

Beyond Strictly Proper Scoring Rules: The Importance of Being Local

Hailiang Du^{1,2}

¹Department of Mathematical Sciences,
Durham University, Durham, DH1 3LE, U.K.

²Centre for the Analysis of Time Series,
London School of Economics, London WC2A 2AE. UK

Email: hailiang.du@durham.ac.uk

December 24, 2020

Abstract

The evaluation of probabilistic forecasts plays a central role both in the interpretation and in the use of forecast systems and their development. Probabilistic scores (scoring rules) provide statistical measures to assess the quality of probabilistic forecasts. Often, many probabilistic forecast systems are available while evaluations of their performance are not standardized, with different scoring rules being used to measure different aspects of forecast performance. Even when the discussion is restricted to strictly proper scoring rules, there remains considerable variability between them; indeed strictly proper scoring rules need not rank competing forecast systems in the same order when none of these systems are perfect. The locality property is explored to further distinguish scoring rules. The nonlocal strictly proper scoring rules considered are shown to have a property that can produce “unfortunate” evaluations. Particularly the fact that Continuous Rank Probability Score prefers the outcome close to the median of the forecast distribution regardless the probability mass assigned to the value at/near the median raises concern to its use. The only local strictly proper scoring rules, the logarithmic score, has direct interpretations in terms of probabilities and bits of information. The nonlocal strictly proper scoring rules, on the other hand, lack meaningful direct interpretation for decision support. The logarithmic score is also shown to be invariant under smooth trans-

formation of the forecast variable, while the nonlocal strictly proper scoring rules considered may, however, change their preferences due to the transformation. It is therefore suggested that the logarithmic score always be included in the evaluation of probabilistic forecasts.

Keywords: scoring rule; forecast evaluation; skill score.

1 Introduction

Forecast evaluation has a long history of being a crucial topic for model development and decision support. The outputs from a stochastic model can be naturally interpreted in the form of probabilistic forecast. Given a deterministic model, uncertainty in the initial state due to the observational noise; limited computational power; and model discrepancy prevent one from making a perfect deterministic forecast of the future or even identifying the Truth in the past. In order to account for all sorts of uncertainties, the model outputs are often interpreted as probabilistic forecasts with the aim of providing useful information for decision support. Probabilistic forecasts have been widely adopted in various fields including meteorology, social science, pharmacology, economics and finance; and have become common in operational forecasting over the last quarter century.

The evaluation of probabilistic forecasts plays a central role both in the interpretation and in the use of forecast systems and their development. Such evaluation has not yet been standardized, with many different probabilistic scoring rules [14, 20, 36, 51] being used. As probabilistic forecasts become more common, the need to select (probabilistic) scoring rule(s) for constructing probabilistic forecasts, calibrating forecast systems, ranking competing forecast systems and quantifying forecast improvement has led to the research work presented in this paper.

The importance of using strictly proper scoring rules has been noted in the literature [6], as only strictly proper scoring rules encourage the forecaster to be honest, i.e. reporting a forecast probability distribution gives an optimal expected score only when the verification is, in fact, drawn from that probability distribution. When the discussion is restricted to strictly proper scoring rules, however, there remains considerable variability between scoring rules (there are, in fact, an infinite number of strictly proper scoring rules). And strictly proper scoring rules need not rank competing forecast systems in the same order when none of these systems are perfect.

The locality property is explored to further distinguish various strictly proper scoring rules. A

property that reflects “unfortunate”¹ evaluations is introduced. Nonlocal strictly proper scoring rules considered are shown to have a mathematical property, named implausible, that could produce “unfortunate” evaluations. A few striking examples of the potential issues that result from the use of nonlocal scoring rules are presented. The only local strictly proper scoring rule, the logarithmic score (also known as Ignorance), has direct interpretations in terms of probabilities and bits of information. The nonlocal strictly proper scoring rules are found to lack meaningful direct interpretation. The logarithmic score is also shown to be invariant under smooth transformation of the forecast variable, while the nonlocal strictly proper scoring rules considered may, however, change their preferences due to the transformation.

This paper emphasizes the fact that being strictly proper is not sufficient in decision support when measuring the difference between imperfect forecast systems and suggests that the only local scoring rule, Ignorance, should always be included in the evaluation of probabilistic forecasts.

The definition of a scoring rule for probabilistic forecast and the importance of using strictly proper scoring rules are presented in Section 2. A number of strictly proper scoring rules are defined in Section 3. A common example of strictly proper scoring rules ranking forecast systems differently without the presence of True underlying distribution is given in Section 4. The locality property is defined and discussed in Section 5. Section 6 introduces a mathematical property that reflects “unfortunate” evaluations and shows nonlocal scoring rules considered have such property. The interpretation of local and nonlocal scoring rules are discussed in Section 7. Section 8 investigates the behavior of proper scoring rules when smooth transformation is applied to the forecast variable. Section 8 provides discussion and conclusions.

2 Probabilistic Scoring Rules and Importance of Being Strictly Proper

While the true value of a forecast is most clearly reflected in its utility to the end user, probabilistic scores are fundamental to the performance analysis of probabilistic forecasts. Ideally they provide a general measure of future forecast quality, independent of any specific end user [6]. A probabilistic score (scoring rule) is a function $S(p(x), Y)$, where $p(x)$ is a probability density

¹Fortuna was the goddess of fortune (luck) in Roman religion. “Unfortunate” in this paper refers to bad advice, which is a disaster in terms of decision support.

function and Y is the outcome. In this paper, probabilistic forecasts in the form of probability density functions (PDFs) $p(x)$ are considered². The notation $p(x)$ denotes the entire function, while $p(Y)$ always denotes the value of the function at the particular outcome Y . By convention, a lower score is taken to reflect a better forecast. Analytically, the *expected score*,

$$E(S(p(x), Y)) = \int S(p(x), Y)Q(Y)dY, \quad (1)$$

which takes the expectation of the scoring rule under the True underlying distribution Q from which the outcome Y is drawn, quantifies the quality of a forecast system. In practice, an archive of forecast-outcome pairs is required to evaluate the quality of a forecast system. It contains a large number N of forecasts $\{p_i(x), i = 1, \dots, N\}$ and corresponding outcomes $\{Y_i, i = 1, \dots, N\}$. The forecast system yields an *empirical score*:

$$S_{emp} = \frac{1}{N} \sum_i^N S(p_i(x), Y_i). \quad (2)$$

Note the size of the forecast archive can play a major role in determining the significance of the result [26], regardless of which scoring rule is employed [6].

Several scoring rules are widely used for the evaluation of probabilistic forecasts [5, 11, 14, 16, 20, 27, 39, 51]; different scoring rules might quantify different attributes of the forecast. Note, however, that Good (1952)'s logarithmic score, also known as Ignorance, (defined below in Section 3d.) is the only scoring rule consistent with the use of (log) likelihoods to evaluate assessors or Bayesian inference [52, 53].

Since any functional form based on $p(x)$ and Y could be considered as a scoring rule, one may introduce and use a scoring rule that favors to particular forecast system which might lead to dishonest and misleading evaluations where the scoring rule encourages the forecaster to select a probabilistic forecast distribution that the forecaster knows is not correct (For example the well-known Finley (1884) tornado forecasts [33, 44]). To avoid such dishonest evaluation, strictly proper [35, 47] scoring rules (defined in the following) are preferred. The term, *proper*, was first introduced by [54], while the general idea goes back to [5] and [16]. A scoring rule, $S(p(x), Y)$, is said to be *proper* if inequality (3) holds for any pair of forecast PDFs, and *strictly proper* when equality implies $p = q$:

$$\int q(z)S(p(x), z)dz \geq \int q(z)S(q(x), z)dz. \quad (3)$$

²Results and conclusions presented in this paper also apply to probabilistic forecasts in the form of probability mass functions in the context of categorical variables.

