be achieved by evolving forward the distribution of possible initial conditions that are consistent with both the observation of the initial condition and with the system dynamics. Some uncertainty regarding the outcome will therefore remain. As a result, the best possible score given the observations available would come from a perfect *probabilistic* forecast rather than a perfect point forecast and thus the optimal score A_p would be unattainable. A more useful quantity, if available, would be

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

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. Since the skill of a perfect probabilistic forecast is never expected to be known, however, SSprob is never available in practice. The proportion of potential skill gained is thus never expected to be available for probabilistic forecasts. This places question marks over the value of the skill score representation of scoring rules.

In some studies and applications, instead of using ensembles to construct probabilistic forecasts, the ensemble mean is calculated and treated as a point forecast. This is arguably ill advised since, under this approach, important information regarding the shape of the distribution is discarded. Worse still is that, when the dynamics of the model are nonlinear, the mean of the forecast distribution is often an unlikely quantity. Consider a forecast distribution of the waiting time between eruptions of 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 short waiting time is likely. A point forecast

based on the mean, on the other hand, would make a prediction somewhere between the two; a relatively unlikely outcome. Nonetheless, it is still worth pointing out that, in such situations, since the outcomes 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, however, it is often the case that statistical models are applied. For example, in linear regression, the resulting forecast is a single point estimate that forms the mean of some Gaussian forecast distribution. In theory, however, the simulation model approach taken in numerical weather prediction could be applied to such cases though, in practice, it may be the case that statistical models are 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 for the outcome of a football match. Generally, forecasting of this kind is done using statistical approaches. In theory, however, if one knew the dynamics of the world perfectly and had perfect observations of its exact state at a given time, it would be possible to make a perfect point forecast of the state of the world [Laplace, 2012] and therefore the outcome of that match. Once the assumption that perfect observations are available is removed, however, the very best forecast of that match would, again, be a probability distribution even with a perfect model [Frigg et al., 2014]. Obtaining perfect observations of the world is, of course, out of the control of the forecaster (and impossible in practice) and thus, as discussed above, skill scores do not represent the proportion of possible skill achieved by the forecaster.

4 Sampling Distributions

Whilst, so far, general properties of scoring rules and their skill score form have been considered, the question of how each behaves in the context of a finite sample is also of importance. In practice, any measure of forecast

accuracy is calculated over a finite sample of forecasts and outcomes. The skill score form of a scoring rule aims to give a direct comparison between the skill of two forecasting systems. It is shown in this section that the skill score form of a measure of forecast accuracy can be biased for finite samples. First, this is demonstrated analytically using a point forecasting example in which the mean squared error is used as the measure of accuracy. This is then demonstrated with 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 of two forecasting systems which is shown to be unbiased.

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

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

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

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

and, when $A_p = 0$,

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

The skill score form of a measure of accuracy is thus a simple transformation of the relative skill. The relative skill is an unbiased estimator of the actual difference in skill. It is shown, however, 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). To demonstrate this, a simple example using point forecasts is presented.

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 $N(Y_i, 1)$ centred on the outcome Y_i. Both forecasting systems are thus expected, on average, to have the same mean squared error. 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 $E(\text{MSESS}) = 1 - \frac{N}{N-2}$. Since both forecasting systems are defined to have the same mean squared error, 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 but it can, however, be estimated. This is now done for both the ignorance and Brier scores in a special case in which both probabilistic forecasting systems are expected to have the same skill. Define $\boldsymbol{p}_f = p_{f,1}, .., p_{f,N}$ and $\boldsymbol{p}_r = p_{r,1}, .., p_{r,N}$ to be two sets of iid random draws from a standard uniform distribution $U(0, 1)$. Let each of $p_{f,i}$ 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 has an equal chance of coinciding with the true probability. Given that it is not known with which 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, both the relative skill and the skill score form of any evaluation measure are both zero. To test whether there is any bias in either measure for finite samples, sets of forecasts and outcomes of size N are randomly drawn and both the relative skill and the skill score calculated. The mean of each is then calculated to give an estimate of the expected value and thus the bias. This is repeated for various values of N. The results of the experiment are shown in figure 1. In the top panel, the estimated bias of the relative skill and the skill score form of the ignorance are shown whilst these are both shown for the Brier score in the lower panel. Here, it is clear that the skill score form of both scoring rules is biased whilst this 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

