Reliability Engineering and System Safety 163 (2017) 109–120

Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress

Cross validation for the classical model of structured expert judgment

Abigail R. Colson[a,b], Roger M. Cooke[c,d,e,*]

[a] Center for Disease Dynamics, Economics & Policy, Washington, DC, USA
[b] University of Strathclyde, Glasgow, UK
[c] Resources for the Future, Washington, DC, USA
[d] University of Strathclyde, Glasgow, UK
[e] TU Delft (ret), Delft, The Netherlands

ARTICLE INFO

Keywords:
Expert judgment
Calibration
Information
Classical model
Out-of-sample validation

ABSTRACT

We update the 2008 TU Delft structured expert judgment database with data from 33 professionally contracted Classical Model studies conducted between 2006 and March 2015 to evaluate its performance relative to other expert aggregation models. We briefly review alternative mathematical aggregation schemes, including harmonic weighting, before focusing on linear pooling of expert judgments with equal weights and performance-based weights. Performance weighting outperforms equal weighting in all but 1 of the 33 studies in-sample. True out-of-sample validation is rarely possible for Classical Model studies, and cross validation techniques that split calibration questions into a training and test set are used instead. Performance weighting incurs an "out-of-sample penalty" and its statistical accuracy out-of-sample is lower than that of equal weighting. However, as a function of training set size, the statistical accuracy of performance-based combinations reaches 75% of the equal weight value when the training set includes 80% of calibration variables. At this point the training set is sufficiently powerful to resolve differences in individual expert performance. The information of performance-based combinations is double that of equal weighting when the training set is at least 50% of the set of calibration variables. Previous out-of-sample validation work used a Total Out-of-Sample Validity Index based on all splits of the calibration questions into training and test subsets, which is expensive to compute and includes small training sets of dubious value. As an alternative, we propose an Out-of-Sample Validity Index based on averaging the product of statistical accuracy and information over all training sets sized at 80% of the calibration set. Performance weighting outperforms equal weighting on this Out-of-Sample Validity Index in 26 of the 33 post-2006 studies; the probability of 26 or more successes on 33 trials if there were no difference between performance weighting and equal weighting is 0.001.

1. Introduction

Structured expert judgment denotes techniques for using expert judgments as scientific data. A recent overview dates its inception to large scale engineering studies from 1975 [9]. Cooke et al. [13] first proposed the use of calibration (here called "statistical accuracy") and information to score experts' performance, and the use of these scores for defining and validating schemes combining experts' judgments is termed the Classical Model [6]. By 2006, analysts had conducted 45 professionally contracted Classical Model studies. Cooke and Goossens [12] summarized and published the results from these studies, and made the data, called the TU Delft database, available to the research community. The studies in the TU Delft database include those from the dawn of the Classical Model, and their study designs differ wildly. The number of experts in a given study ranged from 4 to 77 and the number of calibration variables (i.e., questions from the field for which realizations are known post hoc; these questions are the basis for creating performance-based combinations of the experts' assessments) ranged from 5 to 55.

The TU Delft database allows researchers to explore how experts and different combinations of experts perform on data from real expert judgment applications. Researchers have used this data to investigate how the performance-weight (*PW*) combinations of the Classical Model compare to equal-weight (*EW*) combinations of experts both in-sample and out-of-sample. Cooke and Goossens [12] demonstrated that *PW* is superior to *EW* on in-sample comparisons, in which the same set of data is used to both initialize and validate the model. Clemen [5] first raised the question of the Classical Model's out-of-sample validity, using the TU Delft database to explore if performance-based combinations predict out-of-sample items better than equally weighted combi-

* Corresponding author.
E-mail address: cooke@rff.org (R.M. Cooke).

http://dx.doi.org/10.1016/j.ress.2017.02.003
Received 28 January 2016; Received in revised form 7 February 2017; Accepted 18 February 2017
Available online 24 February 2017

nations of the experts. In recent years other researchers have proposed various methods for validation of the Classical Model and drawn conflicting conclusions.

Since 2006 use of the Classical Model has continued to expand, thanks in large part to high-profile applications (for example, [1]). Over thirty three independent expert judgment studies were performed between 2006 and March 2015. These studies were contracted by a variety of organizations including: Bristol University (UK), the British government, the European Commission, PrioNet (Canada), Public Health Canada, the Robert Wood Johnson Foundation, Sanguin, the US Department of Homeland Security, and the US Environmental Protection Agency. In these recent studies, panels of 4–21 experts assessed between 7 and 17 calibration variables. These studies are generally better resourced, better executed, and better documented than the very early Classical Model applications.

Updating the 2006 database and establishing a baseline for the in- and out-of-sample validation of performance based weighting is timely and important. The recent report of the National Academy of Sciences on the social cost of carbon lends urgency to this effort, noting "*performance-weighted average of distributions usually outperforms the simple average, where performance is again measured again by calibration and informativeness (and is often evaluated on seed variables not used to define the weights, because the value of the quantity of interest in many expert elicitation studies remains unknown)*" [27, p. 339].

Another recent spur is the 5-year forecasting tournament organized by IARPA of which Philip Tetlock's Good Judgment Project was proclaimed the winner. The tournament concerned current events assessed by "ordinary citizens" as opposed to quantification of scientific/engineering uncertainties. Radically down-selecting from a pool of more than 3000,[1] Tetlock's group distilled a small group of "superforecasters" based on their performance. Although very different in purpose and method to the Classical Model, the Good Judgment Project strongly underscores the value of performance based combinations.

In this study we use data from 33 post-2006 studies (Described in Supplementary Online Material 1) to explore the in-sample and out-of-sample validity of the Classical Model. Based on the post-2006 data, we test the null hypothesis that performance-weight (*PW*) combinations of the experts are no better than equal-weight (*EW*) combinations in terms of statistical accuracy and informativeness. Finally, we develop an Out-of-Sample Validity Index (OoSVI) which can be used to validate future Classical Model studies and related research.

The 33 post-2006 studies considered here excludes two sets of post-2006 applications. One concerns an ongoing expert elicitation program at the Montserrat Volcano Observatory that has produced a wealth of data on expert performance [29,3]. The second is a recently completed large scale study by the World Health Organization involving 72 experts spread over 134 distinct panels [2,20]. Since both sets of studies involve heavily overlapping expert panels, they do not lend themselves to the present analysis where the panels are considered independent.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the Classical Model and reviews alternate pooling schemes, comparing their statistical accuracy across the post-2006 data. Section 3 summarizes the in-sample properties of the post-2006 data. Section 4 reviews previous out-of-sample validation research based on the TU Delft database, and Section 5 summarizes the out-of-sample performance of the newly collected post-2006 data. Section 6 provides two detailed case studies that demonstrate good and poor out-of-sample performance. Section 7 evaluates the hypothesis that *PW* is no better than *EW* out-of-sample. Section 8 compares the present results with those of Eggstaff et al. [16] and a final section gathers conclusions.

The Supplementary Online Material (SOM) provides: (1) references and information on the 33 post-2006 applications analyzed here, (2) a detailed description of the Classical Model, (3) more information on quantile averaging in the post-2006 dataset, (4) improved exposition of proofs of the scoring rule properties (adapted from Cooke [6]), (5) additional details on previous cross validation research, and (6) an expanded list of references for applications of the Classical Model.

2. Aggregating expert judgments

2.1. The Classical Model

In the Classical Model, experts quantify their uncertainty regarding two types of questions. The variables of interest are the target of the elicitation; these questions cannot be adequately answered by existing data or models, so expert judgment is needed as additional evidence. Calibration variables (also termed seed variables) are questions from the experts' field which are unknown to the experts at the time of the elicitation, but whose true values will be known post hoc. Experts are scored and weighted according to their calibration and information, and their assessments are combined into a *PW* decision maker, which can be compared to an *EW* decision maker. The calibration and information scores are briefly discussed below, and more detail is available in SOM 2.

In the context of expert judgment, the term "calibration" gives engineers and scientists the false impression that the judgments of experts are being "adjusted," as they would calibrate instruments by adjusting their scales. This is not the case. Since calibration is only loosely defined in decision theory literature, this confusion is best avoided by replacing "calibration" with "statistical accuracy," defined as the P-value at which one would falsely reject the hypotheses that a set of probability assessments were statistically accurate. Very crudely, it answers questions like "how likely is it that at least 7 out of 10 realizations should fall outside an expert's 90% confidence bands, if each value really had an independent 90% chance of falling inside the bands?"

Information is measured as Shannon relative information with respect to a user supplied background measure. Shannon relative information is used because it is scale invariant, tail insensitive, slow, and familiar. The combined score, the product of statistical accuracy and informativeness, satisfies a long run proper scoring rule constraint and involves choosing an optimal statistical accuracy threshold beneath which experts are unweighted. Weights for the *PW* decision maker are based on this combined score, as described in SOM 2.

The Classical Model's performance measures of statistical accuracy and information do not map neatly onto the terms "accuracy" and "precision", which are familiar to social scientists. Accuracy denotes the distance between a true value and a mean or median estimate, and precision denotes a standard deviation. While appropriate for repeated measurements of similar variables, these notions are scale dependent and therefore not useful in aggregating performance across variables on vastly different physical scales. For example, how should one add an error of 10^9 colony forming units of campylobacter infection to an error of 25 micrograms per liter of nitrogen concentration? Expert judgments frequently involve different scales, both within one study and between studies. For this reason, the performance measures in the Classical Model are scale invariant. That said, the exhaustive out-of-sample analysis of Eggstaff et al. [16] (described in Section 4) found that the realizations were closer to the *PW* combination's median than the *EW* combination's median in 74% of the 75 million out-of-sample predictions based on the TU Delft data. Such non-parametric ordinal proximity measures, proposed by Clemen [5] are not used to score expert performance, as the scores strongly depend on the size of the expert panels. Thus, the present study focuses on the standard Classical

[1] Full documentation is not available at this writing and the information here is based on http://www.npr.org/sections/parallels/2014/04/02/297839429/-so-you-think-youre-smarter-than-a-cia-agent accessed 1/12/2017 and [31].

