A Bayesian method for calibration and aggregation of expert judgement

David Hartley and Simon French

Abstract This paper outlines a Bayesian framework for structured expert judgement (SEJ) that can be utilised as an alternative to the traditional non-Bayesian methods (including the commonly used Cooke's Classical model [13]). We provide an overview of the structure of an expert judgement study and outline opinion pooling techniques noting the benefits and limitations of these approaches. Some new tractable Bayesian models are highlighted, before presenting our own model which aims to combine and enhance the best of these existing Bayesian frameworks. In particular: clustering, calibrating and then aggregating the experts' judgements utilising a Supra-Bayesian parameter updating approach combined with either an agglomerative hierarchical clustering or an embedded Dirichlet process mixture model. We illustrate the benefit of our approach by analysing data from a number of existing studies in the healthcare domain, specifically in the two contexts of health insurance and transmission risks for chronic wasting disease.

Key Words: *Structured Expert Judgement, Risk assessment, Bayesian hierarchical model, Calibration, Homogeneity groups.*

1 Introduction

Statistical decision theory outlines how decision makers are regularly required to make judgements in the face of uncertainty. Suppose a decision maker (DM) needed to assess the consequences of their decision based on the outcome of a set of impactful variables, $\mathbf{X}$, known as *target variables*. Let us assume he or she starts by

David Hartley
Department of Statistics, University of Warwick, Coventry, CV4 7AL, e-mail: d.s.hartley@warwick.ac.uk

Simon French
Department of Statistics, University of Warwick, Coventry, CV4 7AL, e-mail: simon.french@warwick.ac.uk

specifically considering the expected outcome of a single random variable $X \in \mathbf{X}$. Furthermore, let us assume that observations of X, denoted x_X, follow an unknown probability distribution $p(x_X)$. In many cases the DM may not have the data to assess $p(x_X)$ robustly and so must reach out to experts to give their assessment of their uncertainty. Let $\mathbf{E}$ define the set of experts, $|\mathbf{E}|$ the number of experts and $e \in \mathbf{E}$ an individual expert. Let us assume the DM held the belief, $\pi_{DM}(x_X)$, prior to talking to the experts and the experts' beliefs are given by $p_E(x_X) = \{p_e(x_X) : e \in \mathbf{E}\}$. The goal of a structured expert judgement model will thus be to build a distribution for the DM's perspective of the uncertainty given the experts' statements, $p_{DM}(x_X)$. We are looking for a function ϕ such that $p_{DM}(x_X) \propto \phi(p_E(x_X), \pi_{DM}(x_X))$ [26]. This problem corresponds to the expert problem as outlined in French [21] and [22].

A simple approach which recognises that the DM may explicitly rely solely on the experts, is for the DM to simply average over the experts' beliefs, i.e. $p_{DM} = \phi(p_E) = \sum_{e \in \mathbf{E}} (1/|\mathbf{E}|) p_e$. Here DM prior beliefs are either ignored or added into the sum by stating that the DM is an additional expert in $\mathbf{E}$. This method can then be easily extended to allow the DM to vary the effect of each expert on the resulting distribution.

$$p_{DM} = \sum_{e \in \mathbf{E}} \omega_e p_e \tag{1}$$

where ω_e is a weighting factor for expert e. These methods are termed linear pooling methods ([31], [25], [10]). Different pooling models then propose different ways to determine the relative weightings ω_e. This can be simple averaging or determined based on some performance evaluation metric. The most utilised version of this linear opinion pool is Cooke's Classical model [13] and Cooke's model is often taken as the benchmark for comparing algorithmic methods.

Key benefits of these pooling methods include; simplicity for DM's to understand, speed of application, and critically, they make the aggregation process transparent and auditable. They have also, in the case of Cooke's Classical model in particular, proven to be robust under significant scrutiny [19]. They are not, however, without issue. They do assess diversity of opinions provided, however, they do not capture any underlying consistency in opinion the experts may be highlighting. Pooling methods also mean that DM's own beliefs are not considered explicitly, (other than in parity with the experts if their own assessment is included in the sum). Similarly, known biases in experts or inter-expert correlation can impact how a DM may wish to update their belief but are not considered. Finally, the key output of a pooling approach may be an unparameterised distribution. Often a parameterised distribution (to feed as a prior into a further model) is desired. If parameterised distributions are elicited directly from experts, or we fit a continuous parametric distribution to elicited quantiles, the output of pooling methods will be a finite mixture of parameterised distributions.

Bayesian models try to tackle these issues whilst retaining some of the advantages of the pooling techniques. A Bayesian expert model flowing directly from Bayes theorem assumes that the posterior the decision maker has on x_X, conditional on the experts' beliefs, $p_{DM}(x_X|p_E)$ is proportional to the likelihood they ascribe to hearing the experts' elicited values of p_E given x_X, $p_{DM}(p_E|x_X)$, multiplied by the

prior belief the DM has, π_{DM}, i.e. $p_{DM}(x_X|p_E) \propto p_{DM}(p_E|x_X) * \pi_{DM}(x_X)$. Thus the DM treats the elicited information as data. Bayesian methods for expert judgement are not new ([38], [40], [37], [20]). They have not been used in practice as much to date as pooling techniques due to an over-sensitivity to input conditions and difficulty in implementation, the key challenge of which is calculating the likelihood function $p_{DM}(p_E|x_X)$. Early Bayesian models treated experts as completely exchangeable and did not consider calibration, this often resulted in very narrow posterior distributions which demonstrate high overconfidence. There is broad recognition that today, pooling models, Bayesian methods and indeed other approaches conceptually may outperform each other in different contexts ([18], [13], [50]).

