Received: 22 February 2024 | Revised: 13 May 2024 | Accepted: 14 May 2024

DOI: 10.1002/ffo2.189

ORIGINAL ARTICLE

WILEY

# Scoring rules and performance, new analysis of expert judgment data

Gabriela F. Nane[1] | Roger M. Cooke[1,2]

[1]Department of Applied Mathematics, Delft University of Technology, Delft, The Netherlands

[2]Resources for the Future, Washington, District of Columbia, USA

**Correspondence**
Gabriela F. Nane
Email: g.f.nane@tudelft.nl

**Funding information**
None

## Abstract

A review of scoring rules highlights the distinction between rewarding honesty and rewarding quality. This motivates the introduction of a scale-invariant version of the Continuous Ranked Probability Score (CRPS) which enables statistical accuracy (SA) testing based on an exact rather than an asymptotic distribution of the density of convolutions. A recent data set of 6761 expert probabilistic forecasts for questions for which the actual values are known is used to compare performance. New insights include that (a) variance due to assessed variables dominates variance due to experts, (b) performance on mean absolute percentage error (MAPE) is weakly related to SA (c) scale-invariant CRPS combinations compete with the Classical Model (CM) on SA and MAPE, and (d) CRPS is more forgiving with regard to SA than the CM as CRPS is insensitive to location bias.

**KEYWORDS**

Brier score, Classical Model, Continuous Ranked Probability Score, expert judgment, geometric probability, location bias, logarithmic score, mean absolute percentage error, overconfidence, probability interval score, scoring rules

## 1 | INTRODUCTION

Continuous Ranked Probability Scores (*CRPSs*), Probability Interval Scores (*PISs*), and scores from the Classical Model (*CM*) have recently captured attention in evaluating COVID-19 probabilistic model predictions (Colonna et al., 2022; Cramer et al., 2022; Ray et al., 2020), fueling the debate on how probabilistic predictions should be evaluated. This article offers insights from a recent expert judgment data set (Cooke et al., 2021) comprised of expert probabilistic predictions over a wide variety of fields for which realizations or true values are also available.

Familiarity with foundations teaches that the problem of combining and evaluating experts' probabilistic predictions is not a purely mathematical problem. The laws of probability even supplemented with Savage's axioms of rational decision theory and the theory of proper scoring rules, will never tell us how best to combine experts' judgments. The problem is more akin to finding an optimal design in engineering. A bicycle after all obeys Newton's laws but does not follow from them. Any working design will involve features motivated by practicalities rather than laws. An example is the measurement of "spread" of a distribution. If our data are measured on different dimensions (e.g., meters, micrograms per cubic meter, etc.) then traditional measures like the standard deviation and prediction intervals are unsuitable because they inherit the physical dimension of the underlying variables: changing meters to kilometers changes some of the numbers. Comparing spreads across variables with different physical dimensions requires a scale-invariant measure. If, as is the case with expert judgment, the tails of the distribution are poorly constrained in the data, we also need it to be "tail insensitive." Agreement of probabilistic predictions with reality, variously called statistical accuracy (SA) or calibration, must accommodate aggregating over variables and combining experts. Such practical issues must be addressed in deciding

how to evaluate probabilistic predictions. The first step is thus to inform our discussion as to available data.

The following sections address the expert data, scoring rules, and performance. Conclusions are drawn in the final section.

## 2 | EXPERT DATA

This paper uses data from 49 studies involving 526 experts assessing in total 580 calibration variables from their fields, for which realizations are known (four experts from the original data were dropped because they did not assess all calibration variables in their respective panels). In total there are 6761 expert probabilistic forecasts of variables from their fields for which true values are known. The data are described and referenced in Cooke et al. (2021). The supplementary information of that article gives a description of the *CM* (Cooke, 1991), which informed the data collection. Its relevant aspects are briefly reviewed in this section.

The number of assessed calibration variables differ per study and Figure 1 provides information about this. Experts assessed at least seven and at most 21 calibration variables in all studies. Ten calibration variables (the default number) were used in 21 of the 49 studies. In the data used for this analysis, expert assessments take the form of fixed percentiles, 5th, 50th, and95th, from the assessor's subjective distribution for a continuously distributed unknown quantity.

*CM* uses weights derived from expert performance on calibration variables to derive combinations of expert distributions (termed Decision Makers [*DMs*] in a linear pool, see [Cooke, 1991] ch.11 for an extended discussion of pooling). In the *CM*, *SA* is measured as the probability of falsely rejecting the hypothesis that a probabilistic assessor is statistically accurate. It is, in other words, the *p* value of rejection for this hypothesis. We hasten to add that *CM* does *not* test and reject expert hypotheses but, in compliance with proper scoring rule theory for sets of assessments (see below), uses this *p* value to measure the degree of correspondence between assessments and data in forming weighted combinations of expert distributions. When $n$ true values for a number of such quantities are observed, we compute the sample distribution $s$ of interquantile relative frequencies and compare this with the theoretical interquantile mass function $p = (0.05, 0.45, 0.45, 0.05)$. The test statistic is $2nI(s, p)$, where $I$ is the Shannon relative information (log-likelihood ratio) and $n$ is the number of calibration variables. Assuming that the realizations are independently sampled from the assessor's distributions, this statistic is asymptotically $\chi^2$ distributed with degrees of freedom equal to the number of assessed percentiles. Thus, *SA* is measured as $1 - F_{\chi^2}(2nI(s, p))$, where $F_{\chi^2}$ is the cumulative distribution function (*CDF*) of a $\chi^2$ distribution with three degrees of freedom. Low scores (near 0) mean it is unlikely that the divergence between $s$ and $p$ should arise by chance. Higher scores (near 1) indicate better agreement between $s$ and $p$. Note that *CM* relies on an asymptotic approximation which for a small number of calibration variables is not very good (Cooke, 2014). Simulations for 10 calibration variables are provided in Hanea and Nane (2021). The approximation is deemed capable of detecting only large differences in experts' performances which are indeed usually present. It is noteworthy that SA uses only the assessed percentiles and does not rely on an interpolated *CDF*.

With the expert-provided assessments for the calibration variables, realizations are expected to fall below experts' 5th percentile 5% of the time, to fall between the 5th and 50th percentile 45% of the time, and so on. The (interpolated) percentiles of the realizations for all expert probabilistic predictions are shown in Figure 2, revealing the realizations' concentration in very low and very high percentiles.

In *CM* informativeness (*Inf*) is measured as the Shannon relative information of the minimal information fit to the experts' percentiles, relative to a user-selected background measure. In this analysis, the background measure per variable is always uniform on an interval 20% larger than the smallest interval containing all experts' assessments and the realization. In this case, the minimal information fit is piecewise

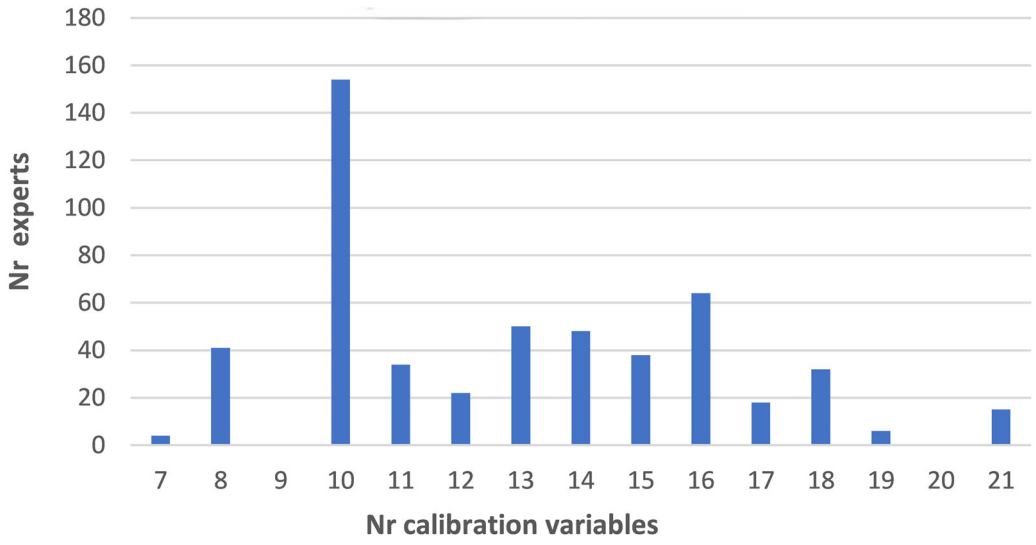


**FIGURE 1** Number of experts per number of assessed calibration variables.

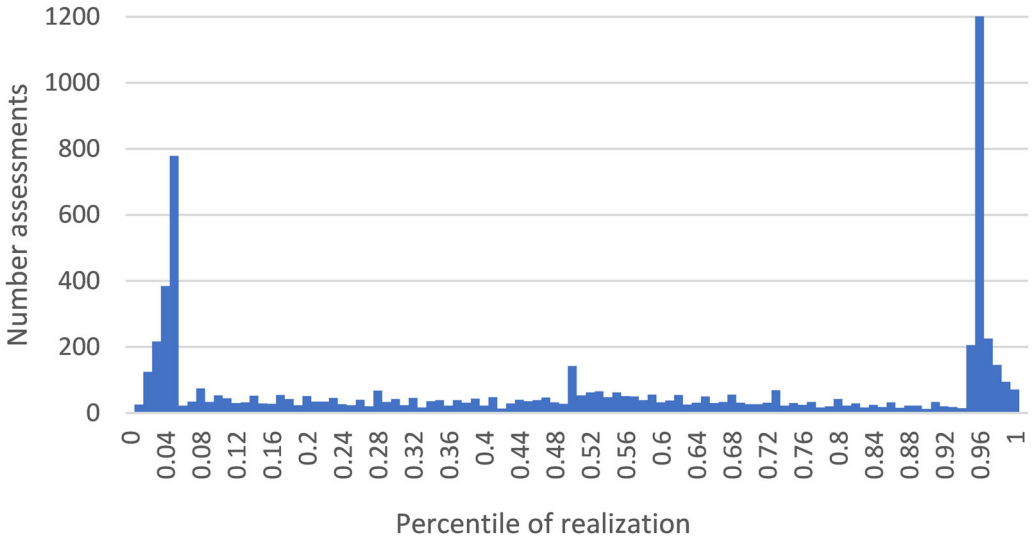


**FIGURE 2** Frequency of percentiles of realizations for 6761 expert probabilistic forecasts from 49 studies.

uniform such that the mass in each interquantile interval complies with the expert's assessment. Relative information is tail insensitive and "slow" so that the experts' information scores are quite insensitive to the size of the background measure. This slowness means that the ratio of *Inf* scores is much less variable over experts than the ratio of *SA* scores, see Cooke et al. (2021) for details.