Model scoring variables: statistical accuracy, informativeness, and the combined score (i.e., the product of statistical accuracy and information).

2.2. Linear, geometric, and harmonic pooling

Both *PW* and *EW* are examples of linear pooling, whereby the combination is a weighted (*PW*) or unweighted (*EW*) average of the experts' distributions. Other pooling schemes have been proposed, and it is appropriate to consider their performance before restricting attention to *PW* and *EW*. Geometric averaging, or geometric weighting (*GW*) has been advocated as being "independence preserving" [22] and "externally Bayesian" [18]. Geometric averaging tends to concentrate mass in regions where the experts agree. Lichtendahl et al. [23] suggest that averaging experts' quantiles might be superior to *EW*. Flandoli et al. [17] also used this technique in their analysis of the Classical Model. Averaging quantiles is easier to compute than averaging distributions, and is frequently employed by unwary practitioners. It was recently used in climate change uncertainty quantification [19] and, unwittingly, in a re-analysis of data from a Classical Model study [28].

As shown by Bamber et al. [4], averaging quantiles is equivalent to harmonically weighting (*HW*) the densities (Box 1). *GW* concentrates mass more aggressively than linear pooling, and *HW* is slightly more extreme. For example, the arithmetic mean of 0.01 and 0.99 is 0.5, the geometric mean is 0.0995, and the harmonic mean is 0.0198.

Given its tendency to over-confidence, it is not surprising that *HW* returns poor statistical performance. Fig. 1, reproduced from Bamber et al. [4], compares the statistical accuracy of the best and worst experts (*BE* and *WE*, respectively), *EW*, and *HW* on the post-2006 dataset. Scores for the same study are on the same vertical line, ordered from the left according to the statistical accuracy of *EW*. Bamber et al. find that on 18 of the 33 studies, the hypothesis that *HW* is statistically accurate would be rejected at the 5% level; on 9 studies it would be rejected at the 0.1% level. The hypothesis that the best expert is statistically accurate is rejected at the 5% level in 7 studies, but it is not rejected at the 0.1% level in any of the 33 studies. The geometric mean of the ratios of *BE* statistical accuracy/*WE* statistical accuracy is 890,000, indicating the wide gap between the best and worst performing experts in these studies.

SOM 3 provides more information on the scoring of *EW* and *HW* and includes a comparison to performance weighting (*PW*). SOM 3 also analyzes the dependence of *EW*, *HW*, and *PW* performance on the number of experts and number of calibration variables in a study.

3. In-sample validation

Before exploring the out-of-sample validity of *PW*, it is useful to first establish its in-sample validity. If *PW* is not superior to *EW* in-sample, there is no motive for studying its out-of-sample performance or using *PW* in practice.

The Classical Model introduces three types of performance weights. SOM 2 provides more detail. First, global weights (*PWg)* use the average information score over all calibration variables, which is proportional to the information in the product of the marginal distributions. Second, item weights (*PWi)* use the information scores on each item to derive item-specific weights. *PWi* is most often used in practice, and it enables the expert to up- or down weight him/herself for variables in which (s)he feels more or less confident. Both *PWg* and *PWi* optimize the choice of a threshold statistical accuracy (*Sa*) value; experts with *Sa* below the cutoff value are unweighted. SOM 4 shows that this yields asymptotically strictly proper scoring rules (without, however, excluding other possibilities). Third, a non-optimized *PW* combination (*NoOp)* forgoes optimization and assigns weights proportional to the combined score based on the global information measure. This latter weight does not satisfy the conditions of SOM 4. These three combinations (together with *EW* and user-specified weighting) are implemented in the freeware EXCALIBUR [14], downloadable at http://www.lighttwist.net/wp/.

The combined scores of *EW*, *PWg*, *PWi,* and *NoOp* are shown in Fig. 2, ordered according to *PWi* scores. More detail is presented in Table 1. These are *in-sample* comparisons, as the statistical accuracy and informativeness of the various combinations are measured on the same calibration variables used to initialize the performance weighting.

The in-sample superiority of the *PW* combinations over *EW*, evident in Fig. 2, is not a foregone conclusion. Table 1 shows that the statistical accuracy of *EW* is better than that of *PW NoOp* in 30% of the cases, and *EW* has the highest combined score in one case (Nebraska). In three cases (CoveringKids, Erie Carps, and Hemophilia) the best expert's combined score is higher than that of the other combinations. However, *PWi* has the highest combined score in 24 of the cases, coinciding with *PWg* in 13 studies and the best expert in 12 studies. In 14 studies the *PWi* combined score is strictly greater than that of *PWg*. Comparing the *NoOp* combined scores with those of *PWg* shows that optimization plays a significant role in improving the performance of the combination of experts.

The geometric mean (or geomean) of the ratios of combined scores from different weighting schemes gives an overall picture of relative in-sample performance. The geomean over all 33 studies of *PWgComb/EWComb* is 3.36 and for *PWiComb/EWComb* it is 3.86. The geomean *BEComb/EWComb* it is 1.84, indicating it would be better simply to use the best expert than to apply equal weighting. The geomean of *PWgComb/BEComb* is 1.83, demonstrating the value of using performance weights on all the experts rather than rely on the best expert.

4. Review of cross validation studies

The previous section established that *PW* outperforms *EW* in the post-2006 data based on an in-sample analysis. A sensible next question, first raised by Clemen [5], is how do *PW* and *EW* compare out-of-sample?

True out-of-sample validation would require observing the variables of interest and then calculating how *PW* and *EW* perform based

Box 1.Bamber et al. [4] explain that averaging quantiles is equivalent to harmonic weighting of the densities.

The following proof is reproduced from Bamber et al. [4]:

Let F and G be CDFs from experts 1 and 2, with densities f, g. Let HW, hw denote respectively the CDF and density of the result of averaging the quantiles of F, G. Then for all $r \in (0, 1)$:

$$HW^{-1}(r) = \tfrac{1}{2}(F^{-1}(r) + G^{-1}(r)). \quad (1)$$

Taking derivatives of both sides:

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

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

Eq. (3) says that hw is the harmonic mean of f and g, evaluated at points corresponding to the r-th quantile of each distribution.

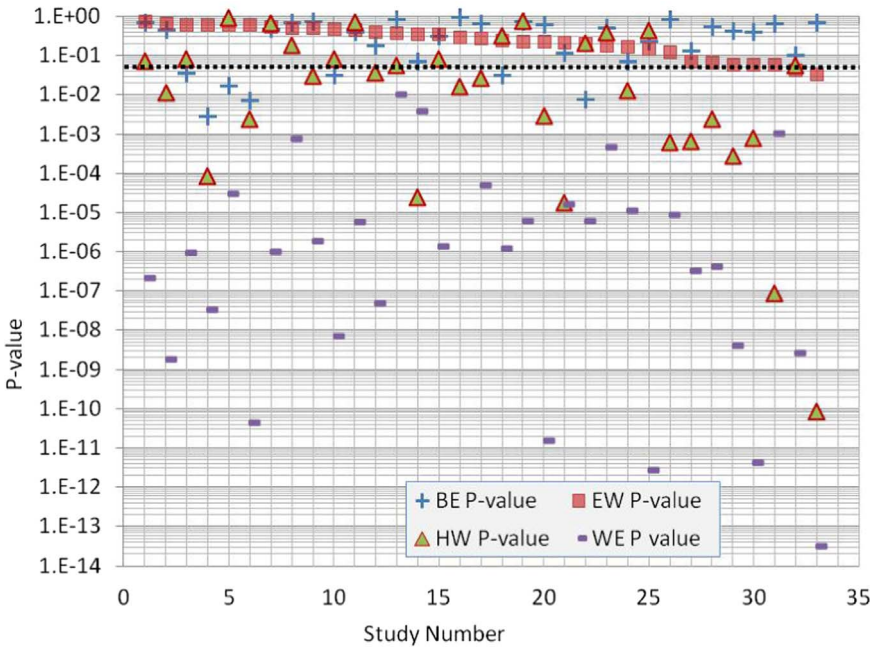

Fig. 1. Statistical accuracy (*p*-value) of the best expert (BE), worst expert (WE), equal weighting (EW), and harmonic weighting (HW) on the post-2006 dataset. Scores for the same study are on the same vertical line, ordered from the left according to the statistical accuracy of EW. The dotted line indicates the 5% rejection threshold. (Reproduced from Bamber et al. [4]).

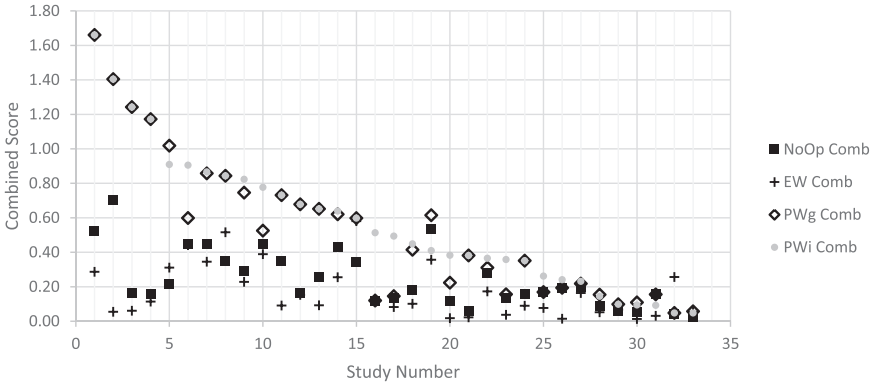

Fig. 2. In-sample validation, combined scores for EW, PWi, PWg, NoOp and EW for 33 post-2006 studies.

on those realizations. The variables of interest in an expert judgment study, however, are rarely observed. The lack of observation is what necessitates the use of expert judgment. Thus, true out-of-sample evaluation is seldom possible, and cross validation techniques based on subsets of the calibration questions are used instead.

