

1 Towards improving the framework for probabilistic  
2 forecast evaluation\*

3 L. A. Smith<sup>1</sup> E. B. Suckling<sup>1</sup> E. L. Thompson<sup>1</sup> T. Maynard<sup>1</sup> H. Du<sup>2</sup>

<sup>1</sup>Centre for the Analysis of Time Series,  
London School of Economics, London WC2A 2AE. UK

<sup>2</sup>Center for Robust Decision Making on Climate and Energy Policy,  
University of Chicago, Chicago 60637. US

4 April 16, 2015

5 **Abstract**

6 The evaluation of forecast performance plays a central role both  
7 in the interpretation and use the forecast system and in their devel-  
8 opment. Different evaluation measures (scores) are available, often  
9 quantifying different characteristics of forecast performance. The prop-  
10 erties of several proper scores for probabilistic forecast evaluation are  
11 contrasted and then used to interpret decadal probability hindcasts of  
12 global mean temperature. The Continuous Ranked Probability Score  
13 (CRPS), Proper Linear (PL) score, and IJ Good's logarithmic score  
14 (also referred to as Ignorance) are compared; although information  
15 from all three may be useful, the logarithmic score has an immediate  
16 interpretation and is not insensitive to forecast busts. Neither CRPS  
17 nor PL is local; this is shown to produce counter intuitive evaluations  
18 by CRPS. Benchmark forecasts from empirical models like Dynamic  
19 Climatology place the scores in context. Comparing scores for fore-  
20 cast systems based on physical models (in this case HadCM3, from  
21 the CMIP5 decadal archive) against such benchmarks is more infor-  
22 mative than internal comparison systems based on similar physical  
23 simulation models with each other. It is shown that a forecast sys-  
24 tem based on HadCM3 out performs Dynamic Climatology in decadal  
25 global mean temperature hindcasts; Dynamic Climatology previously

---

\*Part of the EQUIP special issue of Climatic Change

outperformed a forecast system based upon HadGEM2 and reasons for these results are suggested. Forecasts of aggregate data (5-year means of global mean temperature) are, of course, narrower than forecasts of annual averages due to the suppression of variance; while the average “distance” between the forecasts and a target may be expected to decrease, little if any discernible improvement in probabilistic skill is achieved.

## 1 Introduction

Decision making would profit from reliable, high fidelity probability forecasts for climate variables on decadal to centennial timescales. Many forecast systems are available, but evaluations of their performance are not standardised, with many different scores being used to measure different aspects of performance. These are often not directly comparable across models or across different studies. EQUIP (the ‘End-to-end Quantification of Uncertainty for Impacts Prediction’ consortium project) aimed to provide guidance to users of information at the space and time scales of interest, and to develop approaches to enable evidence-based choice between alternate forecasting methods, based on reliable and informative measures of forecast skill. The intercomparison of simulation models is valuable in many ways; comparison of forecasts from simulation models with empirically-based reference forecasts provides additional information. In particular it aids in distinguishing the case when each forecast system does well; and so the best system cannot be identified (i.e. equifinality) from the case in which each forecast system performs very poorly (i.e. equidismality) [1, 35]. Indeed some climate researchers have required the demonstration of skill against a more easily prepared reference forecast as a condition for accepting any complicated forecasting scheme as useful [34]. This raises the question of how exactly to quantify skill.

Three measures of forecast system performance (hereafter, scores) are studied below and the desirability of their attributes is considered. It is critical to keep in mind that an entire forecast system is evaluated, not merely the model at its core. Each score in turn is then illustrated in the context of decadal forecasts of global mean temperature. Section 2 discusses several measures of forecast system performance, including the logarithmic score (Ignorance) [16, 27], the Continuous Ranked Probability Score (CRPS) [10, 14] and the Proper Linear score (PL) [13]. General considerations for selecting a preferred score are discussed; CRPS is demonstrated capable of misleading behavior. Section 3 then introduces the forecast targets and

64 forecast systems to be considered in this paper. Both empirical and simula-  
65 tion models are identified and the primary target, global mean temperature  
66 (GMT), is discussed. Section 4 considers the performance of probability  
67 forecasts (both empirical and simulation-based) on decadal scales in the  
68 light of each of these scores.

## 69 2 Measuring forecast performance

70 Several scores are available for the evaluation of probabilistic forecasts [4,  
71 14, 23, 21]; each quantifies different attributes of the forecast. While the  
72 importance of using *proper scores* is well recognised [4, 12], researchers of-  
73 ten face requests to present results under a variety of scores. Indeed in the  
74 context of meteorological forecast evaluation there are several recommenda-  
75 tions in the literature [24, 26, 39, 12, 15], although often with little discussion  
76 of which attributes different scores aim to quantify, or their strengths and  
77 weaknesses in a particular forecast setting. By convention, a lower score is  
78 taken to reflect a better forecast.

79 A score is a functional of both the forecast (whose pdfs are denoted by  
80 either  $p$  or  $q$ ) and the observed outcome ( $X$ ). It is useful to speak of the  
81 “True” distribution from which the outcome is drawn (hereafter,  $Q$ ) without  
82 assuming that such a distribution exists in all cases of interest. Given a  
83 proper score, a forecast system providing  $Q$  will be preferred whenever it  
84 is included amongst those under consideration.[4, 12] When this is not the  
85 case, then even proper scores may rank two forecast systems differently,  
86 making it difficult to provide definitive statements about forecast quality.  
87 There are, however, desirable properties of the scores themselves that may  
88 help to narrow down the set of scores appropriate for a given task.

89 A score,  $S(p(x), X)$ , is said to be ‘proper’ if inequality (1) holds for any  
90 pair of forecast pdfs, and ‘strictly proper’ when equality implies  $p = q$ :

$$\int q(z)S(p(z), z)dz \geq \int q(z)S(q(z), z)dz. \quad (1)$$

91 For a given forecast  $p$ , a score is itself a random variable with values that  
92 depend on the observed outcome  $X$ . One can calculate the expected score  
93 of the forecast  $p$  when  $X$  is actually drawn from underlying distribution  $q$ .  
94 A proper score does not, in expectation, judge any other forecast  $p$  to score  
95 better than  $q$  as a forecast of  $q$  itself. The interpretation of proper does  
96 not, however, require one to believe that a “True” distribution  $Q$  exists.  
97 While use of a proper score might be motivated by concerns of hedging [28],

proper scores are preferred even when there is no human in the loop, as in parameter selection [9]. For completeness, and without endorsement, the discussion below is not restricted to proper scores.

