Averaging quantiles, variance shrinkage, and overconfidence

Cooke, Roger M.

DOI
10.1002/ffo2.139

Publication date
2022

Document Version
Final published version

Published in
Futures and Foresight Science

Received: 20 March 2022 | Revised: 20 May 2022 | Accepted: 22 June 2022

DOI: 10.1002/ffo2.139

ORIGINAL ARTICLE

WILEY

Averaging quantiles, variance shrinkage, and overconfidence

Roger M. Cooke

Department of Mathematics, Resources for the Future, TU Delft, Delft, the Netherlands

Correspondence
Roger M. Cooke, Resources for the Future, TU Delft, Delft 2517KH, the Netherlands.
Email: Cooke@rff.org

Abstract

Averaging quantiles as a way of combining experts' judgments is studied both mathematically and empirically. Quantile averaging is equivalent to taking the harmonic mean of densities evaluated at quantile points. A variance shrinkage law is established between equal and harmonic weighting. Data from 49 post-2006 studies are extended to include harmonic weighting in addition to equal and performance-based weighting. It emerges that harmonic weighting has the highest average information and degraded statistical accuracy. The hypothesis that the quantile average is statistically accurate would be rejected at the 5% level in 28 studies and at the 0.1% level in 15 studies. For performance weighting, these numbers are 3 and 1, for equal weighting 2 and 1.

KEYWORDS

averaging distributions, averaging quantiles, combing experts, expert judgment, over-confidence, variance shrinkage

1 | INTRODUCTION

Suppose one elicits cumulative distribution functions (*cdf*'s) $F_1, \ldots F_n$ and/or probability density functions (*pdf*'s) $f_1, \ldots f_n$ from n experts. What should one do with this information? Some argue against combining the distributions unless necessary for policy (Morgan, 2014, Morgan et al., 2009). The equally weighted combination of *cdfs*, $EW(x) = (F_1(x) + \ldots F_n(x))/n$ is the legacy method. Geometric averaging, or Geometric Weighting $GW(x) = \int_{z \le x} \Pi\, f_i(z)^{1/n} dz / \int \Pi\, f_i(u)^{1/n} du$ has been advocated as being "independence preserving" (Laddaga, 1977) and "externally Bayesian" (Genest & Zidek, 1986). Geometric averaging tends to concentrate mass in regions where the experts agree. This tendency is more pronounced with harmonic averaging or harmonic weighting (*HW*). *HW* has found recent adherents who propose quantile averaging as an alternative to *EW*. As shown below, averaging quantiles is equivalent to harmonically averaging densities at the quantile points.

These solutions all require complete *cdf*'s. When only fixed percentiles, or quantiles, of each distribution, say 5, 50, and 95 percentiles, are given, the above solutions require imputing distribution functions based on the elicited quantiles. Popular approaches are fitting a parametric distribution (O'Hagan et al., 2006) or minimizing information subject to quantile constraints relative to a background support (Cooke, 1991). Averaging quantiles is much simpler; one simply averages the 5 percentiles, the 50 percentiles, and the 95 percentiles. There is no need to impute a distribution. Although not attested in any guidance of which the author is aware, it is often employed as a way of summarizing data without introducing additional assumptions. It has been adopted by the COVID-19 ForecastHub (https://covid19forecasthub.org/doc/ensemble/; Cramer et al., 2021; Ray et al., 2020). Examples of others using quantile averaging include (Christensen et al., 2018; De Gooijer & Dawit, 2019; Flandoli et al., 2011; Kim et al., 2021; Sayedi et al., 2020; de Vries & van de Wal, 2015). It has been promoted as an alternative to equal weighting as horizontal averaging as opposed to vertical averaging (Lichtendahl et al., 2013).

Here, mathematical and empirical properties of quantile averaging are examined. The next section shows that quantile averaging of distributions is equivalent to harmonically averaging their densities at

the quantile points (taken from Bamber et al., 2016; Colson & Cooke, 2017). and derives a variance shrinkage law: Defining *Ave Var* = $(1/n)\Sigma\ Var(F_i)$, simple calculations show:

$$Var\ EW\ = Var \text{ of means} + \text{Ave } Var \geq \text{Ave } Var \geq Var\ HW$$

The conditions for equality are different for the two inequalities.

Variance shrinkage raises the question whether *HW* invites overconfidence. A database of 49 post-2006 studies (Cooke et al., 2021) has been extended to include *HW* combinations for each study. Section 3 contains a comparison of *PW* (item-specific performance-based weighting), *EW*, and *HW* at the study level. The following picture emerges: Whereas *PW* and *EW* as statistical hypotheses would be rejected at the 5% level on 3 resp 2 of the 49 studies, HW is rejected on 28 (57%) studies. On 15 (31%) studies rejection is at the 0.1% level. *HW's* informativeness on average exceeds that of *EW* and is comparable to that of *PW*. Section 3 gives results and examines whether study parameters could predict the poor statistical performance of *HW*. Section 4 shows that *HW* is appropriate when interpolating, as opposed to combining, distributions. A final section gathers conclusions. Supporting Information gives mathematical details. All data and code are available from the author on request.