In the first cross validation analysis of the Classical Model, Clemen [5] suggested a remove-one-at-a-time (ROAT) approach in which one item is removed, the performance weights are recalculated, and then performance is evaluated based on predictions for the item that was removed. The predictions originate from different *PW* decision maker combinations but are pooled and compared to the *EW* combination. In a preliminary analysis, Clemen looked at a non-random sample of 14 studies and found that *PW* outperformed *EW* in 9 of them, which was not statistically significant. Lin and Cheng expanded ROAT analysis on the TU Delft database by expanding the pool of studies considered to 28 (2008) and then 40 (2009) of the pre-2006 studies. They found performance of the *PW* decision maker degrades out-of-sample relative to in-sample. In their first analysis, *PW* significantly outperformed *EW* out-of-sample, but they found no significant difference between the two in the second analysis. Lin and Cheng did not report that their code has been vetted against EXCALIBUR, and large differences exist between the values reported in Lin and Cheng [24] and Cooke and Goossens [12]. SOM 5 provides information on these discrepancies.

Although ROAT analysis is a simple and frequently implemented cross validation technique, it suffers from an inherent bias against the *PW* decision maker, as described previously by Cooke [7,8,10]. In ROAT analysis, each calibration variable is predicted by a separate performance-based combination in which experts who assessed the removed item badly are up-weighted, and those who assess the removed item well may be down-weighted. The combination is then scored according to its performance on the removed item. Cooke has previously illustrated this bias with a simple example (2012a; 2014), which, because of its importance, is explained again here.

Suppose Experts 1 and 2 state the probability of flipping heads from a coin as $P_1(Heads)$=0.8 and $P_2(Heads)$=0.2. Suppose the experts' assessments are weighted proportionally to the likelihood of their distributions (given observed data) and combined into a decision maker, such that the decision maker's assessment is $P_{dm} = wP_1 + (1-w)P_2$. Likelihood weights are not proper scoring rules and do not account for information; but a strong analogy links them to the classical model, as the driving term in that model is the likelihood of the hypothesis that an expert is statistically accurate. Moreover, these experts are equally informative.

Table 1
Summary of in-sample results for the 33 post-2006 studies, showing the statistical accuracy (Sa), information (Inf), and combined scores (Comb) for each combination. PW is performance weights, either without optimization, with optimized global weights, or with optimized item weights (see SOM 2 for detail). The best expert in each panel is the expert with the highest combined score. For each study (i.e., each row), the optimal combined scores are in boldface and highlighted.

			Equal Weight			*PW* Non-optimized			*PW* Global			*PW* Item			Best Expert		
Study	**#Exp**	**#Cal**	**Sa**	**Inf**	**Comb**	**Sa**	**Inf**	**Comb**	**Sa**	**Inf**	**Comb**	**Sa**	**Inf**	**Comb**	**Sa**	**Inf**	**Comb**
Arkansas	4	10	0.39	0.20	0.08	0.50	0.34	0.17	0.50	0.34	0.17	0.50	0.52	**0.26**	0.07	0.41	0.03
Arsenic	9	10	0.06	1.10	0.07	0.04	1.68	0.06	0.04	2.74	**0.10**	0.04	2.74	**0.10**	0.04	2.74	**0.10**
ATCEP	5	10	0.12	0.25	0.03	0.68	0.23	**0.16**	0.68	0.23	**0.16**	0.24	0.38	0.09	0.10	0.50	0.05
Biol_Agent	12	12	0.41	0.24	0.10	0.41	0.43	0.18	0.68	0.61	0.41	0.68	0.66	**0.45**	0.31	1.00	0.31
CDC_ROI	20	10	0.23	1.23	0.29	0.39	1.35	0.52	0.72	2.31	**1.66**	0.72	2.31	**1.66**	0.72	2.31	**1.66**
CoveringKids	5	10	0.63	0.27	0.17	0.72	0.38	0.28	0.72	0.43	0.31	0.72	0.51	0.36	0.62	0.89	**0.55**
create-vicki	7	10	0.06	0.21	0.01	0.19	0.27	0.05	0.39	0.28	**0.11**	0.31	0.30	0.09	0.02	0.25	0.00
CWD	14	10	0.47	0.93	0.44	0.47	0.94	0.45	0.49	1.22	0.60	0.68	1.33	**0.90**	0.31	2.19	0.69
Daniela	4	7	0.53	0.17	0.09	0.68	0.23	0.16	0.55	0.63	**0.35**	0.55	0.63	**0.35**	0.55	0.63	**0.35**
DCPN_Fistula	8	10	0.06	0.62	0.04	0.12	1.14	0.14	0.12	1.31	0.16	0.27	1.34	**0.36**	0.01	1.92	0.01
eBPP	14	15	0.36	0.32	0.11	0.36	0.43	0.15	0.83	1.41	**1.17**	0.83	1.41	**1.17**	0.83	1.41	**1.17**
Eff_Erup	14	8	0.29	0.80	0.23	0.29	1.02	0.29	0.66	1.12	0.75	0.66	1.24	**0.82**	0.19	1.80	0.33
Erie_Carp	11	15	0.31	0.29	0.09	0.57	0.45	0.25	0.76	0.86	0.65	0.76	0.86	0.65	0.53	1.29	**0.68**
FCEP	5	8	0.22	0.10	0.02	0.14	0.39	0.06	0.66	0.57	**0.38**	0.66	0.57	**0.38**	0.66	0.57	**0.38**
Florida	7	10	0.76	0.46	0.34	0.56	0.80	0.45	0.76	1.13	0.86	0.76	1.15	**0.87**	0.12	1.74	0.22
Gerstenberger	12	14	0.64	0.48	0.31	0.35	0.61	0.21	0.93	1.10	**1.02**	0.76	1.20	0.91	0.54	1.74	0.93
GL_NIS	9	13	0.04	0.31	0.01	0.93	0.21	0.19	0.93	0.21	0.19	0.93	0.26	**0.24**	0.45	0.27	0.12
Goodheart	6	10	0.55	0.28	0.15	0.47	0.35	0.16	0.71	0.96	**0.68**	0.71	0.96	**0.68**	0.71	0.96	**0.68**
Hemopilia	18	8	0.25	0.20	0.05	0.31	0.27	0.08	0.31	0.49	0.15	0.31	0.46	0.14	0.85	1.07	**0.91**
IceSheets	10	11	0.49	0.52	0.25	0.62	0.70	0.43	0.40	1.55	0.62	0.62	1.04	**0.64**	0.40	1.55	0.62
Illinois	5	10	0.62	0.26	0.16	0.39	0.51	0.20	0.34	0.65	0.22	0.39	0.60	**0.23**	0.13	0.97	0.13
Liander	11	10	0.23	0.48	0.11	0.23	0.50	0.11	0.23	0.52	0.12	0.68	0.75	**0.51**	0.00	0.86	0.00
Nebraska	4	10	0.37	0.70	**0.26**	0.03	1.25	0.04	0.03	1.45	0.05	0.03	1.45	0.05	0.03	1.45	0.05
Obesity	4	10	0.07	0.24	0.02	0.50	0.23	0.12	0.44	0.51	0.22	0.78	0.49	**0.38**	0.44	0.51	0.22
PHAC_T4	10	13	0.27	0.20	0.05	0.09	0.26	0.02	0.14	0.40	0.06	0.10	0.49	**0.05**	0.01	1.25	0.01
San_Diego	8	10	0.33	1.07	0.36	0.78	0.69	0.54	0.88	0.69	**0.61**	0.35	1.19	0.41	0.03	1.12	0.04
Sheep	14	15	0.66	0.78	0.52	0.36	0.98	0.35	0.64	1.31	**0.84**	0.64	1.31	**0.84**	0.64	1.31	**0.84**
SPEED	14	16	0.52	0.75	0.39	0.63	0.71	0.45	0.68	0.78	0.53	0.99	0.78	**0.78**	0.23	0.84	0.19
TDC	18	17	0.17	0.36	0.06	0.30	0.55	0.17	0.99	1.26	**1.24**	0.99	1.26	**1.24**	0.99	1.26	**1.24**
Tobacco	7	10	0.20	0.45	0.09	0.66	0.53	0.35	0.69	1.06	**0.73**	0.69	1.06	**0.73**	0.69	1.06	**0.73**
Topaz	21	16	0.63	0.92	0.58	0.31	1.12	0.34	0.41	1.46	**0.60**	0.41	1.46	**0.60**	0.41	1.46	**0.60**
UMD_NREMOVAL	9	11	0.07	0.80	0.05	0.49	1.43	0.70	0.71	1.99	**1.40**	0.71	1.99	**1.40**	0.71	1.99	**1.40**
Washington	5	10	0.15	0.53	0.08	0.20	0.65	0.13	0.20	0.72	0.14	0.50	0.99	**0.49**	0.06	1.29	0.08

If we observe n Heads and n Tails, the experts' likelihood ratio is:

$$\frac{0.8^n \times 0.2^n}{0.2^n \times 0.8^n} = 1 \tag{4}$$

and each expert receives weight 1/2. Removing one observed Tail changes the likelihood ratio to 0.8/0.2 = 4, so Expert 1 now receives four times the weight of Expert 2 in the combined decision maker. The new decision maker's assessment of the probability of the Heads is [(4/5) × 0.8 + (1/5) × 0.2.] = 0.68 and the probability of Tails is 1 − 0.68 = 0.32. In ROAT cross validation, this model is then evaluated on its ability to predict the Tail that was removed, so the likelihood based on this observation is 0.32. Removing a Head has a similar affect, and swings the decision maker toward Expert 2. If this process is repeating for each of 10 coin tosses, the likelihood for the ROAT model is lower than the likelihood of the original model by a factor of (0.32/0.5)^10 = 0.01.

