Weighting and Aggregating Expert Ecological Judgements

(Running head: Weighting and Aggregating Experts)

Victoria Hemming[1,2,3], Anca M. Hanea[1,2], Terry Walshe[2] , Mark A. Burgman[2,4]

[1] The Centre of Excellence for Biosecurity Risk Analysis, The University of Melbourne, Melbourne, Victoria, Australia. Twitter: @CEBRA_UOM

[2] School of Biosciences, The University of Melbourne, Melbourne, Victoria, Australia.

[3] Department of Forest and Conservation Sciences, The University of British Columbia, Vancouver, Canada.

[4] The Centre for Environmental Policy, Imperial College London, London, United Kingdom @IC_CEP

* Corresponding author: Victoria Hemming (victoria.hemming@ubc.ca, twitter @v_hemming, Department of Forest and Conservation Sciences, The University of British Columbia, Vancouver, Canada). ORCID ID: 0000-0003-3220-6161.

Keywords: Performance Weights, Equal Weights, Aggregation, Expert Judgement, Calibration, Classical Model

Abstract

Performance weighted aggregation of expert judgements, using calibration questions, has been advocated to improve pooled quantitative judgements for ecological questions. However, there is little discussion or practical advice in the ecological literature regarding the application, advantages or challenges of performance weighting.

In this paper we 1) illustrate how the IDEA protocol with four-step question format can be extended to include performance weighted aggregation from the Classical Model, and 2) explore the extent to which this extension improves pooled judgements for a range of performance measures.

Our case study demonstrates that performance weights can improve judgements derived from the IDEA protocol with four-step question format. However, there is no *a-priori* guarantee of improvement. We conclude that the merits of the method lie in demonstrating that the final aggregation of judgements provides the best representation of uncertainty (i.e. validation), whether that be via equally weighted or performance weighted aggregation.

Whether the time and effort entailed in performance weights can be justified is a matter for decision-makers. Our case study outlines the rationale, challenges, and benefits of performance weighted aggregations. It will help to inform decisions about the deployment of performance weighting and avoid common pitfalls in its application.

1 Introduction

Over the past 15 years a considerable body of research has emerged in the ecological literature emphasising the need for more rigorous and structured methods for collecting quantitative expert judgements. The literature has summarised existing structured elicitation protocols and key steps which could be adapted and applied to better suit the practical (e.g. geographically dispersed experts) and financial (lack of funding) constraints of most ecological contexts (Burgman 2004, Low Choy et al. 2009, Kuhnert et al. 2010, Burgman et al. 2011a, Martin et al. 2012, McBride et al. 2012a, McBride et al. 2012b, Drescher et al. 2013).

A common approach that has been advocated is to recruit a diverse group of individuals and take an equally weighted aggregation of their independent judgements (Burgman et al. 2011b, Hemming et al. 2018b). This will often produce point estimates which are at least as accurate (i.e. closer to the truth) and interval judgements which are better calibrated than the median-ranked individual for these scores (Burgman et al. 2011b, Budescu and Chen 2014, Hemming et al. 2018b). While one person can sometimes outperform the group aggregate, rarely can that person be predicted by credentials conventionally associated with expertise such as age, experience, or peer-identification (Aspinall and Cooke 2013, Burgman 2015, Mellers et al. 2015).

The performance of the equal weighted aggregation is largely explained as a statistical phenomenon (Lorenz et al. 2011) in which the judgments of individuals represent random independent samples. If those samples are diverse then not only should the information pool related to the questions increase (Clemen and Winkler 1999), but the errors made by individuals are more likely to cancel (Larrick and Soll 2006, Budescu and Chen 2014). This phenomenon is often termed the ‘wisdom of the crowd’ (Surowiecki 2004), or the ‘staticised group’ (Einhorn et al. 1977, Hogarth 1978). Interestingly, participants need not be experts and can be biased, as long as they have some information related to the questions that can be combined for prediction (Budescu and Chen 2014).

Equal weighting is advantageous as it's relatively simple to apply (Hogarth 1978, Hora 2004, Hemming et al. 2018b). Typically, group sizes of 5-12 participants derive improved judgements, with diminishing returns thereafter (Hogarth 1978, Hora 2004, Hemming et al. 2018b). It requires no additional work to develop questions or performance measures to score and aggregate experts. It can be applied to any type of prediction including point estimates, distributions and probabilities. The simplicity of equal weighting, and its ability to improve a wide range of estimates make it suitable for aggregating judgements under the practical and financial constraints typical of many ecological decisions.

However, despite substantial testing and real-world applications, many people find equal weighting difficult to trust (Weiss and Shanteau 2004). This is partly because the method relies on the recruitment of a diversity of individuals, often including individuals who may normally be excluded from such elicitations because of their perceived limited knowledge (Shanteau et al. 2002, Weiss and Shanteau 2004, Burgman et al. 2011a).

When uncertainty is elicited, the diversity of the group can also increase the uncertainty associated with group judgements, sometimes leading to uninformative judgements (MacDonald et al. 2008, Barons et al. 2018). Occasionally individuals will outperform the group aggregation, and ideally decision-makers would like to restrict elicitation to these better performing individuals, or at least have the judgments of those individuals weigh more than those of lesser performers. Finally, there is no single method for generating an equally weighted aggregation. For example, for point estimates, the arithmetic mean is commonly applied, but one could also use the median, geometric mean or harmonic mean (Armstrong 2001, Colson and Cooke 2017). Rarely is there any validation to support such choices made by the analyst, which can lead to questions about the validity of the specific method chosen, and the influence of analyst's subjective bias. The problems associated with equal weights can serve to undermine the credibility of the final judgements derived and the subsequent decisions and assessment based on such judgements.

Performance weighted aggregation is often suggested as a way to address these challenges and perceived deficiencies (Cooke 1991, Budescu and Chen 2014, Mellers et al. 2015). It involves developing sets of questions related to the main elicitation questions for which the answers can be obtained but are not widely known to experts (Cooke 1991, Goossens et al. 2008, Tetlock and Gardner 2015). These are referred to as test, seed or calibration questions (we use the term calibration questions from hereon). Those who perform better on these questions are afforded more weight in the final aggregation of the main elicitation questions. The method is differentiated from other forms of weighted aggregation such as those based on self-rating, peer-rating, trimming, or representativeness in that weights are obtained via validation of judgements against an external truth (Armstrong 2001, Aspinall and Cooke 2013).

The main reason decision-makers seek to apply performance weights is to create aggregated judgements which are more accurate (for point estimates), or well-calibrated and informative (for interval judgements, probabilities and probability distributions) (Budescu and Chen 2014, Mellers et al. 2015, Colson and Cooke 2017). However, the inclusion of calibration questions is also seen to create a sense of legitimacy. It provides evidence that those who have been included in the final aggregation have some knowledge in the relevant domain, and that they can communicate their knowledge together with their uncertainty in the format required by the analyst (Barons et al. 2018, Quigley et al. 2018). It can also be used to validate assumptions made by the analyst in combining expert judgements.

Despite advocacy, there has been little progress in ecology towards understanding or applying performance weighted aggregation, outside of a few applications (Metcalf and Wallace 2013, Wittmann et al. 2015, Barons et al. 2018). We contend this has led to an under-appreciation of the fundamental requirements of the method in ecology, of how the method can be practically applied more widely in ecology, and the extent to which implementation may improve outcomes.

In this paper we 1) illustrate how the IDEA protocol with four-step question format can be extended to include performance weighted aggregation from the Classical Model, and 2) explore the extent to which this extension improves pooled judgements for a range of performance measures.

We choose the IDEA protocol ("Investigate", "Discuss", "Estimate", and "Aggregate") as it is a structured elicitation protocol that has been tested and refined in the ecological literature and (Hanea et al. 2016, Hemming et al. 2018a). The method involves first recruiting a diverse group of individuals, and allowing each individual to "Investigate" the problem before making a private individual estimate (often termed "Round 1"), following which experts see the judgements of others and then enter into a "Discussion" phase. Experts then provide a final private "Estimate" ("Round 2"). The judgements are "Aggregated", typically using equal weights (Figure 1).

Elicitation in the IDEA protocol can be undertaken remotely (i.e. via email), in a face-to-face workshop, or by combining the two formats. This flexibility provides a practical advantage for ecologists who usually have limited resources to convene experts face-to-face.

Most applications of the IDEA protocol in ecology aim to obtain quantitative judgements together with uncertainty. When doing so, the four-step question format is often deployed

(Speirs-Bridge et al. 2010) (Figure 1). This method derives a credible interval with a 'best' point estimate based on the following questions:

1. Realistically what is the lowest plausible value for x?
2. Realistically what is the highest plausible value for x?
3. Realistically what is your best estimate for x?
4. Looking at your interval from lowest to highest, how confident are you that your interval will capture the realised truth.

The four-step question format has been demonstrated to reduce overconfidence in interval judgements relative to eliciting fixed quantiles (Speirs-Bridge et al. 2010). It has also helped in obtaining quantitative judgements (with uncertainty) from experts who may eschew quantification. Its development and application has improved the quality of information derived from expert elicitation in ecology beyond that of categorial variables and point estimates, which can be imbued with considerable ambiguity or fail to provide crucial information about uncertainty (Wallsten et al. 1986, Gregory and Keeney 2017).

The practical advantages of the IDEA protocol with the four-step question format has seen the adoption of the combined method spread rapidly in ecology (Adams-Hosking et al. 2016, Hudson et al. 2017, Barons et al. 2018, Carwardine et al. 2019, Estévez et al. 2019). However, it has been suggested that the aggregations derived could be further improved by incorporating the performance weighted aggregation (Metcalf and Wallace 2013, Hemming et al. 2018a, Hemming et al. 2018b).

The Classical Model (Cooke 1991) is a method for performance weighted aggregation often cited in the ecological literature as a means to improve uncertain quantitative ecological judgements (Burgman et al. 2011a, Martin et al. 2012, Drescher et al. 2013, Metcalf and Wallace 2013, Hemming et al. 2018a). While it has been applied to a large number of engineering case studies (Cooke and Goossens 2008, Colson and Cooke 2017) we are aware of only two ecological examples, both in the Laurentian Great Lakes (Rothlisberger et al. 2009, Wittmann et al. 2015).

In this this study we apply the Classical Model to a case study in which judgements were elicited using the IDEA protocol (Hemming et al. 2018a) and four-step question format (Speirs-Bridge et al. 2010). In doing so, we address the key aims of this study (outlined above), while providing an insight into key considerations required for the deployment of performance weighted aggregation more broadly.