Recently, Bayesian methods have become more tractable. Albert et al. [2] demonstrated a Supra-Bayesian hierarchical model for the aggregation of expert judgement utilising homogeneity groups. But, (as was discussed in the subsequent comments) did not explicitly tackle how a DM could adjust for known calibration issues of the experts, nor how the homogeneity sets into which experts are grouped could be attained. Similarly, Clemen and Lichtendahl [9] put forward an intriguing hierarchical approach into how one could address calibrating an expert's opinion but not how a DM could then aggregate these recalibrated results together. More recently, Billari et al. [4] proposed a relatively parsimonious Bayesian aggregation method considering mixture models and Perälä et al. [42] proposed an interesting model for calibration utilising Gaussian hierarchical processes. In none of the models outlined however, has calibration, aggregation and homogeneity group definition been attempted within a single framework.

Remark. There are those who would argue this form of aggregation should not occur at all, [36]. They suggest we bypass the issue of combining probability distributions altogether. Firstly, by getting experts to share all the information they have relevant to the issue at hand and then having the group of experts discuss, align and agree on a "single body of evidence." This evidence is then passed through Bayes theorem item by item to arrive at a posterior probability curve. This is arguably a mixture of a behavioural and algorithmic method for aggregation. Other behavioural methods such as the DELPHI *[16] and the* SHEFFIELD *[23] method exist, however, these are not without their own issues and are reviewed in French [22] and EFSA [18].*

Other potential models for analysing and aggregating expert judgement sit outside of traditional probability theory and consider broader concepts inherent within evidence theory or Dempster-Shafer theory ([30], [1], [5], [44]). Problems with significant structural uncertainty will lend themselves to utilise these methods. In this case we will be focussing on areas where the structure is formalised sufficiently that Bayesian methods are applicable. Although we do note, that further work should be done to assess and compare Bayesian methods for SEJ versus those that utilise evidence theory notably, the transferable belief model, [30].

It is also important to highlight the different DM contexts within which SEJ studies are performed. In some studies there is a specific DM who is close enough to the modelling to provide their prior belief, π_{DM}, into the process. In these instances, as outlined previously, the modelling is calculating how the decision maker should

adjust their beliefs given the experts' judgements. In other contexts the DM is farther away from the modelling and the output of the SEJ study is the perspective of a *rational scientist*, [18]. Here the aim is to identify what a rational scientist would believe given the experts' statements. All of the knowledge about the target variables should be encoded in the experts' judgements and the DM prior used in the model should be as uninformative as possible. This second context is very common and the Bayesian approach needs to be applicable in both cases.

There is an important decision to be made about whether experts are making an assertion regarding the uncertainty of an input variable to a decision/model, (i.e. helping with the construction of an informative prior for a larger more complex model) or to the resulting output (supporting the development of the DM's full posterior for the model). This has implications regarding whether experts should provide judgements on parameters or only upon observables. For now we will leave aside this question as in many cases this subtlety does not impact the final result and therefore, unless explicitly stated, we will assume it does not matter.

In practice it can be difficult for experts to think in terms of distributions and therefore it is often wise to simplify the problem by not directly eliciting the distribution or parameters involved. Instead we can extract each expert's perspective on some intuitive points within the broader distribution and then construct a function g_e which represents our best approximation to the expert's beliefs given the elicited data. Cooke [13] outlines the benefits of eliciting in this way, given the challenges experts have in mentally formulating parametric distributions. Typically we would elicit three quantiles[1] from each expert $e \in \mathbf{E}$, L_e, M_e, U_e, associated with three probabilities P_L, P_M, P_U (often the 0.05, 0.50 and 0.95 quantiles) and the full distribution for the expert is approximated by $g_e(\cdot|L_e, M_e, U_e)$. In certain studies, five quantiles may be elicited (often these represent the 0.05, 0.25, 0.5, 0.75 and 0.95 quantiles). In this case, the full distribution g_e is thus conditional accordingly. g_e will often be from a parametric family and as such we encode our model further by utilising L_e, M_e, U_e to infer the parameters γ_e of g_e, that represents the expert e's conceptual model of X. Thus, g_e and γ_e should be chosen in such a way that it closely approximates the expert's beliefs at the elicited quantiles (i.e. $g_e(x|\gamma_e) \equiv p_e(x) + \varepsilon$, with minimal error term ε.). This process should not be done in isolation from the experts and a feedback process is often employed to playback g_e and give an opportunity for refinement. Other methods exist for obtaining g_e more directly from experts and many authors have considered the best elicitation methods for expert judgement models ([3], [7], [29], [28], [23]). The decision maker's prior will often be elicited and parameterised in a similar way.

This paper contributes further to the discussion by outlining a more complete Bayesian model, building on the work of Albert et al. [2] and Clement and Lichtendahl [9], which both calibrates and aggregates the experts' judgements into a suitable parameterised posterior for the decision maker, whilst also assessing probable homogeneity groups for the experts. This assessment is made using a clustering algorithm. Advantages of different clustering techniques are discussed. This method

[1] The elicited quantiles should be equivalent to the expert's true beliefs +/- any elicitation error [45].

is reviewed against both the simplistic expert averaging pooling model and the prevailing SEJ model (Cooke's Classical model) within two historic studies.

In Section 2 we review the commonly used Classical model in addition to the key Bayesian approaches which we will build from. Section 3 will outline the structure of our model before an analysis of results for two historic studies and conclusions in Sections 4 and 5 respectively.

2 Outline of key SEJ methodologies

One of the key issues when considering expert judgement is that the DM is considering the psychology of the experts as much as the quantitative information they provide. There is much literature into the cognitive biases that experts may exhibit, [35], as well as the reasons for discrepancies in the judgements they offer [41]. Two topics that can be useful for highlighting differences between experts' judgements are expert statistical accuracy and expert information [13].

Statistical accuracy assesses how well an expert's forecasts truly represent reality. An expert's judgements on a variable may not reflect truth due to a fundamental misunderstanding of the underlying generative physical model or they may simply be miscalibrated due to a systematic error such as overconfidence or some form of cognitive bias.

Remark. In some of the earlier literature on Cooke's Classical model [13] experts' "statistical accuracy" was referred to as "calibration". In order to avoid confusion with recalibration techniques, discussed later, Cooke has recommended updating the terminology [11].