2 | METHODS

Let *F* and *G* be continuous invertible *cdf*'s from experts 1 and 2, with densities *f*, *g*. Let *HW*, *hw* denote respectively the *cdf* and *pdf* of the result of averaging the quantiles of *F*, *G*:

$$HW^{-1}(r) = 1/2(F^{-1}(r) + G^{-1}(r)) \tag{1}$$

A good intuitive interpretation (Andrea Bevilacqua, personal communication) notes that *HW* takes the average of the experts' median values and a confidence interval (CI) whose width is the average of the experts' CIs. The position of the median within the CI depends on the distributions.

To gain further insight into Equation (1), take derivatives of both sides:

$$1/hw(HW^{-1}(r)) = 1/2(1/f(F^{-1}(r)) + 1/g(G^{-1}(r))) \tag{2}$$

$$hw(HW^{-1}(r)) = \frac{2}{1/f(F^{-1}(r)) + 1/g(G^{-1}(r))} \tag{3}$$

Equation (3) says that *hw* is the harmonic mean of *f* and *g*, evaluated at points corresponding to the *r*-th quantile of each distribution. The harmonic mean of *n* numbers strongly favors the smallest of these numbers: the harmonic mean of 0.01 and 0.99 is 0.0198, the geometric mean is 0.099 and the average is 0.5. To appreciate the effect of this, consider a flexible and tractable class of distributions on the unit interval:

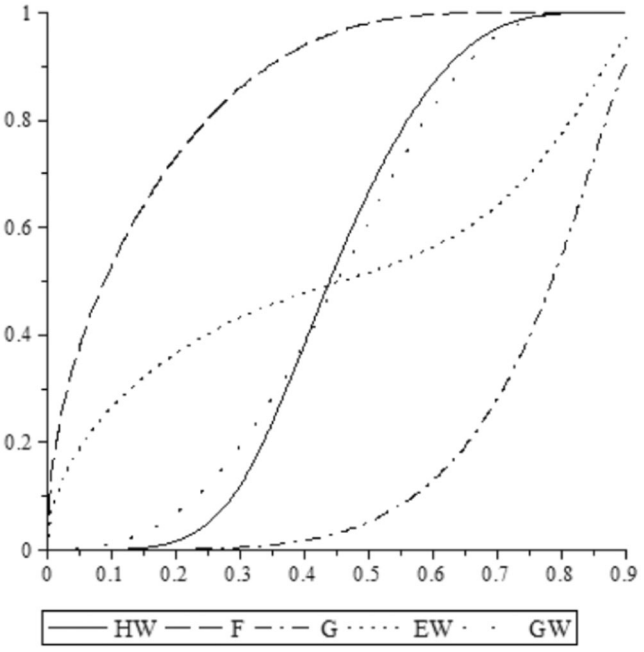

FIGURE 1 *F*(*a* = 5, *b* = 0.5), *G*(*a* = 5, *b* = 5), *HW* = quantile average, *EW* = Arithmetic average of distributions, *GW* = geometric average of distributions.

$$F(x) = 1 - a^{-\frac{x^b}{1-x^b}};\ F^{-1}(r) = \left(-\frac{ln(1-r)}{ln(a)} \cdot \left(1 - \frac{ln(1-r)}{ln(a)}\right)^{-1}\right)^{\frac{1}{b}}$$

$$a > 1;\ b > 0 \tag{4}$$

Figure 1 shows two expert distributions from this class, *F* and *G*, and also shows *HW*, *EW*, and *GW*.

For each *x* on the horizontal axis, the slope of *HW*(*x*) is close to the smaller of the slopes of *F*(*x*) and *G*(*x*); causing *HW*(*x*) to grow slowly for small and large *x*, resulting in a concentrated distribution. *EW* in contrast has a much wider CI. Note that *HW* is more concentrated than *GW*.

Variance shrinkage is based on the Cauchy Schwarz inequality: for any $x,y \in \mathbb{R}^n$, $(\Sigma x_i^2)(\Sigma y_i^2) \geq (\Sigma x_i y_i)^2$ with equality if and only the x_i and y_i are proportional. Putting $y_i = 1$, this says

$$n\sum x_i^2 \geq \left(\sum x_i\right)^2 = \sum_{ij} x_i x_j \tag{5}$$

with equality if and only if the x_i are equal.

The *cdf* of the quantile average of random variables $Y_1, \ldots Y_n$ with continuous invertible *cdf*'s is the *cdf* of $HW = (1/n)\Sigma X_i$ when the X_i has the same *cdf* as Y_i and all X_i have rank or Spearman correlation $r(X_i, X_j) = 1$. The joint distribution of $(X_1, \ldots X_n)$ is such that if values $x_{1, \ldots} x_n$ are sampled and if x_1 realizes the *q*th quantile of X_1, them, since all variables are completely rank correlated x_i realizes the *q*th quantile of X_i, $i = 2, \ldots n$. Hence *HW* averages the quantiles of $Y_1, \ldots Y_n$.

Although the X_i are completely rank correlated, their product moment correlation ρ need not be 1. If $r(X_i, X_j) = 1$ then $X_i = \phi(X_j)$ for some strictly monotonic transformation ϕ, whereas $\rho(X_i, X_j) = 1$ if and only if $X_i = aX_j + b$ for some positive *a* and some $b \in \mathbb{R}$. If *U* is uniform on (0,1),

then $r(U,U^{10}) = 1$ but $\rho(U,U^{10}) = 0.66$. From the Pearson formula[1] relating rank and product normal correlations for two normal variables we infer that $\rho(X_i, X_j) = 1$ if and only if $r(X_i, X_j) = 1$ for normal variables X_i, X_j.

