Forecast of environment systems using expert judgements: performance comparison between the possibilistic and the classical model

Jeremy Rohmer, Eric Chojnacki

Forecast of environment systems using expert judgements: performance comparison between the possibilistic and the classical model

Jeremy Rohmer[1], Eric Chojnacki[2]

[1]{BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 - France}

[2]{Institut de Radioprotection et de Sûreté Nucléaire (IRSN), PSN-RES/SEMIA/LSMA, Cadarache, St Paul-Lez-Durance, 13115, France}

Correspondence to: J. Rohmer (j.rohmer@brgm.fr)

Abstract

Expert judgment is widely used to inform forecasts (e.g. using the 5^{th}, 50^{th} and 95^{th} percentile of some variable of interest) for a large variety of applications related to environment systems. This task can rely on Cooke's classical model (CM) within the probabilistic framework, and consists in combining expert information after a preliminary step where experts are weighted using calibration and informativeness scores estimated using some seed questions for which the answers can be obtained. In the literature, an alternative model (PM) has been proposed using a different framework to process the information supplied by experts, namely possibility theory. In the present study, we assess whether both models perform similarly when the seed questions are different from those used to determine the scores, i.e. by taking the viewpoint of forecast. Using an extensive out-of-sample validation procedure, two aspects are investigated using 33 expert datasets: 1) robustness to the set of calibration questions used to estimate the scores, i.e. whether the best and worst performing expert differs; 2) forecast performance, i.e. the degree of accuracy and informativeness of the derived forecast intervals. Regarding 1), the validation procedure shows that PM is less sensitive. Regarding 2), PM achieves more accuracy but with less informativeness when the averaging operator is used. Interestingly, the differences with CM only remain of moderate magnitude for the considered cases despite the conceptual dissimilarities of both models and their lack of agreement on the selection of the best performing expert.

Keywords: Expert Calibration; Forecast Accuracy; Informativeness; Robustness; Probability; Possibility

1 Introduction

Experts' opinions are key ingredients to support the process of decision making (Sutherland and Burgman 2015; Aspinall 2010) and to inform forecasts for environmental systems especially when data are scarce and incomplete. See Burgman (2005) for an overview and discussion regarding conservation and environmental management, Knol et al. (2010) regarding environmental health impact assessment, Krueger et al. (2012) regarding environmental modelling, Drescher et al. (2013) for ecological research, and Lannoy and Procaccia (2014) from an industrial perspective.

Since the original critiques of the practices (Moshleh and Bier 1988; see also Lin and Bier 2008), a large variety of research studies have been proposed to structure the process of deriving information from experts (see Morgan et al. 1990; Cooke 2008; O'Hagan 2019 among others). The formalized treatment of experts' judgments (or opinions) to inform decisions, forecasts, or predictions is named *expert elicitation*. Among the most popular protocol is the Classical Model (CM), originally developed by Cooke (1991). It is based on performance weighted aggregation, i.e. it proposes to aggregate (combine) experts' opinions about a question of interest by pooling them using performance weights (scores), that are calibrated using the answers given by the experts to questions with answers known to the interviewers (termed as seed or calibration questions). CM has been applied in a large variety of different application domains (Cooke and Goossens 2008) and more specifically for environment systems (see some real case applications by Rothlisberger et al. 2012; Metcalf and Wallace 2013; Wittmann et al. 2015). Besides, the CM performance has been tested during extensive validation exercises (Colson and Cooke 2017, 2018; Eggstaff et al. 2014; Lin and Cheng 2009).

The pillar of CM is the use of probabilistic tools to process the information supplied by the experts. In situations of high degree of data/information scarcity, restricting the analysis to the use of only probabilities has, however, been criticized for inducing an appearance of more refined knowledge with respect to the existing uncertainty than is really present (Klir 1989); one problem being that randomness and lack of information can hardly be distinguished (see a detailed discussion by Dubois, 2010) when using only probabilities. Regarding the specific issue of expert knowledge representation, Dubois and Prade (1994) outline that the probability

setting may be often too "rich" to be currently supplied by individuals, because the identification of the probability distribution requires more information than what an expert is able to supply, which is often restricted to the 0.50 and 0.95 percentiles (or a prescribed mode): there are many probability distributions that have the prescribed percentiles. This means that the expert knowledge is pervaded by incompleteness: this lack of precision should be faithfully captured. Therefore, explicitly accounting for this imprecision has motivated the development of alternative uncertainty theories like Fuzzy sets, Dempster-Shafer theory, Possibility theory (see e.g. Dubois and Guyonnet 2011 and references therein). Some examples in the context of decision-making for environment systems are provided by Tacnet et al. (2014) with applications to natural risks.

Adopting such alternative settings does not mean that probability theory is rejected, but aims at complementing it by leaving room for a flexible representation of imprecision in the supplied data. As outlined in the concluding remarks of Flage et al. (2014), testing different approaches for representing and characterizing uncertainties is of high interest to support decision making, because each method can capture different types of information and knowledge i.e. they can shed light to different aspects of the problem and bring different perspectives, and eventually help the decision making process. This has motivated the present comparative analysis between two distinct formalisms for informing forecast using expert judgements: probabilistic by focusing on CM, and an alternative setting by focusing on the one proposed by Sandri et al. (1995) within the possibility theory (Dubois and Prade 1988); termed as Possibilistic Model (PM). Further details on this type of information processing are provided in Sect. 2.2.

PM has been applied in different contexts, namely for risk analysis of spaceflight systems and of chemical process plant (ESTEC and DSM dataset of the TU Delft expert judgment database, Cooke and Goossens 2008) by Sandri et al. (1995), and for information post-processing of nuclear computer codes (Destercke and Chojnacki, 2008; Baccou and Chojnacki 2014). From these previous studies, the following conclusions have been drawn. PM has shown to provide valuable complementary views on expert knowledge, by highlighting more easily the potential conflict between the experts and by measuring directly the reliability (such an information remains difficult to extract from a probability distribution). From a risk assessment perspective, the PM scores are defined based on concepts that are closely related to best estimate and uncertainty bounds (this is further discussed in Sect. 2.2): these are useful to decision makers in the context of risk assessments to understand most likely scenario and to investigate how

sensitivity their decisions are to different risk attitudes, as outlined by Hemming et al. (2020) for ecological applications. Finally, from a practical point of view, the workload for PM-based evaluation appears to be of reasonable magnitude (as outlined by Destercke and Chojnacki, 2008). Besides, it can easily be checked by the experts, and does not lead to incoherencies as outlined by Sandri et al. (1995). This makes the PM procedure of evaluation easy to integrate in any risk and impact assessments.

To date, comparison exercises have been conducted through a direct application of the respective models and by searching the reasons of the dissimilarities. To clarify the best practices, and to improve recommendations for using these approaches to support efficiently the decision-making process, the viewpoint of forecasts has to be addressed (see for instance the discussion by Rae and Alexander 2017 for safety analysis). Put into other words, examining whether both models, CM and PM, perform similarly when they are tested on questions that are different from the ones used to determine the scores, has to our best knowledge, never been tackled. In the present study, we aim at addressing this question by investigating two aspects: 1) robustness to the set of calibration questions: in both models, experts who perform well on the seed questions are afforded more weights. Thus, we aim at assessing the sensitivity to the set of questions, i.e. whether the same experts are afforded the same weight when modifying the questions; 2) forecast performance, i.e. whether both models lead to as accurate and informative forecasts.

The paper is organized as follows. After providing technical details on both models (Sect. 2), we formalize a comparison exercise based on an out-of-sample validation procedure (Sect. 3), which is applied in Sect. 4 on expert datasets that cover a large variety of situations. The comparison results are then discussed in Sect. 5.

2 Methods

2.1 Classical model

We recall the main principles of the Classical Model (denoted CM). Full details and justification can be found in Cooke (1991). CM consists of two stages: 1) Calibration: experts are asked a set of questions (termed seed or calibration) for which the answers are known to the interviewers. These questions relate to the main questions of interest. Experts are scored based on their performance with respect to the calibration questions; 2) Aggregation: the experts'

opinions are combined (aggregated) to inform the forecast regarding the questions of interest. The experts who performed well on the calibration questions (during the first stage) are afforded more weight (denoted W_{CM}) in the final aggregation for the questions of interest.

Formally, let us consider X the unknown variable, P a probability measure on X. The k^{th} percentile, denoted q_k, is the deterministic value x s.t. $P(X \leq x)=k/100$, where $k \in [0,100]$. If B+1 percentiles values have been given by an expert e (including the lower and upper bound), then the corresponding probability density $\boldsymbol{p}^X=(p_1,...,p_B)$ is a histogram made of B inter-percentiles (the value of an inter-percentile being the difference between two successive q_k values). For each of the seed variable, the expert is generally asked answers in the form of percentiles (typically 5^{th}, 50^{th} and 95^{th} percentiles). At the end of this process, the information provided by each expert e is encoded by an empirical probability distribution denoted f_e (one distribution per seed variable and per expert). The aggregation of the n_e expert assessments is performed via a linear pooling, i.e. the weighted averaging of the probabilities provided by the experts (as recommended by Cooke et al. 2020) as follows: $\mathrm{DM}_{\mathrm{avgCM}} = \frac{\sum_{e=1}^{n_e} W_{\mathrm{CM}}(e).f_e}{\sum_{e=1}^{n_e} W_{\mathrm{CM}}(e)}$.

Two main scoring measures are used to assess the ability of an expert to provide a well-calibrated and informative probability distribution. The first one, referred to as informativeness, (denoted Inf_{CM}) measures the degree to which the distribution $\boldsymbol{p}^X$ provided by the expert for the variable X, is concentrated and to which it deviates from the least informative distribution, i.e. the uniform distribution $\boldsymbol{q}$. It is based on the measure of distance between two probability distributions $\boldsymbol{p}^X$, $\boldsymbol{q}$ given by the relative entropy or KL (Kullback-Leibler) divergence (Kullback and Leibler 1951) formally defined as follows:

$$\mathrm{KL}(\boldsymbol{p}^X, \boldsymbol{q}) = \frac{1}{n}\sum_{i=1}^{n} p_i.\log(\frac{p_i}{q_i}) \quad (1)$$

where n is the number of discretized values.

