Expert forecasting with and without uncertainty quantification and weighting
What do the data say?

Cooke, Roger M.; Marti, Deniz; Mazzuchi, Thomas

DOI
10.1016/j.ijforecast.2020.06.007

Publication date
2020

Document Version
Final published version

Published in
International Journal of Forecasting

International Journal of Forecasting 37 (2021) 378–387

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Expert forecasting with and without uncertainty quantification and weighting: What do the data say?

Roger M. Cooke [a,b,*,1], Deniz Marti [c], Thomas Mazzuchi [c]

[a] Resources for the Future, United States of America
[b] TU Delft, Dept Mathematics, The Netherlands
[c] George Washington University. Department of Engineering Management and Systems Engineering, United States of America

ARTICLE INFO

Keywords:
Calibration
Combining forecasts
Evaluating forecasts
Judgmental forecasting
Panel data
Simulation

ABSTRACT

Post-2006 expert judgment data has been extended to *530* experts assessing *580* calibration variables from their fields. New analysis shows that point predictions as medians of combined expert distributions outperform combined medians, and medians of performance weighted combinations outperform medians of equal weighted combinations. Relative to the equal weight combination of medians, using the medians of performance weighted combinations yields a *65%* improvement. Using the medians of equally weighted combinations yields a *46%* improvement. The *Random Expert Hypothesis* underlying all performance-blind combination schemes, namely that differences in expert performance reflect random stressors and not persistent properties of the experts, is tested by randomly scrambling expert panels. Generating distributions for a full set of performance metrics, the hypotheses that the original panels' performance measures are drawn from distributions produced by random scrambling are rejected at significance levels ranging from $E-6$ to $E-12$. Random stressors cannot produce the variations in performance seen in the original panels. In- and out-of-sample validation results are updated.

1. Introduction

Using expert uncertainty quantification (UQ) as scientific data with traceability and validation dates from (Cooke, 1987, 1991; Cooke et al., 1988) under the name "Classical Model" or "Structured Expert Judgment". The distinguishing features include treating experts as statistical hypotheses and evaluating performance with respect to statistical accuracy and informativeness based on calibration variables (a.k.a. seed variables) from the experts' field to which true values are or become known post-elicitation. The procedure results in combinations of experts' distributions (termed decision makers or *DM*s) using performance-based weighting (*PW*) derived from statistical accuracy and informativeness and using equal weighting (*EW*). The reader is referred to Appendix 1 for specific details. This data enables the study of forecast accuracy, statistical accuracy, and informativeness of experts and of PW and EW combinations. Appendix 2 summarizes and updates in-sample validation results, and Appendix 3 gives information and references for the post 2006 data.

Application highlights involved nuclear safety in the 1990s with the European Union and the United States Nuclear Regulatory Commission, fine particulates with Harvard University and the government of Kuwait in 2004–2005, food-borne diseases for the World Health Organization (WHO) in 2011–2013, ice sheet dynamics for

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

https://doi.org/10.1016/j.ijforecast.2020.06.007

Princeton University, Rutgers University and Resources for the Future (Bamber et al., 2019), and volcanic hazard levels in different parts of the world. The Classical Model was a key decision-support procedure during the prolonged eruption on the island of Montserrat, West Indies, in 1995–2018. Over the same period, expert elicitations using the Classical Model have informed many issues of public health policy, civil aviation safety, fire impacts on engineered structures, and earthquake resistance of critical utility facilities.

Validation is the cornerstone of science. A special issue on expert judgment (Cooke & Goossens, 2008) focused on this issue. Colson and Cooke (2017) gave an extensive review of validation research and applied the cross validation code of Eggstaff et al. (2014) to the *33* professional studies post-2006 available at that time. These studies comprised *320* experts assessing *369* calibration variables. These numbers have since grown to *49* studies involving *530* experts and *580* calibration variables. Recently, Wilson (2017) and Wilson and Farrow (2018) used the pre-2006 data to study expert dependence and its effect on validation both with regard to uncertainty quantification and point predictions. They documented inferior performance of equal weighting but note that very few experts exhibit large error correlations between questions. Their concern that expert performance may not be a persistent property of experts is addressed here with the updated dataset. Their researches into dependence can hopefully be extended with the present data.

The best current summary and applications and validation research on the Classical Model are published by Colson and Cooke (2017, 2018); see esp. online supplements[2]). The reader is referred to these sources for older publications. Post-2006 studies are better resourced, better documented, and more uniform in design than the older studies. They provide a unique data base for comparing expert predictions, and predictions of combinations of experts, with true values from the experts' fields. All data, both pre- and post-2006, are available at http://rogermcooke.net/.

The Classical Model aims at uncertainty quantification, and the underlying performance measures are designed to reward good uncertainty assessors. It is sometimes said that uncertainty quantification is not indicated when the interest is only in point predictions. This unique expert judgment data resource enables that idea to be rigorously tested. In Section 2, it is shown that uncertainty quantification (UQ) improves point predictions in the following sense: combining expert median assessments is inferior to taking the median of a combination of expert UQ's. Equally weighted combination of experts' "best guesses" is the default predictor when UQ is not employed. Against this benchmark (treating experts' medians as "best guesses"), using the medians of performance weighted combinations yields a *65%* reduction in prediction error. Using the medians of equally weighted combinations yields a *46%* reduction.

Section 3 briefly reviews and updates cross validation results. Section 4 addresses the concern of Wilson (2017) and Wilson and Farrow (2018). We test the hypothesis that fluctuations in expert performance can be explained by random stressors during the elicitation, termed the Random Expert Hypothesis (REH). If REH were true, then panel performance should be statistically unaffected by randomly reallocating the individual assessments over the panel members. Note that performance-blind combination methods such as equally weighting expert uncertainties or combining expert quantiles with equal weight are invariant under randomly scrambling the expert assessments. Performance-weighted combinations on the other hand depend on identifying the best performing experts in a panel and assigning them high weight. The hypotheses that panel metrics such as *statistical accuracy of best expert in a panel* and *standard deviation of experts' performance* are statistically unaffected by randomly scrambling expert assessments are rejected at significance levels ranging from $E-6$ to $E-12$. These tests are more powerful than the previous out-of-sample cross validation tests.