2 Methods

2.1 Fundamentals of performance weighting

There is a considerable body of literature describing the application of performance weighting with calibration questions, however, it is spread across a broad range of domains which can be difficult to access and synthesise. We summarise key points to be considered prior to application.

Generating performance weights with calibration questions entails (a) the development of questions for which there are answers unknown to the participants, and (b) the selection of an appropriate scoring rule to measure the performance of expert estimates.

There is little prescriptive guidance as to what makes a good calibration question, although some features are self-evident (Cooke and Goossens 2000, Aspinall and Cooke 2013, Tetlock and Gardner 2015, Quigley et al. 2018). They should relate to the knowledge needed to answer the main elicitation questions (i.e. domain knowledge). They should ask questions about uncertainty to capture an expert's ability to adapt and communicate their knowledge. They should be in a similar format to the main elicitation questions. They should not be questions which can be easily guessed, and not so hard that an expert could not reasonably form a judgement. A substantial number of calibration questions may be required to differentiate luck from good judgement, depending on the scoring rules. Ideally, questions should relate to predictions of events or quantities rather than estimating the outcomes of past events (retrodictions), although this is not always possible. The questions should be reviewed by at least two people with domain knowledge to ensure they provide fair and reasonable assessments of an expert's ability to make good judgements related to the main elicitation questions.

One of the most important aspects of scoring rules is that they should not influence experts in an undesirable way - termed proper scoring rules (Brier 1950). Strictly proper scoring rules are those for which an expert maximises the expected score, if and only if they state their true beliefs (Gneiting and Raftery 2007). There are many methods for scoring and assessing expert judgements (Brier 1950, Cooke 1991, Flandoli et al. 2011, Budescu and Chen 2014, Satopää et al. 2014, Hemming et al. 2018b), which vary depending on the types of judgements elicited (probabilities, intervals, distributions etc). Not all scoring rules are proper scoring rules, and few have been substantially tested and applied in real applications. The Brier Score is an exception and has been used to assess performance of individuals and groups on single event probabilities such as weather forecasts and geopolitical events, but has not been developed into a formal weighting scheme (Brier 1950, Tetlock and Gardner 2015, Barons et al. 2018). The other exception is the scoring rule of the Classical Model (discussed below) (Cooke 1991).

Scoring rules aim to optimise judgements and the way in which they do this depends on their reward structure (Winkler and Murphy 1968, Tetlock 2005). For example, scoring rules for interval judgements may penalise overconfidence (e.g. intervals that are too narrow, which include the truth less often than the purported level of confidence provided by the expert) more than under-confidence (e.g. intervals that are too broad and capture more realisations than the purported level of confidence of the expert). It's therefore important to understand how such transgressions of judgement are handled by a proposed scoring rule, to ensure that the reward structure matches the preferences and needs of the decision-maker and the problem at hand. This of course requires an awareness among decision-makers about what aspects of judgement are most important to them.

Obtaining an understanding of the reward structure can be challenging as research papers outlining the application of scoring rules rarely provide clear examples of how judgements are incorporated and combined. Few adequately discuss their embedded reward structure. A further complication arises in understanding scoring rules because the terms used to describe judgement, such as 'calibration', 'accuracy' and 'overconfidence', are used interchangeably and may refer to different concepts (Lichtenstein and Fischhoff 1977, Lichtendahl Jr et al. 2013, Cooke 2018b, Hemming et al. 2018b).

2.2 The Classical Model

In this paper we choose to investigate the application of the Classical Model (Cooke 1991). The method was developed as a means for reaching rational consensus, which is defined by Cooke and Goossens (2008) as an agreement as to how to derive a consensus distribution from multiple, elicited distributions. Ultimately, it treats expert judgement as a form of empirical data and promotes adherence to four critical elements of scientific inquiry: accountability, empirical control, neutrality, and fairness (Cooke and Goossens 2008).

In elicitations employing the Classical Model, experts are asked a set of calibration questions (usually 10-15), for which the answers can be obtained. As noted above, these questions should relate to the main questions of the elicitation (termed target variables or questions of interest). Unlike the four-step question format commonly used with the IDEA protocol, experts are asked to specify their judgements as quantiles of a continuous non-parametric probability distribution (usually 5th, 50th, and 95th) for both calibration questions and questions of interest. The individual judgements of experts are typically elicited in a face-to-face elicitation with one or more facilitators present (Wittmann et al. 2015). Experts are scored on their performance using two performance measures (see section 2.4 for details): “statistical accuracy” (often termed “calibration”), and “information” (sometimes termed “informativeness”, or “relative information”). These are subsequently multiplied to provide an asymptotically proper scoring rule (the CM Score) (refer to Appendix S1: Section 4.2.3), and to derive differential weights.

Experts who perform well on the calibration questions are afforded more weight in the final aggregations for the questions of interest. Both equally weighted and performance weighted linear pooled aggregations of distributions are then created and subsequently scored on their performance on the calibration questions (i.e. via in-sample validation, where the same questions used to develop the performance weighted aggregations are used to score the aggregations). To achieve rational consensus, experts or decision makers usually agree prior to the elicitation that the aggregation which achieves the highest combined score on the calibration questions will be used to weight expert judgements of the target questions.

The primary purpose of performance weighting and calibration questions in the Classical Model is to come to an unbiased and empirically validated decision on how to combine the expert judgements. This step can help to overcome pre-judgements and exclusion of potentially knowledgeable individuals, as well arbitrary choices by analysts and decision makers about how to weight and aggregate experts. In analyses of 78 case studies using the Classical Model, performance weighted aggregations have outperformed equal weights in 76 studies (in-sample validation), suggesting the method can also be used to optimise aggregated judgements (Cooke and Goossens 2008, Colson and Cooke 2017).

2.3 Case study

To demonstrate how the Classical Model could be applied in ecology, and to investigate potential improvements from its application, we use estimates for ecological questions from a previous case study by Hemming et al. (2018b). In brief, the case study used the IDEA protocol with the four-step question format to elicit judgements for thirteen questions relating to future abiotic and biotic events on the Great Barrier Reef. The elicitation was undertaken via email and the experts volunteered their time. The questions related to the types of events experts may be asked in assessing risk to the Great Barrier Reef (Ward 2014), for example, the percentage cover of coral bleaching that may be detected in the next monitoring event at a specified reef (see Appendix S1: Section 1). The questions related to future monitoring events, so that judgements could be scored against outcomes once monitoring data were collected.

In total, 58 experts completed Round 2 of the elicitation exercise. These 58 individuals had been randomly assigned to one of eight groups within which judgments were aggregated. In Hemming et al. (2018b) the judgements were standardised to 80% credible intervals using linear extrapolation (outlined in Appendix S1: Section 1) and subsequently aggregated using an equal weighted quantile aggregation (taking the arithmetic mean) (refer to Appendix S1: Section 5). The judgements were then scored using performance measures of the IDEA protocol. The study found that 1) the equally weighted aggregate judgements were often more accurate and better calibrated than the median individual, 2) individuals could outperform the aggregation, however, they could not have been selected based on their credentials or demographic data, and 3) discussion and feedback led to improved final judgements (Appendix S1: Section 1). However, it was suggested further improvements may be made via performance weighted aggregation.

2.3.1 Four-step to quantiles

To make responses of the four-step question format compatible with requirements of the Classical Model (quantiles of a continuous non-parametric distribution), individual judgements need to be standardised to 90% credible intervals. We then assume (a) that the best estimate is the 50th percentile (i.e. a median), and, (b) upper and lower estimates represent a *central* credible interval (i.e. whereby the probability mass beyond a judgment's interval is apportioned equally above and below the upper and lower bounds, respectively). We interpret lower bounds as 5th quantiles and upper bounds as 95th quantiles.

In zero-inflated settings it is possible for respondents to provide a judgment of zero for both their 5th and 50th quantile (which occurred in our case study but is not consistent with a continuous distribution - refer to Appendix S1: Section 2). In such cases, a small number may be added or deducted to separate the quantiles. For example, zeros may be replaced by the following numbers depending on where in the estimate the zeros occur (Cooke 2018a):

- Lower / 5th : 0.00001
- Best / 50th : 0.0001
- Upper / 95th: 0.001

In our case study, we also encountered circumstances where the lower estimate, or best estimate reasonably coincided with the upper bounds which led to similar adjustments (see Appendix S1: Sections 2-3).

2.4 Scoring Judgements

Assuming the judgements approximate quantiles of a continuous probability distribution, the judgements can then be scored using the Classical Model's performance measures. There is substantial ambiguity and confusion in the ecological literature as to what the performance measures of the Classical Model actually reward. They have been cited as rewarding 'accuracy'

(Rothlisberger et al. 2009, Burgman et al. 2011a, Martin et al. 2012), which may give the impression they reward the accuracy of point estimates. They have also been noted to score 'calibration' and 'precision' (width) which may give the impression they are designed to assess interval judgements according to definitions that arise in the psychological literature (Lichtenstein and Fischhoff 1977, Yaniv and Foster 1997, Burgman et al. 2011a, Wittmann et al. 2015).

Verbal clarifications contained within the Classical Model literature often fail to clarify the reward structure, which may perpetuate misinterpretations. For example, statistical accuracy has been described as a measure of the likelihood that *"at least 7 out of 10 realisations should fall outside an expert's 90% confidence bands, if each value really had an independent 90% chance of falling inside the bands?"* (Rothlisberger et al. 2009, Colson and Cooke 2017). This may give the impression that it is designed primarily to score the calibration of the 90% credible interval judgements, rather than the calibration of the expert's interquantile ranges.

To better understand the reward structure of the Classical Model so that they are not misapplied we will contrast the performance measures for the Classical Model with those commonly used in the IDEA protocol (Hemming et al. 2018b). We outline these performance measures below. Equations and a worked example are provided in Appendix S1: Section 4.

IDEA performance measures

With the four-step question format in the IDEA protocol, individuals are scored by performance measures of *accuracy*, *calibration* and *informativeness* (Hemming et al. 2018b).