The uniform distribution $\boldsymbol{q}$ is defined on the interval $[l^*, u^*]$ whose bounds can be defined using the overshoot rule, i.e. $l^* = l - k(u-l)/100$, and $u^* = u + k(u-l)/100$, where l, and u are the minimum and maximum values of all answers provided by the experts, and k is a parameter that is here chosen at 10.

The informativeness is then calculated per calibration question, and the score of an expert corresponds to the average information taken across all calibration questions, as follows:

$$\mathrm{Inf}_{CM} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{KL}(\boldsymbol{p}^i, \boldsymbol{q}) \quad (2)$$

where N is the number of calibration questions. Higher numbers represent distributions that show greater departure from a uniform distribution, i.e. they are more informative compared to the uniform distribution.

The second score is the statistical accuracy, denoted Cal_{CM}, (also referred to as calibration score), and compares the adequacy between the information provided by the expert and the known values of the seed variables. Let us consider that the expert has given B percentiles (q_1,…, q_{B}) for N seed variables. On this basis, the following empirical distribution by $\boldsymbol{r}$=(r_1,…,r_{B}, $r_{\mathrm{B+1}}$) can be defined as follows:

- r_{j} is the proportion of seed variables whose values are between q_{j} and $q_{\mathrm{j+1}}$ for for j≠1 and j≠B;
- r_1 (resp. $r_{\mathrm{B+1}}$) is the proportion of seed variables whose values are lower (respectively larger) than the percentile q_1 (respectively q_{B}).

An expert is considered perfectly calibrated if the distribution of the proportions $\boldsymbol{r}$ matches the theoretical distribution derived from the proportions of seed variables within each theoretical inter-quantile range; e.g., for the 5$^{\mathrm{th}}$, 50$^{\mathrm{th}}$ and 95$^{\mathrm{th}}$ percentiles, the theoretical distribution of proportions is $\boldsymbol{r}_{\mathrm{th}}$=(0.05,0.45,0.45,0.05). The comparison between $\boldsymbol{r}_{\mathrm{th}}$ and $\boldsymbol{r}$ can be done using the *KL* distance. The P-value of the chi-square test (with B-1 degrees of freedom) is then used to derive the statistical accuracy, as follows:

$$Cal_{\mathrm{CM}} = 1 - \chi^2_{B-1}(2.N.\mathrm{KL}(\boldsymbol{r}_{th}, \boldsymbol{r})) \quad (3)$$

Higher values indicate an expert's distribution closer to the theoretical distribution, i.e. better calibration.

A global score W_{CM} for each expert is then defined as follows:

$$W_{\mathrm{CM}} = Cal_{\mathrm{CM}} \times Inf_{\mathrm{CM}} \times 1_{\alpha}(Cal_{\mathrm{CM}} \geq \alpha) \quad (4)$$

where $1_\alpha(Cal_{\mathrm{CM}} \geq \alpha)$=1 if $Cal_{\mathrm{CM}} \geq \alpha$, and is zero otherwise. The threshold α is estimated via an optimisation procedure (see Cooke, 1991 for more details), which aims at maximizing the score $Cal_{\mathrm{CM}} \times Inf_{\mathrm{CM}}$ of the "decision-maker" resulting from the linear pooling of all experts $\mathrm{DM}_{\mathrm{avgCM}}$.

2.2 Possibilistic model

2.2.1 Representing expert knowledge

Instead of relying on probabilities to represent expert knowledge, alternative mathematical frameworks rely on the use of an interval-valued representation: when the expert provides a lower and an upper bound of some unknown variable, interval is the simplest approach for representing the pieces of information. In most cases however, experts may provide more information by expressing preferences inside this interval. Such "nuanced" information can be conveyed using the possibility distributions, also referred to as fuzzy intervals or "nested intervals" (Zadeh 1978; Dubois and Prade 1988). A more detailed introduction to possibility theory as a framework for knowledge modelling is provided by Dubois and Prade (2015).

This distribution is formally defined as a mapping $\pi : \mathbb{R} \rightarrow [0; 1]$. The possibility degree $\pi(x)$of a given parameter value *x* is the plausibility of this value given the state of knowledge; if $\pi(x) = 1$, the value is considered totally possible (= plausible); if $\pi(x) = 0$, the value is considered impossible. For instance, say that an expert has provided a best estimate *b* and an interval [*a* ; *c*], where she/he is certain that the true value is located. The preference of the expert is modelled by a degree of possibility ranging from 0 to 1. In practice, the most likely value *b* (referred to as the "core" of π) is assigned a degree of possibility equal to one, whereas the "certain" interval [*a* ; *c*] (referred to as the "support" of π) is assigned a nil degree of possibility, such that values located outside this interval are considered impossible. Linear segments are usually selected for the left and right sides of the possibility distribution, which either correspond to a trapezoidal (or triangular) distribution.

Yet, as outlined by Sandri et al. (1995), a triangular (or trapezoidal) possibility distribution as afore-defined may be too restrictive in practice, because it may overlook the level of confidence that the expert is provided with the forecast interval, i.e. the interval whose bounds are defined by the 5$^{\mathrm{th}}$ and 95$^{\mathrm{th}}$ percentile is assigned a level of confidence of 90%. In the present study, we adopt a more generic definition of possibility distributions by interpreting them from a

probabilistic point of view: possibility distributions can be viewed as a set of nested intervals, each of them being assigned a level of confidence 1-α (Baudrit and Dubois, 2006). These intervals, defined as $\pi_\alpha = \{x, \pi(x) \geq \alpha\}$, are called α-cuts: they contain all the values that have a degree of possibility of at least α (lying between 0 and 1), and they formally correspond to the intervals with a level of confidence 1-α as traditionally defined in the probability theory, i.e. $P(x \in \pi_\alpha) \geq 1 - \alpha$. This means that the level of confidence can be interpreted as the smallest probability that the true value of *X* hits π_α (e.g., from the point of view of the experts, the proportion of cases where $x \in \pi_\alpha$ from her/his experience).

In the situation considered here, the experts provide their answers in the form of percentiles (typically 5^{th}, 50^{th} and 95^{th} percentiles). Based on the approach used by Destercke and Chojnacki (2008), the available knowledge is then represented by a possibility distribution π that is constructed as follows:

- the median value defines the core of π;
- the interval defined by the 5^{th} and 95^{th} percentiles is interpreted as the α-cut, with α=1–0.90=0.10;
- the lower and upper bound (l^*, u^*) define the support of π: these are either provided by the experts or assumed to be linked to the minimum and maximum values of the answers given by the experts (as defined for CM, see Eq. 1);
- linear segments are selected to link the bounds of the support, the 0.10-cut and the core.

Figure 1a provides two examples of possibility distributions constructed based on the (q_5-q_{50}-q_{95}) triplets provided by two experts, namely (0.0, 0.5, 3.0), and (0.5, 3.0, 3.5), for the considered variable *X*. The lower and upper bound of the considered variable respectively reaches -0.5 and 4.0. Figure 1a also provides an illustration on the graphical advantage of this setting: it directly depicts the consensus between both sources of information (i.e. both experts) that is represented by the area where both distributions overlap; the area outside being a representation of the conflict between them (see also a discussion by Baccou and Chojnacki, 2014).

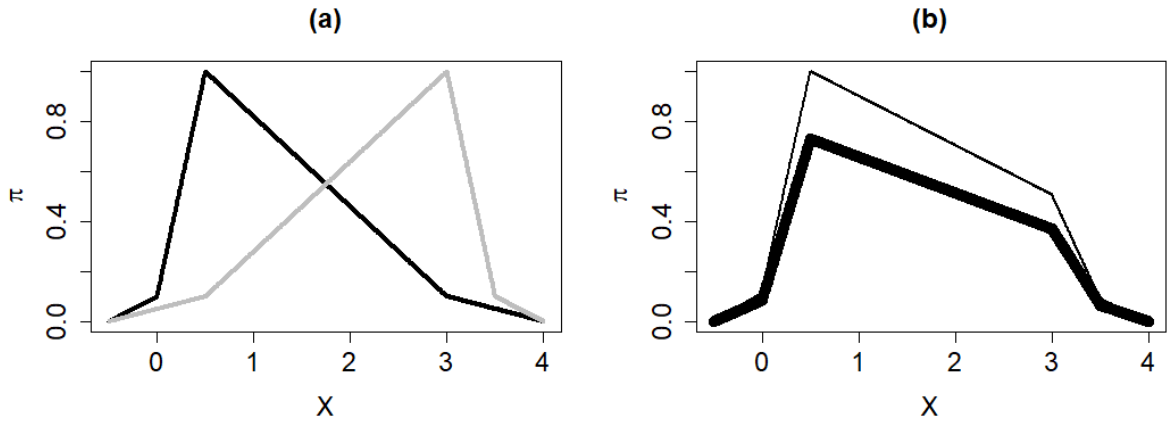

Fig. 1 (a) Examples of two possibility distributions constructed based on the (q_5-q_{50}-q_{95}) percentile triplets provided by two experts, namely (0.0, 0.5, 3.0) in black, and (0.5, 3.0, 3.5) in grey, for the considered variable *X*. The lower and upper bound of the considered variable respectively reaches -0.5 and 4.0. (b) Weighted averaging of the possibility distribution in (a) with weight of 70 and 30% (bold and dotted lines respectively indicate the resulting distribution before and after normalisation between 0 and 1).

2.2.2 Scoring

Similarly as for CM, two main scores are defined (Sandri et al., 1995). Let us consider *X* the variable of interest, and π^X the possibility distribution constructed based on the percentiles supplied by the considered expert. The informativeness is then measured by comparing the imprecision of π^X to the one of the possibility distribution of minimal information (defined as a flat possibility distribution (l^*, u^*) equal to 1.0 between l^* and u^*, and 0.0 outside). A measure of imprecision of π^X is the area $\int_{-\infty}^{+\infty} \pi^X(x)\mathrm{d}x$. The informativeness is then defined as the complement to 1 of the ratio between both areas $I(X) = 1 - \int_{-\infty}^{+\infty} \pi^X(x)\mathrm{d}x \,/(u^* - l^*)$. For the considered expert, the informativeness score Inf_{PM} is estimated by averaging over all calibration questions as follows:

$$Inf_{PM} = \frac{1}{N}\sum_{i=1}^{N} I(X_i) \qquad (5)$$

Figure 2a,b provides two examples, where the blue distribution is the flat possibility distribution, and the triangular ones give the respective information of both experts. In this example, the second expert is less informative than the first one (compare the area in Figure 2b to the one in Figure 2a).