Information, is a measure of how useful an expert's opinion is to a DM. If an expert provides a very vague forecast this is less useful for decision making than if they are able to be more specific (assuming they are statistically accurate). Thus, typically a high level of information would manifest as a very tight distribution (or narrow set of quantiles) elicited from the expert. Information is a concept closely aligned with 'sharpness' of forecasts, outlined by Gneiting et al [27]. Sharpness, in this context, is defined by the concentration of predictive distributions. Similar to information, sharpness is a property of the forecasts only.

Cooke's Classical model utilises these two phenomena, creating a weighting for each expert in an aggregation of the form (1). This weighting is calculated based on the expert's performance over a set of seed variables for which true realisations are known *a priori*. Bayesian models approach these topics differently, trying to adjust the DM's belief given the information they provide. There is a subtle but important distinction here. In the Classical model, experts are included (or excluded) based on their performance relative to these measures. In the Bayesian model, all experts are included equally, however, the uncertainty the DM has regarding the variables,

i.e. the credence he or she puts in what they hear, is adjusted based on experts' performance in these variables.[2]

2.1 Cooke's Classical model

Based on the linear opinion pool, the format of Cooke's Classical model is as equation (1). Weight for expert e, ω_e, is proportional to the product of a statistical accuracy score and an information score, given the data elicited from e. Statistical accuracy is calculated as the p-value of e's assessments for a number of seed variables versus empirical results, measured via a chi-square test. Information is as function only of the judgements themselves and is an assessment of the increased precision e gives versus a background distribution, typically the uniform or log-uniform distribution. Informativeness is measured utilising Shannon relative information. For a full outline of Cooke's Classical model, a review of some of the original literature ([13], [14]) is recommended. It has been shown that Cooke's format preserves the behaviour of an asymptotically proper scoring rule and thus each expert will be rewarded for demonstrating their true beliefs.

The number of practical applications of the Classical model in decision problems is testament to the clear benefits this framework has. This model has proven to be robust; Cooke ([13], [15]) has a database of over 80 studies conducted using this methodology and there are many publications validating the results of the model ([12], [8], [19]). There are however, limitations to this approach in addition to those outlined for pooling techniques in general. In practice, in a large proportion of Cooke's model applications, the majority of experts are set with a weighting of zero and only a few impact the final result (with the exception of tail extremities which still consider all experts). In many cases all of the weight goes to a single expert. French [22], postulates that this approach may be seen as undemocratic as members of the group may be completely excluded from the final decision. The statistical accuracy calculation of the Classical model also relies on a significant number of seed variables, however, ultimately only uses these as a scoring mechanism. If an expert were to display consistent bias, their accuracy score would be low and their impact on the final decision would be minimal. If it were possible to adjust for this bias in a rigorous way however, their full judgement could be utilised to help inform the DM. In many cases the cost of collecting data is high and as such the maximum amount of verifiable information needs to be extracted from each data point.

[2] One philosophical debate concerns whether the DM should be allowed to adjust experts' elicited values. Doing so potentially reduces accountability for the experts. If you do not adjust however, knowing that experts are miscalibrated, then you may be wilfully ignoring useful information. In this paper we will not consider the merits of either philosophy in detail, however, will outline a model which adjusts for known calibration issues.

Bayesian models[3] try to address some of these issues, however, inherently create some of their own. They are by nature more complex than pooling techniques and for a Bayesian approach to be considered a practical choice for a DM it must thus be shown to outperform Cooke's Classical model against some criteria. The appropriate criteria to use here will be discussed later. Within this paper we will demonstrate the impact that considering these elements can have to the final distribution for a DM.

2.2 A Bayesian model for calibration

Consistent with ideas from Cox [17] and Morris [39], Clemen and Lichtendahl [9] developed a model of overconfidence using past data to estimate, what the authors term, "inflation factors" for assessed distributions post hoc. Bayesian hierarchical models are used, allowing experts to be calibrated individually whilst simultaneously capturing inter-expert calibration effects. Before outlining the more complex hierarchical elements of the model however, it is helpful to outline how the inflation factors for a single expert are calculated.

Let us suppose, as per prior notation, a DM has reached out to a group of experts ($\mathbf{E}$) in order to help assess uncertainty for an unknown quantity X. Let us assume further he or she has reason to believe expert $e \in \mathbf{E}$ may be prone to some form of consistent bias which the DM wishes to remove before updating their own belief accordingly. Finally, we assume the DM has asked e to assess three quantiles, denoted by L_e, M_e and U_e, (e.g. 0.05, 0.50 and 0.95), corresponding to lower, middle and upper estimates respectively. We will outline later the impact of other choices here. The goal of the DM is to be able to transform e's responses on the tail quantiles into their unbiased counterparts L_e^* and U_e^*.

Remark: For a three quantile model, it is possible to infer inflation factors for the spread of the distribution (i.e. calculate the unbiased values L_e^* and U_e^*) or to create an inflation factor for the location parameter, $M_e^* = \beta_e M_e$, but not both. To attempt to define all three simultaneously, given only three elicited quantiles, would lead to an overspecified model, almost completely defined by the choice of priors. The original model outlined by Clemen and Lichtendahl [9] attempted to infer all three parameters and so we use a slightly different structure. Experts' miscalibration with respect to spread is typically a more meaningful metric to calculate, thus in a three parameter model we define the median estimate to be its own unbiased counterpart, i.e. $M_e^* = M_e$ and attempt to infer L_e^* and U_e^*. We will later demonstrate an extension to the parameterisation to a situation with five elicited quantiles whereby β_e can be inferred.

Rather than calculating inflation factors directly on the elicited values the bias in the spread is calculated relative to the distance from M_e^*. The theory here is that there exists multiplicative parameters α_{le} and α_{ue} such that $\alpha_{le}(M_e^* - L_e)$ and $\alpha_{ue}(U_e -$

[3] In trying to address some of the more complex elements of SEJ, such as inter-expert correlation, Bayesian approaches may suggest some procedural changes to the structure of an expert judgement study, a selection of these are outlined in [32].