## 2.1 RMSE of the ensemble mean

The Root Mean Squared Error (RMSE) quantifies the distance between the ensemble mean,  $\bar{x}(i)$  of the  $i^{th}$  forecast and the corresponding outcome,  $X(i)$ , defined as,

$$RMSE(\bar{x}, X) = \sqrt{\frac{1}{m} \sum_{i=1}^m (\bar{x}(i) - X(i))^2}, \quad (2)$$

Note that rather than provide a score for a single forecast RMSE summarizes  $m$  forecasts. Any of the wide variety of forecast distributions with the same mean will achieve the same score. An alternative summary score resembling the RMSE can be defined via

$$S_{RMSE}((p_1, \dots, p_m), (X_1, \dots, X_m)) = \sqrt{\frac{1}{m} \sum_{i=1}^m \left( \int_{-\infty}^{\infty} (X_i - z)^2 p(i, z) dz \right)}. \quad (3)$$

The original RMSE re-emerges by setting the forecast  $p$  as a delta function at the ensemble mean. The integral term is sometimes referred to as the Mean Squared Error (MSE). This score is not proper, and the lowest score is attained when the standard deviation of the forecast is zero - an unfortunate incentive for an imperfect probabilistic forecast.

## 2.2 Naive Linear and Proper Linear scores

The Naive Linear (NL) score is not proper. It is defined by:

$$S_{NL}(p(x), X) = -p(X). \quad (4)$$

The NL score can be “made” strictly proper by the addition of an integral term over  $p$  to equation 4,

$$S_{PL}(p(x), X) = -2p(X) + \int_{-\infty}^{\infty} p^2(z) dz, \quad (5)$$

resulting in the Proper Linear (PL) score [13]. The PL score is related to the quadratic score, which is part of the power rule family that contains an

120 infinite number of proper scores [28]. The popular Brier [3] and Continuous  
 121 Ranked Probability scores [10] are also special cases of the quadratic scoring  
 122 rule family [33]. The PL score itself rewards a forecast both for the proba-  
 123 bility placed on the outcome (the first term in equation 5) and for the shape  
 124 of the distribution (the second term in equation 5). Narrower distributions  
 125 are penalised regardless of the outcome. Arguably the second term clouds  
 126 the interpretation of the score, unless one has some particular incentive to  
 127 minimize this integral. This illustrates a case where an intuitive score, the  
 128 probability of the outcome, can be made to be proper at the cost of some  
 129 immediate intuitive appeal. Alternatively, in cases where it is meaningful to  
 130 speak of the distribution from which the outcome is drawn (referred to as  
 131  $Q$  above), then PL is simply related to the integral of the squared difference  
 132 between the forecast  $p(x)$  and  $Q(x)$ . This point is revisited in Section 4.

### 133 2.3 Continuous Ranked Probability

134 The Continuous Ranked Probability Score (CRPS) is the integral of the  
 135 square of the  $L^2$  distance between the cumulative distribution function of  
 136 the forecast  $p$  and a step function at the outcome [10],

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

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

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

138 CRPS can be interpreted as the integral of the Brier score over all threshold  
 139 values; for point forecasts CRPS reduces to the mean absolute error. The  
 140 CRPS rewards a forecast for both its calibration and shape, but unlike the  
 141 PL score they are assessed simultaneously. A decomposition into reliability  
 142 and resolution components is possible [19, 6]. The CRPS is sometimes said  
 143 to assign a value to a raw ensemble of point forecasts [14, 11, 7]<sup>1</sup>; this claim  
 144 is equivalent to interpreting the ensemble members as probability forecast  
 145 consisting of a collection of delta functions. Given that ensemble interpre-  
 146 tation, *any* probability scoring rule can be applied, of course. CRPS is

---

<sup>1</sup>We note there are concerns regarding statistical consistency under this interpreta-  
 tion [7]

147 somewhat more tolerant of weaknesses of this delta function ensemble inter-  
 148 pretation than the other scores discussed here.<sup>2</sup> The authors are unaware  
 149 of an intuitive interpretation of the quantitative values of CRPS.

## 150 2.4 Ignorance

151 The Ignorance score [16, 27] is a strictly proper score defined as,

$$S(p(x), X) = -\log_2(p(X)), \quad (8)$$

152 where  $p(X)$  is the density assigned to the outcome  $X$ . It is the only proper  
 153 local score, rewarding a forecast solely for the probability density placed  
 154 on the observed outcome, rather than for other features of the forecast dis-  
 155 tribution such as its shape. This makes computing the score significantly  
 156 less computationally expensive. The Ignorance score corresponds to the ex-  
 157 pected wealth doubling (or halving) time of a Kelly investment strategy, and  
 158 can be expressed as an effective interest rate [18]. Kelly’s focus [22] was on  
 159 information theory, specifically on providing a context for the mathematical  
 160 results of Shannon while neither of them could define a “communication  
 161 system” precisely. A gambling analogy was selected because it had the es-  
 162 sential features of a communication system. Ignorance emerges as a natural  
 163 measure of information content of probability forecasts in general.

164 Selten [28] objects to the Ignorance score because it severely penalises  
 165 forecasts that place very low probabilities on the observed outcome, and  
 166 indeed Ignorance gives an infinitely bad score if an outcome occurs that the  
 167 forecaster said was impossible. One of the present authors (TM) works in  
 168 the insurance industry, however, and believes this to be a *desirable* property  
 169 of a score – extreme model failure has been one of the key causes of distress in  
 170 the financial services industry. Acknowledging unlikely possibilities as such  
 171 and thereby avoiding the infinite penalty of having stated they were truly  
 172 impossible might be seen as basic good practice (see, however, discussion by  
 173 Borel (1962) regarding vanishingly small probabilities); adopting a minimum  
 174 forecast probability to account for the imperfection of science is perhaps  
 175 akin to adding a margin for safety in engineering terms. In the next section,  
 176 CPRS is shown to be remarkably insensitive to outcomes in regions that

---

<sup>2</sup>At the request of a reviewer we make this tolerance explicit. For a given forecast  $p(x)$ , PL and IGN will give worse scores to an outcome  $X$  when  $p(X)$  is smaller, while CRPS may award its best possible score to an outcome  $X$  which is deemed impossible by the forecast PDF (that is  $p(X) = 0$ ). Scores which systematically prefer forecasts which place a lower probability on the outcome are called perverse.

