Supplementary Online Material for Cross Validation of Classical Model for Structured Expert Judgment

Roger M. Cooke
Resources for the Future
Delft Univ. of Technology Dept Mathematics
November 18, 2016

Table of Contents
1. Post-2006 Data and Applications Documentation ................................................................. 2
2. Classical Model Performance Measures and Combination ....................................................... 4
   2.1 Statistical accuracy .................................................................................................................. 4
   2.2 Information ............................................................................................................................. 5
   2.3 Combination: Decision Maker .............................................................................................. 6
3. Strictly Proper Scoring Rules as Weights .................................................................................. 11
   4.1 Introduction ............................................................................................................................ 11
   4.2 Scoring rules for individual variables .................................................................................... 11
   4.3 Scoring Rules for Average Probabilities / Expected Frequencies ........................................ 12
   4.3.1 Characterization of Strictly Proper Scoring Rules for Average Probabilities ................... 12
   4.4 Asymptotic Properties .......................................................................................................... 16
   4.5 Weights ................................................................................................................................. 18
   Discrepancies in Past Cross Validation Studies ......................................................................... 19
6. Additional Classical Model Applications .................................................................................. 22
   6.1 Nuclear reports Published as a result of the joint ec/usnrc project on uncertainty analysis of probabilistic accident consequence codes (under the Third EC-Framework Programme) ........................................... 22
   6.2 Probabilistic accident consequence uncertainty analysis Reports published on the project uncertainty analysis of the probabilistic accident consequence code cosyma using expert judgement (under the fourth EC- framework programme) ........................................................................................................ 23
   6.3 Ecosystems and public health .............................................................................................. 23
   6.4 Civil aviation and structural reliability ................................................................................ 26
   6.5 Information security .............................................................................................................. 28
   6.6 Natural hazards .................................................................................................................... 28
   6.7 Climate ................................................................................................................................ 30
   6.8 Banking and finance ............................................................................................................. 30
   6.9 Completed, publication in preparation ................................................................................. 30

References ...................................................................................................................................... 32
1. Post-2006 Data and Applications Documentation

Expert judgment materials, including data from expert judgment studies, are available at [http://rogerm Cooke.net/](http://rogerm Cooke.net/). The studies can be read by the expert judgment software EXCALIBUR (Cooke and Solomatine 1992), which is a free downloadable at [http://www.lighttwist.net/wp/](http://www.lighttwist.net/wp/). Summary information is presented in the following table. Full references are available at the end of the Supplementary Online Material.

<table>
<thead>
<tr>
<th>Contracting Party</th>
<th>Performed by</th>
<th>Study name</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robert Wood Johnson Foundation</td>
<td>Center for Disease Dynamics, Economics &amp; Policy</td>
<td>Arkansas</td>
<td>Grant effectiveness and child health insurance enrollment (Laxminarayan et al. 2012)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Florida</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Illinois</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nebraska</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Washington</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CoveringKids</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tobacco</td>
<td>Grant effectiveness and tobacco control (Colson, Cooke, and Laxminarayan 2016)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Obesity</td>
<td>Grant effectiveness, childhood obesity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fistula</td>
<td>Effectiveness of obstetric fistula repair (Colson, Adhikari, et al. 2015)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>San_Diego</td>
<td>Effectiveness of surgical procedures</td>
</tr>
<tr>
<td>Disease Control Priority Project, 3rd Edition</td>
<td>University of Notre Dame</td>
<td>CDC_ROI</td>
<td>Return on investment in public health tracking (Colson, Cohen, et al. 2015)</td>
</tr>
<tr>
<td>Center for Disease Control and Prevention</td>
<td>Vicki Bier</td>
<td>Create</td>
<td>Terrorism</td>
</tr>
<tr>
<td>University of Wisconsin, CREATE</td>
<td>University of Notre Dame</td>
<td>Erie_Carp</td>
<td>Establishment of Asian Carp in Lake Erie (Cooke et al. 2014; Wittmann et al. 2015; Wittmann et al. 2014; Zhang et al. 2016)</td>
</tr>
<tr>
<td>National Oceanic and Atmospheric Administration and Environmental Protection Agency (EPA)</td>
<td>University of Maryland</td>
<td>GL_NIS</td>
<td>Costs of invasive species in Great Lakes (Rothlisberger et al. 2010; Rothlisberger et al. 2012)</td>
</tr>
<tr>
<td>EPA and University of Maryland</td>
<td>University Medical Center Utrecht</td>
<td>UMD_NREMOV AL</td>
<td>Nitrogen removal in Chesapeake Bay watershed (Koch, Febria, et al. 2015; Walsh 2015; Koch, Filoso, et al. 2015)</td>
</tr>
<tr>
<td>University Medical Center Utrecht</td>
<td>TU Delft</td>
<td>Hemophilia</td>
<td>Optimal treatment of patients with severe haemophilia (Fischer, Lewandowski, and Janssen 2013)</td>
</tr>
<tr>
<td>National Institute for Public Health and the Environment</td>
<td>University of Cambridge</td>
<td>ATCEP</td>
<td>Air traffic Controllers Human Error</td>
</tr>
<tr>
<td>Brand Preventie</td>
<td></td>
<td>FCEP</td>
<td>Flight Crew Human Error</td>
</tr>
<tr>
<td>Liander</td>
<td></td>
<td>Daniela</td>
<td>Fire prevention and control (Hanea 2009)</td>
</tr>
<tr>
<td>University of Cambridge</td>
<td></td>
<td>Liander</td>
<td>Underground cast iron gas-lines (Forys, Kurowicka, and Peppelman 2013)</td>
</tr>
<tr>
<td>Health Protection England</td>
<td>Willy Aspinall</td>
<td>Arsenic</td>
<td>Air quality levels for arsenic</td>
</tr>
<tr>
<td>University of Ottawa</td>
<td></td>
<td>Biol_A gent</td>
<td>Human dose-response curves for bioterror agents</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CWD</td>
<td>Infection transmission risks: chronic wasting disease (cwd) from deer to</td>
</tr>
<tr>
<td><strong>PrioNet</strong></td>
<td>humans (Tyshenko et al. 2011; Tyshenko, ElSaadany, Oraby, Darshan, et al. 2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>eBPP</strong></td>
<td>Xenotropic murine leukemia virus-related virus (XMRV) blood/tissue infection transmission risks (Tyshenko, ElSaadany, Oraby, Laderoute, et al. 2012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Eff Erupt</strong></td>
<td>Icelandic fissure eruptions: source characterization (Loughlin et al. 2013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>IceSheets</strong></td>
<td>Contribution to sea level rise from ice sheets melting due to global warming (Bamber and Aspinall 2013; Bamber, Aspinall, and Cooke 2016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PHAC_T4</strong></td>
<td>Additional CWD factors</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sheep</strong></td>
<td>Risk management policy for sheep scab control</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SPEED</strong></td>
<td>Volcano hazards at Vesuvius &amp; Campi Flegrei, Italy (Bevilacqua et al. 2015; Neri et al. 2015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TdC</strong></td>
<td>Volcano hazards (Tristan da Cunha) (Hicks et al. 2014)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>TOPAZ</strong></td>
<td>Tectonic hazards for radwaste siting in Japan (Scourse, Aspinall, and Chapman 2015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Natural Hazards Research Platform/GNS Science</strong></td>
<td>Canterbury Seismic Hazard Model (Gerstenberger et al. 2014; Gerstenberger et al. 2013; Christophersen, Nicol, and Gerstenberger 2011)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Embry-Riddle</strong></td>
<td>Airports safety</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Details from 33 post-2006 applications of the Classical Model.
2. Classical Model Performance Measures and Combination

