Supporting Information 1:

Improving expert forecasts in reliability.

Application and evidence for structured elicitation protocols:

Victoria Hemming, Nicholas Armstrong, Mark A. Burgman, Anca M. Hanea.

Quality and Reliability Engineering International

Table of contents

1 Data cleaning and conversion into non-parametric distributions

Data from Round 1 were cleaned and converted to 90% non-parametric distributions. This first involved converting ratios into their frequency formats. This was possible for all but one calibration question for which the raw values for either the denominator or numerator could not be obtained.

Each expert's bounds derived from the four-step elicitation were then extrapolated to 90% credible bounds using linear extrapolation, in which (equation 1.1 and 1.2):

Lower standardised bound:

$$\ell_e^{n,r} = b_e'^{n,r} - \left((b_e'^{n,r} - \ell_e'^{n,r}) * \left({}^{j}\!/_{m_e'^{n,r}} \right) \right) \tag{A.1}$$

Upper standardised bound:

$$u_e^{n,r} = b_e'^{n,r} + \left((u_e'^{n,r} - b_e'^{n,r}) * \left({}^{j}\!/_{m_e'^{n,r}} \right) \right) \tag{A.2}$$

where, b' =best estimate, ℓ' =lower bound estimate, u' = upper bound estimate, m' = level of confidence given by the participant *e*, in Round *r*, and *j*= the level of confidence each of the intervals was to be standardised to (i.e. 90%).

The lower estimate of the 90% standardised intervals was assumed to represent the 5th quantile, the best estimate the median, and upper estimate the 95th quantile of a non-parametric distribution.

In cases where the adjusted intervals fell outside of reasonable bounds (such as [0,1] for probabilities), distributions were truncated just above their lower extreme, and just below their upper extreme. For example, if zero was specified as the absolute minimum and a participant estimated zero for one or more of their estimates, then a small number was added depending on which estimate the zero had been provided for. A zero in the:

- 5^{th} was converted to 0.00001.
- 50^{th} quantile was converted to 0.0001.
- 95^{th} quantile was converted to 0.001.

In questions for which the experts provided estimates in which the upper, best, and /or lower estimates were equal to one another, a small number was either subtracted from the lower estimate to make it lower than the median, or added to the upper estimate to make it higher than the median. The actual number depended on the number of significant figures of the median for example, if the median was:

- 100, then 1 was removed from the lower estimate, or added to the upper estimate.
- 10, then 0.1was removed from the lower estimate, or added to the upper estimate.
- 1, then 0.01 was removed from the lower estimate, or added to the upper estimate
- 0.1, then 0.001, was removed from the lower estimate, or added to the upper estimate
- 0.01, then 0.0001, was removed from the lower estimate, or added to the upper estimate

This was required to avoid floating point errors in *Excalibur* [1], the program used to score experts, develop weights, and aggregate judgements for the IDEA protocol.

The estimates were then exported into individual *.CSV* files from *R,* and converted to *.DTT* files using the macro developed by Cooke [2]. The files from each participant were entered into Excalibur as two case files, one for each workshop group. Excalibur was then used to generate an equal weight aggregation for each of the calibration questions, and each of the questions of interest. The results from Excalibur were imported into *R*. *RMarkdown* was then used to create feedback documents for each of the workshops, which combined graphs and tables containing the standardized 90% intervals for each question. These feedback documents were sent to experts two days prior to each workshop.

In the Round 2 analysis, data were read into *R* and cleaned as in Round 1. Only nine expert estimates were used to calculate aggregations that could be scored on calibration questions.

The following data tables report summary statistics from the manuscript. Due to the small samples sizes and the non-normal nature of the scoring rules, a non-parametric method was used to calculate the 95% confidence intervals around the median. This involves reporting the value of the upper and lower ranked score. Where n was calculated using the following formula:

$$Rank\ of\ CI_{lower} = \frac{n}{2} - \frac{1.96\sqrt{n}}{2}$$

(Equation A.3)

$$Rank\ of\ CI_{upper} = 1 + \frac{n}{2} + \frac{1.96\sqrt{n}}{2}$$

(Equation A.4)

Where, n=the sample size. The rank order was rounded to the nearest integer (for a sample of nine this meant selecting the 2nd and 8th highest values). When sample sizes were less than or

equal to $n=8$ the confidence intervals correspond to the minimum and maximum values, and therefore were not calculated.

2 Performance measures

The following section is adapted from the author's work uploaded onto the Open Science Framework [3].

2.1 IDEA protocol

2.1.1 ALRE Accuracy

'ALRE Accuracy' is used as a measure of performance for the best estimate (a point estimate). It aims to assess the difference between the prediction *b* (the participant's best estimate) and observed value *x*.

Commonly applied measures of accuracy include Mean Absolute Percentage Error (MAPE), which gives the average percentage difference between the prediction and observed value, and Root Mean Square Percentage Error (RMAPE), which is the square root of the MAPE [4]. Both MAPE and RMAPE are strongly affected by one or a few very divergent responses [4].

To overcome these limitations, Burgman, Carr [5] outlined an alternative approach, which we adopt. The approach involves first standardising the best estimates $b_e'^{n,r}$ from each participant *e*, for each question *n*, in each round *r* (including the realised outcome) by the range of responses for each question. This is termed 'range-coding' and is given by,

$$b_e^n = {(b_e'^n - b_{min}'^n)}\Big/{(b_{max}'^n - b_{min}'^n)}$$

(Equation A.5)

where, b_e^n is the range-coded response for participant e, $b_{max}'^n$ is the maximum best estimate response taken from the pool of responses (best estimates) from all participants for question n, and $b_{min}'^n$ is the minimum best estimate response. Note the realised truth (x'^n) for each question

is also range-coded using Equation A.5. If estimates are to be compared across rounds, then both Round 1 and Round 2 estimates need to be range-coded together.

Range-coding reduces the contribution of the question scales on the results. The range-coded values are then used to calculate performance using the average log-ratio error (ALRE, [5]):

$$ALRE_e = \frac{1}{N}\sum_{n=1}^{N}\left|\log_{10}\left({}^{x^n+1}/_{b_e^n+1}\right)\right|$$

(Equation A.6)

where, N is the number of quantities assessed b_e^n is the range-coded best estimate prediction for question n by expert e, and x^n is the range-coded observed (true) value for question n (range-coded values are derived from Equation A.5, above). A '1' is added to avoid taking the log of zero (which occurs when the realisation is standardised). The $\log_{10}$ ratio provides a measure that emphasises order of magnitude errors rather than linear errors. That is, a prediction that is 10 fold greater than the observed value weighs as heavily as a prediction which is one-tenth the observed value [5]. Smaller ALRE scores indicate more accurate responses. For any given question, the log ratio scores have a maximum possible range of 0.31 (=$\log_{10}(2)$), which occurs when the true answer coincides with either the group minimum or group maximum.

To provide an example, imagine four experts provide the estimates listed in Table A. 1 for a single question. Imagine the realised truth for the question is 30.

Table A. 1 Hypothetical estimates of four experts for a single question

Expert	5th (lower)	50th (best)	95th (upper)
1	2	12	34
2	4	15	50
3	7	9	40
4	20	22	23

We then range-code the best estimates from the experts and include the answer using Equation A.5 above. Column 3 of Table A. 2 provides the output of this range-coding, the numbers underlined represent the minimum and the maximum from the data set.

We can then use these range-coded estimates to calculate the ALRE score, using Equation A.6 above. In this case the experts are only assessed on one question. The ALRE scores for the experts are provided in Table A. 2.

Table A. 2 The range-coded responses and the ALRE score

Expert	50th	Range coded responses (b_e^n , x^n)	ALRE (for 1 question)
1	12	0.14	0.24
2	15	0.29	0.19
3	<u>9</u>	0.00	0.30
4	22	0.62	0.09
Answer	<u>30</u>	1.00	N/A

2.1.2 Calibration

In this study, we refer to 'calibration' in terms of interval judgements in which a judge is considered well-calibrated if over the long run, for all questions answered, the proportion their intervals that capture the realised truth equals the probability assigned [6-9].