177 forecast to have vanishingly small or zero probability. No optimal balance on  
 178 the the appropriate level of sensitivity of scores has been generally agreed.

## 179 2.5 Comparing the behaviour of Ignorance and CRPS

180 The Ignorance and CRPS scores corresponding to a variety of different out-  
 181 comes given two bimodal forecast distributions are shown in Figure 1. Figure  
 182 1a shows distributions with symmetric (thick blue) and asymmetric (thin  
 183 red) shapes. Figure 1c compares the Ignorance ( $y$ ) and CRPS ( $x$ ) scores in  
 184 the case of a symmetric bimodal distribution (the thick blue distribution in  
 185 Figure 1a) as the observed outcome moves across the forecast distribution  
 186 from large negative values of  $x$ , through  $x = 0$ , to large positive values of  $x$ .  
 187 The minimum (best) CRPS score is achieved by an outcome at the median  
 188 of the underlying distribution, that is at  $x = 0$  in the symmetric (thick blue)  
 189 case, and near  $x = 0.7$  in the asymmetric (thin red) case marked as a vertical  
 190 line in Figure 1a and as a black star in Figure 1b. Ignorance is minimised  
 191 when the outcome is at the mode of the forecast distribution (the green star  
 192 in Figure 1b). These two points do not correspond to the same outcome.

193 This example shows that the CRPS score can rate an outcome from a  
 194 structurally flawed forecast system highly even when both (a) the outcome  
 195 is repeatedly observed where the forecast system has assigned a small prob-  
 196 ability and (b) the forecast repeatedly places significant probability mass  
 197 in regions of vanishingly small (or zero) probability of occurring; Ignorance  
 198 would penalise such forecast systems severely. Consider a bimodal fore-  
 199 cast like the thick blue distribution in Figure 1 (for example, strong winds  
 200 forecast from either east or west but the direction is uncertain), and an  
 201 underlying  $Q$  distribution which is unimodal with low variance centred at  
 202 zero. The outcome is almost certainly close to zero, which is in a region  
 203 where the forecast ascribes very low probability density – hence, the Igno-  
 204 rance score will heavily penalise the system producing the bimodal forecast.  
 205 The CRPS however will give the forecast the best possible score when this  
 206 outcome occurs. Figure 1c shows the IGN (thick green) and CRPS (thin  
 207 black) as a function of the outcome corresponding to the asymmetric case in  
 208 Figure 1a above it. IGN( $x$ ) returns large (poor) values for outcome far from  
 209 one or the other mode. CRPS( $x$ ) returns large values for outcome far from  
 210 zero, but for values of  $x$  near zero low (good) scores are returned. Figure  
 211 1d reflects a case similar to 1c, where the width of each mode is halved:  
 212 IGN returns low values on a more narrow range, while CRPS again returns  
 213 a similar (low) score for points in the central low probability region. These  
 214 two scores would give rather different impressions of forecast quality when

215 evaluating this bimodal probability forecast when the outcome was gener-  
 216 ated, say, by a Gaussian distribution, with zero mean. The fact that both  
 217 scores are proper restricts their behaviour to agree when given  $Q$ , but not  
 218 when given an imperfect probability forecast.

219 Return to the symmetric (thick blue PDF) forecast in Figure 1a and  
 220 consider all possible forecasts with this bimodal shape but centered at some  
 221 value of  $x = c$ , where  $c$  need not be zero as it is in Figure 1a. Consider the  
 222 case of an outcome at the origin,  $x = 0$ . Will IGN and CRPS rank members  
 223 of this family of forecasts differently? Yes. IGN (and PL) will favour the  
 224 forecasts that place higher probability on the outcome while CRPS will favor  
 225 forecasts that have low probability on the outcome. In this case, IGN will  
 226 favor (equally) the two forecasts with values of  $c$  such that a mode is at the  
 227 origin, while CRPS will favor the forecast with  $c = 0$  (shown), which has a  
 228 local minimum of probability at the outcome. CRPS expresses a deliberate  
 229 robust behavior scoring this family of forecasts in a way that is unreasonable  
 230 if not unacceptable.

231 Alternatively, one can view this effect in terms of the score as a function  
 232 of the outcome. The thick blue curve in Figure 1b plots the two curves in  
 233 Figure 1c against each other: the thick blue curve in Figure 1b traces the  
 234 trajectory of the point  $(\text{CRPS}(x), \text{IGN}(x))$  as  $x$  goes from  $-10$  to  $+10$ . Note  
 235 that the minimum IGN occurs at a different point along this trajectory than  
 236 the minimum CRPS. Specifically IGN is minimal at  $x = -1$  and  $x = +1$ ,  
 237 CRPS is minimal at  $x = 0$ . The thin red curve traces the trajectory in the  
 238 case when the modes are asymmetric, specifically when they have weights  
 239 .45 (left) and .55 (right). In this case  $\text{IGN}(x)$  is a minimal at  $x = 1$  (the  
 240 unique mode) and CRPS is minimal near  $x = 0.667$ . Thus IGN scores the  
 241 forecast as better when the outcome corresponds to large  $p(x)$  as might be  
 242 deemed desirable; CRPS does not. While it might be possible to construct  
 243 a situation where these behaviors of CPRS are desirable, these examples  
 244 suggest CRPS be interpreted with great caution, if used at all, in normal  
 245 forecast evaluation.

### 246 **3 Contrasting the skill of decadal forecasts under** 247 **different scores**

248 In this section the behaviour and utility of different scores are contrasted  
 249 by evaluating the performance of probabilistic decadal hindcasts of global  
 250 mean temperature (GMT) from a simulation model (HadCM3) and from two  
 251 simple empirical models (Static Climatology and Dynamic Climatology).



Such evaluations allow comparisons of the relative skill of large simulation models against simple, computationally inexpensive, empirical models. The interpretation of that comparison, and its value, will vary with the score used.