If the X_i have means μ_i and variances σ_i^2 it follows that

$$Var(HW) = (1/n^2)\left[\sum\sigma_1^2 + \sum_{i\neq j} C_{ij}\right];\ C_{ij} = Cov(x_i, x_j) \tag{6}$$

Equation (6) entails that *Var(HW)* does not depend on the means and therefore is invariant under adding arbitrary location parameters to the variables. Pithily put, the uncertainty of *HW* does not depend on how near or far apart the variables are.

Proposition 1. $(1/n)\Sigma\sigma_i^2 \geq$ Var(HW) *with equality if and only if the* σ_i^2 *are all equal and* $\rho(X_i, X_j) = 1$.

pf: $(1/n)\sum\sigma_i^2 - Var(HW)$

$$\begin{aligned} &= (1/n)\sum\sigma_1^2 - (1/n^2)\left[\sum\sigma_1^2 + \sum_{i\neq j} C_{ij}\right] \\ &= \left[(n-1)\sum\sigma_1^2 + \sum_{i\neq j} C_{ij}\right]/n^2 = \left[n\sum\sigma_1^2 + \sum_{ij} C_{ij}\right]/n^2 \end{aligned} \tag{7}$$

where $C_{ii} = \sigma_i^2$. $\rho(X_i, X_j) = C_{ij}/\sigma_i\sigma_j \leq 1$ with equality if and only if $X_i = aX_j + b$, $a_i > 0$, $b \in \mathbb{R}$. Therefore, with (5)

$$\sum_{i,j} C_{ij} \leq \sum_{i,j} \sigma_i\sigma_j \leq n\sum\sigma_1^2 \tag{8}$$

so that the shrinkage $[n\Sigma\sigma_i^2 - \Sigma_{ij}C_{ij}]/n^2$ is non-negative. The first inequality in Equation (8) holds with equality if and only if $\rho(X_i, X_j) = 1$ while the second holds if and only if the σ_i are equal. □

For variables with unit product-moment correlation, the first inequality always holds with equality in Equation (8), but not the second. Standardizing a variable by dividing by its standard deviation gives the variable unit variance. Standardized versions of *U* and U^{10} are completely rank correlated but the shrinkage is 17% (see Figure 2 left panel).

A similar shrinkage formula based on the means characterizes the difference between the variance of an equally weighted combination of distributions and the average variance. For variables $X_1, \ldots X_n$, with densities $f_1, \ldots f_n$, variances σ_i^2 and means μ_i let *EW* denote the distribution with density $(1/n)\Sigma f_i$. We have

Proposition 2. $Var(EW) - (1/n)\sum\sigma_i^2 = [n\sum\mu_i^2 - \sum_{ij}\mu_i\mu_j]/n^2 \geq 0$.

pf: $Var(EW) = \int x^2\left(\sum f_i(x)/n\right)dx - \sum\mu_i^2/n + \sum\mu_i^2/n - \left(\sum\mu_i/n\right)^2$

$$\begin{aligned} &= (1/n)\sum\sigma_i^2 + (1/n)\sum\mu_i^2 - \left[\sum_{ij}\mu_i\mu_j\right]/n^2 \\ &= (1/n)\sum\sigma_i^2 + \left[n\sum\mu_i^2 - \sum_{ij}\mu_i\mu_j\right]/n^2 \end{aligned} \tag{9}$$

The last term is non-negative by the Cauchy Schwarz inequality and equals 0 if and only if the μ_i are equal □

We recognize Equation (9) as the mean of the variances of the F_i plus the variance of the means of the F_i. For the special case $n = 2$, Equation (9) becomes

$$Var(EW) = \frac{(\sigma_1^2 + \sigma_2^2)}{2} + \frac{(\mu_1 - \mu_2)^2}{4} \tag{10}$$

Figure 2 compares powers of uniforms with unit variance (left panel) and normals with unit variance (right panel). The shrinkage *Ave Var – Var HW* on the left is due to the differences between rank and product-moment correlation, while that of *Var EW – Ave Var* on the right is due to differences in means. The conditions for equality are different for the above propositions, but we can put them together to define a *total shrinkage*

$$\begin{aligned} \text{total shrinkage} &= Var(EW) - Var(HW) \\ &= \left[n\sum\mu_i^2 - \sum_{ij}\mu_i\mu_j + n\sum\sigma_i^2 - \sum_{ij} C_{ij}\right]/n^2 \end{aligned} \tag{11}$$

Figure 2 suggests that when experts' central masses have little overlap, the shrinkage from Equation (9) can be quite severe.