<u>*Accuracy*</u> is designed to assesses the accuracy of point estimates. It is the difference between b, the expert's best estimate, and the observed value, x. It is measured using the average log ratio error (ALRE) of expert responses. The measure is a relative measure, scale invariant, and emphasizes order of magnitude errors rather than linear errors. Smaller ALRE scores indicate

more accurate responses. For any given question the log ratio score has 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 (Burgman et al. 2011b)

Calibration is the proportion of intervals provided by the experts containing the realised truth relative to their assigned confidence (Lichtenstein and Fischhoff 1977, Lin and Bier 2008). For example, if the expert's intervals are standardised to 90% credible intervals then we expect for a well calibrated expert and 100 questions, that 90 of the realisations will fall between their 5^{th} and 95^{th} quantiles. If they capture fewer realisations, they may be considered overconfident, and if they capture more realisations they may be considered underconfident. The measure is an absolute measure and is scale invariant. If the realisations are equal to the expert's 5^{th} or 95^{th} quantiles, then they are usually assessed as being included within the expert's credible intervals.

Informativeness is used to denote a measure of the width (or precision) of the intervals provided by experts (Yaniv and Foster 1997). It is a relative measure and scale invariant. For each question, the expert's intervals are divided by a background range for the question, where the range is based on all estimates provided by the pool of experts for that question. Answers close to 0 indicate that an expert was highly informative, while a 1 would indicate the expert's uncertainty spanned the entire range of responses for that question. The final score for informativeness for an expert is their average across all questions.

Performance measures of the Classical Model

The Classical Model has two main performance measures that assess the ability of an expert to provide useful probability distributions, statistical accuracy and information.

Statistical accuracy (often referred to as calibration and often denoted by 'C') assesses the ability of experts to answer according to a theoretically optimal multinomial distribution. It assesses the interquantile calibration of experts. For example, over a set of questions for which realisations could be obtained, we would expect for any high performing expert that:

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

The expectation of where the realisations fall in relation to an expert's interquantile ranges can be expressed as a theoretical multinomial distribution $p=(0.05, 0.45, 0.45, 0.05)$ (Bedford and Cooke 2001). Under the Classical Model, the actual proportion of realisations within each inter-quantile range for each expert (or aggregation) e, is tallied to create a multinomial distribution for each expert: $s(e)= (Q_1, Q_2, Q_3, Q_4)$.

The realised distribution is then compared to the theoretical distribution using the Kullback-Leibler (KL) divergence measure and a chi-square test with three degrees of freedom. Statistical Accuracy is the *p*-value of this test. Higher values indicate an expert's distribution more closely matches the theoretical distribution. A statistical accuracy below 0.05 is often used as a cut-off point at which an expert is considered statistically inaccurate (i.e. Bamber et al. (2016), Colson and Cooke (2017)). The 0.05 level is often used in meta-analyses comparing the weighting and aggregation schemes in the Classical Model literature, but can also be used by the analyst as a cut-off point at which zero weight may be assigned to the expert's judgement.

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

We highlight this assumption as (on rare occasions) it can affect the score participants receive. For example, in the unlikely case that a participant was to estimate the median perfectly for 9 of 10 questions, they could obtain a multinomial distribution of $S(e)= (1, 9, 0, 0)$, which when compared to the theoretically optimal multinomial distribution means they would be considered statistically inaccurate at the 0.05 level, despite having perfect calibration and exceptional accuracy under the IDEA protocol scoring rules.

Information (often referred to as relative information, or informativeness) under the Classical Model measures the degree to which the expert's distribution is concentrated and to which it differs from a uniform or log-uniform distribution (which are considered the least informative distributions). It uses the KL divergence measure, which is scale invariant (Quigley et al. 2018). Information is calculated per question and does not depend on the realisation. The final information score of an expert is an average taken across all calibration questions. Larger numbers indicate better performance because they represent distributions which show greater departure from a uniform or log-uniform distribution.

A simple example contrasting the performance measures is provided in Box 1, and Figure 2. In the results section, we plot outcomes for these measures against each other to gain a better understanding of the underlying reward structures.

2.5 Weighting and aggregating

There are notable trade-offs between statistical accuracy and information in the Classical Model. By providing very wide intervals, an expert may achieve near perfect statistical accuracy, but will have low information (Quigley et al. 2018). Likewise, by providing very narrow intervals, they will have a high level of information, but usually at the cost of poor statistical accuracy. Ideally an expert should have both high statistical accuracy and information (Quigley et al. 2018). Therefore, the performance measures of the Classical Model are only proper if they are combined.

Under the Classical Model, the scores for statistical accuracy and information are combined to provide weights for each expert. There are five basic ways in which experts may be weighted and combined (equations provided in the Appendix S1):

Equal Weights (EW): is a linear pool of all expert distributions using the arithmetic mean of their distributions. It affords all experts the same weight regardless of how well they performed on calibration questions. It can be calculated without calibration questions.

Global Weights (GW): is calculated based on the combined statistical accuracy and information scores (CM Score) averaged across all calibration questions. Experts who performed better on the calibration questions are afforded more weight than those who performed poorly.

Itemised Weights (IW): uses the same statistical accuracy scores as Global Weights, however, the weight each expert is awarded will change per question because it considers the information of the expert for each question of interest rather than the average calculated based on all of the calibration questions. This often leads to aggregations with higher information (and informativeness) on average than Global Weights.

Global Weights Optimised (GWO) and *Itemised Weights Optimised (IWO)*: are similar to their un-optimised variants described above (i.e. Global Weights (GW) and Itemised Weights (IW)). However, they optimise the statistical accuracy score by successively raising the level at which an expert is considered statistically inaccurate from an alpha level equal to the lowest calibration score. The weights are calculated and used to generate weighted aggregations that are scored on the calibration questions. The weighted aggregation with the highest performance on the calibration questions is chosen (Quigley et al. 2018). In decisions with one or two well calibrated experts, most or all of the weight may be assigned to those experts with no weight given to the other experts.

For a set of calibration questions, an analyst may create a set of pooled judgements for each question under each weighting scheme. These pooled judgements can then be scored for their statistical accuracy and information (i.e. in-sample validation). These scores are then multiplied to create an overall score, which we term the Classical Model (CM) Score. The aggregation method which produces the highest CM Score on the calibration questions is usually taken as the preferred weighting scheme when combining expert judgements on the questions of interest (for which answers are not known). If two aggregations result in the same statistical accuracy, that with a higher information score is preferred (Bedford and Cooke 2001).

2.5.1 Linear pooling versus quantile aggregation

The Classical Model uses linear pooling of distributions for both equal weighted and performance weighted aggregations, which differs from quantile aggregation commonly used by the IDEA protocol when the four-step question format is used (Hemming et al. 2018a) (refer to Appendix S1: Section 5 for discussion and a worked example).

Quantile aggregation is simple to apply, and entails no additional assumptions about what the estimates represent beyond a best estimate with a credible interval. In general, it provides more accurate and better calibrated judgements compared to the best-regarded experts (Burgman et al. 2011b, Hemming et al. 2018b). However, Bamber et al. (2016) and Colson and Cooke (2017) found that quantile aggregation is much more overconfident than linear pooling (when assessed using the Classical Model's Statistical Accuracy measure). To investigate these findings, we extend our analysis to compare how the two methods of equally weighted aggregation can affect judgements. Henceforth we use the term 'equal weights' (abbreviated to EW) to refer to linear pooling of distributions, and 'quantile aggregation' (abbreviated to QuA) to refer to quantile aggregation.

2.6 Analysis

For the eight groups of experts in our case study, we assessed the six alternative approaches to aggregation (two forms of equal weighted aggregations (EW (Classical Model), QuA (IDEA)), and four forms of performance weighted aggregation from the Classical Model (IW, GW, IWO, GWO) (described in Section 2.5). Individual and group performance was assessed using the five performance measures (statistical accuracy, information, calibration, informativeness, and accuracy) (described in Section 2.4), and the Classical Model scoring rule (CM Score)(described in Section 2.5).

To obtain the performance measures and aggregations associated with the Classical Model, the analyst must enter judgements in software called *Excalibur* (Lightwist 2013, Cooke

2018a)(Appendix S1: Section 3). For measures associated with the IDEA protocol we developed *R*-code (available on the Open Science Framework (Hemming 2019)). More details are available in Appendix S1: Section 3.

To contrast the differences of the aggregations, we use boxplots, constructed in *R* (version 3.4.1 (2017-06-30) -- "Single Candle"), using the *ggplot2* package. The boxes represent the 25th, 50th and 75th percentiles. The whiskers represent the spread of the data referenced on the inter-quartile range, (Q1-1.5*IQR, Q3+1.5*IQR). For normally distributed data this is approximately 2.7 standard deviations, or 99.3% of the data (Krzywinski and Altman 2014).

3 Results

3.1 Comparison of performance measures

In Figure 3, we plot the two performance measures underpinning weights obtained under the Classical Model for the 58 participants. When scored on statistical accuracy, less than half (23) of participants were statistically accurate (obtaining scores higher than 0.05). For 13 questions the highest possible statistically accuracy score would have been 0.93. No individuals achieved this score (highest statistical accuracy score was 0.53).

High statistical accuracy usually came at the expense of lower information. Participants who were statistically accurate were more likely to have lower information scores. Such observations reflect the trade-offs between statistical accuracy and information discussed by Quigley et al. (2018) and re-enforce the need to combine the two measures to derive a proper scoring rule.

Figure 4 shows the scatter between statistical accuracy (the Classical Model) and calibration (IDEA Protocol) for the 58 participants over 13 questions. While the difference in the highest statistical accuracy score possible and that obtained by experts appears large (i.e. a change from

0.93 to 0.53 implies a 43% reduction in statistical accuracy), we can see that this change was due to just one additional realisation falling outside of the experts' credible intervals. Thus, statistical accuracy can be highly sensitive to seemingly small variations in performance.

Figure 4 also shows that while there is a positive correlation between the two measures (Spearman rank correlation= 0.84, 95% CI: 0.74, 0.90) there are also some notable differences. Importantly, an expert may have near perfect calibration under the scoring rules employed by the IDEA protocol, but be statistically inaccurate at the 0.05 level according to the Classical Model. These results further clarify that statistical accuracy does not reward calibration primarily between the expert's 90% credible intervals, on which IDEA's calibration depends.

In Appendix S1: Section 7, we demonstrate that the differences occur because the Classical Model's rules score a multinomial distribution with three degrees of freedom p= (0.05, 0.45, 0.45, 0.05). As such, beyond very low levels of calibration (i.e. <50% calibration for 13 questions), the statistical accuracy measure cannot be used to assess the calibration of 90% credible intervals (i.e. a multinomial distribution with one degree of freedom, or a binomial distribution).