### 3.1 Simulation-based hindcasts

The simulation based forecast system uses simulations from the UK Met Office HadCM3 model [17], which formed part of the CMIP5 decadal hindcast experiment [36]. The forecast archive consists of a series of 10-member initial condition ensembles, launched annually between 1960-2009, and extended out to a lead time of 120 months. This HadCM3 forecast archive was from the CMIP5 library (last downloaded on 07-04-2014). Even so, the small forecast-outcome archive is a limiting factor in the analysis, especially since generating probabilistic forecasts from the ensemble members [5, 35] and the subsequent evaluation must be done in such a way as to avoid using the same information more than once (hereafter, information contamination) [37, 32].

Figure 2 shows the 10-member ensembles of simulated GMT values for every tenth launch year over the full hindcast period; HadCRUT3 observations [8] are shown for comparison. It is clear that the HadCM3 ensemble members are generally cooler than the observed temperatures from HadCRUT3 and would perform poorly if this systematic error were not accounted for. Unless otherwise noted, the ensemble interpretation applied below uses a lead-time dependent offset to account for this systematic error in HadCM3 simulations; the translation of model-values in the simulation into target quantities in the world is an important feature of the forecast system. Unless otherwise stated the ensemble is interpreted as a probability forecast, using the Ignorance score to determine the lead-time-dependent kernel offset and kernel width parameters under cross-validation. This procedure is described further in [5, 35, 32].

### 3.2 The Dynamic Climatology empirical model

The Dynamic Climatology (DC) is an empirical model [31, 35] which uses the observed GMT record. At each launch time, the GMT value is initialised to its observed value from the HadCRUT3 record. An  $\ell$ -step ahead ensemble forecast is generated by adding the set of observed  $\ell^{th}$  differences (across the observed GMT record) to the initialised GMT value at launch, leaving out the period under consideration itself, (that is, adopting a cross-

validation approach). For example a one year ahead forecast made in 1992 for the year 1993 is generated by adding each of the annually averaged consecutive year temperature differences between the years 1960-2012, *except for the 1992-93 difference itself*, to the observed annual average GMT value for the year 1992. Similarly, an  $n$  year ahead forecast is generated from the observed 1992 temperature and all the  $n$  year temperature differences over the hindcast period except for an interval<sup>3</sup> about the point being forecast, this is a direct DC model. In general, one expects the dynamics of uncertainty to vary with initial condition [30], this version of DC does not exploit that expectation: for a given lead time the same distribution of change in GMT is forecast each time. Note that if only non-overlapping intervals are considered, then these ensemble members are independent, as opposed to the HadCM3 ensembles which are ten internally consistent trajectories and are artificially enhanced by access to information from events during that period (volcanos, for example). Generating trajectories from iterated DC models based on a sum of repeated draws from the distribution of one-year differences is also possible; doing so would require assumptions on temporal correlations, and the simpler direct DC scheme is adopted here as it already provides an interesting baseline for comparison with simulation models. A Static Climatology (SC) distribution is also generated as a reference forecast by directly kernel density estimation [29, 5] the observed GMT values over the period 1960-2009.

DC hindcasts are generated for every year in the period 1960-2009 for comparison with HadCM3. HadCM3 ensembles, each with 10 members, are available for every year from 1960 until 2009. Given that a ten year forecast evaluated with the target observed in year  $y$  shares 9 common years with the target in year  $y - 1$  and that in year  $y + 1$ , information contamination is unavoidable if information involving these three years ( $y - 1$ ,  $y$ , and  $y + 1$ ) is treated as independent. For this reason<sup>4</sup>, the experiment was repeated independently starting in 1960, 1961, 1962, 1963, and 1964; for HadCM3 forecast systems the scores shown in the figure 3 reflect the average of the

---

<sup>3</sup>For  $n = 1$ , only the target difference is omitted; for other values of  $n$  the interval is centered on the target difference and ranges from minus  $n_{omit}$  to plus  $n_{omit}$ , where  $n_{omit}$  is the largest integer less than or equal to  $\frac{n}{2}$ .

<sup>4</sup>If a ten year DC forecast launched in 1961 was to include information from a ten year forward difference from 1960, it would be artificially skilful as the temperature difference between 1970 and 1960 is certain to resemble the target difference (between 1971 and 1961). More generally, the score of a ten year forecast for a slowly varying quantity launched in 1960 is not independent of skill of the same forecast system applied to 1961. Even without any direct information contamination from the use of overlapping windows, this serial dependence complicates the interpretation of the cumulative score. [38, 20]

319 result and the max-min range when the vertical bars have no caps. For  
320 the Static Climatology, bootstrap resampling bars are shown, with caps at  
321 the 10% and 90% range (as in Figure 3a). Forecast system under both  
322 approaches are shown from DC in Figure 3; note the results are similar  
323 except for the expected increase due to smaller samples in the independent  
324 experiment case (with caps).

## 325 4 Interpreting probabilistic forecast skill scores

326 In this section, the evaluation of probabilistic hindcasts from the HadCM3  
327 and DC models under different scores are interpreted and contrasted. The  
328 Static Climatology is taken as a reference forecast. Given the evident (phys-  
329 ically expected and causally argued prior to 1960) upward drift in GMT, DC  
330 would be expected to provide a more relevant reference forecast. [35]

331 The top three panels of Figure 3 show skill according to the three different  
332 scores as a function of lead time. Sampling uncertainty in the skill score  
333 (due to the limited number of forecasts considered) is reflected in bootstrap  
334 resampling range (plotted as vertical bars with caps) of the scores for each  
335 lead time, with the 10%-90% resampling intervals. The bootstrap resamples  
336 with replacement from the sample of forecast values; when the sample size  
337 is small these ranges can be large due merely to a few poor forecasts. This  
338 is a property of the size of the forecast-outcome archive, and may happen  
339 even when the outcome is drawn from the forecast distribution (that is,  $Q$   
340 above), although this may be unlikely to happen. These resampling bars  
341 (with caps) are shown in figure 3 for the SC scores (black dotted) and the  
342 traditional unified DC scores (green dashed) [35]; in these cases the sample  
343 size is relatively large. The outcomes of two ten year forecasts initiated  
344 in consecutive years are far from independent (as they have nine years in  
345 common). For this reason five evaluation experiments were considered, with  
346 consecutive initial conditions within each experiment separated by a period  
347 of five years (that is, 1960, 1965, 1970 ...). The vertical bars without caps in  
348 figure 3 reflect the results of repeating the entire forecast evaluation 5 times,  
349 one experiment initialized in each of 1960, 1961, 1962, 1963, and 1964. The  
350 vertical bars (without caps) show the range of these experiments, the solid  
351 line connects their mean.