The variability in expert judgment data is large. Two important sources of variability are the experts themselves and the variables assessed. Different experts have different backgrounds and different heuristics for giving probabilistic predictions. Variables differ in their relation to the experts' knowledge base. Using the realizations of experts' predictions, we can quantify the contributions from these two sources. We illustrate this with the ice sheet study of 2018 (Bamber et al., 2019). Table 1 arranges the 20 experts in rows and the 16 calibration variables in columns. Each cell contains the (interpolated) *CDF* value for the realization for each row expert's assessment of the column variable. The bottom row and, respectively, the rightmost column contain the marginal averages. The eyeball sees that the column averages are more variable than the row averages.

Let *Var* denote the variance in the entire expert-variable matrix and let $Var[E(e|v)]$ denote the variance of the expectations for the experts' probability densities given variable $v$, and so on. The law of total variance gives:

$$Var = Var[E(e|v)] + E[Var(e|v)] = Var[E(v|e)] + E[Var(v|e)].$$

$\frac{Var[E(e \mid v)]}{Var}$ denotes the fraction of the overall variance *Var* that is explained by the variables and similarly $\frac{Var[E[v \mid e)]}{Var}$ is the fraction explained by the experts. For this data $\frac{Var[E(e \mid v])}{Var}$ = 0.553 and $\frac{Var[E(v \mid e)]}{Var}$ = 0.0175. Roughly speaking, this means that much more variability comes from the variables than from the experts. Figure 3 shows this decomposition for the 23 studies with at least 10 experts and 10 calibration variables.

## 3 | SCORING RULES

Scoring rules were introduced by de Finetti in 1937 as tools for encouraging honesty in eliciting subjective probabilities (De Finetti, 1937) and have been further developed by many authors including (Brown, 1974; Gneiting & Raftery, 2007; De Groot & Fienberg, 1983; Murphy, 1977; Savage, 1971; Shuford et al., 1966). The latter reference gives an extensive overview. Carvalho (2016) and Dawid and Musio (2014) review applications of proper scoring rules, and Merkle and Steyvers (2013) investigates how the choice of scoring rules impacts conclusions. An expert receives a score as a function of his/her probability assessment and the realization. The score is strictly proper if the expert maximizes (for negatively sensed rules, minimizes) his/her expected score per item by, and only by, stating his/her true belief. This section discusses scoring rules for individual variables with discrete probabilities, scoring rules for sets of discrete variables, and scoring rules for continuous variables, focusing on rules encountered in practice.

### 3.1 | Scoring rules for individual variables with discrete probabilities

Using a result of Murphy (1977), De Groot and Fienberg (1983) gave an additive decomposition of strictly proper rules into "calibration" and "refinement" terms, thereby replacing Murphy's "resolution" (refinement applies only to well-calibrated experts). In the case of the logarithmic rule, refinement becomes the Kullback-Leibler divergence of the sample distribution of realizations. Some authors (Hersbach, 2000) adopt a framework in which nature picks a distribution for an unknown quantity and forecasters attempt to predict this distribution.

Scoring rules for individual variables were not designed for evaluating or combining experts and are not generally fit for that

**TABLE 1** Variance decomposition for the ice sheet 2018 study (Bamber et al., 2019), with 20 experts and 16 calibration variables.

| %tile rls | vbl1 | vbl2 | vbl3 | vbl4 | vbl5 | vbl6 | vbl7 | vbl8 | vbl9 | vbl10 | vbl11 | vbl12 | vbl13 | vbl14 | vbl15 | vbl16 | $E(v\|e)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| exp1 | 0.572 | 0.035 | 0.905 | 0.043 | 0.048 | 0.199 | 0.045 | 0.380 | 0.082 | 0.954 | 0.050 | 0.045 | 0.951 | 0.951 | 0.950 | 0.985 | 0.450 |
| exp2 | 0.957 | 0.035 | 0.905 | 0.042 | 0.048 | 0.955 | 0.045 | 0.073 | 0.957 | 0.953 | 0.771 | 0.032 | 0.037 | 0.049 | 0.950 | 0.978 | 0.487 |
| exp3 | 0.800 | 0.335 | 0.048 | 0.727 | 0.332 | 0.591 | 0.575 | 0.058 | 0.086 | 0.268 | 0.469 | 0.082 | 0.062 | 0.259 | 0.950 | 0.908 | 0.409 |
| exp4 | 0.952 | 0.039 | 0.050 | 0.502 | 0.049 | 0.957 | 0.122 | 0.061 | 0.029 | 0.952 | 0.906 | 0.548 | 0.078 | 0.049 | 0.975 | 0.963 | 0.452 |
| exp5 | 0.613 | 0.042 | 0.136 | 0.046 | 0.568 | 0.759 | 0.045 | 0.624 | 0.101 | 0.950 | 0.565 | 0.285 | 0.040 | 0.039 | 0.975 | 0.963 | 0.422 |
| exp6 | 0.950 | 0.048 | 0.905 | 0.956 | 0.248 | 0.793 | 0.424 | 0.041 | 0.075 | 0.532 | 0.931 | 0.071 | 0.161 | 0.951 | 0.971 | 0.978 | 0.565 |
| exp7 | 0.170 | 0.253 | 0.924 | 0.838 | 0.958 | 0.955 | 0.098 | 0.018 | 0.101 | 0.532 | 0.437 | 0.328 | 0.078 | 0.207 | 0.974 | 0.806 | 0.480 |
| exp8 | 0.590 | 0.219 | 0.455 | 0.351 | 0.049 | 0.695 | 0.580 | 0.298 | 0.234 | 0.950 | 0.487 | 0.047 | 0.044 | 0.121 | 0.960 | 0.908 | 0.437 |
| exp9 | 0.952 | 0.032 | 0.455 | 0.461 | 0.635 | 0.819 | 0.048 | 0.362 | 0.114 | 0.201 | 0.046 | 0.521 | 0.037 | 0.165 | 0.973 | 0.987 | 0.426 |
| exp10 | 0.320 | 0.042 | 0.950 | 0.164 | 0.400 | 0.955 | 0.122 | 0.298 | 0.040 | 0.952 | 0.758 | 0.048 | 0.124 | 0.049 | 0.976 | 0.963 | 0.448 |
| exp11 | 0.950 | 0.042 | 0.545 | 0.511 | 0.482 | 0.570 | 0.259 | 0.599 | 0.262 | 0.950 | 0.049 | 0.084 | 0.575 | 0.926 | 0.757 | 0.974 | 0.533 |
| exp12 | 0.500 | 0.048 | 0.752 | 0.735 | 0.469 | 0.953 | 0.613 | 0.110 | 0.031 | 0.688 | 0.629 | 0.074 | 0.125 | 0.479 | 0.725 | 0.984 | 0.495 |
| exp13 | 0.950 | 0.031 | 0.455 | 0.043 | 0.049 | 0.955 | 0.311 | 0.073 | 0.101 | 0.950 | 0.375 | 0.113 | 0.029 | 0.021 | 0.950 | 0.963 | 0.398 |
| exp14 | 0.725 | 0.048 | 0.151 | 0.950 | 0.046 | 0.955 | 0.208 | 0.080 | 0.070 | 0.952 | 0.088 | 0.383 | 0.611 | 0.049 | 0.969 | 0.963 | 0.453 |
| exp15 | 0.500 | 0.039 | 0.545 | 0.277 | 0.717 | 0.845 | 0.122 | 0.017 | 0.101 | 0.877 | 0.487 | 0.521 | 0.029 | 0.952 | 0.973 | 0.978 | 0.499 |
| exp16 | 0.500 | 0.037 | 0.095 | 0.046 | 0.182 | 0.956 | 0.424 | 0.951 | 0.075 | 0.954 | 0.717 | 0.045 | 0.106 | 0.049 | 0.978 | 0.758 | 0.430 |
| exp17 | 0.951 | 0.035 | 0.095 | 0.651 | 0.049 | 0.957 | 0.951 | 0.149 | 0.038 | 0.511 | 0.487 | 0.305 | 0.658 | 0.950 | 0.973 | 0.887 | 0.540 |
| exp18 | 0.050 | 0.034 | 0.302 | 0.614 | 0.049 | 0.954 | 0.086 | 0.073 | 0.101 | 0.952 | 0.049 | 0.352 | 0.274 | 0.476 | 0.976 | 0.986 | 0.395 |
| exp19 | 0.950 | 0.088 | 0.132 | 0.247 | 0.049 | 0.533 | 0.406 | 0.298 | 0.040 | 0.950 | 0.771 | 0.071 | 0.512 | 0.239 | 0.975 | 0.978 | 0.452 |
| exp20 | 0.750 | 0.801 | 0.635 | 0.728 | 0.049 | 0.953 | 0.045 | 0.500 | 0.559 | 0.953 | 0.608 | 0.048 | 0.037 | 0.082 | 0.974 | 0.981 | 0.544 |
| $E(e\|v)$ | 0.685 | 0.114 | 0.472 | 0.447 | 0.274 | 0.815 | 0.276 | 0.253 | 0.160 | 0.799 | 0.484 | 0.200 | 0.228 | 0.353 | 0.945 | 0.944 | 0.466 |

*Note*: The percentile which the realization realizes (%ile rls) for expert's *i*'s distribution for the calibration variable *j* is reported in row *i* and column *j*. Averaging these percentiles with respect to each expert ($E(v|e)$) and each variable ($E(v|e)$) is included.

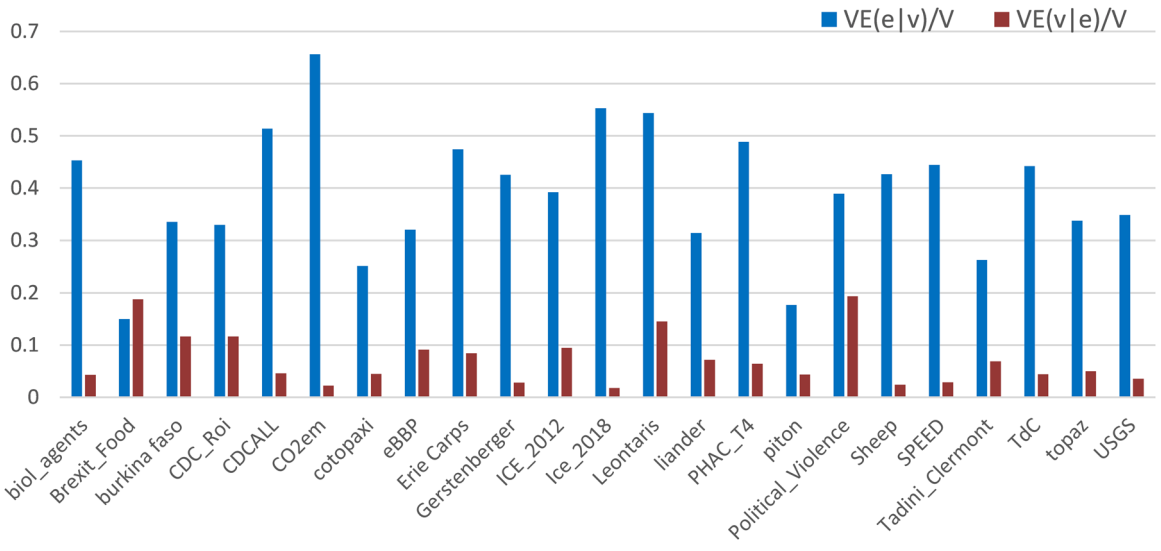


**FIGURE 3** Variance decomposition for 23 studies with at least 10 experts and 10 calibration variables. For each study, the fraction of the overall variance that is explained by the experts ($VE(e|v)/V$, in blue) and the one explained by the variables ($VE(v|e)/V$, in the dark red) are included.

purpose. Indeed, rewarding honesty is not the same as rewarding quality. A simple example illustrates this difference: Consider 100 fair coin tosses. An expert assesses the probability of *heads* on each toss as 1/2. With the standard scoring rules (Brier, logarithmic, spherical, and quadratic), the score for the outcome *heads* is the same as their score for *tails* on each toss. If the score for all 100 assessments is a function of their 100 scores for the individual tosses, then their score for 100 tosses is independent of the outcome sequence; the outcome of 100 *heads* receives the same score as 50 *heads* and 50 *tails*. Equal scores do not imply equal quality.