Figure 5 shows the correlation between information (Classical Model) and informativeness (IDEA Protocol). The two scores are negatively correlated (Spearman rank correlation = -0.69, 95%CI: -0.80, -0.52), an artefact of the scoring rules, whereby under the IDEA protocol participants who receive a low score are more informative (narrower intervals), whereas for the Classical Model a higher score indicates that they provide more information relative to a uniform or log-uniform distribution. Figure 5 demonstrates that information and informativeness are slightly different measures of an expert's judgement. The Classical Model does not only assess the width of intervals, it also accounts for their departure from a uniform distribution. This can mean that higher information score may be obtained in some cases simply by reducing the symmetry of the ranges between an expert's 2^{nd} and 3^{rd} quantiles (i.e. if the median does not fall squarely in the centre of the range then information can be increased).

3.2 Performance of aggregations

Figure 6 shows the CM Score for each of the aggregations. In the Classical Model, this combined score would be used to select the final aggregation for uncertainty by a decision-maker. For this case study, if we were to use the median values of these scores, we would not choose quantile aggregation (QuA) because it has a low CM Score (median value of 0.14). Equally weighted linear pooling employed by the Classical Model does better (EW) (median value of 0.41), and there is some indication that performance weighted aggregation by the optimised variants (IWO, and GWO) may lead to further improvements (median values of 0.60 and 0.50 respectively).

Figures 7a and 7b decompose the CM Score provided in Figure 6, into statistical accuracy and information scores of the Classical Model. While quantile aggregation (QuA) performs well on information (median value of 1.51), it performs poorly in terms of statistical accuracy (median value of 0.10) compared to equal weights (EW) (median value of 0.36) and performance weighted aggregations (IW, IWO, GW, GWO) (all achieving a median value of 0.36, except for IW which achieves 0.27), with two groups considered statistically inaccurate at the 0.05 level. This supports the finding by Bamber et al. (2016) and Colson and Cooke (2017) that quantile aggregation used in the IDEA protocol with four-step question format can be overconfident relative to linear-pooling of distributions when assessed by statistical accuracy.

There is little or no difference in the median performance of equally weighted (EW) and the performance weighted aggregations (IW, IWO, GW, GWO) in terms of statistical accuracy. However, both the optimised aggregations (GWO, and IWO) and itemised weights (IW) have higher information (1.56 and 1.68) than equal weights (EW, median of 1.14) or global weights (GW, median of 1.18), and are equivalent to quantile aggregation (QuA, 1.51) suggesting performance weighting improves estimates in this case study by being more informative than equal weights (EW).

Figures 7c-e assess each of the aggregation methods according to measures commonly used in the IDEA protocol. Even when scored according to calibration between the expert's 90% credible intervals, the study finds that quantile aggregation (QuA) generates more overconfident estimates (median calibration of 0.77), having a lower calibration than all other aggregations (0.85, or on average by one question). It does, however, have a higher level of informativeness (0.25) than all other aggregations (medians ranging between 0.33 and 0.42), including optimised aggregations. The median accuracy of the best estimate is better for all aggregations than the median ranked individual for this measure. However, the optimised aggregations have some groups which perform worse than the median individual. This may not be surprising because (as discussed) the Classical Model was not designed to optimise point estimates.

Quantile aggregation (QuA) performed relatively poorly on statistical accuracy and calibration (Figure 7). Recall that some questions related to count data, and the upper and lower bounds were adjusted so that they did not contain zero. In our case study, the lowest estimate which could be provided by an expert was 0.00001. This adjustment may have led to overconfident judgements for two questions which contained zeros.

To check this, we replaced the answers for these two questions with 0.000011 and re-calculated the calibration and statistical accuracy of judgements of each of the groups (Figure 8, see also Appendix S1: Section 8). The adjustment improved the statistical accuracy of many groups across all aggregations. All but one aggregation (quantile aggregation, QuA) had a median statistical accuracy above 0.53 (Figure 8a). Only one group was considered statistically inaccurate when their judgements were combined via quantile aggregation (QuA).

Quantile aggregation (QuA) was overconfident, even when assessed according to calibration of interval judgements but many groups were less so than prior to accounting for the two questions with zeros (Figure 8b). Group judgements for quantile aggregation achieved good but not perfect median calibration of 0.76, although no group reached perfect calibration when quantile aggregation (QuA) was used. In contrast, each of the linear pooled distributions except for the itemised optimised weights achieved a median group calibration of 0.90 (i.e. perfect calibration). The data adjustments improved calibration and statistical accuracy, but they did not substantially alter the information or informativeness scores which meant that quantile aggregation (QuA) was still substantially more informative that the equal weights (EW) (a median informativeness score of 0.24 compared to 0.42).

4 Discussion

Performance weights have been proposed to improve expert judgements in ecology. However, there have been few applications and little discussion of their strengths and weaknesses in the ecological literature. Here, we outlined the key rationales and theories of performance weights, then described one of the most well-known methods, the Classical Model (Cooke 1991), and examined how it might be applied to improve judgements derived from the IDEA protocol with four-step question format (Hemming et al. 2018a, Hemming et al. 2018b).

This study highlighted how the Classical Model and the IDEA protocol may be integrated, but clarified important differences between them that should be considered before applying performance weights.

The four-step question format needs to first be converted into quantiles of a continuous probability distribution. It may be better to remove these assumptions by eliciting these quantiles directly. However, the four-step question format is often used because it helps to overcome overconfidence relative to eliciting fixed intervals (Speirs-Bridge et al. 2010), and because experts who are unfamiliar with the language of statistical distributions are comfortable in providing quantitative judgements of uncertainty (a problem not only encountered in ecological domains (Walls and Quigley 2001, Hirsch et al. 2004)). These trade-offs need to be considered when deciding how best to elicit estimates. If the four-step question format is to be used with the Classical Model, then we suggest that the assumptions about how the estimates will be interpreted are communicated to experts in introductory material and through the feedback and discussion stages of the IDEA protocol.

Once judgements were converted into quantiles of a continuous probability distribution, we described key steps required to incorporate the judgements into *Excalibur* and to generate scores and aggregations for the Classical Model (outlined in more detail in the Appendix S1: Sections 2-3). These steps have not been substantially documented in the literature, inhibiting use of performance weights. The advice outlined here will make implementation of the method more accessible to those unfamiliar with the Classical Model and improve efficiencies when analysing data.

We then described the performance measures underpinning the Classical Model, noting that there was considerable ambiguity in the literature as to how the Classical Model rewards judgements, with terms such as “calibration”, “accuracy”, information”, and “overconfidence” being differently interpreted (Rothlisberger et al. 2009, Burgman et al. 2011b, Metcalf and Wallace 2013, Wittmann et al. 2015, Colson and Cooke 2017).

Insights from our results emphasise that the Classical Model was designed to assess probability distributions rather than point estimates or interval judgements (as some interpretations suggest). Specifically, ‘statistical accuracy’ measures the degree to which an expert’s multinomial distribution matches a theoretically optimal multinomial distribution, and ‘information’ measures the departure from a uniform or log-uniform background measure. As such the Classical Model is not focused primarily on avoiding surprises outside of the 90% confidence intervals, or the precision of the intervals (as assessed in the IDEA protocol) and may lead to counterintuitive outcomes in settings where this is a primary concern.

The question therefore arises as to when each performance measure may be more appropriate? Calibration, informativeness and accuracy (as scored in the IDEA protocol) tend to be important in the contexts of risk assessments and structured decision-making in which decision-makers are deciding to take action, and are using the best estimate to understand the most likely scenario, or the uncertainty bounds to investigate how sensitive their decisions are to different risk attitudes (Gregory et al. 2012, Addison et al. 2015). In other words, the measures normally associated with IDEA may be most useful when assessing the outputs of a model or risk analysis (Morgan and Henrion (1990), page 78).

On the other hand, it may be more important to understand the calibration within the expert’s interquartile ranges (i.e. the 2^{nd} and 3^{rd} quantiles) (as scored by the Classical Model) when they estimate probability distributions as inputs to a model, for example sampling in Monte Carlo simulations, especially where tail risks are a key concern (Morgan and Henrion (1990), page 78).

While calibration and informativeness of interval judgements may be of interest they have not yet been combined into a proper scoring rule (although telling experts they will be scored on both should minimise gaming behaviour). Our results demonstrate that the Classical Model does not by itself provide this information, which may be disappointing to those who seek to apply the Classical Model to optimise or assess such judgements. However, if this information was of interest the performance measures of the IDEA protocol may be used to provide this information. Agreement as to which performance measures will be used should be made prior to application.

Equal weighted aggregations are often used in ecology when combining expert judgments. However, there are numerous methods by which an equal weighted aggregation can be derived, and not all will perform equally well or have been validated. We contrasted two forms of equal weighted aggregation, quantile aggregation (QuA, used in Hemming et al. 2018), and equal weighting via linear pooling of distributions (EW, used by the Classical Model). We found that both forms of equal weighted aggregation were better than the median ranked individual for each measure of statistical accuracy, calibration, and accuracy. Furthermore, as was demonstrated in Hemming et al. (2018b), while some individuals could outperform the group aggregation they could not be predicted by standard metrics of expertise (years of experience, peer-recommendation, or self-rating). This suggests that taking the equal weighted aggregation is a more robust method than trying to select a single expert with good judgement based on their credentials and status.

Our results corroborate those of the Bamber et al. (2016) and Colson and Cooke (2017), that while quantile aggregation is simpler to apply, and was more informative, it led to overconfident estimates compared to linear pooling of equally weighted distributions, and performance weighted distributions (Figure 7). This was true regardless of whether we assessed the judgements based on calibration or statistical accuracy.

We found that the degree of overconfidence was reduced when we accounted for questions with zeros, and the way in which the Classical Model accounts for realisations which equal a participant's estimates (i.e. if the realisation coincides with the lower bound it will be considered as falling outside of the expert's 90% credible intervals). As these adjustments are not made when the four-step question format is used in the IDEA protocol, the degree of overconfidence from quantile aggregation may not typically be as severe for many applications of the IDEA protocol. Nonetheless, we would suggest these findings warrant further investigation on more case studies with the four-step question-format.