M_e^*) are unbiased. α_{le} and α_{ue} are therefore scale parameters whereby a value, for either, strictly greater than 1 suggests that the expert is overconfident. L_e^* and U_e^* can thus be calculated by:

$$L_e^* = M_e^* - \alpha_{le}(M_e^* - L_e) = (1-\alpha_{le})M_e + \alpha_{le}L_e \tag{2}$$

$$U_e^* = M_e^* + \alpha_{ue}(U_e - M_e^*) = (1-\alpha_{ue})M_e + \alpha_{ue}U_e \tag{3}$$

Having established the relationship between the elicited values and their unbiased counterparts, we need to fit a model, g_e, to approximate the unbiased expert distribution $p_e(x|L_e^*, M_e^*, U_e^*)$. The model Clemen and Lichtendahl propose fits two uniform components on the intervals [L_e^*, M_e^*] and [M_e^*, U_e^*] respectively with exponential tails above U_e^* and below L_e^*. However, this choice is arbitrary and later we will outline another approach. The assumption Clemen and Lichtendahl make is that final results should be largely invariant to these modelling assumptions.

Remark. The formulation of inflation factors in this way makes the assumption that all of the training variables are on the same scale (i.e. if some variables are logarithmic in nature and others are not this would create a challenge in this approach). Wiper and French [50] rescaled judgements through the DM*'s prior to avoid assumptions on the common scale. This is something that could be assessed for our approach, however, it is outside of the scope of this paper.*

Thus, the task now becomes how to assess the unknown parameters α_{le} and α_{ue}. The core premise of calibration is that each of these variables is assumed to be constant for each expert (within the pool of seed and target variables). SEJ studies of this nature are typically one-off activities, and the seed variables are elicited in a single process alongside the target variables. If seed variables were captured longitudinally over time then experts would have the opportunity to learn and adjust and this assumption regarding constant bias will be incorrect.

From standard Bayesian theory, if $Y_1, ..., Y_{|\mathbf{Y}|} \in \mathbf{Y}$ are assumed to be random variables sufficiently similar to X, e has historically made judgements against $\mathbf{Y}$ and the DM holds the observed values $y_1, ..., y_{|\mathbf{Y}|}$, (which can be perceived to be exchangeable). Then the data set compromising $\{y_Y\}$ and $\{L_{Ye}, M_{Ye}, U_{Ye}\}$ for $Y \in \mathbf{Y}$ can be used to discover the posterior distributions for each of the unknown parameters.[4] The set $\mathbf{Y}$ is termed the set of *seed variables*. We can build a model utilising a Markov chain Monte Carlo (MCMC) method such that $\forall e \in \mathbf{E}, Y \in \mathbf{Y}$:

$$y_Y \sim g_e(\cdot|L_{Ye}, M_{Ye}, U_{Ye}, \alpha_{le}, \alpha_{ue}) \tag{4}$$

where L_{Ye}, M_{Ye}, U_{Ye} denotes expert e's elicited quantile for Y and y_Y represents the true realisation of Y, $\forall Y \in \mathbf{Y}$. After a sufficient burn in period the model can outline posterior distributions for the hyperparameters α_{le} and α_{ue} for each expert. Thus,

[4] The subscript Y here, and in all future formulas, is used to denote that these are quantiles elicited for the seed variable $Y \in \mathbf{Y}$. Similarly, from here onwards, the subscript X, denotes a variable relating to a target variable $X \in \mathbf{X}$ and the superscript $*$ denotes an unbiased value calculated post recalibration.

the DM has the scale by which he or she should adjust the expert's elicited opinions on X before updating their own belief.

This calibration approach could be calculated for each expert individually, however given potential common sources for bias across experts, expert to expert correlation should be assessed. Sources of bias common to multiple experts might include mutual experiences or identical literature reviewed. To establish potential correlation here, the model is extended hierachically to capture this behaviour. Let α_{le} be assumed to be a random draw from a gamma distribution:

$$\alpha_{le}|A_l, B_l \sim \Gamma(A_l + 1, B_l) \tag{5}$$

where hyperparameters A_l and B_l are defined by:

$$A_l \sim Pois(a_l) \;\; and \;\; B_l \sim Exp(b_l) \tag{6}$$

If we set a_l and b_l to 2; this results in a relatively diffuse positive prior, with mean near 1. This is the prior as outlined in the original paper, [9]. The gamma distribution was chosen due to its strictly positive shape and the Poisson and exponential forms were selected in order to govern the behaviour of the gamma and ensure that it was diffuse and with a suitable mean. There is no logical interpretation of the forms evident here, they were selected for their shapes. Nonetheless, this is a compelling prior to use as clearly the scale parameters must be greater than zero and we are starting from the premise that experts are likely to be calibrated.

An identical model can then be applied to α_{ue}. The complete parameterisation of this model will result in a set of hyperparameters (A_l, B_l, A_u, B_u) which capture the similarities in behaviour across experts. When there is internal structure within the set of experts, i.e. a subset of experts come from similar backgrounds or schools of thought, experts can be grouped together into what are known as homogeneity groups. Each group can then have its own set of these hyperparameters which infer group behaviour. Implementing this through a hierarchical model will result in the posterior distribution for these hyperparameters in addition to those of each experts' characteristics. The calculation of expert to expert correlations it could be argued is a significant advantage of the Bayesian approach over some of the classical SEJ models.

Experts may be subject to cognitive biases which drive grouping, in addition to their school of thought/background. In the primary definition of our model, only overconfidence bias is addressed (in the calibration step). This mitigates some of the potential issues, but is not sufficient for bias management overall within an SEJ study. We contend that the modelling exercise is not the only mechanism to manage bias. Other biases should be addressed during the elicitation process and be embedded into the facilitation protocol.