Another example concerns the popular Brier score for "rain/no rain" events. This rule is twice the mean squared difference between the forecast probability of rain and the outcome indicator (1 for rain, 0 for no rain), negatively sensed on [0, 2] per forecast.[1] Nearly as popular is the logarithmic rule assigning the score of *ln*(probability of outcome) per variable. It is positively sensed on the range $(-\infty, 0]$ per forecast. Consider 1000 next-day forecasts of rain by two experts. Suppose the experts bin their forecasts as shown below (Cooke, 2014) (Table 2).

Ten probability bins are considered, each associated with a forecast probability of rain. The experts' assessments are equally informative in the sense that they each assign the same probabilities to the same number of days. Expert 1 is statistically perfectly accurate, that is, the empirical relative frequency of actual rainy days from the assessed 100 days is identical with the probability associated with each bin. Expert 2 is massively inaccurate statistically. The sample distributions bear little resemblance to his/her assessed probabilities (5%, …, 95%). Expert 1 has a mean Brier score of 0.34 and Expert 2 is a mean Brier score of 0.18, nearly twice as good. For the logarithmic score, Expert 1's mean score is −0.5, Expert 2's mean score is −0.32, again nearly twice as good. Expert 2 gets better scores because the higher resolution of the overall distribution strongly outweighs statistical inaccuracy, even if the sample distributions per bin bear little resemblance to the forecast probabilities. In fact, if we replace {1, 99} for Expert 2 with {17, 83}, the Brier scores for the experts would be equal (for the Logarithmic rule, achieve this result with {15, 85}). Again equal scores do not imply equal quality. Such examples make it difficult to explain to experts and DMs what the numerical values of these scores mean. For more discussion, see Cooke (1991, 2014). In the context of expert judgment, we would like to reward both honesty *and* quality with scoring rules that

**TABLE 2** 1000 rain/no rain probability forecasts for two experts (upper table), along with Brier (middle table) and Logarithmic scores (lower table).

| Probability bin | 5% | 15% | 25% | 35% | 45% | 55% | 65% | 75% | 85% | 95% | Totals | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Expert 1 | | | | | | | | | | | | |
| Assessed | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1000 | |
| Realized | 5 | 15 | 25 | 35 | 45 | 55 | 65 | 75 | 85 | 95 | 500 | |
| Expert 2 | | | | | | | | | | | | |
| Assessed | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 1000 | |
| Realized | 1 | 1 | 1 | 1 | 1 | 99 | 99 | 99 | 99 | 99 | 500 | |
| Brier Score | (= (1 − Quadratic score); negatively sensed in (0,2)) | | | | | | | | | | | Mean score |
| Expert 1 | | | | | | | | | | | | |
| Calibration | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | |
| Resolution | 9.50 | 25.50 | 37.50 | 45.50 | 49.50 | 49.50 | 45.50 | 37.50 | 25.50 | 9.50 | 335.00 | 0.34 |
| Expert 2 | | | | | | | | | | | | |
| Calibration | 0.32 | 3.92 | 11.52 | 23.12 | 38.72 | 38.72 | 23.12 | 11.52 | 3.92 | 0.32 | 155.20 | 0.18 |
| Resolution | 1.98 | 1.98 | 1.98 | 1.98 | 1.98 | 1.98 | 1.98 | 1.98 | 1.98 | 1.98 | 19.80 | |
| Log score | (=ln(*p*(occur)) × #occur + ln(*P*(not occur)) × #not occur; positively sensed in (−infty, 0]) | | | | | | | | | | | |
| Expert 1 | | | | | | | | | | | | |
| Calibration | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | −0.50 |
| Resolution | −19.85 | −42.27 | −56.23 | −64.74 | −68.81 | −68.81 | −64.74 | −56.23 | −42.27 | −19.85 | −503.83 | |
| Expert 2 | | | | | | | | | | | | |
| Calibration | −2.47 | −12.39 | −24.27 | −38.10 | −54.38 | −54.38 | −38.10 | −24.27 | −12.39 | −2.47 | −263.22 | −0.32 |
| Resolution | −5.60 | −5.60 | −5.60 | −5.60 | −5.60 | −5.60 | −5.60 | −5.60 | −5.60 | −5.60 | −56.00 | |

*Note*: The probability forecasts are binned in 10 equally spaced intervals and their distribution is included in the "assessed" rows. The "realized" rows depict which events have occurred. The contribution of the calibration and resolution with respect to each bin is depicted for the Brier and logarithmic scores.

are intuitive and easily explained. This requires numerical insight into the rules' behavior.

## 3.2 | Scoring rules for average probabilities

Scoring rules for average probabilities were introduced to avoid problems with scores for individual items (Cooke, 1991). Let random variables $X_1, \ldots, X_n$ take outcomes in a finite set, $O$, let $M_O$ be the set of probability measures on $O$, and let $M_n$ be the set of probability measures on $X_1, \ldots, X_n$. For $\Pi \in M_n$, let $\pi$ be the vector of average probabilities, that is, $\pi_i = (1/n)\sum_{j=1,\ldots,n}\Pi\{X_j = i\}$. The vector of average probabilities for the outcomes in $O$ is the vector of expected relative frequencies of these outcomes in $X_1, \ldots, X_n$. Let $s$ be the observed relative frequency of outcomes for realization $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$. A scoring rule for average probabilities assigns a number, $R$, to the pair $(\pi, s)$. $R$ is strictly proper (positively sensed) if, for all $\Pi \in M_n$,

$$\arg\max_{\phi \in M_O} E_\Pi[R(\phi, s)] = \pi. \tag{1}$$

This says, whatever the expert's belief, $\Pi$, about $(X_1, \ldots, X_n)$, (s)he achieves the maximal expected score by stating the probability, $\pi$, over outcomes which corresponds to his/her average probabilities according to $\Pi$. The proofs are a bit more complicated because, "$\forall\ \Pi$," goes over a much larger set than the argmax over $M_O$.

There is a representation theorem in Cooke (1991) for such rules. However, more useful in practice are rules which are asymptotically strictly proper as $n \to \infty$. These rules allow the product form in the *CM*. Where $1_A$ is the indicator function for the set $A$ and $\alpha$ is any number strictly between 0 and 1, it is shown that

$$SA \times Inf \times 1_{SA>\alpha},\ \alpha > 0, \tag{2}$$

is asymptotically strictly proper in the set of all product measures over $(X_1, X_2 \ldots, X_n)$. By design, *SA* is a very fast function and *Inf* very slow. This means that when scores in Equation (1) are normalized to sum to one, *SA* strongly dominates and *Inf* modulates between experts with comparable *SA* scores. The presence of the cutoff indicator $1_{SA>\alpha}$ is imposed by the proper scoring rule requirement. The theory of proper scoring rules does not say what the value of $\alpha$ should be, just that there should be some cutoff on *SA* beneath which an expert is unweighted. The $\alpha$ maximizing the score of the combined expert is termed optimal in the *CM* (see Section 4.4).

## 3.3 | Scoring rules for continuous variables

The *PIS*, the related *CRPS*, and the *CM* have recently been applied to COVID-19 models' probabilistic forecasts. Computable examples highlight issues encountered in Section 3.1, namely trading off calibration with resolution; reminding us that equal scores do not imply equal quality. In addition, the scale dependence of *CRPS* impedes aggregation over variables on different scales. We derive a scale-invariant version of *CRPS* and derive a closed form of its convolution to be used in testing experts' *SA* without recourse to an asymptotic distribution.

### 3.3.1 | PISs

Numerical insight into the behavior of these scores requires a bit of effort. The $(1 - \alpha)$ uncertainty interval with upper (lower) bound H ($L$) and probability $(1 - \alpha)$ of catching the true value, has the *PIS* (negatively sensed) (Aitchison & Dunsmore, 1968) for realization $y$:

$$(H - L) + \frac{2}{\alpha} \times [(L - y)_+ + (y - H)_+],$$

where $X_+ = X$ if $X > 0$ and $X_+ = 0$ otherwise. Note that $2/\alpha$ is the slope of the overconfidence penalty for $y \notin [L, H]$. The length $\|H - L\|$ is called the "sharpness" (the resolution component); small values reward concentrated probability mass. $\frac{2}{\alpha} \times [(L - y)_+ + (y - H)_+]$ measures (mis)calibration.

To better understand the characteristics of *PIS*, consider $Y$ uniformly distributed on the interval $[0, 1]$ (hereafter denoted $Y \sim U[0, 1]$) and the $(1 - \alpha)$ uncertainty interval $[L, H]$. Then

$$\begin{aligned} E_Y[PIS(Y)] &= H - L + \frac{2}{\alpha}\int_0^L (L - x)\,dx + \frac{2}{\alpha}\int_H^1 (x - H)\,dx \\ &= H - L + \frac{1}{\alpha}[L^2 + (1 - H)^2]. \end{aligned}$$

For the central 0.9 interval $[0.05, 0.95]$, the expected PIS is 0.95. The interval $[0, 0.9]$ with the same "coverage" has worse expected score 1. Suppose an expert prefers to give an 80% interval,$[0.1, 0.9]$, then the expected score is 0.9. This is better than 0.95 because the prediction interval is sharper. An expert seeking to optimize (i.e., minimize) his/her expected score might take a central 2% prediction interval $[0.49, 0.51]$ with expected score of 0.51. The way in which the *PIS* trades calibration for sharpness may strike some as counterintuitive. For example, an expert claiming that the degenerate interval $[0.5, 0.5]$ has 40% probability of catching the realization would achieve an expected score of 0.833, better than the score of the 90% central interval. The sharpness of an interval of zero length outweighs the overconfidence of claiming 40% mass at the point 0.5.

