Wisdom or Madness: Expert Data on Wisdom of Crowds

Roger M. Cooke

Dept Mathematics, Delft University of Technology, The Netherlands, cooke@rff.org

From structured expert judgment data with realizations it is concluded that (1) experts' Mean Absolute Percentage Errors are very fat tailed, making convergence problematic, (2) probabilistic proximity of experts' median forecasts to realizations are modestly dependent, whereas experts' abilities to catch realizations in their 90% bands are much less so, (3) expert agreement does not predict expert panel performance, (4) regarding the performance metrics Statistical Accuracy and Mean Absolute Percentage Errors, number of experts is helpful for the first, harmful for the second whereas dependence in placement of medians is harmful for the first, helpful for the second, and (5) following Jensen's inequality, averaging experts' median assessments is slightly better than choosing a random expert but (from a previous publication) much worse than the median of equally weighted or performance weighted combinations of experts' distributions, underscoring the importance of method of aggregation. Probabilistic crowds are wiser than point forecast crowds.

Keywords: Expert Judgment, MAPE, Fat Tails, Agreement, Dependence, Diversity, Explained Variance.

1. Introduction

It seems to have started in 1841 with Charles Mackay's Memoirs Of Extraordinary Popular Delusions and The Madness Of Crowds. Francis Galton parried in 1907 with Cornwall fair goers' average (originally median) estimate of a dead bull's weight which was nearly spot on. James Surowieki's The Wisdom of Crowds (2004) distinguished wise crowds from irrational crowds on five criteria: diversity, independence, decentralization, aggregation and trust. Douglas Murray brought us back to The Madness of Crowds, gender, race and identity (2019). Much is written on the credibility of crowds. Lacking is any scientific use of expert probabilistic forecasting data for which realizations or true values are available. Emphasis is placed on "expert" and "probabilistic" for a number of reasons: (1) the inevitable winnowing of reliable crowds often turns on predicates associated with expertise, (2) experts' scientific training distinguishes knowledge from uncertain guesses, the provenance of forecasting, (3) probabilistic forecasting converts all quantities to a common scale, namely probability, because of which (4) we can develop performance metrics applicable to any forecast situation, and finally (5) we have extensive data from 107 structured expert judgment (SEJ) panels. There is even discussion whether Galton's "vox populi" shouldn't be called "vox expertorum" given the large number of expert butchers and farmers attending these events.

SEJ panels consist, on average, of 11 vetted experts giving 5, 50 and 95 percentiles for uncertain variables from their fields and also for, on average, 14 calibration variables from their fields to which true values are or become known. Performance on these calibration variables is used to construct performance weighted combinations and compare with equally weighed combinations. Expert performance is persistent, performance based combinations are superior to equal weight combinations both in– and out–of sample and have been evaluated in real applications (Cooke et al 2021, Aspinall, 2010).

2. Crowd–Casting Versus SEJ Forecasting

When crowd–casting and expert forecasting mingle, Surowieki's criteria run up against expert communalities. Scientists in an SEJ forecasting panel have similar training, follow the same literature and often know each other. Physicist Max Planck (1950) famously quipped "science advances one funeral at a time". Surowieki opines: "Homogeneous groups, particularly small ones, are

Proceedings of the 35th European Safety and Reliability and the 33th Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage & Marja Ylönen

ISBN: 981-973-0000-00-0 :: doi: 10.3850/981-973-0000-00-0 esrel2022-paper

often victims of what the psychologist Irving Janis called "groupthink."(Surowieki 2004, p.36) "After a survey of expert forecasts and analyses in a wide variety of fields, Wharton professor J. Scott Armstrong wrote,'I could find no studies that showed an important advantage for expertise'." (Ibid p.33). The antidote is crowd size: "...much of what we've seen so far suggests that a large group of diverse individuals will come up with better and more robust forecasts and make more intelligent decisions than even the most skilled "decision maker."(Ibid p.32). Au contraire, says Naomi Oreskes in Why Trust Science (2019): scientific consensus resulting from rigorous peer review provides a basis for trust.

SEJ data is used to examine two pillars of WOC: (1) Is crowd size really beneficial? and (2) Is "diversity" / "independence" beneficial? To address these, 40 forecasting panels with at least 10 experts and at least 10 calibration variables are selected giving 586 forecast variables with realizations, 698 experts and 10,189 expert forecasts (see appendix Table A1). "Beneficial" is measured by two performance metrics.