Let us consider x^* the true (known) value of the variable of interest X. The calibration for PM can be understood as the extent to which the considered expert judges x^* as the plausible true value of X: it is formally estimated as the degree of possibility $\pi^X(x^*)$ at x^*. In our representation of expert knowledge (Sect. 2.2.1), this means that the closer $\pi^X(x^*)$ to one, the closer the core of π^X to x^*. The calibration score is then derived by averaging over all calibration questions as follows:

$$Cal_{PM} = \frac{1}{N}\sum_{i=1}^{N} \pi^X(x_i^*) \quad (6)$$

Figure 2c,d provides two examples of triangular possibility distributions, where the second expert is less calibrated than the first one (compare the degree of possibility in Figure 2d to the one in Figure 2c).

Similarly as for CM, the objective is to pool the answers provided by a panel composed of n_e experts regarding the question of interest and to derive an assessment using the weighted averaging of the possibility distributions $\mathrm{DM}_{\mathrm{avgPM}} = \sum_{e=1}^{n_e} W_{\mathrm{PM}}(e).\,\pi(e) / \sum_{e=1}^{n_e} W_{\mathrm{PM}}(e)$. The global score W_{PM} is calculated by following the same principle of CM (Eq. 4 with an optimisation of the threshold α). Figure 1b provides an example of the possibility distribution derived from the weighted averaging of both possibility distributions of Figure 1a.

2.2.3 Differences between PM and CM scores

Conceptually, PM and CM scores are defined based on different considerations; see also Sandri et al. (1995) and Destercke and Chojnacki (2008) for a detailed analysis. On the one hand, Cal_{PM} measures how close the median value (here interpreted as the "best estimate" of the

expert) is to the true value of the variable of interest, and can be interpreted as an accuracy measure similarly as for metrology. On the other hand, the interpretation of Cal_{CM} is closely related to the statistical interpretation of percentiles: when an expert provides q_5, she/he actually says that there is a 5% chance that the true value is below q_5. By providing the median q_{50}, she/he actually says that a 50% percent chance the true value is below the median, etc. Viewing the expert's assessments as statistical hypothesis (e.g. Colson and Cooke, 2018), Cal_{CM} is the P-value for assessing the goodness of fit between the statistical hypothesis and the data, i.e. it measures the degree to which the statistical hypothesis is supported by the data. In this sense, Cal_{CM} is a measure of "statistical" accuracy. Regarding the implementation, a second difference is that Cal_{PM} is calculated by averaging the degrees of possibility calculated per calibration question (instead of relying on the histogram based on all the answers for CM, see Eq. 3). This has advantages from a practical viewpoint: 1) it is less sensitive the number of calibration questions (contrary to CM as extensively by Eggstaff et al., 2014); 2) information can easily be extracted: checking the results is eased, as well as the interpretation, via the identification of the calibration questions where the considered expert performs well (or badly), as outlined by Destercke and Chojnacki (2008) and Baccou and Chojnacki (2014). Since Cal_{CM} is based on the distributions of the expert answers, such reasoning is more tedious to conduct.

Regarding informativeness, there are some similarities in the score definition, which results in most cases to similar results in terms of ordering for weight attribution (see the application cases of Sandri et al., 1995): both scores measure how precise the experts are from the "least informative" expert, i.e. how far the expert distribution is from a distribution of minimal information. The amount of information is however defined differently, either using the uncertainty range (measured by the area below the possibility distribution for PM) or the relative entropy as a measure of statistical information measure for CM. From a practical perspective, PM appears to have an advantage in terms of result presentation and communication, because an area between distributions can directly be graphically depicted, which can ease the interpretation interpretable by non-specialists.

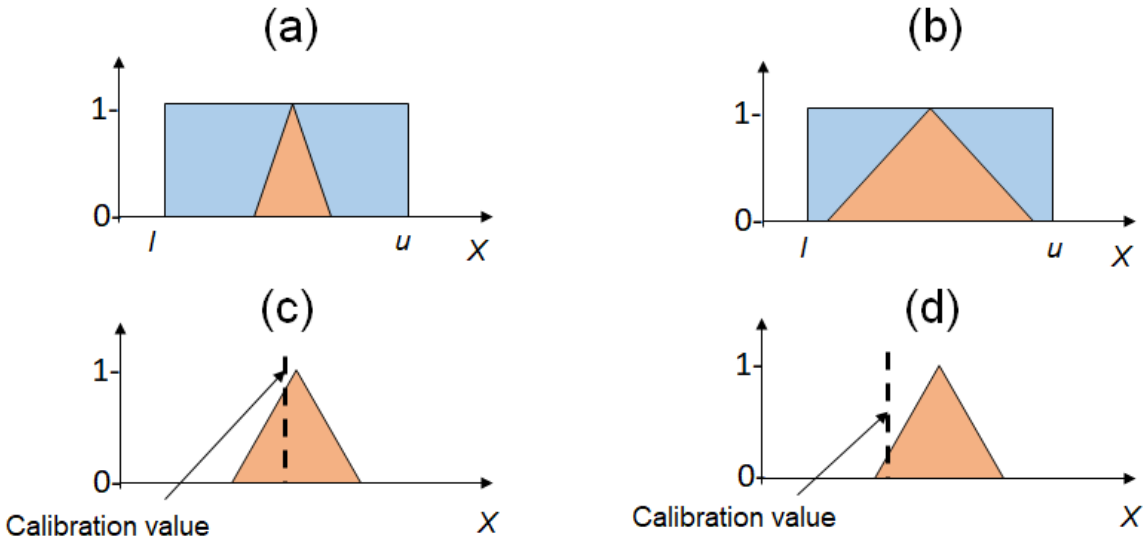

Fig. 2 Examples of expert-based information for a given variable *X* using possibility distribution (a,b): illustration of how informativeness is measured as the area between the blue- and the orange–coloured distribution; here expert (b) is less informative than (a). (c,d): illustration of how calibration is measured as the degree of possibility at the calibration value; here expert (d) is less calibrated than (c).

3 Definition of the comparison exercise

3.1 Procedure

We aim at examining how both models perform when they are tested on out-of-sample data, i.e. questions that are different from the ones used to determine the scores. To do so, we focus on the out-of-sample validation procedure described by Colson and Cooke (2017), which consists in splitting the set of *N* calibration questions into training and test subsets. Colson and Cooke (2017) recommend to size the training subsets at k=80% of the set of calibration questions in order to reach a trade-off between expert performance on the training set and performance of the combinations on the test set. This procedure is performed by considering all combinations of seed questions with a size of 80% of the calibration set, i.e. $N_i = \frac{N!}{(k)!(N-k)!}$ where $N!$ is the factorial of N and equals $N \times (N\text{-}1) \times (N\text{-}2) \ldots 3 \times 2 \times 1$; for instance, an initial set of N=10 calibration questions implies considering N_i=45 different training sets sized at k=8 (and N_S=2 test questions).

We consider two different approaches for aggregating the expert opinions: 1) weighted averaging of the probabilities using the CM (DM_{avgCM}), and the PM scores (DM_{avgPM}) as described in Sect. 2; 2) using the information provided by the best selected expert using the CM (DM_{bestCM}), or the PM (DM_{bestPM}) scores. Though the second approach is less frequently used to inform forecast in practices, this can be informative from a methodological viewpoint regarding our objective of model comparison, because, as indicated by Cooke et al. (2020), the quality of the best expert is the main determinant for the validation procedure of Colson and Cooke (2017).

The comparison is performed by adopting the viewpoint of statistical predictive modelling, and we propose to compare both models by relying on two commonly-used criteria in this domain, namely the stability of the model parameters with respect to changes of the training dataset (i.e. here the sensitivity of the performance-based weights to the calibration phase), and the predictability (here the capability to provide "satisfactory" forecasts); see e.g. Yu and Kumbier (2020). The first aspect, is related to the robustness to the set of calibration questions, and aims at assessing how the weights afforded to the experts are influenced by the set of calibration questions, i.e. the stability of the weights of each expert in the final aggregation depending on the set of questions. In particular, we focus on the worst and best performing expert and assess whether the same expert is systematically selected as the best (or the worst) performing one at each iteration of the validation procedure considering the three scores, calibration, informativeness and global. This is measured by the selection frequency defined as the number of times the considered expert is identified as the best (respectively worst) performing with respect to the considered score. In addition, we analyse whether PM and CM agrees on the best (worst) selected expert, by analysing the agreement frequency, which is defined as the number of times both models provide consistent selection results.

The second aspect relates to the performance of PM and of CM to provide accurate and informative forecasts. Since the different models provide different interpretations and tools for processing the expert knowledge (knowledge representation, see Sect. 2.2.3), we propose to define a common setting of comparison by focusing on three different criteria. We adopt here a pragmatic approach, i.e. the viewpoint of the decision-maker by following the same spirit of the performance measures of the IDEA protocol (e.g., Hemming et al. 2018).