### 3.3.2 | CPRS

Consider $y$ an unknown scalar quantity of interest. Suppose $y$ has a true, unknown *CDF* $F_Y$, characterizing the distribution of a random variable $Y$. An expert provides *CDF* $F_X$ which (s)he believes to be the distribution of $Y$. We assume both $F_Y$ and $F_X$ are continuous and strictly increasing on their support. The CRPS is defined as (Brown, 1974)

$$CRPS(F_X, y) = \int_{-\infty}^{\infty}[F_X(x) - 1_{\{x \geq y\}}]^2 dx. \tag{3}$$

Lower values indicate better performance. CRPS is known to be strictly proper relative to the class of Borel probability measures with

finite first moment (Gneiting & Raftery, 2007). These authors note: "Applications of the *CRPS* have been hampered by a lack of readily computable solutions to the integral Equation (3)".

This section presents computable solutions to this integral which then allow us to study its trade offs between SA and "sharpness." The *CRPS* can be thrown into a scale-invariant form which offers significant advantages. To understand the behavior of the CRPS score, let us consider $X$ uniformly distributed on $[L, H]$ (written hereafter as $X \sim U[L, H]$), with $0 < L < H < 1$. As we will show later, this particular choice of distribution is relevant for the development of our proposed score. For $y \in [0, 1]$:

$$CRPS(F_X, y) = \begin{cases} L - y + \dfrac{H - L}{3} & \text{for } 0 \le y < L, \\ \dfrac{(y - L)^3}{3(H - L)^2} - \dfrac{(y - H)^3}{3(H - L)^2} & \text{for } y \in [L, H], \\ y - H + \dfrac{H - L}{3} & \text{for } H < y \le 1. \end{cases} \quad (4)$$

Figure 4 shows the CRPS score as a function of $y$, for different cases of $L$ and $H$. Note the possible values are scale dependent. Cases when $y$ falls within and outside the $F_X$ support are highlighted.

The expectation of the CRPS score, which may be infinite, is given by

$$E_Y[CRPS(F_X, Y)] = \int_y \int_x [F_X(x) - 1_{\{x \ge y\}}]^2 dx \, dF_Y(y). \quad (5)$$

We discuss some computable solutions for this expectation.

### 3.3.3 | Computable solutions

Consider $Y \sim U[0, 1]$ and an assessment of $Y$'s distribution by an expert as that of random variable $X, X \sim U[0, H], 0 < H < 1$. The expert thinks values greater than $H$ are impossible, although these can in fact arise. The expected CRPS is computed based on the distribution of $Y$. The *CDF* of $X, F(x) = x/H$, for $x \in [0, H]$ and $F_X(x) = 1$,for $x \ge H$, along with the survivor function of $X, S(x) = 1 - F(x)$ are shown in Figure 5 (see also Candille & Talagrand, 2005).

Then $E_Y[CRPS(F_X, Y)] = \int_0^1 \int_0^1 [F(x) - 1_{x \ge y}]^2 dx \, dy$ is computed in two steps:

(A) For $y < H$:

$$\int_0^H \left[ \int_0^y \frac{x^2}{H^2} dx + \int_y^H \left( \frac{H - x}{H} \right)^2 dx \right] dy = \frac{H^2}{6}.$$

(B) For $y > H$:

$$\int_H^1 \left( \int_0^H \frac{x^2}{H^2} dx + \int_H^y dx + \int_y^1 0 \, dx \right) dy = \frac{H(1 - H)}{3} + \frac{(1 - H)^2}{2}.$$

Therefore:

$$E_Y[CRPS(F_X, Y)] = \frac{H^2}{6} + \frac{H(1 - H)}{3} + \frac{(1 - H)^2}{2}. \quad (6)$$

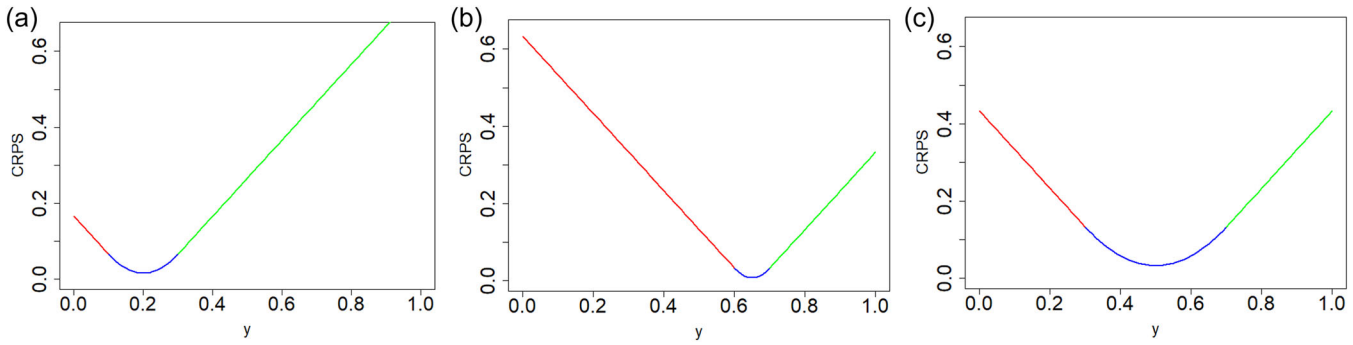


**FIGURE 4** Continuous Ranked Probability Score (CRPS) for $X \sim U[L, H]$ and $y \in [0, 1]$. Differences for $y < L$ (red), for $y \in [L, H]$ (blue), and $y > H$ (green) are highlighted. (a) $L = 0.1$, $H = 0.3$, (b) $L = 0.6$, $H = 0.7$, and (c) $L = 0.3$, $H = 0.7$.

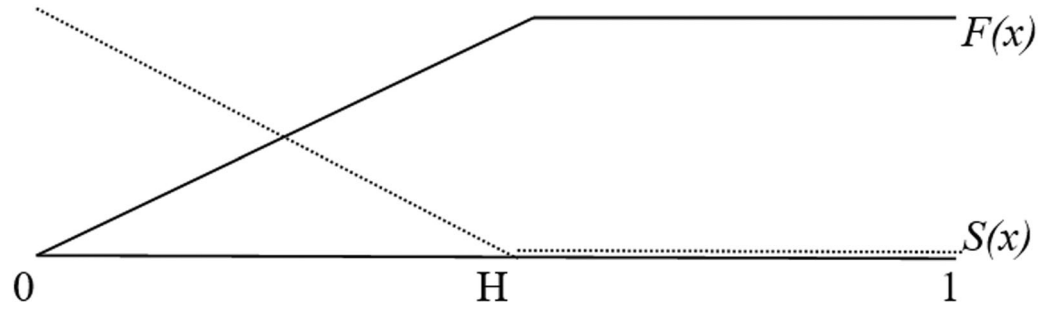


**FIGURE 5** Cumulative distribution function and survivor function of a uniformly distributed random variable on [0,H].

As noted by Hersbach (2000), these results acquire a physical dimension. The result for (A) is the score an expert with $X \sim U[0, H]$ expects, namely $H^2/6$, which has the physical dimension of $H^2$. If $X$ is in meters and changes to centimeters, the expected score increases by a factor $10^4$.

If $X \sim U[L, H]$, for $0 < L < H < 1$, then the same method of calculation applies mutatis mutandis. If $L = 1 - H$, with $H \geq 0.5$, then the contributions from $x < y$ and $y < x$ are equal and we need only to double the contribution from $x < y$. If $y < L$, the contribution from $x < y$ is zero. We compute

$$\begin{aligned}
&\int_L^H \int_L^y F(x)^2\, dx\, dy + \int_H^1 \int_L^y F(x)^2\, dx\, dy \\
&= \int_L^H \int_L^y \frac{(x-L)^2}{(H-L)^2}\, dx\, dy + \int_H^1 \left[ \int_L^H \frac{(x-L)^2}{(H-L)^2}\, dx + \int_H^y dx \right] dy, \\
&= \int_L^H \frac{y^3}{3(H-L)^2}\, dx + \int_H^1 \left[ \frac{(H-L)}{3} + (y-H) \right] dy, \\
&= \frac{(H-L)^2}{12} + \frac{(H-L)(1-H)}{3} + \frac{(1-H)^2}{2}.
\end{aligned}$$

Adding the identical contribution from $y < x$ gives:

$$E_Y[CRPS(F, Y)] = \frac{(H-L)^2}{6} + \frac{2(H-L)(1-H)}{3} + (1-H)^2. \quad (7)$$

Again, the score inherits a physical dimension from $X$. Substituting $L = 1 - H$ in the above equation, we find that $E_Y[CRPS(F, Y)]$ for $X \sim U[1-H, H]$ is equal to $E_Y[CRPS(F, Y)]$ for $X \sim U[0, H]$ from Equation (6). This holds for any $0 < L < H < 1$, with $L = 1 - H$. Hence, for $X \sim U[0, 0.7]$, $E_Y[CRPS(F, Y)] = 0.1966 = E_Y[CRPS(\tilde{F}, Y)]$, for $\tilde{X} \sim U[0.3, 0.7]$. By the same token, $X \sim U[0, 0.5]$ yields the same expected score of 0.25 as $\tilde{X}$ with Dirac distribution $\delta(0.5)$ concentrated at 0.5.

For CDFs $\{F_i\}$ and $F$, if $F_i \to F$, then $E_Y[CRPS(F_i, y)] \to E_Y[CRPS(F, y)]$, by the Helly–Bray theorem (Billingsley, 2013). It follows that, if $F_n \to U[0, 0.5]$ and $\tilde{F}_n \to \delta(0.5)$, then for all $\varepsilon > 0$ and for all sufficiently large $n$, $|E_Y[CPRS(F_n, Y)] - E_Y[CPRS(\tilde{F}_n, Y)]| < \varepsilon$. This illustrates how the *CRPS* compensates loss of SA by a gain in "sharpness," and again illustrates that equal scores do not entail equal quality.

Note that if the probabilistic forecast is the distribution of $Y$ (uniform [0, 1]), then the expected *CRPS* score is 1/6, by Equation (6). For $0 < H < 1$, $E_Y[CRPS(F, Y)] > \frac{1}{6}$, by Equation (6) and for $0.5 < H < 1$, $L = 1 - H$, $E_Y[CRPS(F, Y)] > \frac{1}{6}$, by Equation (7). An expert would receive a better (lower) expected score if their probabilistic forecast were equal to the distribution of $Y$.