Similar expositions on the Classical Model can be found in the online supplementary material of other publications, including (Cooke et al. 2014; Wittmann et al. 2014; Cooke 2015; Koch, Febria, et al. 2015; Wittmann et al. 2015).

There are two generic, quantitative measures of expert performance, *calibration* or *statistical accuracy and information*. Loosely, statistical accuracy measures the statistical likelihood that a set of experimental results correspond, in a statistical sense, with an expert’s assessments. More precisely, it is the p-value at which we would falsely reject the hypothesis that an expert's probability statements were accurate. Suppose the 5-, 50- and 95-percentiles, or quantiles, were elicited from each expert for each of N continuous variables. Each expert effectively divides the range of possible outcomes of each variable into four intervals: less than or equal to the 5 percent value, greater than the 5 percent value and less than or equal to the 50 percent value, etc. The probabilities for these intervals are expressed as a vector

\[ p = (p_1, p_2, p_3, p_4) = (0.05, 0.45, 0.45, 0.05). \]

2.1 Statistical accuracy

If N quantities are assessed, each expert may be regarded as a statistical hypothesis, namely that each realization falls in one of the four inter-quantile intervals with probability vector \( p \). Suppose we have realizations \( x_1, \ldots x_N \) of these quantities. We may then form the sample distribution of the expert's inter quantile intervals as:

- \( s_1(e) = \# \{ i \mid x_i \leq 5\% \text{ quantile} \} \)/N
- \( s_2(e) = \# \{ i \mid 5\% \text{ quantile} < x_i \leq 50\% \text{ quantile} \} \)/N
- \( s_3(e) = \# \{ i \mid 50\% \text{ quantile} < x_i \leq 95\% \text{ quantile} \} \)/N
- \( s_4(e) = \# \{ i \mid 95\% \text{ quantile} < x_i \} \)/N

\( s(e) = (s_1, \ldots s_4) \)

Note that the sample distribution depends on the expert \( e \). If the realizations are indeed drawn independently from a distribution with quantiles as stated by the expert then the quantity

\[ 2N I(s(e) \mid p) = 2N \sum_{i=1,4} s_i \ln(s_i / p_i) \]  

is asymptotically distributed as a chi-square variable with three degrees of freedom. This is the likelihood ratio statistic, and \( I(s \mid p) \) is the relative information of distribution \( s \) with respect to \( p \). Extracting the leading term of the logarithm yields the familiar chi-square test statistic for goodness of fit. There are advantages in using the form in (1) (Cooke 1991, and SOM 4).

If after a few realizations the expert were to see that all realizations fell outside his/her 90 percent central confidence intervals, \( s(e) \) might conclude that these intervals were too narrow and might broaden them on subsequent assessments. This means that for this expert the uncertainty distributions are not independent, and \( s(e) \) learns from the realizations. Expert learning is not a goal of an expert judgment study. Rather, the problem owner wants experts who do not need to
learn from the elicitation. Independence is not an assumption about the expert's distribution but a desideratum of the problem owner. Hence the decision maker (see below) scores expert $e$ as the statistical likelihood of the hypothesis

$$H_e: \ "the \ inter \ quantile \ interval \ containing \ the \ true \ value \ for \ each \ variable \ is \ drawn \ independently \ from \ probability \ vector \ p."$$

A simple test for this hypothesis uses the test statistic (1), and the likelihood, or $p$-value, or calibration score of this hypothesis, is:

$$Cal(e) = p\text{-value}(e) = \text{Prob}\{2NI(s(e) \mid p) \geq r \mid H_e\}$$

where $r$ is the value of (1) based on the observed values $x_1, \ldots, x_N$. It is the probability under hypothesis $H_e$ that a deviation at least as great as $r$ should be observed on $N$ realizations if $H_e$ were true. Calibration scores are absolute and can be compared across studies. However, it is appropriate to equalize the power of the different hypothesis tests by equalizing the effective number of realizations. To compare scores on two data sets with $N$ and $N'$ realizations, we use the minimum of $N$ and $N'$ in (1), without changing the sample distribution $s$.

Although the calibration score uses the language of simple hypothesis testing, it must be emphasized that we are not rejecting expert-hypotheses; rather we are using this language to measure the degree to which the data supports the hypothesis that the expert's probabilities are accurate. Low scores, near zero, mean that it is unlikely that the expert’s probabilities are correct. High scores, near 1, indicate good support.

### 2.2 Information

The second scoring variable is information. Loosely, the information in a distribution is the degree to which the distribution is concentrated. Information cannot be measured absolutely, but only with respect to a background measure. Being concentrated or "spread out" is measured relative to some other distribution.

Measuring information requires associating a density with each assessment of each expert. To do this, we use the unique density that complies with the experts' quantiles and is minimally informative with respect to the background measure. This density can easily be found with the method of Lagrange multipliers. For a uniform background measure, the density is constant between the assessed quantiles. The background measure is not elicited from experts as indeed it must be the same for all experts; instead it is chosen by the analyst.