As the information from the four-step elicitation involves a standardisation of intervals, we use the standardised upper and lower values of those intervals and the standardised level of confidence associated with those intervals. Scoring participants on their standardised intervals is thought to be acceptable as the participants receive feedback on these standardisations between Round 1 and Round 2 and are informed they can (and should) adjust their estimates if

they do not accord with their true beliefs. They are also made aware that this this is how they will be scored.

In this study, we standardised intervals to 90%, therefore a perfectly calibrated individual will capture the realised truth approximately 90% of the time. We calculated the actual number of realisations captured as,

$$C_e = \frac{t}{N} \times 100$$

(Equation A.7)

where, C_e is the score for calibration for participant e, while t is the number of standardised upper and lower intervals provided by the participant which contained the realised truth *x*, and N is the total number of questions answered by the participant.

This scoring rule follows that used by Burgman, Carr [5], Speirs-Bridge, Fidler [10], and McBride, Fidler [11] for evaluating performance of intervals derived from the four-step elicitation. As it is possible for participants to obtain a high calibration by providing very wide (uninformative) intervals, this measure must be considered alongside informativeness (precision) (described below).

2.1.3 Informativeness (termed 'precision' in the study)

'Informativeness' measures the width (or precision) of the of the participant's intervals relative to the total range provided by participants for a question (in accordance with [12]). This differs from the relative information score [13] described by [11], which scores information within, and outside each of the participant's quantiles relative to a uniform or log-uniform distribution

The informativeness of participants is given by the width of standardised intervals (i.e. the 90% credible intervals) supplied by participants for each question in each round:

$$w_e^n = u_e^n - \ell_e^n$$

(Equation A.8)

where, w_e^n is the width of the standardised interval of participant *e* for question *n*, in round *r*, while u_e^n is the upper standardised estimate provided by participant *e* for question *n*, in Round *r*, and ℓ_e^n is the lower standardised estimate provided by participant *e* for question *n*.

For each question, a background range was also calculated

$$w_{max}^n = u_{max}^n - \ell_{min}^n$$

(Equation A.9)

Where, w_{max}^n is the background range created for question n, u_{max}^n is the highest standardised upper bound estimate provided for question n by any participant, and ℓ_{min}^n is the lowest standardised lower bound estimate provided by any participant.

The average informativeness score of each participant can then be calculated as:

$$I_e = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{w_e^n}{w_{max}^n} \right|$$

(Equation A.10)

where I_e^r is the average informativeness of participant e over all questions in Round r, $w_e^{n,r}$ is the width of the interval provided by participant *e* in Round *r* for question n, w_{max}^n is the background range for question *n*, and N is the total number of questions answered.

Scores range between 0 (no uncertainty), to 1 (participant's intervals were always equal to the background range of the questions). Lower scores are better.

Note that the score must be considered in conjunction with calibration as it may reward participants who report no uncertainty. In this case, unless the participant knows the truth with absolute certainty, they would be expected to have poor calibration, which is often weighted higher than informativeness by a decision maker.

To provide an example, we can take the estimates provided by the four experts provided in Table A. 1, and determine the ranges w_e^n for each expert's 90% credible intervals. Table A. 3 provides this information.

We can also calculate a background range w_{max}^n using the upper (50) and lower (2) standardised estimates of the expert's range, in this example the background range w_{max}^n is 50-2= 48.

From this information, we can calculate the informativeness of each expert I_e using Equation A.10 above. Table A. 3 provides the output of these calculations.

Table A. 3 Informativeness for each expert calculated from their 5th and 95th quantiles, and the background range R

Expert	5th	95th	w_e^n	I_e
1	2	34	32	0.67
2	4	50	46	0.96
3	7	40	33	0.69
4	20	23	3	0.06

2.2 The Classical Model

2.2.1 Statistical accuracy

Statistical accuracy (often referred to as calibration and often denoted by 'C') is an absolute value, defined as the p-value of the statistical test which assesses the degree to which the expert answers according to a theoretically optimal multinomial distribution (most often = *p(0.05,0.45,0.45,0.05))*. The resulting p-value can be interpreted as the value at which one would falsely reject the hypotheses that a set of probability assessments of the expert accord with this theoretical multinomial distribution [14].

An explanation of how this is calculated is nicely summarised in Quigley, Colson [15], this section draws extensively from Quigley, Colson [15], to summarise the key points here. Experts are asked to provide judgements as 5^{th}, 50^{th}, and 95^{th} quantiles of a probability distribution. This creates four inter-quantile ranges (<5, 5-50, 50-95, >95). Over a set of questions for which realisations could be obtained, we would expect that for:

- 5% of their judgements the realisations would fall below their 5^{th} quantile. We express the observed proportion as Q_1.
- 45% of their judgements the realisation would fall between their 5^{th} and their 45^{th} quantile. We express the observed proportion as Q_2.
- 45% of their judgements the realisations would fall between their 50^{th} and their 95^{th} quantile. We express the observed proportion as Q_3.
- 5% of their judgements the realisation would fall above their 95^{th} quantile. We express the observed proportion as Q_4.

One thing to note in scoring expert judgements, if the realisations are equal to the values provided by the experts for the 5th, 50th, and 95th quantiles, then the following rules are used to decide which probability bin the realisation should be placed into:

- If the realisation equals the 5th quantile, it is placed in first probability bin (Q_1).
- If the realisation equals the 50th quantile, it is place in the second probability bin (Q_2).
- If the realisation equals the 95th, it is placed in the third probability bin (Q_3).

The expectation of where the realisations will fall in relation to an expert's inter-quantile ranges can be expressed as a theoretically optimal multinomial distribution $p=(0.05, 0.45, 0.45, 0.05)$ [16]. Under the Classical Model, the actual proportion of realisations within each inter-quantile range for each expert (or aggregation) is tallied to create a multinomial distribution for each expert: $s(e)=(Q_1, Q_2, Q_3, Q_4)$.

As an example, if we imagine an expert (e) is asked to answer 10 calibration questions. We subsequently find that across these ten questions one realisation falls below their 5th quantile, four fall between their 5th and 50th quantile, four between their 50th and 95th quantile, and one above their 95th quantile, then the expert will have a distribution as follows: $s(e)=(1/10, 4/10, 4/10, 1/10)$ or $(0.10, 0.40, 0.40, 0.10)$ (Table A. 4).

Table A. 4 **A summary of the performance of a hypothetical expert (e) on ten calibration questions compared to expected performance. Table adapted from Quigley, Colson [15].**

	Inter-quantile intervals			
	Q1 Below 5th	Q2 5th to 50th	Q3 50th to 95th	Q4 Above 95th
Observed proportion of realisations s(e)	0.10	0.40	0.40	0.10
Expected proportion of realisations p	0.05	0.45	0.45	0.05

From this information, the Classical Model then measures how extreme the set of realisations *s(e)* are with respect to the expected realisations, p. This is done using the Kullback-Leibler (KL) divergence measure. The KL measure is a measure of difference between the observed and expected probabilities [15]. The formula is as follows:

$$K(s_e\ p) = \sum_{i=1}^{m} s_{ei} \ln\left(\frac{s_{ei}}{p_i}\right)$$

(Equation A.11)

where: $K(s_e\ p)$ is the KL divergence measure for expert e, s_{ei} is the observed proportion of realisations (out of N questions), in interval i, for expert e, p_i is the expected proportion of realisations in interval i, m is the number inter-quantile intervals (in the Classical Model this is 4).

Applying this to the data in Table A. 4 we obtain:

$$K(s_e\ p) = 0.10\ln\left(\frac{0.10}{0.05}\right) + 0.40\ln\left(\frac{0.40}{0.45}\right) + 0.40\ln\left(\frac{0.40}{0.45}\right) + 0.10\ln\left(\frac{0.10}{0.05}\right)$$

$$= 0.04$$

(Equation A.12)

If the observed proportions perfectly match the expected proportions then the divergence measure would be 0, as the difference grows so does the measure [15].

If an expert's assessments are statistically accurate i.e. the long run observations will equal the expected proportions, then the probability measure will be equivalent to a chi-squared distribution for large sample sizes:

$$Pr\{2\mathrm{N}K(s_e\ p)\ \leq g\}\ \rightarrow \chi^2_{m-1}(g), as\ N \rightarrow \infty$$

(Equation A.13)

where N is the number of calibration questions, and $\chi^2_{m-1}(g)$ is the Cumulative Distribution Function (CDF) of the χ^2 distribution with m-1 degrees of freedom evaluated at x, where m represents the number of inter-quantile intervals assessed.