The absolute percentage error for forecast f with realization r is $|(f–r)/r|$ and is unstable for r close to zero. Absolute percentage error is scale invariant, so scores for different forecasts and different realizations can be averaged, yielding the Mean Absolute Percentage Error (MAPE). We can average over all calibration variables for each expert to form an expert MAPE, we can average the expert MAPEs for all experts in a panel to arrive at an expert panel MAPE which is the expected MAPE of a randomly chosen expert. Invoking the Wisdom Of Crowds (WOC) we can first average experts' median forecasts and then compute the WOC MAPE, per variable and per panel. By Jensen's inequality WOC MAPE is always less or equal to expert panel MAPE, though the mean difference per panel is a mere *0.009* in this case (see appendix).

Statistical Accuracy (SA) is based on the relative frequency with which the realizations of independent calibration variables fall inside the forecaster's four inter–quantile intervals. SA is the probability that these relative frequencies should differ from the theoretical inter quantile probabilities (5%, 45%, 45%, 5%) by at least the observed amount. Low values near zero mean that it is very unlikely that the forecaster's probabilities are statistically accurate, high values, near 1, indicate good agreement between observed and expected relative frequencies.

3. Tail Size[a]

Participants in WOC discussions need to appreciate how much the discussion has been constrained by statistical assumptions, and how fragile these assumptions really are. If we sample a set of numbers from some distribution, we can always compute the average of these numbers as well as the variance, standard deviation, correlations with other sets of numbers etc. But if we sample more numbers or sample a like-sized second batch, do these averages, variances and correlations tend to agree? The law of large numbers says that averages, variances, and correlations stabilize as we draw ever larger samples; however this law applies only if the distribution from which the numbers are drawn is "thin tailed". If the distribution is "fat tailed" then none of this holds.

Pictures give a better idea than formal mathematical definitions. The left panel of Figure 1 gives running averages (average the first two, then the first three, etc) of 1000 independent samples from a uniform distribution on the [0, 1] interval. The horizontal axis gives the size over which the average is taken. On the vertical we plot the running average. At the horizontal value 1000 we average all 1000 samples. In the right panel we do the same, with the same numbers, except that these numbers are now inverted; 0.1 becomes 10, etc. The inverse of the uniform distribution is a very fat tailed distribution. With thin tailed distributions running averages converge, with fat tailed distributions they do not. Very large values keep popping up at a rate which prevents convergence. Fat tailed distributions are not exotic, but are not common knowledge.

[a] This is a non-technical introduction to fat tailed distributions. Many text books give a full mathematical treatment, Cooke et al (2014) is directed to numerate non-specialists.

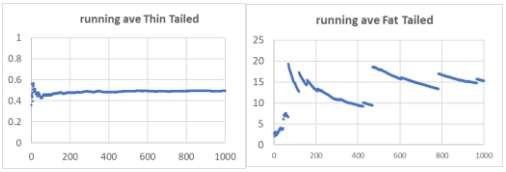


Fig. 1: Thin tailed (left) and fat tailed (right) running averages. Horizonal axes denote the number of samples over which we average.

If we draw repeated samples of size 1000, the thin tailed running averages will differ a bit at the beginning but quickly settle into the pattern. Figure 2 shows what happens with three samples of 1000 from the from the distributions in Figure 1. Notice the changing scale on the vertical axis for the fat tailed distributions; these samples do not settle into a pattern, they are dissipative.

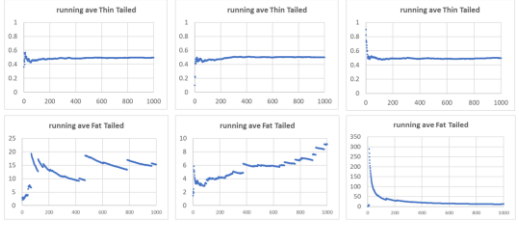


Fig. 2: Three Repeated samples with running averages

Statisticians don't like fat tails, as they prevent application of the familiar statistical methods. It's easy to delete a single large value as an "outlier" so that the rest "look normal". However, when one looks at larger samples from a fat tailed distribution, one realizes that the large values are characteristic of the whole distribution. If we order the sample from smallest to largest values, we see that the distance between adjacent samples just gets larger as the sample values get larger. As we gather more samples, the average of the whole sample tends to resemble the largest sample in the set. In fact, this is a defining feature of the "subexponential" class of fat tailed distributions.

Once we look, we can see fat tailed distributions everywhere - damages from natural disasters, crop insurance claims, citation scores, flood damages, income distributions, hospital discharge rates etc. Loss distributions in risk analysis are often fat tailed (Cooke et al 2014). What about experts?