The driver behind virtually all these applications is the validational aspect of the Classical Model. Although the psychological community has long drawn attention to cognitive biases inherent in expert UQ, and despite a robust interest in validation research, there are barriers to the use of performance measures. In the conclusion, we speculate on possible explanations for this.

2. Point prediction; a co-benefit of expert UQ

Statistical accuracy and informativeness are performance metrics for quantifying uncertainty. There is nothing in these metrics that rewards proximity of the medians to the true values. If these performance metrics enable more accurate predictions of the true values, then this is a collateral benefit, or co-benefit of performance weighting. The Median Distance for variable i from a distribution with $Median_i$ is defined as $MD_i = |(Median_i - true\ value_i)|$, where $true\ value_i$ is the true value of the calibration variable i.

MD_i is dependent on the scale of variable i; for example, changing from meters to kilometers will affect the value of MD_i. To aggregate over variables with different scales, the scale dependence must be removed. To compare the proximity of the medians of *PW* and *EW* to the realizations, taking the ratio of *MD* for *EW* and *PW* (denoted as *EWMD* and *PWMD*, respectively) per variable removes the scale dependence. These ratios are then aggregated over all variables in a study by taking the geometric mean (geomean):

$$\frac{EWMD}{PWMD} = \left[\prod_{i=1}^{N} \frac{EWMD_i}{PWMD_i}\right]^{1/N}$$

where N is the number of calibration variables.

The geomean is appropriate for aggregating ratios as the geomean of inverse ratios is the inverse of the ratios' geomean and the geomean of ratios is the ratio of geomeans. However, for *10* of the *580* predictions, the item specific PW (PW_i, see Appendix 1) median was actually equal to the realization, making the above ratio infinite (this does not happen for other predictors). These

[2] See https://www.sciencedirect.com/science/article/pii/S0951832017302090?via%3Dihub#s0065

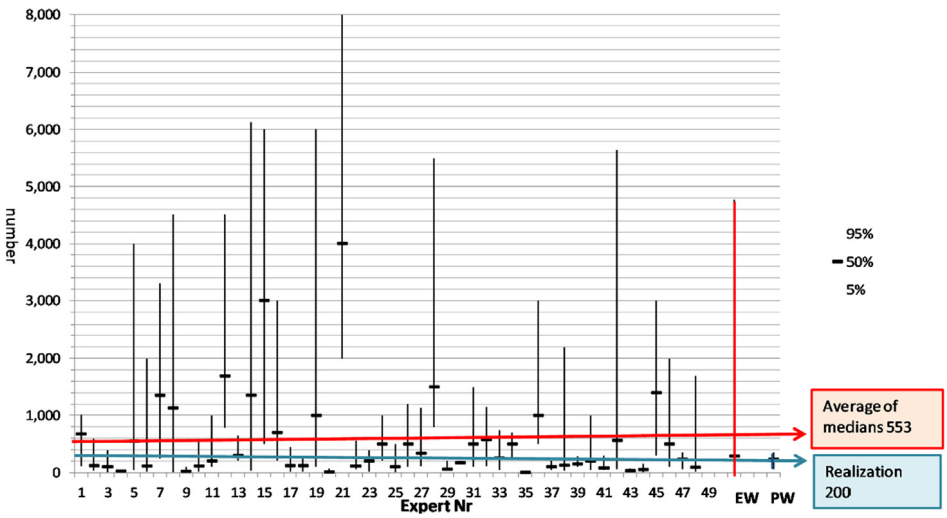

Fig. 1. Responses of 48 experts in a recent CDC elicitation, showing the average of medians (533), the true value or Realization (200), and medians of EW and PW combinations.

10 variables were therefore excluded from the comparisons, although this is prejudicial to PW_i. For a few other variables, $|PW_i - true\ value_i|$ is so small as to render the above ratio very large. While these variables would dominate the average of this ratio, the geomean is relatively insensitive to large outliers.

While the mean of a linear combination of distributions is the linear combination of their means, the same does not hold for the median. A recent elicitation on food borne illness pathways at the CDC illustrates this difference. One of the *14* calibration questions was: "*Between Jan. 1, 2016, and Dec. 31, 2016, a total of 11,277 samples of raw ground beef from 1,193 establishments were tested for Salmonella spp. Of these samples, how many tested positive for Salmonella spp?*" The *5, 50*, and *95* percentiles of the *48* participating experts, and EW and PW combinations are shown in Fig. 1, as is the true value *200*. The medians of the EW and PW combinations are quite close to *200*, whereas the average of medians is *533*.

Although quantile averaging is usually done unawares, Lichtendahl et al. (2013) have recommend this as opposed to "averaging probabilities". This has been shown to produce highly overconfident results (Colson & Cooke, 2017). A brief explanation suffices; consider two experts with *5th* and *95th* percentiles of *[0,1]* and *[10, 11]*, respectively. Averaging their percentiles yields a *90%* confidence interval of *[5, 6]*; equally narrow but disjoint from each expert's confidence interval. Experience in expert judgment shows that such situations are not uncommon. "Averaging the probabilities" requires knowledge of the entire distributions. It is more complex but produces distributions more evenly spread over the interval *[0, 11]*. For purposes of combining distributions, averaging quantiles is a very common and rather severe mistake.