In the Classical Model *m* is equal to 4 so we have a chi-squared distribution with three degrees of freedom. The expert had a divergence measure of 0.04 (Equation A.12), so $2NK(s_e, p)$ is equal to 0.88. The statistical accuracy (SA) of the expert can then calculated as:

$$SA = 1 - \chi^2_3(0.88)$$

$$= 0.83$$

(Equation A.14)

While this is calculated in Excalibur, it can also be calculated using the divergence measure and the *pchisq()* function in R.

A multinomial distribution provided by the expert *s(e)* which is equal to the distribution *p* will receive a score close to or equal to 1. A distribution which is dramatically different from the distribution p will result in a lower score [16, 17]. Scores below and alpha level of 0.05 are usually interpreted as statistically inaccurate, in other words there is less than a 5% chance that one would falsely reject the hypothesis that the expert is answering according to the p-distribution [16]. We recommend referring to Quigley, Colson [15] for a more thorough explanation.

2.2.2 Information

'Information' (often referred to as informativeness but differentiated here from the IDEA protocol) under the Classical Model (CM) measures the degree to which the distribution

supplied is both concentrated and the level of departure from a uniform or log-uniform distribution (which are considered the least informative distributions).

To calculated this, the Classical Model again uses the Kullback-Leibler (KL) divergence measure, as this measure is scale invariant [15]. The spread of the expert's distributions is assessed relative to an intrinsic range (Z^n). This is calculated by assessing the spread of all expert's distributions for each question, determining the lowest estimate and the highest estimate provided to create a range (i.e. w_{max}^n), and adding a 10% overshoot (O) to the lowest ℓ_{min}^n and highest u_{max}^n estimates of this range to create a maximum O_{max}^n and minimum values O_{min}^n for range. See Equations A.15 – A.17 below.

The informativeness of an expert's probability distribution is then assessed using the KL divergence measure relative to a uniform distribution applied over the intrinsic range. The uniform distribution is chosen because it's the least informative distribution (note that a log-uniform background measure is used across very wide ranges).

On any question, an uninformative expert is one whose estimates accord to a uniform distribution in relation to the intrinsic range (i.e. 5% of the range accords with their 5th quantile, 50% below their 50th quantile). As expert's estimates provide ranges which are narrower and / or demonstrate more departure from a uniform distribution, their information score increases. The lowest score for information is 0, which coincides to a uniform distribution of the intrinsic range, while the score is theoretically unbounded above.

To provide an example, revisit the estimates provided by four experts in Table A. 1 for a single question. We have already determined that the range R^n for the estimates is 48.

From this data, we then add the 10% overshoot to the range using the following formulas:

Lower bound:

$$O^n_{min} = \ell^n_{min} - (w^n_{max} * .10)$$

(Equation A.15)

Upper bound:

$$O^n_{max} = u^n_{max} + (w^n_{max} * .10)$$

(Equation A.16)

Intrinsic Range:

$$Z^n_{max} = O^n_{max} - O^n_{min}$$

(Equation A.17)

Where, O^n_{min} is the lower bound of the intrinsic range (Z^n_{max}), for question *n*, ℓ^n_{min} is the lowest estimate provided by the group of experts, O^n_{max} is the upper bound of the intrinsic range, u^n_{max} is the upper estimate provided by the group of experts, w^n_{max} is the range (without the overshoot) provided by the group of experts for question *n* (i.e. See Equation A.9 above).

When we apply these equations to the estimates provided in Table A 1, it gives us a lower bound of -2.8 and an upper bound of 54.8, thus the intrinsic range (Z^n_{max}) for this question (*n*) is 57.6.

We then calculate the range ($A^n_{i,e}$) for each expert's inter-quantile intervals (Q_i), for question *n* of *N* and determine the proportion of the intrinsic range Z^n_{max} that is captured in each $A^n_{i,e}$ using the following formula:

$$Z^n_{i,e} = \frac{|A^n_{i,e}|}{Z^n_{max}}$$

(Equation A.18)

Where m relates to the i^{th} interquantile interval (Q), where i takes a number between 1 and m. Where the range of:

- Q_1 extends from the lower limit of the intrinsic range O_{min}^n to the 5th quantile (lower standardised estimate) provided by the expert (i.e. $\ell_e^{n,r}$),
- Q_2 extends from the 5th quantile (lower standardised estimate) provided by the expert (i.e. $\ell_e^{n,r}$) to the 50th quantile (i.e. their best estimate $b_e'^{n,r}$).
- Q_3 extends from the 50th quantile (i.e. their best estimate $b_e'^{n,r}$) to their 95th quantile (i.e. their upper standardised estimate $u_e^{n,r}$).
- Q_4 extends from the 95th quantile (i.e. their upper standardised estimate $u_e^{n,r}$) to the upper limit of the intrinsic range O_{max}^n . the 5th to the 50th quantile, and X_4 from the 95th to the upper limit of the intrinsic range.

We provide a worked example for Expert 1 in Table A. 5, and the results for all experts in Table A. 6.

Table A. 5 Calculating the proportion of the intrinsic range captured by the expert's inter-quantile ranges

Expert	$Z_{1,1}^n$	$Z_{2,1}^n$	$Z_{3,1}^n$	$Z_{4,1}^n$
1	$\left(\frac{\lvert 2 - -2.8\rvert}{57.6}\right)$	$\left(\frac{\lvert 12 - 2\rvert}{57.6}\right)$	$\left(\frac{\lvert 34 - 2\rvert}{57.6}\right)$	$\left(\frac{\lvert 54.8 - 34\rvert}{57.6}\right)$

Table A. 6 The proportion of the intrinsic range captured within each of the expert's inter-quantile ranges.

Expert	$Z_{1,e}^{n}$	$Z_{2,e}^{n}$	$Z_{3,e}^{n}$	$Z_{4,e}^{n}$
1	0.08	0.17	0.38	0.36
2	0.12	0.19	0.61	0.08
3	0.17	0.03	0.54	0.26
4	0.40	0.03	0.02	0.55
Background	0.08	0.17	0.38	0.36

The information from Table A. 6 can be used to create another multinomial distribution for each expert. For Expert 1 the multinomial distribution would be *s(e)= (0.08, 0.17, 0.38, 0.36).* This can then be compared against the expectation had the range been a uniform distribution, *p=(0.05, 0.45, 0.45, 0.05),* to obtain the relative information score using the following formula, noting that the expectation, p, is the numerator which differs from Equations A.11 and A.12 above:

$$RI(p, s_e) = \sum_{i=1}^{m} p_i \ln\left(\frac{p_i}{s_{ei}}\right)$$

(Equation A.19)

If we expand this for Expert 1 then we get the following score for relative Information (RI),

$$RI(p, s_e) = .05\ln\left(\frac{.05}{.08}\right) + 0.45\ln\left(\frac{.45}{.17}\right) + 0.45\ln\left(\frac{.45}{.38}\right).05\ln\left(\frac{.05}{.36}\right)$$

$$= 0.38$$

(Equation A.20)

Applying this formula to all experts we get the relative information scores provided in Table A. 7.

Table A. 7 The relative information for each inter-quantile range provided by four experts, and their average relative information (RI).

	$RI(Q_1)$	$RI(Q_2)$	$RI(Q_3)$	$RI(Q_4)$	***I***
Expert 1	-0.03	0.43	0.07	-0.10	**0.38**
Expert 2	-0.04	0.39	-0.14	-0.03	**0.18**
Expert 3	-0.06	1.15	-0.08	-0.08	**0.93**
Expert 4	-0.10	1.15	1.46	-0.12	**2.39**

2.2.3 The Classical Model score: CM Score

In the Classical Model, the scores for experts are obtained by multiplying the information (RI) Score by the statistical accuracy (SA) Score:

$$\text{CM Score} = \text{SA} \times \text{RI}.$$

(Equation A.21)

The Classical Model scoring rule is termed an asymptotically proper scoring rule. It is asymptotical because the distribution of the test statistic used to calculate statistical accuracy has an asymptotic distribution. Theoretically this means that you would have infinitely many calibration questions then the scoring rule will be a proper scoring rule. However, these asymptotic properties are well approximated by 10-15 questions.