4. Crowd Size

Do averages of ever more forecast errors trend down? There is an antecedent question: Do such averages converge at all? Figure 3 shows running averages of US damages in excess of 10$M due to natural disasters (left) and absolute percentage error in 10,198 expert forecasts (right). To be sure, these experts assess different quantities, but their absolute percentage errors can be plotted on an absolute scale reflecting the factor by which the forecast differs from the realization in absolute value. Such graphs depend on the ordering but random re–orderings will exhibit the same key feature: ever larger values keep popping up that prevent convergence.

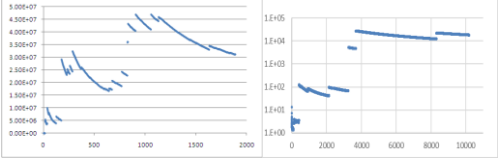


Fig. 3: Running averages for US damages in excess of 10$M (right, Cooke et al 2014) and running averages for 10,189 expert absolute percentage errors.

Experts' absolute percentage errors in aggregate are very fat tailed and averages do not appear to converge. Does this also apply to WOC forecasts? Figure 4 (left) shows 586 realizations in ascending order plotted with their WOC forecasts. Note the very small realization with forecast differing by 8 orders of magnitude. To avoid instabilities due to small realizations, we subset the 535 forecasts for which the realizations are greater or equal to 0.1. The running averages are shown in Figure 4 (right); the averages of ever larger sets of absolute percentage forecast errors just keeps growing.

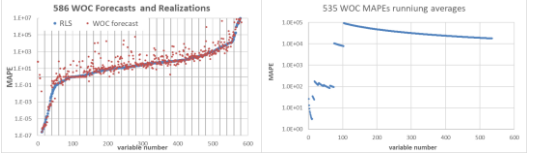


Figure 4: 586 WOC forecasts against realizations (left) and running averages of 535 WOC MAPEs (right)

The panel sizes in our dataset do not support tail analysis per panel, but the effect of number of forecasters can be seen in other ways. Figure 5 plots all 586 WOC MAPEs against the panel size over which the median forecasts are averaged. The rank correlation in Figure 5 is weakly positive. WOC panel MAPES are not decreasing in panel size.

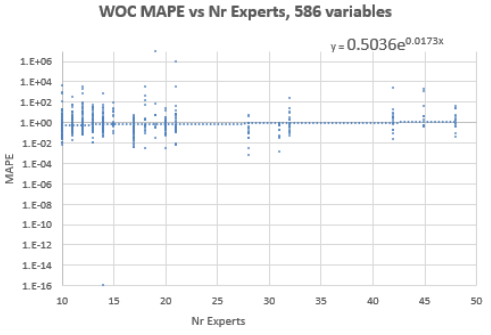


Fig. 5: WOC MAPE against number of empanelled experts for 586 variables.

5. Dependence / Diversity

Increasing crowd size can have no effect if people all say the same thing. Diversity and dependence must be addressed. If we look only at expert forecasts in a panel, then of course they will be dependent for the simple reason that they are forecasting the same quantity. All forecasts of a volcanic eruption in m^3 will be large, all forecasts of the weekly growth of the dome in *m* will be small. Apparently that's not the right question. We should be asking about dependence in experts' forecast *errors* and error must be relative to the realization. Diversity usually refers to something like 'different world views'. This is operationalized here as the amount of (dis)agreement in a panel.

To capture diversity and dependence, we construct four dependence matrices for each panel. The *density@realization* matrix assigns each (expert, variable) the (interpolated) percentile of the expert's probability distribution realized by the true value. We can compute correlations variable–wise or expert–wise. It emerges that the mean correlation expert–wise is *0.39* whereas for variables it is *0.06* (see appendix Table A1). Because of this, *40%* of the total variance is due to the variables, and *8%* is due to the experts (positive correlation reduces explanatory power). The *HiLo diversity matrix* assigns the value *–1* to an (expert, variable) if the expert's point forecast (median) is above the realization, and assigns *1* otherwise. Total agreement (minimal diversity) entails that expert medians are either all above or all below the realizations. The *Tail diversity matrix* assigns *–1* if the realization falls outside the expert's *90%* confidence band, and *1* otherwise. Total agreement (minimal diversity) means that all experts' confidence bands catch or all fail to catch the realizations. The net agreement (agreements – disagreements) for *{–1,1}* matrices can be computed from the covariance matrices (see appendix with illustrative calculation in Table A2). To compare across studies, the net agreement is divided by the numbers of experts and variables. The *MAPE matrix* assigns the absolute percentage error to each (expert, variable).