We consider that the answers to the seed questions are range-coded i.e. answers are scaled between 0 and 1 using the lower and upper bound (l^*, u^*) of the considered variable, that are

defined as for computing Eq. 1 in Sect. 2.1. From the viewpoint of the decision-maker, we define the following three performance criteria that are estimated based on the forecast intervals at each of the N_i iterations of the validation procedure:

- *Accuracy of the forecast best estimate.* It is intuitively understood as the degree to which predictions correspond with observed experimental results. To measure the accuracy of the considered expert, we rely on the average log-ratio error (McBride et al., 2012: Eq. 5) as follows:

$$ALRE = \frac{1}{N_i.N_S}\sum_{i=1}^{N_S}\sum_{j=1}^{N_i}\left|\log_{10}\left(\frac{x_{i,j}+1}{b_{i,j}+1}\right)\right| \quad (7)$$

where N_S is the number of calibration (seed) questions in the test dataset at the jth iteration of the validation procedure; $x_{i,j}$ is the true answer and $b_{i,j}$ is the best estimate provided by the expert (assumed to be the median for CM and the value with maximum degree of possibility for PM) for the ith calibration question of the test dataset at the jth iteration of the validation procedure. The ratio $\frac{x_{i,j}+1}{b_{i,j}+1}$ is termed relative error. The lower *ALRE*, the more accurate the considered expert;

- *Imprecision of the forecast interval:* it relates to the width *w* of the expert interval. For CM, this interval is defined by the lower and upper percentile (e.g. q_{95}-q_{05}). For PM, it relates to the width of the α-cut of the possibility distribution, with the α value chosen to be consistent with the probabilistic approach, see Sect. 2.2 (e.g. α=10% when the forecast interval with 90% confidence is provided by the experts). Imprecision is measured using the average score defined by Hemming et al. (2018): Eq. 7, as follows:

$$IMP = \frac{1}{N_i.N_S}\sum_{i=1}^{N_S}\sum_{j=1}^{N_i}\left|\frac{w_{i,j}}{w_{\mathrm{i,j,max}}}\right| \quad (8)$$

where $w_{max}=u^*-l^*$. The ratio $\frac{w_{\mathrm{i,j}}}{w_{\mathrm{i,j,max}}}$ is termed relative interval width for the ith calibration question of the validation dataset at the jth iteration of the validation procedure. The lower *IMP*, the higher the informativeness of the considered expert;

- *Likelihood to miss the true seed value.* It is understood as the degree to which the expert interval contains the true answer. We define the criterion *MISS* as one minus the frequency (considering all N_S calibration questions at all N_i iterations of the validation

procedure) that the true answer falls within the bounds of the forecast interval provided by the expert.

3.2 Data

We use the expert datasets from the post-2006 database[1] analysed by Colson and Cooke (2017) by focusing on the datasets for which the experts provide a triplet of answers (e.g., 5th, 50th and 95th percentiles). A total of 33 datasets are analysed; see a summary in Table 1. They cover a large spectrum of domains of application, namely natural hazards, environmental impact, climatic change, health risk, etc. These datasets are also diverse regarding the number of seed variables (with median value of 13 and inter-quartile of 5) and of experts (with median value of 11 and inter-quartile of 5), hence allowing to tackle a broad range of situations.

Table 1 Description of the expert databases used in the comparison exercise

N°	Expert dataset	Number of seed variables	Number of experts
1	all_CDC	14	48
2	ArsenicD-R	10	9
3	ATCEP_Error	10	5
4	Biol_agents	12	12
5	Brexit-Food	10	10
6	CREATE	10	7
7	CWD	10	14
8	Daniela	7	4
9	eBBP	15	14

[1] Available at http://rogermcooke.net/rogermcooke_files/POST2006EJSTUDIES.ZIP

10	EffusiveErupt	8	14
11	Erie_Carps	15	11
12	FCEP_Error	8	5
13	Gerstenberger	14	12
14	GL-NIS	13	9
15	Goodheart	10	6
16	Hemophilia	8	18
17	ICE_US+EU_June_22_2018	16	20
18	IceSheet2012	11	10
19	liander	10	11
20	p6r	14	21
21	PHAC_2009_T4	13	10
22	PoliticalViolence_March17_CW	21	16
23	puig-gdp	13	9
24	puig-oil	20	8
25	Raveem	18	8
26	Sheep_Scab	15	14
27	SPEED	16	14
28	Tadini_Clermont_anon	13	12
29	Tadini_Quito_anon	13	8
30	TdC	17	18
31	Topaz	16	21
32	umd_nremoval	11	9
33	USGSfinal	18	32

4 Application

4.1 Detailed analysis of one expert dataset

The expert dataset is provided by Gerstenberger et al. (2016) in the domain of probabilistic seismic-hazard assessment. It is available under the title "*Gerstenberger*" within the post-2006 database. The dataset is composed of 14 calibration questions. Examples are provided in the electronic supplementary materials of Gerstenberger et al. (2016). The validation procedure described in Sect. 3.1 considers here $N_i = \frac{N!}{(k)!(N-k)!} = \frac{14!}{(11)!(14-11)!} = 364$ training subsets with size at k=11 (80% of N=14). For each of the training subset, three questions are thus used to evaluate the forecast performance. A panel of 12 experts is considered. The experts are asked to provide the median and the 10th and 90th percentile.

Let us first analyse the 364th iteration of the validation procedure for which the three first seed variables are used as test dataset, namely: the subducted distance, the rock uplift and the surface rupture length. Figure 3 provides an overview of the different answers (q_{10}, q_{50}, q_{90}) using the probabilistic and the possibilistic representation.

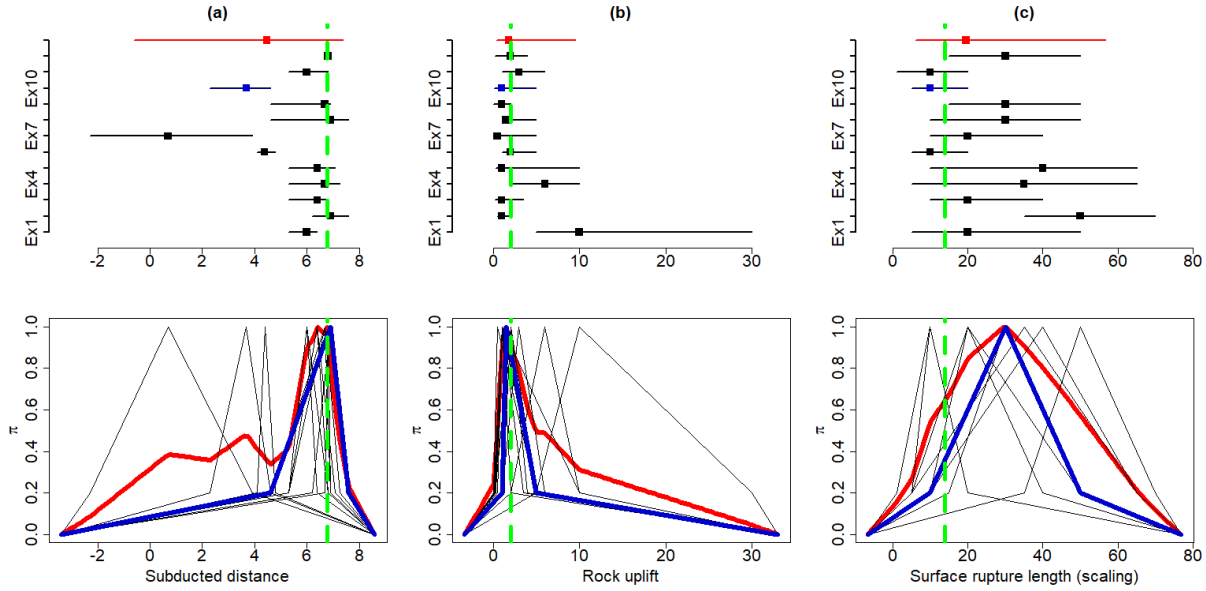

Fig. 3 (Top) Percentiles (q_{10}, q_{50}, q_{90}) provided by the different experts (denoted Ex) for three seed questions of the *Gerstenberger* expert dataset (a,b,c). The blue and red lines respectively represent the information provided by the best performing expert according to the CM global score (Ex 10) and by the CM-based weighted averaging of all experts' information. The

vertical green dashed line indicates the true value of the seed variable; (Bottom) Possibilistic representation of the expert answers. The blue distribution represents the information provided by the best performing expert (Ex 8). The red distribution represents the distribution resulting from the weighted averaging.

Both models, CM and PM, were applied to estimate the calibration, informativeness and global scores (Fig. 4). This shows that the best expert according to CM and PM differ depending on the type of score: for calibration, CM identifies experts Ex 7 and Ex 10 as both leading to the maximum score value, whereas PM identifies Ex 5 as the best calibrated expert. Both models, however, agrees on the least calibrated expert, namely Ex 6 (Fig. 4a); for informativeness, CM and PM respectively identifies Ex 10 and Ex 2 as the most informative expert, and Ex 9 and Ex 5 as the least informative expert (Fig. 4b). Finally, the analysis of the global score shows that both models agree on the least performing expert (Ex 6), but differ on the selection of the best performing expert, i.e. Ex 10 for CM and Ex 8 for PM (the corresponding distributions are outlined in blue in Fig. 3); interestingly, PM score for Ex 10 remains moderate-to-high.