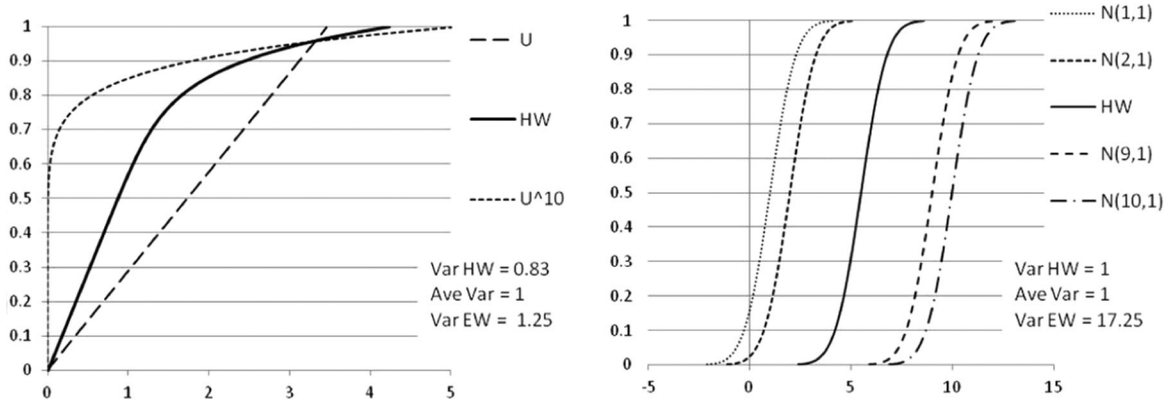

FIGURE 2 Left panel, *cdf*'s of powers of uniform variables standardized to have unit variance; Right panel, normal variables $N(\mu,\sigma^2)$ with unit variance. The Quantile average *HW* is shown on both panels.

TABLE 1 Results from 49 post-2006 structured expert judgment studies

	PW			EW			HW				
	SA	inf	comb	SA	inf	comb	SA	inf	comb	#calib vbls	#exprts
Arkansas	0.50	0.34	0.17	0.39	0.20	0.08	5.55E−02	0.64	3.55E−02	10	4
Arsenic D-R	0.04	2.74	0.10	0.06	1.10	0.07	7.99E−04	1.32	1.06E−03	10	9
ATCEP Error	0.68	0.23	0.16	0.12	0.25	0.03	5.99E−04	1.07	6.38E−04	10	5
Biol agents	0.68	0.61	0.41	0.41	0.24	0.10	3.60E−02	0.88	3.18E−02	12	12
CDC ROI	0.72	2.31	1.66	0.23	1.23	0.29	7.56E−01	1.57	1.18E+00	10	20
CoveringKids	0.72	0.43	0.31	0.63	0.27	0.17	9.03E−01	0.60	5.38E−01	10	5
CREATE	0.39	0.28	0.11	0.06	0.21	0.01	2.77E−04	0.52	1.44E−04	10	7
CWD	0.49	1.22	0.60	0.47	0.93	0.44	7.07E−01	1.49	1.06E+00	10	14
Daniela	0.55	0.63	0.35	0.53	0.17	0.09	1.82E−01	0.52	9.48E−02	7	4
dcpn_fistula	0.12	1.31	0.16	0.06	0.62	0.04	8.78E−08	1.13	9.88E−08	10	8
eBBP	0.83	1.41	1.17	0.36	0.32	0.11	8.04E−02	0.95	7.67E−02	15	14
EffusiveErupt	0.66	1.12	0.75	0.29	0.80	0.23	2.65E−02	1.51	3.99E−02	8	14
Erie Carps*	0.66	0.86	0.57	0.18	0.28	0.05	3.87E−01	0.75	2.92E−01	15	10
FCEP Error	0.66	0.57	0.38	0.22	0.10	0.02	1.75E−05	0.77	1.35E−05	8	5
Florida	0.76	1.13	0.86	0.76	0.46	0.34	6.98E−02	0.88	6.15E−02	10	7
GL-NIS	0.93	0.21	0.19	0.04	0.31	0.01	5.53E−02	0.84	4.66E−02	13	9
Gerstenberger	0.93	1.10	1.02	0.64	0.48	0.31	8.10E−02	0.97	7.82E−02	14	12
Goodheart	0.71	0.96	0.68	0.55	0.28	0.15	6.83E−01	0.89	6.07E−01	10	6
Hemophilia	0.31	0.49	0.15	0.25	0.20	0.05	3.12E−01	0.78	2.43E−01	8	18
IceSheet2012	0.40	1.55	0.62	0.49	0.52	0.25	7.96E−02	1.20	9.56E−02	11	10
Illinois	0.34	0.65	0.22	0.62	0.26	0.16	2.37E−03	0.79	1.88E−03	10	5
Liander	0.23	0.52	0.12	0.23	0.48	0.11	2.81E−03	1.20	3.36E−03	10	11
Nebraska	0.03	1.45	0.05	0.37	0.70	0.26	2.40E−05	1.19	2.86E−05	10	4
Obesity	0.44	0.51	0.22	0.07	0.24	0.02	6.68E−04	0.74	4.98E−04	10	4
PHAC T4	0.18	0.35	0.06	0.30	0.21	0.06	2.02E−02	0.70	1.41E−02	13	10
San Diego*	0.15	0.76	0.12	0.15	1.01	0.15	3.02E−03	1.58	3.32E−02	10	8
Sheep Scab	0.64	1.31	0.84	0.66	0.78	0.52	1.15E−02	1.41	1.63E−02	15	14
SPEED	0.68	0.78	0.53	0.52	0.75	0.39	2.97E−02	1.17	3.46E−02	16	14
TdC	0.99	1.26	1.24	0.17	0.36	0.06	1.24E−02	1.08	1.34E−02	17	18
Tobacco	0.69	1.06	0.73	0.20	0.45	0.09	2.11E−01	0.71	1.49E−01	10	7
Topaz	0.41	1.46	0.60	0.63	0.92	0.58	8.66E−05	1.53	1.32E−04	16	21
umd_nremoval	0.71	1.99	1.40	0.07	0.80	0.05	2.40E−03	1.22	2.93E−03	11	9
Washington	0.20	0.72	0.14	0.15	0.53	0.08	4.21E−01	0.86	3.63E−01	10	5
GeoPol	0.42	1.15	0.49	0.20	0.56	0.11	5.02E−06	1.28	6.43E−05	16	9
BFIQ	0.69	0.57	0.40	0.42	0.29	0.12	1.15E−02	0.67	7.78E−03	11	7
IQEarn	0.70	0.62	0.44	0.70	0.57	0.41	4.54E−01	0.90	4.09E−01	11	8
USGS	0.51	1.51	0.77	0.06	0.80	0.05	4.49E−04	1.54	6.90E−04	18	32
UK	0.22	0.66	0.14	0.13	0.33	0.04	1.19E−01	0.78	9.31E−02	10	6