In addition to this bias against *PW*, ROAT is a problematic method for cross validation because removing one calibration variable can influence an individual expert's statistical accuracy by a factor of three or more. Statistical accuracy is a "fast" function, meaning it commonly varies by several orders of magnitude over experts in a given study. To illustrate the variation from removing one item, Fig. 3 shows the weights of five experts in the European Union-United States Nuclear Regulatory Commission (EU-USNRC) atmospheric dispersion study [21] as each of 23 calibration variables is removed one-at-a-time.

ROAT analysis is based primarily on the same scoring rule used by the Classical Model, i.e. a combination score that is the product of statistical accuracy and information. Researchers have also suggested cross validation for the Classical Model should be based on performance measures different from those that underlie it. Clemen [5] proposed evaluating the Classical Model based on the distance of the *PW* decision maker's median to the realization. Others [24,25,16] have also used that method, and, as mentioned in Section 2, *PW* outperformed *EW* in 74% of the 75 million out-of-sample predictions considered by Eggstaff et al. However, obtaining an accurate median estimate is a different objective from the Classical Model's goal of informative and statistically accurate uncertainty assessments.

Lin and Huang [26] conducted ROAT analysis with the Brier score, which is related to the quadratic scoring rule. They followed Winkler [30], who first proposed strictly proper scoring rules for individual variables to score experts. A score is assigned to each expert's probability assessment for each calibration variable based on each realization, and the scores are summed over the set of calibration variables. This idea has been strongly discouraged [10,6]. Cooke [10] provided a simple example as a counter-argument to this approach. If an expert assesses that the probability of flipping Heads with a coin of unknown composition is 1/2, the score for each toss is the same for either Heads or Tails. If these individual scores are summed to create an overall score, then the overall score is independent of the actual observed outcomes. Observing 50 Heads and 50 Tails yields the same overall expert score as observing 100 Heads. This is why scoring rules in the Classical Model are asymptotically strictly proper scoring rules for expected relative frequencies. Instead of assessing scores per calibration variable and summing over all calibration variables, sets of assessments are scored by comparing expected and observed relative frequencies (for detail see SOM 4).

Although ROAT has been the predominant approach to cross

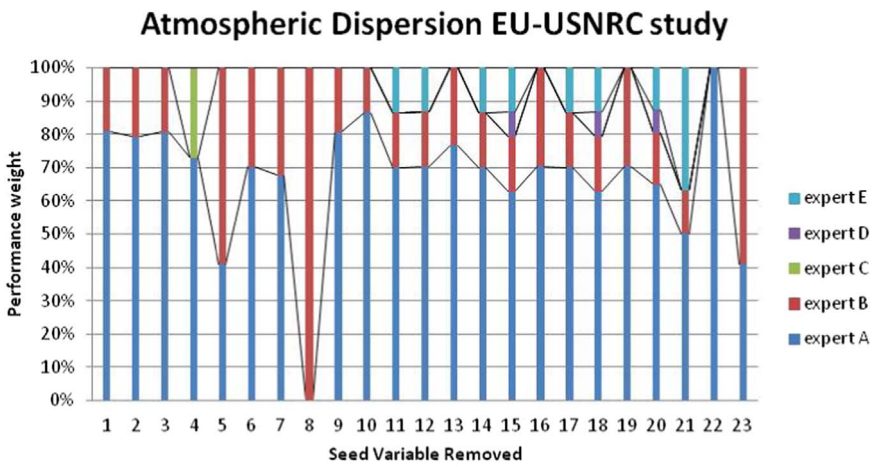

Fig. 3. Variation of expert weights when calibration variables are removed one-at-a-time.

validation of the Classical Model, other approaches have also been considered. These approaches split the calibration variables into two complementary sets: the training set is used to calculate the performance-based weights, and the remaining variables comprise a test set that is used for validation. Cooke [7] focused on the panels in the TU Delft database with 16 or more calibration questions. He split the 13 such panels into half, with each set serving as the training set to predict the other half. Each of the included panels thus provided 2 data points. Cooke found *PW* outperformed *EW* on 20 of the 26 comparisons. Flandoli et al. [17] examined five datasets, choosing 30% of the number of calibration variables as the size of the test set, provided this number was at least 8, otherwise the test set was 8. For each of the five studies, they sampled training sets from all possible combinations of calibration variables and found that generally the Classical Model's *PW* outperformed *EW* or their alternative Expected Relative Frequency model in terms of statistical accuracy. They recoded the Classical Model in R, and their results for the cases studied do not agree with EXCALIBUR (SOM 5).

The most extensive cross validation study is Eggstaff et al. [16], which performed cross validation on all possible training/test set combinations (except the empty set and the full set) for the 62 studies available at the inception of their work. Studies with large numbers of calibration variables were split into separate studies to suppress combinatoric explosion. In this comprehensive analysis, they found *PW* significantly outperforms *EW*. Eggstaff et al. only consider global weighting, as it is easier to implement than item weighting. Eggstaff's code excludes experts who assess less than the full set of calibration variables, whereas EXCALIBUR includes these experts and reduces the power of the statistical accuracy score to equal that of the expert with the fewest assessed calibration variables. Eggstaff's approach is reasonable for the purpose of cross validation, but it can produce differences with EXCALIBUR.

The results of one choice of training/test set is termed a *split*. Using the exhaustive cross validation approach of Eggstaff et al., excluding the empty set and the entire set, a study with 10 calibration variables produces $2^{10}-2 = 1022$ splits. There are 10 splits with training size 1 and 252 splits with training size 5. A study with 17 calibration variables produces 131,070 splits, 24,310 of which have training size 8. Simply aggregating splits would strongly overweight the mid-sized training sets. Each additional calibration variable doubles the computation time until memory constraints become binding. Computing all splits for a study with 17 calibration variables takes over 24 h on a fast PC.

Although EXCALIBUR does not perform cross validation, it can be used to spot check cross validations. The cross validation code of Eggstaff et al. [16] was checked extensively against EXCALIBUR after publication, and some errors were corrected which affected a few cases. After correcting these, exact agreement with EXCALIBUR was achieved. This is the only cross validation code that has been vetted in this way.

5. Out-of-sample cross validation of the post-2006 studies

The present study builds on the approach of Eggstaff et al. [16], and uses the out-of-sample validation code which Lt. Col. Eggstaff graciously provided. We apply their comprehensive cross validation technique to the 33 post-2006 studies.

To compare studies and test the effectiveness of performance based combination, the scores must be rendered comparable. For a given study, scores for a fixed training set size can be averaged, as they are in Eggstaff et al. [16]. Whatever the size of the training set, the *EW* combination is always the same. A small training set means that testing the hypothesis that an expert is statistically accurate has low power, and *PW* is less able to resolve differences in expert performance. At the same time, the ability to distinguish *PW* and *EW* performance has greater power. The converse holds for large training sets: *PW* is better able to resolve experts' statistical accuracy, but the test set is less able to resolve differences in the statistical accuracy of *PW* and *EW*.

For the rest of this study, *PW* denotes *PWglobal*. *PWSa*, *PWInf* and *PWComb* denote the statistical accuracy, informativeness and combined scores of *PWglobal* respectively. Similar abbreviations apply for *EW*.

Fig. 4 shows the statistical accuracy scores *PWSa* and *EWSa* first averaged within a study over each training set size (e.g., all training sets of 8 calibration variables), then averaged across studies for each percentage size (e.g., all training sets consisting of 80% of the calibration variables, including, for example, training sets of 8 of 10 calibration variables and 11 of 14 calibration variables). Each training

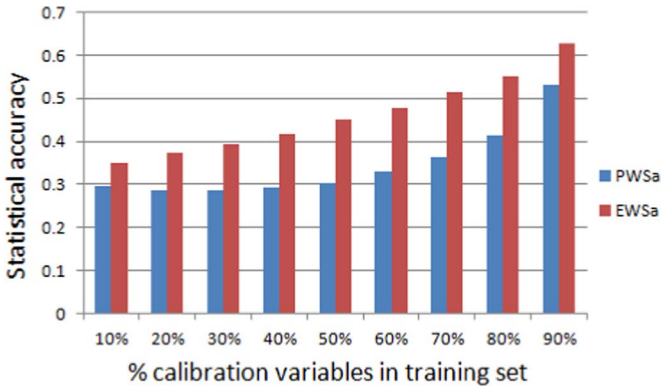

Fig. 4. Average over all studies per training set size percentage of the average PWSa and EWSa; higher values are better.

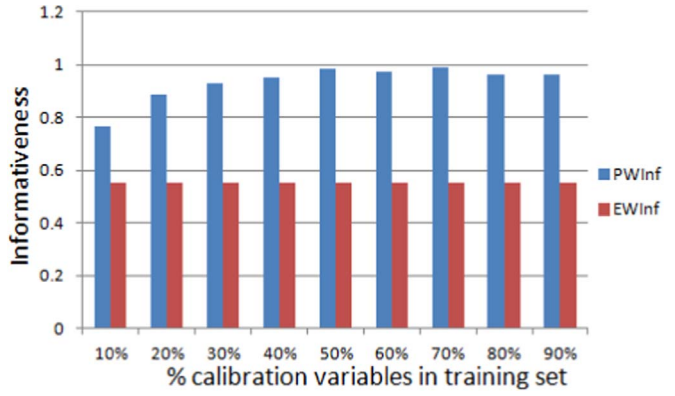

Fig. 5. Average over all studies per training set size percentage of the average PWInf and average EWInf; higher values are better.

set percentage includes studies for which the training sets and test sets are different; increasing the calibration set increases both the number of variables in the training set of size *x*% and the corresponding test set of size 100−*x*%. For a fixed training and test set size, the statistical accuracy scores are comparable.