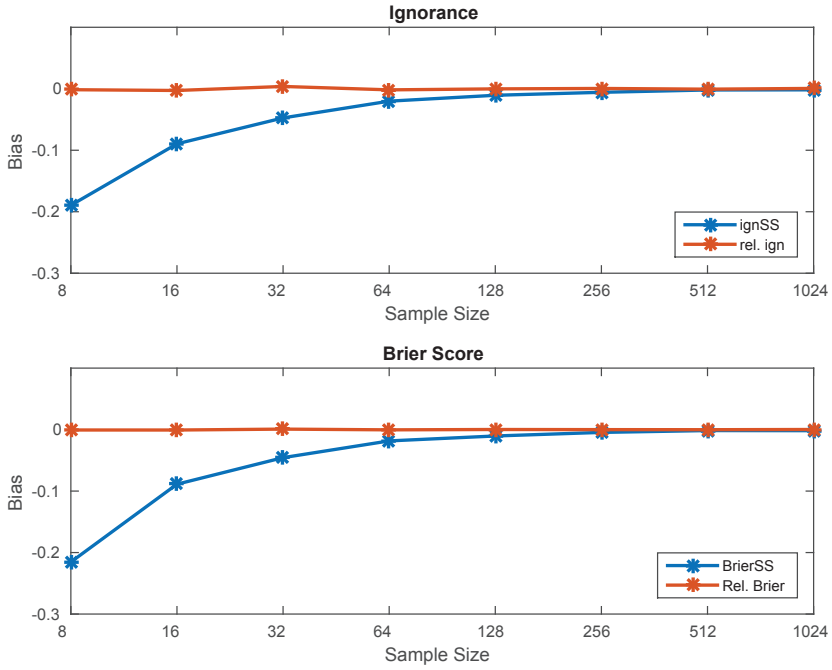

Fig. 1: The estimated bias of the relative skill (red) and the skill score form (blue) of the ignorance (top panel) and the Brier Score (lower panel) as a function of the sample size.

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 would mean 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. The skill score form of a measure of accuracy, however, as described in equation 6, was first suggested for the evaluation of point forecasts

[Murphy and Daan, 1985]. Although this paper is mostly concerned with skill scores in the context of probabilistic forecasts, for comparison, it is useful to illustrate the intended interpretation of skill scores in the context of point forecasts. Consider the mean squared error and mean absolute error, defined in section 2.1. In both cases, the value A_p in equation 6 is zero since, for an optimal point forecast, the forecast and the outcome would coincide. If the mean absolute error achieved from the forecasting system of interest and the reference forecasting system is 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 error of the forecast system of interest and reference forecasting system is, again, 3 and 4 respectively. The mean squared error skill score would then also be 0.25. However, a 25 percent reduction in the mean squared error is harder to interpret than a 25 percent reduction in the mean absolute error. This is arguably because the mean squared error has a less intuitive interpretation in the first place and so forcing it into skill score form adds little or nothing of value and arguably makes it less intuitive.
The nature of probabilistic forecasts means that the evaluation techniques described above can not be applied (without compromising the information content of the forecast) and thus a scoring rule is required. Whilst, as previously discussed, the skill score form of some distance metrics like the mean absolute error can have a simple and useful interpretation, this is not necessarily the case for scoring rules. Consider the ignorance score for example. 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 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. The proportion of possible bits gained, however, should not be considered a linear gain in value. In fact, using the skill score form alone, the gain in probability or density placed on the outcome by using one forecasting system over another cannot be recovered. Arguably, then, it does not 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. Arguably, a weakness of the Brier score over the ignorance score is that the former has a far less clear interpretation and, if there is no useful interpretation in the score in the first place, transforming it into skill score form will not create one.
Scoring rules such as the Ignorance and the Brier Score are clear, mathematically precise and informative scores. Forcing them into skill score form destroys this utility. There are no persuasive arguments establishing that there is any benefit in doing so. From the above, it should also be clear that skill scores based on different measures of accuracy cannot be directly compared, even though they are forced to be on the same scale.