352 It is clear that the different scores lead to different estimates of the rel-  
353 ative skill provided by the alternate models. When the multiple-realization  
354 bars (no caps) overlap, then there is at least one set of experiments in which,  
355 at that lead time, the forecast system judged better on average performs less

356 well than the forecast system which does less well when the results are av-  
 357 eraged. Overlap between HadCM3 and DC is common under each score.  
 358 Looking at the relative Ignorance directly (Figure 3d) shows that HadCM3  
 359 outperforms DC in every individual case for lead times of 1, 2, 3 and 4 years.  
 360 The extent to which the absolute values are meaningful varies with the score  
 361 considered. In the case of the Ignorance score, the difference between two  
 362 forecast systems reflects the number of additional “bits of information” in  
 363 the better forecast: a difference of 2 bits corresponds to the better forecast  
 364 system placing (on average,  $2^2 =$ ) 4 times more probability on the outcome  
 365 than the alternative forecast system, while a relative IGN of 4 bits would  
 366 correspond to a factor of 16 and a difference of 0.5 a factor of roughly 1.41  
 367 (that is  $2^{1/2}$ ), in other words half a bit corresponds to a gain of about 41%.  
 368 For the other scores, the authors are not aware of any clear interpretation of  
 369 the absolute value of the score. In some cases it makes sense to consider an  
 370 integration over the “True” distribution ( $Q$ , above); in that case the expect-  
 371 ation of the PL is the mean square difference between the forecast density  $p$   
 372 and the density from which the outcome is drawn  $Q$ . The interpretation of  
 373 the expectation with respect to  $Q$  is cloudy in weather-like forecasting sce-  
 374 narios, where the same  $Q$  distribution is never seen twice over the lifetime  
 375 of the system.<sup>5</sup> The Proper Linear score could be interpreted in cases where  
 376 the second term in its definition (equation 5) is motivated by the application  
 377 (not merely for the sake of “making” the naive linear score proper).

378 Each score considered indicates that HadCM3 and DC consistently out-  
 379 perform the Static Climatology. The Ignorance score allows the simple in-  
 380 terpretation of Figure 3d that on average the HadCM3 ensemble decadal  
 381 forecasts place about 70% more probability on the outcome as DC in year  
 382 one, then just over half a bit ( $\sim 41\%$  more) at longer lead times. Figure 3c  
 383 shows that both the HadCM3 and DC models consistently place significantly  
 384 more probability on the outcome than the Static Climatology.

385 Note that SC is roughly constant across lead times, which is to be ex-  
 386 pected as the same forecast distributions is issued (ignoring cross validation  
 387 changes and the effect of the trend) for all lead times. Note also that this  
 388 HadCM3 forecast system outperforms DC, while the HadGEM2 forecast sys-  
 389 tem reported in [35] did not outperform DC. Detailed reasons why this is the  
 390 case are beyond the scope of this paper, nevertheless note (i) the HadCM3  
 391 system considered in this paper had ten ensemble members launched annu-  
 392 ally; whereas the HadGEM2 forecast system had only 3 members launched

---

<sup>5</sup>We thank an anonymous reviewer for stressing the relevance of this interpretation.  
 The result follows from a calculation similar to that found in [6].

every 5 years. (ii) some<sup>6</sup> CMIP5 models are forced by major volcanos, while the DC is not (the hindcasts for the GCMs include specific information on specific years, this version of DC does not), (iii) the multiple-realization bars (no caps) of HadCM3 and DC often overlap in CRPS and PL while the relative IGN in panel d shows a clear separation out to lead time five years or more; on average HadCM3 consistently scores just over half a bit better than DC.

One expects that as simulations, observations, models and ensemble experimental designs improve, the simulation forecast systems will outperform DC even more clearly. Future work will consider the design of better benchmark empirical models, accounting for (and quantifying) the false skill in forecast systems based upon CMIP simulations arising from their foreknowledge of events (volcano-like information), and relative skill in higher resolution targets (finer resolution in space and/or time).

Climate models are sometimes said to show more skill over longer temporal averages; the basis of this claim is unclear. Forecasts of five-year time averages of GMT from the HadCM3 and DC models (not shown) have similar levels of relative probabilistic skill to those of one-year averaged forecasts. The variance in “temperature” decreases when five year means are taken, and the apparent RMS error may appear “smaller”. Note, however, that the metric has changed as well, hence the scare quotes. The probability of the outcome in the two cases changes only slightly, indicating that in this case at least, the suggested gain in skill is a chimera.

## 5 Conclusions

Measures of skill play a critical role in the development, deployment and application of probability forecasts. The choice of score quite literally determines what can be seen in the forecasts, influencing not only forecast system design and model development, but also decisions on whether or not to purchase forecasts from that forecast system or invest in accordance with the probabilities from a forecast system.

The properties of some common skill scores have been discussed and illustrated. Even when the discussion is restricted to proper scores, there remains considerable variability between scores in terms of their sensitivity to outcomes in regions of low (or vanishing) probability; proper scores need not rank competing forecast systems in the same order when each forecast

---

<sup>6</sup>A comparison contrasting forecast systems which include this information from those which do not will be reported elsewhere.

428 system is imperfect. In general, the Continuous Ranked Probability Score  
 429 can define the best forecast system to be one which consistently assigns zero  
 430 probability to the observed outcome, while the Ignorance score will assign an  
 431 infinite penalty to an outcome which falls in a region the forecast states to  
 432 be impossible; such issues should be considered when deciding which score  
 433 is appropriate for a specific task. Ensemble interpretations [5] which inter-  
 434 pret a probability forecast as a single delta function (such as the ensemble  
 435 mean) or as a collection of delta functions (reflecting, for example, the posi-  
 436 tion of each ensemble member) rather than considering all the probabilistic  
 437 information available may provide misleading estimates of skill in nonlinear  
 438 systems. Scores can be used for a variety of different aims, of course. The  
 439 properties desired of a score for parameter selection [25, 9] can be rather  
 440 different from those desired in evaluating an operational forecast system.

