Futures & Foresight Science

WILEY

ORIGINAL ARTICLE OPEN ACCESS

Continuous Distributions and Measures of Statistical Accuracy for Structured Expert Judgment

Guus Rongen[1,2] | Gabriela F. Nane[3] | Oswaldo Morales-Napoles[1] | Roger M. Cooke[4]

[1]Civil Engineering and Geosciences, Delft University of Technology, Delft, the Netherlands | [2]Pattle Delamore Partners Ltd, Ōtautahi Christchurch, Aotearoa New Zealand | [3]Delft Institute of Applied Mathematics, Delft University of Technology, Delft, the Netherlands | [4]Resources For the Future, Washington, Washington DC, USA

Correspondence: Guus Rongen (g.w.f.rongen@tudelft.nl)

Received: 12 October 2024 | **Revised:** 5 April 2025 | **Accepted:** 11 April 2025

Funding: This study was funded by the TKI project EMU-FD. This study project is funded by Rijkswaterstaat, Deltares and HKV consultants.

Keywords: classical model | expert judgment | metalog | scoring rule | statistical accuracy

ABSTRACT

This study evaluates five scoring rules, or measures of statistical accuracy, for assessing uncertainty estimates from expert judgment studies and model forecasts. These rules — the Continuously Ranked Probability Score (*CRPS*), Kolmogorov-Smirnov (*KS*), Cramer-von Mises (*CvM*), Anderson Darling (*AD*), and chi-square test — were applied to 6864 expert uncertainty estimates from 49 Classical Model (CM) studies. We compared their sensitivity to various biases and their ability to serve as performance-based weight for expert estimates. Additionally, the piecewise uniform and Metalog distribution were evaluated for their representation of expert estimates because four of the five rules require interpolating the experts' estimates. Simulating biased estimates reveals varying sensitivity of the considered test statistics to these biases. Expert weights derived using one measure of statistical accuracy were evaluated with other measures to assess their performance. The main conclusions are (1) *CRPS* overlooks important biases, while chi-square and *AD* behave similarly, as do *KS* and *CvM*. (2) All measures except *CRPS* agree that performance weighting is superior to equal weighting with respect to statistical accuracy. (3) Neither distributions can effectively predict the position of a removed quantile estimate. These insights show the behavior of different scoring rules for combining uncertainty estimates from expert or models, and extent the knowledge for best-practices.

1 | Introduction

Uncertainties are both widespread and influential in many fields, from climate modeling and economic forecasting to engineering design and legal decisions. The ability to accurately quantify uncertainties is important for informed decision-making, and it will often increase the value and usefulness to the outcomes. However, constraints such as limited data availability, problem complexity, or financial or even ethical restrictions can limit possibilities to accurately quantify these uncertainties. Expert judgment is a method to quantify uncertainty for variables whose uncertainty is difficult to quantify through other means. Expert judgment provides data in settings where a statistical or physics-based model would require assumptions or an extrapolation. It can take informal forms, such as asking an experienced person for their expectations. This might be fine for noncritical issues but high-stakes situations demand a more structured approach, one that is replicable, subject to review, and that could be assessed for potential biases, ensuring reliability and integrity.

The classical model (CM) is such an approach, which formalizes the process of expert judgment elicitation in such a way that the resulting uncertainty estimates can be treated as scientific data. It combines expert estimates using weights that are based on comparing uncertainty estimates to known outcomes

of a number of check (or calibration/seed) questions. Colonna et al. (2022) recently applied the method to combine COVID-forecasting models. They interpreted the different models as experts and their forecasts as estimates. The Classical Model was used to evaluate and combine these forecasting models, by comparing them to the actual course of events. This shows the value of CM outside the typical field of expert judgment.

First presented in (Roger M. Cooke 1991), the Classical Model has been widely applied and data from these applications have been made available to researchers, first in (Roger M. Cooke and Goossens 2008) and most recently in (Roger M. Cooke et al. 2021). The latter reference also gives a light exposition of the CM and introduces the expert judgment data used in this study. A special issue of Reliability Engineering and System Safety hosting the first publication of expert data (Roger M. Cooke and Goossens 2008) also contained contributions from many statisticians, risk analysts and practitioners who raised issues regarding CM[1]. Some of these issues, such as in-sample validation and overconfidence were amply addressed in the discussion papers of that special issue. Other concerns, most notably out-of-sample validation, persistence of performance and point value forecast performance spawned a stream of research. Much of this is summarized in (Roger M. Cooke et al. 2021) and in the references therein.

The main characteristics of the CM are:

- The experts estimate uncertainty by assigning values to the 5th, (25th), 50th, (75th), and 95th percentiles of their subjective probability distribution for each variable.
- The expert estimates are tested with calibration (aka seed) questions from the experts' fields. Their performance is measured in two dimensions. *Statistical Accuracy* and *Informativeness*.
- *Statistical accuracy* (SA) compares the expert's distributions with the actual values of the calibration questions. It is measured assuming only that the realizations are independent. From this it follows that the familiar chi-square goodness of fit statistic based on the interquantile realizations is asymptotically chi-square under the hypothesis that the expert is statistically accurate.
- *Informativeness* measures the degree to which a distribution is concentrated. It is measured by fitting a continuous CDF to an expert's elicited quantiles that minimizes the Shannon relative information of this fitted distribution with respect to an analyst selected background measure while complying with the expert's quantile assessments. When the background measure is uniform, the fitted CDF is piecewise uniform (for more information see (Roger M. Cooke et al. 2021). In this study, we assume the background measure is uniform. Combining information scores across variables again invokes independence assumptions.

In the Classical Model, the product of statistical accuracy and informativeness gives the (dimensionless) combined score, which after normalization gives the experts' weights. These weights are used to combine experts' distributions resulting in a so-called decision maker (DM). A subsequent option is to include only experts with an SA above a cutoff level (typically 5%), or to optimize the decision maker. Optimization involves consecutively excluding the expert with the lowest statistical accuracy, until the (renormalized) weighted estimates of the remaining experts form the decision maker with the highest combined score. Optimization often reduces the number of weighted experts. Cross validation research has shown that performance weighting increases the decision makers' out-of-sample informativeness without sacrificing statistical accuracy (Colson and Cooke 2017). Recently, Roger M. Cooke et al. (2021) demonstrated that randomly interchanging expert assessments within a study does indeed significantly affect expert performance (rejecting the so-called Random Expert Hypothesis). This highlights the value of performance weighting.

By design, SA dominates the CM performance-based expert weights. The informativeness score then modulates between experts with similar SA scores. For this reason, the CM measure of SA is 'assumption lean' and an interpolated continuous CDF is used only for the measure of informativeness. Researchers have drawn attention to two features of this design choice: (1) the measure of statistical accuracy depends only on the interquantile intervals in which realizations fall, and not on the relative position within these intervals, and (2) the SA measurement depends on a chi-square (χ^2) approximation which for the typical number of calibration variables and elicited quantiles is not very accurate (Roger M. Cooke 2014, see also Figure 7).

An important aspect of a structured expert judgment exercise is the understanding of the sensitivity of its results to the number of experts and questions. The effort required to elicit information from experts means that we are never fully certain of each expert's statistical accuracy, underscoring the need for measures of statistical accuracy that utilize the information provided by the experts in the best way. In other words, we need to determine expert weights accurately based on a limited data set, such that they reflect an expert's relative weight within a panel. To this end, this study aims to explore three main questions: 1) How do different goodness of fit tests, each with known asymptotic or exact distributions, compare in evaluating expert estimates? 2) How do two approaches of interpolating a continuous CDF compare in representing expert estimates? And, finally 3) what do the findings of 1 and 2 mean for the best-practice of eliciting structured expert judgments?

Five goodness of fit tests are considered. In addition to the standard χ^2 test in the CM, we consider three test-statistics that are commonly used to compare samples to continuous distributions, the Kolmogorov-Smirnov (*KS*), Anderson-Darling (*AD*), and Cramer-von Mises (*CvM*) test. Finally, we consider the Continuous Ranked Probability Score (*CRPS*). Recent work by Nane and Cooke (2024) uses the *CRPS* in CM. They present a *CRPS*-based score that assigns a scalar value to each assessment *cum* realization. Under suitable transformation, these scores for individual variables can be summed such that the exact distribution of the sum is available in closed form. This yields a measure of SA which appeals to an interpolated CDF but not to an asymptotic distribution.

All five tests except χ^2 compare quantiles to continuous CDFs and therefore require a distribution for transforming realizations to quantiles, using the expert estimates. For this we use two classes of distributions. The first is the piecewise uniform (PWU) distribution corresponding to the minimum information assumption in the Classical Model. The second is the Metalogistic, or Metalog, distribution (Keelin 2016). This recently introduced distribution offers great shape-flexibility, which helps with fitting a probability distribution to the large variety of expert quantile-estimates. Low parameter probability distributions often yield poor fits in these cases.

The five measures of statistical accuracy and two classes of distribution are compared in a variety of analyses, based on two different data sets. All analyses were done using Anduryl, an open-source Python-module and graphical user interface (Guus Rongen et al. 2020). Metalog calculations were conducted using (Adamczewski 2023). We used 49 expert judgment studies from the past decades, described in (Roger M. Cooke et al. 2021), comprising 530 experts and 580 calibration variables. Recently published structured expert judgment studies, such as (G. Rongen et al. 2022) and (Ren et al. 2024), were not considered because they have not yet been described and compared in an overview study. Additional to the published studies, we simulate expert estimates from distributions with a specific bias for a more clinical comparison. The analyses show 1) the statistical accuracy results from each score, 2) the ability of each measure of statistical accuracy to detect different biases, and 3) how the weights from each measure of statistical accuracy perform when used to create a DM that is evaluated with another measure. Results are presented for the PWU distribution and the Metalog distribution. In a final analysis, we consider the case-studies with 5 percentile estimates, removing 2 of these 5 percentiles, and see how well both distributions are able to estimate the position of the missing percentile.

2 | Methods

2.1 | Measures of Statistical Accuracy

To determine the statistical accuracy of an expert or forecast based on a set of seed questions, the questions' realizations or observations y are compared to the estimated distributions $F(x)$, resulting in a set of quantiles. We test the hypothesis whether these quantiles $F(\mathrm{x})$ are drawn from a uniform distribution $U[0-1]$. If so, the realizations appear to be drawn from the expert distributions or forecasts, which can consequently be considered accurate. The manner in which each of the above introduced test-statistics does this is explained below.

2.1.1 | Kolmogorov-Smirnov

The Kolmogorov-Smirnov (KS) test compares two samples (two-sided test) or a sample with a distribution (one-sided test) by using the supremum distance (equation 1) between (empirical) cumulative distribution functions (Kolmogorov 1933; Smirnoff 1939)

$$D_n = \sup_x |F_n(x) - F(x)| \tag{1}$$

In the context of the CM, a perfectly statistically accurate expert is one for whom the quantiles of the realizations for the calibration questions are uniformly distributed. An expert's statistical accuracy is thus tested by comparing these quantiles to a uniform distribution using the one-sided KS-test. The arrow in Figure 1 illustrates the KS test-statistic. In the KS test, the largest difference tends to be found near the median. Consequently, the statistic is relatively insensitive to deviations in the tail, which, when applied to expert judgments, typically corresponds with overconfidence.

For hypothesis testing, the KS distance is used to investigate the probability that the sample comes from the tested distribution. For this, an exact distribution is approximated using the method proposed by Simard and L'Ecuyer (2011). In classical statistics, a probability lower than 0.05 (i.e., the significance level) leads to rejecting the hypothesis that the data is independently sampled from the distribution of interest.

2.1.2 | Cramer-von Mises and Anderson-Darling

The Cramer-von Mises (CvM) statistic is the area between the empirical CDF and target CDF (Cramér 1928; Von Mises 1928), illustrated by the hatched area in Figure 1. In contrast to the KS-test, CvM considers the full distribution rather than the distance at a single point.

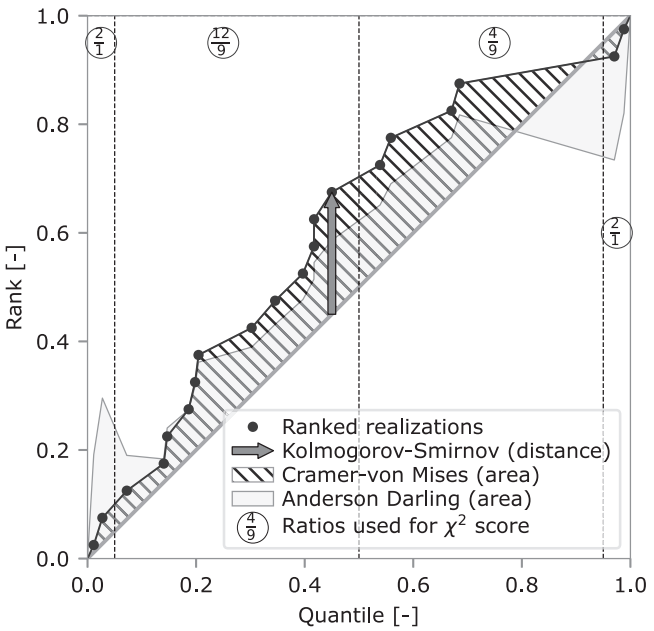

FIGURE 1 | Illustration of KS, CvM, and AD test statistics for a sample from a uniform distribution. The sample is plotted by their ranks (the connected dots). The arrow indicates the Kolmogorov-Smirnov (KS) statistic, the hatched area Cramer-von Mises (CvM), and the filled area (a weighted version of the hatched area) the Anderson-Darling (AD) statistic.