TABLE 1 (Continued)

	PW			EW			HW				
	SA	inf	comb	SA	inf	comb	SA	inf	comb	#calib vbls	#exprts
Spain	3.59E−05	0.69	0.00	1.22E−05	0.23	0.00	1.96E−08	0.80	1.56E−08	10	5
Italy	0.45	0.47	0.21	0.22	0.20	0.04	1.25E−01	0.49	6.11E−02	10	4
France	0.65	1.96	1.28	0.08	0.43	0.03	2.66E−02	0.92	2.44E−02	10	5
all_CDC	0.97	2.54	2.46	0.25	1.08	0.27	2.06E−04	1.74	3.58E−04	14	48
Puig-GDP	0.93	0.99	0.92	0.06	0.43	0.03	5.41E−04	1.25	6.75E−04	13	9
Puig-oil*	0.13	0.61	0.08	0.88	0.20	0.18	2.23E−10	1.07	2.38E−10	20	6
PoliticalViolence*	0.13	1.82	0.23	0.44	1.05	0.46	1.73E−07	1.73	8.19E−16	21	16
Brexit food	0.55	0.84	0.46	0.11	0.27	0.03	7.07E−01	1.26	8.88E−01	10	10
Tadini Quito	0.93	0.85	0.79	0.42	0.23	0.10	2.02E−02	0.95	1.92E−02	13	8
Tadini Clermont	0.75	1.14	0.86	0.33	0.28	0.09	9.28E−01	0.28	2.63E−01	13	12
ICE_2018	0.94	0.93	0.87	0.13	0.55	0.07	8.97E−02	1.22	0.11	16	20
Ave	**0.54**	**1.01**	**0.55**	**0.31**	**0.49**	**0.15**	**0.16**	**1.03**	**0.14**		
Geomean	**0.37**			**0.19**			**5.1E−03**				
#SA < 0.05	**3**			**2**			**28**				
#SA < 0.001	**1**			**1**			**15**				
# Best			**40**			**5**			**4**		

Note: "SA" denotes statistical accuracy, "Inf" denotes informativeness, and "comb" denotes the product of these two. Statistical accuracy is the *p*-value at which the hypothesis of statistical accuracy would be falsely rejected. Informativeness is Shannon relative information with respect to a background measure. The product of these two is an asymptotic strictly proper scoring rule for average probabilities. Details for scoring are in Colson and Cooke (2017) and Cooke et al. (2021). Numbers of experts and calibration variables are shown. Asterisks denote studies in which one or more experts did not assess all calibration variables. Studies with bolded names were the 33 studies analyzed in detail in Colson and Cooke (2017).

3 | RESULTS

The TU Delft expert judgment database contains 49 studies since 2006 involving 530 experts assessing, in addition to the variables of interest, 580 calibration variables from their field to which true values were known. Of these, 140 experts (26%) would not be rejected as statistical hypotheses at the traditional 5% level. The study compares *EW* and performance-weighted combinations (*PW*) in which experts' distributions are weighted according to their statistical accuracy and informativeness (see Cooke [1991]; an updated exposition is in Colson and Cooke [2017]; for references see Cooke et al. [2021] and Supporting Information). For the present study, the *HW* combinations have been added for each study. Four studies (with asterisks in Table 1) involved experts who did not answer all calibration variables. These experts were dropped, causing the numbers in those studies to differ somewhat from those in Cooke et al. (2021). For comparing the three combination schemes *PW*, *EW*, and *HW* this is immaterial.

The mean statistical accuracy scores of all three combinations are above the traditional 5% rejection threshold for simple hypothesis testing (for the geomean or geometrical average this holds only for *PW* and *EW*). In 28 of the 49 studies (57%), *HW* would be rejected at the 5% level, and on 15 (31%), rejection would be at the 0.1% level. This contrasts with *EW* and *PW* where 2 resp. 3 combinations would be rejected at the 5% level. On average, *HW's* informativeness was substantially greater than *EW's* and slightly better than *PW's*. *PW* has the highest combined score (the product of statistical score and informativeness) in 40 studies, *EW* in 5 studies, and *HW* in 4 studies (this is an in-sample comparison with *PW*, for out-of-sample comparisons, see Colson and Cooke [2017]; Cooke et al. [2021] and Supporting Information). The combined score is an asymptotic strictly proper scoring rule for average probabilities.