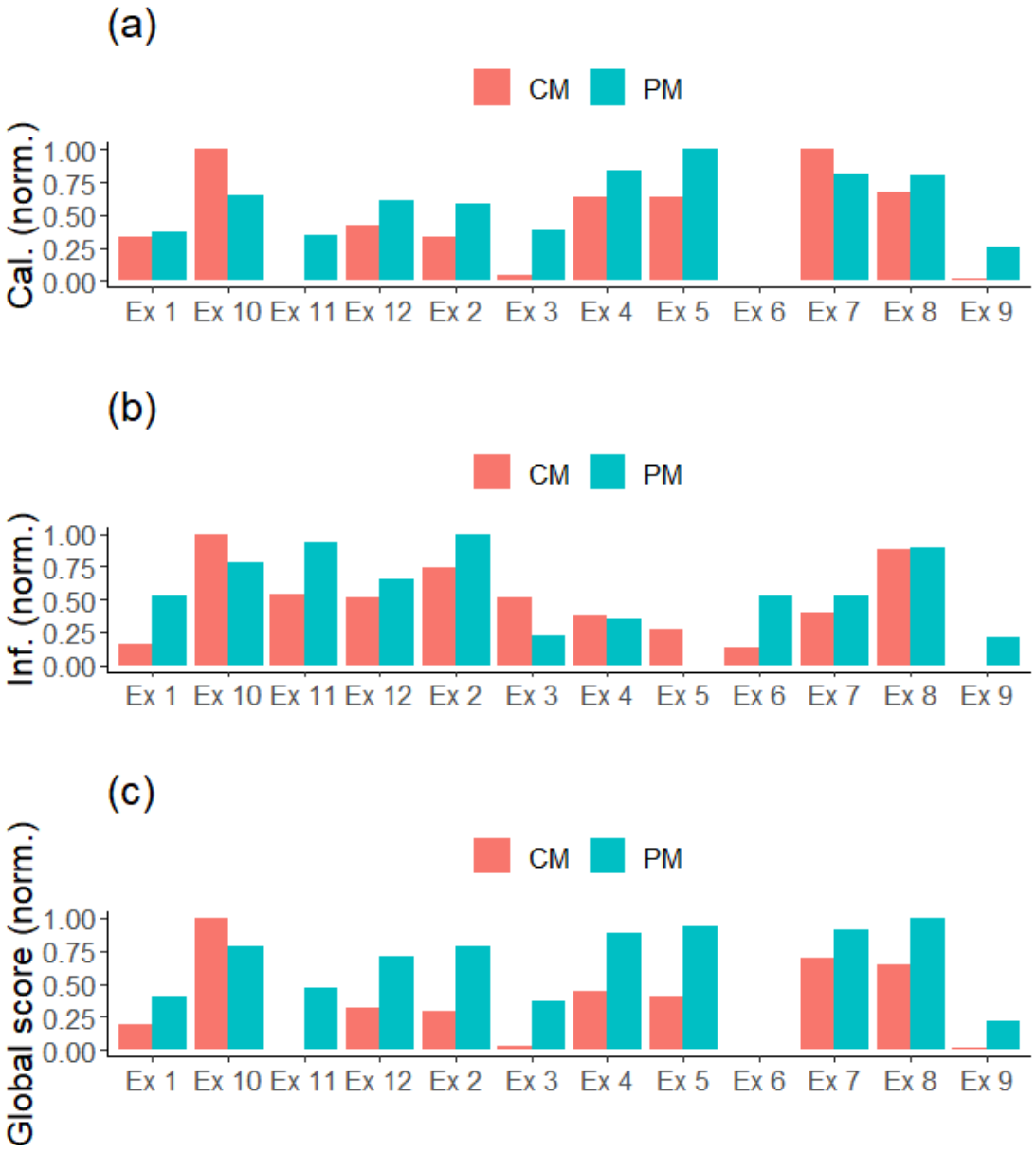

Fig. 4. Performance score (normalized between 0 and 1) for each expert (denoted Ex) and both models, CM and PM, considering the *Gerstenberger* expert dataset: (a) Calibration; (b) Informativeness; (c) Global.

Using the derived global scores, the information provided by the experts are aggregated via a weighted averaging procedure (by following the approach of DM_{avgCM} and DM_{avgPM}), resulting in the red-coloured distribution in Fig. 3a and Fig. 3b for CM and PM respectively. On this basis, we analyse the three aspects of forecast performance (see Sect. 3.1). The analysis of the relative errors with respect to the true seed value (vertical green-coloured dashed line in Fig. 3), considering the four approaches for expert aggregation, DM_{avgPM}, DM_{avgCM}, DM_{bestPM}, and

DM_{bestCM}, shows that the three first types of forecasts are approximately as accurate with a minimum *ALRE* value (calculated for the three considered questions) of 0.022 for DM_{avgPM}, and a maximum one of 0.030 for DM_{bestCM}. The analysis of the relative interval width shows that for this forecast, selecting the best expert, whatever the model CM or PM, is the most informative, with *IMP* (calculated for the three considered questions) of the order of 0.2, but the forecast interval based on DM_{bestCM} fails to contain one of the three true seed values as shown in Fig. 3a. The averaging approach, whatever the model CM or PM, both leads to similarly informative prediction interval (with *IMP* of the order of 0.6) without missing to contain the true (as shown by the red distributions in Fig. 3).

Table 2 Expert selected as best and worst performing for the three scores considering each model, CM or PM, for the *Gerstenberger* expert dataset. The number in brackets provides the number of times the expert is selected for the 364 iterations of the validation procedure.

	PM – Best	CM – Best	PM – Worst	CM – Worst
Calibration	Ex 5 (49%)	Ex 5 (24%)	Ex 9 (38%)	Ex 6 (72%)
Informativeness	Ex 2 (60%)	Ex 10 (78%)	Ex 5 (89%)	Ex 1 (41%)
Global	Ex 8 (69%)	Ex 10 (42%)	Ex 9 (46%)	Ex 6 (76%)

The afore-described analysis is re-conducted 364 times by following the validation procedure of Sect. 3.1. We first analyse the robustness of the selection, i.e. the degree to which the expert identified as best (respectively worst) performing, differs across the different validation iterations. Table 2 shows that PM presents the lower sensitivity to the training dataset for the selection of the best expert (with respect to calibration and global score), whereas it is CM for the most informative one. This is reversed for the selection of the worst performing expert. We note that both models rarely agrees on the identification of the experts with the highest selection probability - expect for the best calibrated expert.

This result is supported by the analysis in Table 3, which shows that the agreement between PM and CM on the selection of the best expert is low-to-moderate: for about 25% of the 364 iterations, PM and CM both agree considering the calibration and the global score. For the worst

performing expert, the agreement is higher of 40-50%. For informativeness, the agreement between PM and CM remains low.

Table 3 Agreement frequency between PM and CM given the best and worst performing expert considering the three scores for the *Gerstenberger* expert dataset.

	Best	Worst
Calibration	25%	52%
Informativeness	16%	10%
Global	24%	42%

Second, we analyse the forecast performance considering all 364 iterations of the validation procedure (Fig. 5). Several observations can be made:

- Fig. 5a shows that, in average, DM_{avgCM} and DM_{avgPM} both lead to quasi-similar relative errors (see red-coloured dot in Fig. 5a), but with a less disperse distribution for DM_{avgPM} (as shown by the boxplots). The evaluation of *ALRE* ~~using Eq. 3~~ shows that DM_{avgPM} is the more accurate with an *ALRE* value of 0.021, but the difference with *ALRE* of DM_{avgCM} remains moderate (<0.003). The approach DM_{bestCM} is the least accurate (with *ALRE* value of 0.039 to be compared to 0.034 for DM_{bestPM});
- Fig. 5b shows that DM_{bestPM}, though the more accurate, leads to the least informative forecasts (with *IMP*=0.67 to be compared to the *IMP* ranging from 0.22 to 0.59 for the alternative approaches);
- Finally, Fig. 5c shows that the forecast interval derived from DM_{avgPM} includes the true seed values ~99% of the investigated cases. Fig. 5c also shows that there is a non-negligible likelihood that the approach based on selecting the best expert misses the true seed value, which lies only 35 and 42% of the times within the forecast interval provided by DM_{avgCM} and DM_{bestCM} respectively.

These results are valid for the considered expert judgement database and their generalisation to other situations is further investigated in Sect. 4.2.

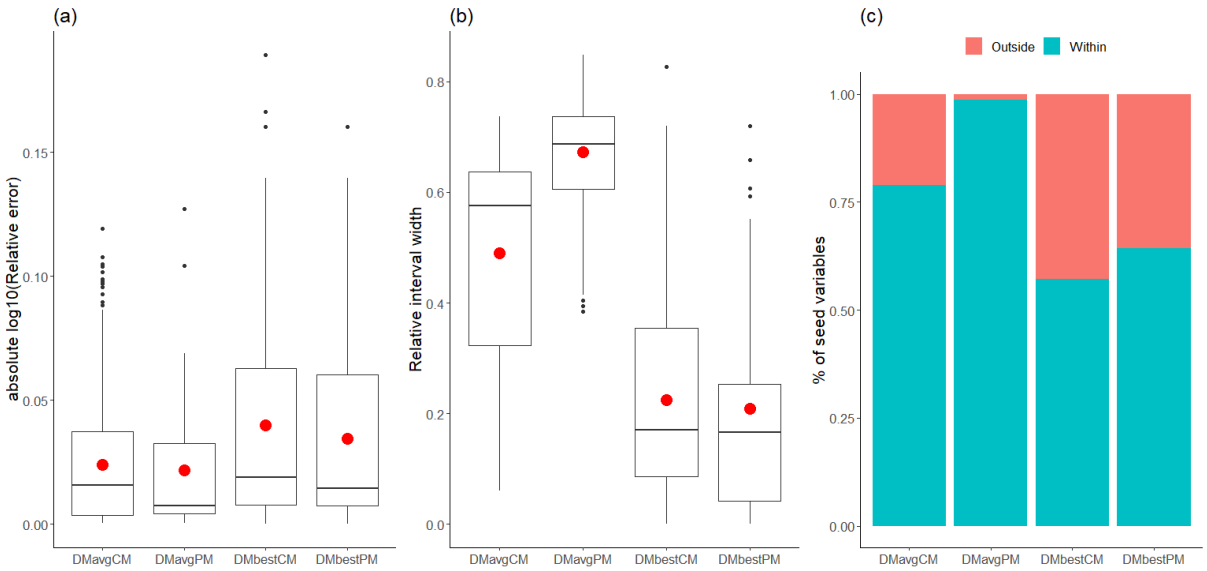

Fig. 5 Forecast performance derived from the validation procedure applied to the *Gerstenberger* expert dataset. (a) Absolute value of the logarithm (base 10) of the relative error (see Eq. 7); (b) Relative interval width (see Eq. 8); (c) Number of times (over all validation iterations) that the true value of the seed variable is within (or outside) the forecast interval.

4.2 Global analysis of multiple expert datasets

The validation procedure described in Sect. 3.1 is applied to all of the 33 expert judgement databases described in Sect. 3.2. Regarding the calibration and the global score, Fig. 6 gives insights into the robustness to the training dataset by showing that PM ~~systematically~~ leads to higher selection frequency of the best performing expert; in particular the median value of the selection frequency exceeds 75%, hence showing that PM-based selection of the best expert is the least sensitive to the changes in the seed variables' values. Considering the worst performing expert, this tendency is inverted and it is CM that leads to higher selection frequency, but the differences with PM appear to be smaller than those for the best performing expert (compare in particular the differences between the left and right pairs of boxplots in Fig. 6c). Regarding informativeness, the robustness of PM and of CM appears to be equivalent with high selection frequencies (larger than 75%).