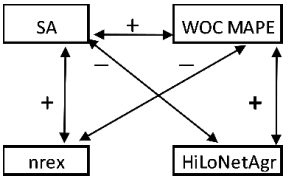
The following picture emerges: regarding the *density@realization*, the experts display a moderate, not extreme, tendency to cluster. The variance decompositions of *HiLo* and *density@realization* are quite similar. *MAPE* differs in that both experts and variables have less explanatory power. The *Tail matrix* reverses this relation between experts and variables. This is a strong signal in the data, see appendix Fig A1. The experts cluster moderately regarding the placement of medians but not regarding the uncertainty bands. Nevertheless, the assumption of no dependence is statistically rejected for most experts (see Table A3). Table 1 gives rank correlations between the two performance metrics with other study covariates. Statistical significance is based on the Student T approximation to the distribution of rank correlation (Kendall et al 1939). Significance level 0.05 is 'significant', those in (0.05, 0.2] are labelled 'indicative', others are "too weak". Table 1 also indicates whether increasing values of the covariate is helpful(+), harmful(–) or too

weak(?) for each performance metric. Negative correlations are helpful for MAPE, positive correlations are helpful for SA.

The negative rank correlation between *HiLoNetAgr* and mean *SA* is significant: more DISagreement corresponds with higher *SA* in line with the diversity theme that agreement is harmful. However, its negative correlation with *WOC* panel *MAPE* argues that more agreement is helpful. The number of experts in a panel positively correlates with *WOC* panel *MAPE*; more experts tends to raise the mean absolute percentage error (harmful). This reflects the fat tails in the MAPE distribution: as forecast errors are averaged over more experts, the average error gets larger. The positive correlation with Mean *SA* means that adding experts to the panel tends to raise the mean *SA* of experts (helpful). Higher *SA* is helpful for *WOC* Panel *MAPE*.

Table 1: 40 studies, rank correlations of covariates with performance metrics(left), diagram of significant or indicative rank correlations (right).'+' means helpful, '–' means harmful. The rank correlation between nrex and HiLoNetAgr is -0.2 (not shown).

40 studies; Spearman correlation with							
	Mean SA	AveCorVbl	AveCorExp	Tai Net Agr	HiLo Net Agr	nrex	nrv/o
WOC Panel MAPE	-0.20	0.08	-0.10	-0.05	-0.15	0.19	0.04
significant at	*0.11*	*>0.27*	*>0.27*	*>0.27*	*0.18*	*0.12*	*>0.27*
Mean SA		0.18	-0.07	-0.07	-0.26	0.17	0.09
significant at		*0.13*	*>0.27*	*>0.27*	*0.05*	*0.15*	*>0.27*
Qualitative relations: helpful (+), harmful (-) too weak (?)							
	Mean SA	AveCorVbl	AveCorExp	Tai Net Agr	HiLo Net Agr	nrex	nrv/o
WOC Panel MAPE	+	?	?	?	+	–	–
Mean SA		+	?	?	–	+	?



There are no strong relationships in Table 1. However, if we focus on TailNetAgr then a sharp relationship emerges. TailNetAgr correlates strongly positive with SA in some studies and strongly negative in others. As shown in Figure 6, this is driven by the mean statistical accuracy of all experts in a panel. The rank correlation in Figure 6 is 0.68.

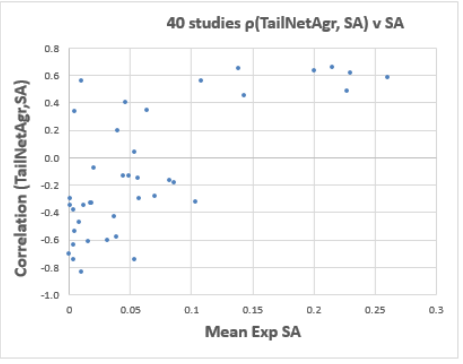


Figure 6: Correlation of Tail Net Agreement and statistical accuracy against mean expert statistical accuracy.