6 Discussion

The skill score form of a forecast evaluation metric is designed to give an intuitive measure of the gain in skill achieved by using one forecasting system over another. Whilst a skill score represents the mean gain in accuracy 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 in interpreting this as the gain in skill that is achievable. When observational noise is present in the initial condition, even a perfect forecasting system cannot yield a perfect point forecast and thus the skill score does not represent the proportion of possible skill that could be gained by improving the forecasting system. In addition, even if it were possible for the forecasting system to yield a perfect point forecast, it is unclear that the proportion of potential skill gained represents a useful indication of the proportion of actual value 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 from using one forecasting system over another. In many cases, the intended interpretation is arguably misleading as a measure of the proportion of possible skill gained. Combined with the fact that the skill score form can introduce a bias to the score, it should be treated

with caution. Pressure to express forecast system evaluation in terms of skill scores is found to be misplaced; in the lack of better motivation, forecast evaluation can be more effective when considered in terms of raw scores with more meaningful units.

Acknowledgements

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

References

References

[Alkema et al., 2015] Alkema, L., Gerland, P., Raftery, A., and Wilmoth, J. (2015). The united nations probabilistic population projections: an introduction to demographic forecasting with uncertainty. *Foresight (Colchester, Vt.)*, 2015(37):19.

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

[Böse et al., 2017] Böse, J.-H., Flunkert, V., Gasthaus, J., Januschowski, T., Lange, D., Salinas, D., Schelter, S., Seeger, M., and Wang, Y. (2017). Probabilistic demand forecasting at scale. *Proc. VLDB Endow.*, 10(12):1694–1705.

[Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1.

[Bröcker and Smith, 2008] Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4):663–678.

[Christensen et al.,] Christensen, H. M., Moroz, I. M., and Palmer, T. N. Evaluation of ensemble forecast uncertainty using a new proper score: Ap-

plication to medium-range and seasonal forecasts. *Quarterly Journal of the Royal Meteorological Society*, 141(687).

[Constantinou et al., 2012] Constantinou, A. C., Fenton, N. E., and Neil, M. (2012). pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322 – 339.

[ECMWF,] ECMWF. Quality of our forecasts. `https://www.ecmwf.int/en/forecasts/quality-our-forecasts`. Accessed: 29/09/2018.

[Frigg et al., 2014] Frigg, R., Bradley, S., Du, H., and Smith, L. A. (2014). Laplaces demon and the adventures of his apprentices. *Philosophy of Science*, 81(1):31–59.

[Good, 1952] Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14:107–114.

[Haave and Høiland, 2017] Haave, H. S. and Høiland, H. (2017). Evaluating association football player performances using markov models. Master's thesis, NTNU.

[Hansen et al., 2011] Hansen, J. W., Mason, S. J., Sun, L., and Tall, A. (2011). Review of seasonal climate forecasting for agriculture in sub-saharan africa. *Experimental Agriculture*, 47(2):205–240.

[Hastings et al., 1993] Hastings, A., Hom, C. L., Ellner, S., Turchin, P., and 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 et al., 2017] Karoly, P. J., Ung, H., Grayden, D. B., Kuhlmann, L., Leyde, K., Cook, M. J., and Freestone, D. R. (2017). The circadian profile of epilepsy improves seizure forecasting. *Brain*, 140(8):2169–2182.

[Katz and Lazo, 2011] Katz, R. W. and Lazo, J. K. (2011). Economic value of weather and climate forecasts. In *The Oxford Handbook of Economic Forecasting*.