Whereas in-sample *PWSa* is usually greater than *EWSa* (see Table 1), Fig. 4 shows that *PWSa* degrades out-of-sample relative to *EWSa*. There is indeed an 'out-of-sample penalty' for *PWSa*. Statistical accuracy is a very fast function, typically varying over 4 orders of magnitude in a panel of 5 experts with 10 calibration variables. A difference between a P-value of 0.60 or 0.50, like those observed in Fig. 4, is quite small by comparison. Cooke [10] considers the small sample behavior of the statistical accuracy statistic. All these *Sa* scores increase with training set size, reflecting the loss of statistical power as the test set size decreases. *PWSa* increases faster than *EWSa* for larger training sets. For small training sets with low power to resolve differences between the experts, in-sample statistical accuracy scores tend to be more equal. *PW* is therefore less able to distinguish more and less statistically accurate experts, and *PW* is similar to *EW*. Not until the training set exceeds 70% of the calibration set does *PW* consistently identify the more accurate experts and accord them more weight. The differences between *PWSa* and *EWSa* then start to close.

Information shows a different pattern (Fig. 5). As in Fig. 4, per study *PWInf* and *EWInf* are averaged for each training set size for each study, and these averages are then averaged per percentage size over all studies. As described, informativeness is scored as Shannon relative information with respect to an analyst-defined background measure. Per variable, this background measure is by default uniform or loguniform on the smallest interval containing all expert quantiles and the realization, if available (i.e., for calibration questions), plus a 10% overshoot. Thus, expert information scores are directly comparable within a study but not between studies. For a given study, *EWInf* differs for each individual training set, but the average over all training sets of a given size always equals the in-sample *EWinf* values in Table 1. Hence, for each training set size, averaging over all studies returns the average of the in-sample *EW* information scores (column *EWInf* of Table 1), or 0.499. Per training set size, the heterogeneity across studies is the same. *PWInf* is lowest for small training sets, reflecting the fact that *PW* is more similar to *EW*, but increases quickly to twice *EWInf*. Unlike statistical accuracy, informativeness is a slow function and a difference of a factor 2 is noteworthy.

Both *PWComb* and *EWComb* increase with training set size due to loss of statistical power in the test set. We may anticipate that *PWComb* should grow more quickly.

To isolate the growth of *PWComb* that is not due to decreasing statistical power, we must articulate the notation a bit. Let *PWComb(t,s)* denote the *PW* combined score on training set *t* of study *s*. Let $Av_{\#t=k}$ *PWComb(t,s)* denote the average of *PWComb(t,s)* over all training sets of size *k* of study *s*. Similar notation applies for *EWComb*. Fixing *s* and fixing training size |*t*| *EWSa(t,s)* and *EWInf(t,s)* are nearly independent: The average of their product (the average of combined scores) is indistinguishable from the product of their averages. More exactly, the mean and standard deviation over all studies and all training percentage sizes of $Av_{\#t=k}$*EWComb(t,s)*−*[*$Av_{\#t=k}$*EW Sa(t,s)*×$Av_{\#t=k}$*EWInf(t,s)]* are respectively −4.3E-4 and 6.5E-4. Therefore, for all *s*

$$Av_{\#t=k}\frac{PWSA(t,s)}{Av_{\#t=k}EWSa(t,s)}\times\frac{PWInf(t,s)}{Av_{\#t=k}EWInf(t,s)} = \frac{Av_{\#t=k}PWComb(t,s)}{Av_{\#t=k}EWSa(t,s)\,'Av_{\#t=k}EWInf(t,s)}$$

Because of independence, the right hand side differs very little from

$$\frac{Av_{\#t=k}PWComb(t,s)}{Av_{\#t=k}EWComb(t,s)}.$$

The latter quantity is taken to represent the out-of-sample performance of *PWComb* for study *s* and training set size *k* which is not conflated with the loss of statistical power in the test set. An increase or decrease of this quantity as *k* varies represents a real change in *PWComb* relative to *EWComb* that does not depend on statistical power of the test set.

When combining ratios, we must take the geomean to insure that the combination of the reciprocals is the reciprocal of the combination. Geo_s denotes the geometric average over all studies *s*. Fig. 6 plots Geo_s $Av_{\#t=\%k}$ *PWComb(t,s)* and Geo_s $Av_{\#t=\%k}$ *EWComb(t,s)*, where *%k* denotes the *kth* percentage of the calibration set. Their ratio in Fig. 7 shows the growth in Geo_s $Av_{\#t=\%k}$ *PWComb(t,s)* which does not depend on statistical power loss. The ratio does not grow until the training sizes exceed 50% of the calibration set.

The geomeans in Fig. 7 are all greater than 1, indicating that *PW* outperforms *EW* on training sets of all percentages. However, they are less than the in-sample geomean of *PWComb* / *EWComb* (3.36), demonstrating the out-of-sample penalty.

The expert weights are much more volatile for small training sizes, as these weights are based on statistical accuracy measured with only a few calibration variables. Fig. 8 plots the weighted average variance of the expert weights. More precisely, (a) we compute the in-sample combined score of each expert in each study for every training / test split, (b) we compute the variance of each expert's combined score per training set size, (c) we take a weighted average of the experts' variances per training set size (weighted using the experts' average combined scores per training set size), and finally (d) we average the weighted average variance over all studies per percentage training set size. The result is a picture of the overall volatility in expert weights expressed as a function of training set percentage size. This volatility declines sharply up to sizes of 70% after which the differences are less than 0.005.

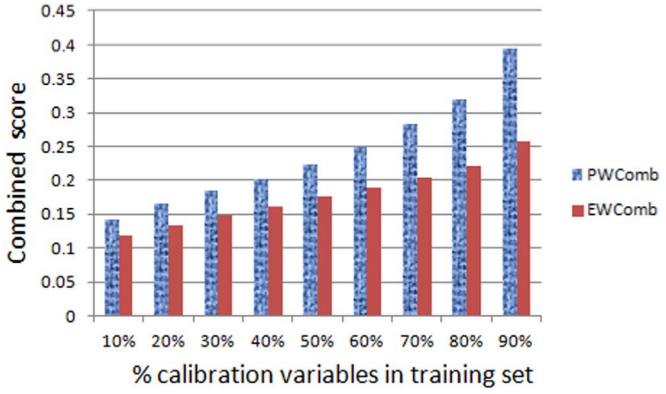

Fig. 6. PWComb and EWComb averaged over training sets of same size, and geo-averaged over studies per training set size percentage.

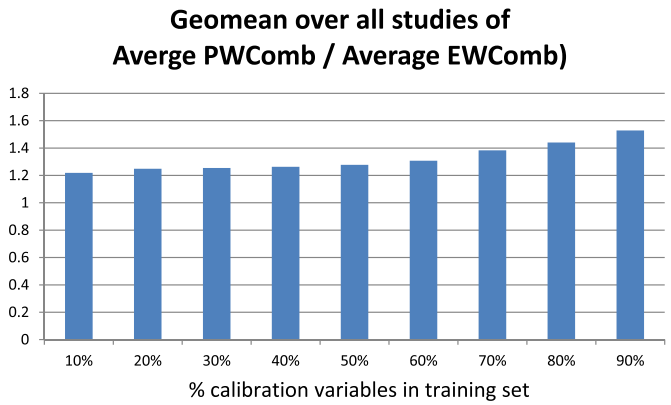

Fig. 7. The ratio of average PWComb and average EWComb, geo-averaged over studies per training set size percentage.

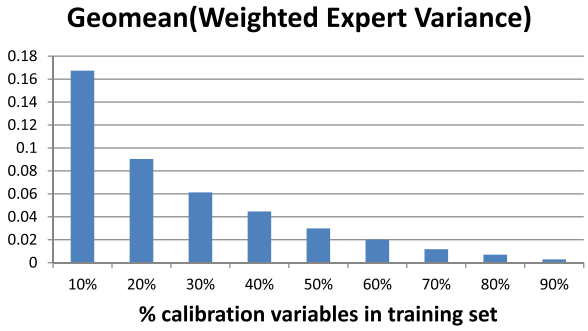

Fig. 8. Geomean of weighted average of variances in experts' in-sample combined scores (un-normalized weights), geomean taken over all studies per training set percentage size.

6. Detailed data for two studies

It is helpful to look at detailed data for studies showing "good" (Biol_agents) and "bad" (San Diego) out-of-sample characteristics. In both cases *PWComb* exceeds *EWComb* in-sample (see Table 1).

Starting with the "good," Fig. 9 shows *PWComb - EWComb* for each test set (left panel) and the averages of these scores over training set sizes (right panel). Both *PWComb* and *EWComb* increase with training set size. The right panel shows that *PWComb* increases more rapidly, hence the difference between *PWComb* and *EWComb* (left panel) also tends to increase. This indicates that, as the training set increases, *PW* is improving at a rate greater than the loss of power in the test set.

Fig. 10 shows the variance of the experts' un-normalized weights in Biol_agents as a function of training set size. The variance declines for all experts as training set size increases.

This pattern is by no means universal. The poorest out-of-sample performance is found in the San Diego study, shown in Fig. 11. For all training set sizes *PWComb* is worse than *EWComb* (right panel). The variance in experts' combined scores decrease very quickly (Fig. 12), from much higher initial values than in Biol_agents in Fig. 10.