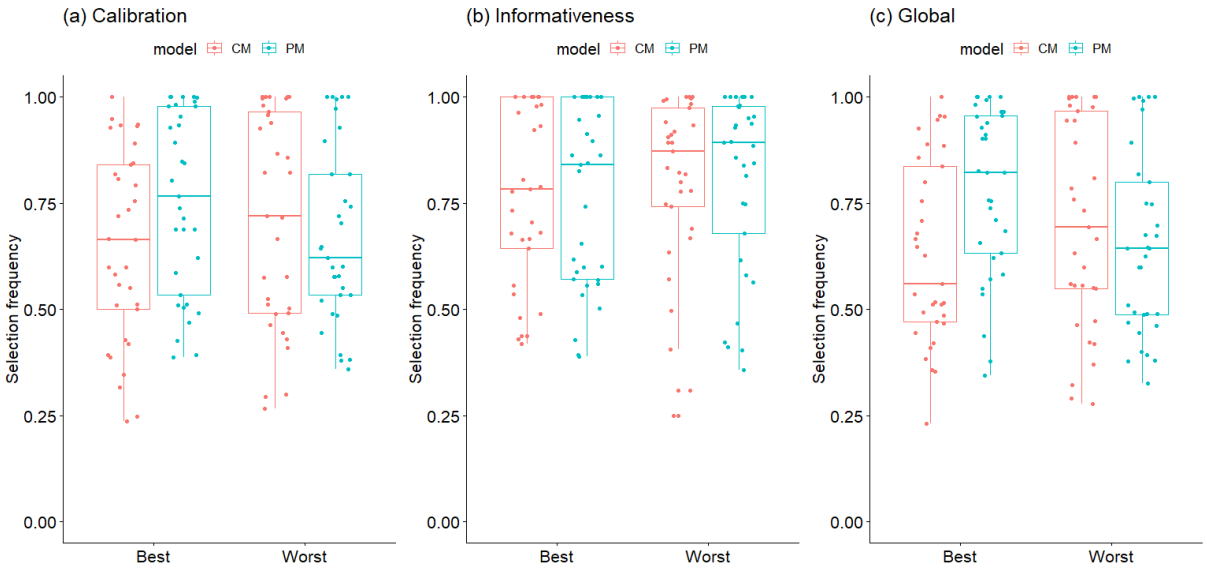

Fig. 6 Selection frequency of the best and worst performing expert considering the different scores: Calibration (a); Informativeness (b); Global (c).

Regarding calibration and global score, the analysis of the agreement frequency in Fig. 7 shows that PM and CM often disagree on the selection of the best and worst performing expert; compare for instance the median values of agreement frequency, which is ~40% for the best expert, and ~65% for the worst one (respectively blue and red boxplot in Fig. 7). Regarding informativeness, both models lead to more consistent results with median values of agreement frequency of >65%, and ~70% regarding the best and worst expert, respectively.

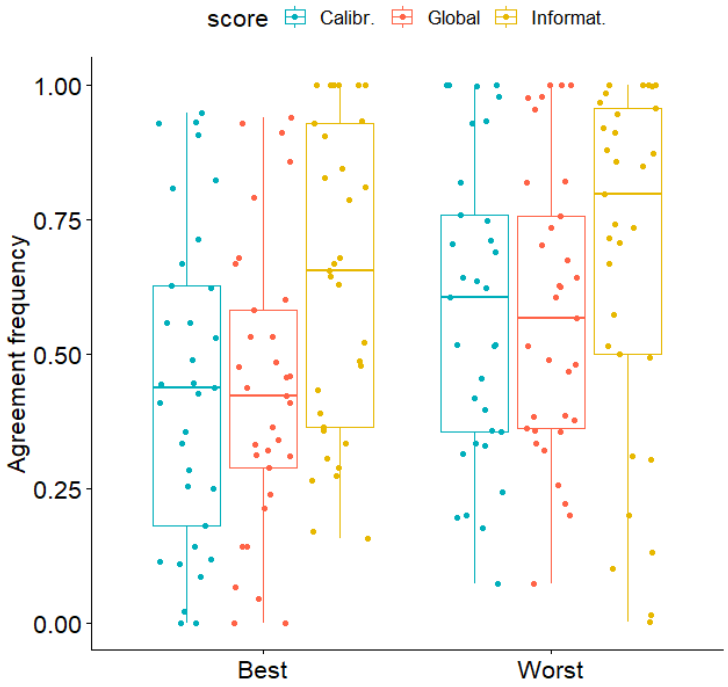

Fig. 7 Agreement frequency between PM and CM given the best and worst performing expert considering the three scores: Calibration (a); Informativeness (b); Global (c).

Second, we analyse the criteria of forecast performance described in Sect. 3.1. Fig. 8 shows the differences between PM and CM for the three criteria. Several observations can be made:

- PM leads to more accurate forecasts (Fig. 8a) whatever the aggregation operator, i.e. weighted averaging DM_{avg} or based on the best selected expert DM_{best}, with 70% (i.e. 23 cases) and 54% (i.e. 18 cases) of the cases with negative *ALRE* differences;
- The differences in accuracy remain however of moderate magnitude: the median value is close to zero (reaching -0.0035 and -0.002 for DM_{avg} and DM_{best} respectively);
- DM_{avgPM} leads to the least informative forecasts, with all the cases leading to positive *IMP* differences (Fig. 8b-left) with large differences (median value ~0.25);
- Though the median value of the *IMP* differences is almost zero, we can note that DM_{bestPM} can result in highly informative forecasts as shown by the long lower tail of the distribution in Fig. 8b-right;
- Fig. 8c shows the higher likelihood for the DM_{avgPM} forecast intervals to cover the true seed values (Fig. 8c-left), and to a lesser extent for DM_{bestPM} as well.

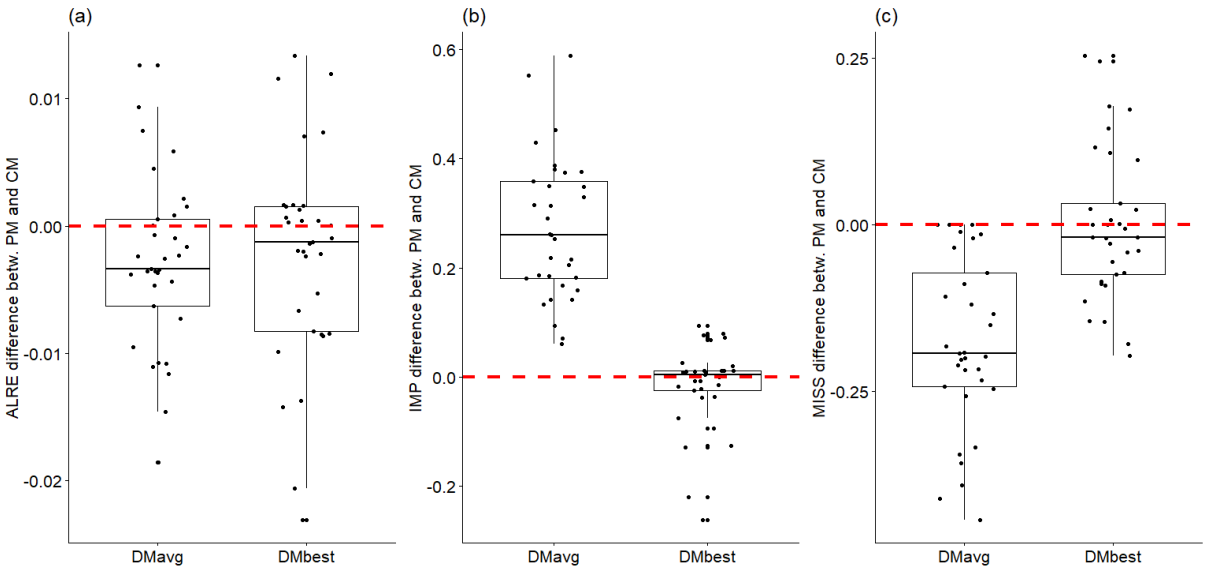

Fig. 8 Boxplots showing the differences between PM and CM considering the forecast performance criteria: (a) accuracy measured by *ALRE*; (b) imprecision of the forecast intervals measured by *IMP*; (c) likelihood to miss the true seed values measured by *MISS*. The mean and standard deviation are shown by the error-bars.

5 Discussion

Table 4 summarizes the main results of the comparison exercise.

Table 4 Summary of the main results

Criterion	Main result
Selection stability	PM is more stable regarding the selection of the best performing expert considering the calibration and the global score. Both models are similarly highly stable considering informativeness.
Agreement	CM and PM only moderately agrees on the selection of the best and worst performing expert regarding the calibration and the global score. Agreement is higher for the selection of the more (or the least) informative expert.
Forecast accuracy	PM, whatever the aggregation approach, leads to more accurate forecast, but the difference with CM remains of moderate magnitude.

Forecast interval's imprecision	PM leads to more imprecise forecasts when the weighted averaging is used.
Likelihood to miss the true value	PM-based forecast intervals almost systematically contain the true value.

The first result of this comparison exercise is the tendency of CM to be less stable than PM regarding the selection of the best performing expert. This result appears to be little influenced by the expert judgement database's characteristics (number of calibration questions, and of experts): the P-value of the linear correlation between the selection frequency and the characteristics is well above the significance threshold at 5% (Table 5). This also suggests that neither increasing the number of calibration questions, nor the number of experts, might improve CM robustness. Finding alternative option for improving this aspect is here identified as a key aspect for further investigation in the future.

Table 5 Linear (Pearson's) correlation coefficient between the performance criteria and the characteristics of the expert judgement database's characteristics. The number in brackets is the P-value of the test of for significance of the correlation coefficient. The numbers outlined in bold indicate that the P-value is below the significance threshold at 5%.