Just like *KS*, the *CvM* test-statistics is relatively insensitive to deviations in the tail. The Anderson Darling (*AD*) statistic, based on *CvM*, compensates this by adding more weight to the tails of the distribution (Anderson and Darling 1952). The equation for both statistics is:

$$n\int_{-\infty}^{\infty}(F_n(x)-F(x))^2 w(x)dF(x), \tag{2}$$

where n is the sample size, $F(x)$ the hypothesized distribution (uniform, in this study), and $F_n(x)$ is the empirical cumulative distribution function (the expert's percentile points under the assumed probability distribution). The weight $w(x)$ differs for *CvM* and *AD*. In *CvM*, all realizations x have weight 1.0. For *AD*, more weight is assigned to both tails of the distribution:

$$w(x)=[F(x)(1-F(x))]^{-1}. \tag{3}$$

By assigning a large weight to the deviation of quantile points in the tail, *AD* compensates *CvM*'s insensitivity to overconfidence. This is shown by the filled area in Figure 1, which, compared to the hatched area, has a larger distance to the diagonal at the edges. Distributions from (Csörgő and Faraway 1996) are used to convert the *CvM* test statistic to a *p*-value. An approximation of the distribution for the *AD* test statistic, for a uniform distribution, is given by (Marsaglia and Marsaglia 2004) and (Grace and Wood 2012). Marsaglia and Marsaglia (2004) cover the full range $a\in(0,\infty)$ (with a being the *AD*-statistic), while the approximation from (Grace and Wood 2012) is specified only for $a\in[3,\infty)$. The latter is more accurate for high values of a, which is why we apply (Marsaglia and Marsaglia 2004) for $a\in(0,3)$, (Grace and Wood 2012) for $a\in(4,\infty)$, and linearly interpolate between the two for $a\in[3,4]$ to ensure a smooth transition.

Another test-statistic that is often used for normality testing is Shapiro-Wilk (Shapiro and Wilk 1965). This statistic tests whether a sample is normally distributed with any mean and variance. It does not test whether a sample is standard normal distributed (i.e., $N(0,1)$), so neither can it be used to test whether a sample is uniformly distributed between 0 and 1. Therefore, it was not used in this study.

2.1.3 | CRPS

The Continuously Ranked Probability score (*CRPS*) is a measure for comparing forecast or estimates to realizations (Brown 1974). For a given (expert's) distribution F for a random variable X and realization y, the CRPS is defined by

$$CRPS(F,y)=\int_{-\infty}^{\infty}[F(x)-1_{\{x\geq y\}}]^2\mathrm{d}x. \tag{4}$$

In which 1 is the Heaviside step function. *CRPS* thus compares a probabilistic forecast to a scalar realization, integrating the squared difference between $F(x)$ and a step function at the realization. This is illustrated in Figure 2 a. The *CRPS* tests the agreement between y and $F(x)$ for a single variable on the variable's scale. To be used in the classical model, the *CRPS* needs to be derived for multiple realizations y to obtain a sufficiently high degree of confidence in the statistical accuracy. To compare different variables on the same scale instead of the quantities' scales, a scale-invariant *CRPS* is needed. This allows combining the different estimates.

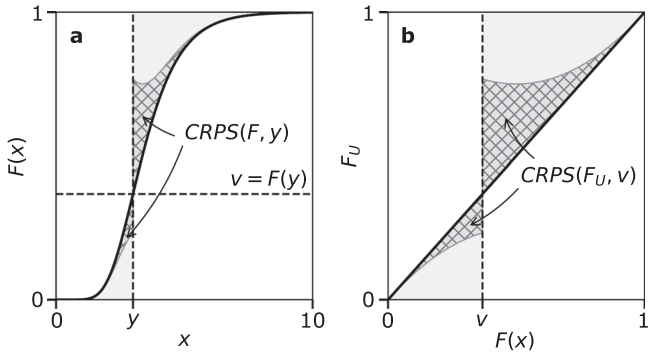

FIGURE 2 | Illustration of regular CRPS (a) and scale invariant CRPS (b). The CRPS is calculated by summing the squared difference between the CDF and step function, which is illustrated with the crossed area.

Nane and Cooke (2024) present a method for this. To assess an expert's statistical accuracy, they test the hypothesis that $F(x)\sim U[0,1]$. For this, realization y is transformed to its quantile v, using expert estimate $F(x)$ for variable x. With this transformation, the score is no longer a function of $F(x)$ and y, but of the standard uniform distribution F_U and v. This transformation is displayed in Figure 2 b. Through this, the transformed *CRPS* score becomes scale invariant. Furthermore, if we assume n independent variables, then the distribution of the convoluted transformed *CRPS* score follows an exact rather than an asymptotic distribution. The details of computing the transformed *CRPS* score are found in (Nane and Cooke 2024). Throughout the manuscript, we refer to this transformed *CRPS* instead of the original *CRPS* score presented by (Brown 1974).

Note that while both ***CRPS*** and ***CvM*** integrate the difference between an observed and estimated CDF, CRPS does this per item individually and evaluates the resulting metric against a distribution (a closed form convolution of the scores) based on ***N*** items. ***CvM***, on the other hand, combines the quantiles of all observations into a single empirical CDF, and compare this to a uniform distribution.

2.1.4 | Classical Model – Chi-Square

The χ^2-based test statistic is the standard statistic used in the Classical Model. The statistic evaluates observations based on *which* quantile-interval they are in, and not *where* in this interval. This in contrary to the measures presented above, which use the quantile-position and therefore require an assumed distribution.

χ^2's statistical accuracy is calculated as follows: If k quantiles are assessed, with n the number of calibration variables assessed by an expert and n_i the number of realizations falling

in the i-th interquantile interval, then $\mathrm{s} = (s_1, ..., s_{k+1})$, where $s_i = \frac{n_i}{n}$ is the sample distribution for the expert. The vector $\mathrm{p} = (p_1, ..., p_{k+1})$ is the expected relative frequency of interquantile realizations, thus if the 5%, 50%, 95% quantiles are elicited, then $p = (0.05, 0.45, 0.45, 0.05)$. Under the hypothesis that the realizations are independently drawn from p, the quantity $2nI(\mathrm{s}|\mathrm{p})$ is asymptotically chi-square (χ^2) distributed with k degrees of freedom, where $I(\mathrm{s}|\mathrm{p})$ is the Shannon relative information of s with respect to p (Roger M. Cooke 1991). Using this distribution, the p-value, or probability that an expert is statistically accurate, can be calculated through the ratio s to p. This is illustrated in Figure 1 with the circled fractions. For example, the first bin contains two observations where one (out of twenty) is to be expected. This gives a fraction of $\frac{s_1}{p_1} = \frac{2}{1} = \frac{0.10}{0.05}$. The p-value resulting from the test is used as statistical accuracy. Although this p-value is calculated using the χ^2-distribution, the scoring rule used in the Classical Model is different from the commonly known χ^2-test.

2.2 | Metalog Distribution

The Metalogistic, or Metalog, distribution is a continuous univariate probability distribution with high shape flexibility introduced by Keelin (2016). It accommodates bounded, semi-bounded, and unbounded distributions. This makes it an appealing choice for fitting empirical data (e.g., as a continuous replacement for a histogram), but also for modeling expert estimates. The Metalog is a generalized form of the logistic distribution, achieved by substituting the mean and standard deviation in the quantile function of the logistic distribution using series expansion. In this study, the three and five term functions $M_3(y)$ and $M_5(y)$ are used:

$$\begin{aligned} M_3(y) &= a_1 + a_2 \ln\frac{y}{1-y} + a_3(y-0.5)\ln\frac{y}{1-y} \\ M_5(y) &= M_3(y) + a_4(y-0.5) + a_5(y-0.5)^2 \end{aligned} \quad (5)$$

Here, y denotes the cumulative probability and a_i the constants.

Because the distribution is defined using its quantile function, a unique vector of n size a can be fitted to any set of n percentiles. This is a useful property for resembling experts' quantile estimates without changing their estimates. However, the quantile function $M_n(y)$ should be strictly increasing for all $y \in (0, 1)$. This is not necessarily the case for all sets of n percentiles, resulting in invalid or infeasible distributions with negative probability density.

Figure 3 shows eight examples of Metalog distributions (smooth grey curves) and piecewise uniform (PWU) distributions (stepped black curves) fitted to either three-percentile estimates (a, b, c) or five-percentile estimates (d, e, f, g, h). For three-percentile estimates, an infeasible a-vector can be resolved by imposing a lower or upper bound. This introduces a fourth parameter, making the solution overdetermined. We address this by selecting the bound such that it minimizes the maximum probability density, resulting in the least informative distribution.

For five-percentile estimates, many expert estimates (combinations of five quantiles) lead to infeasible distributions. To be able to process the results for case studies with five quantiles as well, the infeasible estimates are split in two by the median, resulting in two three-quantile estimates (0.05, 0.25, 0.50, and 0.50, 0.75, 0.95). This gives a step in density at the median, as shown by the solid line in Figure 3h. Optionally, this can be resolved by imposing a bound on the distribution with the lowest density at the median such as shown by the dashed line. However, this is primarily an aesthetic solution, which is why we chose not to do this. Note that the Metalog distribution can also fit the quantile estimates with a feasible (non-negative) distribution, but this requires adding more terms to the a-vector than there are quantiles. Further details on the fitting procedure can be found in Appendix B.

All tests except χ^2 use quantile points of realizations for evaluating statistical accuracy, rather than the quantile intervals in which the realizations fall. A fitted Metalog distribution provides these quantile points based on the expert estimates. In Section 3.4 we examine whether the realization quantiles from the Metalog are a better representation of expert estimates than the realization quantiles generated by a PWU distribution. Applying the Metalog also changes the calculation of informativeness in the CM since the default approach is based on the piecewise uniform assumption. For further explanation on how this calculation is performed for the Metalog, please refer to Appendix B.

2.3 | Comparing Measures of Statistical Accuracy

The five test statistics detailed in Section 2.1 evaluate statistical accuracy in different ways, leading to different test scores. The sensitivity to biases is explored using the method presented in Section 2.3.1. Section 2.3.2 explains how the statistical accuracy from the different tests are compared given their difference in values. Finally, the method for comparing the quantile estimates from PWU and Metalog is outlined in Section 2.3.3.

2.3.1 | Scores' Sensitivity to Detect Biases

To assess the ability of the measures of statistical accuracy to detect biases in experts' individual assessments, we introduce criteria for location bias and underconfidence or overconfidence. Location bias is defined as the absolute difference between the fraction of realizations below the median estimate and 0.5, or

$$\left| \frac{\sum_{i=1}^{n}\left(x_i < F_{e,i}^{-1}(0.5)\right)}{n} - 0.5 \right|, \quad (6)$$

with n being the number of items, x_i the realization for item i and $F_{e,i}^{-1}(0.5)$ expert e's median estimate for item i.

Overconfidence and underconfidence are quantified by the number of realizations below and above the lowest and highest estimated quantile, divided by the expected number. Let LQ and UQ be the lower and upper quantile (typically 0.05 and 0.95, but

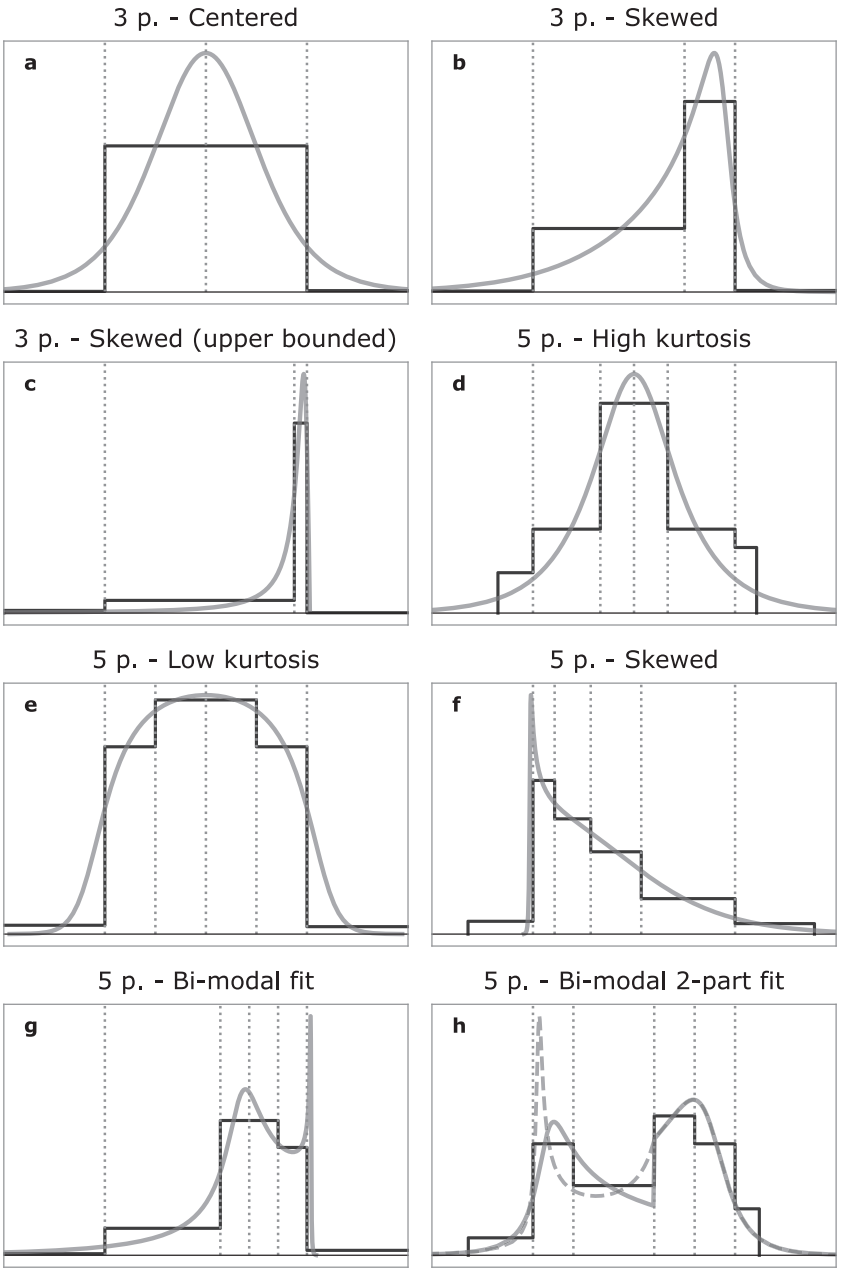

FIGURE 3 | 8 examples of Metalog distributions and piecewise uniform distributions fitted to 3-percentile (a–c) and 5-percentile (d–h) expert estimates. The estimates are indicated with the vertical dashed lines.