For a given forecast p , a scoring rule evaluated at the outcome is a random variable with values that depend on the outcome Y . Note being strictly proper is a property of the functional form of the scoring rule alone, not of the particular distribution $p(x)$ or $q(x)$. Strictly proper scoring rules give a probabilistic forecast distribution an optimal expected score only when the outcome is, in fact, drawn from that probability distribution [6]. In expectation, a strictly proper scoring rule does not judge any other forecast p to score better than q as a forecast of q itself. Note that the interpretation of strictly proper does not, however, require one to believe that the True underlying distribution Q exists. Strictly proper is a property of the scoring rule; it is neither necessary to assume that Y is drawn from any kind of True distribution nor that any kind of data is to hand.

The question of whether the employed scoring rule is strictly proper or not can be answered independently of any data being considered [6]. Although concerns of hedging are often mentioned[39], strictly proper scoring rules are preferred even when there is no human involvement, as in parameter selection [9].

While the importance of using strictly proper scoring rules is well recognized [6, 7, 12], researchers often face requests to present results under a variety of scoring rules, both proper and nonproper scoring rules. The fact that nonproper scoring rules like Root Mean Squared Error (RMSE) are still widely used in forecast evaluation often leads to confusion and poorly optimized forecast systems. There have been many discussions regarding the evil of RMSE in the literature (see [6, 30, 41, 51]), therefore RMSE, which is in fact not a strictly proper scoring rule, will not be considered in this paper.

3 Strictly Proper Scoring Rules

A variety of strictly proper scoring rules have been introduced since the 1950s. Some of those widely used are listed below:

3.1 Energy Score Family

[14] introduced the Energy Score family based on [45] statistical energy perspective. The Energy Score family S_{ES} , is defined as follows³:

$$S_{ES}(p(x), Y) = E_p \|x - Y\|^\beta - \frac{1}{2} E_p \|x - x'\|^\beta, \quad (4)$$

where $\beta \in (0, 2)$ is a real number; x and x' are independent copies of a random vector with distribution p ; and $\|\cdot\|$ denotes the Euclidean norm. [45] and [14] show that the Energy Score is strictly proper relative to the class \mathbb{P}_β , where \mathbb{P}_β denotes the class of the Borel probability measures p such that $E_p \|x\|^\beta$ is finite. When $\beta = 1$, one obtains:

$$S_{CRPS}(p(x), Y) = E_p \|x - Y\| - \frac{1}{2} E_p \|x - x'\|. \quad (5)$$

This is equivalent to the well-known Continuous Ranked Probability Score⁴ (CRPS) [28, 49] (and see [1, 14, 46] for the proof of equivalence), where it is the integral of the square of the L^2 distance between the cumulative distribution function (CDF) of the forecast p and a step function at the outcome [11],

$$S_{CRPS}(p(x), Y) = \int \left(\int_{-\infty}^x p(z) dz - H(x - Y) \right)^2 dx, \quad (6)$$

where the Heaviside (step) function H is defined as follows:

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \quad (7)$$

The CRPS was to our knowledge first published by [8]. It can also be considered as a generalization of the Brier Score [5] (the Brier Score only applies to binary outcomes [28, 34]). For a point forecast, the CRPS is equal to the mean absolute error. In the past decade, the CRPS has been widely used by the atmospheric sciences community [15, 38, 55].

3.2 Power Score Family

Let α be a real number with $\alpha > 1$. The Power Score family [39] S_{PS} , is defined as follows:

$$S_{PS}(p(x), Y) = -\alpha p(Y)^{\alpha-1} + (\alpha - 1) \int p^\alpha(z) dz, \quad (8)$$

³Negative orientation is applied to the original Energy Score defined by [14] so that it is consistent with the convention that a lower score reflects a better forecast.

⁴The Continuous Ranked Probability Score is generalized from the Ranked Probability Score (RPS) [11, 31] which is widely used to evaluate discrete (categorical) probabilistic forecasts.

The Power Score family is also strictly proper; this can simply be derived from the derivatives of the expected score [39]. When $\alpha = 2$, one obtains the Proper Linear Score (PLS) (also called the Quadratic Score [5]):

$$S_{PLS}(p(x), Y) = -2p(Y) + \int p^2(z)dz, \quad (9)$$

PLS derives from the (Naive) Linear Score [43], $S_{LS}(p(x), Y) = -p(Y)$, which is not a proper scoring rule as the (Naive) Linear Score favors a $p(x)$ featuring a very small spread and which is centered at the point x^* for which $Q(x^*)$ is very large [6].

3.3 Pseudo-spherical Score Family

[17] introduced the Pseudo-spherical Score family S_{PSS} (β is a real number with $\beta > 1$), defined as follows:

$$S_{PSS}(p(x), Y) = -\frac{p(Y)^{\beta-1}}{(\int p^\beta(z)dz)^{1/\beta}}. \quad (10)$$

The Pseudo-spherical Score family is strictly proper; this can be derived using Hölder and Minkowski inequality. When $\beta = 2$, one obtains the traditional Spherical Score (SPS):

$$S_{SPS}(p(x), Y) = -\frac{p(Y)}{(\int p^2(z)dz)^{1/2}}. \quad (11)$$

3.4 Ignorance

[16] introduced the logarithmic score (also known as Ignorance [36]) given by⁵:

$$S(p(x), Y) = -\log_2(p(Y)), \quad (12)$$

where $p(Y)$ is the density assigned to the outcome Y . Ignorance (IGN) is a strictly proper scoring rule; this can be derived using Kullback-Leibler inequality [24]. The expected (with respect to p) IGN is also a famous information measure, Shannon entropy. In addition, the expected IGN of p relative to a distribution q becomes the classical Kullback-Leibler divergence [23].

⁵Note that defining the logarithmic score in terms of \log_2 is equivalent to the alternative definition in terms of \ln up to the factor $1/\ln 2$ which does not affect rankings of different forecast systems.

4 Different Strictly Proper Scoring Rules Rank Forecast Systems Differently

Obviously the Energy Score family, Power Score family and Pseudo-spherical Score family contain an infinite number of strictly proper scoring rules. [50] have introduced weighted scoring rules by blending the Power Score with the Pseudo-spherical Score; the weighted scoring rule is shown to be strictly proper too. Furthermore, [47] proved that a linear transformation of a strictly proper scoring rule is also strictly proper. Given a strictly proper scoring rule, a forecast system providing Q will always be preferred whenever it is included amongst those under consideration. When none of the competing forecast systems are perfect, then even strictly proper scoring rules may rank two forecast systems differently, making it impossible to provide definitive statements regarding the relative merit of imperfect forecast systems without considering an additional measure of forecast quality.