### 3.3.4 | Scale invariant *CRPS* and a new test for SA

Scale invariance in Bayesian decision analysis was introduced by Morris (1974) and discussed in Morris (1977) and Wiper and French (1995). Similar to Morris (1974), we introduce a reparametrization of *CRPS* by transforming the realizations according to the probability integral transformation of an expert's assessed *CDF*. An expert with *CDF* $F_X$ for continuous variable $X$ is scored not with respect to the realization $y$ but with $F_X(y)$, the quantile of the distribution of $X$ realized by $y$. The proposed CRPS transformation has several advantages:

(i) The transformed *CRPS* becomes scale-invariant.
(ii) The expert's sampling distribution of transformed *CRPS* can be expressed in closed form.
(iii) The density of convolutions of transformed *CRPS* scores for independent variables is available in closed form.
(iv) Transformed *CRPS* can then be used to test the expert's SA without recourse to an asymptotic distribution.

On the downside, *CRPS* is insensitive to location bias (see below). If the experts assess only certain quantiles, a second downside is that continuous *CDFs* must be interpolated before applying *CRPS*.

To motivate this transformation, suppose we would like to test the hypothesis that $Y$ follows the expert's assessed distribution $F_X$. Applying the probability integral transformation, let $U = F_X(X)$ and define $V = F_X(Y)$. Then $F_U(u) = u$. The hypothesis that $F_X = F_Y$ is equivalent to the hypothesis

$$H_0 : V \sim U[0, 1].$$

In this case *CRPS* can be written, for the realization $v$ and for $U$ uniformly distributed on [0, 1], as

$$\begin{aligned}
CRPS(F_U, v) &= \int_{-\infty}^{\infty} [u - 1_{\{u \geq v\}}]^2 du, \\
&= \int_{-\infty}^{0} (0-0)^2 du + \int_0^v u^2 du + \int_v^1 (u-1)^2 du \\
&\quad + \int_1^{\infty} (1-1)^2 du, \\
&= \frac{v^3}{3} - \frac{(v-1)^3}{3}.
\end{aligned} \quad (8)$$

The range of the $CRPS(F_U, v)$ is $\left[\frac{1}{12}, \frac{1}{3}\right]$, for $v \in [0, 1]$. The distribution of *CRPS* is the distribution of the random variable

$$\frac{1}{3}[V^3 - (V-1)^3] = \frac{1}{3} - V + V^2,$$

taking values in $[\frac{1}{12}, \frac{1}{3}]$ (lower values are better). Under the null hypothesis, $V$ is uniform [0, 1]. For fixed $Q \in [\frac{1}{12}, \frac{1}{3}]$, to find the probability that $CRPS(F_U, v) \leq Q$, under the null hypothesis, we find the roots of $V^2 - V + (\frac{1}{3} - Q) = 0$:

$$v_{1,2} = \frac{1 \pm \sqrt{1 - 4(\frac{1}{3} - Q)}}{2} = \frac{1 \pm \sqrt{4Q - \frac{1}{3}}}{2}.$$

Collecting the mass between the two roots, we obtain the CDF

$$P(CRPS(F_U, v) \leq x) = \sqrt{4x - \frac{1}{3}},$$

with density

$$\frac{2}{\sqrt{4x-\frac{1}{3}}}, x \in \left(\frac{1}{12}, \frac{1}{3}\right). \tag{9}$$

Figure 6 shows $CRPS(F_U, v)$, together with its *CDF* and density under the null hypothesis $H_0$.

From Figure 6, it is evident that the score $CRPS(F_U, v)$ is symmetric around the value $v = 0.5$. This is different from the behavior of $CRPS(F_U, v)$ exhibited in Figure 4a or 4b, and it illustrates a feature of the scale-invariant version of *CRPS*.

From Equation (9), we can easily compute

$$\begin{aligned} E_V[CRPS(F_U, V)] &= \frac{1}{6}, \\ E_V[CRPS^2(F_U, Vt)] &= \frac{1}{30}, \\ Var_V(CRPS(F_U, V)) &= 0.0055555. \end{aligned} \tag{10}$$

It is handier to consider the following transformation

$$Z(U, V) = 4\ CRPS(F_U, V) - \frac{1}{3}. \tag{11}$$

Then $Z$ has *CDF* and density

$$F_Z(x) = \sqrt{x}, \text{ and } \quad f_Z(x) = \frac{1}{2\sqrt{x}},\ x \in [0, 1].$$

Note that $f_Z$ is the density of $U^2$, where $U \sim U[0, 1]$. So far, only one unknown scalar quantity of interest, and expert's resulting *CRPS* score, have been considered.

Suppose an expert provides uncertainty assessments for $n$ random variables. The emerging question is how to aggregate the *CRPS* scores of each of the $n$ variables? The transformation Equation (11), and the observation that the density $f_Z$ is the density of a squared uniform random variable are again handy.

If we assume the $n$ variables to be independent, then we need to consider $Z_1, \ldots, Z_n$ independent variables, each with density $f_Z$. For these, we need to find the density of $Z^{(n)} = Z_1 + \cdots + Z_n$. Or, in terms of the squared uniform random variables, we need to find the density of $S_n = U_1^2 + \cdots + U_n^2$.

Weissman (2017) provides closed-form distributions for $S_n$, for $n = 3, 4, 5, 6, 8, 10, 12$ and their graphical representations. A connection is also made with a topic of geometrical probability, that is, finding the *CDF* of $S_n$, $P(S_n \le s)$, is equivalent to finding the volume of the intersection between the unit $n$-cube and the ball of radius $\sqrt{s}$, in $\mathbb{R}^n$, when both are centered at the origin. In his comment to Weissman (2017), Forrester (2018) observes that the more generic volume problem posed by Xu (1996), of finding the volume of the intersection of a cube and a ball in $n$-space has already been solved by Tibken and Constales (Rousseau & Ruehr, 1997). Weissman (2017) reports that Constales' solution to the volume problem involves a method based on Fourier series and implies that, for general $n$,

$$F_n(s) = P(S_n \le s) = \frac{1}{6} + \frac{s}{n} + \frac{1}{\pi}\mathbf{Im}\sum_{k=1}^{\infty} \left[\left(\frac{C(2\sqrt{k/n}) - iS(2\sqrt{k/n})}{2\sqrt{k/n}}\right)^n \frac{e^{2\pi i k s/n}}{k}\right], \tag{12}$$

where $S(x) = \int_0^x \sin(t^2)\, dt$ and $C(x) = \int_0^x \cos(t^2)\, dt$ denote the Fresnel integrals and **Im** is the imaginary part of a complex number.

Figure 7 graphically compares the above cumulative distribution for $n = 2$ (left) and $n = 10$ (right) with the empirical distribution function of the corresponding sum of squared uniform observations. 100 observations were sampled for both empirical distribution functions. The CDF was implemented in `R`, by making use of functions implementing the Fresnel integrals in the `pracma` package (Borchers & Borchers, 2022).

Consider $n$ observations of continuous variables assessed by an expert $e$. The following procedure calculates *SA* for the *CRPS* statistic

(1) For each realization $y_i = 1, \ldots, n$, compute $q_e^i = F_{i,e}(y_i)$, the quantile of $y_i$ in expert $e$'s *CDF* $F_{i,e}$.
(2) The *SA* hypothesis $H_0$ entails that these quantiles are independent samples from a uniform distribution. Under this hypothesis, $CRPS_i(e) = CRPS(F_U, q_e^i)$ can be computed from Equation (8).

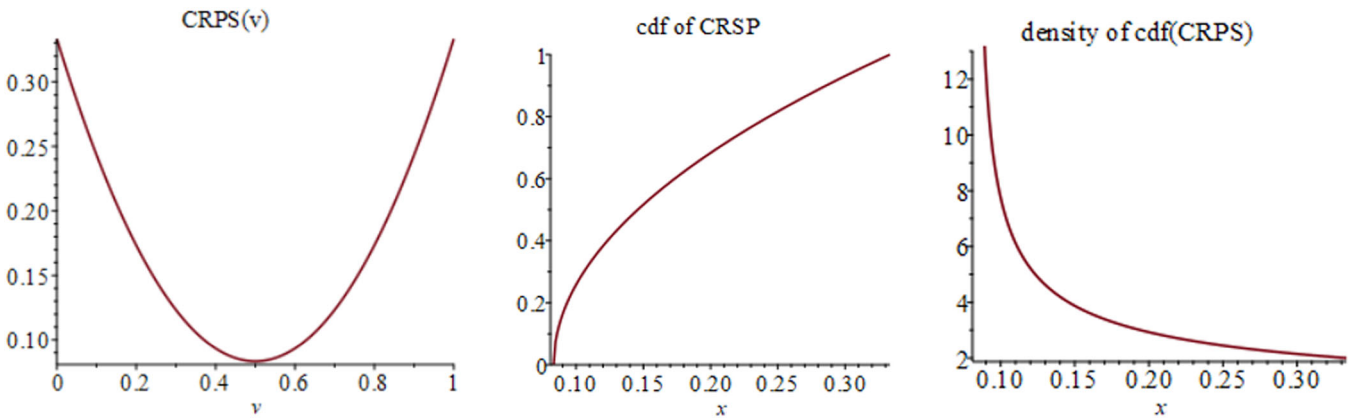


**FIGURE 6** Continuous Ranked Probability Score (CRPS) function, for $U \sim U[0, 1]$ and $v \in [0, 1]$, together with its cumulative and density functions, under the null hypothesis $H_0$.

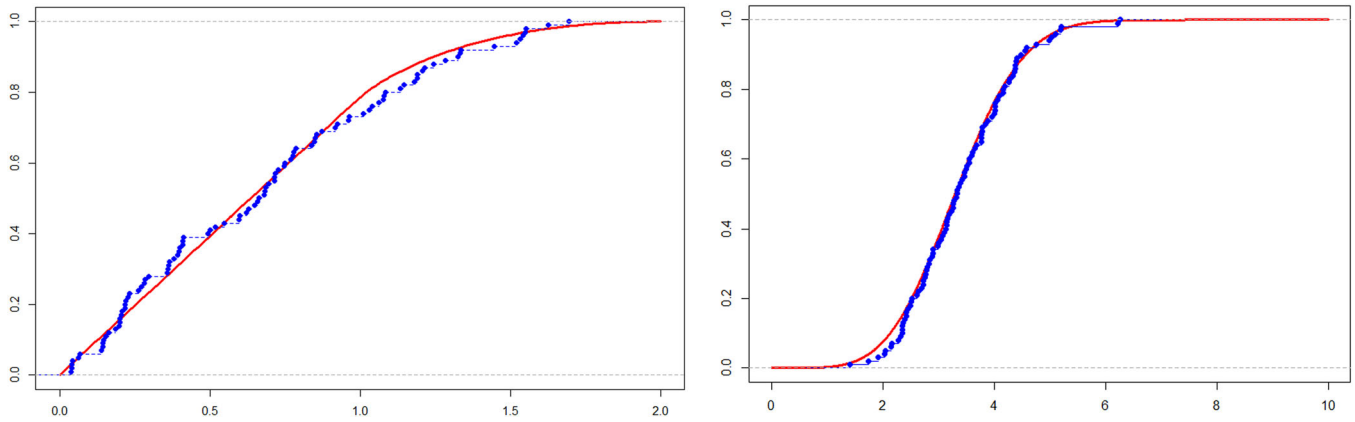


**FIGURE 7** The exact distribution (red) and empirical distribution function for 100 samples of the sum of squared simulated uniform observations (blue), for $n = 2$ (left) and $n = 10$ (right).