[Lahiri and Wang, 2013] Lahiri, K. and Wang, J. G. (2013). Evaluating probability forecasts for gdp declines using alternative methodologies. *International Journal of Forecasting*, 29(1):175–190.

[Laplace, 2012] Laplace, P.-S. (2012). *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media.

[Lorenz, 1963] Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, 20(2):130–141.

[Met Office,] Met Office. 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`. Accessed: 29/09/2018.

[Murphy and Epstein, 1989] Murphy, A. and Epstein, E. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117(3):572–582.

[Murphy and Daan, 1985] Murphy, H. and Daan, H. (1985). Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences*, pages 379–437.

[Richards, 2014] Richards, J. A. (2014). Probabilities of victory in head-to-head team matchups. *Fall 2014 Baseball Research Journal*, 43(2).

[Rinehart, 1969] Rinehart, J. (1969). Old faithful geyser performance 1870 through 1966. *Bulletin Volcanologique*, 33(1):153–163.

[Roulston and Smith, 2002] Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660.

[Siegert et al., 2011] Siegert, S., Bröcker, J., and Kantz, H. (2011). Predicting outliers in ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 137(660):1887–1897.

[Smith and Judd, 2001] Smith, L. and Judd, K. (2001). Indistinguishable states i. perfect model scenario. *Physica D*, 151:125–141.

[Smith and Judd, 2004] Smith, L. and Judd, K. (2004). Indistinguishable states ii. imperfect model scenario. *Physica D*, 196:224–242.

[Smith, 1993] Smith, L. K. (1993). The influence of weather and climate on recreation and tourism. *Weather*, 48(12):398–404.

[Smith and Beven, 2014] Smith, P. J. and Beven, K. J. (2014). When to issue a flood warning: towards a risk-based approach based on real time probabilistic forecasts. In *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, pages 1395–1404.

[Strogatz, 2018] Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press.

[Tödter and Ahrens, 2012] Tödter, J. and Ahrens, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review*, 140(6):2005–2017.

[Weigel et al., 2007] Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2007). The discrete brier and ranked probability skill scores. *Monthly Weather Review*, 135(1):118–124.

[Wilks, 2001] Wilks, D. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8(2):209–219.

[Ziel and Steinert, 2018] Ziel, F. and Steinert, R. (2018). Probabilistic mid- and long-term electricity price forecasting. *Renewable and Sustainable Energy Reviews*, 94:251–266.

References

[Alkema et al., 2015] Alkema, L., Gerland, P., Raftery, A., and Wilmoth, J. (2015). The united nations probabilistic population projections: an introduction to demographic forecasting with uncertainty. *Foresight (Colchester, Vt.)*, 2015(37):19.

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

[Böse et al., 2017] Böse, J.-H., Flunkert, V., Gasthaus, J., Januschowski, T., Lange, D., Salinas, D., Schelter, S., Seeger, M., and Wang, Y. (2017). Probabilistic demand forecasting at scale. *Proc. VLDB Endow.*, 10(12):1694–1705.

[Brier, 1950] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78:1.

[Bröcker and Smith, 2008] Bröcker, J. and Smith, L. A. (2008). From ensemble forecasts to predictive distribution functions. *Tellus A*, 60(4):663–678.

[Christensen et al.,] Christensen, H. M., Moroz, I. M., and Palmer, T. N. 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).

[Constantinou et al., 2012] Constantinou, A. C., Fenton, N. E., and Neil, M. (2012). pi-football: A bayesian network model for forecasting association football match outcomes. *Knowledge-Based Systems*, 36:322 – 339.

[ECMWF,] ECMWF. Quality of our forecasts. `https://www.ecmwf.int/en/forecasts/quality-our-forecasts`. Accessed: 29/09/2018.

[Frigg et al., 2014] Frigg, R., Bradley, S., Du, H., and Smith, L. A. (2014). Laplaces demon and the adventures of his apprentices. *Philosophy of Science*, 81(1):31–59.