Consider the case where outcomes are independent random draws from a standard Gaussian distribution. Two forecast systems are constructed, where forecast system A uses $N(0, \sigma^2)$ and forecast system B uses $N(0, 1/\sigma^2)$ where $\sigma > 1$. Obviously neither of the forecast systems is perfect; forecast system A represents a wider distribution around 0 with larger standard deviation while forecast system B represents a narrower distribution with smaller standard deviation. Figure 1 shows the expectation (under the True distribution, $N(0, 1)$) of various scoring rules (Ignorance⁶, Continuous Rank Probability Score, Proper Linear Score and Spherical Score) of forecast system A relative to forecast system B as a function of σ . If the relative score (also known as *skill score*) is negative, it indicates forecast system A outperforms forecast system B. Both IGN and PLS prefer wider⁷ forecast distribution (forecast system A) than narrower forecast distribution (forecast system B). The CRPS, on the contrary, ranks forecast system B higher than A. Interestingly SPS considers both imperfect forecast system as having the same forecast quality with the expected relative score being zero. (Note given any finite sample of forecasts, there is a 50% chance that the empirical SPS prefers forecast A (or B) to the other.) A more thorough investigation in contrasting how certain scoring rules would rank competing forecasts of specified departures from

⁶Ignorance is downscaled by a factor of 20 in order to have a similar scale with other scoring rules in Figure 1.

⁷This particular example, based on Gaussian distributions, is designed to show that different scoring rules may rank forecast systems differently and not to indicate whether each of the scoring rules considered here prefers wider or narrower forecast distributions in general, which is not true either.

the target distribution can be found in [25].

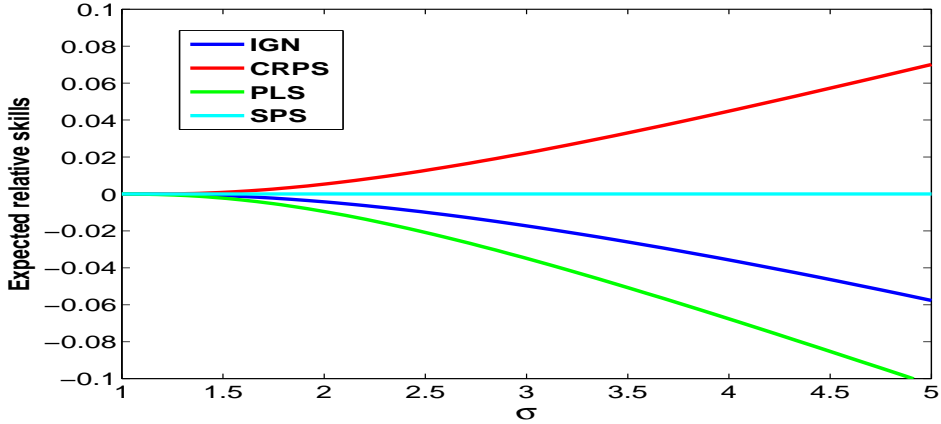


Figure 1: The expectation of various scoring rules of forecast system A, $N(0, \sigma^2)$, relative to forecast system B, $N(0, 1/\sigma^2)$, where the outcome is drawn from a standard Gaussian distribution. A negative relative score suggests system A outperforms system B.

5 Locality

To distinguish between strictly proper scoring rules, the locality property is explored here. A scoring rule is *local* if the probabilistic forecast is evaluated only at the actual outcome, which means that the scoring rule depends solely on the probability assigned to the outcome, rather than being rewarded for other features of the forecast distribution, such as its shape. [10] and [2] show that every local, smooth and proper scoring rule for continuous variables is equivalent to (an affine function of) IGN, which makes IGN the only proper local scoring rule for continuous variables. Thus all other proper scoring rules, including those listed in Section 3, are nonlocal. The locality property itself does not suggest whether local or nonlocal scoring rules should be preferred, although it might seem unreasonable that features of the forecast other than the value it assigned to the outcome should matter at all. In the following sections, the preference of a local scoring rule is supported based on both mathematical properties and interpretation of the scoring rule.

6 Implausible

A mathematical property called implausible is introduced in this section. Nonlocal scoring rules listed in Section 3 are shown to have such undesirable property; striking “unfortunate” evaluation examples that result from the use of nonlocal scoring rules are presented below.

A scoring rule is *implausible*⁸, if for ANY $r > 1, r \in \mathbb{R}$, there exist two forecast systems $p_1(x)$ and $p_2(x)$, and Y , where $p_1(Y)/p_2(Y) = r$ while $S(p_1, Y) > S(p_2, Y)$. In other words, an implausible scoring rule means that for all $r > 1$ it is possible to find $p_1(x)$, $p_2(x)$ and Y (which may all vary with r) such that $p_1(Y)/p_2(Y) = r$ and $S(p_1, Y) > S(p_2, Y)$. Ignorance is clearly not implausible as given $p_1(Y)/p_2(Y) = r$, $S(p_1, Y)$ would always be smaller than $S(p_2, Y)$ by $\log_2 r$. The Energy Score family is implausible; this can be shown via investigating an undesirable mathematical property of the Energy Score. Take the derivative of the Energy Score respect to the outcome Y (where Y is a realization of the random variable x):

$$\frac{\partial S_{ES}(p(x), Y)}{\partial Y} = \int_{-\infty}^Y \beta(Y - x)^{\beta-1} p(x) dx - \int_Y^{\infty} \beta(x - Y)^{\beta-1} p(x) dx \quad (13)$$

The zero solution of the RHS of Eq. 13 only relies on the location of Y regardless the value of $p(Y)$. For the CRPS, where $\beta = 1$, $\min_Y S(p, Y)$ is achieved when $\int_{-\infty}^Y p(x) dx - \int_Y^{\infty} p(x) dx = 0$ which gives Y as being the median of $p(x)$. Such mathematical property may lead to “unfortunate” results as illustrated in Figure 2.⁹ The blue line and red line represent two forecast systems A and B (each based on a Bimodal distribution with the same shape but different centers). Intuitively, one would expect that if the outcome lands between -0.5 and 0.5 (or more generally that the outcome is drawn from some PDF which is bounded between -0.5 and 0.5) forecast system B shall be preferred as system B would assign significantly more probability mass to the outcome than system A (especially when the outcome lands around 0); similarly if the outcome lands between 0.5 and 1.5 forecast system A shall be preferred. The green line represents the CRPS of system A relative to system B, a negative (below the dotted zero line) relative score suggests system A outperforms system B according to the CRPS. It appears that if the outcome lands between -0.5 and 0.5, the CRPS would prefer system A over B even when system B assigns significantly more

⁸[41] defined “perverse” scoring rules to be those which systematically prefer forecasts which place a lower probability on the outcome. [29] considered a scoring rule to be “not feasible” when a probable event scores worse than an improbable one. These definitions are elevated here.

⁹It is usual to analyze how scores change as the forecast varies, while this and the following examples investigate how the values of scores change as the outcome varies.

probability mass to the outcome than system A. This is due to the fact that the CRPS prefers the outcome to be close to the median of the forecast distribution no matter how much probability mass is around the median. Obviously the CRPS is implausible, as shown in Figure 2, when $Y = 1$, $p_A(Y)/p_B(Y) = \infty$ while $S_{CRPS}(p_A, Y) > S_{CRPS}(p_B, Y)$ (similar examples can be found to show all members of the Energy Score family are implausible). Ironically, if the forecast system A is delivered to the user, the developer of forecast system A would hope the outcome lands at 0 in order to achieve the best CRPS despite the fact the forecast system A assigns 0 probability to the outcome. Considering a parameter estimation scenario, if the observed outcomes are drawn from a delta function or a sharp Gaussian distribution centered at 0 and the forecast distribution is a Bimodal distribution with its center to be tuned, tuning the parameter based on the CRPS would converge to a Bimodal distribution centered at 0 where the probability mass assign to the outcome would always be near 0.