The uniform and log-uniform background measures require an intrinsic range on which these measures are concentrated. The classical model implements the so-called “$k\%$ overshoot rule”:

for each item with uniform background we consider the smallest interval $I = [L, U]$ containing all the assessed quantiles of all experts and the realization, if known. This interval is extended to

$$I^* = [L^*, U^*]; L^* = L - k(U-L)/100; U^* = U + k(U-L)/100.$$
The value of $k$ is chosen by the analyst. For a log uniform background, the same procedure is applied to logged values. A large value of $k$ tends to make all experts look quite informative, and tends to suppress the relative differences in information scores. The information score of expert $e$ on assessments for uncertain quantities 1…N is

$$\text{Inf} (e) = \text{Average Relative information w.r.t. Background} = \frac{1}{N} \sum_{i=1}^{N} I(f_{e,i} | g_i)$$

where $g_i$ is the background density for variable $i$ and $f_{e,i}$ is expert $e$'s density for item $i$. This is proportional to the relative information of the expert's joint distribution given the background, under the assumption that the variables are independent. As with calibration, the assumption of independence here reflects a desideratum of the decision maker and not an elicited feature of the expert's joint distribution. The information score does not depend on the realizations. An expert can give her/himself a high information score by choosing quantiles very close together. The information score of $e$ depends on the intrinsic range which depends on the assessments of the other experts. Hence, information scores cannot be exactly compared across studies.

The above information score is chosen because it is familiar, tail insensitive, scale invariant and "slow." The latter property means that relative information is a slow function; large changes in the expert assessments produce only modest changes in the information score. This contrasts with the likelihood function in the calibration score, which is a very "fast" function. This causes the normalized product of calibration and information to be driven by the calibration score.

### 2.3 Combination: Decision Maker

The combined score of expert $e$ will serve as an unnormalized weight for $e$:

$$w_d(e) = \text{Cal} (e) \times \text{Inf} (e) \times \mathbb{1}_{d}(\text{Cal}(e) \geq \alpha),$$  \hspace{1cm} (2)

where $\mathbb{1}_{d}(\text{Cal}(e) \geq \alpha) = 1$ if $\text{Cal}(e) \geq \alpha$, and is zero otherwise. The combined score thus depends on $\alpha$; if $\text{Cal}(e)$ falls below cut-off level $\alpha$, expert $e$ is unweighted. The presence of a cut-off level is imposed by the requirement that the combined score be an asymptotically strictly proper scoring rule. That is, an expert maximizes his/her long run expected score by and only by ensuring that his probabilities $p = (0.05, 0.45, 0.45, 0.05)$ correspond to his true beliefs (Cooke, 1991, section 6). $\alpha$ is similar to a significance level in simple hypothesis testing, but its origin is to measure “goodness” and not to reject hypotheses.

A combination of expert assessments is called a "decision maker" (DM). All decision makers discussed here are examples of linear pooling; the classical model is essentially a method for deriving weights in a linear pool. "Good expertise" corresponds to good calibration (high statistical likelihood, high p-value) and high information. Weights that reward good expertise and pass these virtues on to the decision maker are desired.

The reward aspect of weights is very important. We could simply solve the following optimization problem: find a set of weights such that the linear pool under these weights maximizes the product of calibration and information. Solving this problem on real data, one finds that the weights do not generally reflect the performance of the individual experts. As an expert's influence on the decision maker should not appear haphazard, and "gaming" the system
with assessments tilted to achieve a desired outcome should be discouraged, we must impose a strictly scoring rule constraint on the weighting scheme.

The scoring rule constraint requires the term \( \mathbb{I}_\alpha (\text{Cal}(e) \geq \alpha) \) in eq (2), but does not indicate what value of \( \alpha \) we should choose. Therefore, we choose \( \alpha \) to maximize the combined score of the resulting decision maker. Let \( DM_\alpha(i) \) be the result of linear pooling for any item \( i \) with weights proportional to (2):

\[
DM_\alpha(i) = \frac{\sum_{e=1}^E w_\alpha(e) f_{e,i}}{\sum_{e=1}^E w_\alpha(e)}
\]

(3)

The optimized global weight \( DM \) is \( DM_{\alpha^*} \) where \( \alpha^* \) maximizes

\[
\text{calibration score}(DM_{\alpha^*}) \times \text{information score}(DM_{\alpha^*}).
\]

(4)

This weight is termed global because the information score is based on all the assessed calibration variables. A variation on this scheme allows a different set of weights to be used for each item. This is accomplished by using information scores for each item rather than the average information score:

\[
w_\alpha(e,i) = \mathbb{I}_\alpha(\text{calibration score}) \times \text{calibration score}(e) \times I(f_{e,i} \mid g_i)
\]

(5)

For each \( \alpha \) we define the item weight \( IDM_\alpha \) for item \( i \) as

\[
IDM_\alpha(i) = \frac{\sum_{e=1}^E w_\alpha(e,i) f_{e,i}}{\sum_{e=1}^E w_\alpha(e,i)}
\]

(6)

The optimized item weight \( DM \) is \( IDM_{\alpha^*} \) where \( \alpha^* \) maximizes

\[
\text{calibration score}(IDM_{\alpha^*}) \times \text{information score}(IDM_{\alpha^*}).
\]

(7)

The non-optimized versions of the global and item weight DM’s are obtained by setting \( \alpha = 0 \).

The optimization in (5) and (7) often causes experts to be unweighted, even experts with good scores. Such experts are not “rejected;” unweighting simply means that their input is already captured by a smaller subset of experts. Their value to the whole study is brought out in studying the robustness of the optimal \( DM \) under loss of experts.