If we project all 40 points on the vertical axis of Figure 4, their average rank correlation with *SA* is –0.08. However, plotting these values against the Mean *SA* reveals the following: if the panel on the whole is statistically accurate, then high Tail agreement corresponds to high expert *SA*, otherwise high Tail agreement corresponds to low *SA*. In simplistic consensus terms, if the consensus is right then high Tail agreement predicts good *SA*, if the consensus is wrong then low Tail agreement predicts good *SA*. For other forms of agreement the signal is similar but weaker, also for absolute percentage error. On reflection this result is hardly surprising, but probably not the result those claiming scientific consensus confers credibility hoped to hear.

6. Conclusions

Both expert *MAPE*s and *WOC MAPE*s appear to be very fat tailed, raising doubt about *WOC MAPE* convergence. *WOC MAPE* is minimally better than picking a random expert. Crowd size and diversity (*HiLoNetAgr*) work in opposite directions on the two performance metric and also opposite to each other. Dependence is not simply good or bad, its complicated.

(Cooke et al 2021) compared aggregation schemes in *49* post 2006 studies including *22* smaller panels. For the these *49* panels, the *WOC MAPE* was *1,472.3*. For the *40* panels studied here *WOC MAPE* was *23,220.86*, underscoring the DIS advantage of larger panels. Instead of

averaging the experts' medians per panel, if we choose the expert with the best *SA*, the *MAPE* over the 40 panels would be 6.3 (appendix table A4).

Instead of averaging medians, (Cooke et al 2021) took the medians of combined expert distributions. Equally weighted combinations yielded a *MAPE* of 3.8 and performance weighted combinations yielded 2.2. Method of aggregation is perhaps the most important contributor to wisdom of crowds and adding uncertainty quantification to point forecasts enables better aggregation methods. Probabilistic crowds are wiser than point forecast crowds. In short, if you want better forecasts look for better experts not bigger crowds.

Appendix A

Table A1 Studies & important covariates; Expert *MAPE* is the average of experts' *MAPE*s, each expert's *MAPE* is the average absolute percentage error over all variables. *WOC MAPE* is the average *MAPE* of the average of experts' medians for each variable. *WOC MAPE* is less or equal to Expert *MAPE* by Jensen's inequality. Net agreement is per expert variable. *IndEx'dTailNetAgr* is the expected net *Tail agreement* if the experts' 1s and –1s were randomly distributed; if the experts were all statistically accurate the probability of *1* would be *0.9* and the expected agreement would be *0.64*. For *HiLo* the probability of *1* would be *0.5* and the independent expected agreement would be *0*.