Clearly, there are differences in studies that are not revealed by the in-sample performance scores. Future research will focus on impacts of study parameters on cross-validation. Without going deeply into the causes of the differences in these two cases, we many note from Table 1 that the best expert in Biol_agents coincides with the *PW*, whereas in San Diego, the best expert scores well below *PW*. San Diego is also unusual in that *EW* is more informative than *PW* in Table 1.

7. Statistical test of *PW* versus *EW* out-of-sample

Previous publications [10,15,16] have used a "total out-of-sample validity index" based on all training/test set splits defined per study as follows: (a) take the ratio *Average PWComb/Average EWComb* per training set size (b) take the geomean of these ratios over all training set sizes. The main justification for this is that it leaves nothing out; however, it includes splits with very low power in the training or test sets, is computationally too heavy for real time deployment, and involves training sets where the expert weights have high volatility.

We propose an "Out of Sample Validity Index" (OoSVI) defined by step (a) above applied only to training sets whose size is 80% of the entire set of calibration variables. The reasons for this choice are:

1. The expert weights used to construct *PW* have relatively low volatility at 80%
2. The expert weights at 80% more closely resemble the weights used in the actual study based on all calibration variables
3. For studies assessing 5-, 50- and 95-percentiles on 10 calibration variables, the possible statistical accuracy scores range over a factor of 31, which is ample for distinguishing *EWSa* and *PWSa*.
4. This OoSVI can be computed quickly and processed with the primary study results, even for large numbers of calibration variables. With 22 calibration variables (the largest number in Eggstaff's study), evaluating all splits with 80% in the training set involves evaluating 7315 splits, for 70% the number is 170,544.

The test for statistical accuracy for a 20% test size has greatly reduced power, but this applies equally to *EW* and *PW* without prejudicing the ratio *PWComb/EWComb*.

The simplest test for the hypothesis that *PW* and *EW* are indistinguishable considers an indicator for each study which takes the value 1 if *PW* outperforms *EW* and takes the value 0 otherwise. The null hypothesis assigns such an indicator the distribution (1/2, 1/2). Any column of Table 2 might be chosen with the indicator taking "1" if the row value is greater than 1 ("success"), and "0" ("failure") otherwise. Using the 80% column, we find 26 "successes" in 33 trials. The probability of seeing at least 26 successes if there were no difference between *PW* and *EW* is 0.001. Had we used the geomean over all training set sizes (last column) with 23 "successes" the

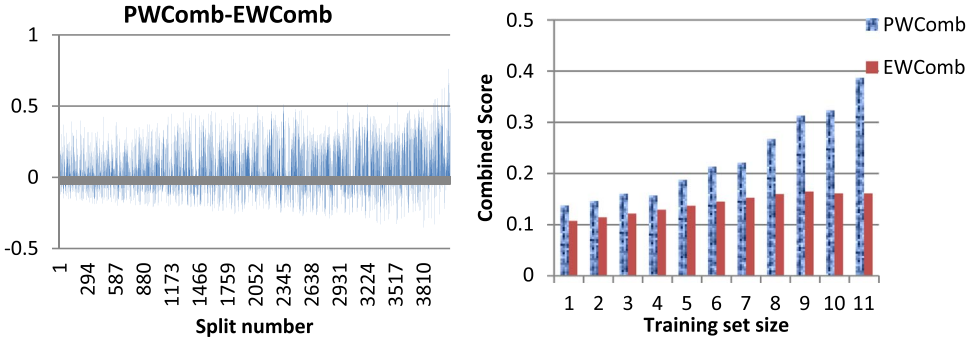

Fig. 9. Data from the Biol_Agents study. Left, differences of combined test set scores for PW and EW for all training splits; training set size increases from left to right, from size 1 to size 11. Right, combined scores of PW and EW averaged per training set size.

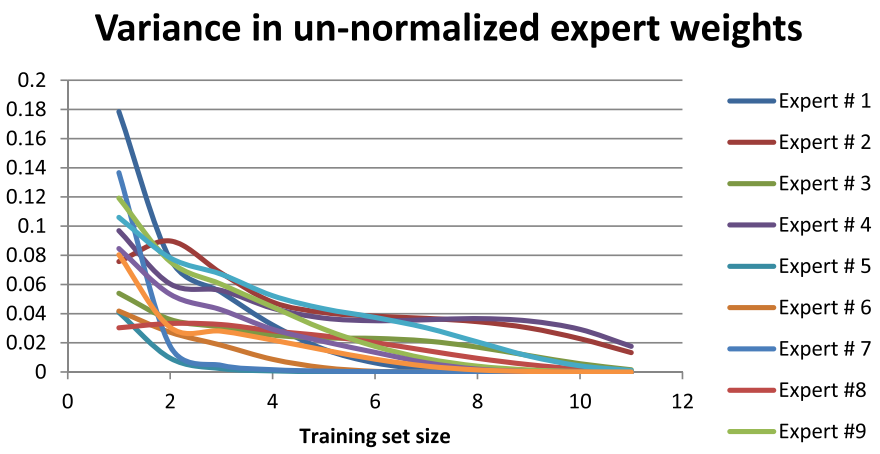

Fig. 10. Variance of experts' combined scores on the training sets (un-normalized weights) per training set size, for Biol_Agent studys.

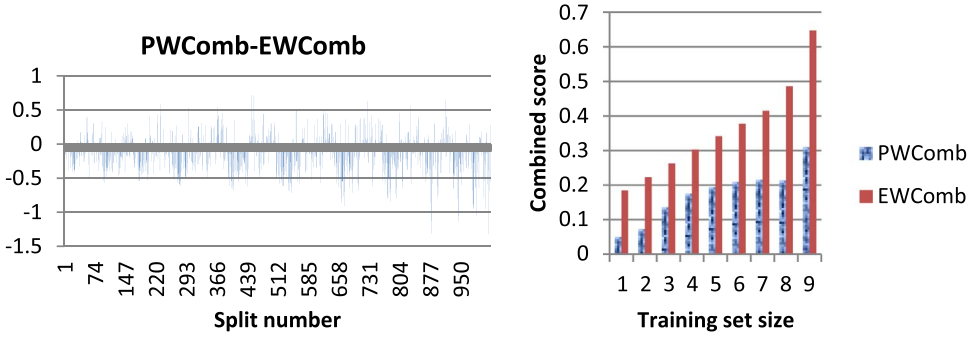

Fig. 11. Data from the San Diego study. Left, differences of combined test set scores for PW and EW for all training splits; training set size increases from left to right, from size 1 to size 9. Right, combined scores of PW and EW averaged per training set size.

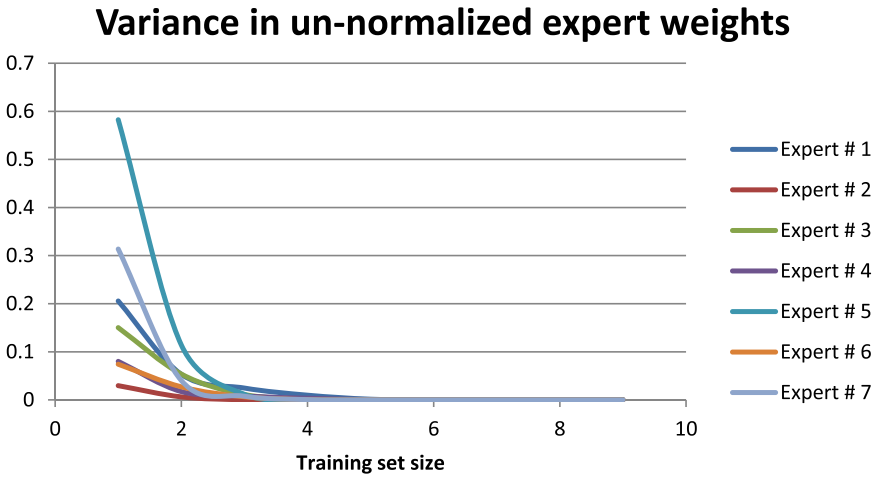

Fig. 12. Variance of experts' combined scores on the training sets (*un*-normalized weights) per training set size, for study San Diego study.

exceedance probability would be 0.018. For each percentage split of 50% or more, the null hypothesis would be rejected at the 5% level.

Table 3 shows the correlation between various study characteristics and in-sample performance measures and the OoSVI. If study characteristics, such as the number of experts or seed questions included, were correlated with OoSVI, that could guide future elicitation practice. This preliminary analysis, though, suggests none of these study characteristics is correlated with OoSVI. OoSVI is most strongly correlated with the statistical accuracy of the best and second best expert, indicating that identifying good experts is the crux of the method's performance, both in- and out-of-sample. The geomean of OoSVI for studies with a best expert whose *Sa* is above 0.05 is 1.54; the geomean for studies with a best expert whose *Sa* is below 0.05 falls to 1.14. For the *Sa* of the second best expert, the geomeans are 1.64 and 1.17 respectively. /

8. Discussion

The present results may be compared with the results of Eggstaff et al. [16], which analyzed out-of-sample validation for 62 studies available at the inception of their research. Those results are also reported in Cooke [10,11]. The latter sources give the *Total OoSVI*, which corresponds to the last column of Table 2. Eggstaff's data records 45 successes (*Total OoSVI* > 1) out of 62 trials, or 72%. This study finds 23 successes out of 33 trials, or 70%. The value of *Total OoSVI* for Eggstaff et al. [16] is 2.25, which is higher than the comparable value

Table 2
Average *PWComb*/Average *EWComb* for training sets sized as percent of all calibration variables.