(3) For each $i = 1, \ldots, n$, compute $z_i(e) = 4CRPS_i(e) - 1/3$, which is given in Equation (11).
(4) Compute $s(e) = F_n(\sum_{i=1}^{n} z_i(e))$, where $F_n$ is the exact distribution of the sum of $n$ independent squared uniform variables, given in Equation (12).

Note that the procedure can be applied for continuous and invertible CDFs. *CRPS* uses an exact instead of an asymptotic distribution for the convolution of these *CDF*s. From Figure 6, it is evident that the score $CRPS(v)$ for value $v$ is symmetric around the value 0.5. The distribution of the sum of such variables is insensitive to location bias in the following sense: the score for $2n$ observations of 0.4 is the same as for $n$ observations of 0.4 and $n$ observations of 0.6. Figure 6 also shows that *CRPS* is insensitive to underconfidence: An expert whose probability transformed realizations are all 0.5 scores better than one for whom the hypothesis of Section 3.3.4 that $F_X = F_Y$ holds. Underconfidence is relatively rare with expert judgment. Overconfidence, on the other hand, is not rare (see Figure 2 ) and the *CRPS* score is sensitive to overconfidence.

## 4 | PERFORMANCE RESULTS

As our expert data is in the form of fixed quantiles of continuous distributions, we focus on performance metrics from *CM* and *CRPS*. The performance metrics of interest are *SA*, *Inf*, and mean absolute percentage error (*MAPE*), both for the experts themselves and for combinations of experts (called *DMs*). Before doing this, we first address the issue of persistence. This notion was introduced in Cooke et al. (2021), and we extend those results by examining the persistence of *MAPE*.

### 4.1 | Persistence of *MAPE*

*MAPE* is a scale-invariant measure of the proximity of a point forecast to the realization. In our case, the forecast is taken as the median. There are many measures for such proximity (Gneiting et al., 2007; Morley et al., 2018), each with benefits and drawbacks. *MAPE* is perhaps the most popular, defined for forecasts $x_i$ and realization $r_i$, for $i = 1 \ldots n$ as

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - r_i}{r_i}\right|.$$

This is evidently unstable for very small $r_i$. Instability arises on this data set, as the largest *MAPE* is over one million. *MAPE* is not used in computing DMs in the *CM*, as it does not comport with *CM*'s scoring rule properties. It is nonetheless an important performance metric. We first address the persistence of *MAPE*.

Consider a panel-wide performance metric in a panel of experts assessing variables from their field for which true values are known, for example, the maximum value of *SA* in the panel. Suppose that the experts are really equivalent and that observed differences in *SA* were simply due to "random stressors" during the elicitation. In contrast to persistent influences like knowledge, experience, and intuition, random stressors might be things like fatigue, mood, and distraction. It is not possible to observe or measure these influences. We can, however, test the claim that differences are just "noise." The null hypothesis is that the experts' responses for each variable are independently sampled from the same distribution (overdistributions). If the experts were re-elicited a short time later, then their responses would be independent resamples from this distribution. This is the operational meaning of the statement that expert differences are not persistent. That means that, for example, Expert 1's elicited distribution for variable 1 could just as well have come from Expert 2 and conversely. Nonpersistence is a statistical hypothesis that can be tested by randomly scrambling the original expert assessments. Thus, a new Expert 1 chooses assessments for each variable from the assessments of all experts for that variable. Expert 2 randomly chooses assessments not already chosen by Expert 1, and so on. We repeat this process, say 1000 times, to generate a distribution of the performance metric values in which any "expert effect" has been wiped out. If the maximum *SA* value were

not persistent, then the original value could be regarded as a random sample from the scrambled distribution. Equivalently, the percentile in the scrambled distribution realized by the original maximum *SA* would be uniformly distributed. With 49 studies, we would have 49 percentiles which would be uniformly distributed under the hypothesis that the performance metric were not persistent in this set of studies.

As described in Cooke et al. (2021) this hypothesis of uniformity can be tested either with a simple binomial test or with a normal test where the sum of the 49 percentiles is approximated as normal under the null hypothesis. For positively (negatively) sensed metrics high (low) percentiles are critical. The results from (Cooke et al., 2021) are extended to include the minimal *MAPE* value in each panel. Along with MAPE, Table 3 includes statistics regarding SA and combined score, which is the product of the SA and information score.

We see, from Table 3, that the persistence of the panel minimum *MAPE* is on a par with the spread of *SA* scores as reflected in their standard deviation. Roughly summarized, over the set of 49 studies, the ability to put one's median assessment close to the realization is a property of the experts. That said, the Pearson correlations of *MAPE* with *SA* and with *Inf* are negligible (−0.02 and 0.03, respectively). The rank correlations however are −0.25 and 0.23. The negative rank correlation −0.25 means that good SA is weakly correlated with low percentage error. Putting the median near the realization seems to be a different skill from giving statistically accurate and informative probabilistic forecasts.

## 4.2 | Results for SA with *CM* and *CRPS*

As mentioned, *SA* can be scored either with interquantile hit rates and the asymptotic $\chi^2$ *CDF*, or with *CRPS* using the exact distribution of the sums of *CRPS* scores. For the purposes of this section, we distinguish these two as *CM SA* and *CRPS SA*. When applied to this data, *CRPS* needs to interpolate the quantile of the realization. It was also noted that *CRPS* is insensitive to location and under confidence bias.

Figure 8 plots the SA scores for 526 experts based on *CM* and on *CRPS*, ordered by the *CRPS* scores. Although the drift of the two scores is similar, there is substantial scatter. *CRPS*'s log geometric mean *SA* score is −2.76 while that of *CM* is −3.47; in this sense, *CRPS* is more forgiving.

We define an expert's location bias as the absolute difference between the percent of realizations above the medians and 50%. Location bias of 50% means that all realizations are above or all realizations are below the medians.

Figure 9 black circles the *CM* scores of those experts for whom the location bias is greater or equal to 20%. *CM SA* scores of experts for whom the location bias is 50% (*all* realizations were either below or *all* above the medians) are red circled. The location bias of these circled experts is missed by *CRPS* and helps explain some of the downward scatter.

For 70 experts, the location bias is 0%. These are termed experts without location bias (though of course there could be location bias in the lowest and highest interquantile intervals). Figure 10 plots these 70 *CM SA* scores against all the *CRPS SA* scores. On this subset, *CRPS*'s log geometric mean *SA* score is −2.50, while that of *CM* is −2.68.

The number of calibration variables assessed in the 49 panels ranges from 7 to 21, see Figure 1. This influences *CM SA* in two ways, a larger number (i) increases the accuracy of the $\chi^2$ approximation and (ii) tends to lower SA scores of poorly calibrated experts as the test of SA has greater power. The first should decrease the differences between *CM SA* and *CRPS SA*, at least for location-unbiased experts, whereas the second enables a greater range of *CM SA* scores and might, therefore, increase the differences. A multiple regression of $\log\left(\frac{CRPS\ SA}{CM\ SA}\right)$ against location bias and number of calibration variables explains 30% of the variance (adjusted $R^2$) and both explanatory variables have a significant positive effect on the dependent variable. Thus, the influence of (ii) exceeds that of (i). The Pearson correlations of the dependent variable with location bias and with number of calibration variables are 0.41 and 0.29 respectively. The correlation of the two explanatory variables is −0.17. Under confidence

**TABLE 3** Persistence of performance metrics in 49 studies.

| | Mean SA | StDev SA | Max SA | Mean CS | Stdev CS | Max CS | Min CS | Min MAPE |
|---|---|---|---|---|---|---|---|---|
| Count(Orig,Med) | 42 | 38 | 36 | 39 | 40 | 38 | 40 | 38 |
| Binom | 1.81E−07 | 7.10E−05 | 7.01E−04 | 1.92E−05 | 4.63E−06 | 7.10E−05 | 4.63E−06 | 7.10E−05 |
| Sum percentiles | 36.33 | 34.27 | 31.29 | 33.84 | 32.99 | 31.92 | 36.40 | 34.1 |
| Normal | 2.40E−09 | 6.61E−07 | 3.91E−04 | 1.89E−06 | 1.33E−05 | 1.21E−04 | 1.95E−09 | 6.03E−07 |

*Note*: For each study, 1000 random scrambles of expert assessments are drawn. Statistics of the original panel and the scrambled panels are computed: the average statistical accuracy (Mean SA) for all experts, the standard deviation SA (StDev SA), the maximum SA (Max SA) among all experts, the average, standard deviation, maximum and minimum combined score (Mean CS, StDev CS, Max CS, Min CS) and minimum mean absolute percentage error (Min MAPE). The statistics from the original panel are compared with the empirical median of the corresponding statistics from the random scrambles. Count (Orig, Med) counts for how many studies the original statistic is higher than the corresponding median. For Min CS and Min MAPE, Count(Orig, Med) counts for how many studies the original statistic is lower than the corresponding median. Binomial denotes the probability under the hypothesis of nonpersistence that the number of percentiles above the median should equal or exceed the indicated number. Sum percentiles denote the sum of percentiles the original panel statistic realizes in the random scrambles. Normal is the probability under the nonpersistence hypothesis that the sum of percentiles should be at least as great as that indicated.