STUDY	Mean SA	AveCorVbl	AveCorExp	Tai Net Agr	HiLo Net Agr	IndEx'dTailNetAgr	IndEx'dHiLoNetAgr	nrexp	nrvb	expert MAPE	WOCMAPE
biol_agents	0.06	0.00	0.43	0.14	0.29	0.00	0.00	12	12	5.34	5.01
Brexit_Food	0.01	0.13	0.08	0.11	0.11	0.09	0.09	10	10	1.47	0.83
burkina faso	0.00	0.05	0.33	0.22	0.24	0.22	0.05	12	10	310.07	309.65
CDC_ROI	0.11	0.09	0.35	0.06	0.23	0.00	0.02	20	10	3.43	3.05
CDCall	0.14	0.03	0.55	0.23	0.38	0.06	0.01	48	14	6.89	6.71
COZem	0.14	-0.07	0.65	0.17	0.40	0.01	0.02	10	11	0.90	0.82
cotopaxi	0.23	0.00	0.23	0.15	0.20	0.12	0.01	20	14	1.66	1.36
eBBP	0.20	0.08	0.32	0.20	0.17	0.09	0.01	14	15	0.70	0.52
Erie Carps	0.23	0.06	0.49	0.30	0.16	0.15	0.01	10	15	1.22	1.05
Gerstenberger	0.06	-0.01	0.39	0.14	0.20	0.06	0.00	12	13	1.40	1.02
ICE_2012	0.09	0.05	0.37	0.08	0.30	0.00	0.01	10	11	1.03	0.82
ICE_2018	0.07	-0.03	0.55	0.12	0.41	0.02	0.01	20	16	1.54	1.04
Leontaris	0.05	0.30	0.60	0.08	0.37	0.01	0.12	11	14	3.32	3.06
liander	0.00	-0.02	0.28	0.20	0.35	0.17	0.13	11	10	1.54	1.35
PHAC_T4	0.01	0.09	0.47	0.21	0.33	0.02	0.00	10	12	56.85	56.57
piton	0.04	-0.02	0.15	0.04	0.18	0.00	0.00	28	15	1.04	0.74
Political_Violence	0.00	0.34	0.44	0.27	0.30	0.25	0.03	15	21	7.74	7.35
Sheep	0.06	-0.03	0.40	0.06	0.26	0.00	0.01	14	15	3.39	3.09
SPEED	0.02	0.00	0.42	0.10	0.37	0.02	0.08	14	16	2.14	1.81
Tadini_Clermont	0.21	0.02	0.23	0.18	0.23	0.18	0.02	12	13	2.13	1.94
TdC	0.10	0.02	0.43	0.00	0.29	0.00	0.01	18	17	216.32	215.98
topaz	0.03	0.02	0.32	0.12	0.30	0.01	0.00	21	16	49778.26	49777.81
usgs	0.00	0.00	0.34	0.10	0.28	0.02	0.04	32	18	15.55	15.14
ethiopia	0.00	0.01	0.39	0.23	0.24	0.19	0.03	11	10	37.88	37.58
stromboli	0.04	-0.02	0.27	0.08	0.22	0.00	0.00	21	16	3.77	3.47
RIVM2023	0.08	0.11	0.52	0.16	0.34	0.00	0.00	42	15	183.60	183.26
Kulumbo	0.02	0.02	0.42	0.17	0.25	0.02	0.01	13	19	3.67	3.31
TUDDISPR	0.00	0.22	0.67	0.10	0.40	0.04	0.00	11	36	0.57	0.45
Gas_3rd_party	0.00	0.17	0.51	0.18	0.43	0.09	0.21	14	17	5.56	5.33
GAS95(corros)	0.05	0.25	0.51	0.21	0.39	0.05	0.21	13	11	5.75	5.50
Dykering	0.05	0.15	0.89	0.37	0.80	0.06	0.01	17	47	0.50	0.48
CARMA	0.26	0.15	0.41	0.06	0.24	0.03	0.02	12	10	248.88	248.26
OPRISKBANK	0.02	0.06	0.26	0.02	0.14	0.02	0.00	10	16	332.02	331.55
INFOSEC	0.05	0.05	0.52	0.13	0.35	0.05	0.00	13	10	12.24	11.82
DAMS	0.00	-0.04	0.35	0.20	0.25	0.05	0.01	11	11	6.56	6.25
PILOTS	0.04	0.02	0.39	0.13	0.31	0.01	0.02	31	10	0.51	0.29
SETE_CIDADES	0.02	0.04	0.25	0.19	0.16	0.08	0.01	19	10	877202.92	877202.18
TEIDEMAY_05	0.01	0.02	0.13	0.06	0.19	0.04	0.12	17	10	1.49	0.79
VESUVIO	0.01	-0.01	0.23	-0.01	0.17	0.01	0.03	14	10	62.36	61.80
Volcano_risk	0.04	0.00	0.22	0.05	0.19	0.04	0.03	45	10	316.14	315.45
AVERAGE	**0.06**	**0.06**	**0.39**	**0.14**	**0.29**	**0.06**	**0.04**	**17.45**	**14.65**	**23221.21**	**23220.86**

Application of Jensen's inequality: $|\ [(\Sigma_{i=1..n}\, fi/n) - r]/r\ | = (1/|r|)\ |(1/n)[(\Sigma_{i=1..n}\, f_i) - nr]| = (1/|r|)\ |\ (1/n)\ \Sigma_{i=1..n}\ (f_i - r)\ | \leq (1/(|r|n)\ )\ \Sigma_{i=1..n}\ |(fi - r)| = (1/n)\ \Sigma_{i=1..n}\ |(fi - r)/r|$.

Matrix M = $M(ex \times vb)$ of $\{-1,1\}$; nx = number of experts, nv = number of variables. Net Agreement for expert $x_1 = NA(1) =$ *#Agreements−#Disagreements* for x_1. $x_1 := x \bullet l = \Sigma_{i-1\ldots mv}\ x_i$. *#agreements* – *#disagreements* for $x_1 = NA(1) = \Sigma_{i>1}\ x_1 \bullet x_i$. $C_1 := \Sigma_{i=1\ldots nx}\ Cov(x_1, x_i)$.

Lemma: $NA(1) = nv\ C_1 + (x_1/nv)(\Sigma_{i \geq 1}\ x_i) - nv$.