Finding simple point predictions does not require combining distributions. One could simply take a simple linear combination of the experts' medians rather than first combining their distributions. People will continue combining quantiles in any case; therefore, it is useful to assess the ensuing loss of performance. *PWMDQ* and *EWMDQ* denote the performance weighted and equally weighted combinations of the experts' medians ("Q" denotes quantile averaging), respectively. *PWiMD* represents the "high end" predictor based on item specific performance weights (see Appendix 1). EWMDQ is the "low end predictor". It is the predictor most often applied for expert judgment and is therefore chosen as benchmark. Eliciting only a "best guess" should be discouraged. This invites confusion as to whether the mode, median, or mean is intended, and there is no reason to think that performance of "best guesses" in point prediction is any better than *EWMDQ*.

Fig. 2 plots *EWMDQ/EWMD* and *EWMDQ/PWiMD* per study. Values greater than *1* indicate superior predictions relative to *EWMDQ*. The geomean over all 570 predictions of the prediction error ratio *EWMDQ/PWiMD* is *1.65* and that of *EWMDQ/EWMD* is *1.46*. On aggregate, predictions of *PWiMD* are *65%* closer to the truth than those of *EWMDQ*, and the improvement for *EWMD* is *46%*. Even if one is only interested in expert based point predictions, it is preferable for the experts to quantify their uncertainty and to combine their distributions with equal weights. It is better still to measure performance as probability assessors and form weighted combinations of their distributions. These studies have no common experts and may be considered independent. However, taking the geomean of forecast errors per study and then taking the geomean of these geomeans over all studies over-weights the studies with smaller numbers of calibration variables and would result in slightly lower improvements (1.60 and 1.41, respectively).

To compare individual forecasts with realizations, the absolute forecast error must be made scale independent. Table 1 gives the absolute dimensionless forecast errors $|(forecast - true\ value)/true\ value|$ for each of the 569 forecasts for which the realization is non zero. A perfect score would be zero. Averages and standard deviations are strongly affected by large outliers. The geometric average

Table 1
Average and standard deviation of absolute dimensionless forecast errors for item specific performance weights (PWi), global performance weights (PWg), non-optimized global performance weights (PWn), equal weights (EW), performance weighted average of medians (PWQ), and equal weighted average of medians (EWQ) (for definitions see Appendix 1). "rls" denotes "true value" or realization.

	|(PWi - rls)/rls|	|PWg - rls|/rls	|PWn - rls|/rls	|(EW - rls)/rls|	|PWQ - rls|/rls	|(EWQ - rls)/rls|
Ave	2.2	2.7	2.3	3.8	278.6	1472.3
Stdev	11.8	16.0	14.7	45.2	5646.8	33299.8
Geomean	0.38	0.40	0.37	0.43	0.42	0.63

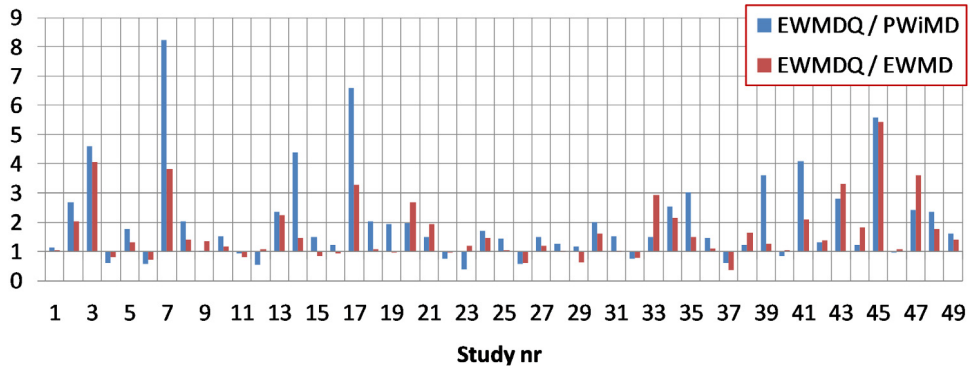

Fig. 2. Ratios of EWMDQ/EWMD (red) and EWMDQ/PWiMD (blue) per study. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(geomean) is insensitive to large outliers. Depending on how one values large outliers, the laurels go to medians of item specific performance weighting *(PWi)* or medians of non-optimized global performance weighting (*PWn)*. In any event, the benchmark of equal weighted combinations of experts' medians is the poorest performer. If one eschews UQ, one is condemned to quantile averaging with a marked penalty in forecast error. UQ without performance weighting avoids the effort of finding suitable calibration variables but suffers a performance penalty.

The mass functions for forecasts whose dimensionless absolute error is less than *5* are shown in Fig. 3. For PW_i, *94* of the *569* forecasts have an absolute dimensionless forecast error less than 0.1, and this number is *81* for EW.

3. Out-of-sample cross validation

Considerable effort has been devoted to validating expert judgments. Unless the variables of interest can be observed shortly after completion of a study, out-of-sample validation comes down to cross validation. The calibration variables are split into a training set for initializing the *PW* and a test set for comparing *PW* and *EW*. The sets on which the performance weights are derived and evaluated are thus disjoint. Early efforts with different methods produced uneven results as documented by Colson and Cooke (2017). Col. Eggstaff and colleagues produced an out-of-sample validation code and applied this to all pre-2008 expert data. To our knowledge, it is the only code that has been benchmarked against the standard expert judgment software. This code was applied to 33 post-2006 studies by Colson and Cooke (2017) and is here extended to the current set of 49 post-2006 studies. Understanding the strengths and limitations of cross validation is essential to appreciate the contributions of testing the REH.