We then examined how performance weighting could be used to improve aggregated judgements. We found that there was little difference in the calibration or statistical accuracy of performance weighting and equal weighted linear pooled distributions. However, performance weighting produced more informative bounds than equal weighted linear pooling (by 10% of the background range when measured according to informativeness). These results suggest that if the aim is to reduce arbitrary uncertainty while achieving well-calibrated intervals, then performance weights can better achieve this.

In our study, we demonstrated a modest improvement by performance weighted aggregation. However, we note that there is no guarantee that performance weighted aggregation will lead to improvements in all cases. However, a clear advantage of the Classical Model, and other methods which utilise calibration questions is that they provide empirical evidence for the legitimacy of final aggregations (often lacking in studies that use expert judgement). This is especially important because decisions regarding who should be included in an elicitation and how to aggregate these judgements may exclude potentially knowledgeable individuals, and often lack validation.

Whether or not the decision context justifies (or can afford) the additional time and expense ultimately depends on the context of the case study, the decision-maker and the value of additional information. Wittmann et al. (2015) and Rothlisberger et al. (2009) justify their application based on the immense value of fisheries to the Great Lakes and the possibility of litigation following mismanagement. This suggests that there are contexts in ecology in which this additional time and expense can be justified. If resources are not available to deploy calibration questions and performance weighted aggregation, then our study shows that an equal weighted aggregation (i.e. quantile aggregation or linear pooling of distributions) provides an effective means to improve judgements relative to selecting a single seemingly well-credentialed expert.

However, there are obstacles to wider uptake of performance weighting and lines for further research. We found it difficult to develop questions about future events on the Great Barrier Reef for which we could obtain data in a reasonable time (3-6 months). Despite the substantial amount of monitoring which takes place there (GBRMPA 2014). Others have noted problems in obtaining access to ecological datasets (Meek et al. 2015). It may be possible to use existing datasets to generate calibration questions. However, especially with remote elicitation, there will always be a risk that experts discover the sources of the data when forming their judgements (as occurred in (Hemming et al. 2019a)).

We found that questions relating to count data (particularly where the realisations are often zero inflated) should be avoided when using the Classical Model. In ecology, zero inflated count data are common (Martin et al. 2005).

Calibration questions should be related to target variables, for which the answer is known or will become known (Cooke and Goossens 2000). However, ascertaining whether or not a question is *relevant* in many domains may be difficult because domains are often ill-defined, making the selection of *relevant* questions a subjective decision (Colyvan and Ginzburg 2003). If datasets are difficult to obtain, then the analyst may need to rely on past questions for which the data are available, or questions which are less relevant to the questions of interest. It would be useful to understand at what point calibration questions become so distantly related to target questions that in-sample validation is not a good predictor of performance.

We used *Excalibur* to generate aggregations and score experts, however, the program was challenging to use. The analysis was time consuming and it was difficult to provide a reproducible workflow for our analysis. The methods of aggregation and the scoring rules should be simple enough to re-code in *R* and other freely available software (we note that recently they have been re-coded in *MATLAB* (Leontaris and Morales-Nápoles 2018)). A revision of *Excalibur* could help to increase adoption of the method.

Our study explored the effect of performance weights using in-sample validation (i.e. on the same questions used to score experts and generate aggregations) for one case study. However, the ideal test is how well it performs out-of-sample (i.e. on questions not used in the training set) (Clemen 2008). This has not been addressed by this study. When Colson and Cooke (2017) addressed this question they found some differences in out-of-sample performance that were not revealed by in-sample validation and suggested this would be the focus of further research.

The scoring rules and aggregation methods of the Classical Model may not always be well-understood. To avoid confusion, we suggest that in future, statistical accuracy scores should be accompanied by their corresponding multinomial distributions. We provide *R* and *MATLAB* code for this (Hemming et al. 2019b). While, it's less easy to convey the reward structure of the information score, we believe it would be useful to display the intervals of the aggregations so that the relative improvements can be compared (this is already often presented in applications of the Classical Model).

5 Conclusions

Performance weighted aggregations with calibration questions has been proposed as a means to improve expert judgements in ecology, however, applications have been scarce. We explored how the Classical Model could be applied to the IDEA protocol with four-step question format.

Our study found that the Classical Model could be applied to the IDEA protocol with four-step question format provided the values of the four-step elicitation can be assumed to represent quantiles of a continuous distribution. A key finding of this paper is that the reward structures embedded in the performance measures of the two approaches to elicitation are often confused and differ in important ways. This should be understood prior to application to ensure that the methods for optimisation match the decision-maker's preferences and problem setting.

We demonstrated that equal weighted aggregations can achieve relatively well-calibrated aggregated judgements. However, linear pooling of distributions may produce better calibrated but less informative distributions than quantile aggregation as found by Bamber et al. (2016) and Colson and Cooke (2017). We found that performance weighted aggregations can outperform equal weighted aggregations, in our case by providing more informative judgments, however, we emphasise that there is no guarantee they will do so in every case. The main reason that the candidate alternatives for aggregation should be explored is to ensure the final representation of uncertainty is the best possible (whether that be via equal weights or performance weights).

Whether the time and investment in applying performance weights is worth the benefits is ultimately a matter of context. Our example illustrates that there are contexts in which this additional time and effort may be justified.

Our paper will help ecologists to better understand the fundamental steps, challenges, and advantages involved in deploying performance weighted aggregation, and to avoid common pitfalls which may arise. We welcome more research to understand how these methods could be adapted to better suit the practical and financial constraints of a wider range of ecological applications and estimates (i.e. point estimates, interval judgements, and single event probabilities).

6 Acknowledgements

The authors would like to thank the experts who volunteered their time for the case study presented. VH received funding to draft this publication by the Australian Research Training Program, and the David Hay Memorial Fund. VH and AH were funded by the Australian Centre of Excellence for Biosecurity Risk Analysis, VH, AH, TW and MB were funded by the School of BioSciences at the University of Melbourne, MB was also funded by Centre for Environmental Policy, Imperial College London.

7 Literature citations

Adams-Hosking, C., M. F. McBride, G. Baxter, M. Burgman, D. de Villiers, R. Kavanagh, I. Lawler, D. Lunney, A. Melzer, P. Menkhorst, R. Molsher, B. D. Moore, D. Phalen, J. R. Rhodes, C. Todd, D. Whisson, and C. A. McAlpine. 2016. Use of expert knowledge to elicit population trends for the koala (*Phascolarctos cinereus*). Diversity and Distributions **22**:249-262.

Addison, P. F. E., K. de Bie, and L. Rumpff. 2015. Setting conservation management thresholds using a novel participatory modeling approach. Conservation Biology **29**:1411-1422.

Armstrong, J. S. 2001. Combining forecasts. Pages 417-439 *in* J. S. Armstrong, editor. Principles of forecasting: A handbook for researchers and practitioners. Springer US, Boston, United States of America.

Aspinall, W. P., and R. M. Cooke. 2013. Quantifying scientific uncertainty from expert judgement elicitation. Pages 64-99 *in* J. Rougier, S. Sparks, and L. Hill, editors. Risk and Uncertainty Assessment for Natural Hazards. Cambridge University Press, Cambridge, United Kingdom.

Bamber, J., W. Aspinall, and R. Cooke. 2016. A commentary on “how to interpret expert judgment assessments of twenty-first century sea-level rise” by Hylke de Vries and Roderik SW van de Wal. Climatic Change **137**:321-328.

Barons, M. J., A. M. Hanea, S. K. Wright, K. C. Baldock, L. Wilfert, D. Chandler, S. Datta, J. Fannon, C. Hartfield, and A. Lucas. 2018. Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation. Journal of Apicultural Research:1-12.

Bedford, T., and R. M. Cooke. 2001. Mathematical tools for probabilistic risk analysis. Cambridge University Press, Cambridge, United Kingdom.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. Monthly weather review **78**:1-3.

Budescu, D. V., and E. Chen. 2014. Identifying expertise to extract the wisdom of crowds. Management Science **61**:267-280.

Burgman, M., A. Carr, L. Godden, R. Gregory, M. McBride, L. Flander, and L. Maguire. 2011a. Redefining expertise and improving ecological judgment. Conservation Letters **4**:81-87.

Burgman, M. A. 2004. Expert frailties in conservation risk assessment and listing decisions. Pages 20-29 *in* P. Hutchings, D. Lunney, and C. Dickman, editors. Threatened species legislation: is it just an Act? Royal Zoological Society, Mosman, NSW, Australia.

Burgman, M. A. 2015. Trusting Judgements: How to get the best out of experts. Cambridge University Press, Cambridge, United Kingdom.

Burgman, M. A., M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle, F. Fidler, L. Rumpff, and C. Twardy. 2011b. Expert status and performance. PLoS One **6**:1-7.

Carwardine, J., T. G. Martin, J. Firn, R. P. Reyes, S. Nicol, A. Reeson, H. S. Grantham, D. Stratford, L. Kehoe, and I. Chadès. 2019. Priority Threat Management for biodiversity conservation: A handbook. Journal of Applied Ecology **56**:481-490.

Clemen, R. T. 2008. Comment on Cooke's classical method. Reliability Engineering & System Safety **93**:760-765.

Clemen, R. T., and R. L. Winkler. 1999. Combining probability distributions from experts in risk analysis. Risk Analysis **19**:187-203.

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.

Colyvan, M., and L. R. Ginzburg. 2003. Laws of nature and laws of ecology. Oikos **101**:649-653.

Cooke, R. 2018a. Macro converting XL file to EXCALIBUR dtt file.*in* R. Cooke, editor. http://rogermcooke.net/

Cooke, R., and L. Goossens. 2000. Procedures guide for structural expert judgement in accident consequence modelling. Radiation Protection Dosimetry **90**:303-309.

Cooke, R. M. 1991. Experts in uncertainty: Opinion and subjective probability in science. Oxford University Press, New York.

Cooke, R. M. 2018b. Validation in the Classical Model. Pages 37-59 *in* L. C. Dias, A. Morton, and J. Quigley, editors. Elicitation: The science and art of structuring judgement. Springer International Publishing, Cham, Switzerland.

Cooke, R. M., and L. L. Goossens. 2008. TU Delft expert judgment data base. Reliability Engineering & System Safety **93**:657-674.

Drescher, M., A. Perera, C. Johnson, L. Buse, C. Drew, and M. Burgman. 2013. Toward rigorous use of expert knowledge in ecological research. Ecosphere **4**:1-26.

Einhorn, H. J., R. M. Hogarth, and E. Klempner. 1977. Quality of group judgment. Psychological Bulletin **84**:158.