Item weights are potentially more attractive as they allow an expert to up- or down-weight her/himself for individual items according to how much (s)he feels (s)he knows about that item. "Knowing less" means choosing quantiles farther apart and lowering the information score for that item. Of course, good performance of item weights requires that experts successfully perform this up-down weighting. Anecdotal evidence suggests that item weights improve over global weights as the experts receive more training in probabilistic assessment. Both item and global weights can be described as optimal weights under a strictly proper scoring rule constraint. With both global and item weights, calibration strongly dominates over information, and information serves to modulate between more or less equally well calibrated experts.
Since any combination of expert distributions yields assessments for the seed variables, any combination can be evaluated on the seed variables. In particular, we can compute the calibration and the information of any proposed decision maker. We should hope that the "performance based decision maker" would perform better than the result of simple averaging, and we should also hope that the proposed DM is not worse than the best expert in the panel.
2. Mathematical Pooling: Harmonic, Geometric and Arithmetic Means

This analysis builds on the material in Bamber et al. (2016) and the main manuscript. Using the 33 professional expert judgment studies performed since 2006, it is possible to compare HW, EW and performance weighting (PW). To facilitate third party checks of the results, for this comparison PW is based on global weights, and experts who assessed less than the full set of calibration variables are excluded. This causes the PW and EW solutions used here to differ slightly from the solutions published elsewhere, but the integrity of the present comparison is not affected.

<table>
<thead>
<tr>
<th>PW</th>
<th>EW</th>
<th>HW</th>
<th>#seeds</th>
<th>#experts</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="" /></td>
<td><img src="image" alt="" /></td>
<td><img src="image" alt="" /></td>
<td><img src="image" alt="" /></td>
<td><img src="image" alt="" /></td>
</tr>
</tbody>
</table>

Table 2: Performance of PW, EW and HW. “#seeds” denotes the number of calibration variables used in each study, “#experts” denotes the number of experts who assessed all calibration variables in each study.

The performance of HW, EW and PW are compared with regard to statistical accuracy, informativeness and the combined score (the product of the former two) (Table 2). HW is the best (as determined by the combined score) in four of the thirty-three cases. HW’s informativeness is slightly higher than that of PW and substantially higher than EW. The statistical accuracy of HW is substantially below that of EW and PW. As reported in Bamber et al. (2016), in 18 cases (55 percent) the hypothesis that HW is statistically accurate would be rejected at the 5 percent level. In nine cases rejection would be at the 0.001 level.
Figure 1: Number of calibration variables and number of experts against P-values for HW, EW and PW.

These data provide evidence on how performance is affected by the number of experts and number of calibration variables. Figure 1 graphs the number of calibration variables and number of experts against the statistical accuracy scores for HW, EW, and PW. HW degrades as the number of calibration variables increases, whereas EW is unaffected, and PW actually improves. The statistical power of the measure of statistical accuracy increases with the number of calibration variables, and this would tend to suppress statistical accuracy scores of all experts and combinations alike. However, no such tendency is observed for EW or PW (the effect would presumably be observed if greater numbers of calibration variables were available). The number of experts does not have a marked effect on any of the combinations.
3. Strictly Proper Scoring Rules as Weights

This builds on the original material from (Cooke 1991).

4.1 Introduction

This chapter covers the mathematics underlying the classical model in full detail. This material was published in 1991, and the original exposition has been simplified and improved for courses given to NASA. The original numbering of theorems and propositions from (Cooke 1991) is preserved for reference. The main simplification concerns the characterization of scoring rules for average probabilities. This improved rendition is published here for the first time.

4.2 Scoring rules for individual variables

Scoring rules were introduced by L. J. Savage and B. DeFinetti for elicitation. Where \( S^n \) is the simplex on \( n \) dimensions, i.e. the set of non negative \( n \)-vectors whose components sum to one, suppose expert is asked to state \( p \in S^n \). When outcome \( i \) is known, the expert is paid \( R(p,i) \).

**Definition:** \( R(p,i) \) is strictly proper positive (negative) sensed if for all \( q \in S^n \),
\[
\arg\max_p (\arg\min_p) E_q R(p,i) = q.
\]

**Direct Rule:** \( R_d(p,i) = Kp^i \) (reward proportional to probability of observed outcome)
\[
E_q(R_d(p,i)) = \sum_i q_i R_d(p,i) = K \sum_i q_i p_i.
\]

**Lemma:** \( \arg\max_p E_q(R_d(p,i)) = (0,0...0, 1, 0...0) \), where "1" is at the position of the largest component of \( q \).

**Quadratic rule:** \( R_Q(p,i) = 2p^i - \sum_j p_j^2 \).

**Spherical rule:** \( R_S(p,i) = p^i / (\sum_j p_j^2)^{1/2} \).

**Logarithmic rule:** \( R_L(p,i) = \ln(p^i) \).

**Brier score** \( R_B(p,i) = (p - i)^2; i \in \{0,1\} \) \((= (1 - R_Q)/2)\) (used in rain forecasting)

Shuford, Albert and Massengil (1966) prove that \( R_L \) is the only rule which depends only on the probability of the observed outcome, for \( n > 2 \).

**DeGroot Feinberg decomposition**
Imagine that an expert assigns variables to different “probability bins” where the events assigned to the same bin are assessed to have the same distribution. The variables are subsequently observed and a sample distribution is found for each bin:

Let \( p = p_1, ..., p_B \in (S^n)^B \) (a vector of probability vectors);

\( s = s_1, ..., s_B \in (S^n)^B \) (vector of sample distributions)

\( n = n_1, ..., n_B \in \mathbb{N}^B \) (occupation vector, \( n_i = \# vbls assigned to bin i \)).
For \( r, t \in S^n \), write: \( E_t(R(r)) = \Sigma_{i=1..n} t_i R(r, i) \). The DeGroot, Feinberg decomposition of scoring rule \( R(p, n, s) \) summed over bins \( i = 1...B \) is (note that \( s_i, p_i \in S^n, i = 1...B \)):

\[
R(p, n, s) = \Sigma_{i=1..B} n_i [E_{si}(R(p_i)) - E_{si}(R(s_i))] + \Sigma_{i=1..B} n_i E_{si}(R(s_i)) .
\]

- **“Calibration term”**
- **“Resolution term”**

**Lemma:** For the logarithmic rule,

\[
R(p, n, s) = - \Sigma_{i=1..B} n_i [I(s_i|p_i) + H(s_i)]
\]

where \( H(p) = -\Sigma p_i \ln p_i \) is called the entropy of \( p \in S^n \).