A graph of the full calibration model is outlined in Fig. 1. Each graph outlined in this paper has three components; nodes, edges and plates. The nodes represent the quantities in the statistical model. Rectangular nodes denote constants and elliptical nodes are stochastic variables. Deterministic nodes are logical functions of other nodes and will also be elliptical. An edge defined by a solid arrow indicates

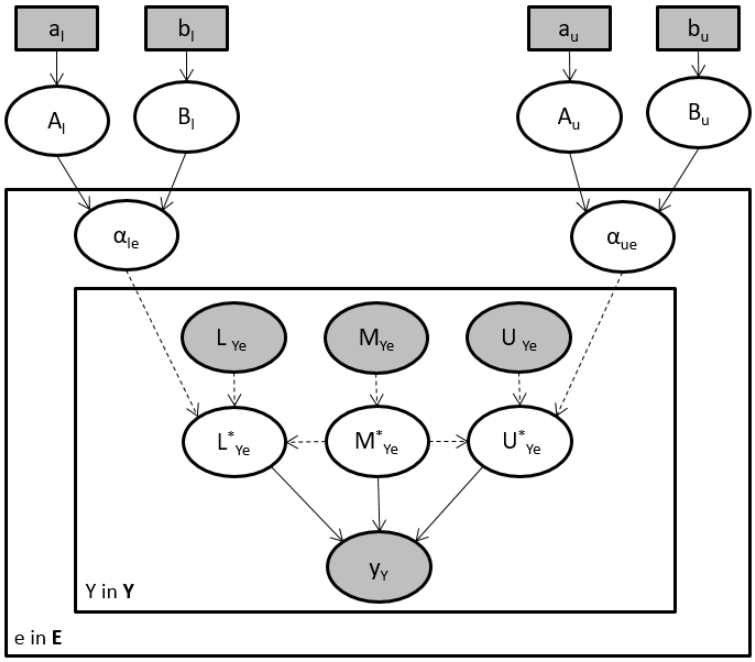

Fig. 1 Hierarchical model for expert calibration utilising standard plate notation, grey ellipses represent known values, white ellipses unknown variables and smaller squares indicate fixed model parameters. True seed variable realisations y_Y are assumed to be random draws from a distribution determined by expert e's unbiased quantiles $L^*_{Ye}, M^*_{Ye}, U^*_{Ye}$. These unbiased quantiles are a logical function of the elicited quantiles, L_{Ye}, M_{Ye}, U_{Ye} and the inflation factors α_{le}, α_{ue} calculating the expert's overconfidence.

stochastic dependence between the variables. An edge with a dotted arrow indicates a logical function. Grey shapes are known values. In order to keep the size of the graph small, repeated parts of the graph are represented using a plate (a large rectangular box). The plate will contain an index in the bottom left corner which will denote the element that is repeated.

If experts operate as coherent subjective Bayesians, certain forms of recalibration drive philosophical mathematical inconsistencies [34]. The exact form of calibration we are employing is explicitly excluded from the mathematical analysis in Kadane and Fischhoff [34]. In large data sets, such as Cooke's Delft database there is also evidence of incoherence among expert judgements, even on small numbers of elicited quantiles.

It is important to note that this formulation of expert recalibration is reliant on a certain level of mathematical coherence in expert's responses. Whilst it manages for lack of statistical accuracy it does not adjust for incoherence. If, for example, an expert had judgements in which the lower quantile is greater than the mid quantile or in the extreme case an upper quantile less than the lower quantile then this approach will not work. The authors recommend that these forms of incoherence should be

challenged and managed as part of the elicitation process rather than the modelling process.

It would be possible, if desired, to create a simpler parameterisation here. Rather than having different hyper-parameters for the upper and lower inflation factors A_l, B_l, A_u, B_u, we assume that α_{ue} and α_{le} are random draws from a single distribution determined by just two hyper-parameters, A and B. This would minimise the number of elements that need to be specified, but would put constraints on the relationship between the two inflation factors. Leaving these separate, allows for freedom in the model for these to be unrelated and potentially in opposing directions, i.e. over-confident in the lower quantile and under-confident in the upper quantile.

2.3 A model for aggregation

The traditional Bayesian approach treats the elicited information from experts as data and updates the DM's prior via Bayes formula [21]. The aggregation model, taken from Albert et al. [2], utilises a Supra-Bayesian parameter updating approach for combining indirect elicitation across multiple experts. Here we use the term indirect elicitation as rather than eliciting parameters from experts directly, experts' knowledge is elicited on more intuitive observables and the (hyper)parameters then inferred. Similar to Clemen and Lichtendahl's model, this method is generic and can be utilised with a multitude of parameterisations.

The aggregation model starts with the clustering of experts into homogeneity groups. Let us assume the experts are broken into a set of homogeneity groups $\mathbf{H}$, comprising of groups $h \in \mathbf{H}$, each of size $|h|$ such that $|\cup h| = |\mathbf{E}|, h \in \mathbf{H}$. The aim of the model will be to assess the variation both between and within these homogeneity classes. Homogeneity classes effectively use weighting to adjust for dependence between experts rather than by attempting to elicit some form of correlation structure. The ability to account for inter-expert dependence is important to ensure uncertainty is not understated and is one of the advantages of Bayesian approaches [48],[49],[32]. Selecting the right homogeneity classes is imperative. The guidance from Albert et al. is for experts within a class to be selected "corresponding to similar backgrounds or schools of thought." When this assignment is not trivial, or there are multiple potential grouping choices, a protocol for defining the groups can be useful. Later we will touch on how algorithmic approaches may be used to create $\mathbf{H}$.