0.10 and 0.90 in two of the 49 cases). The ratio of tail realizations is calculated with

$$\frac{\sum_{i=1}^{n}(x_i < F_{e,i}(LQ)) + \sum_{i=1}^{n}(x_i > F_{e,i}(UQ))}{n(LQ + (1 - UQ))}. \quad (7)$$

A value greater than 1.0 indicates overconfidence, a value less than 1.0 indicates underconfidence.

2.3.2 | Comparing Measures of Statistical Accuracy Through Decision Makers

As discussed in Section 2.1.4, the Classical Model (CM) relies on the χ^2 statistical accuracy in combination with the information score to assign weights to each expert. These weights are then used to aggregate experts' distributions into a decision maker (DM) using the global weights algorithm with optimization (GLopt) or without (GL). Whenever comparing global weights in the analyses, the values of the weights are the normalized product of statistical accuracy and informativeness. Within this, the measures of statistical accuracy introduced in Section 2.1 serve as alternatives for the statistical accuracy term. Consequently, different measures assign different weights to the experts. Moreover, applying optimization can lead to DMs composed of the estimates of different sets of experts. In addition to the global weights, we consider the equal weight DM (EQ) which assigns the same weight to every expert. We did not consider using item weight, which involves varying weights per item based on the expert's informativeness for that item.

We are interested in comparing the effects of applying different measures of SA within the CM on the decision maker's SA. For this, we cannot simply compare the SA of the DM calculated using each measure, because some measures give on average higher scores than others. For example, *KS* and *CvM* are less sensitive to overconfidence, a prevailing bias in CM studies, and therefore give higher SA scores. This does not mean the experts are actually statistically more accurate. To compare the measures, we consider the weights from each measure, evaluated with each of the measures of SA. Both with and without optimization, and for both the PWU and Metalog distribution. As an example, we list the steps in comparing the *KS* and *CRPS* weights according to the χ^2 measure of SA:

1. First, experts' weights, as the normalized product of statistical accuracy and informativeness, are calculated based on *KS* and *CRPS* measures of SA.
2. Decision makers distributions are obtained for each set of those weights, which we will refer as DM_{KS} and DM_{CRPS}.
3. The χ^2 measure of statistical accuracy is then calculated for the DM_{KS} and DM_{CRPS}.
4. This is repeated for all 49 studies. Ranking these SAs gives a 2 sets of 49 ranked SA scores.
5. Using the Mann-Whitney rank sum test (Mann and Whitney 1947), we test whether the DM_{KS} and DM_{CRPS} ranks are statistically equivalent, or whether one is lower (or higher) than the other according to the χ^2 measure of statistical accuracy.

We do this for all combinations of SA measures, such that each pair of DMs is compared with respect to each of the five measures of SA.

2.3.3 | Determining the Metalog's and PWU's Ability to Predict Missing Quantiles

The choice of the Metalog distribution to represent the probability density between percentile estimates is rooted in the hypothesis that it better aligns with the distribution perceived by experts. This is due to its smooth curve without abrupt changes in probability density at estimated percentiles. To test this hypothesis, we remove the second and fourth percentile from the case-studies involving five elicited percentiles. The removed percentiles are then estimated using both the PWU and Metalog distributions. By comparing the difference between the estimated percentile point and the removed value (e.g., $F^{-1}(0.25) - x_{0.25}$ for the 25^{th} percentile) we can determine which distribution more accurately predicts the location of the removed percentiles.

3 | Results

We use expert data from 49 studies that have been previously used and explained (Roger M. Cooke et al. 2021). The data comprise 6864 individual expert assessments. The different measures of statistical accuracy (SA) are calculated for the global weights decision maker (DM) with and without optimization, and the equal weights DM. These results allow us to compare SA across different measures of SA (Section 3.1), assess their sensitivity to different biases (Section 3.2), cross compare their performance (Section 3.3), and evaluate the ability of the Metalog and PWU distributions to predict missing quantiles (Section 3.4).

3.1 | Individual Experts' Measures of Statistical Accuracy

The PWU and Metalog distribution's quantile estimates of the realization, for the 6864 individual experts, are shown in Figure 4. Overconfidence is signaled by the high number of realizations in the tails. The quantile positions for realizations outside the [0.05, 0.95] interval differ most between Metalog and piecewise uniform (PWU). The standard (PWU) approach requires an assumption on the probability density in the [0.0, 0.05] and [0.95, 1.0] range as the 0.0 and 1.0 quantiles are not elicited. These lower and upper bounds are usually placed at the minimum and maximum of *all* experts' estimates and the realization, extended by (typically) a 10% overshoot of this total range. Since only one expert gives the lowest or highest estimate, the estimates of the other experts end up being extended by (much) more than 10%. This leads to the position of a realization in the (often very wide) tails being relatively close to the elicited outer quantiles (e.g., 0.05 and 0.95). The Metalog distribution does not require an assumption on the tails, except for very skewed estimates. Therefore, the realizations are placed based on the fitted distribution and experts are judged on their own overconfidence. This tends to result in quantiles closer to 0.0 and 1.0.

Figure 5 shows the statistical accuracy (SA) for the five considered measures, for each of the 530 experts. Histograms of the SA for each individual measure are displayed on the diagonal. The first bin covers the first 5%, that is, the significance level commonly used in simple hypothesis testing. For an expert with SA less than 5%, the hypothesis that the expert is statistically accurate would be rejected at the 5% level. The dashed lines in the scatter plots also indicate this 5% significance level. The scatter plots in the lower left triangle are obtained under the PWU assumption, those in the upper right triangle under the Metalog assumption. Many scatters are overlapping in the < 0.05 corner, Figure A3 in Appendix A shows more clearly how the measures compare in that range.

The scores χ^2, *CRPS*, *KS*, *CvM* and *AD*, assign a significance level above $> 5\%$ to 27%, 32%, 58%, 62%, and 46% of the experts when assuming PWU, and 27%, 18%, 49%, 48%, and 17% for Metalog. For all but χ^2, assuming a Metalog distribution leads to lower SAs relative to a PWU distribution. This is because using PWU results in the realizations being at quantiles closer to the 5th and 95th (as illustrated in Figure 4). χ^2 relies on quantile intervals, such that the choice of the distribution interpolation these quantiles does not affect statistical accuracy.

The rank correlations between the different measures of SA are high. However, when considering only experts with SA greater than 0.05 for both measures, the correlation is generally low for

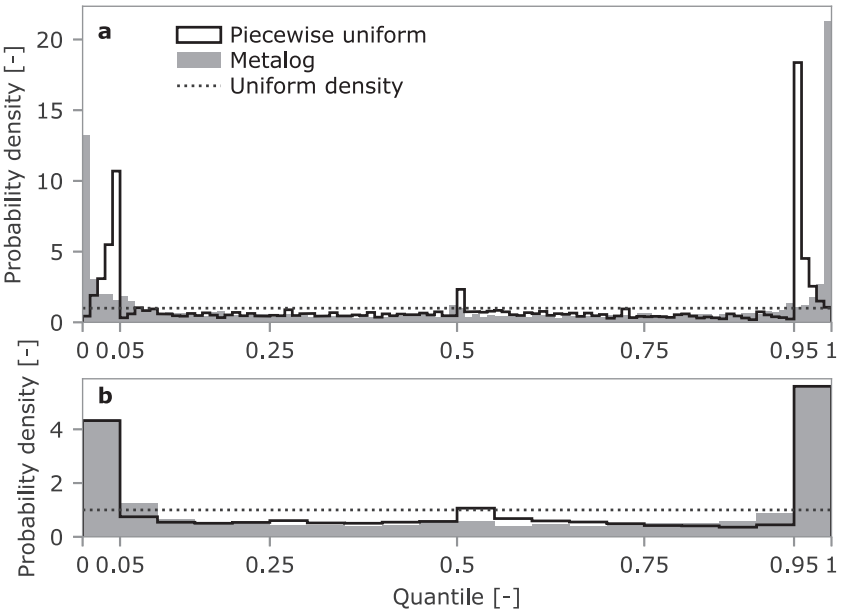

FIGURE 4 | Quantiles of realizations for 6864 individual expert quantile assessments with respect to the fitted piecewise uniform (white) and Metalog (grey) distributions. The difference between (a) with 100 bins and (b) with 20 bins shows the effect of assuming PWU or Metalog on the tail quantiles.

all combinations except between *KS*, *CvM*, and *AD*. Under the PWU assumption, *CvM* and *KS* are relatively similar to each other and to *AD*. When assuming the Metalog distribution, *AD* gives significantly lower scores than to *KS* and *CvM*. This is due to the extra weight assigned to realizations in the tail (see Figure 1). A high SA for *AD* does however still ensure high SA with *KS* and *CvM*.

Figure A3 in Appendix A shows the same results plotted on a logarithmic scale, demonstrating again that *KS*, *CvM*, and *AD* are relatively similar measures of statistical accuracy. *CRPS* and χ^2 also show some resemblance due to their high sensitivity to overconfidence. Moreover, *KS* and *CvM* are less likely to yield very low ($< 10^{-5}$) scores, while χ^2 tends to give the lowest scores. In terms of expert weights, the linear representation is more relevant, as it will generally hardly matter for the DM whether an expert gets a 10^{-3} or 10^{-10} score.

3.2 | Analysis of Sensitivity to Biases

We analyze the sensitivity of measures of statistical accuracy to under- and overconfidence and location bias (i.e., overestimating or underestimating). The method for calculating the biases was explained in Section 2.3.1. First, we examine the results for individual expert assessments, as depicted in Figure 6. *CRPS* is known to be location bias insensitive. However, the *CRPS* location-bias scatter plot does not show a very different pattern from the other measures, indicating that experts who score high with *CRPS* are not strongly location-biased. χ^2, *CRPS*, and *AD* (under Metalog assumption) are most sensitive to overconfidence, while *KS*, *CvM*, and *AD* (under PWU assumption) are least sensitive to overconfidence. All scores are sensitive to underconfidence, however *CRPS* actually rewards it, which is further discussed in Section 4.1.

Another approach to assess the sensitivity of measures of statistical accuracy to biases is by sampling from known distributions. We simulate four experts with different biases, 1) the perfectly statistically accurate (no bias), 2) the overconfident, 3) the underconfident, and 4) the location-biased (overestimating) expert.

The results of the simulation are shown in the Figure 7. The four columns correspond to the four experts, with the top row showing the beta-distribution from which realization quantiles are sampled. For each expert, 5 to 50 values are sampled from the distribution. Repeating this process 10,000 times gives the distribution of *p*-values, indicated with the colored bands.

CRPS shows the highest sensitivity to overconfidence, followed by χ^2 and *AD*, and finally *KS* and *CvM*. For underconfidence, a similar sensitivity pattern emerges, except that *CRPS* rewards rather than penalizes underconfidence. An expert with location bias gets the lowest *p*-values from *AD*, *KS*, and *CvM*, followed by χ^2. In general, *KS* and *CvM* respond similar to biases and *AD* and χ^2 do as well. *CRPS* does not pick up location bias (as explained by Nane and Cooke (2024)).

For the perfectly calibrated expert, all test statistics produce a uniform distribution for the *p*-value, which aligns with the asymptotic or exact distribution. χ^2 requires more realizations to reach this uniform result because the χ^2 distribution is an asymptotic rather than an exact distribution of the χ^2 test-statistic. For this reason, a *p*-value equal to 1 is only possible with a multiple of 20 calibration variables, when eliciting the 5th, 50th and 95th percentile. Note that this is of little consequence for the Classical Model; using 10 variables was deemed sufficient to select a statistically accurate expert over an overconfident expert, and that the mean χ^2 score for 10 experts using the asymptotic distribution is not 0.50 but 0.40 (Roger M. Cooke 2014, see Figure 7).