### 4.3 Scoring Rules for Average Probabilities / Expected Frequencies

- **\( O = \{1,...m\} \): outcomes**
- **\( M(O) = \{ p \in S^n \mid p_i > 0 \} \) non-degenerate probability vectors**
- **\( X_i : \Omega \to O \): Uncertain quantity (finite valued)**
- \( X = X_1,...,X_n \)
- **\( K_i^n = \Sigma_{j=1..n} I(X_j = i) \): frequency counter for outcome \( i \) in variables \( 1...n \).**
- **\( S_i(n) = K_i^n/n \): relative frequency for outcome \( i \).**
- **\( S(n) = S_1(n),...S_m(n) \): sample distribution over \( X_1...X_n \).**
- **\( M(X) = non-degenerate measures on X \); \( P \in M(X) \Rightarrow \forall \)words \( W, P(W) > 0 \)**
- **\( Q \in M(X), q_i(n) = (1/n) \Sigma_{j=1..n} Q(X_j = i) \): Assessed distribution.**
- **\( q(n) = q_1(n)...q_m(n) \): vector of average probabilities.**

**Definition:** For \( M \subseteq M(X) \), \( R(p, s(n)) \) is a positively sensed \( M \) strictly proper scoring rule for average probabilities (SPSRAP) if

\[
\forall Q \in M, 
argmax_{p \in S^n} E_Q(R(p, s(n)) = q(n).
\]

**Note:** the quantifier is over a different (and potentially much larger) set than the argmax.

### 4.3.1 Characterization of Strictly Proper Scoring Rules for Average Probabilities

Numbered as in (Cooke 1991, chap. 9).
THEOREM 9.1 (Cooke 1991): If \( R(p, s(n)) \) is differentiable in \( p \), then the following are equivalent:

1. \( \forall Q \in M(X), \quad \forall_p E_Q(R(p, s(n)) \mid p=q(n)) = 0 \)
2. \( \forall i,j,k \in (1,..m-1) \exists \) integrable functions \( g_i, g_{ijk} \) such that

\[
\partial R/\partial p_i = g_i(p, n)(p_i - s_i(n)) + \sum_{k<j} g_{ijk}(p, n)(s_j(n)p_k - s_k(n)p_j).
\]

PROOF: We use the following lemmata:

**Lemma 9.1** Where \( F \) is a field of events, \( A_1...A_N \in F \) : \( (1/N) \sum_{j=1...N} Q(1Aj=1) = (1/N) \sum_{k=1...N} kQ(k \text{ of } A_1..A_N \text{ occur}) \) (average probability = expected relative frequency).

**Proof:** The expected relative frequency of occurrence of \( A_1...A_N \), \( (1/N) \sum_{k} kQ(k \text{ of } A_1..A_N \text{ occur}), \) may be written \( E((1/N) \sum_{j=1...N} 1A_j) = (1/N) \sum_{j=1...N} E(1A_j) = (1/N) \sum_{j=1...N} Q(I_{A_j} = 1). \)

**Lemma 9.2** \( \forall A \in M(L\times N) \) (set of \( L\times N \) matrices) with \( \text{Rank} \ (A) = L < N \); and

\[
V = \{ x \in \mathbb{R}^N \mid Ax = b: b \neq 0; x_i > 0, i = 1,...N \}
\]

If \( \dim V > 0 \), then \( \dim V = N - L + 1 \).

**Proof:** The null space of \( A \) has dimension \( N - L \), if \( V \neq \emptyset \) then there is \( x \in V \) which is not in the null space of \( A \). The difference of any two vectors in \( V \) is in the null space of \( A \). Hence, \( \dim V = N - L + 1 \).

**Proof strategy:**

Statement (1) says that for all \( i \), the following two vectors are orthogonal:

\[
\partial_i (R(p, k^{N,1})) \ldots \partial_i (R(p, k^{N,a})) \perp Q(k^{N,1}) \ldots \ Q(k^{N,a});
\]

where \( k^{N,1} \ldots k^{N,a} \) is the set of all possible frequency vectors on \( X_1...X_N \). We will count the dimensions of the subspaces generated by these vectors, and show that scoring rules satisfying statement (2) of the theorem span the subspace of vectors orthogonal to \( Q(k^{N,1})...Q(k^{N,a}) \). We simplify this notation in the following definition:

Define \( W = \{ k \in \mathbb{N}^{m-1} \mid k_i \geq 0; \ i=1...m-1, \ \sum_{i=1...m-1} k_i \leq N \} \) (set of frequency vectors over outcomes \( 1...m \) \( \not\equiv 1 \))

Use \( k \in W \) to index \( \mathbb{R}^{|W|} \): \( k \in W, \ Q_k = Q(k) \) (assessed prob. of frequency vector \( k \in W \))

\( Q = (Q_1,...,Q_{|W|}) \in \mathbb{R}^{|W|}. \)

\( A(p) = \text{subspace generated by } \{ Q \in \mathbb{R}^{|W|} \mid q(N) = p \}; \) (\( q(N) \) is the vector of average probabilities associated with \( Q \))
\[ R(p, i)_k = (\frac{\partial}{\partial p_i}) R(p, k/N); \]

\[ B(p,i) = \text{subspace of } \mathbb{R}^{|W|} \text{ generated by vectors } R(p,i) \text{ where scoring rule } R(p, k/N) \text{ is differentiable in } p \text{ and satisfies statement (2) of the theorem.} \]

**Lemma 9.3** \[ Dim A(p)^{\perp} \leq Dim B(p,i); \]

**Proof of lemma 9.3**

\[ Q \in A(p) \text{ satisfies: } \sum_{k \in W} Q(k) = 1, \sum_{k \in W} k_i Q_k / N = p_i, i = 1,..m \nsubseteq 1. \]

These are \( m \) independent equations so \( Dim(A(p)) = |W| - (m - 1) \) (Lemma 9.2)

\[ \Rightarrow Dim(A(p)^{\perp}) = m - 1. \]

We must show \( Dim B(p,i) \geq m - 1 \). It suffices to find \( m - 1 \) linearly independent vectors satisfying (2). In fact, it suffices to find \( m - 1 \) vectors of the form (2) which are independent on the \( m - 1 \) components of \( |W|: \)

\[ K^{(1)} = N,0,0,..; \]
\[ K^{(2)} = 0,N,0,..; \]
\[ ... \]
\[ K^{(m-1)} = 0,0,...N. \]

Choose:

\[ R^{(i)}(p, s(N)) = \frac{1}{2}p_i^2 - p_is_i \]