Statistical accuracy and informativeness are metrics for measuring performance as uncertainty assessors. Forecast accuracy based on medians is also important. The relative forecast error of various combination schemes was extensively studied by Cooke et al. (2021) from which the following information is extracted (Table 2). The variations of performance weighted combinations are explained in the Supporting Information.

As quantile averaging is often used without calibration variables, it could be of interest to anticipate poor statistical performance of quantile averaging based only on study characteristics without reference to the true values. The variance shrinkage laws are suggestive but when variables are measured in different physical units, scale-invariant tools are required. The Spearman rank correlation matrix of *HW* statistical accuracy with study characteristics (Table 3) does not show strong relationships. The number of experts

and number of calibration variables are rank correlated in this data set at 0.53; indeed, studies with modest budgets tend to follow the guidance of 10 calibration variables and at least 4, preferably six experts. Better resourced studies can afford to raise both numbers.

From Table 3, neither the number of calibration variables nor the number of experts exerts a strong influence on the statistical accuracy of the quantile average. However, each tends to have a negative impact on *HW*'s statistical accuracy. A possible explanation is that harmonic averaging leans heavily towards the smallest value of the densities. This would explain the negative correlation with Max Inf as this concentrates the mass of *HW* in a smaller region. Adding more experts increases the chance that one will have very high information and that will shrink the bands of *HW*. Both Max Inf and #experts correlate positively with #calib vbls.

To appreciate the problems, Figures 3 and 4 show range graphs for two studies. For each calibration variable, the experts' 90% CIs are shown as horizontal lines and the medians as dots. The bottom CIs are those of *HW*. The realization is shown as a red vertical line. *IQearn* has one of the best performances for *HWSA* whereas *puig-oil* has one of the poorest. In both studies, the *PW* and *EW* have good statistical accuracy (see captions). Both studies have nonoverlapping confidence bounds. This has the effect of increasing the support of the uniform background measures relative to the size of the CIs and thus increasing the average informativeness of the experts. Indeed, a CI of [5, 6] looks more informative against a background of [1, 100] than against [1, 10]. The average information for *IQearn* is 1.29 while that of *puig-oil* experts is 1.25. The key difference is the placement of the realization (vertical red line) relative to the experts' assessments. That, of course, cannot be inferred from study characteristics. Without knowing the realizations, it is impossible to anticipate the poor performance of *HW* for *puig-oil*.

TABLE 2 Average and standard deviation of absolute dimensionless forecast errors for item-specific performance weights (PWi), global performance weights (PWg), non-optimized global performance weights (PWn), equal weights (EW), performance weighted average of medians (PWQ), and equal-weighted average of medians (EWQ), and corresponds to *HW*

	|(PWi − rls)/rls|	|PWg − rls|/rls	|PWn − rls|/rls	|(EW − rls)/rls|	|PWQ − rls|/rls	|(EWQ − rls)/rls|
Ave	2.2	2.7	2.3	3.8	278.6	1472.3
Stdev	11.8	16.0	14.7	45.2	5646.8	33,299.8
Geomean	0.38	0.40	0.37	0.43	0.42	0.63

Note: "rls" denotes "realization," the true values of the random variables.

TABLE 3 Rank correlation matrix for harmonic weighting. Max Inf is the maximal information score of an expert in a panel.

Spearman rank correlation matrix *HW*			
	#calib vbls	#experts	Max Inf
HW Stat. accuracy	−0.15	−0.09	−0.25
#calib vbls		0.53	0.38
#experts			0.62

4 | WHEN QUANTILE AVERAGING IS APPROPRIATE: INTERPOLATING VERSUS COMBINING

Rather than combining distributions over a single uncertain variable, we are often confronted with situations in which we must interpolate distributions at different values of some underlying