Estévez, R. A., F. O. Mardones, F. Álamos, G. Arriagada, J. Carey, C. Correa, J. Escobar-Dodero, Á. Gaete, A. Gallardo, and R. Ibarra. 2019. Eliciting expert judgements to estimate risk and protective factors for Piscirickettsiosis in Chilean salmon farming. Aquaculture.

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

GBRMPA. 2014. Great Barrier Reef Outlook Report. Townsville, Australia.

Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association **102**:359-378.

Goossens, L. H. J., R. M. Cooke, A. R. Hale, and L. Rodić-Wiersma. 2008. Fifteen years of expert judgement at TUDelft. Safety Science **46**:234-244.

Gregory, R., L. Failing, M. Harstone, G. Long, T. McDaniels, and D. Ohlson. 2012. Structured Decision Making: a practical guide to environmental management choices, Chichester, West Sussex.

Gregory, R., and R. L. Keeney. 2017. A Practical Approach to Address Uncertainty in Stakeholder Deliberations. Risk Analysis **37**:487-501.

Hanea, A., M. McBride, M. Burgman, B. Wintle, F. Fidler, L. Flander, B. Manning, and S. Mascaro 2016. InvestigateDiscussEstimateAggregate for structured expert judgement. International journal of forecasting **33**:267-269.

Hemming, V. 2019. Code: Weighting and Aggregating Expert Ecological Judgements. The Open Science Framework. DOI 10.17605/OSF.IO/FXQVK. http://osf.io/fxqvk

Hemming, V., N. Armstrong, M. A. Burgman, and A. M. Hanea. 2019a. Improving expert forecasts in reliability: Application and evidence for structured elicitation protocols. Quality and Reliability Engineering International **n/a**.

Hemming, V., M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle. 2018a. A practical guide to structured expert elicitation using the IDEA protocol. Methods in Ecology and Evolution **9**:169-181.

Hemming, V., S. Lane, and A. Hanea. 2019b. Classical Model Calculator. The Open Science Framework. DOI 10.17605/OSF.IO/BGYZU. http://osf.io/bgyzu

Hemming, V., T. V. Walshe, A. M. Hanea, F. Fidler, and M. A. Burgman. 2018b. Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. PLoS One **13**:e0198468.

Hirsch, K. G., J. J. Podur, R. F. Janser, R. S. McAlpine, and D. L. Martell. 2004. Productivity of Ontario initial-attack fire crews: results of an expert-judgement elicitation study. Canadian Journal of Forest Research **34**:705-715.

Hogarth, R. M. 1978. A note on aggregating opinions. Organizational Behavior and Human Performance **21**:40-46.

Hora, S. C. 2004. Probability judgments for continuous quantities: Linear combinations and calibration. Management Science **50**:597-604.

Hudson, E. G., V. J. Brookes, and M. P. Ward. 2017. Assessing the risk of a canine rabies incursion in Northern Australia. Frontiers in Veterinary Science **4**:141.

Krzywinski, M., and N. Altman. 2014. Points of Significance: Visualizing samples with box plots. Nat Meth **11**:119-120.

Kuhnert, P. M., T. G. Martin, and S. P. Griffiths. 2010. A guide to eliciting and using expert knowledge in Bayesian ecological models. Ecology Letters **13**:900-914.

Larrick, R. P., and J. B. Soll. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. Management Science **52**:111-127.

Leontaris, G., and O. Morales-Nápoles. 2018. ANDURIL — A MATLAB toolbox for ANalysis and Decisions with UnceRtaInty: Learning from expert judgments. SoftwareX **7**:313-317.

Lichtendahl Jr, K. C., Y. Grushka-Cockayne, and R. L. Winkler. 2013. Is it better to average probabilities or quantiles? Management Science **59**:1594-1611.

Lichtenstein, S., and B. Fischhoff. 1977. Do those who know more also know more about how much they know? Organizational Behavior and Human Performance **20**:159-183.

Lightwist. 2013. Excalibur. http://www.lighttwist.net/wp/excalibur

Lin, S.-W., and V. M. Bier. 2008. A study of expert overconfidence. Reliability Engineering & System Safety **93**:711-721.

Lorenz, J., H. Rauhut, F. Schweitzer, and D. Helbing. 2011. How social influence can undermine the wisdom of crowd effect. Proceedings of the National Academy of Sciences **108**:9020-9025.

Low Choy, S., R. O'Leary, and K. Mengersen. 2009. Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models. Ecology **90**:265-277.

MacDonald, J. A., M. J. Small, and M. Morgan. 2008. Explosion probability of unexploded ordnance: expert beliefs. Risk Analysis **28**:825-841.

Martin, T. G., M. A. Burgman, F. Fidler, P. M. Kuhnert, S. Low-Choy, M. McBride, and K. Mengersen. 2012. Eliciting expert knowledge in conservation science. Conservation Biology **26**:29-38.

Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters **8**:1235-1246.

McBride, M. F., F. Fidler, and M. A. Burgman. 2012a. Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research. Diversity and Distributions **18**:782-794.

McBride, M. F., S. T. Garnett, J. K. Szabo, A. H. Burbidge, S. H. Butchart, L. Christidis, G. Dutson, H. A. Ford, R. H. Loyn, and D. M. Watson. 2012b. Structured elicitation of expert judgments for threatened species assessment: a case study on a continental scale using email. Methods in Ecology and Evolution **3**:906-920.

Meek, M. H., C. Wells, K. M. Tomalty, J. Ashander, E. M. Cole, D. A. Gille, B. J. Putman, J. P. Rose, M. S. Savoca, and L. Yamane. 2015. Fear of failure in conservation: the problem and potential solutions to aid conservation of extremely small populations. Biological Conservation **184**:209-217.

Mellers, B., E. Stone, T. Murray, A. Minster, N. Rohrbaugh, M. Bishop, E. Chen, J. Baker, Y. Hou, and M. Horowitz. 2015. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science **10**:267-281.

Metcalf, S. J., and K. J. Wallace. 2013. Ranking biodiversity risk factors using expert groups – Treating linguistic uncertainty and documenting epistemic uncertainty. Biological Conservation **162**:1-8.

Morgan, M. G., and M. Henrion. 1990. Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis Cambridge University Press. New York, NY, United States of America.

Quigley, J., A. Colson, W. Aspinall, and R. M. Cooke. 2018. Elicitation in the Classical Model. Pages 15-36 *in* L. C. Dias, A. Morton, and J. Quigley, editors. Elicitation: The science and art of structuring judgement. Springer International Publishing, Cham, Switzerland.

Rothlisberger, J. D., D. M. Lodge, R. M. Cooke, and D. C. Finnoff. 2009. Future declines of the binational Laurentian Great Lakes fisheries: the importance of environmental and cultural change. Frontiers in Ecology and the Environment **8**:239-244.

Satopää, V. A., J. Baron, D. P. Foster, B. A. Mellers, P. E. Tetlock, and L. H. Ungar. 2014. Combining multiple probability predictions using a simple logit model. International journal of forecasting **30**:344-356.

Shanteau, J., D. J. Weiss, R. P. Thomas, and J. C. Pounds. 2002. Performance-based assessment of expertise: How to decide if someone is an expert or not. European Journal of Operational Research **136**:253-263.

Speirs-Bridge, A., F. Fidler, M. McBride, L. Flander, G. Cumming, and M. Burgman. 2010. Reducing overconfidence in the interval judgments of experts. Risk Analysis **30**:512-523.

Surowiecki, J. 2004. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations. Little, Brown, London, United Kingdom.

Tetlock, P. 2005. Expert political judgement: How good is it? How can we know? Princetown University Press, Princetown, United States of America.

Tetlock, P., and D. Gardner. 2015. Superforecasting: The art and science of prediction. Random House, New York.

Walls, L., and J. Quigley. 2001. Building prior distributions to support Bayesian reliability growth modelling using expert judgement. Reliability Engineering & System Safety **74**:117-128.

Wallsten, T. S., D. V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth. 1986. Measuring the vague meanings of probability terms. Journal of Experimental Psychology: General **115**:348.

Ward, T. 2014. The rapid assessment workshop to elicit expert consensus to inform the development of the Great Barrier Reef Outlook Report 2014. Townsville.

Weiss, D. J., and J. Shanteau. 2004. The vice of consensus and the virtue of consistency. Psychological investigations of competent decision making:226-240.

Winkler, R. L., and A. H. Murphy. 1968. "Good" probability assessors. Journal of applied Meteorology **7**:751-758.

Wittmann, M. E., R. M. Cooke, J. D. Rothlisberger, E. S. Rutherford, H. Zhang, D. M. Mason, and D. M. Lodge. 2015. Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. Conservation Biology **29**:187-197.

Yaniv, I., and D. P. Foster. 1997. Precision and accuracy of judgmental estimation. Journal of Behavioral Decision Making **10**:21-32.

8 Data availability statement

Open Access: All data and code for the analyses presented in this paper are available on the Open Science Framework (Hemming 2019).

Box 1 Scoring rules IDEA protocol vs. The Classical Model

In Figure 2, two experts have been asked to provide their estimates for 10 calibration questions. They have then been scored on their performance using the scoring rules outlined Section 2.4 from the Classical Model and the IDEA protocol.

Statistical accuracy (Classical Model) vs Calibration (IDEA)

Expert A, has an inter-quantile distribution of *s(A)= (0.10, 0.40, 0.40, 0.10)*, that is, over 10 questions one realisation fell below their 5^{th} interval, four between their 5^{th} and their 50^{th}, four between their 50^{th} and their 95^{th}, and one above their 95^{th}. When compared to the theoretically optimal inter-quantile distribution of *p= (0.05, 0.45, 0.45, 0.05),* using a chi-squared test with three-degrees of freedom they receive a statistical accuracy (SA) of 0.83, which is the highest statistical accuracy that can be achieved on 10 questions.

Expert B, provides a theoretical distribution *s(B)= (0.10, 0.90, 0.0, 0.0)*, which is quite different to the theoretically optimal inter-quantile distribution *p*. Their statistical accuracy is low, 0.003. Having a statistical accuracy below 0.05 they would be deemed statistically inaccurate under the Classical Model.

In contrast, when scored using calibration (CA) from the IDEA protocol, Expert B would be perfectly calibrated having nine of their ten 90% credible intervals capturing the realised truth. Expert A would also be considered well-calibrated, but less so than Expert B, only capturing eight out of 10 realisations in their 90% credible intervals.