For \( j = 1...m \nsubseteq 1; j \neq i \), choose:

\[ R^{(j)}(p,s(N)) = -s_i \ln p_i - s_j \ln p_j - (1-s_i-s_j) \ln (1-p_i-p_j). \]

**Verify:**

\[ \partial_i R^{(i)} = p_i - s_i \]

\[ \partial_i R^{(j)} = \frac{p_i - s_i + s_j p_j - p_i s_j}{p_i (1-p_i-p_j)} \]

So; filling in \( s = K^{(h)}/N: \)

\[ j \neq i: \]
\[ \partial_i R^{(j)}(p, k^{(h)}/N) = \frac{p_i}{p_i(l - p_i - p_j)}; \text{ if } h \neq i, h \neq j; \]

\[ \partial_i R^{(j)}(p, k^{(h)}/N) = 0; \text{ if } h = j; \]

\[ \partial_i R^{(j)}(p, k^{(h)}/N) = \frac{p_i - 1 + p_j}{p_i(l - p_i - p_j)}; \text{ if } h = i; \]

\[ \partial_i R^{(i)}(p, k^{(h)}/N) = p_i; h \neq i \]

\[ \partial_i R^{(i)}(p, k^{(h)}/N) = p_i - 1; h = i \]

Multiply row \( j \) by \( p_i(l - p_i - p_j), j \neq i: \)

| \( k(1) \) | \( k(2) \) | \( \ldots \ldots \) | \( k(i) \) | \( k(i+1) \) | \( \ldots \ldots \) | \( k(m-1) \) |
| \( \partial_i R^{(1)} \) | 0 | \( p_i \) | \( p_i + p_i - 1 \) | \( p_i \) | \( p_i \) | \( p_i \) |
| \( \partial_i R^{(2)} \) | \( p_i \) | 0 | \( p_i + p_i - 1 \) | \( p_i \) | \( p_i \) | \( p_i \) |
| \( \partial_i R^{(3)} \) | \( p_i \) | \( p_i \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) |
| \( \partial_i R^{(i)} \) | \( p_i \) | \( p_i \) | \( \ldots \ldots \) | \( p_i - 1 \) | \( p_i \) | \( p_i \) | \( p_i \) |
| \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) |
| \( \partial_i R^{(m-1)} \) | \( p_i \) | \( p_i \) | \( p_i \) | \( p_i + p_{m-1} - 1 \) | \( p_i \) | \( p_i \) | \( 0 \) |

Subtract the \( i \)-th row from each row:

| \( k(1) \) | \( k(2) \) | \( \ldots \ldots \) | \( K(i) \) | \( k(i+1) \) | \( \ldots \ldots \) | \( K(m-1) \) |
| \( \partial_i R^{(1)} \) | \( -p_i \) | 0 | 0 | \( p_1 \) | 0 | 0 | 0 |
| \( \partial_i R^{(2)} \) | 0 | \( -p_i \) | 0 | \( p_2 \) | 0 | 0 | 0 |
| \( \partial_i R^{(3)} \) | 0 | 0 | \( -p_i \) | \( p_3 \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) |
| \( \partial_i R^{(i)} \) | \( p_i \) | \( p_i \) | \( \ldots \ldots \) | \( p_i - 1 \) | \( p_i \) | \( p_i \) | \( 0 \) |
| \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) | \( \ldots \ldots \) |
| \( \partial_i R^{(m-1)} \) | 0 | 0 | 0 | \( p_{m-1} \) | 0 | 0 | \( -p_i \) |

These are linearly dependent only if some linear combination of the rows = (0, 0, ..., 0). That can only be the sum of all rows. However, for column \( k^i \) the sum is \( p_1 + p_2 + ... + p_{m-1} - 1 \). This equals 0 only if \( p_m = 0 \), but we assume \( p_m > 0 \). □
Proof of Theorem

(2) ⇒ (1)

\[ \partial_i E_Q R = \sum_{k \in W} Q(k) \partial_i R = \sum_{k \in W} Q(k) g_i (p, N)(p_i - k_i/N) + \sum_{j \neq h} g_{ijh}(p, N) (p_h k_j/N - p_j k_h/N) \]

\[ = g_i (p, n) \times (p_i - q_i(N)) + g_{ijh}(p, n) \times (p_i q_h(N) - p_h q_j(N)) \]

\[ = 0 \text{ if } p = q(N). \]

(1) ⇒ (2):

From (2) ⇒ (1) we have that \( B(q(N),i) \subseteq A(p)^{\perp} \).

But also \( \text{Dim } A(p)^{\perp} \leq \text{Dim } B(q, i) \) (Lemma 9.3)

Therefore \( A(p)^{\perp} = B(q, i) \), and any \( R \) satisfying (1) satisfies (2). □

Examples

where \( s, p \in S^n \), strictly positive, and \( I(s | p) = \sum_{i=1...n} s_i \ln(s_i / p_i) \):

- \( I(s(n) | p) \) is \( M(X) \) spsrap
- \( I(p | s(n)) \) is NOT \( M(X) \) spsrap
- \( \sum_{i=1...n} (s_i - p_i)^2 \) is \( M(X) \) spsrap (quadratic loss)
- \( \sum_{i=1...n} (s_i - p_i)^2/p_i \) is NOT \( M(X) \) spsrap (chi-square goodness of fit, also leading term of \( I(s | p) \).

4.4 Asymptotic Properties

Definition:
For \( M \subseteq M(X) \), \( R(p,s(n)) \) is strongly asymptotic \( M \) – strictly proper scoring rule for average probabilities if

\[ \forall Q \in M \]

\[ \text{Argmax}_p E_Q (R(p,s(n))) = p(n), \text{ and } q(n) \to r \text{ as } n \to \infty \Rightarrow p(n) \to r. \]

For \( M \subseteq M(X) \), \( R(p,s(n)) \) is weakly asymptotic \( M \) – strictly proper scoring rule for average probabilities (WAM SPSRAP) if

\[ \forall Q \in M \text{ whenever } q(n) \to r \text{ and } r' \in M(O), r' \neq r, \]

\[ \Rightarrow p(n) \to r. \]
then there exists $N' \in \mathbb{N}$ such that for all $n > N'$,
\[ E_Q(R(r, s(n)) > E_Q(R(r', s(n)). \]

**Relative information score:** $RI(p, s(n)) = 2n(I(s | p))$.

**Hypothesis test score** $w_t(p, s(n)) =
\begin{align*}
1 & \quad \text{if } RI(p, s(n)) \leq t \\
0 & \quad \text{if } RI(p, s(n)) > t
\end{align*}$

**Calibration Score:** $W_t(p, s(n)) = (1 - \chi^2_{m-1}(RI)) \times w_t$

$\chi^2_{m-1}$ is the cdf of a chi square variable with $m-1$ degrees of freedom.