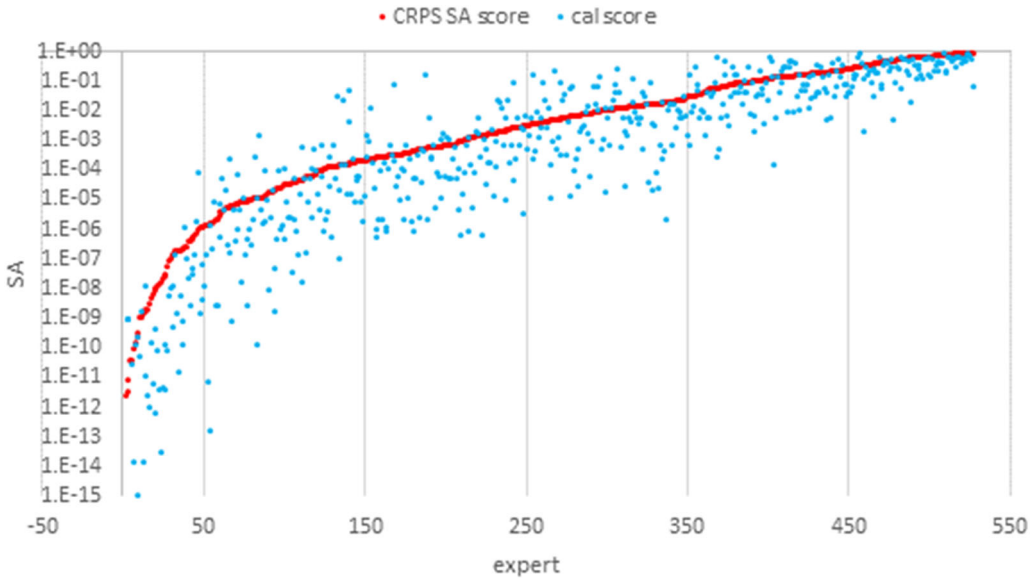


**FIGURE 8** Statistical accuracy of 526 experts with respect to Continuous Ranked Probability Score (CRPS) (red) and Classical Model (blue). Statistical accuracy (SA) scores are ordered by CRPS SA scores.

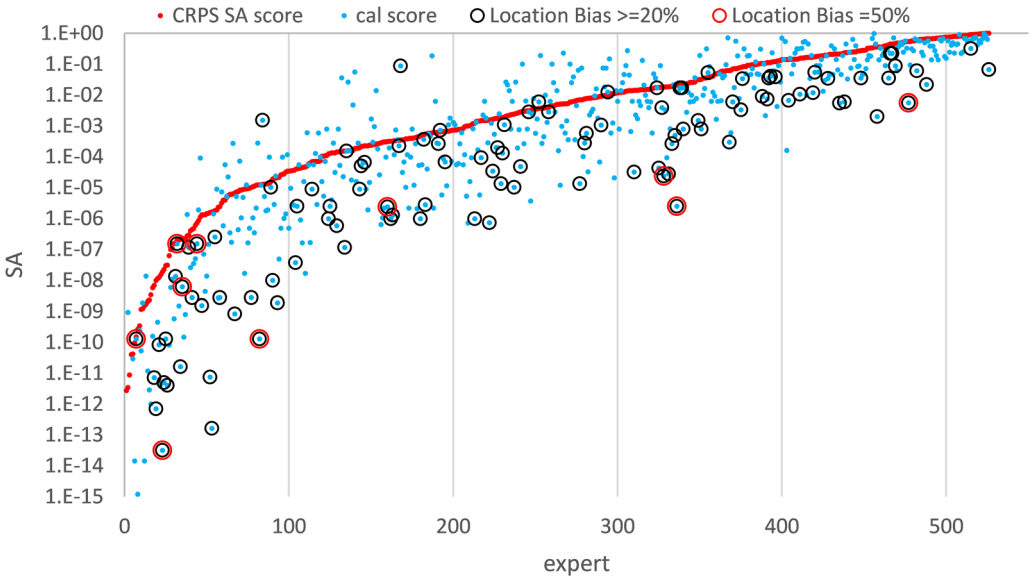


**FIGURE 9** Statistical accuracy of 526 experts with Classical Model (CM) and Continuous Ranked Probability Score (CRPS) with location biases circled. Location bias is the absolute difference between the percentage of medians above the realizations and 50%. Black circles denote *CM* scores of those experts for whom the location bias is greater or equal to 20%. Red circles denote location bias of 50%.

did not have a significant effect because of the low number of under confident experts. A more detailed analysis might better explain the differences in the two *SA* scores but at this point it appears that location bias is the major factor.

## 4.3 | MAPE

SA is not the only scoring variable of interest; proximity of the median to the realization is also important.

The data set contains many very high *MAPE* values. For visualization, we select 326 experts whose *MAPE*s were less than 2. Figure 11 plots these *MAPE* scores (left axis) and also plots the corresponding values of *CM SA* and *CRPS SA* (right axis). Although not overwhelmingly clear in the figure, the *CRPS SA* scores tend to be higher than those of *CM SA*, especially for very low *MAPE*s (see trend lines). The Spearman correlation of *CM SA* and *MAPE* on this data subset is −0.15, while that of *CRPS SA* and *MAPE* is −0.26. This results from the fact that *CRPS* uses the (interpolated) *CDF* whereas *CM* is based on interquantile hit-rates. It is reasonable to expect that

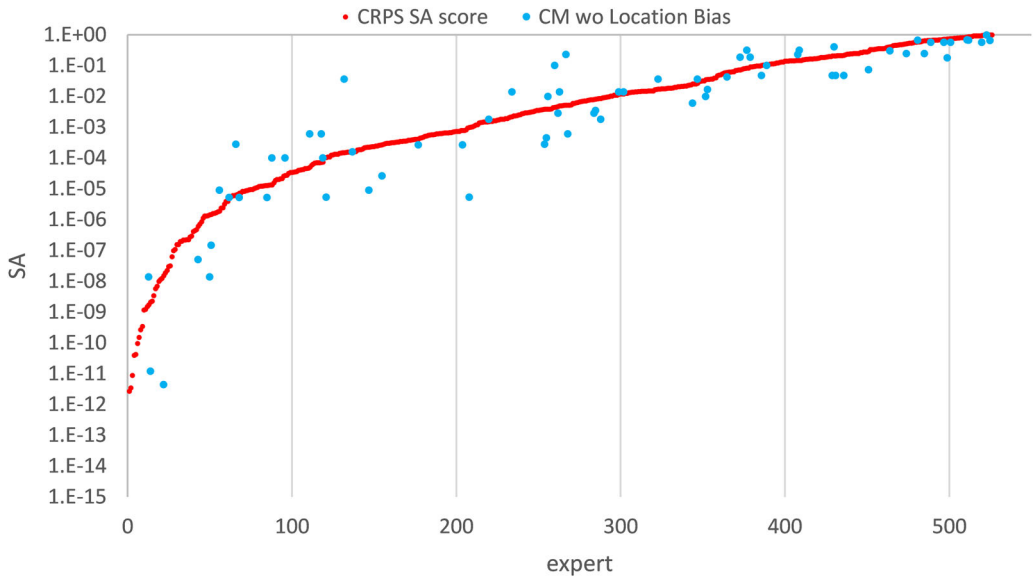


**FIGURE 10** Statistical accuracy (SA) of 526 experts with Continuous Ranked Probability Score (CRPS) and 70 experts with Classical Model (CM) without location bias.

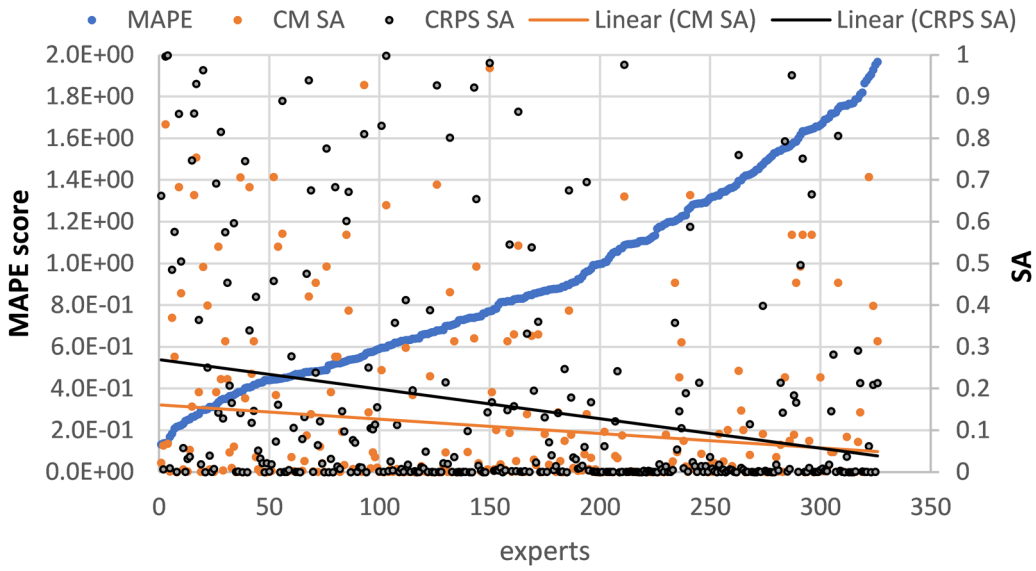


**FIGURE 11** 326 mean absolute percentage error (MAPE) scores <2 (left axis) and Statistical accuracy (SA) for Classical Model (CM) and Continuous Ranked Probability Score (CRPS) (right axis), with trend lines.

weighing experts according to *CRPS* scores will produce better *MAPE* values for the combination of experts than *CM*. Other researchers Flandoli et al. (2011) have used likelihood scores based on interpolated *CDF*s and achieved better *MAPE* performance than with *CM*. Although *CRPS* is a strictly proper scoring rule, its scale-invariant version is not proper.[2] The test statistic based on this score is unaffected by this. The true distribution of variable *Z* in Equation (11) is known, therefore, for large sample sizes, an alternative is to use the Glivenko Cantelli theorem on the convergence of the empirical to the theoretical distribution in analogy to the asymptotic propriety of *CM*. These are some possibilities to be explored in future work.

## 4.4 | DMs

We now turn to *DM*s based on *CM* and *CRPS*. For comparison, we introduce a new *DM* being the single expert in each panel with the smallest *MAPE*. *CM* comes in two flavors. The Global Weight *DM* (GWDM) uses Equation (1) where the *Inf* score is the average inf over all calibration variables. The Item Weight *DM* (IWDM) assigns a weight per variable based on each expert's inf for that variable. This is better in principle as it allows an expert to up- or downweight him/herself item-wise according to his/her self-assessed knowledge of each item. In practice, however, experts are not always able to do this

up-down weighting successfully. As mentioned, scoring rule theory does not determine the value of the cutoff $\alpha$ and in most applications this is done by choosing the value which optimizes the performance of the combination (GWDM_opt and IWDM_opt). As a foil for gauging the effect optimization, the *DM*s without optimization, setting $\alpha = 0$, are also computed, notated simply as GWDM and IWDM. These *DM*'s do not satisfy the proper scoring rule constraint. We note that the optimized *DM*s are optimized on the data used in Tables 4 and 5. True out-of-sample validation for these *DM*s is a complex undertaking out of scope for this study (Colson & Cooke, 2017). In most cases, there is a modest out-of-sample penalty for *SA*. This does not affect *MAPE* and does not affect the nonoptimized *DM*s. Additionally, we consider 'Equal Weights *DM* (EWDM), which assigns weights equally to all experts.