The example shown in Figure 2. contradicts the claim [4, 22, 48] that the CRPS/RPS gives credit for assigning high probabilities to the values near but not identical to the outcome. This kind of claims is mostly originated from [42], where Staël von Holstein shown that the RPS is “sensitive to distance” from the “true” outcome. Actually the “sensitive to distance” defined by Staël von Holstein is based on his definition of “more distant from the true event” (P360 of [42]), which is, however, NOT equivalent to assigning high probabilities to the values near but not identical to the outcome. It was, in fact, noted by Staël von Holstein himself (section 6 of [42]) that his definition of “more distant from the true event” is rather restrictive and changing the definition to an alternative definition [32] will lead to the RPS not being “sensitive to distance”, which is consistent with the example shown in Figure 2.

The Power Score and Spherical Score are also implausible. This can be shown in the case, where $p_1(x)$ and $p_2(x)$ are both Gaussian distributions.

Let $p_1(x)$ be a Gaussian distribution with mean u_1 and standard deviation σ_1 , then

$$\begin{aligned} \int_{-\infty}^{\infty} p_1^\alpha(z) dz &= \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(z-u_1)^2}{2\sigma_1^2}} \right)^\alpha dz \\ &= (2\pi)^{\frac{1-\alpha}{2}} \alpha^{-\frac{1}{2}} \sigma_1^{1-\alpha} \end{aligned} \tag{14}$$

Let $p_2(x)$ be a Gaussian distribution with mean u_2 and standard deviation σ_2 . To prove the Power Score family is implausible, one needs to find $p_1(\cdot)$, $p_2(\cdot)$ and Y so that $p_1(Y) = r p_2(Y)$

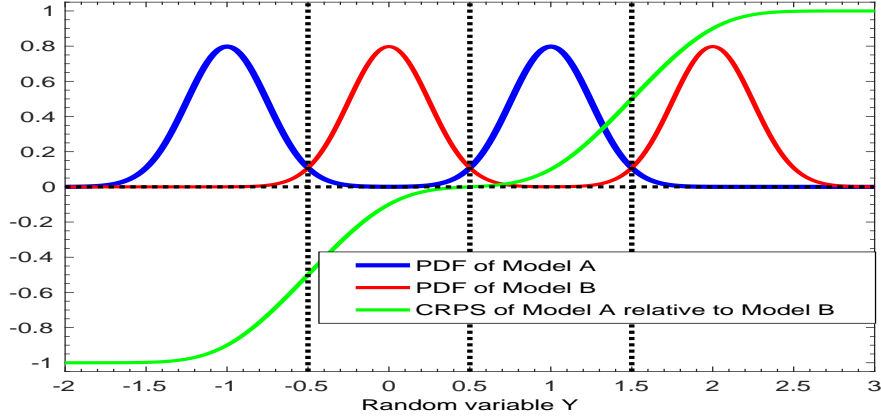


Figure 2: Example showing that the Continuous Rank Probability Score produces “unfortunate” results. The blue line and the red line represent PDFs of forecast systems A and B based on Bimodal distributions with the same shape but different centers. The green line represents the CRPS of system A relative to system B. A negative relative score suggests system A outperforms system B. The dashed vertical lines enclose the regions where “unfortunate” results occur.

but $S_{PS}(p_1(x), Y) > S_{PS}(p_2(x), Y)$, which requires:

$$\begin{aligned}
& -\alpha p_1(Y)^{\alpha-1} + (\alpha-1) \int_{-\infty}^{\infty} p_1^{\alpha}(z) dz > -\alpha p_2(Y)^{\alpha-1} + (\alpha-1) \int_{-\infty}^{\infty} p_2^{\alpha}(z) dz \\
& -\alpha p_1(Y)^{\alpha-1} + (\alpha-1)(2\pi)^{\frac{1-\alpha}{2}} \alpha^{-\frac{1}{2}} \sigma_1^{1-\alpha} > -\alpha p_2(Y)^{\alpha-1} + (\alpha-1)(2\pi)^{\frac{1-\alpha}{2}} \alpha^{-\frac{1}{2}} \sigma_2^{1-\alpha} \quad (15)
\end{aligned}$$

Note that even if $p_2(Y) = 0$, it is still possible that $S_{PS}(p_1(x), Y) > S_{PS}(p_2(x), Y)$, as long as $S_{PS}(p_1(x), Y) > 0$, as one can always find σ_2 large enough so that $(\alpha-1) \int_{-\infty}^{\infty} p_2^{\alpha}(z) dz$ is smaller than $S_{PS}(p_1(x), Y)$. To have $S_{PS}(p_1(x), Y) > 0$:

$$\begin{aligned}
p_1(Y) & < (\alpha-1)^{\frac{1}{\alpha-1}} \alpha^{-\frac{3}{2(\alpha-1)}} \frac{1}{\sqrt{2\pi}\sigma_1} \\
p_1(Y) & < (\alpha-1)^{\frac{1}{\alpha-1}} \alpha^{-\frac{3}{2(\alpha-1)}} p_1(u_1) \quad (16)
\end{aligned}$$

This condition also defines a vulnerable subspace where the evaluation using Power Scores might be misinformative. Figure 3 gives an example where PLS may produce “unfortunate” results. The blue line and red line represent the PDFs of two forecast systems A and B. Intuitively one would expect that if the outcome is less than -4 (or between -2 and -1), system A shall be preferred as system A would assign significantly more probability mass around the outcome than

system B. On the contrary, the relative PLS (the green line) prefers system B instead as positive relative PLS would be observed as shown in Figure 3. Ironically, if the forecast systems A and B are delivered to the user, the developer of forecast system B would hope for the outcome being smaller than -4 in order to “outperform” forecast system A by achieving better PLS despite the fact that forecast system B assigns ~ 0 probability to the outcome.

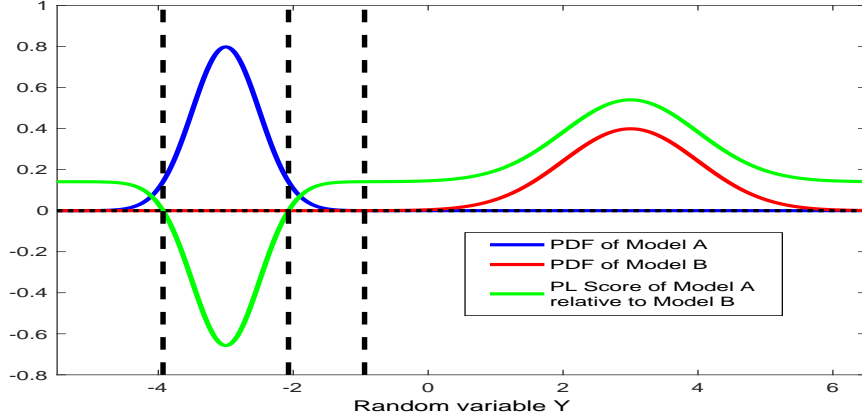


Figure 3: Example showing that the Proper Linear Score produces “unfortunate” result due to the fact it is implausible. The blue line and the red line represent PDFs of forecast system A ($N(-3, 0.5^2)$) and B ($N(3, 1^2)$). The green line represents the PLS of system A relative to system B. A negative relative score suggests system A outperforms system B. The left side of the dashed vertical line at -4 and the region between dashed vertical lines at -2 and -1 define the regions where “unfortunate” results occur.