Let γ_{eh} be a parameterisation, such that $g_e(\cdot|\gamma_{eh})$ represents the conceptual model about X held by expert e who is a member of homogeneity group h. Subscript eh is used from here on to denote this membership. The authors suggest the following hierarchical model to group experts:

$$\begin{aligned} \gamma_{eh} &\sim f(\cdot|\gamma_h, \rho_h) \quad \forall e \in \mathbf{E} \\ \gamma_h &\sim f(\cdot|\gamma, \rho) \quad \forall h \in \mathbf{H} \\ \gamma &\sim \pi_{DM} \end{aligned} \tag{7}$$

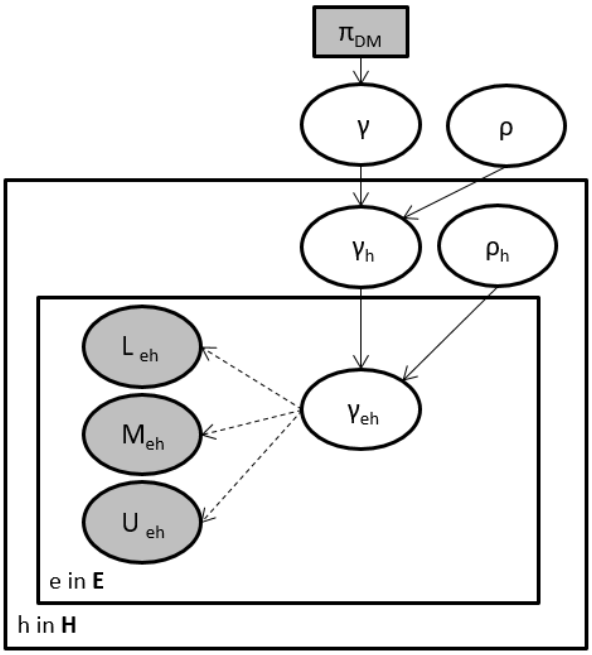

Fig. 2 Hierarchical model for expert aggregation. Experts' opinions, which are elicited as quantiles, L_{eh}, M_{eh}, U_{eh} are related, via a logical determination, to a distribution parameterised by γ_{eh}. The γ_{eh} are considered samples from homogeneity groups with location parameters γ_h. Each γ_h is in turn, drawn from an overarching distribution, with parameters γ, representing the DM's aggregate belief. Model dispersion parameters ρ_h and ρ are also random variables. The number and structure of priors for these parameters will be determined by the choice of distribution f and so for simplicity are not included in the diagram.

Here, experts within a single homogeneity group have parameters drawn from a consistent distribution $f(\cdot|\gamma_h, \rho_h)$. Each of the homogeneity groups has their parameters drawn from a single distribution $f(\cdot|\gamma, \rho)$. The ρ terms here represent dispersion parameters and the γ terms represent location parameters. *Note. the term γ, represents the output of the model, or more explicitly, the agreement of the experts given the decision maker's prior.* Fig. 2 outlines a graphical view for the aggregation model in standard plate notation. The functional form of f will be dependent on the choice of parameterisation g_e.

A simple parameterisation of (7) would be the two parameter model such that $\gamma = (\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\tau = 1/\sigma^2 > 0$. The complete model would thus be:

$$\begin{aligned} \mu_{eh}|\mu_h, \rho_h &\sim \mathcal{N}(\mu_h, \rho_h) & \frac{\tau_h}{\tau_{eh}}|\tau_h, \xi_h &\sim \Gamma(\xi_h, \xi_h) & & \\ \mu_h|\mu, \rho &\sim \mathcal{N}(\mu, \rho) & \frac{\tau}{\tau_h}|\tau, \xi &\sim \Gamma(\xi, \xi) & \forall e &\in \mathbf{E} \\ \mu &\sim \mathcal{N}(\mu_{DM}, \rho_0) & \tau^{-1} &\sim \tau_0^{-1}\Gamma(a, a) & \forall h &\in \mathbf{H} \end{aligned} \tag{8}$$

In this model, (μ, τ) represent the target consensus values; (μ_h, τ_h) are the homogeneity classes values and the (μ_{eh}, τ_{eh}) represent the parameterisation of the views of the individual experts, i.e. $g_e() = \mathcal{N}(\mu_{eh}, \tau_{eh})$. The expert level location parameters, μ_{eh} are assumed to be random draws from a normal distribution with a location parameter defined by the homogeneity group within which the expert sits. The ρ values represent the dispersion of these distributions. The ratio between the expert level dispersion parameter and the homogeneity group dispersion parameter, will be strictly positive and is determined by a gamma distribution, with parameters ξ_h. The homogeneity group parameters are linked to the global parameters in a similar way. ξ and a represent the parameters of the distribution of the global to homogeneity group dispersion ratio and the dispersion prior respectively. When we build this out further in section 3.2, we will use a split normal and thus γ will be extended to include a third parameter.

With suitably diffuse priors selected by the decision maker (specifics outlined later) the full posterior of this model can be calculated utilising Gibbs sampling. The above has demonstrated the appropriate model for aggregation of a single target variable. Fig 2. (and 8) could trivially be extended to the whole set **X** (with a plate around the whole diagram and the appropriate subscripts), as each aggregation is independent.

3 Structure of the model

To get to our final Bayesian model for structured expert judgement, as outlined in the introduction, it is critical to perform both the calibration steps and the aggregation steps on the experts' opinions. It is also necessary to give the DM a more specific approach for calculating the homogeneity groups required by these hierarchical approaches.

Let the base calibration and aggregation models be defined as above and assume all prior notation outlined remains constant. We will first turn our attention to how we would use this information to calculate the homogeneity groups and then will outline how the grouping, calibration and aggregation steps are all linked by describing the method in full.

3.1 Calculation of homogeneity groups

Suppose that two experts exist within a single homogeneity group, i.e. come from a similar background and school of thought. Given the structure of the hierarchical model, the parameters (γ_{eh}) which represent their beliefs about X are drawn from a single distribution $f(\cdot|\gamma_h, \rho_h)$. It is reasonable for the DM to suppose therefore, that the experts' beliefs about the seed variables **Y** are also drawn from similar distributions i.e. experts belong in the same homogeneity group for all elicited variables. If

we consider a single expert e that belong to one of the homogeneity groups h then this expert will be similarly grouped in the homogeneity sets for both the target and the seed variables. i.e.

$$e \in \mathbf{E} \;:\; \{e \in h, h \in \mathbf{H}\}_{\mathbf{X}} \implies \{e \in h, h \in \mathbf{H}\}_{\mathbf{Y}} \tag{9}$$

here $\{\cdot\}_{\mathbf{X}}$ refers to the set of homogeneity groups formed when considering the output variables, $\{\cdot\}_{\mathbf{Y}}$ is similarly defined for the seed variables.