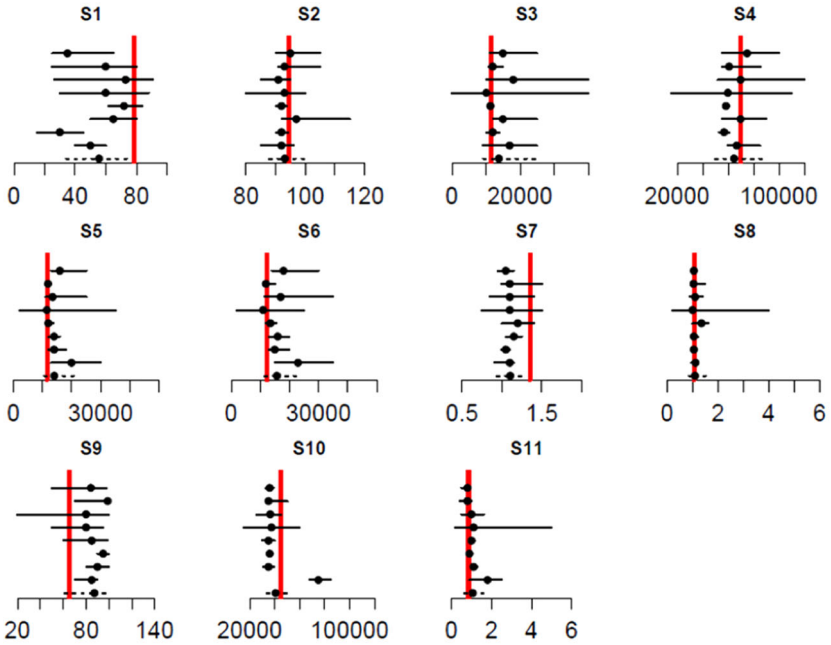

FIGURE 3 Range graphs for the case *IQearn*. Experts' [5%, 95%] confidence intervals are given as horizontal lines, medians as dots, and the realization is given as a red vertical line. *HW* is added as ninth expert at the bottom of each graph. In this case, the statistical accuracies are: *PW* = 0.7, *EW* = 0.7, *HW* = 0.45. The experts' average information with respect to the uniform background is 1.29.

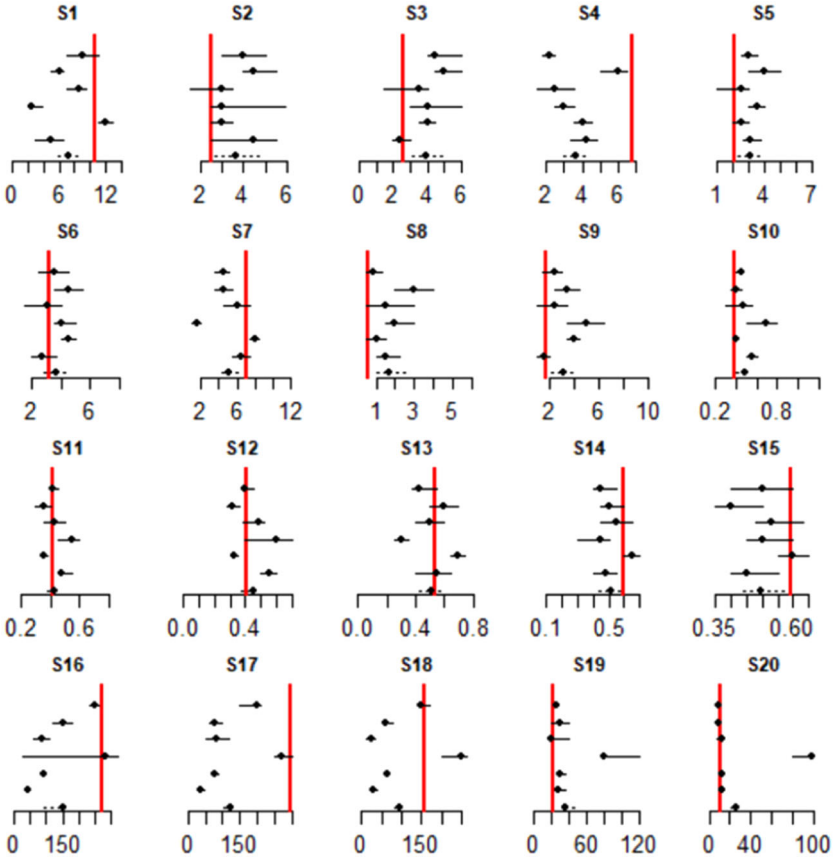

FIGURE 4 Range graphs for the case *puig-oil*. Experts' [5%, 95%] confidence intervals are given as horizontal lines, medians as dots, and the realization is given as a red vertical line. *HW* is added as seventh expert at the bottom of each graph. In this case, the statistical accuracies were: *PW* = 0.13, *EW* = 0.88, *HW* = 2.23E−10. The experts' average information with respect to the uniform background is 1.25.

parameter. Oppenheimer et al. (2016) discuss an application in which experts quantify uncertainty in crosswind dispersion of an airborne pollutant for different downwind distances. According to the standard Gaussian plume model, the crosswind standard deviation of the time-integrated concentration at downwind distance x is $\sigma_c(x) = ax^b$ for (poorly constrained) constants a, b, (a, $b > 0$). Suppose experts quantify their uncertainty in $\sigma_c(x)$ for $x = 10$ km, and 20 km. Barring exceptional circumstances, the uncertainty $\sigma_c(x)$ increases with x.

Suppose we want the distribution for $\sigma_c(15)$. If we take an equal weight combination of the distributions of $\sigma(10)$ and $\sigma(20)$ we may well find that the result has greater variance than that of $\sigma(20)$. The variance shrinkage laws allow us to see exactly when that happens. Put $n = 2$, $Var(\sigma(10)) = V_1$, $Var(\sigma(20)) = V_2$, with means μ_1, μ_2. For the equal weight combination of the uncertainties in $\sigma(10)$ and $\sigma(20)$ Equation (10) says:

$$Var(EW) = (V_1 + V_2)/2 + (\mu_1 - \mu_2)^2/4 > V_2 \Leftrightarrow V_1 + (\mu_1 - \mu_2)^2/2 > V_2 \quad (12)$$

Such an outcome would be unacceptable. By the same token, Equation (7) says that the variance of *HW* is always less than or equal to the average of the variances of $\sigma_c(10)$ and $\sigma_c(20)$ with equality holding in case these distributions are normal with the same variance. These remarks apply mutatis mutandis when interpolating at other distances between 10 and 20 km. In cases of interpolation like the above, quantile averaging provides a reasonable solution, whereas equal weighting of distributions does not.