**Proposition 9.4** Let $M = \{ \prod p | p \in M(O) \}$ (set of product measures on $X$) For any $t \in (o, \infty)$, $w_t$ is WAM SPSRAP.

**Proof:** choose $Q \in M$, Then $\lim_{n \to \infty} E_Q w_t(q, s(n)) = \chi^2_{m-1} (t)$. Choose $r \in S^m$, $r \neq q$. By the strong law of large numbers, $s(n) \to q$, $Q$ a.s. and by Egoroff’s theorem, for every $d > 0$ the convergence is uniform on a set of probability greater than $1 - d$. Choose $d < \chi^2_{m-1} (t)$, for some $k_d > 0$ we can find $N_d$ such that $\forall n > N_d$, $I(s(n) | r) > k_d$, with probability $1 - d$. For $n > \max{N_d, 2(t/k_d)}$, on this set
\[ 2NI(s(n) | r) > 2(t/k_d)k_d = 2t > t. \]

Hence, for $n$ sufficiently large, $Q[w_t(r, s(n)) = 0] > 1 - d$ and
\[ E_Q w_t(r, s(n)) < d < \chi^2_{m-1} (t). \]

**Remark** This also shows that $E_Q w_t(r, s(n)) \to 0$ as $n \to \infty$ if $r \neq q$.

**Lemma:** For any CDF, $F$:
\[ 0 \leq \int_{-\infty}^z F(x) dF(x) - F(z)^2/2 \leq \max_x [F(x) - F.(x)] \]

**Proposition 9.5:** Let $M = \{ \prod p | p \in M(O) \}$ (set of product measures on $X$) For any $t \in (o, \infty)$, $W_t$ is WAM SPSRAP. For any function
\( f: M(O) \times N \rightarrow [a, b] ; 0 < a < b < \infty \)

\( W, f(p,n) \) is WAM SPSRAP

**Proof:** Let \( Q_n \) denote expert’s CDF for \( RI(q, s(n)) \).

\[
E_Q W_t(q,s(n)) = \int_0^t [1 - \chi^2_{m-1} (x)] d Q_n(x) \\
= E_Q W_t(q,s(n)) - \int_0^t \chi^2_{m-1} (x)d Q_n(x) .
\]

\( Q_n \rightarrow \chi^2_{m-1} \). \( \chi^2_{m-1} \) is continuous and bounded so the Helly Bray theorem together with the above Lemma gives, as \( n \rightarrow \infty \)

\[
= \int_0^t \chi^2_{m-1} (x) d Q_n(x) \rightarrow \chi^2_{m-1}(t)^2/2 > 0 .
\]

Suppose \( r \in M(O) \), with \( r \neq q \). Since \( 1 - \chi^2_{m-1} (x) < 1 \), it follows from the proof of 9.4 that \( \lim_{n \rightarrow \infty} E_Q W_t (r, s(n)) = 0. \)

**4.5 Weights**

Collecting all desiderata: Weights should
1. reward good statistical likelihood and high relative information
2. be asymptotically strictly proper for average probabilities
3. be meaningful, familiar, easy to explain
4. allow likelihood to dominate over informativeness

The weights in SOM 2 satisfy these desiderata.
5. Additional Review of Expert Judgment Cross Validation Research

Discrepancies in Past Cross Validation Studies

EXCALIBUR (Cooke and Solomatine 1992) is the standard software for implementing the Classical Model. It is not possible to use EXCALIBUR for cross validation, but it can be used to spot-check the results of code that implements cross validation. Lin and Cheng (2008; 2009), Flandoli et al. (2011), and Eggstaff et al. (2014) all developed their own code for cross validation, but only Eggstaff et al. report their code and results have been vetted against EXCALIBUR, a process that required extensive coordination.

Large differences exist between the values reported in Lin and Cheng (2008) and Cooke and Goossens (2008). Table 3 details those differences and provides a strong argument for communicating with the authors of the data set before publishing results. The numbers of Lin and Cheng bear little resemblance to those of Cooke and Goossens. In 6 of the 28 studies, the study name and number of experts and seed variables are so divergent that it is not possible to determine which study from the TU Delft database is referenced.

The values in Cooke and Goossens (2008) are published values, some of which were computed with archaic MS-DOS code. That code had a crude method for estimating the tail of a chi square distribution, leading to poor resolution below 1E–4. For large numbers of calibration variables (e.g., as in study 24), this problem could be acute. It was addressed in Cooke and Goossens (2008) by reducing the statistical power to a default value of 10. This might explain part of the discrepancy in study 24. In their later cross validation analysis of more studies, Lin and Cheng (2009) report that they deleted questions that were not answered by many experts and experts who did not answer many questions. If they also did this for their 2008 analysis, that could explain differences in the number of seed variables reported in Lin and Cheng (2008) and Cooke and Goossens (2008). For the other studies, however, no explanation suggests itself for the observed differences in EW and PW combined scores.