Many issues involved in choosing the training and test set sizes are discussed in (Colson & Cooke, 2017), to which we refer the interested reader. The upshot is that using *80%* of the calibration variables as a training set balances best the competing goals of resolving expert performance on the training set and resolving the performance of combinations on the test set. The training set then has enough statistical power to reduce the variance in the expert weights, thereby rendering the performance weights similar to the weights based on all calibration variables. The test set loses statistical power for resolving the *PW* and *EW DM*s, but with *10* calibration variables, statistical accuracy scores for assessments of *5th, 50*th, and *95*th percentiles still vary by a factor of *31*. Moreover, higher resolution is of no value if the *PW DM* is very volatile and unlike the *PW DM* of the full study. Of course, the actual sizes of the training and test sets vary with the total number of calibration variables. The *80%* split makes it easier to pool the results from all studies. With *10* calibration variables, there are *45* distinct *8*-tuples of calibration variables to be used as training sets. Performance is scored on the *2* remaining variables. The statistical accuracy, the informativeness, and the combined score (the product of the former two) are averaged over the *45* different test sets. Colson and Cooke (2017) show that the average ratio of combined scores for *PW* and *EW* is indistinguishable from the ratio of average combined scores for fixed training set size. The ratio of combined scores based on 80% of the calibration variables is called the *"Out of sample Validity Index (OoSVI)"*.

For *42* of the *49* studies, the ratio of *PW* comb / *EW* comb is greater than *1*. Under the null hypothesis that there

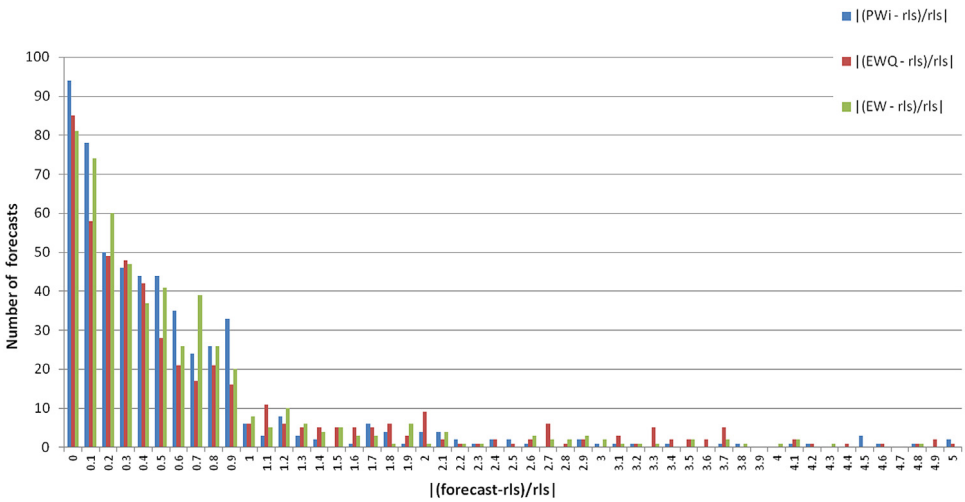

Fig. 3. Dimensionless absolute forecast errors less than 5 and PWi (541 forecasts) EW (534 forecasts) and EWQ (498 forecasts). "rls" denotes realization or true value.

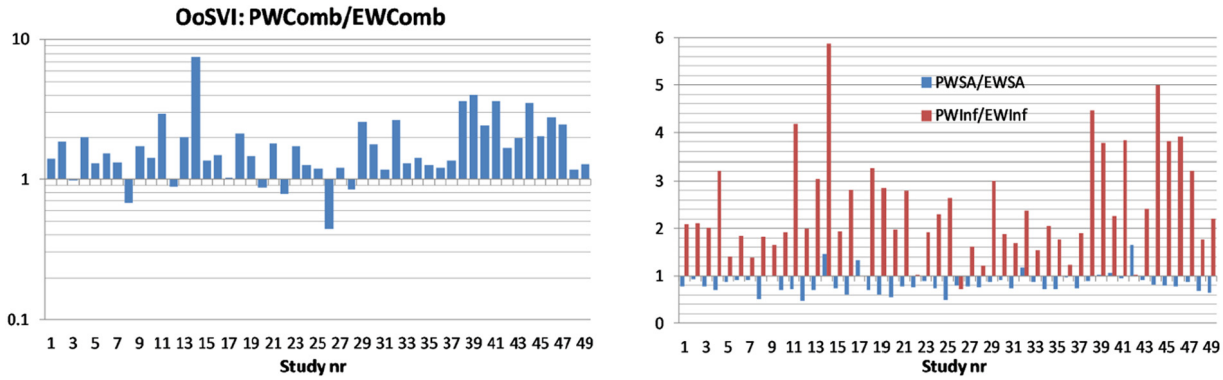

Fig. 4. Ratios of combined scores PWcomb/EWcomb averaged over all training sets sized at 80% of the calibration variables for 49 post-2006 studies (left). Combined scores factored into ratios of statistical accuracy (SA) and informativeness (Inf) (Right).

is no difference between *PW* and *EW*, the probability of seeing *42* or more ratios greater than *1* is *1.8E−7*. The right panel of Fig. 4 shows that *PW* suffers a modest out-of-sample penalty in statistical accuracy, which is more than compensated by a boost of informativeness. Lest this boost seem "small", it is well to note that a factor 2 increase in informativeness corresponds roughly to halving the length of the 90% confidence bands. If an assessor is perfectly statistically accurate, his/her statistical accuracy score is uniformly distributed on the interval [0, 1] with expected value 0.5. The mean out-of-sample statistical accuracy score (p value) for EW is 0.54, while that of PW is 0.43. In a nutshell, CM is able to boost informativeness substantially without sacrificing statistical accuracy.

As in (Colson & Cooke, 2017), the features which best explained the differences in the *OoSVI* were studied. The results echo those earlier findings; if *PW* concentrates all the weight in the best expert *(BE)*, the overall geomean of *OoSVI* for all studies *(1.63)* splits into *2.0 (PW = BE)* and *1.4 (PW ≠ BE)*. Similar results are obtained by splitting into studies in which BE's statistical accuracy is above 0.5 *(2.1)* or below 0.5 *(1.2)* (see Fig. 5). Other features such as number of experts, number of calibration variables, and plenary versus individual elicitation had less effect. The quality of the best expert is the main determinant for *OoSVI*. The rank correlation between *OoSVI* and the in-sample ratio *(PW/EW)* of combined scores is *0.5*; these measures are related but not identical for reasons addressed in the following paragraph.