Similarly to prove the Pseudo-Spherical Score family is implausible, one needs to find $p_1(\cdot)$, $p_2(\cdot)$ and Y so that $p_1(Y) = rp_2(Y)$ but $S_{PSS}(p_1(x), Y) > S_{PSS}(p_2(x), Y)$, which requires:

$$\begin{aligned}
& -\frac{p_1(Y)^{\beta-1}}{(\int_{-\infty}^{\infty} p_1^{\beta}(z)dz)^{1/\beta}} > -\frac{p_2(Y)^{\beta-1}}{(\int_{-\infty}^{\infty} p_2^{\beta}(z)dz)^{1/\beta}} \\
& -\frac{p_1(Y)^{\beta-1}}{(2\pi)^{\frac{1-\beta}{2\beta}} \beta^{-\frac{1}{2\beta}} \sigma_1^{\frac{1-\beta}{\beta}}} > -\frac{p_2(Y)^{\beta-1}}{(2\pi)^{\frac{1-\beta}{2\beta}} \beta^{-\frac{1}{2\beta}} \sigma_2^{\frac{1-\beta}{\beta}}} \\
& -\frac{(rp_2(Y))^{\beta-1}}{\sigma_1^{\frac{1-\beta}{\beta}}} > -\frac{p_2(Y)^{\beta-1}}{\sigma_2^{\frac{1-\beta}{\beta}}} \tag{17} \\
& \sigma_2 > r^{\beta} \sigma_1
\end{aligned}$$

Note the condition in Eq. 17 also places a restriction onto Y , as σ_2 gets larger, the maximum

value of $p_2(x)$ can be smaller than $\frac{p_1(Y)}{r}$. Therefore Y has to be chosen so that $\frac{p_1(Y)}{r} \leq p_2(u_2)$, i.e. $\frac{p_1(Y)}{r} \leq \frac{1}{\sqrt{2\pi}\sigma_2}$ and as $\sigma_2 > r^\beta \sigma_1$, it requires $p_1(Y) < \frac{1}{\sqrt{2\pi}\sigma_1} r^{1-\beta}$. This condition also defines a vulnerable subspace (given $r > 1$) where the evaluation using the Pseudo-Spherical Score might be misinformative.

Figure 4 gives an example where SPS may produce “unfortunate” results. Consider two forecast systems based on Gaussian distributions, where the PDF of system A (blue line) is standard Gaussian and system B (red line) being $N(0, 5^2)$. Intuitively, one would expect that if the outcome lands in the two regions bounded by the black dashed vertical lines, system A shall be preferred as system A would assign significantly more probability mass around the outcome than system B. On the contrary, the relative SPS (the green line) prefers system B instead as positive relative SPS would be observed in both regions.

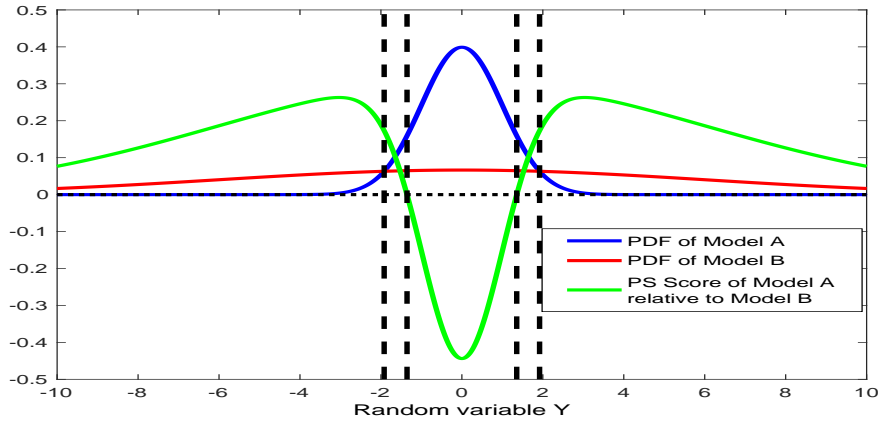


Figure 4: Example showing that the Spherical Score produces “unfortunate” result due to the fact it is implausible. The blue line and the red line represent PDFs of forecast system A (standard Gaussian) and B ($N(0, 5^2)$). The green line represents the SPS of system A relative to system B. A negative relative score suggests system A outperforms system B. The dashed vertical lines enclose the regions where “unfortunate” results occur.

7 Score Interpretation

The difference between two forecast systems is reflected by the difference between their scores. This provides a rank ordering, and thus a preference. Without any reference, a single score of a forecast system hardly provides any evaluation information, which is why score interpretation

should be considered base on the relative score between forecast systems. It is also helpful if the relative score has some meaningful interpretation that relates to the benefit of the users. Otherwise it only indicates which forecast system is better, without answering the question of how much better one is in a meaningful way that adds value to decision support. For example, Figure 1 does show the IGN, PLS and CRPS' preferences between the two forecast systems; however, the interpretation of the relative score (the y-axis in Figure 1) is also important to decision-makers.

A number of meaningful interpretations to proper scoring rules have been identified in the literature. IGN can be interpreted in terms of gambling returns [16, 18, 21, 36]. Under a Kelly betting scenario¹⁰, IGN describes the rate at which the forecaster's fortune increases with time. A house setting fair odds [13] based on a forecast system with a lower value of a nonlocal scoring rule is expected to lose money to a gambler who places bets based on a different forecast system with a lower IGN. Through its close relation to Shannon's information entropy, IGN is related to the amount of information expected from a forecast [36]. IGN can also be easily communicated as an effective interest rate [18]. Jose and Winkler (2008) show that the Pseudo-spherical score and power score families can be interpreted as profits in certain decision problems. Note all the interpretations listed above are based on some specific scenario in which one can, in fact, define a corresponding utility function to replace the scoring rule. For example in a Kelly betting contest, one can define a utility function that reflects the rate (at which the forecasters' fortune increases with time) and use such utility function to replace IGN for forecast evaluation. In practice, it is usually not easy to define a relevant utility function based on probabilistic forecasts for the use of decision support. It is therefore desirable for a scoring rule to have a rather direct and generic interpretation.

The expected IGN can be written as:

$$E(S_{IGN}(p(x), Y)) = \int [-\log_2 p(Y)] Q(Y) dY \quad (18)$$

And the expected relative IGN between two probabilistic forecast system p_1 and p_2 is:

$$\int [-\log_2 \frac{p_1(Y)}{p_2(Y)}] Q(Y) dY \quad (19)$$

Therefore the empirical relative IGN score, $\frac{1}{N} \sum -\log_2 \frac{p_1(Y)}{p_2(Y)}$, reflects the (average) increase in

¹⁰In a Kelly betting contest [21], one bets all of one's wealth on every outcome in proportion to the forecast probability of that outcome. More precisely, a fraction ω_i of ones wealth, where ω_i is the forecast probability of event E_i occurring, should be wagered on the i^{th} outcome.

probability mass that the model forecast p_1 placed on the outcome relative to that of the reference forecast p_2 . Note that although p_1 and p_2 are probability density functions, $-\log_2 \frac{p_1(Y)}{p_2(Y)}$ can be interpreted as increase/decrease in probability¹¹, which gives the Ignorance score a meaningful direct interpretation¹². The relative IGN of two forecast systems also quantifies the information gain (in terms of bits) the model forecast system provides over the reference system. A relative IGN of 1 bit means that, on average, forecasts from the system assign twice the probability to the outcome compared to the reference forecast [36].