<table>
<thead>
<tr>
<th>Study name (from Lin and Cheng)</th>
<th># of experts</th>
<th># of seed variables</th>
<th>Lin and Cheng 2008, Table 1 &quot;within sample&quot;</th>
<th>Cooke and Goossens 2008 Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>PWComb</td>
<td>EWComb</td>
</tr>
<tr>
<td>1 Acrylonitrile</td>
<td>7</td>
<td>10</td>
<td>0.47</td>
<td>0.44</td>
</tr>
<tr>
<td>2 Option trading(^1)</td>
<td>5</td>
<td>34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Dike ring</td>
<td>17</td>
<td>47</td>
<td>0.42</td>
<td>0.03</td>
</tr>
<tr>
<td>4 Flanges</td>
<td>10</td>
<td>8</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>5 Crane risk(^2)</td>
<td>8</td>
<td>10</td>
<td>0.93</td>
<td>0.28</td>
</tr>
<tr>
<td>6 Groundwater</td>
<td>7</td>
<td>10</td>
<td>0.95</td>
<td>0.05</td>
</tr>
<tr>
<td>7 Space debris(^3)</td>
<td>7</td>
<td>26</td>
<td>6.0E-06</td>
<td>0.13</td>
</tr>
<tr>
<td>8 Composite materials</td>
<td>6</td>
<td>12</td>
<td>0.55</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>Radiation in food</td>
<td>7</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------------</td>
<td>---</td>
<td>---</td>
<td>-----</td>
</tr>
<tr>
<td>10</td>
<td>Dry deposition</td>
<td>8</td>
<td>14</td>
<td>0.48</td>
</tr>
<tr>
<td>11</td>
<td>Atmospheric dispersion</td>
<td>8</td>
<td>23</td>
<td>0.38</td>
</tr>
<tr>
<td>12</td>
<td>Early health effects</td>
<td>7/9</td>
<td>15</td>
<td>0.06</td>
</tr>
<tr>
<td>13</td>
<td>Radiation dosimetry</td>
<td>5</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Soil transfer</td>
<td>4</td>
<td>31</td>
<td>1.0E-06</td>
</tr>
<tr>
<td>15</td>
<td>Wet deposition</td>
<td>7</td>
<td>19</td>
<td>0.11</td>
</tr>
<tr>
<td>16</td>
<td>Gas pipelines1</td>
<td>16</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>MONTSE 11</td>
<td>9</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>MOTHER51</td>
<td>5</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>Montserrat1</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Movable barriers</td>
<td>8</td>
<td>14</td>
<td>0.06</td>
</tr>
<tr>
<td>21</td>
<td>Real estate</td>
<td>5</td>
<td>31</td>
<td>0.7</td>
</tr>
<tr>
<td>22</td>
<td>River dredging</td>
<td>6</td>
<td>8</td>
<td>0.54</td>
</tr>
<tr>
<td>23</td>
<td>Sulphur trioxide</td>
<td>4</td>
<td>7</td>
<td>2.53</td>
</tr>
<tr>
<td>24</td>
<td>Building temperature4</td>
<td>6</td>
<td>48</td>
<td>0.002</td>
</tr>
<tr>
<td>25</td>
<td>Atmospheric dispersion (TNO)</td>
<td>7</td>
<td>36</td>
<td>0.09</td>
</tr>
<tr>
<td>26</td>
<td>Radioactive deposition (Delft)5</td>
<td>4</td>
<td>24</td>
<td>0.3</td>
</tr>
<tr>
<td>27</td>
<td>Atmospheric dispersion (Delft)</td>
<td>11</td>
<td>36</td>
<td>0.14</td>
</tr>
<tr>
<td>28</td>
<td>Water pollution6</td>
<td>11</td>
<td>9</td>
<td>0.62</td>
</tr>
</tbody>
</table>

**Table 3:** Comparison of combined scores for performance based and equal weight of Lin and Cheng (2008) and Cooke and Goossens (2008).

**Notes:**

The effective number of seed variables is the smallest number assessed by an expert in the study. The number of total and effective seeds reported in Lin and Cheng do not match Cooke and Goossens for several studies (see below). The statistical accuracy is powered to the effective number of seed variables by EXCALIBUR. In some cases, (e.g., study 24) scores are powered down because of numerical limitations of archaic code.

1. Cooke and Goossens report this study had 12 total seeds and 11 effective seeds.
2. Cooke and Goossens report this study had 18 effective seeds.
3. Cooke and Goossens report this study had 10 effective seeds.
4. Cooke and Goossens report this study had 22 effective seeds.
5. Cooke and Goossens report this study had 11 total and 10 effective seeds.

The out-of-sample code of Flandoli et al. (2011) has been reviewed and found to optimize incorrectly and to conflate uniform and loguniform background measures. Two of the four cases reported had 15 and 16 calibration variables, enabling direct comparison with results from the verified Eggstaff et al. code (2014). The other cases are too large, as Eggstaff split studies with large numbers of seed variables into two separate sets for analysis. Flandoli et al. draw 500 random samples from training sets of fixed size and compute the scores on the complementary test set. Table 4 compares the results from the Flandoli sampling with the complete set using the Eggstaff code.
<table>
<thead>
<tr>
<th></th>
<th>PW</th>
<th></th>
<th>EW</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SA</td>
<td>Inf</td>
<td>Comb</td>
<td>SA</td>
</tr>
<tr>
<td><strong>Pbearl</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 training,</td>
<td>Eggstaff</td>
<td>0.149</td>
<td>0.617</td>
<td>0.072</td>
</tr>
<tr>
<td>8 test</td>
<td>Flandoli Table 8</td>
<td>0.229</td>
<td>0.407</td>
<td>0.093</td>
</tr>
<tr>
<td><strong>Vesuvius</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 training</td>
<td>Eggstaff</td>
<td>0.277</td>
<td>1.176</td>
<td>0.231</td>
</tr>
<tr>
<td>8 test</td>
<td>Flandoli Table 4</td>
<td>0.449</td>
<td>0.896</td>
<td>0.377</td>
</tr>
</tbody>
</table>

Table 4: *Results of Flandoli et al. (2011) based on 500 samples compared with the vetted code of Eggstaff based on the complete set.*
6. Additional Classical Model Applications

This list is current as of November 2016.

6.1. **Nuclear reports** Published as a result of the joint EC/USNRC project on uncertainty analysis of probabilistic accident consequence codes (under the Third EC-Framework Programme)


6.2. **Probabilistic accident consequence uncertainty analysis** Reports published on the project uncertainty analysis of the probabilistic accident consequence code cosyma using expert judgement (under the fourth EC-framework programme)


6.3. Ecosystems and public health


6.3.13. Sarah J. Teck, Benjamin S. Halpern, Carrie V. Kappel, Fiorenza Micheli, Kimberly A. Selkoe, Caitlin M. Crain, Rebecca Martone, Christine Shearer, Joe Arvai, Baruch Fischhoff, Grant Murray, Rabin Neslo, and Roger Cooke (2010) Using expert judgment to estimate marine ecosystem vulnerability in the


6.4. Civil aviation and structural reliability


6.5. Information security


6.6. Natural hazards


6.7. Climate

6.8. Banking and finance

6.9. Completed, publication in preparation


References