Cross validation is essential for demonstrating the value of *PW* relative to *EW* for out-of-sample prediction. The necessity of splitting the calibration set into training and test sets exacts a toll that is illustrated with the "Ice Sheet 2018" study ('ICE_2018;' see Appendix 3) involving *20* experts and *16* calibration variables. With a training set of *13* (*80%* of *16*), there are *560* distinct training sets, and *8* of the *20* experts were weighted on at least one of these sets. For *7* of these *8*, the difference between their maximal and minimal weight was *1*; that is, their weights vacillated between *0* and *1*. The *PW* combinations evaluated on the 3 test variables still exhibit volatility and deviate from the *PW* of the entire study. The most definite

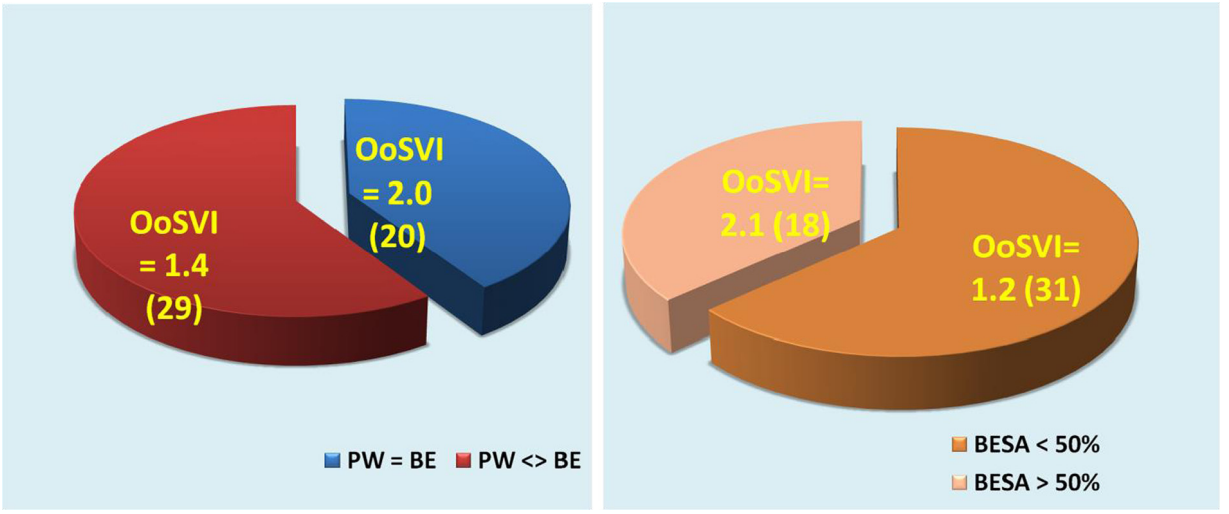

Fig. 5. Breakdown of OoSVI scores in Fig. 4 according to whether PW was identical with the best expert (PW = BE; left) and whether BE's statistical accuracy was greater than 50% (BESA>50%; right). The Geomeans of the 49 studies are shown for the of ratios of combined scores of PWcomb/EWcomb evaluated study-wise for training sets sized at 80% of the calibration variables. *Note.* OoSVI refers to out-of-sample-validity index. BE denotes the best expert, BESA denotes the statistical accuracy of the best expert.

assertion that can be made is that the *OoSVI* compares the score of *EW* with the scores of a swarm of *PWs*, which loosely resembles the *PW* of the full study. The cross-validation data validates the performance weighting method, but not a specific *PW* combination.

4. The random expert hypothesis

Expert judgment studies differ from simple surveys in that experts undergo a credentialing process before being empanelled. It is natural to believe that, once empanelled, experts should have an equal share in determining the panel output. This in turn leads naturally to aggregation schemes based on equal weighting. Against this viewpoint is a mounting body of evidence that, with regard to uncertainty quantification, experts' performances are very uneven. About three-quarter of the *530* experts reviewed in Appendix 2, considered as statistical hypotheses, would be rejected at the 5% level. From this perspective, avowals of degrees of confidence in outcomes based on experts' judgments are susceptible to and in need of validation. This is hardly new. The US Defense Intelligence Agency adopted expert subjective probabilities in 1976 and dropped them a few years later for lack of validation.[3] Nonetheless, the theoretical debate over whether and how to validate expert subjective probabilities continues.

Recently, a new approach to validation has emerged (Marti et al., 2019). Whereas cross validation is hampered by the two-sided loss of statistical power caused by splitting calibration variables into training and test sets, the new approach does not focus on the performance of a combination of experts. Instead, it focuses on the expert performances themselves and investigates the assumption underlying all performance-blind approaches; namely that performance measures are unable to make meaningful distinctions in experts' performances (the Random Expert Hypothesis—*REH)*. This may be because the experts are all equally good or equally bad. It may also be that any putative differences are swamped by the noise inherent in expert judgment—experts are influenced by random stressors, they may get distracted or fatigued, their performance is affected by the particular choice of calibration variables, etc. Concerns in this regard have been raised by Wilson (2017) and Winkler et al. (2018).