Another property which makes the scoring rule a proper scoring rule is that a cut-off is imposed for statistical accuracy (alpha > 0). However, all possible combinations of multinomial distributions will have an alpha > 0 (they just might be very small numbers).

2.2.4 Calculating performance weights

In the Classical Model, the scores for statistical accuracy (SA) and relative information (RI) are then used by 'Excalibur', to develop weights, these weights are used to combine judgements of experts (termed Decision-makers, or DM).

There are five types of aggregations that are calculated each of which were compared in this paper.

Equal Weights: The equally weighted group aggregation is simply a linear pool of all expert distributions using the arithmetic mean of their distributions. It affords all experts the same weight in the group aggregation regardless of how well they perform on test questions. It can be calculated without calibration questions.

Global Weights: considers the statistical accuracy and the relative information of the experts. Those who performed better on the calibration questions are afforded more weight than those who performed poorly. An expert may also receive little weight in the group aggregation if they are as statistically accurate but less informative to another expert.

To calculate global weights an average information score across the set of calibration questions is calculated:

$$\mathrm{RI}_e = \frac{\sum_{\mathrm{n}=1}^{\mathrm{N}} RI(p,\ s_e)}{\mathrm{N}}$$

(Equation A.22)

Raw weights are then calculated removing any expert who has a statistical accuracy below acceptable levels, typically below 0.01 (Quigley et al. 2018).

$$v'_{\mathrm{e}} = \mathrm{SA}_e RI_e \times 1_{\alpha}(\mathrm{SA}_e)$$

(Equation A. 23)

Where $1_{\alpha}(\mathrm{SA}_e)$ is 1 if the statistical accuracy SA_e exceeds the cutoff for α. The use of $1_{\alpha}(\mathrm{SA}_e)$ is imposed by the requirement that the weights w'_{e} should be an asymptotically proper scoring rule: where an expert maximises their long run expected weight if and only if their quantile assessments correspond to their true belief (Quigley et al. 2018). Weights are then normalised across all experts (Quigley et al. 2018).

$$\mathrm{v_e} = \frac{\mathrm{v}'_{\mathrm{e}}}{\sum_{\forall k} v'_k}$$

(Equation A. 24)

Itemised Weights: uses the same SA scores as Global Weights, however, the weight each expert is awarded will change per question because it takes into account the informativeness of the expert for each particular question rather than the average one calculated based on all questions. This often leads to a slightly more informative decision-maker on average than Global Weights.

To calculate itemised weights the raw weight for expert *e* on question *n* is calculated as follows:

$$\mathrm{v}'_{\mathrm{en}} = \mathrm{SA}_e RI_{en} \times 1_{\alpha}(\mathrm{SA}_e)$$

(Equation A. 25)

Noticing that we have two subscripts, one for the expert and one for the question. If there is little relative information in the information score between experts across questions then the assessment will be similar to Global Weights (Quigley et al. 2018).

Optimised Global Weights and Optimised Itemised Weights: these aggregations are similar to their un-optimised variants described above (i.e. Global Weights and Itemised Weights). However, unlike the Global Weights and Itemised Weights, they look to optimise the calibration of the weighted aggregations by successively raising the level of α from $\alpha = 0$. The subsequent weights are then recalculated, and the DM with the highest Performance weight is chosen (Quigley et al. 2018). In decisions where there are clearly one or two well calibrated experts, this can lead to most or all of the weight being afforded to those experts and no weight being afforded to the other experts.

3 Expert estimates and aggregations for questions

Figure A 1**Error! Reference source not found.** shows the quantitative estimates (best estimates and 90% credible intervals) of individual experts and the resulting aggregations on a single calibration question in Round 1 and Round 2.

The red horizontal line shows the realised truth for this question. The aggregations only include estimates from experts who took part in both Round 1 and Round 2 and did not look at any of the answers for the realisations. As can be seen, there was a high degree of variation between experts in their Round 1 (dashed lines) and Round 2 (bold lines) estimates in terms of the width of their intervals and to a lesser degree the distance of their best estimate from the truth. For this question, when experts updated their estimates, their best estimate moved closer to the truth and many became more precise in their 90% credible intervals.

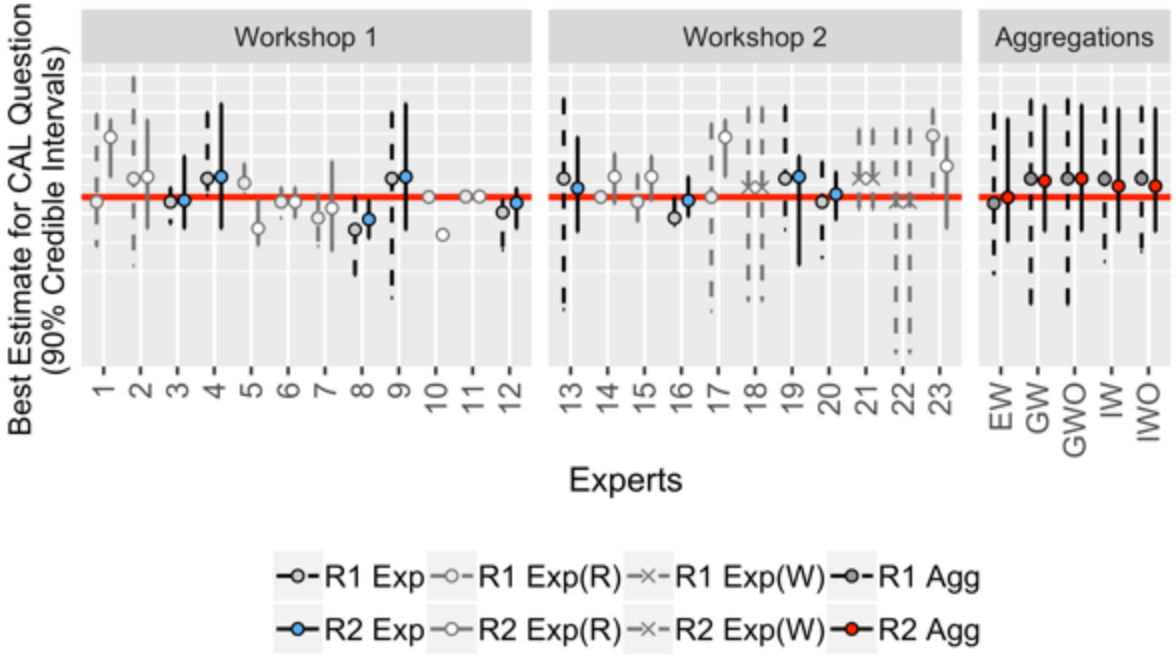

Figure A 1 Round 1 and Round 2 estimates for Question 2, a calibration question (due to confidentiality the Y-axis (plotted on a square root scale) cannot be displayed). The graph shows estimates for each Expert (Exp) and Aggregation (Agg).

Dashed lines indicate Round 1 estimates (R1), whilst solid lines are the revised Round 2 estimates (R2). The aggregations do not include those who withdrew (Exp (W)) or who looked at calibration questions (Exp (R)). Key: EW= Equal Weights, IT=Item Weights, IWO= Item Weights Optimised, GW=Global Weights, GWO=Global Weights Optimised.

Figure A 2 demonstrates the difference between equal-weighted (EW (9 participants), X.EW (all 20 participants)) and performance-weighted aggregations (GW, IW, GWO, IWO, 9 participants) for three of the 17 questions of interest. The orange line represents a point at which Australia would lose an equivalent number of air-vehicles as the Foreign Air Force (FAF) for a particular difference in configuration, role and environment. Numbers above the orange line represent a belief that Australia would lose more air-vehicles than the FAF, and numbers below the line represent a belief Australia would lose less air-vehicles than the FAF. The median estimate for all aggregations in Round 2 on these questions is closely centred on the same number of losses as the FAF, suggesting that for these questions the expert's best estimate was that there would be little difference between Australia and the FAF. However, for each question the 90% credible intervals extend above and below the orange line indicating experts thought there were factors which could make the attrition rates much higher or lower than the FAF.

There was little difference between the two different equal weighted aggregations (EW, X.EW) in their estimates.