	Number of calibration questions		Number of experts	
	CM	PM	CM	PM
ALRE - DM_{avg}	-0.30 (0.08)	-0.23 (0.19)	**-0.40** **(0.02)**	**-0.39** **(0.02)**
IMP - DM_{avg}	**-0.33** **(0.05)**	-0.09 (0.59)	-0.24 (0.17)	+0.06 (0.71)
MISS - DM_{avg}	-0.02 (0.91)	-0.29 (0.10)	+0.03 (0.89)	**-0.38** **(0.03)**
Selection frequency of the best	+0.17 (0.35)	-0.02 (0.93)	-0.28 (0.11)	-0.30 (0.08)

performing expert using global scores				
Agreement frequency of the best performing expert using global scores	**+0.40 (0.02)**		+0.02 (0.92)	

The second result relates to the agreement of PM and CM. On the one hand, this appears to be high regarding informativeness: this is consistent with the fact that the mathematical models used to represent PM and CM informativeness share some similarities despite the differences in the theoretical backgrounds (as discussed in Sect. 2.2.3). On the other hand, the agreement remains minor-to-moderate for calibration. This was expected, as studied by Sandri et al. (1995) and Destercke and Chojnacki (2008), due to the manner that each score has been defined (as discussed in Sect. 2.2.3). By construction, PM-based calibration score is mainly focused on measuring the deviation from a reference value (i.e. a best estimate), whereas CM-based calibration score is mainly focused on the statistical distribution of seed values in relation to the inter-percentiles given by the expert. As an illustration, let us use the example in Sect. 4.1: Fig. 9a shows the number of times the true seed value falls within the inter-percentile intervals considering experts Ex 7 and 10 (identified as the best calibrated one for CM), and expert Ex 5 (identified as the best calibrated one for PM) considering the *Gerstenberger* expert dataset: for Ex 7 and 10, the proportions within the inter-percentile [10, 50%] and [50, 90%] are closer to the theoretical value of 40% and the proportions within the inter-percentile [0, 10%] and [90, 100%] are closer to 10%. Fig. 9b shows the PM-based calibration scores (i.e. the degree of possibility), which indicates here that Ex 5 is better calibrated when using PM (compare the horizontal coloured lines showing the mean value for each expert). A possible solution to improve the low agreement on calibration is identified by analysing the linear correlation with the expert judgement database's characteristics: the statistically significant positive linear coefficient of +0.40 between the agreement frequency and the number of seed questions (Table 5) clearly indicates that increasing this number can lead to more consistent results between CM and PM.

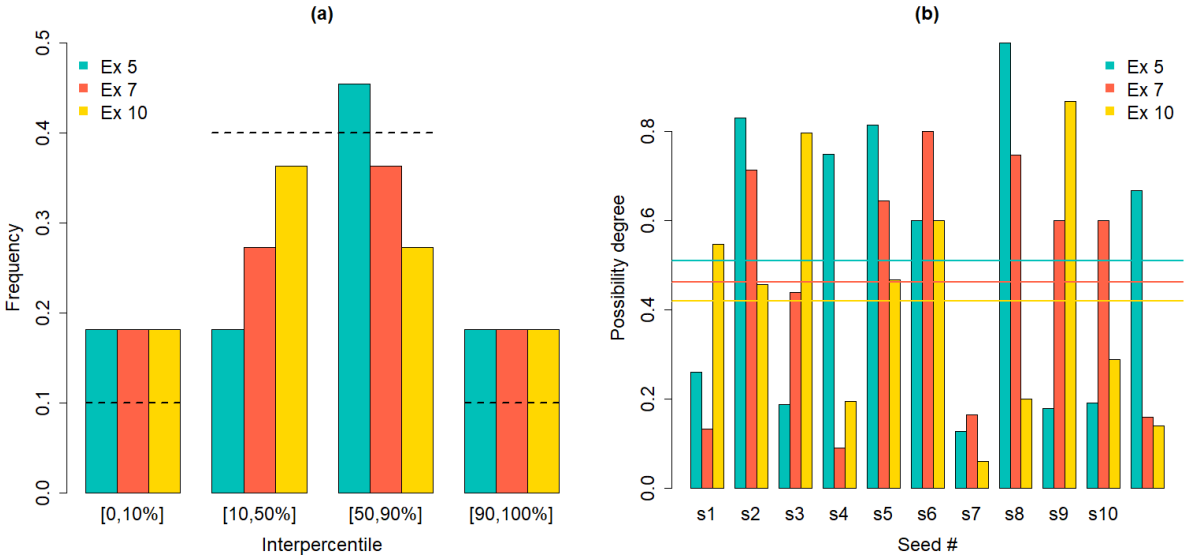

Fig. 9 (a) Proportion of seed variables falling within the inter-percentile intervals considering experts Ex 5, 7 and 10 for the *Gerstenberger* expert dataset. The horizontal dashed lines indicate the theoretical values; (b) Comparison of the degrees of possibility for each of the seed variables (denoted *s*) for experts Ex 5, 7 and 10. The mean value for each expert (i.e. PM calibration score, see Eq. 6) is outlined by the horizontal line.

Finally, the comparison exercise also shows the higher accuracy of PM forecasts (especially when considering the weighted averaging procedure DM_{avgPM}). This appears to be in agreement with the results of Flandoli et al. (2011), which compared CM with their approach named Expected Relative Frequency model, which rewards model ability in central value estimation accuracy (as measured here by *ALRE*). As afore-mentioned, PM-based calibration score should also improve this ability, since it is mainly focused on measuring the deviation from a reference value (i.e. a best estimate). This higher accuracy is however not systematic: PM with weighted averaging leads to more accurate forecasts considering 70% of the total number of cases. The accuracy measured by *ALRE* appears here to be anti-correlated with the number of experts (with a statistically significant linear coefficient of -0.39; see Table 5). This means that the PM-derived forecasts become more accurate when increasing the size of the expert panel. This result related to the "wisdom of the crowd" effect raises however the question of the characteristics of the experts composing the panel (e.g. past experience, degree of expertise, etc.; see e.g. Burgman, 2015), which is out of scope of the present study.

The higher accuracy comes at the expense of a widening of the PM-derived forecasts as indicated by the high *IMP* values (see also the examples of aggregated possibility distributions in Fig. 3b). The advantage is that the true seed value systematically lies within the PM-derived forecast interval (as shown by the third performance criterion *MISS*), i.e. the decision-makers take "less risk" by relying on PM-based forecasts. The drawback of this "safer attitude" is clearly a loss of informativeness; a phenomenon closely related to the so-called "accuracy-informativeness trade-off" (Yaniv and Foster, 1995). When the concern is not point forecasts (i.e. decision making based on best estimate), but interval forecasts, one danger is that the PM-based bounds might be regarded as less useful or less meaningful by the decision-makers (Bolger and Onkal-Atay, 2004).

Solutions to improve PM informativeness cannot be found in the characteristics of the expert judgement databases (number of experts and calibration questions): the examination of the linear correlation shows that PM informativeness is little influenced by these characteristics: the P-value remains well above the significance threshold of 5% (Table 5). Another explanation may be found by analysing how DM_{avgPM} distributes weights to the expert answers in the aggregation. Considering the 33 tested expert datasets, Fig. 10 provides the percentage of experts (averaged over the validation iterations) whose global scores (before applying the thresholding approach) are within the range [0.5, 1.0]. This shows that a larger number of experts are assigned a moderate-to-high global score, i.e. a moderate-to-high contribution in the aggregated forecasts intervals. This is an indication of the lower discriminative capability of the PM global scores. Applying the thresholding approach of Sect. 2.2.2 minimizes this aspect, because some expert answers are discarded in the aggregated forecast interval, but does not fully solve the problem: alternative procedures should then be explored for instance by taking advantage of alternative aggregation operators (see e.g., Baccou and Chojnacki, 2014).

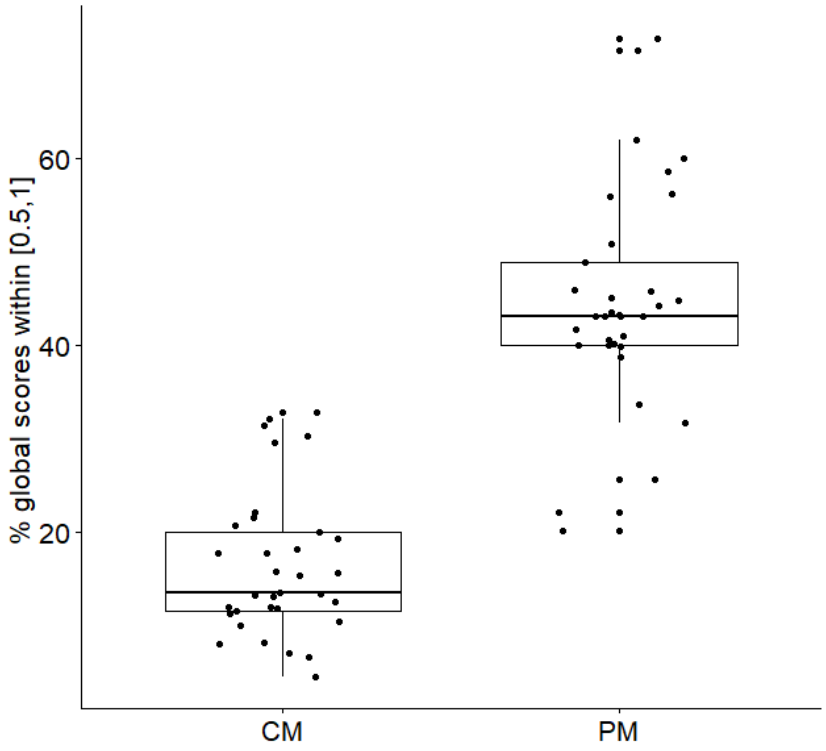

Fig. 10 Percentage of experts (averaged over the validation iterations) whose global scores are within the range [0.5, 1.0] considering PM and CM.

Though CM calibration score is, by construction, focused on statistical accuracy (see Sect. 2.2.3), CM appears, considering the results of our comparison exercise, to achieve a more satisfactory trade-off between both countervailing objectives (accuracy and informativeness), i.e. the forecast intervals are more informative (compared to PM) with a satisfactory level of accuracy (as indicated by the moderate differences with PM-based forecast intervals). Clearly, the price to pay is the higher likelihood to miss the true value. The analysis of the linear correlation (Table 5) shows that the number of seed variables and of experts are anti-correlated (with high statistical significance) with the informativeness and with the accuracy of DM_{avgPM} (with a coefficient value of -0.33 and of -0.40), respectively: this provides evidence of a possible way for improving even more CM performance.