To this pantheon, we add two new *DM*s. CRPS uses weights determined by *CRPS SA*. No further information component is involved. Finally, MinMAPE denotes the *DM*'s obtained by giving weight one to the expert with the smallest *MAPE* value in each panel. Table 4 gives the *SA* scores for all these *DM*s. MinMAPE scores poorly for *SA*, as could be expected. The median *SA* score is 0.02 meaning that half of these *DM's* SA hypotheses would be rejected at the 2% level. EW and CRPS are roughly comparable and the optimized *CM DM*s return the best performance.

Table 5 gives the *Inf* scores for all these *DM*s. MinMAPE scores highest for *Inf*, neither surprising nor self-evident. Its median *Inf* score is almost twice the *CM* scores. Recall that *Inf* is a slow function. Halving the *Inf* corresponds roughy to doubling the width of the 90% confidence interval. EWDM is the least informative, not unexpected. CPRS is comparable to GWDM.

Table 6 gives the *MAPE* scores for all these *DM*s. MinMAPE scores the best (lowest) for *MAPE*, as expected, but the difference with the other *DM* is notable. Of course, this is purchased with poor *SA*. CRPS and IWDM_opt are roughly comparable; EWDM is wiping up the rear.

**TABLE 4** Statistical accuracy for decision makers.

| | EWDM | GWDM | $GWDM_{opt}$ | IWDM | $IWDM_{opt}$ | CRPS | Best MAPE expert |
|---|---|---|---|---|---|---|---|
| 5% | 0.04 | 0.02 | 0.02 | 0.01 | 0.01 | 0.01 | 0.00 |
| 50% | 0.29 | 0.39 | 0.55 | 0.49 | 0.64 | 0.34 | 0.02 |
| 95% | 0.65 | 0.66 | 0.93 | 0.83 | 0.96 | 0.65 | 0.70 |
| Mean | 0.31 | 0.37 | 0.50 | 0.43 | 0.54 | 0.35 | 0.15 |
| Geomean | 0.18 | 0.23 | 0.30 | 0.24 | 0.32 | 0.21 | 0.00 |

Abbreviation: MAPE, mean absolute percentage error.

**TABLE 5** Informativeness for decision makers.

| | EWDM | GWDM | $GWDM_{opt}$ | IWDM | $IWDM_{opt}$ | CRPS | Best MAPE expert |
|---|---|---|---|---|---|---|---|
| 5% | 0.16 | 0.23 | 0.41 | 0.35 | 0.38 | 0.22 | 0.60 |
| 50% | 0.49 | 0.71 | 1.00 | 0.94 | 1.09 | 0.75 | 1.85 |
| 95% | 1.29 | 1.90 | 2.70 | 1.98 | 2.70 | 1.89 | 3.32 |
| Mean | 0.60 | 0.86 | 1.26 | 1.03 | 1.31 | 0.87 | 1.96 |
| Geomean | 0.46 | 0.71 | 1.06 | 0.88 | 1.11 | 0.70 | 1.72 |

Abbreviation: MAPE, mean absolute percentage error.

**TABLE 6** MAPE scores for decision makers.

| | EWDM | GWDM | $GWDM_{opt}$ | IWDM | $IWDM_{opt}$ | CRPS | Best MAPE expert |
|---|---|---|---|---|---|---|---|
| 5% | 0.25 | 0.19 | 0.20 | 0.22 | 0.18 | 0.22 | 0.21 |
| 50% | 0.84 | 0.65 | 0.74 | 0.55 | 0.60 | 0.65 | 0.54 |
| 95% | 6.23 | 6.23 | 6.47 | 5.73 | 6.47 | 6.33 | 1.93 |
| Mean | 3.64 | 2.33 | 2.50 | 1.64 | 1.90 | 2.05 | 0.94 |
| Geomean | 0.95 | 0.90 | 0.95 | 0.73 | 0.84 | 0.82 | 0.58 |

Abbreviation: MAPE, mean absolute percentage error.

## 5 | CONCLUSION

Rewarding honesty in expert elicitation is not the same as rewarding quality in expert probabilistic assessments. This is manifested when numerically equal scores mask large differences in quality. In traditional proper scoring rules, SA and some measure of inf (sharpness, resolution, refinement, information) are hard-wired such that very high sharpness can buy off an attendant very poor SA. In the CM these are measured separately and combined in a product form with SA strongly dominating. A scale-invariant version of *CRPS* isolates the SA component and can be combined with inf as in the *CM*. This is applied to an expert judgment data base involving 49 studies, 526 experts, and 580 calibration variables from their fields. With a closed-form convolution of independent *CRPS* scores, the transformed *CRPS* yields a score for individual variables together with a test for experts' SA on sets of variables without recourse to an asymptotic distribution. This may enable applications with fewer calibration variables. Compared to the SA test used in the CM it has the advantage of better rewarding proximity of a median point forecast to the realization. On the other hand, it is insensitive to location and underconfidence bias. The feature of scoring individual variables might prove useful for screening calibration variables for outliers.

New insights include that (a) variance due to assessed variables dominates variance due to experts, (b) performance on MAPE is weakly related to SA, (c) scale-invariant CRPS combinations compete with the CM on SA and MAPE, and (d) CRPS is more forgiving with regard to SA than the CM as CRPS is insensitive to location bias.

Further analysis on combinations of experts' judgments, comparing the performance of *CRPS* with other tests based on the $\chi^2$, the Kolmogorov–Smirnov, and the Cramer–Von Mises statistics are the subject of a companion study in preparation (Rongen et al., 2024). At this point, we can conclude that the scale-invariant *CRPS* offers an alternative to *CM* with a different palette of pro's and con's. In any event, the ability to score individual variables, rather than sets of variables, secures it a place in the tool box.

## ACKNOWLEDGMENTS

We thank Johannes Bracher for in-depth discussions regarding the proper scoring property of the scale-invariant CRPS score. We thank Ernani Choma for pointing us to *CRPS* and Guus Rongen and Oswaldo Morales for exploring interpolation issues. We also thank the reviewers for their comments, which helped improve the presentation of results and readability of the article.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Roger Cooke's website at https://rogermcooke.net/.

## ORCID

*Gabriela F. Nane* http://orcid.org/0000-0002-3614-1820
*Roger M. Cooke* http://orcid.org/0000-0003-0643-1971

## ENDNOTES

[1] This is the original Brier formulation which is twice the formulation commonly found in rain/no rain discussions. The original formulation is more convenient for the calibration resolution decomposition.

[2] We thank Johannes Bracher for in-depth discussions regarding the proper scoring property of the scale-invariant CRPS score. It became apparent that the scale-invariant version is not proper.

## REFERENCES

Aitchison, J., & Dunsmore, I. (1968). Linear-loss interval estimation of location and scale parameters. *Biometrica*, *55*(1), 141–148.

Bamber, J. L., Oppenheimer, M., Kopp, R. E., Aspinall, W., & Cooke, R. M. (2019). Ice sheet contributions to future sea level rise from structured expert judgement. PNAS.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Borchers, H. W., & Borchers, M. H. W. (2022). Package 'pracma'. *Practical numerical math functions, version*. *2*(5).

Brown, T. A. (1974). *Admissible scoring systems for continuous distributions*. RAND Corporation.

Candille, G., & Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, *131*(609), 2131–2150.

Carvalho, A. (2016). An overview of applications of proper scoring rules. *Decision Analysis*, *13*(4), 223–242.

Colonna, K. J., Nane, G. F., Gabriela, F., Choma, E. F., Cooke, R. M., & Evans, J. S. (2022). A retrospective assessment of COVID-19 model performance in the USA. *Royal Society Open Science*, *9*, 9220021.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety*, *163*, 109–120.

Cooke, R. M. (1991). *Experts in uncertainty: Opinion and subjective probability in science*. Oxford University Press.

Cooke, R. M. (2014). Validating expert judgment with the classical model. In *Experts and consensus in social science* (pp. 191–212). Springer International Publishing.

Cooke, R. M., Marti, D., & Mazzuchi, T. (2021). Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *International Journal of Forecasting*, *37*(1), 378–387.

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., … Reich, N. G. (2022). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proceedings of the National Academy of Sciences USA*, *119*(15), e2113561119.

Dawid, A. P., & Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, *72*(2), 169–183.

De Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de lainstitut Henri Poincaré*, *7*, 1–68.

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety*, *96*(10), 1292–1310.

Forrester, P. J. (2018). Comment on "sum of squares of uniform random variables by i. Weissman". *Statistics & Probability Letters*, *142*, 118–122.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc. B Part 2 69*, 243–368.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

De Groot, M. H., & Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *32*(1–2), 12–22.

Hanea, A. M., & Nane, G. F. (2021). An in-depth perspective on the classical model. *Expert Judgement in Risk and Decision Analysis*, 225–256.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*, 559–570.

Merkle, E. C., & Steyvers, M. (2013). Choosing a strictly proper scoring rule. *Decision Analysis*, *10*(4), 292–304.

Morley, S. K., Brito, T. V., & Welling, D. T. (2018). Measures of model performance based on the log accuracy ratio. *Space Weather*, *16*(1), 69–88.

Morris, P. (1974). Decision analysis expert use. *Management Science*, *20*(9), 1233–1241.

Morris, P. (1977). Combining expert judgments. *Management Science*, *23*(7), 679–693.

Murphy, A. H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Monthly Weather Review*, *105*(7), 803–816.

Ray, E. L., Wattanachit, N., Niemi, J., Kanji, A. H., House, K., Cramer, E. Y., Bracher, J., Zheng, A., Yamana, T. K., Xiong, X., Woody, S., Wang, Y., Wang, L., Walraven, R. L., Tomar, V., Sherratt, K., Sheldon, D., Reiner, Jr, R. C., Prakash, B. A., … Reich, N. G. (2020). Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the us. *medRXiv*.

Rongen, G., Nane, G., Morales-Nápoles, O., & Cooke, R. (2024). *Continuous distributions and measures of statistical accuracy for structured expert judgment*. Manuscript submitted for publication.

Rousseau, C., & Ruehr, O. (1997). Problems and solutions. subsection: The volume of the intersection of a cube and a ball in n-space. Two solutions by bernd tibken and denis constales. *SIAM Review*, *39*(4), 779–786.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, *66*(336), 783–801.

Shuford, E. H., Albert, A., & EdwardMassengill, H. (1966). Admissible probability measurement procedures. *Psychometrika*, *31*(2), 125–145.

Weissman, I. (2017). Sum of squares of uniform random variables. *Statistics & Probability Letters*, *129*, 147–154.

Wiper, M. P., & French, S. (1995). Combining experts' opinions using a normal-wishart model. *Journal of Forecasting*, *14*, 25–34.

Xu, L. (1996). The volume of the intersection of a cube and a ball in n-space. *SIAM Review*, *38*(4), 669.

**How to cite this article:** Nane, G. F., Cooke, R. M. (2024). Scoring rules and performance, new analysis of expert judgment data. *Futures & Foresight Science*, e189. https://doi.org/10.1002/ffo2.189