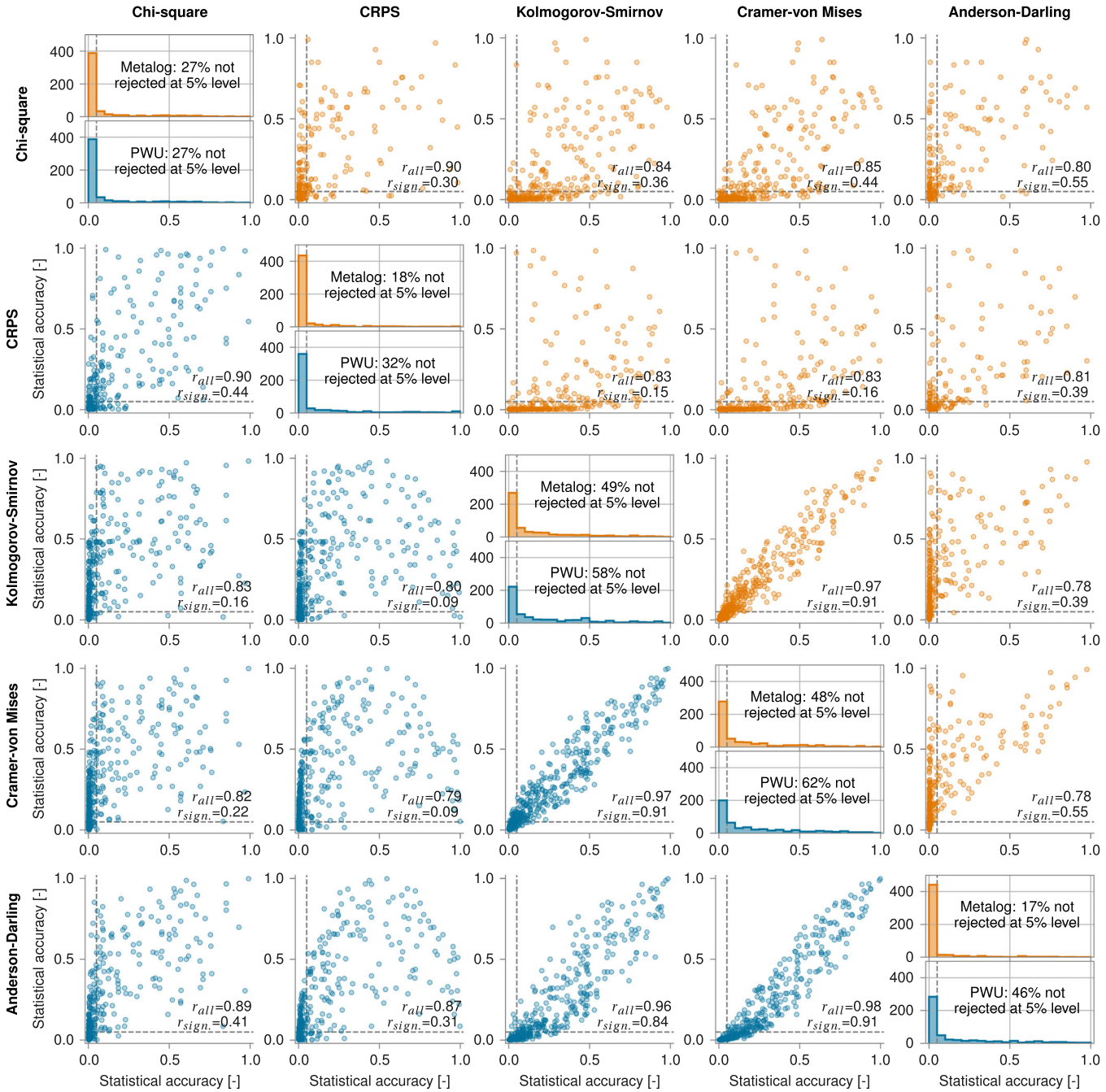

FIGURE 5 | Statistical accuracy for the 530 experts based on their quantile assessments in 49 case studies, using the Metalog distribution (upper right panels) and piecewise uniform (lower left panels). The dashed line represents the 5% significance level. The two numbers in the lower right of each panel are the rank correlation between all experts (above), and the rank correlation between all experts with a greater than 0.05 SA in both tests (below). Diagonal plots present the histogram of each measure's statistical accuracy for all 530 experts (i.e., the marginal distribution of each SA measure). In each histogram, the percentage of experts with a > 5% significance level is reported.

The continuous measures use an assumed distribution for the expert estimate to assign a realization to a quantile. The uncertainty introduced by this assumption is illustrated in this article through the differences between PWU and Metalog results. One example of this is the difference between the fractions < 5% in the histograms in Figure A2. These area equal for χ^2 but different for all other measures. This is a consequence of the assumed distribution that the expert envisioned for their estimate. The χ^2 score does not have uncertainty introduced by this.

3.3 | Comparison of Decision Maker Statistical Accuracy

The previous sections presented results and sensitivities for measures of SA individually. This section compares the weights derived using the different measures and the resulting statistical accuracy, following the procedure set out in Section 2.3.2. Appendix A presents illustrations of the intermediate steps that are followed in deriving the results presented in this section.

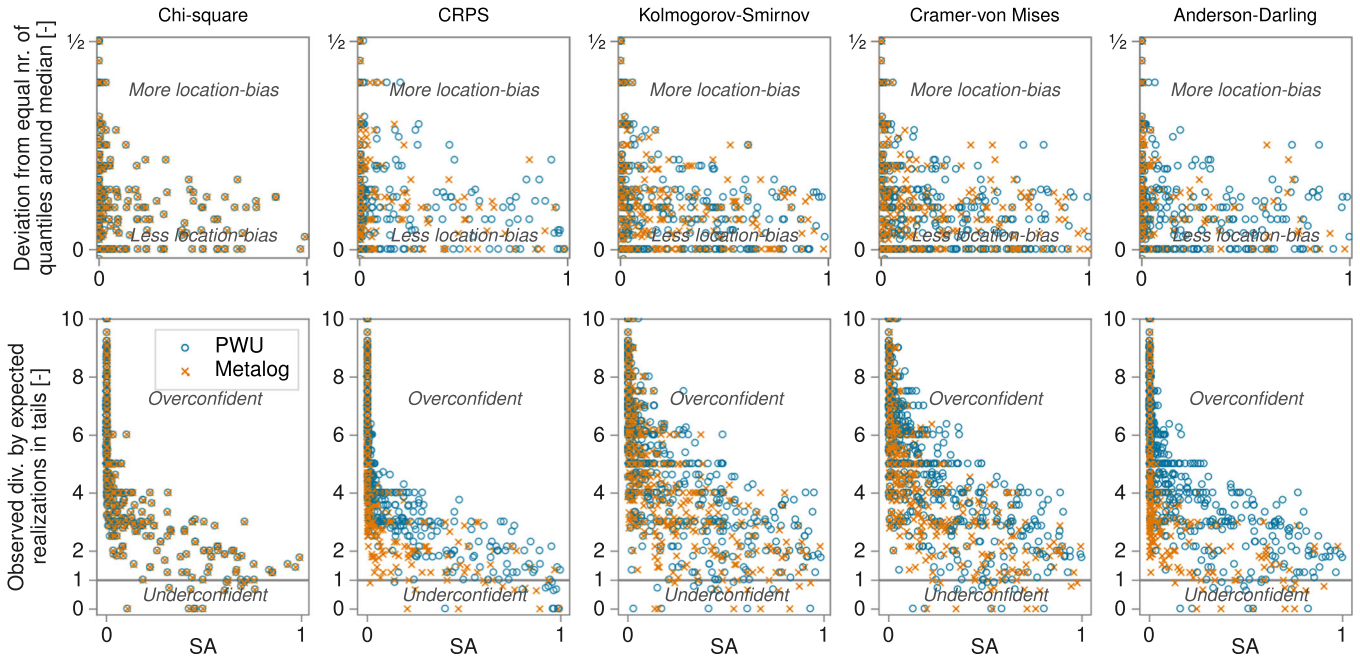

FIGURE 6 | Sensitivity of the measures of statistical accuracy to biases. The top row shows sensitivity to under- and overestimating experts (location-bias), calculated using Eq. 6. The bottom row shows under- and overconfidence, calculated using Eq. 7. Orange crosses indicate statistical accuracy calculated for the Metalog, blue circles for the piecewise uniform distribution.

The measures of statistical accuracy are compared by calculating decision maker weights using each measure of SA (the weight SA, recall that the weights are the normalized product of SA and informativeness), and evaluating the DM's statistical accuracy with, again, each of the measures of SA (the score SA). This yields 49 values per combination of weight SA and score SA, whose means are shown in Table 1. For most combinations, the values on the diagonal are highest, which is where the same measure of SA is used for weights and score. This difference is larger for the GLopt DM than for the GL DM. Note that all SA measures except *CRPS* have a "low opinion of equal weighting" with χ^2 having the lowest. *CRPS* DM's statistical accuracy is actually slightly higher for equal weighting than for *CRPS* weighting.

Table 2 displays the *p*-values of the Mann-Whitney test that compares whether the ranks of SA (for which the means are shown in Table 1) are significantly different from each other. The top five rows compare the ranks under the piecewise uniform assumption and the global weights DM. The only significant number is the 0.025 between *CRPS* (row) and *CvM* (column). This suggests that the SA calculated with DM weights from *CvM* evaluated with χ^2 is significantly higher than the SA calculated with DM weights from *CRPS* evaluated with χ^2. Or, $P\left(SA_{DM_{CvM}|\chi^2} > r\right)$ is greater than $P\left(SA_{DM_{CRPS}|\chi^2} > r\right)$ for all r in (0, 1). Table 2 compares the sets of ranks when evaluated with χ^2. Table A1 shows the same table but now with ranks compared using each of the other four measures of statistical accuracy.

Table 2 and Table A1 show that for the global weights DM without optimization the differences are mostly not significant. The exception are the weights from *CRPS*, which often score significantly lower, especially when evaluated according to the *KS*, *CvM*, or *AD* test. For the global weights DM *with* optimization, the χ^2 SA calculated with weights from every measure other than χ^2 itself, are significantly lower (see the first column in Table 2, rows corresponding to GLopt). Similarly, the *CRPS* SA calculated with global optimized weights from the other measures is considered significantly lower as well (see Table A1). And again, *KS*, *CvM*, and *AD* behave similarly as a group; the SAs calculated with weights from χ^2 and *CRPS* are significantly lower than the SAs calculated with weights from the measures themselves, but the SA from weights in between *KS*, *CvM*, and *AD* are not significantly lower (or higher). For the Metalog distribution, *AD*'s sensitivity to overconfidence makes it behave more similar to χ^2 and *CRPS*, and less similar to *KS* and *CvM*.

Based on the comparison in this section, and the analyses from Section 3.2, the measures of statistical accuracy can be divided into three categories with similar response to specific characteristics of expert assessments:

1. *KS*, *CvM*, and *AD* value quantiles close to their ranked position.
2. χ^2 values a proportional number of quantiles in bins.
3. *CRPS* values the median estimate close to the realization's quantile.

Figure 5 showed that while the correlation between each of these three categories is high, the correlation for the experts that score $SA > 0.05$ in both test is mostly low. Using the global weight algorithm gives enough spread in weight for the differences between *KS*, *CvM*, *AD* on one side, and χ^2 on the other, to be (mostly) not significant. This is because all four measures give a high statistical accuracy to a close to uniform distribution of quantiles between 0 and 1. However, *CRPS* responds to a much different characteristic, which makes the difference