Note that if experts' differences in performance are caused by factors local to the elicitation of calibration variables, then the cross-validation results of the previous section might not extend to the variables of interest. Further, the low power of the test set means that the *DMs*' statistical accuracy scores are poorly resolved and thus extra susceptible to random stressors. The new approach to validation focuses directly on *REH* without the intermediary of performance-based combinations. Indeed, the *REH* itself carries a heavy proof burden and can be tested using expert performance provided by our *49* post-2006 studies. Intuitively, if performance differences are the result of noise, then randomly reallocating the experts' assessments among the panel members will randomly redistribute the random stressors. The fluctuations in performance produced in this way should envelope the performance in the original panels. It emerges that tests of *REH* are much more powerful than the cross-validation tests.

To make this idea precise, consider a *random scramble* of an expert panel composed of *15* experts and *10* calibration variables. '*Scrambled expert 1*' is created by randomly choosing an assessment without replacement from one of the *15* experts for the first variable, a second random draw without replacement gives the second

[3] See https://www.resourcesmag.org/common-resources/iceman-cometh/.

assessment for '*scrambled expert 1*', and so on. '*Scrambled expert 2*' chooses assessments in a similar way from the assessments not chosen by '*scrambled expert 1*'. The final scrambled expert, '*scrambled expert 15*', gets the leftovers. In this scrambled panel, we can measure the statistical accuracy (*SA*) and informativeness (*Inf*) of each expert, the combined scores ($SA \times Inf$), the average scores, and the maximum, minimum, and standard deviation of the scores.

For each study, we repeat the scrambling *1000* times and build up a distribution for each performance metric. This distribution reflects the variation we should see in that performance metric if experts' performance differences were only due to random stressors. Suppose we compute the average *SA* for experts in each of the *1000* scrambled panels for a given study. The *REH* now asserts that the average *SA* in the original panel could just as well be any of 1000 averages in the scrambled panels. There should be a *50%* chance that the original average *SA* is above the median of the *REH* distribution, a *5%* chance that it is above the *95th* percentile of the *REH*, etc. Thus, *REH* expects that in *2.45* of the *49* studies, the original average *SA* should fall above the 95th percentile of the *REH* distribution. In fact, this happens in *20* of the *49* studies. The probability of *20* or more studies falling above the 95th percentile if *REH* were true is *6.6E−14*. *REH* fails if the differences in the experts themselves in the original panel are greater than what can be produced by scrambling the experts.

Note that random scrambling will have no effect on the *EW* combination. Assuming that *EW* is at least as good as *PW* implies *REH*. In consequence (modus tollens), if *REH* is (statistically) rejected, then so is the assumption that *EW* is at least as good as *PW*. In this sense, *REH* provides a more powerful test of the assumption underlying the use of *EW*. The same holds for the "averaging quantile" approaches (Lichtendahl et al., 2013) or indeed any approach which is performance-blind. If all experts in a panel are "equally good" or "equally bad", then *REH* may actually be true for that panel. The use of *PW* depends on the fact that such panels are in the minority. Testing *REH* on a set of cases allows us to gauge the size of that minority.

The data has been standardized in ways that do not affect the *REH*; experts who did not assess all calibration variables were dropped, reducing the number of experts from 530 in Figure A2.1 to 526. All background measures are converted to uniform (most calibration variables already have uniform backgrounds). Whereas some studies assessed the 25th and 75th percentiles in addition to the *5*th, *50*th, and *95*th percentiles, Marti et al. (2019) showed that this had no effect on the REH, and so only the *5*th, *50*th, and *95*th percentiles are used.

For each of the *49* studies, the following eight performance metrics shown in Fig. 6 are computed for the original panel and for each of the *1000* scrambled panels:

1. Panel Average Statistical Accuracy
2. Panel Max Statistical Accuracy
3. Panel Standard Deviation of Statistical Accuracy
4. Panel Min Statistical Accuracy
5. Panel Average Combined Score
6. Panel Max Combined Score
7. Panel Standard Deviation of Combined Score
8. Panel Min Combined Score

For metrics not involving the Min, we are interested in the quantile of the *REH* distribution realized by the metric in the original panel. For metrics 4 and 8 we are interested in the complimentary quantile, that is, the fraction of the *1000* scrambled panels in which the original minimum is *lower* than the scrambled minima. This is done so that all metrics have the same sense; numbers close to 1 are favorable for *PW*. We test the *REH* against the alternative that experts' performance differences are too large to be explained by random stressors and that high values are critical.

If *REH* were true, that is, if the original panel's metrics were really drawn from the *REH* distributions, then the quantiles in Fig. 6 should be uniformly distributed on the interval *[0, 1]*. The number of bars above the value *0.5* should be statistically equal to the number below *0.5*. The "amount of color" above *0.5* should statistically equal the amount below *0.5*.

There are two simple tests for the *REH* hypothesis. The binomial test simply counts the number of values greater than *0.5* for each metric and reports the *p* value for the corresponding null hypothesis: *the probability that half of the random panels outperform the original panel metric is 0.5*. The binomial test does not consider how far above or below *0.5* the metrics are. The *sum test* simply adds the *49* original panel quantiles for each metric. Under *REH*, this sum should be (very nearly) normally distributed with mean $49/2 = 24.5$ and standard deviation $(49/12)^{1/2} = 2.02$. For example, 'Average Statistical Accuracy' in the original panel exceeds the median of the *REH* distribution for 'Average Statistical Accuracy' in *42* of the *49* studies. If the probability of exceeding the median were really *0.5*, the probability of seeing *42* or more "successes" would be *1.81E−7*. Summing the original panels' *49* realized quantiles in the *REH* distribution for 'Average Statistical Accuracy' yields *38.36*. The probability that a normal variable with mean *24.5* and standard deviation *2.02* exceeds *38.36* is *3.48 E-12*. The sum test is much more powerful than the binomial test. Table 2 collects the results for the binomial and sum tests. Interestingly, departures from *REH* seem more severe for the panel minima than for the panel maxima. In other words, REH has more difficulty reproducing the very low scores than the very high scores. Of course, both departures are highly significant. Suppose we reject *REH* for each of the *49* studies. The sum of the *p* values (1 − percentile of original panel) gives the expected number of false rejections. This number is $49 - 38.36 = 10.64$. We might expect that REH is true in one-fifth of the studies.