441 A general methodology has been applied for probabilistic forecast eval-  
 442 uation, contrasting the properties of several proper scores when evaluating  
 443 forecast systems of decadal ensemble hindcasts of global mean temperature  
 444 from the HadCM3 model (part of the CMIP5 decadal archive). Each of  
 445 the three proper scores in Section 2 were considered for evaluation of the  
 446 results. The Ignorance score was shown to best discriminate between the  
 447 performance of the different models. In addition, the Ignorance score can be  
 448 interpreted directly, indicating, for example, that on average the HadCM3  
 449 forecast system places about 40% more probability on the outcome (half  
 450 a bit) than DC. Observations like these illustrate the advantages of scores  
 451 which allow intuitive interpretation of relative forecast merits.

452 Enhanced use of empirical benchmark models in forecast evaluation and  
 453 in deployment can motivate a deeper evaluation of simulation models. The  
 454 use of empirical models as benchmarks allows the comparison of skill be-  
 455 tween forecast systems based upon state-of-the-art simulation models and  
 456 those using simpler, inexpensive alternatives. As models evolve and improve,  
 457 such benchmarks allow one to quantify this improvement: the HadCM3 fore-  
 458 cast system in this paper out-performs DC, whereas a HadGEM2 forecast  
 459 system (with its smaller ensemble size) did not [35]. This cannot be done  
 460 purely through the intercomparison of an (evolving) set of state-of-the-art  
 461 models. The use of task-appropriate scores can better convey the informa-  
 462 tion available from near-term (decadal) forecasts to inform decision making.  
 463 It can also be of use in judging limits on the likely fidelity of centennial  
 464 forecasts. Ideally, identifying where the most reliable decadal information  
 465 lies today, and communicating the limits in the fidelity expected from the  
 466 best available probability forecasts, can both improve decision making and  
 467 strengthen the credibility of science in support of policy making.

## 468 acknowledgements

469 This research was funded as part of the NERC EQUIP project (NE/H003479/1);  
470 it was also supported by the LSE's Grantham Research Institute on Climate  
471 Change and the Environment, and the ESRC Centre for Climate Change  
472 Economics and Policy, funded by the Economic and Social Research Coun-  
473 cil and Munich Re. L.A.S. gratefully acknowledges support from Pembroke  
474 College, Oxford. T.M. gratefully acknowledges Lloyd's of London for sup-  
475 port for his graduate work. L.A.S. and H.D. acknowledge the support of  
476 RDCEP under NSF grant No. 0951576 and EPSRC grant EP/K013661/1  
477 for the completion of this work.

## 478 References

- 479 [1] K. J. Beven, A Manifesto for the Equifinality Thesis. *Journal of Hy-*  
480 *drology*, 320 (1-2), 18-36, (2006).
- 481 [2] E. Borel, Probabilities and life, Dover, New York, (1962).
- 482 [3] G. W. Brier, Verification of forecasts expressed in terms of probability.  
483 *Monthly Weather Review* 78:1-3 (1950).
- 484 [4] J. Bröcker and L. A. Smith, Scoring probabilistic forecasts: The im-  
485 portance of being proper. *Weather and Forecasting* 22:382-388 (2007).
- 486 [5] J. Bröcker and L. A. Smith, From ensemble forecasts to predictive  
487 distribution functions. *Tellus A* 60:663-678 (2008).
- 488 [6] J. Bröcker, Reliability, sufficiency and the decomposition of proper  
489 scores. *Quarterly Journal of the Royal Meteorological Society* 135:1512-  
490 1519 (2009).
- 491 [7] J. Bröcker, Evaluating raw ensembles with the continuous ranked  
492 probability score. *Quarterly Journal of the Royal Meteorological Soci-*  
493 *ety* 138:1611-1617 (2012).
- 494 [8] P. Brohan, J. J. Kennedy, I. Harris, S. F. B. Tett and P. D. Jones,  
495 Uncertainty estimates in regional and global observed temperature  
496 changes: a new dataset from 1850. *Journal of Geophysical Research*  
497 111, D12106 (2006).
- 498 [9] H. Du and L. A. Smith, Parameter estimation using ignorance. *Phys-*  
499 *ical Review E* 86, 016213 (2012).