6 Concluding remarks and further work

The objective of the present study is to compare Cooke's classical model with the possibility model to inform forecasts by testing how modifying the set of seed questions affect their performance (robustness and forecast). Given the conceptual dissimilarities of both models (and more particularly regarding calibration), a priori differences in the results were expected. The out-of-sample validation performed on 33 expert datasets confirms it regarding the robustness to the training dataset, which appears to be higher for the possibility model. Regarding forecast performance, the possibility model achieves more accuracy but with less informativeness when the averaging operator is used. Interestingly, these differences only remain of moderate magnitude for the considered cases, and their performance can be viewed as equally satisfactory: this suggests that there is an interest of mixing the forecasts from both models; in particular to shed light on different aspects of the problem like maximizing deviation from a reference value or statistical accuracy. The question of combining the forecast intervals derived from different models constitutes a line for further investigations in the future, in particular by potentially incorporating complementary alternatives, like the likelihood-based of Flandoli et al. (2011) or the new measure of experts' calibration by Hanea and Nane (2019).

The comparison exercise was conducted by making widely-used assumptions related to the parametrisation of both models. We acknowledge that there is room for improvement; in particular a more careful attention should be paid to improving the weighting aggregation of the experts using the derived performance scores. In the aggregation, discarding some experts was performed via a thresholding approach based on performance maximisation (e.g., Colson and Cooke 2017). An alternative procedure may focus on criteria related to how knowledge is represented; for instance the aggregation procedure introduced by Pichon et al. (2014) iteratively selects the expert information based on the concepts of consistency/conflict and specificity. Further work could take advantage of the flexibility brought by the large spectrum of aggregation operators that the possibilistic framework offers (Dubois et al. 2016).

Finally, we have translated, in the present study, experts' answers about percentiles using degrees of possibility by using the links that exist with the probabilistic framework. Though valid and directly applicable to the existing expert judgement databases, feedback from decision makers about the operational use of this interpretation is currently lacking. The operational definition of possibilities, i.e. an explanation in natural language to a decision maker of the concepts, is a key ingredient for Possibility theory to reach an operative state, and future effort

should be intensified in the direction, similarly as has been done for probabilities (see the discussion by Cooke, 2004).

Acknowledgements

This study has been carried out within the NARSIS project, which has received funding from the European Union's Horizon 2020 Euratom programme under grant agreement no. 755439. We are grateful to the three anonymous reviewers whose comments led to great improvement of the article.

References

Aspinall W (2010) A route to more tractable expert advice. Nature 463(7279):294-295.

Baccou J, Chojnacki E (2014) A practical methodology for information fusion in presence of uncertainty: application to the analysis of a nuclear benchmark. Environment Systems and Decisions 34(2):237-248.

Baudrit C, Dubois D (2006) Practical representation of incomplete probabilistic information, Computational Statistics and Data Analysis 51(1):86–108.

Bolger F, Onkal-Atay D (2004) The effects of feedback on judgmental interval predictions. International Journal of Forecasting 20:29-39.

Burgman MA (2005) Risks and decisions for conservation and environmental management. Cambridge University Press.

Burgman MA (2015) Trusting Judgements: How to Get the Best Out of Experts. Cambridge University Press.

Colson AR, Cooke RM (2017) Cross validation for the classical model of structured expert judgment. Reliability Engineering & System Safety 163:109-120.

Colson AR, Cooke, RM, 2018. Expert elicitation: using the classical model to validate experts' judgments. Review of Environmental Economics and Policy 12(1):113-132.

Cooke RM (1991) Experts in Uncertainty. Oxford University Press, NewYork.

Cooke RM (2004) The anatomy of the squizzel: the role of operational definitions in representing uncertainty. Reliability Engineering & System Safety 85(1-3):313-319.

Cooke RM (2008) Special issue on expert judgement. Reliability Engineering and System Safety 93(5):655-656.

Cooke RM, Goossens LLHJ (2008) TU Delft expert judgement data base. Reliability Engineering & System Safety 935:657–674.

Cooke RM, Marti D, Mazzuchi T (2020) Expert forecasting with and without uncertainty quantification and weighting: What do the data say?. International Journal of Forecasting, in press.

Destercke S, Chojnacki E (2008) Methods for the evaluation and synthesis of multiple sources of information applied to nuclear computer codes. Nuclear engineering and design 238(9): 2484-2493.

Drescher M, Perera AH, Johnson CJ, Buse LJ, Drew CA, Burgman MA (2013) Toward rigorous use of expert knowledge in ecological research. Ecosphere 4(7):1-26.

Dubois, D (2010) Representation, propagation, and decision issues in risk analysis under incomplete probabilistic information. Risk Analysis: An International Journal 30(3):361-368.

Dubois D, Guyonnet D (2011) Risk-informed decision-making in the presence of epistemic uncertainty. International Journal of General Systems 40(02):145-167.

Dubois D, Prade H (1988) Possibility Theory: An Approach to Computerized Processing of Uncertainty. Plenum Press, New York.

Dubois D, Prade H (1994) Possibility theory and data fusion in poorly informed environments. Control Engineering Practice 2(5):811-823.

Dubois D, Prade H (2015) Possibility theory and its applications: where do we stand? In: Kacprzyk J, Pedrycz W (eds) Springer handbook of computational intelligence. Springer, Berlin, pp 31–60.

Dubois D, Liu W, Ma J, Prade H (2016) The basic principles of uncertain information fusion. An organised review of merging rules in different representation frameworks. Information Fusion 32:12-39.

Eggstaff JW, Mazzuchi TA, Sarkani S (2014) The effect of the number of seed variables on the performance of Cooke's classical model. Reliability Engineering & System Safety 121:72–82.

Flandoli F, Giorgi E, Aspinall WP, Neri A (2011) Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. Reliability Engineering & System Safety 96(10):1292-1310.

Flage R, Aven T, Zio E, Baraldi P (2014) Concerns, challenges, and directions of development for the issue of representing uncertainty in risk assessment. Risk Analysis 34(7):1196-1207.

Hanea AM, Nane GF (2019) Calibrating experts' probabilistic assessments for improved probabilistic predictions. Safety science 118:763-771.

Hemming V, Hanea AM, Walshe T, Burgman MA (2020) Weighting and aggregating expert ecological judgments. Ecological Applications 30(4), e02075.

Hemming V, Walshe TV, Hanea AM, Fidler F, Burgman MA (2018) Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. PloS one 13(6).

Knol AB, Slottje P, van der Sluijs JP, Lebret E (2010) The use of expert elicitation in environmental health impact assessment: a seven step procedure. Environmental Health 9(1): 1-16.

Klir GJ (1989) Is there more to uncertainty than some probability theorists might have us believe?. International Journal of General System 15(4):347–378.

Krueger T, Page T, Hubacek K, Smith L, Hiscock K (2012) The role of expert opinion in environmental modelling. Environmental Modelling & Software 36:4-18.

Lannoy A, Procaccia H (2014) Expertise, safety, reliability, and decision making: practical industrial experience. Environment Systems and Decisions 34(2):259-276.

Lin SW, Bier VM (2008) A study of expert overconfidence. Reliability Engineering & System Safety 93(5):711-721.

Lin S-W, Cheng C-H (2009) The reliability of aggregated probability judgments obtained through Cooke's classical model. Journal of Modelling in Management 42:149–61.

McBride MF, Fidler F, Burgman MA (2012) Evaluating the accuracy and calibration of expert predictions under uncertainty: predicting the outcomes of ecological research. Diversity and Distributions 18(8):782-794.

Metcalf SJ, Wallace KJ (2013) Ranking biodiversity risk factors using expert groups–Treating linguistic uncertainty and documenting epistemic uncertainty. Biological Conservation 162:1-8.

Mosleh A, Bier VM, Apostolakis G (1988) A critique of current practice for the use of expert opinions in probabilistic risk assessment. Reliability Engineering & System Safety 20(1):63-85.

Morgan MG, Henrion M, Small M (1990) Uncertainty: a guide to dealing with uncertainty in quantitative risk and policy analysis. Cambridge university press.

O'Hagan A (2019) Expert knowledge elicitation: subjective but scientific. The American Statistician 73(sup1):69-81.

Pichon F, Destercke S, Burger T (2014) A consistency-specificity trade-off to select source behavior in information fusion. IEEE transactions on cybernetics 45(4):598-609.

Rae A, Alexander R (2017) Forecasts or fortune-telling: When are expert judgements of safety risk valid? Safety Science 99:156-165.

Rothlisberger JD, Finnoff DC, Cooke RM, Lodge DM (2012) Ship-borne nonindigenous species diminish Great Lakes ecosystem services. Ecosystems 15(3):1-15.

Sandri SA, Dubois D, Kalfsbeek HW (1995) Elicitation, assessment, and pooling of expert judgments using possibility theory. IEEE transactions on fuzzy systems 3(3):313-335.

Sutherland WJ, Burgman M (2015) Policy advice: use experts wisely. Nature 526(7573):317-318.

Tacnet JM, Dezert J, Curt C, Batton-Hubert M, Chojnacki E (2014) How to manage natural risks in mountain areas in a context of imperfect information? New frameworks and paradigms for expert assessments and decision-making. Environment Systems and Decisions 34(2):288-311.

Wittmann ME, Cooke RM, Rothlisberger JD, Rutherford ES, Zhang H, Mason DM, Lodge DM (2015) Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. Conservation Biology 29(1):187-197.

Yaniv I, Foster DP (1995) Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. Journal of Experimental Psychology: General 124:424 – 432.

Yu B, Kumbier K (2020) Veridical data science. Proceedings of the National Academy of Sciences 117(8):3920-3929.

Zadeh L (1978) Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems 1:3–28.

Declarations

Funding

This study has been carried out within the NARSIS project, which has received funding from the European Union's Horizon 2020 Euratom programme under grant agreement no. 755439.

Conflicts of interest/Competing interests

The authors wish to declare neither conflicts of interest nor competing interests.

Availability of data and material

Expert datasets are available at http://rogermcooke.net/rogermcooke_files/POST2006EJSTUDIES.ZIP

Code availability

R scripts for computing the performance scores are available here: https://github.com/rohmerj/ExpertScoring. More detailed versions of the scripts for conducting the cross-validation exercise are available upon request to the corresponding author.

Authors' contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Jeremy Rohmer. The first draft of the manuscript was written by Jeremy Rohmer and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.