Information (Classical Model) vs Informativeness (Four-step question format)

Expert A and B provide intervals which are exactly the same width for each question. However, Expert B consistently provides a median close to the tails. This means the mass of their intervals departs from a uniform distribution whereby we would expect 5% of the total width of their interval to fall below their 5^{th} quantile, 45% between their 5^{th} and 50^{th}, and again between their 50^{th} and 95^{th}, and 5% above their 95^{th} quantile. Assuming this is the only difference in their intervals, Expert B would achieve a higher information score under the Classical Model than Expert A. However, as experts have intervals that are the same width, both experts would receive the same score for informativeness under the IDEA protocol.

Figure 1 Key steps of the IDEA protocol (figure from Hemming et al. (2018b)). The four-step question format (Speirs-Bridge et al. 2010) (depicted in Step 2) is commonly used to derive a best estimate and credible interval in Round 1 and Round 2.

Figure 2 Judgements provided by two hypothetical experts over 10 questions. The blue lines represent their 90% credible intervals, the blue dots their 'best estimate' or their 'median'. The crosses represent where the realisation fell in relation to their estimates. To calculate statistical accuracy (SA) according to the Classical Model, the proportion of questions answered where realisations fell, 1) below their lowest interval (i.e. 5^{th} quantile), 2) between their lowest estimate and their best estimate / median, 3) between their best estimate / median and their upper estimate / 95^{th} quantile, and 4) above their upper / 95^{th} quantile is calculated and compared to a theoretically optimal distribution *p=(0.05, 0.45, 0.45, 0.05)*. CA refers to calibration as calculated according to the IDEA protocol, which is defined as the proportion of credible intervals capturing the realisation.

Figure 3 The statistical accuracy and information of *n* = 58 participants. A trade-off exists between the two measures used by the Classical Model. Those who are statistically accurate (above 0.05, red horizontal line) often have a lower information score than the median score for individuals (grey vertical line). The blue dashed line shows the highest statistical accuracy score possible for 13 questions (0.93), and the black line shows the highest score obtained by individuals in the elicitation (0.53).

Figure 4 Statistical accuracy of the Classical Model (CM) compared to IDEA calibration for *n* = 58 participants. The graph shows that participants with perfect calibration when assessed by the IDEA protocol, can have poor statistical accuracy for the Classical Model. On the righthand side, we show where the realisations fell in each of

the expert's multinomial distributions (used to calculate statistical accuracy), and contrast this with how many realisations fell within the participant's 90% credible intervals (calibration). Bold numbers indicate the highest scores possible for statistical accuracy and calibration.

Figure 5 The spearman correlation between information calculated for the Classical Model, and informativeness calculated for the IDEA protocol for n = 58 participants. The shaded area represents a 95% confidence interval.

Figure 6 CM Scores derived for each aggregation.

Figure 7 Component performance measures of the Classical Model (CM) and IDEA protocol for n = 8 groups under six alternative procedures for aggregation. a) Statistical accuracy, the red-dashed line represents the 0.05 threshold for statistically inaccurate scores, (Classical Model), the blue dashed line represents a perfect statistical accuracy score for 13 questions, and the black dashed line represents the highest score obtained by any individual, b) information score (Classical Model), the red line represents the median information of an individual c) calibration, (IDEA) the red line represents perfect calibration (0.90), d) informativeness (IDEA), the red line represents the informativeness of the median individual, e) accuracy (IDEA) of the best estimate, the red line represents the accuracy of the median individual.

Figure 8 The scores of aggregations under the Classical Model and the IDEA protocol when adjustments are made to correct for questions for which the realised truth had been zero. a) Statistical accuracy, the red-dashed line represents the 0.05 threshold for statistically inaccurate scores, (Classical Model), the blue dashed line represents a perfect statistical accuracy score for 13 questions, and the black dashed line represents

the highest score obtained by any individual prior to the adjustment; b) calibration, (IDEA) the red line represents perfect calibration (0.90).

Figure 1

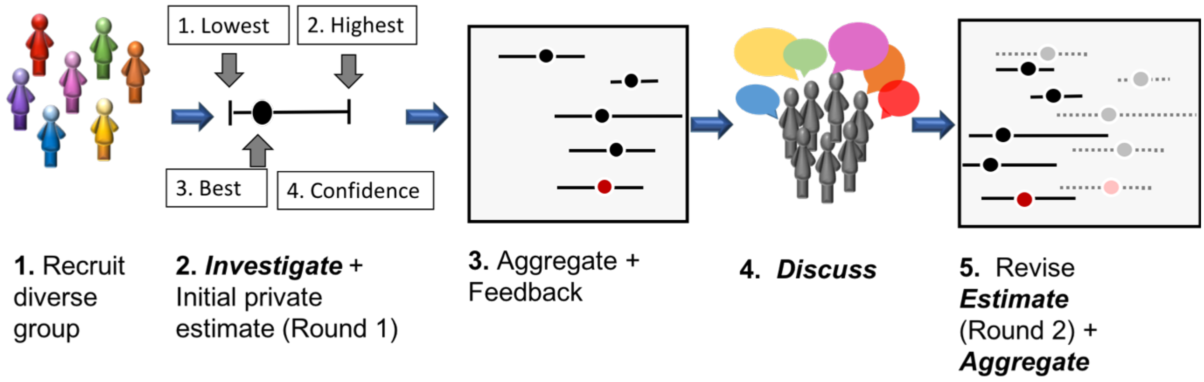
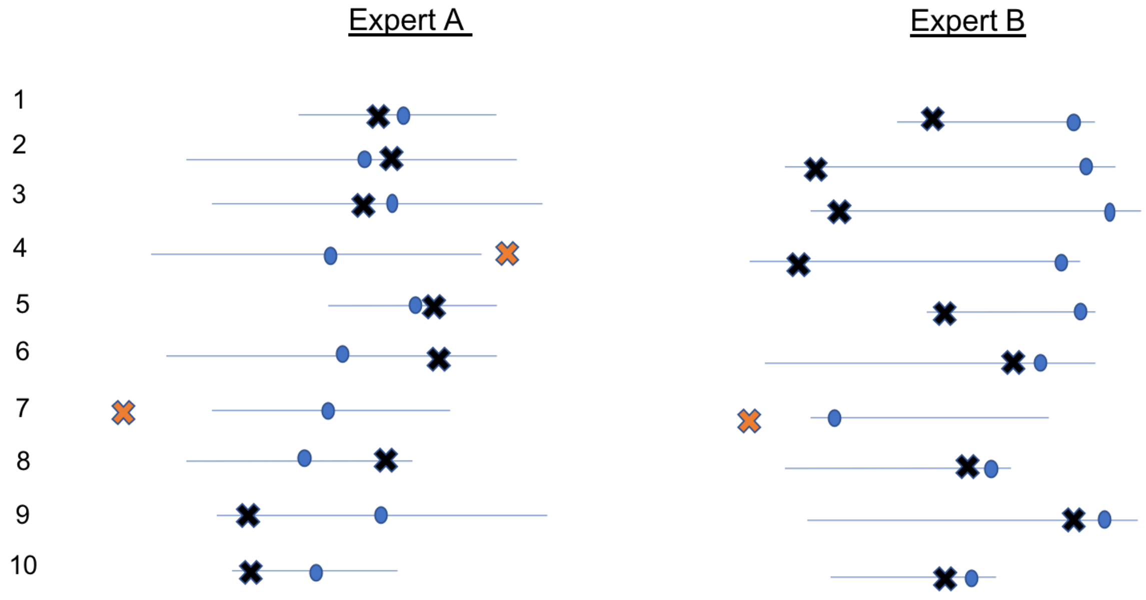

1. Lowest
2. Highest
3. Best
4. Confidence
1. Recruit diverse group
2. Investigate + Initial private estimate (Round 1)
3. Aggregate + Feedback
4. Discuss
5. Revise Estimate (Round 2) + Aggregate

Figure 2



<5	5-50	50-95	>95	SA	CA
1	4	4	1	0.83	0.80

<5	5-50	50-95	>95	SA	CA
1	9	0	0	0.003	0.90

Figure 3

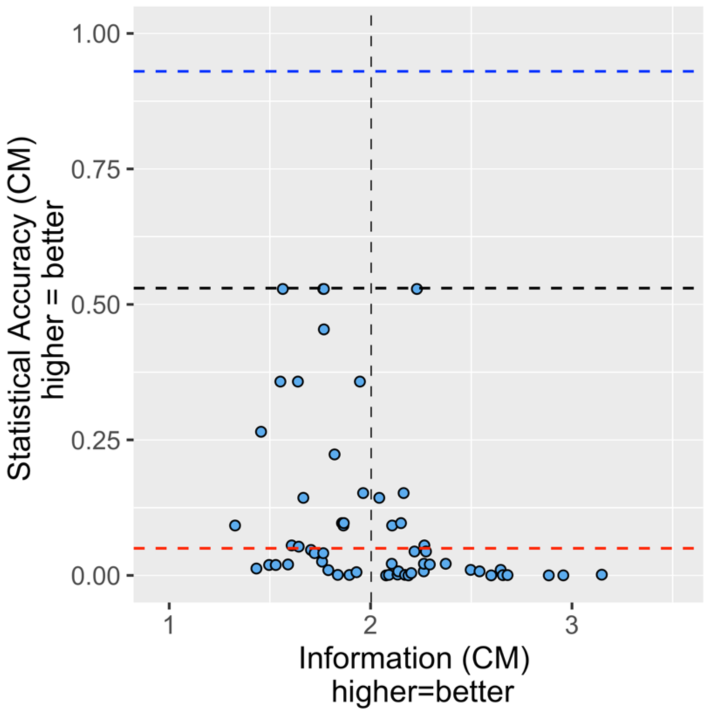
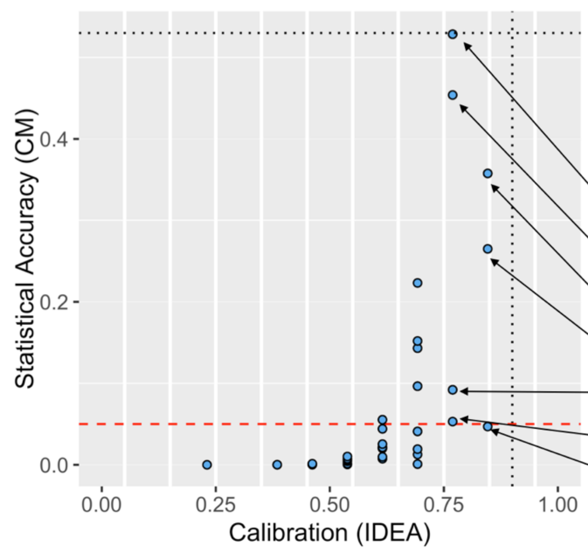
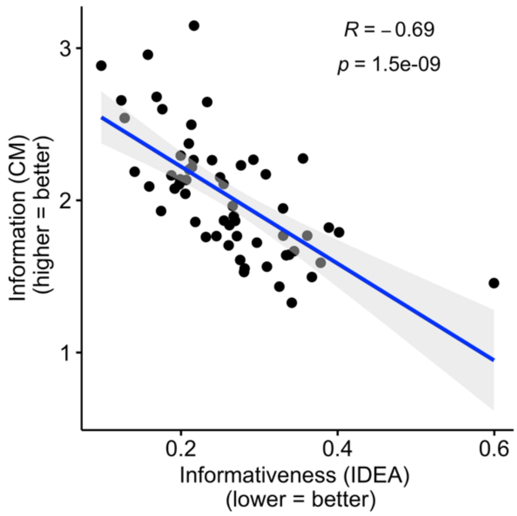