Whichever test we use, the notion that putative differences in expert performance are due to random stressors is overwhelmingly rejected. Table 3 examines the influence of the number of experts and number of calibration variables on the performance metrics.

With *49* samples, a rank correlation of 0.24 is significant at the 5% level. As seen in Table 3, the number of experts is not strongly associated with any of the metrics. The number of calibration variables does appear to

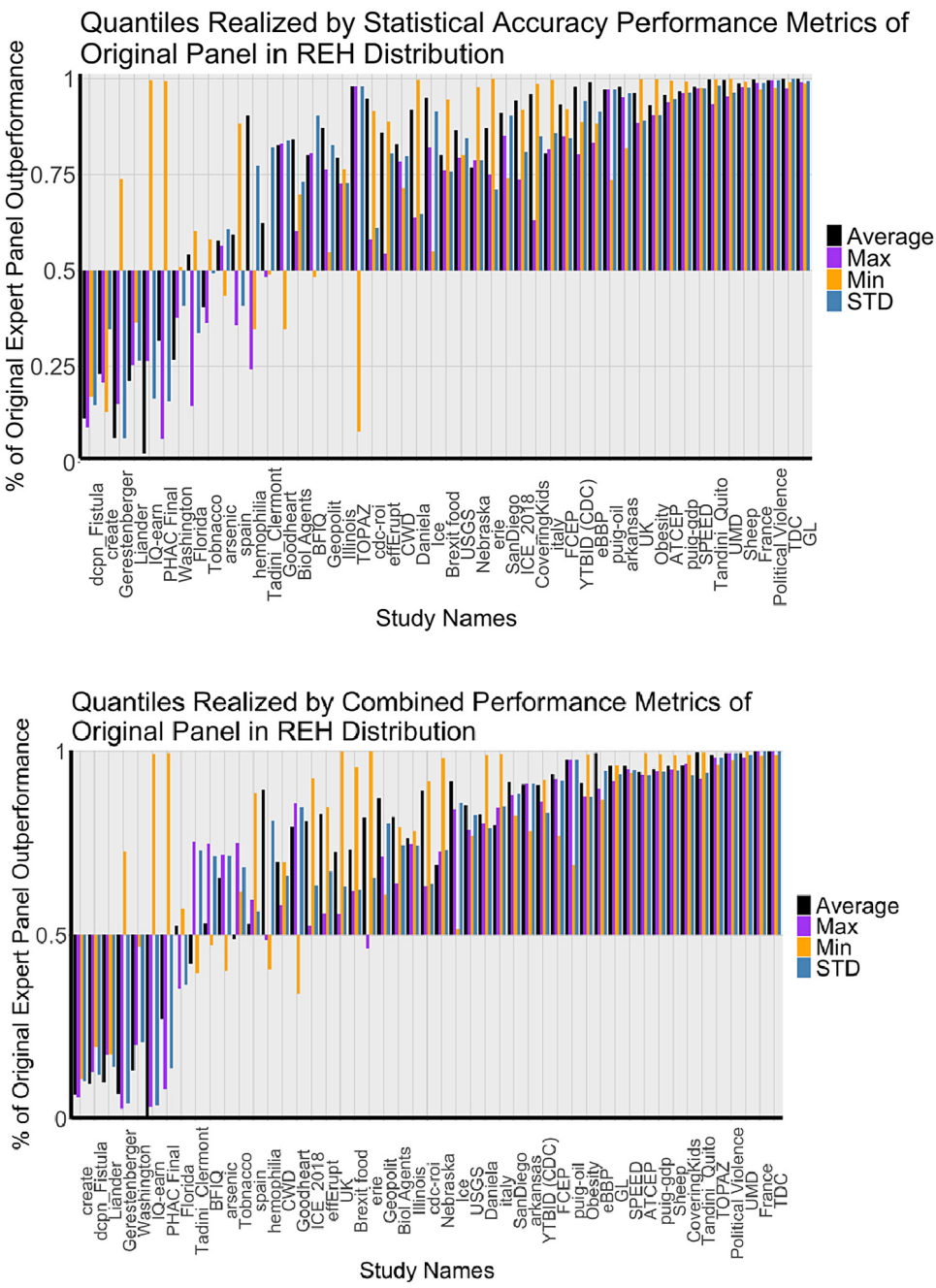

Fig. 6. Quantiles of the REH distributions realized by the performance metrics for Statistical Accuracy (top) and Combined Score (bottom) by the original expert panels, for 49 studies.

exert some influence. Table 3 implies that more calibration variables tend to make the differences between the performance of original experts and randomly scrambled experts greater.

5. Conclusions

Experts exhibit undeniable differences in their ability to quantify uncertainty. These differences can readily be measured and used to improve uncertainty quantification and point predictions. The evidence is overwhelming. However, there are two significant hurdles to applying performance-based combinations: (1) time and effort required for performance measurement and (2) numeracy demands on the analyst.

In most applications, the greatest time and effort is spent in formulating clear questions with operational meaning which address the issues at hand. It is useful to

Table 2
p values at which REH is rejected for the seven performance metrics.

	p values for tests of *REH*	
	Binomial test	Sum test
Average statistical accuracy	1.81E−07	3.49E−12
Standard deviation of statistical accuracy	4.63E−06	6.76E−10
Maximum statistical accuracy	2.35E−04	2.85E−06
Minimum statistical accuracy	4.63E−06	2.99E−11
Average combined score	9.82E−07	3.21E−09
Standard deviation of combined score	9.82E−07	3.96E−08
Maximum combined score	1.92E−05	5.85E−07
Minimum combined score	4.63E−06	6.67E−12

Table 3
Spearman's rank correlation between number of experts and number of calibration variables and percentile scores.