Note. Assuming that the seed variables $\mathbf{Y}$ *are sufficiently similar to the output variables* $\mathbf{X}$*, (9) makes intuitive sense. The drivers of expert homogeneity groups we are trying to build, such as common schools of thought, should not vary between variables. If seed variables significantly different to the target variables are chosen, however, then an individual expert may come from many different schools of thought and therefore this assumption, and the resulting algorithmic approach would breakdown. Although, in this instance the seed variables would add little value to any* SEJ *approach that the* DM *may be using.*

Following from (9); given that $\{h \in \mathbf{H}\}_{Y_1} \equiv \{h \in \mathbf{H}\}_{Y_2} \equiv \cdots \equiv \{h \in \mathbf{H}\}_{Y_{|\mathbf{Y}|}}$, let us consider the $|\mathbf{Y}|$ dimensional space $\in \mathbb{R}^{|\mathbf{Y}|}$ formed by responses to the $|\mathbf{Y}|$ seed variables, which are linearly scaled to the unit interval. Each expert's response for the mid-quantile defines their position along the relevant axis, and therefore each expert is represented in this space by a single point $\mathbf{Y}_e$. It is reasonable to assume from here that experts from a single homogeneity group have responses clustered in some sense within this space. Thus, assigning experts to homogeneity groups simplifies to identifying clusters in the seed variable space and creating an index for each expert based on the cluster within which their seed variable responses sit.

Note. Here we are defining clusters only by expert's responses to the seed variables, $\boldsymbol{Y}$*, we do not consider their responses to the target variables,* $\boldsymbol{X}$*, when determining their homogeneity groupings. Whilst it would be in principle feasible to perform the exercise over the bigger space* $\in \mathbb{R}^{|\mathbf{Y}|+|\mathbf{X}|}$*, determined by the experts mid quantile responses to both seeds and targets, this is not recommended. As the number of dimensions increases the elements will become sparser and the clustering will be less evident. More research is recommended to understand the impact of clustering over the seed variables, the seed and the target variables or indeed the target variables alone, however this is not covered within the scope of this paper.*

Clustering is an exploratory data analysis technique and there is no single definition of what constitutes a cluster, nor how rigidly items must be allocated into these clusters. Given that experts cannot be in multiple homogeneity groups, each expert's responses must belong to exactly one cluster and therefore a *hard clustering* is needed. Furthermore, the set of homogeneity groups must be a covering of the experts, therefore we are ultimately looking for a *strict partition clustering* of the space.

One approach to defining the clusters is to do this through visual inspection. In low dimensional data sets, appropriate clustering can often be determined simply by looking at a plot of the elements in either the x, y plane or the x, y, z cube. Visual inspection is often not feasible in SEJ studies as there are often many more than three

seed variables. This creates a high dimensional space which cannot be visualised easily.

To overcome these challenges we recommend algorithmic cluster determination followed by targeted visual inspection, for validation. The algorithmic approach ensures that the full dimensionality of the data is considered, and provides a mechanism which removes as much subjectivity as possible in the cluster definitions. Visual inspection provides an opportunity for the rationale behind the clusters proposed by the algorithmic approach to be made clear. This can enhance buy-in and allow for adjustment if there is staunch disagreement or further knowledge to be embedded.

There are a multitude of algorithms that could be considered to estimate the cluster structure (and therefore the underlying homogeneity groupings). Agglomerative hierarchical clustering algorithms are an appealing method to use as they are easy to conduct pre-analysis, easy to understand and provide a nice visual way for a DM to review the clusterings that will ultimately impact the model. They have also been shown to work well over sparse data sets.

The hierarchical clustering process is an iterative algorithm which initially puts each element into its own cluster and then merges clusters together based on their distance apart and a linkage criteria. This process is repeated until all of the elements are merged into a single group. Each step in the process creates a different potential set of clusters. In order to arrive at our final strict partition a cut of the clusters which exist at one step in the process is taken. This cut can be created either visually by looking at a *dendogram* of the hierarchical clustering and considering the distance shift at each merge or by conducting the clustering using a library which considers many potential metrics and determines the cut for you, e.g. NbClust in R.

Note. The maximum number of feasible clusters is simply the number of experts present within the study. If each expert sits within their own cluster, and therefore their own homogeneity group, the middle step in the hierarchical aggregation model becomes redundant. In this instance the model can be thought of as only having two levels, an expert level and a total global level. For utilising the three levels in the hierarchical aggregation process, therefore, we need to consider groupings whereby at least two experts are combined, i.e. where the maximum number of clusters is $|\boldsymbol{E}| - 1$.

The disadvantage of a hierarchical clustering approach is that it is not possible to integrate it fully into the Bayesian model (clustering would need to be processed first and then included). In this way, we will have a two-stage method. This will result in the seed variable data being used twice and completely independently, once for clustering and once for calibration, which is unappealing.

This two-stage method therefore gives results which are an approximation to a fully Bayesian method. Using mixture models it is possible to take a further step closer to a fully Bayesian model by integrating the clustering directly into the MCMC. With sufficient data, the number of clusters can be inferred by extending to Dirichlet process mixture models (DPMM). SEJ studies are often not large enough to necessitate this method and simpler hierarchical clustering will suffice. For completeness, the fully Bayesian approach is outlined in the Results section and, for

one study, compared to the hierarchical approach, to determine how reasonable an approximation the two-stage method is.

As clustering is an exploratory process, when used post an algorithmic determination, visual inspection gives the opportunity for validation (and if necessary tweaking) of the clusters defined. It can help both in ensuring that recommended clusters are appropriate and in getting buy-in from DMs and other stakeholders to the choices made. As the seed variable space is high dimensional some processing of the data is required in order to create visuals which can be analysed easily. We recommend running a principal component analysis (PCA) over the data set to reduce dimensionality.