Nonlocal scoring rules include contributions from the entire PDF; the scoring rule may be largely determined by outcomes that did NOT occur, making a meaningful direct interpretation somewhat challenging. For example, the empirical relative PLS between two forecast system p_1 and p_2 based on a large number N of forecast-outcome pairs is:

$$\left[\int p_1^2(z)dz - \int p_2^2(z)dz \right] + \frac{2}{N} \sum_i^N [p_2(Y_i) - p_1(Y_i)] \quad (20)$$

The interpretation of Eq. 20 is clearly more sophisticated than that of the relative IGN. In the second term of Eq. 20, $p_2(Y) - p_1(Y)$, which ranges $(-\infty, \infty)$, is the difference between two probability density functions rather than two probabilities. In the context of decision support, it is unclear how to interpret the probability density function(s) meaningfully other than by using $\log p_2(Y) - \log p_1(Y)$ to reflect the increase/decrease in probability mass placed on Y (this is in fact the approach used by relative IGN). The first term of Eq. 20 being a function of the entire PDF of forecast systems (not depending on the outcome Y) clearly makes it even more challenge for interpretation. Similar interpretation challenges applies to the CRPS and SPS. There are better ways to interpret these nonlocal scoring rules by using True underlying distribution as a reference.

For example, assuming the True underlying distribution Q exists then the expectation of PLS is:

$$E(S_{PLS}(p(x), Y)) = \int [-2p(Y) + \int p^2(z)dz] Q(Y) dY \quad (21)$$

PLS is based on the idea that the scoring rule should reflect “nearness” of the predicted probability distribution to the True underlying distribution. By straightforward manipulation, it comes to

¹¹For small values of δ , one can write $P(x < X \leq x + \delta) \approx p(X)\delta$.

¹²Note the interpretation of the empirical relative Ignorance score does NOT require the knowledge of the True underlying distribution Q .

the following representation:

$$E(S_{PLS}(p(x), Y)) = \int [Q(Y) - p(Y)]^2 dY - \int Q^2(Y) dY \quad (22)$$

The second term in the RHS of Eq. 21 will vanish when comparing two forecast systems using the expected relative score, which gives the expected relative PLS between two probabilistic forecast system p_1 and p_2 :

$$\int [Q(Y) - p_1(Y)]^2 dY - \int [Q(Y) - p_2(Y)]^2 dY \quad (23)$$

Therefore the expected relative PLS between two forecast systems can be interpreted with regard to the mean square difference between the forecast distribution and the True underlying distribution Q .

Similarly, the expectation of CRPS can be written as:

$$E(S_{CRPS}(p(x), Y)) = \int [G(Y) - F(Y)]^2 dY - \int G(Y)(1 - G(Y)) dY, \quad (24)$$

where $F(\cdot)$ is the CDF of the forecast distribution and $G(\cdot)$ is the True underlying CDF. The expectation of the relative CRPS between two forecast systems can be interpreted with regard to the mean square difference between the forecast CDF and the CDF of the Truth.

The expected SPS can be written as:

$$E(S_{SPS}(p(x), Y)) = \left(\int Q(Y)^2 dY \right)^{1/2} \frac{\int p(Y)Q(Y) dY}{\left(\int Q(Y)^2 dY \right)^{1/2} \left(\int p(Y)^2 dY \right)^{1/2}} \quad (25)$$

It can be interpreted regarding the interior angle of deviation between the forecast distribution p and the True underlying distribution Q .

In some cases it makes sense to consider an integration over the True underlying distribution Q . The interpretation of the expected relative score with respect to Q is cloudy in reality, for example in weather-like forecasting scenarios, where the same Q distribution is never seen twice over the lifetime of the system. In practice, the True underlying distribution is rarely (if ever) available to provide such an interpretation, and were it to be the use of imperfect probabilistic forecast is mute. Furthermore, functions of the entire forecast distribution, as in Eq 22. 24. & 25. can hardly be interpreted in a meaningful way for decision support. Therefore, even if the True underlying distribution were available, it is unclear the interpretations of relative scores derived from Eq 23-25 are informative to the decision maker except providing their preference between two forecast systems.

8 Scoring Rules Under Transformation

In practice, it is common that the variable of interest is not the variable observed but a function of the observed variable. For example, wind power is a function of wind speed cubed; wave power is principally a function of wave height squared [37]. It is desirable for a scoring rule to provide coherent evaluations before and after a smooth transformation being applied to the forecast variable. Consider $x^* = \phi(x)$ as a smooth one-to-one (transformation) function of a random variable x . The forecast PDF of x , $p(x)$, becomes $p(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}$ for the random variable x^* after the transformation and the scoring rule $S(p(x), Y)$ becomes $S(p(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}, Y^*)$, where $Y^* = \phi(Y)$. It is almost certain that the value of a scoring rule will change after the transformation. Note score interpretation should always be based on the relative score between forecast systems instead of a single score of a forecast system in order to provide useful information for decision support. It is therefore of interest to investigate whether the relative score will change after taking the transformation and if so, will the scoring rule's preference change as well. Given the relative score between two probabilistic forecast system p_1 and p_2 by:

$$S(p_1(x), Y) - S(p_2(x), Y), \quad (26)$$

after taking a transformation $\phi(x)$ it becomes

$$S(p_1(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}, Y^*) - S(p_2(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}, Y^*). \quad (27)$$

Note that Y and Y^* are one-to-one and $p(x)$ and $p(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}$ reflect the same information of a forecast system. Therefore if (26) does not equal to (27) based on some scoring rule S , the scoring rule will have a non-unique interpretation of the relative skill between two competing forecast systems. Furthermore if $(26) \times (27) < 0$ for some Y and ϕ , it indicates such scoring rule might also change its preference due to the transformation, then the use of such scoring rule as an evaluation tool for decision support is questionable.

Ignorance score is *invariant* under smooth transformation as (26) and (27) are equal for any smooth transformation, proved in the following:

$$\begin{aligned} & S(p_1(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}, Y^*) - S(p_2(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*}, Y^*) \\ &= -\log p_1(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*} \Big|_{Y^*} + \log p_2(\phi^{-1}(x^*))\frac{d\phi^{-1}(x^*)}{dx^*} \Big|_{Y^*} \\ &= -\log p_1(Y) + \log p_2(Y) = S(p_1(x), Y) - S(p_2(x), Y). \end{aligned} \quad (28)$$

For nonlocal scoring rules like Proper Linear Score, Spheric Score and Continuous Rank Probability score, smooth transformation not only have impact on the value of the relative scores but also may cause the change of their preference. Figure 5 gives an example where the CRPS may contradict itself by changing its preference under transformation (similar examples can be found for PLS and SPS). Following Figure 2, Figure 5(a) compares two forecast systems based on a Bimodal distribution with the same shape but different centers. The green line represents the CRPS of system A relative to system B, a negative relative score suggests system A outperforms system B according to the CRPS. The black dashed vertical line in Figure 5(a) corresponds to the threshold $Y = 11.5$, where the CRPS prefers forecast system A when $Y < 11.5$ and prefers forecast system B when $Y > 11.5$. Figure 5(b) compares the same two forecast systems after cubic transformation being applied to the forecast variable. Clearly the relative CRPS has changed after the cubic transformation. Let x refers to wind speed, then x^3 reflects wind power. When the observed wind speed is 10, the relative CRPS (forecast system A relative to forecast system B) is roughly -0.9 as in Figure 5(a). Comparing the same¹³ forecast system A and B in terms of wind power under the same observation (wind speed 10 corresponds to wind power 1000), however, the relative CRPS¹⁴ as in Figure 5(b) becomes roughly 340, which indicates the interpretation of CRPS evaluation for comparing forecast system A and B based on a unique observed wind speed is not unique. Furthermore, the CRPS may even change its preference after the transformation as the threshold (the black solid vertical line in Figure 5(b)) that distinguish the CRPS preference is $Y^3 = 1700$ rather than $Y^3 = 11.5^3 = 1520.875$ (dashed vertical line). If the underlying distribution of the wind speed were bounded between the black dashed line and black solid line, the CRPS would prefer forecast system A before the cubic transformation when the wind speed is evaluated directly, while it prefer forecast system B after the cubic transformation when the wind power (which corresponds to the same wind speed) is evaluated.