5 | CONCLUSION

If all experts say the same thing, then the three schemes considered here are all equivalent. Data show, however, that there is a great deal of variation in experts' assessments and in their performance. Accordingly, there is great variation in the performance of expert combinations. Cherry-picked studies can produce very different conclusions. Reliable conclusions should therefore be based on a large set of studies of known provenance. With regard to *HW*, we may conclude that it achieves higher informativeness at the expense of statistical accuracy. In 57% of the studies, this results in over-confidence, in 31% the overconfidence is severe. The forecast error of averaging medians is, in aggregate, much larger than that of *EW* or *PW*. However, when we are interpolating between distributions, rather than combining them, quantile averaging would seem appropriate.

ACKNOWLEDGMENTS

The author gratefully acknowledges discussions with Prof. Tina Nane and many improvements suggested by anonymous referees. All expert judgment data are freely available at http://rogermcooke.net/

CONFLICT OF INTEREST

The author declares no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Roger M. Cooke http://orcid.org/0000-0003-0643-1971

ENDNOTE

[1] For normal variables $\rho = 2\times sin(r \times \pi/6)$.

REFERENCES

Bamber, J. L., Aspinall, W. J., & Cooke, R. M. (2016). "A commentary on 'how to interpret expert judgment assessments of 21st century sea-level rise' by Hylke de Vries and Roderik SW van de Wal". *Climatic Change*, *137*, 321–328. https://doi.org/10.1007/s10584-016-1672-7

Christensen, P., Gillingham, K., & Nordhaus, W. (2018). Uncertainty in forecasts of long-run economic growth. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(21), 5409–5414. https://doi.org/10.1073/pnas.1713628115

Colson, A., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering and System Safety*, *163*, 109–120. https://doi.org/10.1016/j.ress.2017.02.003

Cooke, R. M. (1991). *Experts in uncertainty; Opinion and subjective probability in science*. Oxford University Press. 321 pages.

Cooke, R. M., Marti, D., & Mazzuchi, T. A. (2021). Expert forecasting with and without uncertainty quantification and weighting: What do the data say? *International Journal of Forecasting*, *published online*, *37*(1), 378–387. https://doi.org/10.1016/j.ijforecast.2020.06.007

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., & Jayawardena, D. (2021). Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the U.S. medRxiv. https://doi.org/10.1101/2021.02.03.21250974

De Gooijer, J. G., & Dawit, Z. (2019). Semiparametric quantile averaging in the presence of high-dimensional predictors. *International Journal of Forecasting*, *35*(3), 891–909.

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering and System Safety*, *96*, 1292–1310. https://doi.org/10.1016/j.ress.2011.05.012

Genest, C., & Zidek, J. (1986). Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, *1*(1), 114–148.

Kim, T., Fakoor, R., Mueller, J., Smola, A., & Tibshirani, R. J. (2021). *Deep quantile aggregation*. arXiv:2103.00083v2 [stat.ML]. March 16, 2021.

Laddaga, R. (1977). Lehrer and the consensus proposal. *Synthese*, *36*, 473–477.

Lichtendahl, K. C., Jr. Grushka-Cockayne, Y., & Winkler, R. L. (2013). Is it better to average probabilities or quantiles? *Management Science*, *59*(7), 1594–1611. https://doi.org/10.1287/mnsc.1120.1667

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20), 7176–7184. https://doi.org/10.1073/pnas.1319946111

Morgan, M. G., Dowlatabadi, H., Henrion, M., Keith, D., Lempert, R., McBride, S., Small, M., & Wilbanks, T. (2009). "Best practice approaches for characterizing, communicating, and incorporating scientific uncertainty in climate decision making" U.S. Climate Change Science Program. Synthesis and Assessment Product 5.2, January 2009.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., & Rakow, T. (2006). *Uncertain judgements; Eliciting expert's probabilities*. Wiley.

Oppenheimer, M., Little, C. M., & Cooke, R. M. (2016). Expert judgment and uncertainty quantification for climate change. *Nature*, *6*, 445–451. https://doi.org/10.1038/NCLIMATE2959

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, R. C., Prakash, B. A., Jr., & Reich, N. G. (2020). Ensemble forecasts of coronavirus disease (COVID-19) in the U.S. *medRxiv*. https://doi.org/10.1101/2020.08.19.20177493

Sayedi, S. S., Abbott, B., Thornton, B., Frederick, J., Vonk, J., Overduin, P., Schädel, C., Schuur, E., Bourbonnais, A., Demidov, N., Gavrilov, A., He, S., Hugelius, G., Jakobsson, M., Jones, M., Joung, D. J., Kraev, G., Macdonald, R., McGuire, A., & Frei, R. (2020). Subsea permafrost carbon stocks and climate change sensitivity estimated by expert assessment. *Environmental Research Letters*, *15*, 124075. https://doi.org/10.1088/1748-9326/abcc29

de Vries, H., & van de Wal, R. S. W. (2015). How to interpret expert judgment assessments of twenty-first century sea level rise. *Climate Change*, *130*, 87–100. https://doi.org/10.1007/s10584-015-1346-x

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Cooke, R. M. (2023). Averaging quantiles, variance shrinkage, and overconfidence. *Futures & Foresight Science*, 5, e139. https://doi.org/10.1002/ffo2.139