Pf: $C_1 = Cov(x_1, \Sigma_{i=1\ldots nx}\ x_i) = Cov(x_1, \Sigma_{i>1}\ x_i) + VAR(x_1) = \Sigma_{i>1}\ x_1 \bullet x_i\ /nv - E(x_1)E(\Sigma_{i>1}\ x_i) + VAR(x_1)$. $E(x_1^2) = 1$; $\Sigma_{i>1}\ x_1 \bullet x_i = nv\ C_1 + nv\ E(x_1)E(\Sigma_{i>1}\ x_i) - nv\ [E(x_1^2) - (E(x_1))^2] = nv\ C_1 + nv\ E(x_1)\ \Sigma_{i \geq 1}\ E(x_i) - nv = nv\ C_1 + (x_1)\ (\Sigma_{i>1}\ x_i)\ /nv - nv$. □

Remark 1: The net agreement for expert *1* per variable is *NA(1)/nv*. The total net agreement for matrix *M* per *(expert,variable)* is *NA(Mnxnv) = [$\Sigma_{i=1..nx}$ NA(i)]/(nx×nv).*

Remark 2: Setting *q=1 – p*, the expected Net Agreement of independent *{ 1,−1}* variables with *P(1) = p is*

$p^2+q^2-2pq = (p-q)^2 = (2p-1)^2$.

With $(p+q)^2 = 1$, the variance of Net Agreement is

$VAR = 1-(2p-1)^4 = \ldots. = 2pq(1-4pq)$.

The Null Hypothesis is that for all cells in M the probability of 1 is constant $\in (0, 1)$ and independent for each cell. Hence, *M* consists of *nx×nv* Bernoulli variables with probability p of 'success'. For expert 1, the agreement / disagreement with each other expert for each variable constitute *(nx–1)×nv* independent *{–1, 1}* variables, each with mean $(2p–1)^2$ and variance *2pq(1–4pq).* The sum of such variables is approximately normal with mean $nv×(nx-1)×(2p-1)^2$ and standard deviation $[nv×(nx-1)×2pq×(1-4pq)]^{1/2}$.

Graphs for Variance Decomposition

Let *e* denote experts and *v* denote variables. *E(e|v)* denotes conditional expectation of *e* given *v* and *V(e|v)* denotes conditional variance of e given v. The Law of Total Variance states that the overall variance *V* satisfies

$V = V(E(v|e)) + E(V(v|e)) = V(E(e|v)) + E(V(e|v))$.

V(E(v|e)/V is the fraction of *V* explained by variation over experts, *V(E(e|v))/V* is the fraction of *V* explained by variation over variables. The variance decomposition for the matrices dens@rls, HiLo and MAPE are similar and indicate moderate clustering of experts. For Tails the pattern is reversed, more variance is explained by experts indicating clustering of variables.

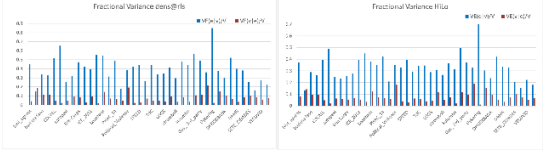


Table A3 Net agreement for each of the 10 experts in CO2m.

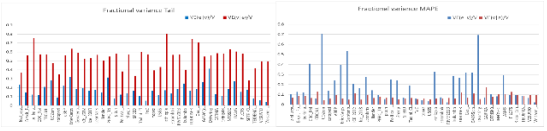


Fig. A1 Fractional variance for variables (blue) and experts (red)

For HiLo and Tail net agreement, we compare the experts' net agreement per expert–variable with the expected net agreement if the *1*'s and *–1*'s were distributed independently over the matrix with the observed frequency of occurrence. For HiLo the net agreement is much lager than expected if the distribution of *{1, – 1}* were independent. For Tail the difference is smaller, though not uniformly.

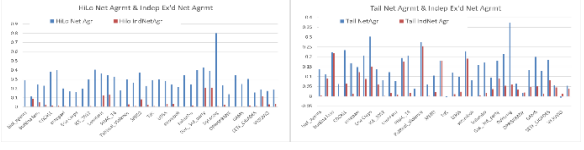


Fig. A2 Net agreement per expert variable compared with net agreement of independent *{1, –1}* to the matrix cells with equal probabilities of *{1,–1}*, for HiLo (left) and Tail (right).

Table A2 shows the HiLo matrix for CO2em (Rennert et al 2022) (left) and the corresponding expert-wise covariance matrix (right). The calculation of Net Agreement for each expert is illustrated.