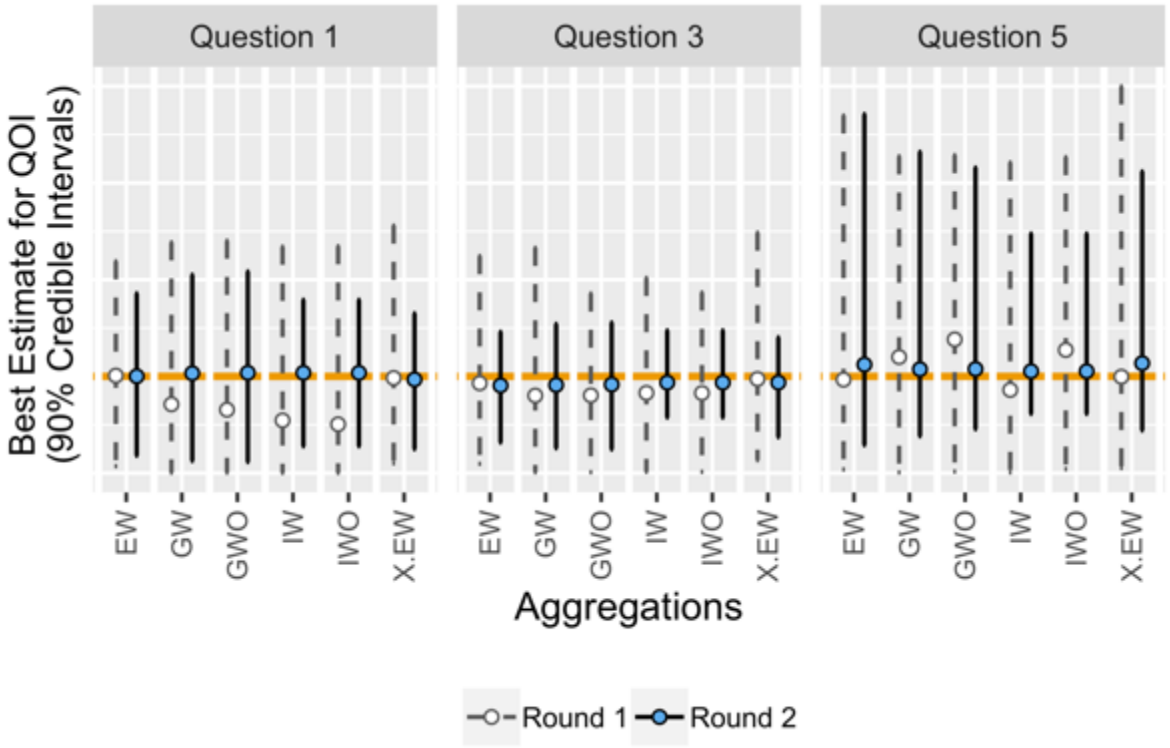

Figure A 2. The graph shows each of the aggregations for three questions of interest, the Y-axis cannot be displayed, but all graphs are on the same linear scale. The white and blue dots represent the best estimate (taken to represent a median), the vertical line represents the 90% credible intervals. The orange line represents Australia losing the equivalent number of air-vehicles as the FAF. Key: EW= Equal Weights, IT=Item Weights, IWO= Item Weights Optimised, GW=Global Weights, GWO=Global Weights Optimised, X.EW=Equal weighted aggregation with all individuals who took part in Round 1 and Round 2.

4 Scores of experts compared to years of experience and self-rating

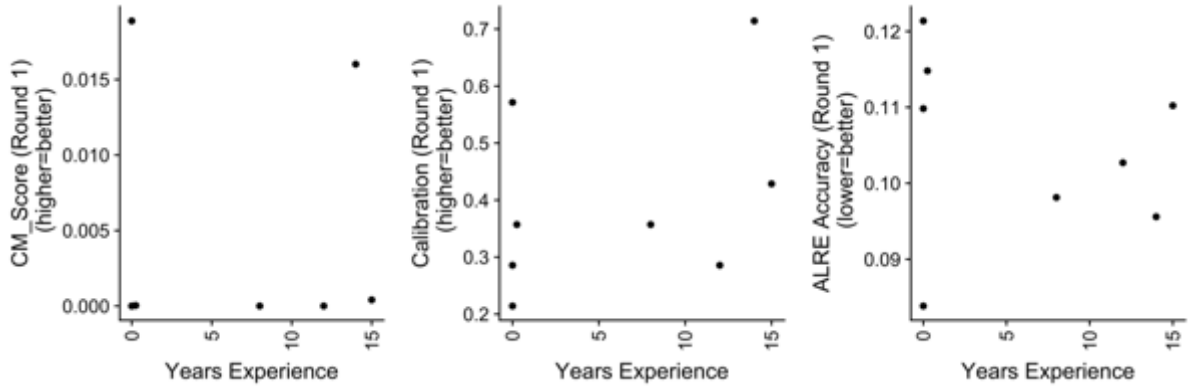

Figure A 3 **Years of experience relevant to the Skua and performance in relation to the CM_Score, Calibration of 90% credible intervals, and accuracy of the best estimate (ALRE Accuracy in Round 1). One expert failed to provide demographic information therefore the graph only demonstrates the scores of eight experts.**

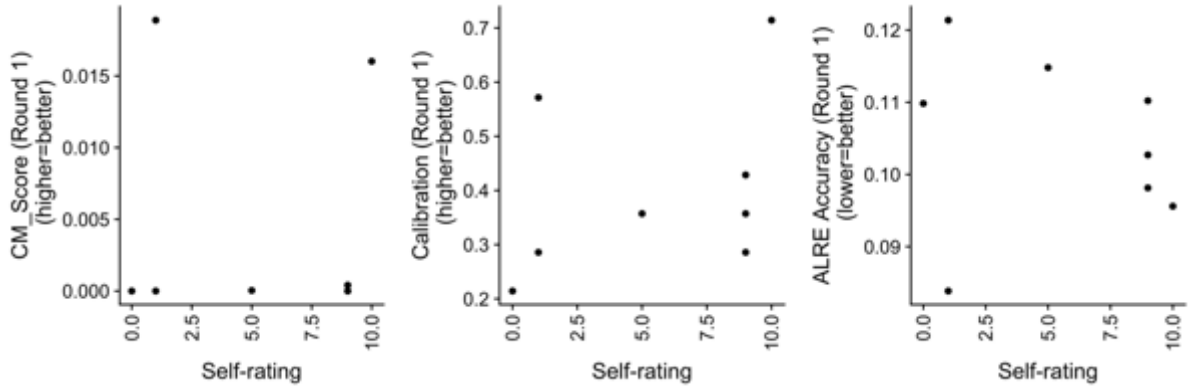

Figure A 4 **Self-rating relevant to the Skua and performance in relation to the CM_Score, Calibration of 90% credible intervals, and accuracy of the best estimate (ALRE Accuracy in Round 1). One expert failed to provide demographic information therefore the graph only demonstrates the scores of eight experts. Experts were enabled to rate themselves on a continuous scale (0= no prior knowledge of the Skua, 1= Basic understanding (e.g. have read a report, or news article, but have no direct experience), 5= Intermediate experience (e.g. relevant experience gained through work, study, hobbies, or lay knowledge), 10=specialist**

understanding (e.g. regularly collect data, prepare or sign off reports, or provide advice on this topic).