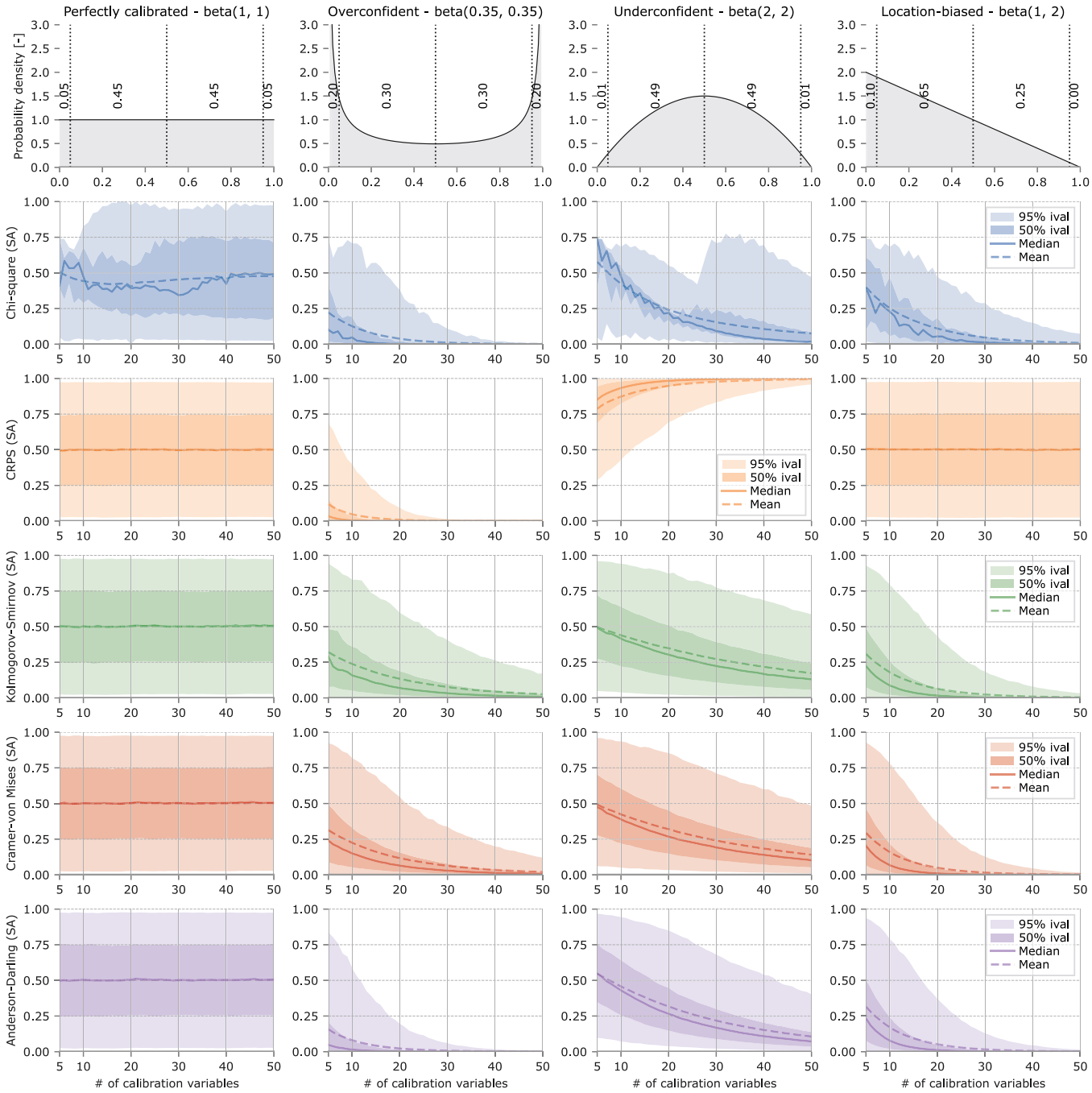

FIGURE 7 | Distribution of the DM's statistical accuracy for the different measures of statistical accuracy, resulting from drawing 10,000 samples of 5 up to 50 realization quantiles with different biases and evaluating their statistical accuracy. The four (non-)biases are: perfectly calibrated, overconfident, underconfident, and location-biased (overestimating).

between it and the other measures significant also under global weights.

When applying optimization, the weight is further concentrated onto a few experts. Referring again to Figure 5, if the weights are assigned to one (or a few) of the experts with a high SA for χ^2, it may not result in a high SA for one of the other measures. This is expressed by the significantly lower p-values in Tables 2 and 3, when applying optimized weights from another measure to the measure under evaluation itself. These lower p-values are to be expected because the different measures of SA assign high scores to different characteristics in the expert estimates (hence the spread in Figure A2.) Consequently, optimizing based on one characteristic reduces the performance when compared to a different uncorrelated characteristic.

This leaves the question why χ^2 behaves differently from KS, CvM and AD (i.e., why the correlation for $SA > 0.05$ in Figure 5 is low). Although all reward a uniform interquantile distribution, this likely results from evaluating quantiles on a continuous scale against evaluating them in bins. When realizations are close to the elicited quantile, a small difference in weight might cause a shift to another interval, which can cause a large difference in χ^2 SA score. Furthermore, the distribution of quantiles within an interquantile

TABLE 1 | Average DM statistical accuracy for the 49 case studies, calculated with weights from the measures of SA on the columns "SA (w.)" and Equal weights "EQ", and scored using the measures of SA on the rows "SA (sc.)". The results for the GL and GLopt decision maker, as well as the PWU and Metalog distribution are shown.

Dist.	DM	SA (w.) SA (sc.)	χ^2	CRPS	KS	CvM	AD	EQ
PWU	GL	χ^2	0.40	0.36	0.43	0.46	0.42	0.32
		CRPS	0.63	0.67	0.67	0.67	0.67	0.69
		KS	0.52	0.41	0.57	0.55	0.54	0.39
		CvM	0.51	0.40	0.56	0.57	0.55	0.38
		AD	0.51	0.39	0.56	0.57	0.55	0.39
	GLopt	χ^2	0.55	0.38	0.39	0.38	0.43	0.32
		CRPS	0.50	0.62	0.35	0.36	0.42	0.69
		KS	0.49	0.42	0.71	0.67	0.66	0.39
		CvM	0.49	0.43	0.69	0.72	0.68	0.38
		AD	0.49	0.44	0.63	0.66	0.65	0.39
Metalog	GL	χ^2	0.39	0.33	0.46	0.44	0.37	0.32
		CRPS	0.48	0.46	0.54	0.56	0.40	0.59
		KS	0.51	0.43	0.58	0.58	0.46	0.40
		CvM	0.50	0.43	0.59	0.60	0.45	0.39
		AD	0.45	0.36	0.53	0.54	0.36	0.40
	GLopt	χ^2	0.52	0.32	0.42	0.43	0.40	0.32
		CRPS	0.34	0.46	0.27	0.31	0.33	0.59
		KS	0.44	0.44	0.70	0.70	0.49	0.40
		CvM	0.46	0.44	0.67	0.71	0.50	0.39
		AD	0.34	0.36	0.39	0.47	0.41	0.40

interval does not matter for χ^2, while *KS*, *CvM* and *AD* reward them being spread out. GL combines all experts so that the CDF has gradient changes at the quantile values for each expert. For 10 experts there are 30 of these changes, which means the CDF looks rather continuous and the effect of interpolating a continuous CDF is attenuated. GLopt, on the other hand, typically weights only one or two experts and hence has much fewer gradient changes. This might amplify the distortions introduced by interpolating a continuous CDF. While this section explored some of the aspects that cause the differences in behavior, the last word on this has not been said.

Finally, the behavior of *AD* heavily depends on assuming the PWU or Metalog distribution. *AD* penalizes quantiles close to 0.0 or 1.0 much more than quantiles close to 0.05 and 0.95 (refer to the weight in equation 3). This means that assuming the Metalog distribution will make *AD* (much) more sensitive to overconfidence, and therefore behave more similar to χ^2 and *CRPS* than to *KS* and *CvM*.

3.4 | Assessing Accuracy of Interpolated Quantiles

Before the elicitation, an analyst chooses whether to elicit five or three quantiles. This choice is made by the analyst for the whole study, not per variable. Opting for three percentiles lowers the elicitation burden, whereas five percentiles more accurately represent the experts' distributions. A distribution that would accurately describe the expert envisioned distribution, could potentially give the accuracy of more than three percentiles, while eliciting only three. The reason for considering the Metalog distribution for interpolating between quantiles, is the hypothesis that it more accurately describes the distribution envisioned by experts. Unlike the PWU distribution, the Metalog distribution lacks discontinuities in probability density at estimated percentiles and, in its the three-percentile version, resembles a bell-shaped curve that is commonly observed in natural samples. To test this hypothesis, we consider the cases with five elicited percentiles, remove the second and fourth quantile, and estimate their position using both distributions (see Section 2.3.3).

The results are illustrated in Figure 8, which displays the difference between distribution-estimated percentiles and expert-estimated percentiles. Panel a shows this for $F(X_{q=0.25}) - 0.25$, with F being the CDF of Metalog or PWU and $X_{q=0.25}$ the expert estimate for the 25th percentile. The dashed line represents a perfect prediction by the distribution. For the 25^{th} percentile, values to the left indicate that the distribution assigned a lower percentile to the experts' estimate. Conversely, values to the right indicate that the distribution assigns a higher percentile than the experts.

TABLE 2 | p values for the Mann-Whitney rank-sum test. A p values less than 0.05 (red) suggests that the ranks of the SAs calculated with the measure of SA in the row is less than the ranks of the SAs calculated with the measure of SA in the column. The SA-values are ranked according to the 49 χ^2 DM's SAs. Both the ranks using PWU and Metalog distribution, as well as using GL and GLopt decision maker, are compared.

Dist	DM	SA	χ^2	CRPS	KS	CvM	AD
PWU	GL	χ^2		0.818	0.217	0.121	0.341
		CRPS	0.184		0.055	0.025	0.101
		KS	0.785	0.945		0.345	0.649
		CvM	0.881	0.976	0.657		0.809
		AD	0.661	0.900	0.353	0.193	
	GLopt	χ^2		1.000	0.998	0.999	0.987
		CRPS	0.001		0.460	0.582	0.182
		KS	0.002	0.542		0.602	0.230
		CvM	0.001	0.421	0.400		0.157
		AD	0.013	0.820	0.772	0.844	
Metalog	GL	χ^2		0.880	0.093	0.198	0.731
		CRPS	0.121		0.009	0.027	0.275
		KS	0.908	0.991		0.660	0.962
		CvM	0.804	0.974	0.343		0.919
		AD	0.272	0.727	0.039	0.082	
	GLopt	χ^2		0.999	0.959	0.963	0.981
		CRPS	0.001		0.040	0.027	0.117
		KS	0.042	0.961		0.409	0.662
		CvM	0.038	0.974	0.594		0.702
		AD	0.020	0.884	0.340	0.301	

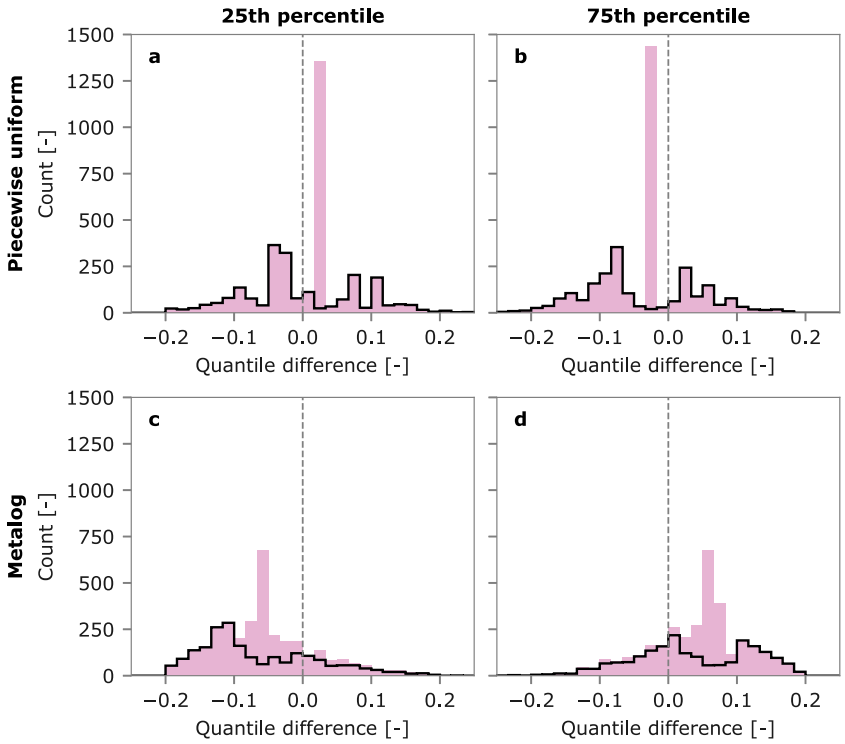

FIGURE 8 | Difference between the experts' estimated 25th percentile (a, c) and 75th percentile (b, d), and the positions according to the PWU distribution (a, b) and Metalog distribution (c, d) fitted to the five-percentile cases with the 25th and 75th percentile removed.

Panels *a* and *b* show high bin values at 0.025. These values correspond to cases where an expert assigned a value to the 25^{th} percentile precisely between the 5^{th} and 50^{th} percentiles' values. This situation accounts for approximately 20% of expert estimates. Filtering these estimates results in the black outlined histogram. On average, PWU performs better than Metalog. Nevertheless, both distributions show significant deviations when estimating the missing percentiles. It seems that the distributions both lack predictive power for the missing percentiles.

Based on this analysis, the Metalog does not offer a better representation of experts' estimates than PWU. The smooth and more informative distribution is too precise (i.e., it concentrates probability density more than experts appear to do). Consequently, the best approach to obtain a more accurate representation of experts' probability density functions (PDFs) is to elicit more percentiles.

Note that adding percentiles does not seem to be needed for increasing the decision maker's statistical accuracy. To assess this, we compared the decision maker's SA with weight from the 3-percentile estimates (i.e., with removed percentiles) to the decision makers constructed using weights from the 5-percentile case. Weights were calculated using χ^2 and the PWU distribution. The SAs for both decision makers were then calculated based on 3-percentile estimates (to ensure a fair comparison). Over the 17 studies, the DM constructed with weights from the 5-percentile estimates scored on average 0.008 higher with a standard deviation of 0.09; a negligible difference.

4 | Discussion and Conclusions

This study set out to test five different measures of statistical accuracy for scoring experts in an expert judgment study. The results are applicable to evaluating and combining uncertainty estimates in a broader context as well, such as in forecasting. The newly evaluated test statistics — the Continuous Ranked Probability Score (*CRPS*), Kolmogorov-Smirnov (*KS*), Cramer-von Mises (*CvM*), and Anderson-Darling (*AD*) — were assessed as alternative to relative information based chi-square test (χ^2) used in the Classical Model. Where χ^2 interprets and scores the estimates through discrete quantile intervals, the four alternatives map realizations on a continuous CDF-scale and accordingly calculate the statistical accuracy based on a continuous distribution. This makes the assumed distribution that connects the expert estimated percentiles relevant. In this context, the Metalog distribution was explored as an alternative to the piecewise uniform (PWU) assumption that is typically employed to model the estimated probability density in the Classical Model. The test statistics were assessed through 49 expert judgment studies from the last decades, and by sampling estimates from distributions with a specific bias. The study's findings are discussed in two parts: the performance of various test statistics (Section 4.1) and the performance of the Metalog distribution (Section 4.2). Finally, Section 4.3 discusses the implications for SEJ practitioners.