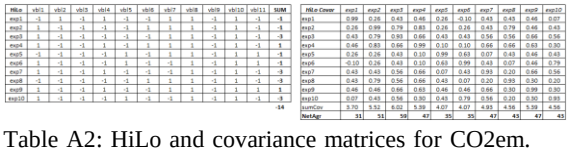


Table A2: HiLo and covariance matrices for CO2em. The net agreement per expert is computed as *nrvb*sumCov +expsum HiLo*sumHiLoMatrix / nrvb – nrvb*. For expert 1 this is *11× 3.702479339+(–1)×(–14)/11 – 11=31*.

Table A3 shows the net agreement for each of the 10 experts in CO2em. For HiLo two experts' net agreement falls within the 95% central range of the distribution under the null hypothesis, for Tail net agreement, only one.

CO2em, 10 experts, 11 variables						
expert	HiLo NetAgr	*ExpSumNA Hilo*	Normal CDF	*Tail NetAgr*	*ExpSumNA Tail*	Normal CDF
1	0.28	-1	4.82E-04	0.34	3	9.99E-01
2	0.46	-1	4.82E-04	0.11	-5	1.28E-26
3	0.54	-3	2.67E-09	0.00	-7	3.86E-45
4	0.43	1	2.22E-01	-0.07	-3	3.54E-13
5	0.32	-1	4.82E-04	0.01	-1	9.13E-05
6	0.32	-1	4.82E-04	0.34	5	1.00E+00
7	0.43	-3	2.67E-09	0.16	1	3.80E-01
8	0.39	-3	2.67E-09	0.31	5	1.00E+00
9	0.43	1	2.22E-01	0.34	5	1.00E+00
10	0.39	-3	2.67E-09	0.16	9	1.00E+00

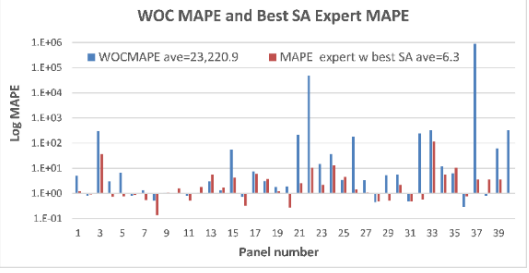


Figure A3: WOC MAPE over 40 panels and MAPE of expert with best SA per panel. MAPE is on log scale.

Mackay, Charles (1841) *Memoirs Of Extraordinary Popular Delusions and The Madness Of Crowds.* Vol. I (1 ed.). London: Richard Bentley.

Surowieki's, James, (2019) *The Wisdom of Crowds* (2005), Anchor; Reprint edition (August 16, 2005)

Murray, Douglas (2019) *The Madness of Crowds, gender, race and identity* (2019),

Planck, Max K. (1950). *Scientific Autobiography and Other Papers.* New York: Philosophical library.

Oreskes, Naomi (2019) *Why Trust Science*, Princeton University Press

Cooke, R.M., Nieboer, D. Misiewicz, J. (2014) *Fat Tailed distributions, data, diagnostics and dependence*, Wiley, London.

Cooke, Roger M., Marti, Deniz and Mazzuchi, Thomas A., (2021) Expert Forecasting with and without Uncertainty Quantification and Weighting: What Do the Data Say? *International Journal of Forecasting*, published online July 25, 2020,

Kendall, M. G., Kendall Sheila F. H. and Babington Smith B, (1939) The Distribution of Spearman's Coefficient of Rank Correlation in a Universe in which all Rankings Occur an Equal Number of Times: *Biometrika* , Jan., 1939, Vol. 30, No. 3/4 (Jan., 1939), pp. 251–273 Published by: *Oxford University Press* on behalf of Biometrika Trust.

Rennert, Kevin, Frank Errickson, Brian C. Prest, Lisa Rennels, Richard G. Newell, William Pizer, Cora Kingdon, Jordan Wingenroth, Roger Cooke, Bryan Parthum, David Smith, Kevin Cromar, Delavane Diaz, Frances Moore, Ulrich K. Müller, Richard Plevin, Adrian E. Raftery, Hana Ševčíková, Hannah Sheets, James H. Stock, Tammy Tan, Mark Watson, Tony Wong, David Anthoff (2022) Comprehensive Evidence Implies a Higher Social Cost of CO2. *Nature*.

Aspinall, W.P. (2010) "A route to more tractable expert advice" *Nature*, vol. 463, 21 January, 2010.

Proceedings of the 35th European Safety and Reliability and the 33th Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage & Marja Ylönen

ISBN: 981-973-0000-00-0 :: doi: 10.3850/981-973-0000-00-0 esrel2022-paper