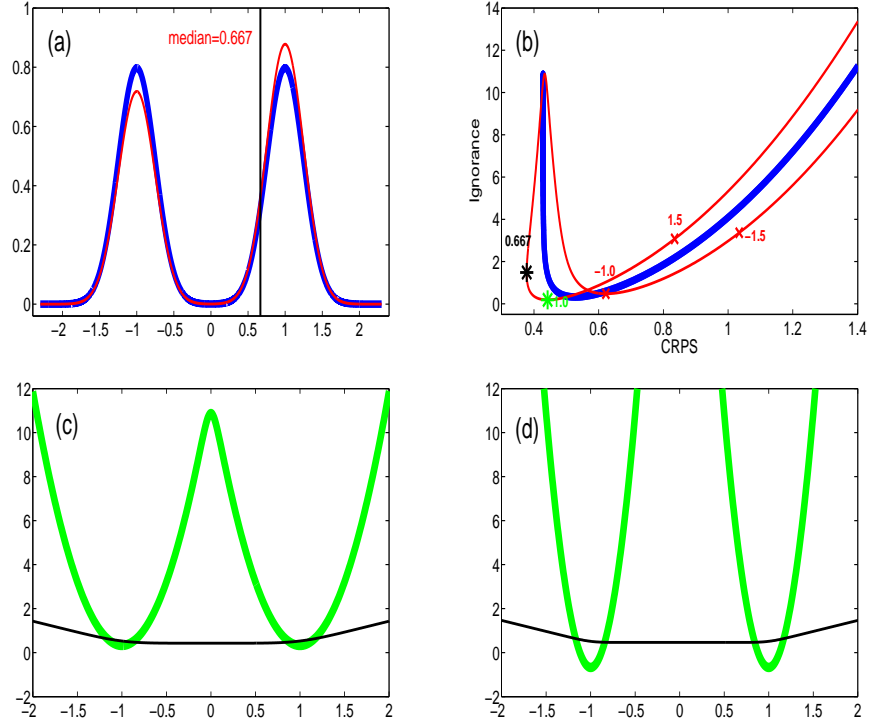


- 500 [10] E. S. Epstein, A scoring system for probability forecasts of ranked  
501 categories. *Journal of Applied Meteorology* 8:985-987 (1969).
- 502 [11] C. A. T. Ferro, D. S. Richardson and A. P. Weigel, On the effect of  
503 ensemble size on the discrete and continuous ranked probability scores.  
504 *Meteorological Applications* 15:19-24 (2008).
- 505 [12] T. E. Fricker, C. A. T. Ferro and D. B. Stephenson, Three recommen-  
506 dations for evaluating climate predictions. *Meteorological Applications*  
507 20, 2:246-255 (2013).
- 508 [13] D. Friedman, Effective scoring rules for probabilistic forecasts. *Man-  
509 agement Science* 78 1:1-3 (1983).
- 510 [14] T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction  
511 and estimation. *Journal of the American Statistical Association* 102,  
512 477:359-378 (2007).
- 513 [15] L. Goddard *et al.*, A verification framework for interannual-to-decadal  
514 predictions experiments. *Climate Dynamics* 40:245-272 (2013).
- 515 [16] I. J. Good, Rational decisions. *Journal of the Royal Statistical Society*  
516 XIV(1):107C114 (1952).
- 517 [17] C. Gordon *et al.*, The simulation of SST, sea ice extents and ocean heat  
518 transports in a version of the Hadley Centre coupled model without  
519 flux adjustments. *Climate Dynamics* 16:147-168 (2000).
- 520 [18] R. Hagedorn and L. A. Smith, Communicating the value of proba-  
521 bilistic forecasts with weather roulette. *Meteorological Applications* 16  
522 2:143-155 (2009).
- 523 [19] H. Hersbach, Decomposition of the continuous ranked probabil-  
524 ity score for ensemble prediction systems. *Weather and Forecasting*  
525 15:559-570 (2000).
- 526 [20] A. Jarman, On the provision, reliability, and use of hurricane forecasts  
527 on various timescales, PhD thesis, The London School of Economics  
528 and Political Science, (2014).
- 529 [21] I. T. Jolliffe and D. B. Stephenson, Forecast verification: a practi-  
530 tioner's guide in atmospheric science. *2nd Ed. John Wiley & Sons*  
531 *Ltd., Hoboken, NJ* (2012).

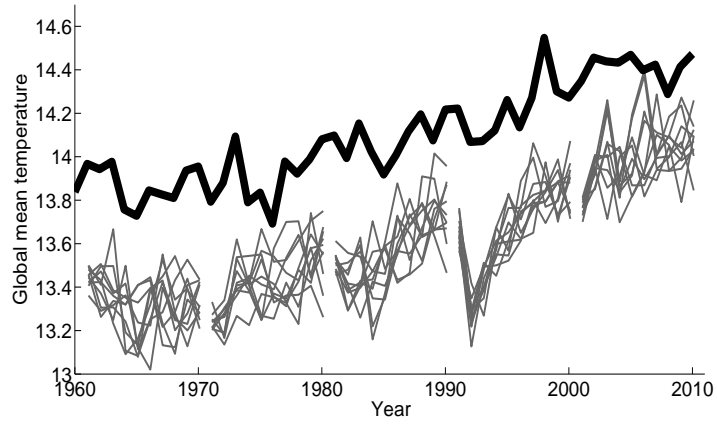


- 532 [22] J. Kelly, A new interpretation of information rate. *Bell Systems Tech-*  
533 *nical Journal* 35:916-926 (1956).
- 534 [23] S. J. Mason and A. P. Weigel, A generic forecast verification frame-  
535 work for administrative purposes. *Monthly Weather Review* 137:331-  
536 349 (2009).
- 537 [24] P. Nurmi, Recommendations on the verification of local weather fore-  
538 casts. *ECMWF Technical Memoranda, Reading, UK* 430 (2003).
- 539 [25] V. F. Pisarenko and D. Sornette, Statistical methods of parameter  
540 estimation for deterministically chaotic time series, *Physical Review E*  
541 69, 036122 (2004).
- 542 [26] D. A. Randall *et al.* Climate models and their evaluation. *Contri-*  
543 *bution of Working Group I to the Fourth Assessment Report of the*  
544 *Intergovernmental Panel on Climate Change; The Physical Science*  
545 *Basis* 589-662 (2007).
- 546 [27] M. S. Roulston and L. A. Smith, Evaluating probabilistic forecasts  
547 using information theory *Monthly Weather Review* 130 6:1653-1660  
548 (2002).
- 549 [28] R. Selten, Axiomatic characterization of the quadratic scoring rule.  
550 *Experimental Economics* 1:43-62 (1998).
- 551 [29] B. W. Silverman, Density Estimation for Statistics and Data Analysis,  
552 Chapman & Hall, (1998).
- 553 [30] L. A. Smith, Local optimal prediction: exploiting strangeness and the  
554 variation of sensitivity to initial condition, *Phil. Trans. Royal Soc.*  
555 *Lond. A*, 348 (1688): 371-381, (1994).
- 556 [31] L. A. Smith, The maintenance of uncertainty *Proceedings Interna-*  
557 *tional School of Physics "Enrico Fermi" Course CXXXIII* 177-246  
558 *Societ'a Italiana de Fisica, Italy* (1997).
- 559 [32] L. A. Smith, H. Du, E. B. Suckling and F. Niehörster, Probabilistic  
560 skill in ensemble seasonal forecasts *Quarterly Journal of the Royal*  
561 *Meteorological Society*, 10.1002/qj.2403.
- 562 [33] C.-A. S. Staël von Holstein, The family of quadratic scoring rules  
563 *Monthly Weather Review* 106 7:917-924 (1978).