¹³In fact the information presented by forecast systems A and B remain the same although their pdf changed when the forecast variable wind speed transformed to wind power.

¹⁴The relative CRPS (the green curve) in Figure 5(b) is scaled down by 1.6×10^5 in order to have similar magnitude as the PDFs.

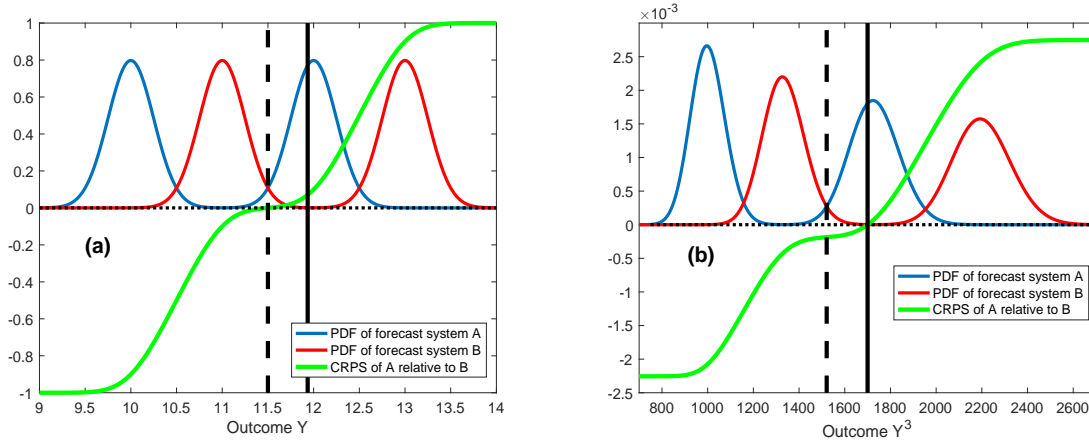


Figure 5: Example showing that the Continuous Rank Probability Score changes its preference under transformation. The blue line and the red line represent PDFs of forecast systems A and B. The green line represents the CRPS of system A relative to system B. A negative relative score suggests system A outperforms system B. (a) before the cubic transformation; (b) after the cubic transformation. The black dashed vertical line and solid line in (a), where $Y = 11.5$ and $Y = 11.935$ respectively, corresponds to those in (b), where $Y = 11.5^3$ and $Y = 11.935^3$.

9 Discussion and Summary

Measures of skill play a critical role in the development, deployment and application of probabilistic forecasts. The property of some common strictly proper scoring rules have been discussed. Given a strictly proper scoring rule, the True forecast system will always be preferred whenever it is included amongst those under consideration. In practice, to correctly measure the difference between imperfect forecast schemes, being strictly proper is not enough, as strictly proper scoring rules need not rank competing forecast systems in the same order when none of the forecast systems are perfect. In general, any scoring rules can be presented with the form:

$$S(p(x), Y) = s_1(p(x)) + s_2(p(x), Y) + s_3(p(Y)). \quad (29)$$

For local scoring rules, the first two terms in the RHS of Eq. 28 are both zero with only the presence of $s_3(p(Y))$, for example the only local proper scoring rule, the logarithmic score (Ignorance). Nonlocal scoring rules contain at least one of the first two terms, for example the Energy Scores consist of s_1 and s_2 , the Power Scores s_1 and s_3 and the Pseudo-sphere Scores only s_2 . The presence of s_1 or s_2 or both allows the scoring rule to give extra credit to the structure of the

forecast PDF. Note such extra credit is not necessarily given for assigning high probabilities to the values near the outcome as the examples in Figures 2-4 show. Without knowing the True underlying distribution, the justification of giving such extra credit is untenable. Nonlocal strictly proper scoring rules considered¹⁵ are shown to have property that can produce “unfortunate” evaluations due to the fact that contributions from the entire shape of the PDF may overwhelm that from the probability assigned to the outcome. Particularly the fact that Continuous Rank Probability Score prefers the outcome close to the median of the forecast distribution regardless the probability mass assigned to the value at/near the median raises concern to the use of Continuous Rank Probability Score. Ignorance has direct interpretations in terms of probabilities and bits of information while the direct interpretation of nonlocal strictly proper scoring rules on the other hand relies on information regarding the unknown (if it even exists) True underlying distribution as a reference. The nonlocal strictly proper scoring rules considered may also contradict themselves when a smooth transformation is applied to the forecast variable while IGN is shown to be invariant under smooth transformation. It is suggested that Ignorance should always be included in the evaluation of probabilistic forecasts.

One of the reasons for using nonlocal scoring rules is to address particular problems where a local scoring rule is not considered “suitable”. For example, Ignorance is infinity if the forecast assigns vanishing probability to an event that obtains. [39] emphasizes that the use of Ignorance implies the value judgment that small differences between small probabilities should be taken very seriously and that wrongly describing something extremely improbable as having zero probability is “an unforgivable sin”. [36] pointed out that forecasters should replace zero forecast probabilities with small probabilities based on the uncertainties in the forecast PDF. Not to do so means reporting the improbable as the impossible. Within the Bayesian framework, Cromwell’s rule states that the use of prior probabilities of 0 or 1 should be avoided. Assigning zero probability to events that are possible also contradicts to Laplace’s rule of succession [19]. In the insurance sector, the premium is inversely proportional to the probability of an event occurring; zero probability would suggest free insurance.

In this manuscript, the value outcome is assumed to be certain. In the presence of uncertainty

¹⁵The author does believe that all non-local scoring rules are implausible, but as there is no general mathematical function of non-local scoring rules, only the popular nonlocal scoring rules (and their families) are considered and shown to be implausible in this manuscript.

in the value of the outcome (for example due to measurement error), one may obtain benefit by assigning probability to the events that do not match the outcome exactly. Note again this does not imply one should use nonlocal scoring rules, as for nonlocal scoring rules the contributions from the entire shape of the PDF are not designed to account for the uncertainty in the value of the outcome. For a local scoring rule, the evaluation can still be considered over the observational uncertainty distribution of the outcome, for example by coupling the forecast distribution with the distribution of observational noise [6].

Scoring rules are designed to assess (probabilistic) forecast performance, which hopefully leads to better decision making. [3] argue that a local proper score should be preferred for ‘pure inference’ problems in which the outcome is the sole arbiter of forecast quality, yet there are other forms of scoring rules that would typically be appropriate in more directly practical contexts (see stock control example in [3]). Note that in such ‘more directly practical contexts’, if a utility function based on probabilistic forecasts can be conveniently defined according to the practical objective (which often is not the case in practice), there is no need for any kind of scoring rules (using the utility function directly will serve the purpose of forecast evaluation sufficiently). Any scoring rules can be directly considered as a utility function, yet the meaning of the corresponding utility function relies on the direct interpretation of the skill of the scoring rule. It is questionable whether nonlocal scoring rules can provide any meaningful direct interpretation. [40] claims that interpretation is a critical aspect in accepting a scoring rule for use in practice; he uses valued property of probabilistic forecasts to support this assertion.