5 Improvements Round 2 Calibration Questions

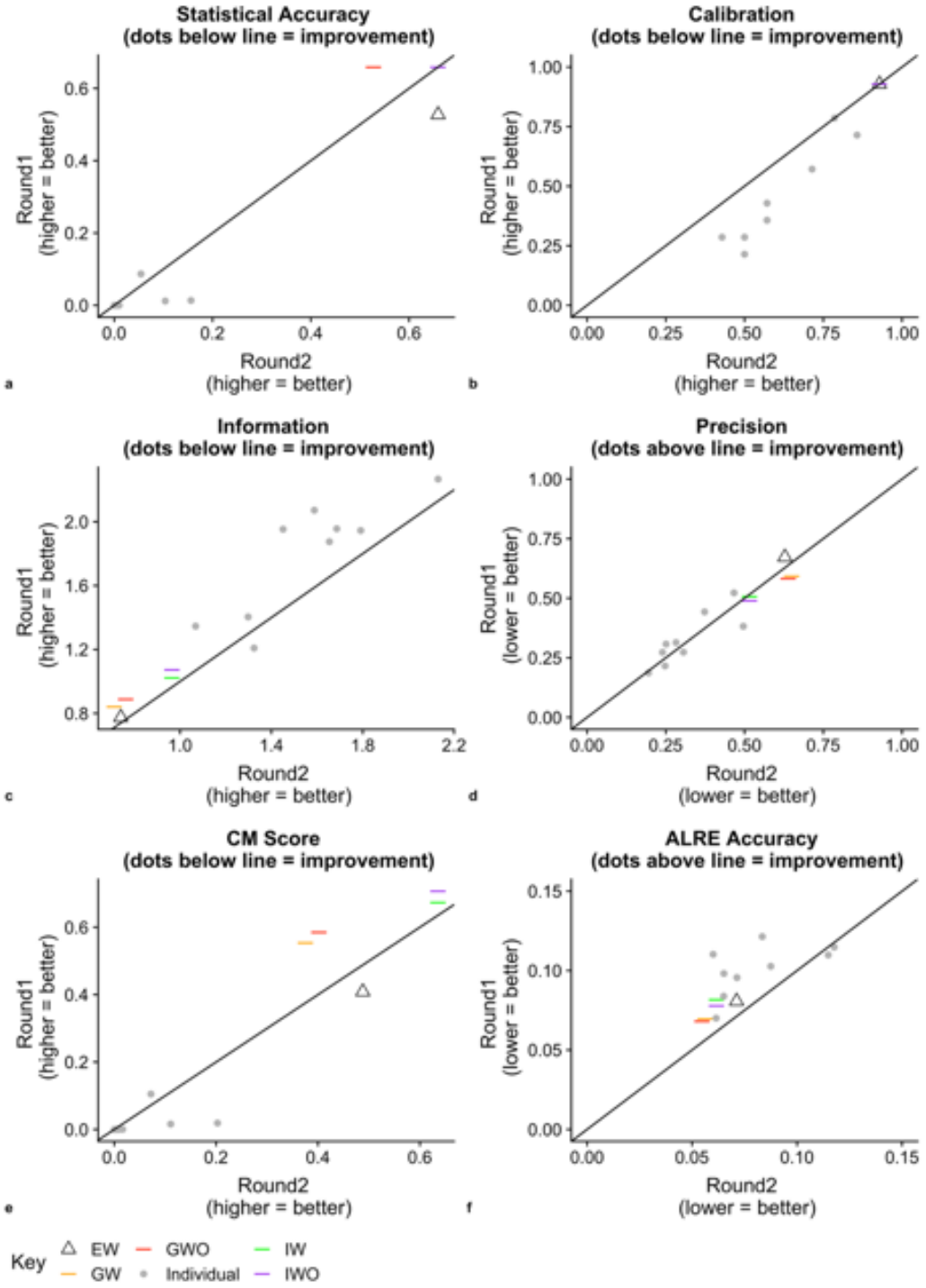

Figure A 5 **Changes in performance on calibration questions between Round 1 and Round 2 for each of the individuals and each of the aggregations. GWO (visible) and GW (hidden) had the same statistical accuracy in Round 1 and Round 2, IWO (visible) and IW (hidden) also had the same statistical accuracy. All aggregations**

had the same calibration in Round 1 and Round 2. Key: EW= Equal Weights, IT=Item Weights, IWO= Item Weights Optimised, GW=Global Weights, GWO=Global Weights Optimised.

6 Improvements Round 2 Questions of Interest

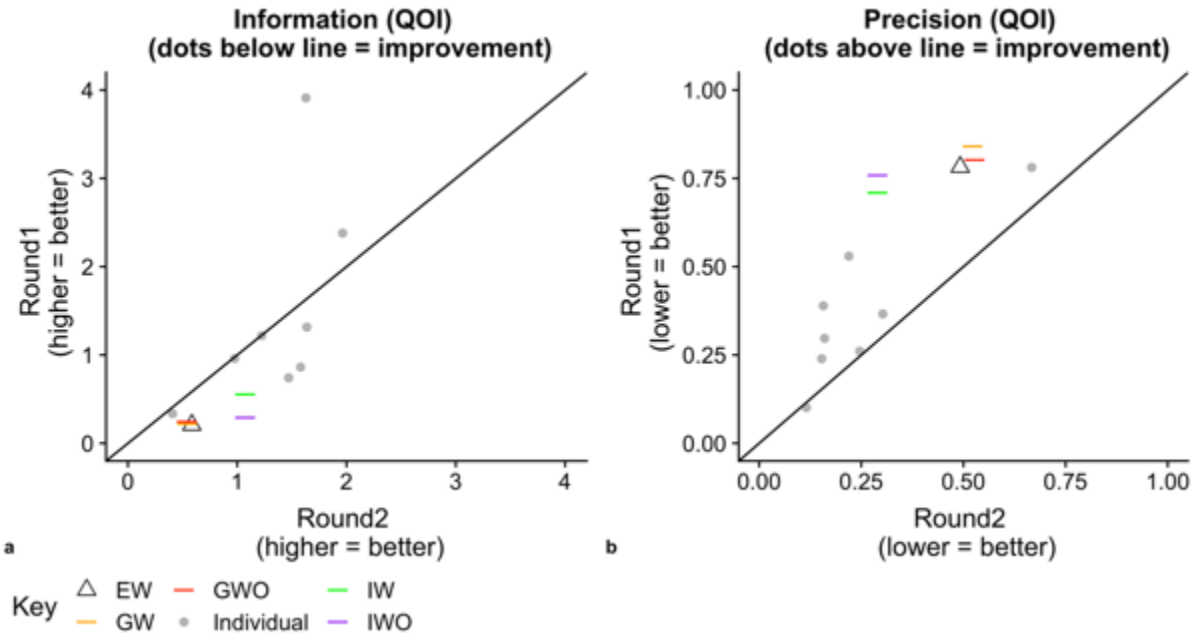

Figure A 6 Changes in information and precision on questions of interest between Round 1 and Round 2 for each of the individuals and each of the aggregations. Key: EW= Equal Weights, IT=Item Weights, IWO= Item Weights Optimised, GW=Global Weights, GWO=Global Weights Optimised.

7 Weights provided to each expert

Table A. 8 provide the weights (Normalised Weight) that each expert received in each aggregation method in Round 1 and Round 2. Note that the calibration and relative information scores will differ from those displayed in the report because they represent the weights for Round 1 and Round 2 calculated separately. These weights were used to develop each of the aggregations. Once the aggregated quantiles for each aggregation were derived for both Round 1 and Round 2 (separately), then the estimates for the experts and each of the aggregations in Round 1 and Round 2 (combined) were scored (i.e. the scores reported in the manuscript). This was required because relative information is a relative score.

Also note that itemised weights (IW, and IWO) do not provide the normalised weight that each expert receives because it takes into account information contained in their estimates on each of the questions of interest (therefore their weight changes per question).

Table A. 8 Change Round 1 and Round 2 Calibration Questions.