	Training set size as percent of calibration variables									Row Geomean
	10%	20%	30%	40%	50%	60%	70%	80%	90%	
Arkansas	1.132	1.256	1.523	1.382	1.423	1.322	1.306	1.404	1.718	**1.376**
Arsenic	1.039	1.035	1.073	1.117	1.293	1.431	1.722	1.874	1.961	**1.352**
ATCEP	1.955	1.677	1.380	1.455	1.166	1.149	1.156	0.997	0.799	**1.262**
Biol_Agent	1.278	1.280	1.217	1.373	1.477	0.453	1.676	2.008	2.405	**1.536**
CDC_ROI	1.006	1.199	1.112	1.229	1.004	1.109	1.107	1.305	1.399	**1.157**
CoveringKids	1.032	1.510	1.407	1.478	1.487	1.463	1.517	1.538	1.427	**1.420**
Create	0.890	0.789	0.763	0.817	1.046	1.277	1.278	1.331	1.142	**1.014**
CWD	1.328	0.980	1.031	0.907	0.812	0.729	0.708	0.680	0.756	**0.862**
Daniela	1.051	1.051	1.086	1.099	1.137	1.137	1.815	1.721	1.721	**1.279**
Fistula	0.262	0.964	0.918	1.039	1.147	1.354	1.362	1.426	1.910	**1.037**
eBPP	1.859	1.844	2.027	1.778	2.402	2.576	2.727	2.958	4.033	**2.384**
Eff_Erup	0.965	0.903	0.903	0.796	0.651	0.664	0.892	0.892	0.919	**0.835**
Erie_Carp	2.920	2.612	2.684	2.567	1.856	1.787	2.047	2.017	2.909	**2.339**
FCEP	3.843	7.704	7.704	7.908	8.897	8.826	7.485	7.485	5.713	**7.091**
Florida	0.920	0.445	0.657	0.695	0.750	0.886	0.979	1.364	1.412	**0.851**
Gerstenberger	1.056	1.183	1.152	1.683	1.651	1.670	1.562	1.501	1.604	**1.431**
GL_NIS	2.177	1.847	1.672	1.477	1.186	1.134	1.066	1.024	0.809	**1.316**
Goodheart	1.180	1.291	1.595	1.441	1.366	1.480	1.611	2.136	2.607	**1.586**
Hemopilia	1.638	2.019	2.019	1.862	2.938	1.534	1.476	10476	2.808	**1.913**
IceSheets	1.266	0.861	0.867	0.814	0.850	0.779	0.807	0.880	0.903	**0.883**
Illinois	0.671	0.697	0.821	0.798	0.867	1.126	1.407	1.800	2.484	**1.073**
Liander	0.881	0.746	0.488	0.614	0.669	0.780	0.870	0.788	0.575	**0.700**
Nebraska	0.559	0.340	0.389	0.517	0.692	0.978	1.393	1.733	1.892	**0.789**
Obesity	3.569	2.383	2.430	2.105	1.842	1.586	1.361	1.267	1.815	**1.945**
PHAC_T4	1.057	0.833	0.709	0.650	0.853	0.974	1.106	1.195	1.180	**0.931**
San_Diego	0.273	0.327	0.516	0.578	0.569	0.555	0.519	0.439	0.478	**0.460**
Sheep	0.772	0.866	0.828	0.902	1.018	1.033	1.119	1.204	1.432	**1.001**
SPEED	0.575	0.661	0.595	0.614	0.632	0.750	0.783	0.844	0.835	**0.692**
TDC	3.413	3.850	4.207	3.416	2.794	2.754	2.667	2.573	2.557	**3.088**
Tobacco	2.179	2.167	2.019	1.965	1.861	1.928	1.830	1.778	10.472	**1.900**
Topaz	0.863	0.860	0.860	0.941	0.966	1.050	1.119	1.178	1.182	**0.994**
UMD_NREMOVAL	1.882	5.221	4.555	4.508	3.634	3.340	3.192	2.654	2.236	**3.300**
Washington	3.614	2.148	1.726	1.370	1.119	1.119	1.142	1.308	1.334	**1.529**
Column Geomean	**1.219**	**1.249**	**1.254**	**1.263**	**1.277**	**1.308**	**1.383**	**1.440**	**1.528**	**1.321**
number > 1	**22**	**19**	**20**	**20**	**22**	**24**	**26**	**26**	**25**	**23**
P(this many success or more in 33 trials) on null hypothesis	0.040	0.243	0.148	0.148	0.040	0.007	0.0007	0.001	0.002	0.018

Table 3
Correlation between the Out-of-Sample Validity Index (OOSVI) and various study characteristics and in-sample performance measures in the post-2006 studies. The p-value is the probability of seeing the observed correlation or stronger if no correlation exists.

Variable	Spearman's rank correlation coefficient	P-value
Study characteristics		
Number of experts	−0.19	0.28
Number of calibration variables	0.01	0.94
Three quantiles (vs. five)	0.02	0.91
Plenary interviews (vs. one-on-one)	−0.17	0.35
In-sample performance		
EW statistical accuracy	0.00	0.99
PW (global) statistical accuracy	0.31	0.08
Best expert statistical accuracy	0.50	< 0.01
Second best expert statistical accuracy	0.32	0.07

from Table 2: 1.32. We note that Eggstaff's data set included more studies with a high number of calibration variables (Fig. 13):

Although the percentage of studies in which *Total OoSVI* > 1 in Eggstaff et al. [16] is similar to the present study, its statistical significance is greater owing to the larger number of studies: the P-value for falsely rejecting the null hypothesis is 0.0002, assuming independence. Eggstaff split the studies with more than 22 calibration variables into two or more sub-studies because of the computational burden. The splitting was not done randomly, split studies use the same experts, and performance on different sub-studies sometimes varies widely. Excluding all these split studies and excluding studies also present in the current set, we retain 40 of the original 62 studies, of which the *Total OoSVI* exceeded 1 in 31 cases. These may be combined with the current study yielding 23 (this study)+31 (Eggstaff)=54 successes on 33 (this study)+40 (Eggstaff)=73 trials. The null hypothesis that *PW* is no better than *EW* is now rejected at the 2.5E-5 level. Our analysis strongly supports the value of *PW* over *EW*, based on both in-sample and out-of-sample performance.

Whereas this study groups training/test splits according to the percentage of all calibration variables in the training set, Eggstaff et al. group splits by the difference: training set size − test set size. For all differences, the ratio (# studies with *PWComb* dominating/# studies with *EWComb* dominating) is greater or equal to one. The ratio decreases as the training set grows larger than the test set. If the goal were to choose a training set size to maximize the probability that *PWComb* > *EWComb,* the advice would be that there should be five more variables in the test set than in the training set. On Eggstaff's dataset, the size of the training set in this case would vary from 1 to 8, and would involve training sets of very different statistical power. Eggstaff also parsed their out-of-sample results by the size of the training set, and noted a relative decline in *PWComb* for very large training sets. The number of studies with very large calibration sets is quite small raising questions of statistically stability. Finally, we note

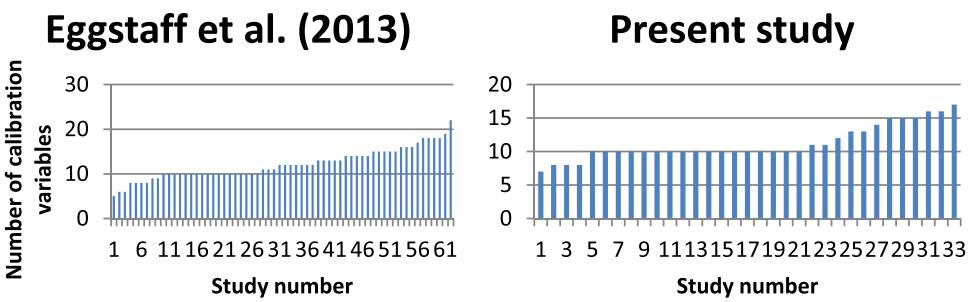

Fig. 13. Distribution of number of calibration variables (vertical axis) in [16] and the present study. The horizontal axis is study number.

that Eggstaff's results have not been recalculated after removing the coding errors (albeit minor) that came to light after publication.

Based on an analysis of a few recent studies, Cooke et al. [15] found that performance averaged over one or two calibration variables presaged the overall performance of *PWComb* relative to *EWComb*. Indeed Fig. 6 could be interpreted as sanctioning a smaller number of calibration variables. However, the large variance in expert weights based on a small number of calibration variables depicted in Fig. 8 counsels caution.

The method employed in the present calculations uses the code of Eggstaff et al. [16], correcting bugs discovered after publication. This is the only cross-validation code verified to have perfect agreement with EXCALIBUR. There are two respects in which these calculations differ from those used in the EXCALIBUR code: First only global performance weights are used as they are easier to implement, whereas item specific performance weights are superior to global weights in 58% of our post-2006 cases (Table 1) and more often used in practice. Second, Eggstaff's code discards experts who assessed less than the full set of calibration variables. It is not uncommon in practice that an expert declines to assess a few calibration variables; this happened in 2 of the 33 post-2006 cases. EXCALIBUR adjusts all statistical accuracy scores to have the statistical power of the smallest number of assessed calibration variables.

9. Conclusion

Society faces consequential decisions that must be taken before the attendant uncertainties can be resolved. Recent emphasis on performance based combinations of uncertainties is found both in the IARPA forecasting tournament and in the recent NAS report on the social cost of carbon. Methods for science based quantification of uncertainty require reliable data on expert performance in the public domain and a critical analysis of the performance of various combination methods.

Although extensive data on expert performance has been available since 2008, it has been largely ignored and the fruits of performance based analysis have largely remained on the vine. Thus harmonic weighting, or "averaging quantiles" is still used by unwary practitioners and even advocated in scientific journals, while an elementary performance analysis could easily predict its strong penchant for overconfidence (as confirmed by the data in Section 2). The notion that performance of expert probability assessors can and should be objectively measured still encounters (mostly passive) resistence.