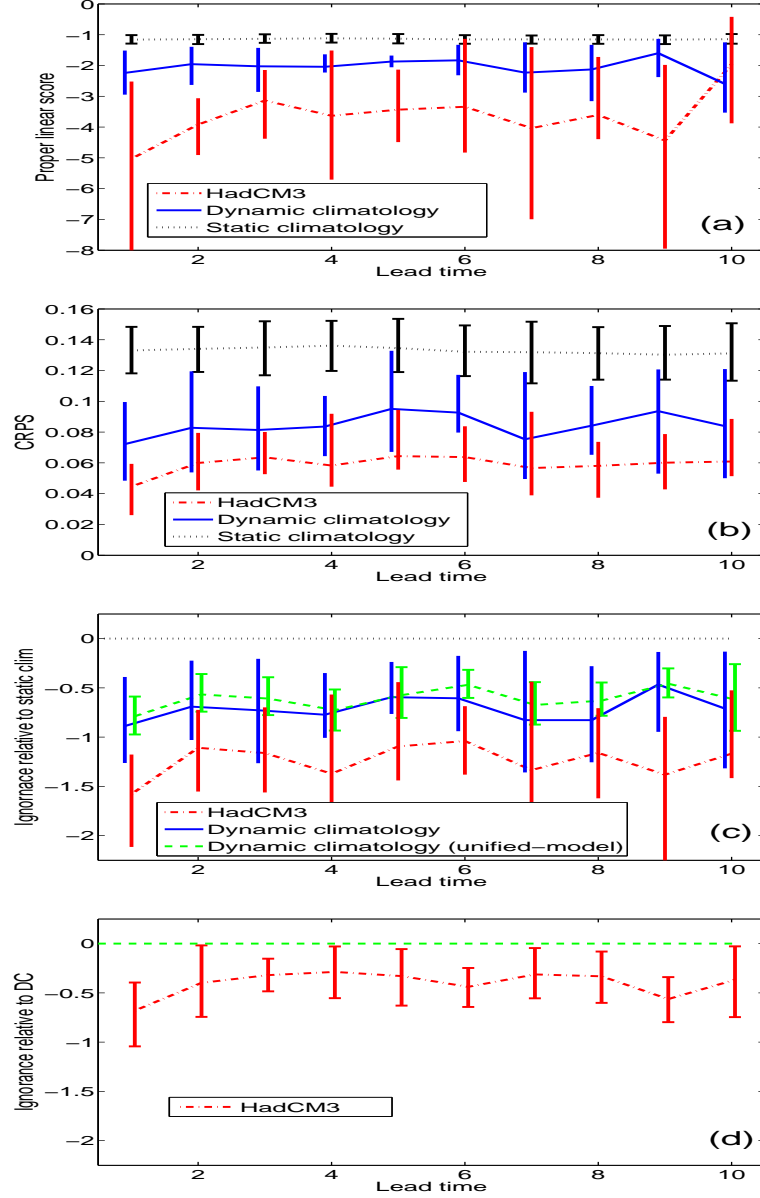
- 564 [34] H. von Storch and F. W. Zwiers, Statistical analysis in climate research  
565 *Cambridge University Press, Cambridge* (1999).
- 566 [35] E. B. Suckling and L. A. Smith, An evaluation of decadal probability  
567 forecasts from state-of-the-art climate models *accepted for publication*  
568 *in Journal of Climate* (2013).
- 569 [36] K. E. Taylor, R. J. Stouffer and G. A. Meehl, An overview of CMIP5  
570 and the experimental design. *Bulletin of the American Meteorological*  
571 *Society* 93 4:485-498 (2008).
- 572 [37] D. S. Wilks. Statistical Methods in the Atmospheric Sciences, Vol-  
573 ume 91, Second Edition (International Geophysics), Academic Press,  
574 2 edition, (2005).
- 575 [38] D. S. Wilks, Sampling distributions of the brier score and brier skill  
576 score under serial dependence. *Quarterly Journal of the Royal Meteorological*  
577 *Society*, 136(653):21092118 (2010).
- 578 [39] World Meteorological Organization, Recommendations for the veri-  
579 fication and intercomparison of QPFs and PQPFs from operational  
580 NWP models. *World Meteorological Organization: Geneva, Switzerland*  
581 (2008).



**Figure 1:** An example comparing the sensitivity of IGN (thick green) and CRPS (thin black) scores for outcomes in different regions of a forecast probability distribution. (a) Two bimodal forecast distributions, one symmetric (thick blue) and one asymmetric (thin red). (b) The Ignorance (y-axis) and CRPS (x-axis) scores given to each forecast distribution as the observed outcome moves across the range of each distribution. Note that minimal (best) scores occur for CRPS when the outcome falls at the median of the forecast distribution, while Ignorance is minimal when the outcome falls at a mode of the forecast distribution. Panels (c) and (d) show the Ignorance score (thick green) and CRPS score (thin black) as a function of the outcome given a symmetric bimodal forecast distribution. All forecast distributions consist of the sum of two Gaussian distributions, one centred at  $-1$ , the other at  $+1$ . Panels (a), (b) and (c) reflect the results where each component has a standard deviation of  $0.25$ . In panel (d) each component has a standard deviation of  $0.125$ . In the symmetric forecasts, each component is equally weighted, while in the asymmetric forecast (reflected in the thin red curves of panel (a) and (b)) the left component has weight  $0.45$  and the right  $0.55$ .



**Figure 2:** Individual HadCM3 ensemble members (thin grey) and HadCRUT3 observations (thick black) of global mean temperature (GMT) between 1960 and 2010. For clarity, only every tenth launch date of the HadCM3 simulations are shown.



**Figure 3:** Performance of HadCM3 and DC forecast systems as a function of lead time under different skill scores: (a) PL score, (b) CRPS, (c) IGN relative to the Static Climatology and (d) IGN relative to DC. In panels (a), (b), and (c) the Static Climatology (SC) is shown for comparison; in panel (c) both HadCM3 and DC perform substantially better than SC on average; multiple-realization sample bars (vertical bars, no caps) show that this is the case in almost every realization. A unified DC forecast system (green dashed) is shown for comparison; traditional (10%-90%) bootstrap resample ranges (green dashed, with caps) reveal a similar result with somewhat improved sampling uncertainty. In Panel (d) the red dash-dotted line fluctuates between -0.25 bit to -0.75 bit indicating that on average the HadCM3 forecast system clearly outperforms the unified DC, placing between ~20% and 60% more probability on the outcome than DC at various lead times. Some of the multiple-realization sample bars (no caps) reach zero in panel (d), indicating that in some realizations the DC outperforms HadCM3.