Round	Weighting	Aggregation	Unique Identifier	Calibration	Relative Information	Number of realisations	Un-normalised weight	Normalised weight
R1	EW	Individual	22	2.95E-08	1.82	14	5.38E-08	0.11
R1	EW	Individual	20	5.80E-07	1.69	14	9.81E-07	0.11
R1	EW	Individual	5	6.38E-07	1.71	14	1.09E-06	0.11
R1	EW	Individual	12	5.80E-07	2.02	14	1.17E-06	0.11
R1	EW	Individual	23	1.90E-05	1.7	14	3.24E-05	0.11
R1	EW	Individual	7	2.16E-04	1.62	14	3.51E-04	0.11
R1	EW	Individual	2	1.19E-02	1.11	14	0.01	0.11
R1	EW	Individual	8	1.35E-02	1.16	14	0.02	0.11
R1	EW	Individual	17	8.69E-02	0.97	14	0.08	0.11
R1	EW	Excalibur	EW	0.53	0.53	14	0.28	
R1	GW	Individual	22	2.95E-08	1.82	14	5.38E-08	4.74E-07
R1	GW	Individual	20	5.80E-07	1.69	14	9.81E-07	8.66E-06
R1	GW	Individual	5	6.38E-07	1.71	14	1.09E-06	9.60E-06

Round	Weighting	Aggregation	Unique Identifier	Calibration	Relative Information	Number of realisations	Un-normalised weight	Normalised weight
R1	GW	Individual	12	5.80E-07	2.02	14	1.17E-06	1.03E-05
R1	GW	Individual	23	1.90E-05	1.7	14	3.24E-05	2.86E-04
R1	GW	Individual	7	2.16E-04	1.62	14	3.51E-04	3.10E-03
R1	GW	Individual	2	1.19E-02	1.11	14	0.01	0.12
R1	GW	Individual	8	1.35E-02	1.16	14	0.02	0.14
R1	GW	Individual	17	8.69E-02	0.97	14	0.08	0.74
R1	GW	Excalibur	GW	0.66	0.6	14	0.4	
R1	GWO	Individual	23	1.90E-05	1.7	14	0	0
R1	GWO	Individual	22	2.95E-08	1.82	14	0	0
R1	GWO	Individual	7	2.16E-04	1.62	14	0	0
R1	GWO	Individual	12	5.80E-07	2.02	14	0	0
R1	GWO	Individual	20	5.80E-07	1.69	14	0	0
R1	GWO	Individual	5	6.38E-07	1.71	14	0	0
R1	GWO	Individual	2	1.19E-02	1.11	14	0	0
R1	GWO	Individual	8	1.35E-02	1.16	14	0.02	0.16
R1	GWO	Individual	17	8.69E-02	0.97	14	0.08	0.84
R1	GWO	Excalibur	GWO	0.66	0.65	14	0.43	
R1	IW	Individual	22	2.95E-08	1.82	14	5.38E-08	
R1	IW	Individual	20	5.80E-07	1.69	14	9.81E-07	
R1	IW	Individual	5	6.38E-07	1.71	14	1.09E-06	
R1	IW	Individual	12	5.80E-07	2.02	14	1.17E-06	
R1	IW	Individual	23	1.90E-05	1.7	14	3.24E-05	
R1	IW	Individual	7	2.16E-04	1.62	14	3.51E-04	
R1	IW	Individual	2	1.19E-02	1.11	14	0.01	
R1	IW	Individual	8	1.35E-02	1.16	14	0.02	
R1	IW	Individual	17	8.69E-02	0.97	14	0.08	
R1	IW	Excalibur	IT	0.66	0.78	14	0.51	
R1	IWO	Individual	23	1.90E-05	1.7	14	0	
R1	IWO	Individual	22	2.95E-08	1.82	14	0	
R1	IWO	Individual	7	2.16E-04	1.62	14	0	
R1	IWO	Individual	12	5.80E-07	2.02	14	0	
R1	IWO	Individual	20	5.80E-07	1.69	14	0	
R1	IWO	Individual	5	6.38E-07	1.71	14	0	
R1	IWO	Individual	2	1.19E-02	1.11	14	0	
R1	IWO	Individual	8	1.35E-02	1.16	14	0.02	
R1	IWO	Individual	17	8.69E-02	0.97	14	0.08	
R1	IWO	Excalibur	IWO	0.66	0.83	14	0.55	
R2	EW	Individual	12	1.23E-05	1.85	14	2.27E-05	0.11

Round	Weighting	Aggregation	Unique Identifier	Calibration	Relative Information	Number of realisations	Un-normalised weight	Normalised weight
R2	EW	Individual	20	1.26E-03	1.51	14	1.91E-03	0.11
R2	EW	Individual	22	2.20E-03	1.31	14	2.89E-03	0.11
R2	EW	Individual	5	2.20E-03	1.41	14	3.10E-03	0.11
R2	EW	Individual	23	6.46E-03	1.17	14	0.01	0.11
R2	EW	Individual	7	9.84E-03	1.37	14	0.01	0.11
R2	EW	Individual	17	5.43E-02	1.04	14	0.06	0.11
R2	EW	Individual	2	0.1	0.8	14	0.08	0.11
R2	EW	Individual	8	0.16	1.02	14	0.16	0.11
R2	EW	Excalibur	EW	0.66	0.47	14	0.31	
R2	GW	Individual	12	1.23E-05	1.85	14	2.27E-05	6.93E-05
R2	GW	Individual	20	1.26E-03	1.51	14	1.91E-03	5.82E-03
R2	GW	Individual	22	2.20E-03	1.31	14	2.89E-03	8.81E-03
R2	GW	Individual	5	2.20E-03	1.41	14	3.10E-03	9.46E-03
R2	GW	Individual	23	6.46E-03	1.17	14	0.01	2.31E-02
R2	GW	Individual	7	9.84E-03	1.37	14	0.01	4.11E-02
R2	GW	Individual	17	5.43E-02	1.04	14	0.06	0.17
R2	GW	Individual	2	0.1	0.8	14	0.08	0.25
R2	GW	Individual	8	0.16	1.02	14	0.16	0.49
R2	GW	Excalibur	GW	0.53	0.44	14	0.23	
R2	GWO	Individual	22	2.20E-03	1.31	14	0	0
R2	GWO	Individual	23	6.46E-03	1.17	14	0	0
R2	GWO	Individual	12	1.23E-05	1.85	14	0	0
R2	GWO	Individual	7	9.84E-03	1.37	14	0	0
R2	GWO	Individual	20	1.26E-03	1.51	14	0	0
R2	GWO	Individual	5	2.20E-03	1.41	14	0	0
R2	GWO	Individual	17	5.43E-02	1.04	14	0.06	0.19
R2	GWO	Individual	2	0.1	0.8	14	0.08	0.28
R2	GWO	Individual	8	0.16	1.02	14	0.16	0.53
R2	GWO	Excalibur	GWO	0.53	0.49	14	0.26	
R2	IW	Individual	12	1.23E-05	1.85	14	2.27E-05	
R2	IW	Individual	20	1.26E-03	1.51	14	1.91E-03	
R2	IW	Individual	22	2.20E-03	1.31	14	2.89E-03	
R2	IW	Individual	5	2.20E-03	1.41	14	3.10E-03	
R2	IW	Individual	23	6.46E-03	1.17	14	0.01	
R2	IW	Individual	7	9.84E-03	1.37	14	0.01	
R2	IW	Individual	17	5.43E-02	1.04	14	0.06	
R2	IW	Individual	2	0.1	0.8	14	0.08	
R2	IW	Individual	8	0.16	1.02	14	0.16	

Round	Weighting	Aggregation	Unique Identifier	Calibration	Relative Information	Number of realisations	Un-normalised weight	Normalised weight
R2	IW	Excalibur	IT	0.66	0.69	14	0.45	
R2	IWO	Individual	12	1.23E-05	1.85	14	0	
R2	IWO	Individual	20	1.26E-03	1.51	14	1.91E-03	
R2	IWO	Individual	22	2.20E-03	1.31	14	2.89E-03	
R2	IWO	Individual	5	2.20E-03	1.41	14	3.10E-03	
R2	IWO	Individual	23	6.46E-03	1.17	14	0.01	
R2	IWO	Individual	7	9.84E-03	1.37	14	0.01	
R2	IWO	Individual	17	5.43E-02	1.04	14	0.06	
R2	IWO	Individual	2	0.1	0.8	14	0.08	
R2	IWO	Individual	8	0.16	1.02	14	0.16	
R2	IWO	Excalibur	IWO	0.66	0.69	14	0.45	

8 Table of results

Table A. 9 Change Round 1 and Round 2 Calibration Questions.