In cases where cross validation has been undertaken, the methods and results to date lack consistency. As reviewed in Section 4 and detailed in SOM5, individual codes used for cross validation of the classical model show disturbing inconsistencies. Building and vetting a cross validation code is time consuming yet absolutely essential for progress in this field. With such codes in hand, the exhaustive cross validation in sections7 and 8 shows that performance based weighting is superior to equal weighting at the 2.2E-5 significance level. This result is echoed in a very different domain by the results of the Good Judgment Project. Performance based selection of "superforecasters" effectively assigns weight zero to 98% of the project participants.

To make out-of-sample validation practical, methods must be developed which can be computed quickly and compared across studies. The OoSVI based on all training/test sets splits in which 80% of the calibration variables are in the training set offers a number of advantages. First and most importantly, the PW on each such split resembles the PW of the whole study. Second, it can be used to improve the design of Classical Model studies by studying the impact of study parameters on OoSVI. Third, incorporating the OoSVI into the front line processing codes could aid in choosing among the combination schemes. For a given application, if the *PW* combination were superior to the *EW* combination in-sample but not out-of-sample, this might motivate the choice of *EW* in that particular case. Finally, OoSVI could itself be used as a scoring variable for the individual experts, and by extension, as a weighting scheme for crafting better and more robust performance-based combinations. Future research can explore to what extent features of the study, such as the number of experts or method of elicitation (e.g., one-on-one versus group sessions) explain in-sample and out-of-sample performance. A cross validation study of item weights, which are the most common weights used in practice in Classical Model applications, and the best expert would also be a worthwhile endeavor.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ress.2017.02.003.

References

[1] Aspinall Willy P. A route to more tractable expert advice. Nature 2010;463(7279):294–5. http://dx.doi.org/10.1038/463294a.
[2] Aspinall Willy P, Cooke Roger M, Havelaar Arie H, Hoffmann Sandra, Hald Tine. Evaluation of a performance-based expert elicitation: WHO global attribution of foodborne diseases. PLoS ONE 2016;11(3):e0149817. http://dx.doi.org/10.1371/journal.pone.0149817.
[3] Aspinall Willy P, Loughlin SC, Michael FV, Miller AD, Norton GE, Rowley KC, Sparks RSJ, Young SR. The montserrat volcano observatory: its evolution, organization, role and activities. Memoirs, 21. *London: Geological Society*; 2002. p. 71–91. http://dx.doi.org/10.1144/GSL.MEM.2002.021.01.04.
[4] Bamber JL, Aspinall WP, Cooke RM. 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. Clim Change 2016;137(3–4):321–8. http://dx.doi.org/10.1007/s10584-016-1672-7.
[5] Clemen Robert T. Comment on Cooke's classical method. Reliab Eng Syst Saf, Expert Judgement 2008;93(5):760–5. http://dx.doi.org/10.1016/j.ress.2008.02.003.
[6] Cooke Roger M. Experts in uncertainty: opinion and subjective probability in science. New York: Oxford University Press; 1991.
[7] Cooke Roger M. Discussion: response to discussants. Reliab Eng Syst Saf, Expert Judgement 2008;93(5):775–7. http://dx.doi.org/10.1016/j.ress.2008.02.006.
[8] Cooke Roger M. Pitfalls of ROAT cross-validation: comment on effects of over-confidence and dependence on aggregated probability judgments. J Model Manag 2012;7(1):20–2.
[9] Cooke Roger M. Uncertainty analysis comes to integrated assessment models for climate change...and conversely. Clim Change 2012;117(3):467–79. http://dx.doi.org/10.1007/s10584-012-0634-y.
[10] Cooke Roger M. Validating expert judgment with the classical model [Ethical Economy 50]. In: Martini Carlo, Boumans Marcel, editors. . Experts and consensus in social science: critical perspectives from economics, 191–212. Springer International Publishing; 2014, [Ethical Economy 50] ⟨http://link.springer.com/

chapter/10.1007/978-3-319-08551-7_10⟩.
[11] Cooke Roger M. Messaging climate change uncertainty. Nat Clim Change 2015;5(1):8–10. http://dx.doi.org/10.1038/nclimate2466.
[12] Cooke Roger M, Goossens Louis LHJ. TU Delft expert judgment data base. *Reliab Eng Syst Saf*, Expert Judgement 2008;93(5):657–74. http://dx.doi.org/10.1016/j.ress.2007.03.005.
[13] Cooke Roger M, Mendel Max, Thijs Wim. Calibration and information in expert resolution; a classical approach. Automatica 1988;24(1):87–93. http://dx.doi.org/10.1016/0005-1098(88)90011-8.
[14] Cooke Roger M, Solomatine D. EXCALIBUR–integrated system for processing expert judgments, user's manual version 3.0. Delft, The Netherlands: Delft University of Technology and SoLogic Delft; 1992.
[15] Cooke Roger M, Marion E Wittmann, Lodge David M, Rothlisberger John D, Rutherford Edward S, Zhang Hongyan, Mason Doran M. Out-of-sample validation for structured expert judgment of Asian carp establishment in Lake Erie. Integr Environ Assess Manag 2014;10(4):522–8. http://dx.doi.org/10.1002/ieam.1559.
[16] Eggstaff Justin W, Thomas A Mazzuchi, Sarkani Shahram. The effect of the number of seed variables on the performance of Cooke's classical model. Reliab Eng Syst Saf 2014;121:72–82. http://dx.doi.org/10.1016/j.ress.2013.07.015.
[17] Flandoli F, Giorgi E, Aspinall Willy P, Neri A. Comparison of a new expert elicitation model with the classical model, equal weights and single experts, using a cross-validation technique. Reliab Eng Syst Saf 2011;96(10):1292–310. http://dx.doi.org/10.1016/j.ress.2011.05.012.
[18] Genest Christian, Zidek James V. Combining probability distributions: a critique and an annotated bibliography. Stat Sci 1986;1(1):114–35.
[19] Gillingham Kenneth, Nordhaus William, David Anthoff, Geoffrey Blanford, Bosetti Valentina, Christensen Peter, Haewon McJeon, Reilly John, Sztorc Paul. Modeling uncertainty in climate change: a multi-model comparison. Cowles Found Discuss Pap New Haven, CT: Cowles Found Res Econ Yale Univ 2015;2022 ⟨http://cowles.yale.edu/sites/default/files/files/pub/d20/d2022.pdf⟩.
[20] Hald Tine, Willy P Aspinall, Devleesschauwer Brecht, Cooke Roger, Corrigan Tim, Havelaar Arie H, Gibb Herman J, et al. World health organization estimates of the relative contributions of food to the burden of disease due to selected foodborne hazards: a structured expert elicitation. PLoS ONE 2016;11(1):e0145839. http://dx.doi.org/10.1371/journal.pone.0145839.
[21] Harper FT, Goossens LHJ, Cooke Roger M, Hora Stephen C, Young ML, Päsler-Sauer J, Miller LA, et al. Probabilistic Accid Conséq Uncertain Study: Dispers Depos Uncertain Assess 1994;Volume I, [Main report, Volume II: Appendices A and B, Volume III: Appendices C, D, E, F, G, H NUREG/CR-6255, EUR 15855 EN, SAND94-1453. Washington, USA and Brussels-Luxembourg: Prepared for the U.S. Nuclear Regulatory Commission and Commission of European Communities].
[22] Laddaga Robert. Lehrer and the consensus proposal. Synthese 1977;36(4):473–7.
[23] Lichtendahl Kenneth C, Grushka-Cockayne Yael, Winkler Robert L. Is It better to average probabilities or quantiles?. Manag Sci 2013;59(7):1594–611. http://dx.doi.org/10.1287/mnsc.1120.1667.
[24] Lin Shi-Woei, Cheng Chih-Hsing. Can Cooke's model sift out better experts and produce well-calibrated aggregated probabilities? [In]. IEEE Int Conf Ind Eng Eng Manag, 2008 IEEM 2008;2008:425–9. http://dx.doi.org/10.1109/IEEM.2008.4737904.
[25] Lin Shi-Woei, Cheng Chih-Hsing. The reliability of aggregated probability judgments obtained through Cooke's classical model. J Model Manag 2009;4(2):149–61. http://dx.doi.org/10.1108/17465660910973961.
[26] Lin Shi-Woei, Huang Ssu-Wei. Effects of overconfidence and dependence on aggregated probability judgments. J Model Manag 2012;7(1):6–22. http://dx.doi.org/10.1108/17465661211208785.
[27] NAS National Academies of Sciences, Engineering, and Medicine . Valuing climate damages: updating estimation of the social cost of carbon dioxide. Washington, DC: The National Academies Press; 2017. http://dx.doi.org/10.17226/24651.
[28] Vries Hylke de, van de Wal Roderik SW. How to interpret expert judgment assessments of 21st century sea-level rise. Clim Change 2015;130(2):87–100. http://dx.doi.org/10.1007/s10584-015-1346-x.
[29] Wadge G, Aspinall Willy P. A review of volcanic hazard and risk-assessment Praxis at the Soufrière Hills Volcano, Montserrat from 1997 to 2011. Memoirs, 39. London: Geological Society; 2014. p. 439–56. http://dx.doi.org/10.1144/M39.24.
[30] Winkler Robert L. Scoring rules and the evaluation of probability assessors. J Am Stat Assoc 1969;64(327):1073–8. http://dx.doi.org/10.2307/2283486.
[31] Ungar L, Mellers B, Satopää V, Tetlock P, Baron J. The good judgment project: a large scale test of different methods of combining expert predictions. 2012 AAAI Fall Symposium Series; 2012.