[Good, 1952] Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society: Series B*, 14:107–114.

[Haave and Høiland, 2017] Haave, H. S. and Høiland, H. (2017). Evaluating association football player performances using markov models. Master's thesis, NTNU.

[Hansen et al., 2011] Hansen, J. W., Mason, S. J., Sun, L., and Tall, A. (2011). Review of seasonal climate forecasting for agriculture in sub-saharan africa. *Experimental Agriculture*, 47(2):205–240.

[Hastings et al., 1993] Hastings, A., Hom, C. L., Ellner, S., Turchin, P., and 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 et al., 2017] Karoly, P. J., Ung, H., Grayden, D. B., Kuhlmann, L., Leyde, K., Cook, M. J., and Freestone, D. R. (2017). The circadian profile of epilepsy improves seizure forecasting. *Brain*, 140(8):2169–2182.

[Katz and Lazo, 2011] Katz, R. W. and Lazo, J. K. (2011). Economic value of weather and climate forecasts. In *The Oxford Handbook of Economic Forecasting*.

[Lahiri and Wang, 2013] Lahiri, K. and Wang, J. G. (2013). Evaluating probability forecasts for gdp declines using alternative methodologies. *International Journal of Forecasting*, 29(1):175–190.

[Laplace, 2012] Laplace, P.-S. (2012). *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*, volume 13. Springer Science & Business Media.

[Lorenz, 1963] Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *J. Atmos. Sci.*, 20(2):130–141.

[Met Office,] Met Office. 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`. Accessed: 29/09/2018.

[Murphy and Epstein, 1989] Murphy, A. and Epstein, E. (1989). Skill scores and correlation coefficients in model verification. *Monthly Weather Review*, 117(3):572–582.

[Murphy and Daan, 1985] Murphy, H. and Daan, H. (1985). Forecast evaluation. *Probability, Statistics and Decision Making in the Atmospheric Sciences*, pages 379–437.

[Richards, 2014] Richards, J. A. (2014). Probabilities of victory in head-to-head team matchups. *Fall 2014 Baseball Research Journal*, 43(2).

[Rinehart, 1969] Rinehart, J. (1969). Old faithful geyser performance 1870 through 1966. *Bulletin Volcanologique*, 33(1):153–163.

[Roulston and Smith, 2002] Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660.

[Siegert et al., 2011] Siegert, S., Bröcker, J., and Kantz, H. (2011). Predicting outliers in ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 137(660):1887–1897.

[Smith and Judd, 2001] Smith, L. and Judd, K. (2001). Indistinguishable states i. perfect model scenario. *Physica D*, 151:125–141.

[Smith and Judd, 2004] Smith, L. and Judd, K. (2004). Indistinguishable states ii. imperfect model scenario. *Physica D*, 196:224–242.

[Smith, 1993] Smith, L. K. (1993). The influence of weather and climate on recreation and tourism. *Weather*, 48(12):398–404.

[Smith and Beven, 2014] Smith, P. J. and Beven, K. J. (2014). When to issue a flood warning: towards a risk-based approach based on real time probabilistic forecasts. In *Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management*, pages 1395–1404.

[Strogatz, 2018] Strogatz, S. H. (2018). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC Press.

[Tödter and Ahrens, 2012] Tödter, J. and Ahrens, B. (2012). Generalization of the ignorance score: Continuous ranked version and its decomposition. *Monthly Weather Review*, 140(6):2005–2017.

[Weigel et al., 2007] Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2007). The discrete brier and ranked probability skill scores. *Monthly Weather Review*, 135(1):118–124.

[Wilks, 2001] Wilks, D. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8(2):209–219.

[Ziel and Steinert, 2018] Ziel, F. and Steinert, R. (2018). Probabilistic mid- and long-term electricity price forecasting. *Renewable and Sustainable Energy Reviews*, 94:251–266.