4.1 | Performance of Different Test Statistics

Comparing the five measures of statistical accuracy revealed varying sensitivity to different biases. χ^2 is sensitive to overconfidence, underconfidence and location bias. *KS* and *CvM* are sensitive to location bias and underconfidence but less sensitive to overconfidence. *AD* performs relatively similar to *KS* and *CvM* when assuming a PWU distribution. The assumed overconfidence related to the Metalog distribution however makes the *AD* weights much stricter on overconfidence. Because of this, *AD* behaves more similar to χ^2 under the Metalog assumption.

The new scale-invariant *CRPS* is sensitive to overconfidence but insensitive to location bias and it rewards underconfidence. A relationship between *CRPS* and underconfidence is illustrated in Figure 9. The figure shows the relationship between the average quantile distance to the median and the statistical accuracy (panel a) or combined score (panel b).

This distance is calculated as $\sum_{i=1}^{N}|F_{i,j}(x_i) - 0.5|/N$, in which $F_{i,j}$ is expert j's estimate for item i with realization x_i. An unbiased, uniform, quantile distribution would have an expected distance of 0.25. The figure shows that there is a strong relationship between *CRPS* and distance to median. This is because 1) the scale-invariant CRPS compares a realization to a uniform distribution, resulting in the minimal difference area while aligning the median

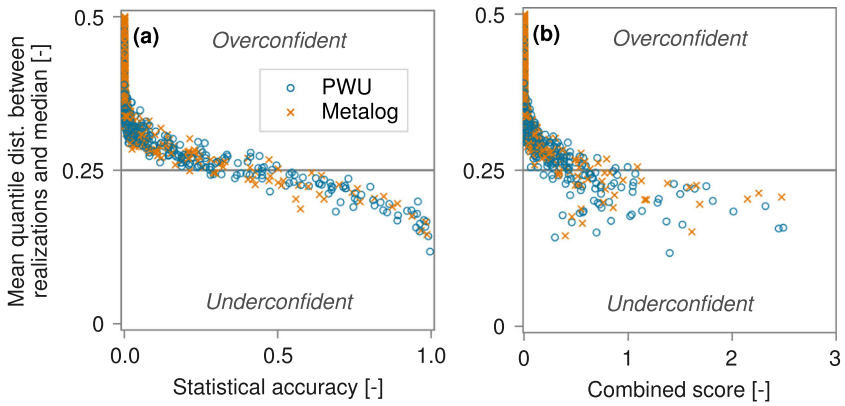

FIGURE 9 | Mean absolute difference between realization's quantile and median for CRPS. Plotted against statistical accuracy (a), and plotted against combined score (b).

estimate with the realization (see Figures 2), 2) The CRPS considers every estimate individually, so the maximum score is achieved when all are close to the median. This distinguishes it from *CvM* and *AD*, where a uniform distribution of all quantiles scores best. The relationship is weakened by considering the combined score which including informativeness. Because the lower informativeness of underconfident experts does not offset the high statistical accuracy, underconfident experts still achieve high combined scores. This means that a perfectly accurate expert could achieve a higher statistical accuracy, and likely a higher weight as well, when deliberately making underconfident estimates. Therefore, using *CRPS* does not encourage experts to state their true, unbiased, beliefs. This means the scale-invariant *CRPS* is not a proper scoring rule, as also pointed out by Bracher and acknowledged in a footnote in (Nane and Cooke 2024).

Measures of statistical accuracy were compared by calculating decision maker (DM) weights with one measure and evaluating them against DMs from the other measures. The results show that for global weights without optimization, the differences between SA scores based on weights from different measures are mostly insignificant. Except for *CRPS*, whose weights lead to significantly lower SA when evaluated using other measures. Applying *optimized* global weights further concentrates weights on experts with a high SA in the specific test-statistic, for which reason they perform worse when applied to most other measures of statistical accuracy.

Under the PWU assumption, SAs calculated with DM-weights from *AD* are considered not statistically significantly different when evaluated by *KS* and *CvM* (see Table A1). However, when calculated under the Metalog distribution, the SA from *AD* weights rank significantly lower when evaluated using *KS* and *CvM*. The inverse is not the case, weights from *KS* and *CvM* do not rank significantly different when evaluated by *AD* under either PWU or Metalog. This indicates that penalizing overconfidence (as *AD* does under Metalog) leads to significantly lower SA, while being less strict on overconfidence (i.e., using *KS* or *CvM* weights under Metalog) does not significantly reduce SA.

4.2 | Performance of the Metalog Distribution

The Metalog distribution (Keelin 2016) was explored as an alternative to the piecewise uniform (PWU) distribution that is commonly used in CM. *CRPS*, *KS*, *CvM*, and *AD* require the realizations' quantile positions rather than intervals, and the Metalog offers a flexible distribution for placing those. The most significant differences between the Metalog and PWU are the quantile positions for realizations that fall outside the [0.05, 0.95] interval, as shown in Figure 4. Using the Metalog means assuming a higher degree of overconfidence, yielding worse results for measures of statistical accuracy that penalize this. However, this issue is not intrinsic to the Metalog distribution itself, but resides in the (overshoot) range that is assumed for the PWU distribution.

The appealing feature of the Metalog is its smooth, bell-shaped curve, which may be more intuitive to experts (a 'soft' argument, but nonetheless relevant in the field of expert judgment). Typically, a naturally observed continuous variable does not exhibit steps in probability density at (estimated) percentiles, as the PWU examples in 3 show. Assuming a bell-shaped curve increases probability density closer to the median while reducing it toward the tails (see Figure 3 a). Paradoxically, when removing the second and fourth percentile from the five-percentile cases and estimating their removed position, the Metalog distribution mostly overestimates the probability density within the 25th to 75th percentile interval. This implies that, in these case studies, experts more often estimated a platykurtic (negative kurtosis, thin-tailed) than a leptokurtic (positive kurtosis, fat-tailed) distribution. The PWU distribution also underperformed in this experiment, highlighting that the best approach for an analyst to obtain an accurate representation of the full distribution is to assess more percentiles.

Fitting a Metalog distribution to all expert estimates in the 49 case-studies proved challenging. While the Metalog distribution offers high shape flexibility, it could not accommodate highly skewed three-percentile estimates without imposing bounds. Additionally, many five-percentile estimates could not be fitted without dividing the distribution into two three-percentile Metalog distribution parts (such as shown in Figure 3g).

4.3 | Final Remarks

The comparisons between the different measures and distributions provide insight into their behavior, but also into the CM in general. Therefore, we reflect on the result in the context of implications to the practitioners, discussing whether to use performance weighting and optimization, and which distribution and measure of SA to use.

Should you use performance weighting? This study shows that all measures except *CRPS* agree that performance weighting is superior to equal weighting with respect to statistical accuracy. Note that is based on 'in-sample' comparisons. Colson and Cooke (2017) show that in out-of-sample comparisons performance weighting increased informativeness without sacrificing accuracy. Therefore, we recommend using performance weighting despite the additional effort involved in collecting and eliciting seed questions.

Which distribution to use, PWU or Metalog? Neither the PWU nor Metalog distribution did a good job of predicting missing quantile assessments. The Metalog has the advantage over PWU that no overshoot assumption is needed to construct a distribution from the estimated percentiles. However, using the Metalog gave issues with consistently fitting the distribution to expert estimates. In our view, the minimal assumptions in the PWU distribution result in wide applicability and ease of use, and therefore should be the standard approach in the CM. However, this does not mean that an analyst could not decide otherwise; in some applications (e.g., when using gradient-based sampling) a smooth Metalog distribution might be preferable over a stepped PDF.

Which measure of SA should you use? All five investigated measures of statistical accuracy have different effects on the resulting combined estimate (DM), such as the number of experts included with significant weight, the sensitivity to different biases, and the assumptions required for calculating the weights. When assuming a piecewise uniform distribution, *KS*, *CvM*, and *AD* behave similarly, but different from χ^2, which again behaves differently to *CRPS*. When assuming the Metalog,

the higher degree of assumed overconfidence makes *AD* behave more similar to χ^2 and *CRPS*, and less to *KS* and *CvM*.

Based on our findings, we would advise not to use *CRPS*, as it performs worse than the others and encourages underconfidence. If the goal is to increase the number of experts with a significant weight, *KS* and *CvM* are both options. However, *AD* was developed as a 'better' version of *CvM*, by being more sensitive to observations in the tails of the distributions. Within the CM, this means a better ability to pick up on the overconfidence bias. This leaves the *AD* and χ^2 tests. χ^2 does not use quantile positions, so it does not need an assumed distribution. The effect of assuming a distribution is significant; it makes the *AD* score behave either like χ^2 (with Metalog) or like *KS* and *CvM* (with PWU). An advantage of the *AD* test might be that it evaluates the positions of each realization, that is, it compares the realizations to a uniform distribution, rather than in four bins. However, whether this provides a significant advantage over χ^2 is unclear from this study, and should be subject of further research. Until then, we advice to use the χ^2 as the default option in the Classical Model.

Should you use optimization in CM? Non-optimized weights derived with one of *KS*, *CvM*, or *AD* and evaluated with χ^2 or vice versa, do not provide significantly worse DMs. However, when optimization is used they do perform significantly worse when evaluated with the other. This is to be expected, as optimization concentrates weight on experts or forecasts that score high according to the chosen measure of SA, and different measures of SA do not assign high weights to the same experts or forecasts (see Figure A2.) However, this considers only the statistical accuracy, and weights in the CM are assigned based on the combination of SA and informativeness. The purpose of optimization in the Classical Model is to increase the informativeness of the DM without decreasing its SA.

In many cases (e.g., Bamber et al. 2019) the difference between the optimized DM and the DM with a 5% SA cutoff are very small and preference is given to the added robustness of weighting more experts above a very small improvement in DM performance. It is not uncommon in such situations to give in on the optimum to gain robustness, as one is always balancing multiple objectives.

Ultimately, the choice of using optimization ties back to the confidence in the chosen measure of SA. Further research on the differences between the different measures, in particular χ^2 and *AD*, including the information score, could provide more insight into this. We recommend redoing some of the original analyses on the Classical Model with different measures of SA and in particular *AD*, the best performing of the alternative scoring rules. Until that has provided more conclusive results, having various options of distributions as well as different measures of statistical accuracy provides analysts with flexibility to tailor the approach to their specific study. These options are available through the open-source Anduryl software, as detailed in 'code and availability' below.

Acknowledgments

This study was funded by the TKI project EMU-FD. This study project is funded by Rijkswaterstaat, Deltares and HKV consultants.

Data Availability Statement

The data and the code used to process the presented results are openly available under the GNU license through: https://github.com/grongen/classical_model_sa_measures_distributions. The version of Anduryl that facilitates the different measures of SA and distributions, is available through: https://github.com/grongen/anduryl/tree/metalog.

Endnotes

[1]The contributions of Bram Wisse, Tim Bedford, John Quigley, Sandra Hoffmann, Paul Fischbeck, Alan Krupnick, Michael McWilliams, O. Morales, D. Kurowicka, A. Roelen, Shi-Woei Lin, Vicki M. Bier, Thomas A. Mazzuchi, William G. Linzey, Armin Brunin, Jouni T. Tuomisto, Andrew Wilson, John S. Evans, Marko Tainio, Roger Cooke, ElSaadany, Xinzheng Huang, Robert Clemen, Anthony O'Hagan and Simon French are gratefully acknowledged.

References

Adamczewski, T. 2023. "The Metalog Distribution." GitHub Repository. https://github.com/Arturus/metalogistic.

Anderson, T. W., and D. A. Darling. 1952. "Asymptotic Theory of Certain' Goodness of Fit' Criteria Based on Stochastic Processes." *The Annals of Mathematical Statistics* 23: 193–212.

Bamber, J. L., M. Oppenheimer, R. E. Kopp, W. P. Aspinall, and R. M. Cooke. 2019. "Ice Sheet Contributions to Future Sea-Level Rise From Structured Expert Judgment." *Proceedings of the National Academy of Sciences* 116, no. 23: 11195–11200.

Brown, T. A. 1974. "Admissible Scoring Systems for Continuous Distributions." *RAND Corporation* 1: 1.

Colonna, K. J., G. F. Nane, E. F. Choma, R. M. Cooke, and J. S. Evans. 2022. "A Retrospective Assessment of COVID-19 Model Performance in the USA." *Royal Society Open Science* 9, no. 10: 220021.

Colson, A. R., and R. M. Cooke. 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." *Experts and Consensus in Social Science* 1: 191–212.

Cooke, R. M., D. Marti, and T. Mazzuchi. 2021. "Expert Forecasting With and Without Uncertainty Quantification and Weighting: What Do the Data Say?" *International Journal of Forecasting* 37, no. 1: 378–387.

Cooke, R. M., and L. L. H. J. Goossens. 2008. "TU Delft Expert Judgment Data Base." *Reliability Engineering & System Safety* 93, no. 5: 657–674. https://doi.org/10.1016/j.ress.2007.03.005.