	Quantile of Avg *SA*	Quantile of STD *SA*	Quantile of Max *SA*	Quantile of Avg. Comb	Quantile of STD Comb	Quantile of Max Comb
Rank correlation to # experts	0.24	0.13	0.03	0.18	0.05	−0.04
Rank correlation to # variables	0.40	0.39	0.39	0.31	0.34	0.31

Note. Avg denotes average, *SA* is for Statistical Accuracy, STD is the standard deviation, Comb. is for combined score, Max is for maximum.

think of expert judgment as a way of obtaining (probabilistic) results from experiments or measurements which are feasible in principle but not in practice. Describing a "thought experiment" is the best way of making clear what one is asking. These efforts should be made in any case, but the need for operational meaning is easier to ignore if there are no calibration questions. Having formulated questions with clear operational meaning facilitates finding calibration variables from the experts' field. The impractical experiments or measurements often suggest experiments and or measurements which are already performed though not published. Often, as in the recent 'Ice Sheet 2018' study, more time is spent agreeing on the best set of calibration variables than in generating them. That said, finding good calibration variables does require a deep dive into the subject matter and is greatly aided by having a domain expert on the analysis team. Quigley et al. (2018) provide guidance on finding calibration variables.

One-on-one interviews cost time and money, although good online meeting tools bring these costs down significantly. One-on-one elicitation enables the analysts to better plumb experts' reasoning. Supervised plenary elicitation in which experts meet, discuss, and then individually perform the elicitation offer advantages of speed and disadvantages in loss of individual engagement.

Sending a questionnaire in the mail to a large set of experts in the hope that a fair number will respond is discouraged for purposes of uncertainty quantification. Expert surveys should be sharply distinguished from structured expert judgment.

Despite all this, the most difficult hurdle is the second: finding qualified analysts. Mathematicians, statisticians, engineers, and scientists know that the Classical Model is not a heavy lift.[4] Many have conducted successful studies in their chosen fields. The analyst must be able to explain the method to the experts and to the problem owners, so that they in turn can explain it up the chain. If a problem owner is unable to explain the method to his/her superiors, (s)he is unlikely to adopt it. The analyst must be comfortable with certain relevant concepts such as statistical likelihood, *p* values, Shannon information, scoring rules, distributions, densities, quantiles, etc. Some knowledge of foundations is needed to explain why uncertainty is represented as subjective probability and not as fuzziness, imprecision, degree of possibility, or certainty factors, to name a few.[5] Writing up the results in a clear an accurate fashion requires more than a nodding acquaintance with all these concepts.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijforecast.2020.06.007.

References

Bamber, J. L., Oppenheimer, Kopp, R. E., Aspinall, W. P., & Cooke, Roger M. (2019). Ice sheet contributions to future sea level rise from structured expert judgement. *Proceedings of the National Academy of Sciences of the United States of America*, https://doi.org/10.1073/pnas.1817205116. https://www.pnas.org/content/early/2019/05/14/1817205116.

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

[4] Cooke (2015) esp. the supplementary online information is written to bring neophytes up to speed. The TU Delft launched a free online course on expert judgment in October 2019.

[5] See (Cooke, 2015) for a foundational discussion targeting potential analysts.

Colson, A., & Cooke, R. M. (2018). Expert elicitation: Using the classical model to validate experts' judgments. *Review of Environmental Economics and Policy*, *12*(1), 113–132, https://doi.org/10.1093/reep/rex022. https://academic.oup.com/reep/article/12/1/113/4835830.

Cooke, Roger M. (1987). *A theory of weights for combining expert opinions*: *Report 87-25*, Dept. of Mathematics, Delft University of Technology.

Cooke, Roger M. (1991). *Experts in uncertainty; opinion and subjective probability in science* (p. 321). New York Oxford: Oxford University Press, ISBN: 0-19-506465-8.

Cooke, Roger M. (2015). Messaging climate change uncertainty with supplementary online material. *Nature Climate Change*, *5*, 8–10, http://dx.doi.org/10.1038/nclimate2466, Published online 18 December 2014 http://www.nature.com/nclimate/journal/v5/n1/full/nclimate2466.html.

Cooke, Roger M., & Goossens, L. H. J. (2008). Special issue on expert judgment. *Reliability Engineering & System Safety*, *93*, 657–674, Available online 12 May 2007, Issue 5, 2008.

Cooke, Roger M., Mendel, M., & Thijs, W. (1988). Calibration and information in expert resolution. *Automatica*, *24*(1), 87–94.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's classical model. *Reliability Engineering & System Safety*, *121*(2014), 72–82. http://dx.doi.org/10.1016/j.ress.2013.07.015.

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

Marti, H. D., Mazzuchi, T. A., & Cooke, R. M. (2019). *Are Performance Weights Beneficial? Investigating the Random Expert Hypothesis*. appearing in Hanea, Nane, French and Bedford.

Quigley, J., Colson, A., Aspinall, W., & Cooke, R. M. (2018). Elicitation in the classical model. In L. K. Dias, A. Morton, & J. Quigley (Eds.), *Elicitation the science and art of structuring judgement* (pp. 15–36). Springer.

Wilson, K. J. (2017). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*, *33*(1), 325–336. http://dx.doi.org/10.1016/j.ijforecast.2015.11.014.

Wilson, K., & Farrow, M. (2018). Combining judgements from correlated experts. In A. Dias, & J. Quigley (Eds.), *Elicitation the science and art of structuring judgement* (pp. 211–240). Springer.

Winkler, R. L., Grushka-Cockayne, Y., K.C., Lichtendahl Jr., & Jose, V. R. R. (2018). Averaging Probability Forecasts: Back to the Future, Working Paper | HBS Working Paper Series | 2018.