Acknowledgment

This research was supported the EPSRC-funded Uncertainty analysis of hierarchical energy systems models: Models versus real energy systems (EP/K03832X/1) and Centre for Energy Systems Integration (EP/P001173/1). Additional support was also provided by Evaluating Probability Scores for the Insurance Sector funded by LSE KEI and Lighthill Risk Network. The author would like to thank Leonard A. Smith and Edward Wheatcroft for reading earlier versions of this article and giving useful feedback; and two anonymous reviewers whose comments and suggestions helped improve and clarify this article.

References

- [1] Baringhaus, L., and C. Franz, 2004: On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88 (1), 190-206.
- [2] J. M. Bernardo. Expected information as expected utility. *Ann. Stat.*, 7:686-690, 1979.
- [3] Bernardo, J. M., and A. F. M. Smith, *Bayesian Theory*. Wiley, 2000.
- [4] Boero, G., J. Smith, and K. F. Wallis, 2011: Scoring rules and survey density forecasts. *International Journal of Forecasting*, 27, 379-393.
- [5] Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, 78, 502 1-3.
- [6] J. Brocker and L.A. Smith. Scoring probabilistic forecasts: On the importance of being proper. *Wea. Forecasting*, 22:382-388, 2007.
- [7] Brown, T. A., 1970: Probabilistic forecasts and reproducing scoring systems. Technical Report RM-6299-ARPA, RAND Corporation.
- [8] Brown, T. A., 1974: Admissible scoring systems for continuous distributions. Manuscript P-5235, The Rand Corporation.
- [9] Du, H., and L. A. Smith, 2012: Parameter estimation using ignorance. *Physical Review E*, 86, 016213.
- [10] E. H. Shuford, H. E. M., A. Albert, 1966: Admissible probability measurement procedures. *Psychometrika*, 31.
- [11] Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8, 985-987.
- [12] Fricker, T. E., C. A. T. Ferro, and D. B. Stephenson, 2013: Three recommendations for evaluating climate prediction. *Meteorological Applications*, 20, 246-255.
- [13] Frigg, R., S. Bradley, H. Du, and L. Smith, 2014: Laplace’s demon and the adventures of his apprentices. *Philosophy of Science*, 81, 31-59.

- [14] Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 477:359-378.
- [15] Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, 40, 1-2: 245-272.
- [16] I. J. Good. Rational decisions. *Journal of the Royal Statistical Society*, XIV(1), 1952.
- [17] Good, I. J., 1971: Comment on 'Measuring information and uncertainty' by Robert J. Buehler, 337-339. Holt, Rinehart and Winston, Toronto.
- [18] R. Hagedorn and L. A. Smith. Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.*, 16:143155, 2009.
- [19] Jaynes, E. T., 2003: *Probability Theory: The Logic of Science*. Cambridge University Press.
- [20] Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley.
- [21] Kelly, J. L., 1956: A new interpretation of information rate. *Bell System Technical Journal*, 35, 917-926.
- [22] Kohonen, J., and J. Suomela, 2006: Lessons learned in the challenge: Making predictions and scoring them. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, J. Quinero-Candela, I. Dagan, B. Magnini, and F. d'AlcheBuc, Eds., Springer Berlin Heidelberg, Berlin, Heidelberg, 95-116.
- [23] Kullback, S., 1959: *Information Theory and Statistics*. Wiley.
- [24] Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *Ann.Math. Stat.*, 22, 79-86.
- [25] Machete, R., 2013: Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, 143.
- [26] Machete, R., and L. Smith, 2016: Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A*, 68, 28 393.

- [27] Mason, S. J., and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Monthly Weather Review*, 137, 331-349.
- [28] Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management Science*, 22, 1087-1096.
- [29] Maynard, T., 2016: Extreme insurance and the dynamics of risk. Ph.D. thesis, The London School of Economics and Political Science, London, UK.
- [30] McSharry, P., and L. A. Smith, 1999: Better nonlinear models from noisy data: attractors with maximum likelihood. *Physical Review Letters*, 83.
- [31] Murphy, A. H., 1969: On the 'Ranked Probability Score'. *Journal of Applied Meteorology*, 8 (6), 988-989.
- [32] Murphy, A. H., 1970: The Ranked Probability Score and The Probability Score: A Comparison. *Monthly Weather Review*, 98 (12), 917-924.
- [33] Murphy, A. H., 1996: The finley affair: A signal event in the history of forecast verification. *Weather and Forecasting*, 11 (1), 3-20.
- [34] Murphy, A. H., and R. L. Winkler, 1970: Scoring rules in probability assessment and evaluation. *Acta. Psychol.*, 34, 273-286.
- [35] Roby, T. B., 1965: Belief states: A preliminary empirical study. *Behavioral Science*, 10.
- [36] M. S. Roulston and L. A. Smith. Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, 130:1653-1660, 2002.
- [37] Savenkov, M., 2009: On the truncated weibull distribution and its usefulness in evaluating the theoretical capacity factor of potential wind (or wave) energy sites. *University Journal of Engineering and Technology*, 1, 21-25.
- [38] Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140, 1086-1096.
- [39] Selten, R., 1998: Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1.

- [40] Smith, L., 2020: Necessary conditions for scoring probability forecast system. in preparation.
- [41] Smith, L. A., E. B. Suckling, E. L. Thompson, T. Maynard, and H. Du, 2015: Towards improving the framework for probabilistic forecast evaluation. *Climatic Change*, 132 (1), 31-45.
- [42] Stael von Holstein, C.-A. S., 1970a: A family of strictly proper scoring rules which are sensitive to distance. *Journal of Applied Meteorology*, 9 (3), 360-364.
- [43] Stael von Holstein, C.-A. S., 1970b: Measurement of subjective probability. *Acta Psychologica*, 34, 146-159.
- [44] Stephenson, D. B., 2000: Use of the 'odds ratio' for diagnosing forecast skill. *Weather and Forecasting*, 15 (2), 221-232.
- [45] Szekely, G. J., 2003: E-statistics: The energy of statistical samples. Technical Report 2003-16, Bowling Green State University.
- [46] Szekely, G. J., and M. L. Rizzo, 2005: A new test for multivariate normality. *Journal of Multivariate Analysis*, 93 (1), 58-80.
- [47] Toda, M., 1963: Measurement of subjective probability distributions. esd-tdr-63-407. Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command.
- [48] Todter, J., and B. Ahrens, 2012: Generalization of the Ignorance Score: Continuous Ranked Version and Its Decomposition. *Monthly Weather Review*, 140 (6), 2005-2017.
- [49] Unger, D. A., 1995: A method to estimate the continuous ranked probability score. Proceedings of the ninth conference on probability and statistics, American Meteorological Society, Boston, USA, 206-213.
- [50] V. R. R. Jose, R. F. N., and R. L. Winkler, 2008: Scoring rules, generalized entropy, and utility maximization. *Operations Research*, 56.
- [51] Wilks, D. S., 1995: Statistical Methods in the Atmospheric Sciences: An Introduction. International Geophysics Series. Academic Press.
- [52] Winkler, R. L., 1969: Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.*, 64.

- [53] Winkler, R. L., 1996: Scoring rules and the evaluation of probabilities. *Test*, 5.
- [54] Winkler, R. L., and A. H. Murphy, 1968: 'good' probability assessors. *Journal of Applied Meteorology*, 7 (5), 751-758.
- [55] Zhang, Y., J. Wang, and X. Wang, 2014: Review on probabilistic forecasting of wind power generation. *Renewable and Sustainable Energy Reviews*, 32, 255-270.