Cramér, H. 1928. "On the Composition of Elementary Errors: First Paper: Mathematical Deductions." *Scandinavian Actuarial Journal* 1928, no. 1: 13–74.

Csörgő, S., and J. J. Faraway. 1996. "The Exact and Asymptotic Distributions of Cramér-Von Mises Statistics." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58, no. 1: 221–234.

Grace, A. W., and I. A. Wood. 2012. "Approximating the Tail of the Anderson–Darling Distribution." *Computational Statistics & Data Analysis* 56, no. 12: 4301–4311.

Keelin, T. W. 2016. "The Metalog Distributions." *Decision Analysis* 13, no. 4: 243–277.

Kolmogorov, A. N. 1933. "Sulla Determinazione Empirica Di Una Legge Didistribuzione." *Giorn Dell'inst Ital Degli Att* 4: 89–91.

Mann, H. B., and D. R. Whitney. 1947. "On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other." *The Annals of Mathematical Statistics* 18: 50–60.

Marsaglia, G., and J. Marsaglia. 2004. "Evaluating the Anderson-Darling Distribution." *Journal of Statistical Software* 9: 1–5.

Von Mises, R. 1928. *Julius Springer*. Statistik und wahrheitUnd Wahrheit 20.

Nane, G. F., and R. M. Cooke. 2024. "Scoring Rules and Performance, New Analysis of Expert Judgment Data." *Futures & Foresight Science* 1: e189.

Ren, X., G. F. Nane, K. C. Terwel, and P. H. A. J. M. van Gelder. 2024. "Measuring the Impacts of Human and Organizational Factors on Human Errors in the Dutch Construction Industry Using Structured Expert Judgement." *Reliability Engineering & System Safety* 244: 109959. https://doi.org/10.1016/j.ress.2024.109959.

Rongen, G., C. M. P. Hart, G. Leontaris, and O. Morales-Nápoles. 2020. "Update (1.2) to Anduril and Anduryl: Performance Improvements and a Graphical User Interface." *SoftwareX* 12: 100497. https://doi.org/10.1016/j.softx.2020.100497.

Rongen, G., O. Morales-Nápoles, and M. Kok. 2022. "Expert Judgment-Based Reliability Analysis of the Dutch Flood Defense System." *Reliability Engineering & System Safety* 224: 108535. https://doi.org/10.1016/j.ress.2022.108535.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52, no. 3/4: 591–611.

Simard, R., and P. L'Ecuyer. 2011. "Computing the Two-Sided Kolmogorov-Smirnov Distribution." *Journal of Statistical Software* 39: 1–18.

Smirnoff, N. 1939. "Sur Les Écarts De La Courbe De Distribution Empirique." *Matematicheskii Sbornik* 48, no. 1: 3–26.

Appendix A
Background on comparison between measures of statistical accuracy

The procedure for cross-comparing the test statistics was explained in Section 2.3.2, and the results were presented in Section 3.3. This appendix section gives additional explanation and illustrations on the intermediate steps.

When applying the different measures of statistical accuracy to the 49 studies, some measures give higher statistical accuracy (on average) than others. Comparing the obtained statistical accuracies is therefore not unbiased. To overcome this, empirical distributions were derived for each measure (recall step 4 in the list in Section 2.3.2). These empirical distributions are shown in Figure A1. Each panel contains five curves, of which each is constructed by calculating the decision maker in the 49 studies and ranking the resulting SA. For example, the orange *CRPS* distribution was constructed by calculating the *CRPS* statistical accuracy for all experts and combining that with the informativeness to derive weights. These weights were combined into a decision maker, for which the *CRPS* SA was calculated. This gives one of the 49 markers in the empirical distribution. Repeating this for all cases using the global weight DM with and without optimization, and the piecewise uniform and Metalog distribution, results in the five distributions in each of the four panels of Figure A1.

The empirical distributions of each method's DM-scores are primarily used for comparing statistical accuracy. However, they also show that:

1. Measures of statistical accuracy that are less sensitive to overconfidence (i.e., *KS*, *CvM*) tend to return higher SA scores for DM, especially under the Metalog assumption. This is because overconfidence is a prevailing bias in expert judgment studies.
2. Optimization results in higher DM SA. Note that this is not necessarily the primary goal of optimization, which typically results in increased informativeness while not decreasing the SA.
3. Assuming the Metalog distribution gives lower scores than assuming PWU for *AD* and *CvM*, due to the measures' sensitivity to overconfidence. χ^2 is sensitive to overconfidence as well, but unaffected by the choice of distribution because it utilizes interquantile intervals. At the same time, it profits from the higher

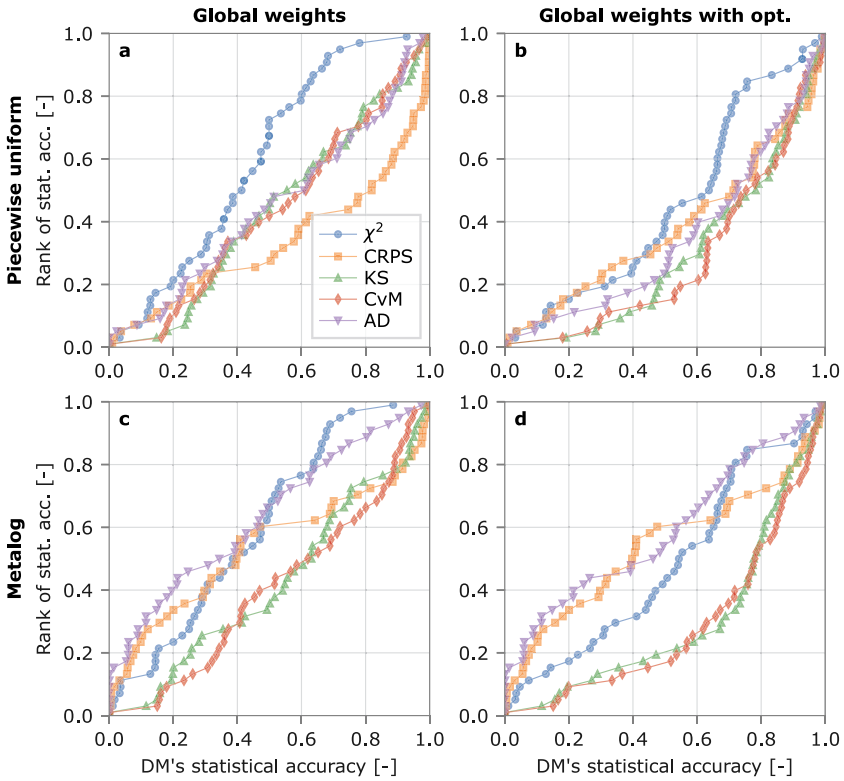

FIGURE A1 | Empirical distributions of the decision makers' statistical accuracies in the 49 case-studies. (a) and (b) for piecewise uniform distribution, (c) and (d) for the Metalog; (a) and (c) for the global weights DM without optimization, and (b) and (d) with optimization.

informativeness of the Metalog distribution (the same reason why *KS* and *CvM* score higher with Metalog as well.)

Using the empirical distributions, a χ^2 SA calculated with DM weights from, for example, the *CRPS*, can be compared to the χ^2 SA calculated with DM weights from χ^2 itself. Doing this for all 49 cases results in a set of ranks per measure of SA. Figure A2 displays these ranks for SA calculated with weights from each measure of SA and evaluated using each measure of SA. For example:

- Figure A2 a., first box plot from the left, shows the χ^2 SA calculated with global DM_{χ^2} weights, and compared to the empirical distribution of DM_{χ^2} SA. This is in fact comparing the χ^2 ranks to the χ^2 ranks itself, resulting in a uniform distribution.
- Figure A2 a., second box plot from the left, shows the χ^2 SA calculated with global DM_{CRPS} weights, and compared to the empirical distribution of DM_{χ^2} SA. The ranks for DM_{CRPS} are on average slightly lower.
- Figure A2 f., second box plot from the left, shows the χ^2 SA calculated with global DM_{CRPS} *optimized* weights, and compared to the empirical distribution of *optimized* DM_{χ^2} SA. Now the *CRPS* weights give substantially lower statistical accuracy.

Whether substantially lower is also significantly lower is tested using the Mann Whitney test. The resulting *p*-values of those tests, for the DM weights evaluated using χ^2, were presented in Table 2. The first five rows in the table

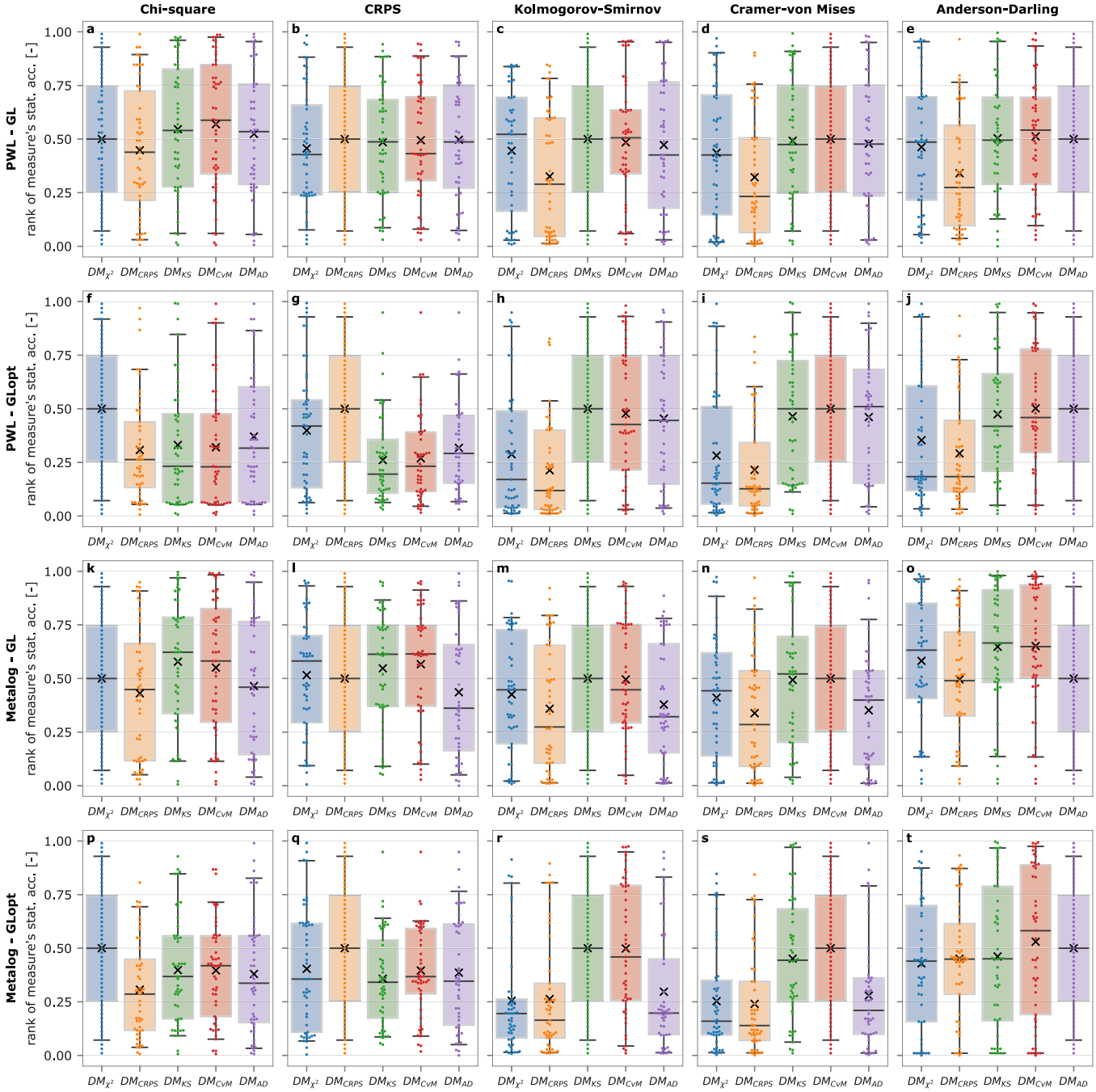

FIGURE A2 | Comparison of all measures of statistical accuracy. DMs composed using each measures' weight (each box plot) are compared to the empirical distribution of every measure (each column). The comparison is done with and without optimization, and for the PWU and Metalog distribution. Each box indicates the 25th to 75th percentile range, with the horizontal line being the median. The fliers indicate the 5th and 95th percentile, and the cross is positioned at the mean.

TABLE A1 | *p*-values for the Mann-Whitney rank-sum test evaluated using the measures of SA: CRPS, KS, CvM, and AD. A *p*-value less than 0.05 (red) suggests that the ranks of the SAs calculated with the measure of SA in the row is less than the ranks of the SAs calculated with the measure of SA in the column. The SA-values are ranked according to the empirical distributions of DM SA displayed in Figure A1. Both the ranks using PWU and Metalog distribution, as well as using GL and GLopt decision maker, are compared. SA (w.) indicates the measure of SA used to calculate the weights, SA (sc.) is the measure used to score the weights.