Statistical Accuracy (CM)
higher = better
1.00
0.75
0.50
0.25
0.00
1
2
3
Information (CM)
higher=better

Figure 4



Classical Model		IDEA protocol	
Statistical Accuracy	Realisations in inter-quantile ranges 5^{th} 50^{th} 95^{th}	Calibration	Realisations captured within 90% credible intervals In | Out
0.53	(2 | 5 | 5 | 1)	0.77	(10 | 3)
0.45	(2 | 6 | 4 | 1)	0.77	(10 | 3)
0.36	(2 | 5 | 6 | 0)	**0.85**	(11 | 2)
0.26	(2 | 7 | 4 | 0)	**0.85**	(11 | 2)
0.092	(3 | 6 | 4 | 0)	0.77	(10 | 3)
0.053	(3 | 7 | 3 | 0)	0.77	(10 | 3)
0.046	(2 | 9 | 2 | 0)	**0.85**	(11 | 2)

Figure 5



R = −0.69
p = 1.5e-09
Information (CM)
(higher = better)
3
2
1
0.2
0.4
0.6
Informativeness (IDEA)
(lower = better)

Figure 6

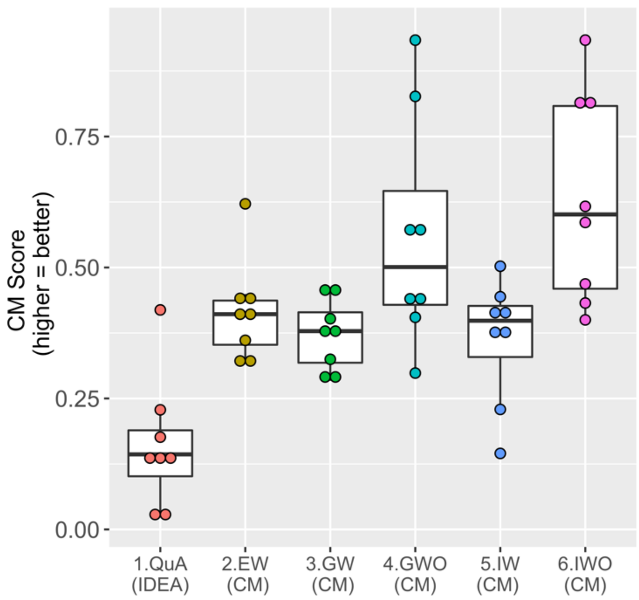

CM Score
(higher = better)
0.00
0.25
0.50
0.75
1.QuA
(IDEA)
2.EW
(CM)
3.GW
(CM)
4.GWO
(CM)
5.IW
(CM)
6.IWO
(CM)

Figure 7

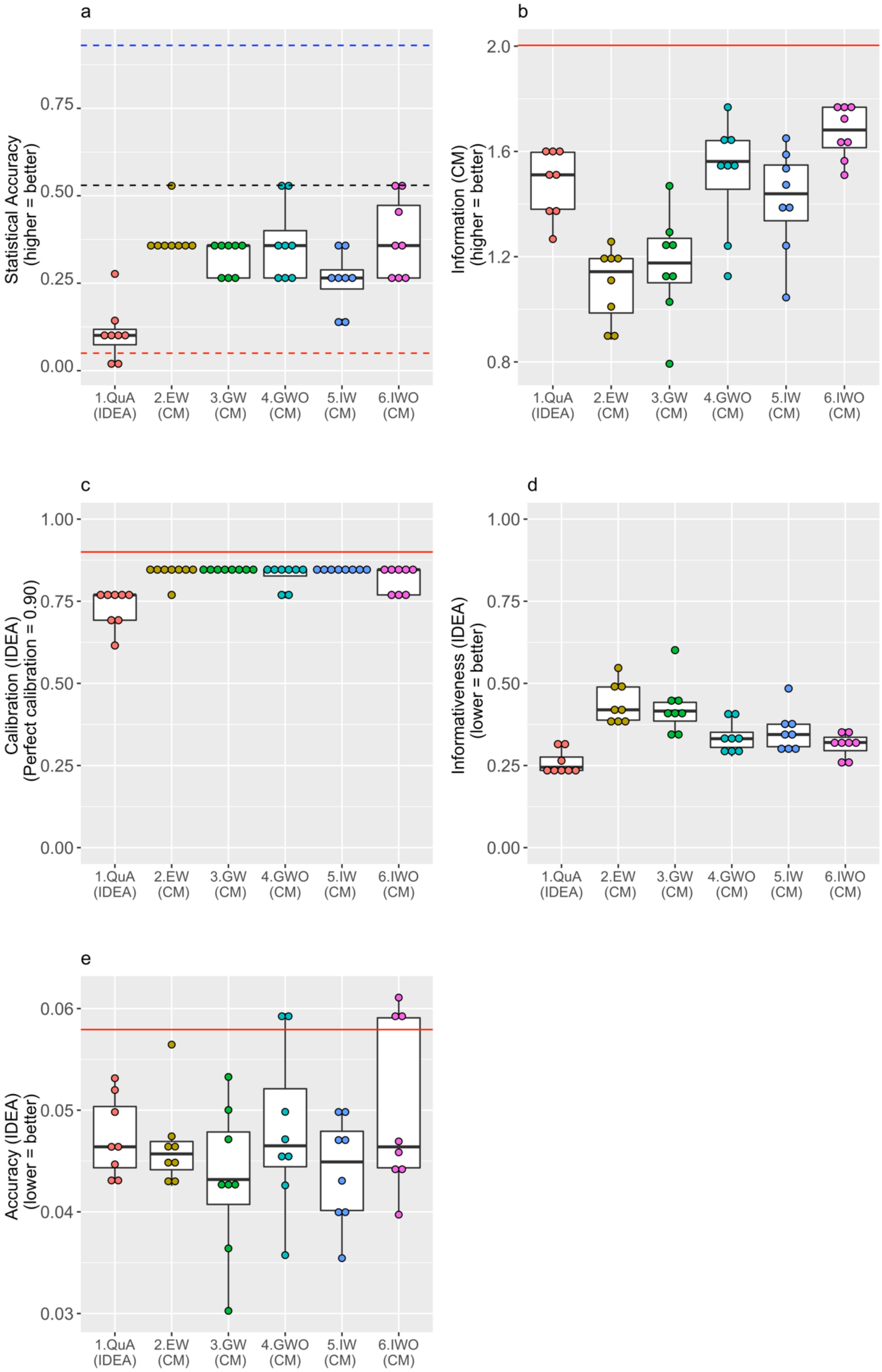
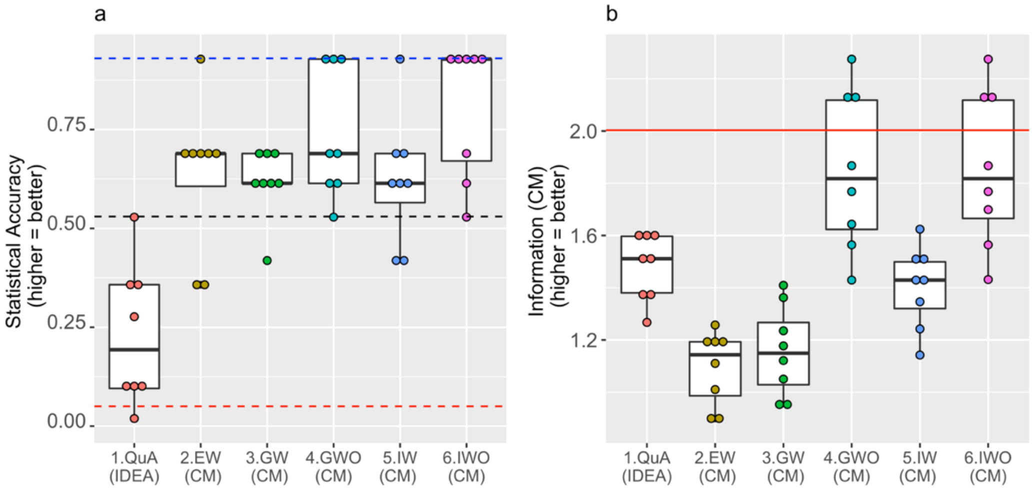

a
Statistical Accuracy
(higher = better)
b
Information (CM)
(higher = better)
c
Calibration (IDEA)
(Perfect calibration = 0.90)
d
Informativeness (IDEA)
(lower = better)
e
Accuracy (IDEA)
(lower = better)
1.QuA (IDEA)
2.EW (CM)
3.GW (CM)
4.GWO (CM)
5.IW (CM)
6.IWO (CM)

Figure 8


a
Statistical Accuracy
(higher = better)
0.00
0.25
0.50
0.75
1.QuA
(IDEA)
2.EW
(CM)
3.GW
(CM)
4.GWO
(CM)
5.IW
(CM)
6.IWO
(CM)
b
Information (CM)
(higher = better)
1.2
1.6
2.0
1.QuA
(IDEA)
2.EW
(CM)
3.GW
(CM)
4.GWO
(CM)
5.IW
(CM)
6.IWO
(CM)