Grouping	Improvement	Median	n	25th	75th	Min	Max	Change in
EW	increase	0.13	1	0.13	0.13	0.13	0.13	SA (CM)
Excalibur	decrease	0.13	2	0.13	0.13	0.13	0.13	SA (CM)
Excalibur	no change	0.00	2	0.00	0.00	0.00	0.00	SA (CM)
Individual	decrease	0.03	1	0.03	0.03	0.03	0.03	SA (CM)
Individual	increase	4.32E-03	8	0.03	1.97E-03	1.17E-05	0.14	Change SA (CM)
EW	decrease	0.03	1	0.03	0.03	0.03	0.03	Information (CM) CAL
Excalibur	decrease	0.12	4	0.13	0.10	0.06	0.13	Information (CM) CAL
Individual	decrease	0.25	8	0.33	0.15	0.11	0.50	Information (CM) CAL
Individual	increase	0.12	1	0.12	0.12	0.12	0.12	Information (CM) CAL
EW	increase	0.08	1	0.08	0.08	0.08	0.08	Cooke CM
Excalibur	decrease	0.12	4	0.18	0.06	0.04	0.18	Cooke CM
Individual	decrease	0.03	1	0.03	0.03	0.03	0.03	Cooke CM
Individual	increase	0.01	8	0.04	3.19E-03	2.49E-05	0.18	Cooke CM
EW	no change	0.00	1	0.00	0.00	0.00	0.00	Calibration (IDEA)
Excalibur	no change	0.00	4	0.00	0.00	0.00	0.00	Calibration (IDEA)

Grouping	Improvement	Median	n	25th	75th	Min	Max	Change in
Individual	increase	0.18	8	0.21	0.14	0.14	0.29	Calibration (IDEA)
Individual	no change	0.00	1	0.00	0.00	0.00	0.00	Calibration (IDEA)
EW	increase	0.01	1	0.01	0.01	0.01	0.01	ALRE (IDEA)
Excalibur	increase	0.01	4	0.02	0.01	0.01	0.02	ALRE (IDEA)
Individual	decrease	3.95E-03	2	0.00	3.42E-03	2.89E-03	5.00E-03	ALRE (IDEA)
Individual	increase	0.02	7	0.04	0.02	0.01	0.05	ALRE (IDEA)
EW	increase	0.04	1	0.04	0.04	0.04	0.04	Precision (IDEA) CAL
Excalibur	decrease	0.04	4	0.06	0.02	0.01	0.06	Precision (IDEA) CAL
Individual	decrease	0.03	4	0.05	0.03	0.01	0.11	Precision (IDEA) CAL
Individual	increase	0.06	5	0.06	0.03	0.03	0.07	Precision (IDEA) CAL

Table A. 10 Change Round 1 and Round 2 Questions of Interest

Grouping	Improvement	Median	n	25th	75th	Min	Max	Change in
EW	increase	0.38	1	0.38	0.38	0.38	0.38	Information (CM) QOI
Excalibur	increase	0.42	4	0.58	0.32	0.30	0.78	Information (CM) QOI
Individual	decrease	1.35	2	1.82	0.88	0.42	2.28	Information (CM) QOI
Individual	increase	0.20	6	0.62	0.03	0.00	0.73	Information (CM) QOI
EW	increase	0.29	1	0.29	0.29	0.29	0.29	Precision (IDEA) QOI
Excalibur	increase	0.37	4	0.43	0.31	0.28	0.47	Precision (IDEA) QOI
Individual	decrease	0.01	1	0.01	0.01	0.01	0.01	Precision (IDEA) QOI
Individual	increase	0.11	7	0.19	0.08	0.02	0.31	Precision (IDEA) QOI

Table A. 11 Scores of individuals and aggregations

Name	Aggregation	CM Score	Stat. Acc	Cal	R.Inf	Inf	ALRE Acc	Round
12	Individual	1.25E-06	6.38E-07	0.36	1.96	0.31	0.10	1
13	Individual	0.11	0.087	0.79	1.21	0.44	0.07	1
16	Individual	1.13E-06	5.80E-07	0.29	1.95	0.19	0.10	1
19	Individual	6.11E-08	2.95E-08	0.21	2.07	0.22	0.11	1
20	Individual	3.71E-05	1.90E-05	0.36	1.95	0.27	0.11	1
3	Individual	4.06E-04	2.16E-04	0.43	1.88	0.31	0.11	1
4	Individual	0.02	0.013	0.57	1.40	0.52	0.08	1
8	Individual	1.31E-06	5.80E-07	0.29	2.27	0.27	0.12	1
9	Individual	0.02	0.012	0.71	1.35	0.38	0.10	1
EW	EW	0.41	0.527	0.93	0.77	0.67	0.08	1
GW	Excalibur	0.55	0.659	0.93	0.84	0.59	0.07	1
GWO	Excalibur	0.58	0.659	0.93	0.89	0.58	0.07	1
IW	Excalibur	0.67	0.659	0.93	1.02	0.51	0.08	1
IWO	Excalibur	0.71	0.659	0.93	1.07	0.49	0.08	1
12	Individual	3.71E-03	2.20E-03	0.57	1.69	0.28	0.07	2
13	Individual	0.07	0.054	0.79	1.32	0.37	0.06	2
16	Individual	2.26E-03	1.26E-03	0.50	1.79	0.19	0.09	2
19	Individual	3.50E-03	2.20E-03	0.50	1.59	0.25	0.11	2
20	Individual	9.37E-03	6.46E-03	0.57	1.45	0.31	0.12	2
3	Individual	0.02	9.84E-03	0.57	1.65	0.25	0.06	2
4	Individual	0.20	0.156	0.71	1.30	0.47	0.07	2
8	Individual	2.62E-05	1.23E-05	0.43	2.13	0.24	0.08	2
9	Individual	0.11	0.104	0.86	1.07	0.50	0.07	2
EW	EW	0.49	0.659	0.93	0.74	0.63	0.07	2
GW	Excalibur	0.38	0.527	0.93	0.71	0.65	0.06	2
GWO	Excalibur	0.40	0.527	0.93	0.76	0.64	0.05	2
IW	Excalibur	0.64	0.659	0.93	0.96	0.52	0.06	2
IWO	Excalibur	0.64	0.659	0.93	0.96	0.52	0.06	2

9 References

[1] Lightwist. Excalibur. 2013.

[2] Cooke R. Macro converting XL file to EXCALIBUR dtt file. In: Cooke R, editor.2018.

[3] Hemming V. Appendix 9 Equations for Scoring and Weighting Judgements (appendix to Chapter 4, 5 and 6). In: Framework OS, editor.2019.

[4] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. International journal of forecasting. 2006;22:679-88.

[5] Burgman M, Carr A, Godden L, Gregory R, McBride M, Flander L, et al. Redefining expertise and improving ecological judgment. Conservation Letters. 2011;4:81-7.

[6] Lichtenstein S, Fischhoff B, Phillips LD. Calibration of probabilities: The state of the art. Decision making and change in human affairs: Springer; 1977. p. 275-324.

[7] Lin S-W, Bier VM. A study of expert overconfidence. Reliability Engineering & System Safety. 2008;93:711-21.

[8] Teigen KH, JØrgensen M. When 90% confidence intervals are 50% certain: on the credibility of credible intervals. Applied Cognitive Psychology. 2005;19:455-75.

[9] Soll JB, Klayman J. Overconfidence in interval estimates. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2004;30:299.

[10] Speirs-Bridge A, Fidler F, McBride M, Flander L, Cumming G, Burgman M. Reducing overconfidence in the interval judgments of experts. Risk Analysis. 2010;30:512-23.

[11] McBride MF, Fidler F, Burgman MA. Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research. Diversity and Distributions. 2012;18:782-94.

[12] Yaniv I, Foster DP. Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. Journal of Experimental Psychology: General. 1995;124:424.

[13] Cooke RM. Experts in uncertainty: Opinion and subjective probability in science. New York: Oxford University Press; 1991.

[14] Colson AR, Cooke RM. Cross validation for the classical model of structured expert judgment. Reliability Engineering & System Safety. 2017;163:109-20.

[15] Quigley J, Colson A, Aspinall W, Cooke RM. Elicitation in the Classical Model. In: Dias LC, Morton A, Quigley J, editors. Elicitation: The science and art of structuring judgement. Cham, Switzerland: Springer International Publishing; 2018. p. 15-36.

[16] Bedford T, Cooke RM. Mathematical tools for probabilistic risk analysis. Cambridge, United Kingdom: Cambridge University Press; 2001.

[17] Aspinall WP, Cooke RM. Quantifying scientific uncertainty from expert judgement elicitation. In: Rougier J, Sparks S, Hill L, editors. Risk and uncertainty assessment for natural hazards. Cambridge, United Kingdom: Cambridge University Press; 2013. p. 64-99.