		SA (sc.)	CRPS					Kolmogorov-Smirnov					Cramer-von Mises					Anderson-Darling				
Dist	**DM**	**SA (w.)**	χ^2	**CRPS**	**KS**	**CvM**	**AD**	χ^2	**CRPS**	**KS**	**CvM**	**AD**	χ^2	**CRPS**	**KS**	**CvM**	**AD**	χ^2	**CRPS**	**KS**	**CvM**	**AD**
PWU	GL	χ^2		0.23	0.23	0.22	0.23		0.97	0.18	0.31	0.32		0.96	0.16	0.14	0.23		0.98	0.20	0.17	0.26
		CRPS	0.77		0.61	0.54	0.52	0.03		0.00	0.00	0.01	0.04		0.00	0.00	0.00	0.02		0.00	0.00	0.00
		KS	0.77	0.40		0.46	0.44	0.82	1.00		0.60	0.69	0.84	1.00		0.45	0.62	0.80	1.00		0.40	0.52
		CvM	0.78	0.47	0.54		0.49	0.69	1.00	0.41		0.56	0.86	1.00	0.55		0.65	0.83	1.00	0.60		0.60
		AD	0.77	0.48	0.57	0.52		0.68	0.99	0.32	0.45		0.77	1.00	0.39	0.36		0.75	1.00	0.48	0.40	
	GLopt	χ^2		0.04	0.99	0.98	0.87		0.86	0.00	0.00	0.00		0.86	0.00	0.00	0.00		0.83	0.02	0.00	0.01
		CRPS	0.96		1.00	1.00	1.00	0.15		0.00	0.00	0.00	0.14		0.00	0.00	0.00	0.17		0.00	0.00	0.00
		KS	0.01	0.00		0.42	0.08	1.00	1.00		0.65	0.79	1.00	1.00		0.27	0.57	0.98	1.00		0.32	0.33
		CvM	0.02	0.00	0.58		0.10	1.00	1.00	0.36		0.61	1.00	1.00	0.74		0.75	1.00	1.00	0.68		0.51
		AD	0.13	0.00	0.92	0.90		1.00	1.00	0.22	0.39		1.00	1.00	0.44	0.25		0.99	1.00	0.67	0.49	
Metalog	GL	χ^2		0.61	0.25	0.16	0.91		0.86	0.10	0.15	0.82		0.89	0.09	0.07	0.84		0.94	0.11	0.11	0.92
		CRPS	0.39		0.21	0.13	0.86	0.14		0.01	0.01	0.37	0.11		0.00	0.00	0.39	0.06		0.00	0.00	0.47
		KS	0.76	0.79		0.39	0.97	0.90	0.99		0.53	0.98	0.91	1.00		0.45	0.99	0.90	1.00		0.46	0.99
		CvM	0.85	0.87	0.62		0.98	0.85	0.99	0.47		0.98	0.93	1.00	0.55		0.99	0.89	1.00	0.54		1.00
		AD	0.09	0.14	0.03	0.02		0.18	0.63	0.02	0.02		0.16	0.62	0.01	0.01		0.08	0.53	0.01	0.00	
	GLopt	χ^2		0.05	0.72	0.44	0.56		0.55	0.00	0.00	0.28		0.68	0.00	0.00	0.28		0.30	0.28	0.06	0.12
		CRPS	0.95		0.99	0.96	0.97	0.45		0.00	0.00	0.26	0.32		0.00	0.00	0.17	0.70		0.46	0.10	0.20
		KS	0.28	0.01		0.15	0.34	1.00	1.00		0.50	1.00	1.00	1.00		0.20	1.00	0.72	0.54		0.19	0.26
		CvM	0.57	0.04	0.85		0.62	1.00	1.00	0.50		1.00	1.00	1.00	0.80		1.00	0.94	0.91	0.81		0.71
		AD	0.44	0.03	0.66	0.38		0.72	0.74	0.00	0.00		0.73	0.83	0.00	0.00		0.88	0.80	0.74	0.29	

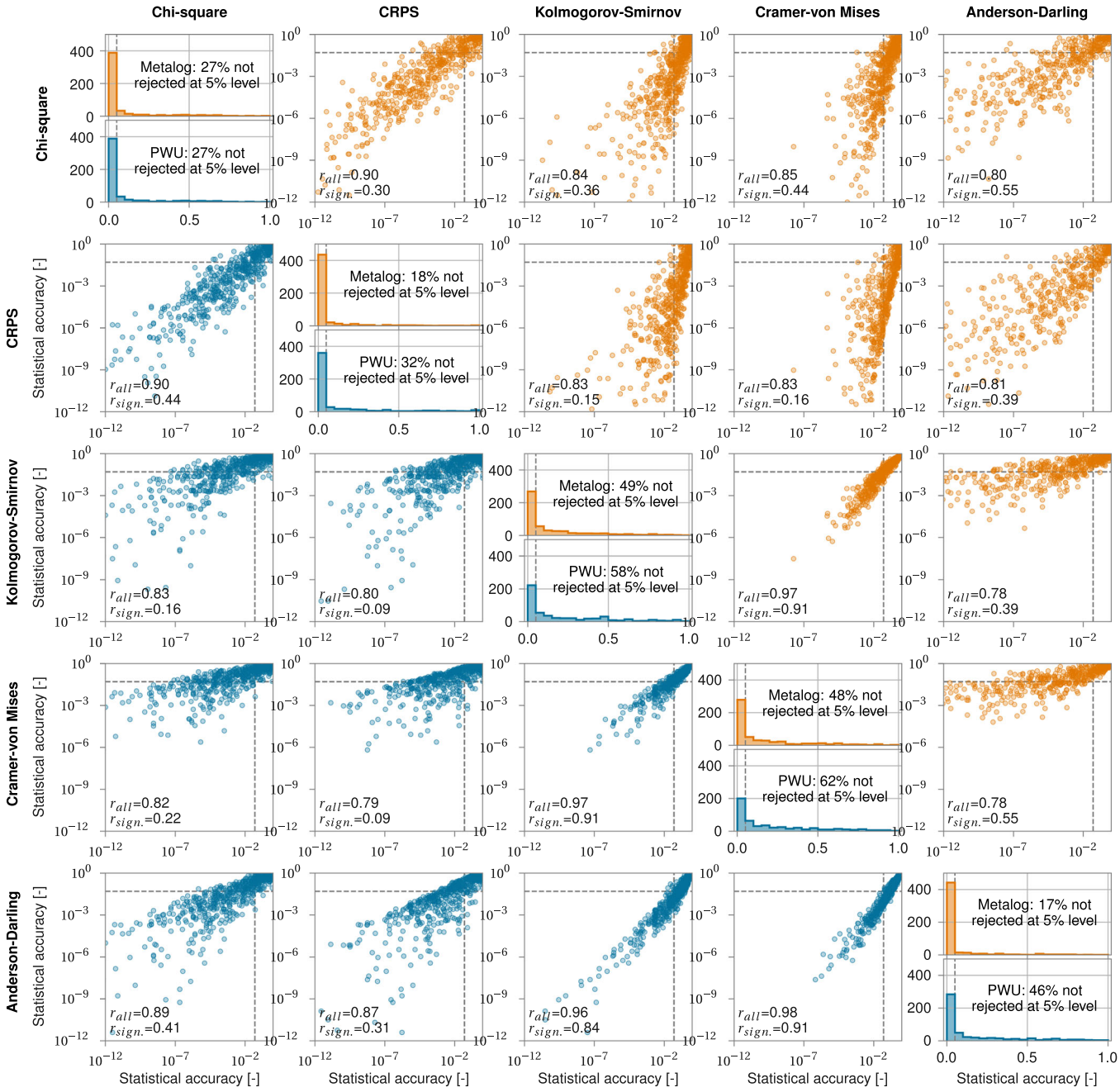

FIGURE A3 | Statistical accuracies for the 530 experts based on their quantile assessments in 49 case studies, using the Metalog distribution (upper right panel) and piecewise uniform (lower left panels). Similar to Figure 5 but on a logarithmic scale. The axes are limited to 10^{-12}, not showing SA values below this limit. The two numbers in the lower left of each panel are the rank correlation between all experts, and the rank correlation between all experts with a greater than 0.05 SA in both test. Diagonal plots present the histogram of each measure's statistical accuracy for all 530 experts (i.e., the marginal of each scatter plot). In each histogram, the percentage of experts with a > 5% significance level is reported.

correspond to Figure A2a. It shows the probability (*p*-value) that each of the samples is lower than the other sample. For example, the statistically significant *p*-value of 0.025 for $DM_{CRPS} < DM_{CvM}$ is represented by the *CvM* (red) box plot showing higher values than the *CRPS* (orange) box plot. Figure A2 f, k, and p correspond to the remaining rows in Table 2.

The *p*-values of the rank comparison evaluated for the other measures of statistical accuracy are shown in Table A1. These correspond to the remaining panels in Figure A2.

Finally, Figure A3 is equal to Figure 5 but on a logarithmic scale. This shows the behavior of the measures to experts that score a very low statistical accuracy.

Appendix B
Metalog distribution

1. Information score for the Metalog distribution

In the Classical Model, the information score compares the probability density distribution $f_{e,i}$ for item i elicited from expert e, to a background density g_i. This background density is uniform across the intrinsic range $[L^*, U^*]$. This range is defined for each item by collecting all experts' estimates and the realization. The minimum and maximum of this set form the lower bound L and upper bound U. An overshoot k (typically 0.1) is than added to obtain $[L^*, U^*] = [L - k(U - L)/100, U + k(U - L)/100]$. Typically, the 5th,

50th, and 95th percentiles are elicited. This creates a probability vector with 4 quantile intervals, $p = (0.05, 0.45, 0.45, 0.05)$. When assuming the piecewise uniform distribution, the expert density $f(e, i)$ is composed of a uniform distribution between each subsequent pair of values in the vector X: $(L^*, x_{0.05}, x_{0.50}, x_{0.95}, U^*)$. The information score is calculated by comparing the interquantile range to the bin size p_i:

$$I_{PWU}(e, i) = \log(U^* - L^*) + \sum_{i=1}^{4} \left[p_i \cdot \log\left(\frac{p_i}{x_{i+1} - x_i} \right) \right] \quad (8)$$

If an expert would estimate values for the three percentiles that result in a uniform distribution on $[L^*, U^*]$ the information score would be zero. Any deviation from this results in an information score above zero.

When assuming the Metalog distribution, the interquantile probability is not uniformly distributed. The information score is therefore calculated by integrating the expert density $f(e, i)$ over the range $[L^*, U^*]$:

$$I_{ML}(e, i) = \log(U^* - L^*) + \int_{x=L^*}^{U^*} \left[p(x) \cdot \log\left(\frac{p(x)}{dx} \right) \right] dx \quad (9)$$

In which $p(x)$ is the uniform background probability density in the range $[L^*, U^*]$. However, the unbounded version of the Metalog ranges from $-\infty$ to ∞ and is thus not limited to $[L^*, U^*]$. Using the infinity range would lead to a probability density of zero for the background probability and an infinite informativeness for the expert. Therefore, only the range $[L^*, U^*]$ is considered for calculating the informativeness. Because of the need for limits on the background probability density, a choice for the overshoot is still required when using the Metalog to interpolate expert percentile estimates.

2. Method for fitting a distribution to varying percentiles

Not all three-percentile or five-percentile expert estimates result in feasible Metalog distributions (i.e., $f(x) > 0$ for all x, with $f(x)$ being the PDF of x). For symmetric three-percentiles cases, the constraints for a feasible Metalog are given by $a_2 > 0$ and $|a_3|/a_2 < 1.66711$. For an unbounded Metalog distribution with a 5th percentile value of 10, and a 95th percentile value of 90, the median should be in between 20 and 80. If this is not the case, a feasible distribution can be achieved be imposing a lower or upper bound such that the constraints are met. This leads to highly skewed distributions, with one bound and one very thick tail. For the 3-percentile case, the steps in fitting a Metalog distribution are:

1. Check whether the linear least squares fitted a-vector is feasible.
2. If not, determine whether the expert estimate is left or right skewed. Find the lower and upper limit of the of the lower bound (left skewed) or upper bound (left skewed) that meets the constraints (i.e., $a_2 > 0$ and $|a_3|/a_2 < 1.66711$).
3. Iterate towards the bound that results in the distribution with the lowest maximum probability density (i.e., the least informative distribution).

For the five-percentile case it is more difficult to obtain a feasible fit. This is partly due to many expert estimates being more or less uniform in between the 5th and 95th percentile, such as shown in Figure 3e. For nonsymmetric cases this often leads to not finding a feasible solution using a 5-term a-vector. Figure 3h. is an example of this. In some cases, a solution is to impose a bound on one side just like for the skewed three-percentile cases. However, this does not work in all cases. Another solution is to use two three-percentile Metalog distribution, one representing the 5th, 25 h and 50th percentiles, and one representing the 50th, 75th and 95th percentiles. This leads to a discontinuity in probability density at the median, which can be removed by imposing bounds on one (or both) of the 3-percentile Metalog parts (e.g., adding an upper bound to the left part in Figure 3g). We chose not to do this as it is only an aesthetic solution, which affects the tail probabilities and tends to create spikes in the probability density functions. The steps for fitting a five-percentile Metalog are therefore:

1. Check whether the linear least squares fitted a-vector is feasible.
2. If not, find two feasible three-percentile Metalog distributions (following the steps above).
3. Optionally, remove the discontinuity in probability density at the median, by iteratively shifting the lower or upper bound of the right or left distribution (whichever has the lowest probability density at the median) until the gap is removed.