Principal component analysis essentially solves an eigenvalue/eigenvector problem to change the coordinate system and create new uncorrelated variables that maximise variance. In doing this the originally high dimensional space can be described as a space with a small number of meaningful dimensions, known as principal components. Each principal component captures a certain percentage of the variance between elements that existed in the original description. When applied to the SEJ data, the first few principal components can be visualised pairwise in two dimensions and the rationale for the clusterings created by the algorithms easily spotted. A scree plot which highlights the cumulative variance of the principal components can help build confidence that by visualising only this small segment of the total space a significant portion of the variance is being explained.

Once the proposed clusters are reviewed and agreed, this uniquely determines the homogeneity groups, $\mathbf{H}$, used to determine the index h assigned to each expert in the first and second line of equation (7).

3.2 Full Method

With the homogeneity group algorithm identified, it is possible to outline our full model. Building on the aggregation/calibration elements previously outlined, we can create an algorithmic approach which defines homogeneity groups, calibrates and finally aggregates. To save the reader from the simple but lengthy algebraic form, we describe the full method utilising a descriptive process. For the more masochistically minded the full algebraic and graphical forms of the model are outlined in subsections A.1, A.2 and A.3 of the appendix.

Posterior distributions for target variables are created utilising the following descriptive process:

1. **Cluster:** Rescale experts' elicited seed variable quantiles onto the unit interval. Run an agglomerative hierarchical clustering algorithm over the $|\mathbf{Y}|$ dimensional seed variable space to create a dendogram of potential homogeneity groups. Define homogeneity group assignments for each expert by creating a cut of the dendogram, either manually or through a tree cutting algorithm. Principal component analysis is conducted on the seed variable space. The first two or three

principal components are visualised pairwise and the clustering proposed by the algorithm are reviewed, discussed and approved.
2. **Calibrate:** Infer inflation factors α_{le} and α_{ue} for each expert $e \in \mathbf{E}$. Here, for each $Y \in \mathbf{Y}$ the realised value for seed variable Y is a random draw from a distribution defined by e's elicited quantiles for Y, (L_{Ye}, M_{Ye}, U_{Ye}), and the inflation factors which define e's overconfidence (α_{le} and α_{ue}). α_{le} and α_{ue} are random draws from homogeneity group hyper-parameters $(A_{lh}, B_{lh}, A_{uh}, B_{uh})$ which capture within homogeneity group inflation factor dependence.
3. **Aggregate:** For each target variable $X \in \mathbf{X}$, elicited quantiles $(L_{Xeh}, M_{Xeh}, U_{Xeh})$ for each expert $e \in E$ are assumed to be a function of unbiased quantiles $(L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh})$ adjusted according to e's inferred inflation factors (α_{le} and α_{ue}) generated in step 2. Here, h denotes the homogeneity group assignment ascribed to e in step 1. These unbiased quantiles represent specific points in expert e's underlying distribution of X, characterised by (now unbiased) parameters γ^*_{Xeh}. Parameters γ^*_{Xeh} are random draws from a distribution with location parameters γ_{Xh} defined by the homogeneity group, h, within which e sits. Homogeneity group distributions are in turn random draws from a global distribution with location parameters γ_X, which are informed by the prior belief of the decision maker, π_{DM_X}. As previously highlighted, X has been included as a subscript to all variables here to denote that they are unique to each target variable under consideration. The posterior of the global distribution γ_X represents the final aggregate for target variable X.

In practice, whilst described sequentially, both steps 2 and 3 are calculated for each step in the MCMC. Thus these two steps can be represented within a single directed acyclic graph (DAG). The full graph is outlined in Fig. 12 in the appendix. The aggregation model is updated to replace the elicited expert values with their unbiased counterparts.

Thus with the full model linked in it's generic form, it is important to define the model parameterisation we shall use within our sample cases. Here we will outline one option for model parameterisation however there are many feasible alternatives.

3.3 Model parameterisation

To parameterise our model correctly, we need first to define the generic distributions g_e and the corresponding unbiased parameters γ^*_{Xeh} $\forall X \in \mathbf{X}$ and $\forall e \in \mathbf{E}$. These choices, at an expert level, define the form of the random draw in the calibration model given in equation (4) and the first line in the aggregation step given in equation (7). Suppose, as proposed in the calibration section, that the experts have provided 3 quantiles (0.05,0.50 and 0.95) for each $X \in \mathbf{X}$ and $Y \in \mathbf{Y}$; the parameterisation we choose should preserve all of the information that the experts have provided. The natural first choice would often be a Gaussian model whereby a distribution $\mathcal{N}(\mu^*_{Xeh}, \sigma^{*2}_{Xeh})$ is selected $\forall X \in \mathbf{X}$,(or similarly for $Y \in \mathbf{Y}$) where $\mu^*_{Xeh}, \sigma^{*2}_{Xeh}$ are chosen such that $\mu^*_{Xeh} = M^*_{Xeh} = M_{Xeh}$ and σ^{*2}_{Xeh} is defined to minimise the error be-

tween the probability density function (p.d.f) at 5% and 95% and the elicited values from the expert. In this context the Gaussian model would certainly simplify the computation required, however it would make a very strong assumption that experts true beliefs are symmetric around the mid quantile (even when their elicited values are not). Often this is not the case and therefore utilising this parameterisation would immediately distort the elicited data.

A second choice for the parameterisation; as initially proposed by Clemen and Lichtendahl [9] is to model the experts' beliefs with two uniform components and exponential tails. We define the subscript Xeh to denotes values for the variable $X \in \mathbf{X}$, from expert $e \in \mathbf{E}$ who is a member of homogeneity group $h \in \mathbf{H}$ and P_L, P_M, P_U, as before, denote the probabilities which were originally elicited against. In this way:

$$g_e(x|L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}) = \begin{cases} P_L \lambda_L e^{-\lambda_L (L^*_{Xeh} - x)} & if \;\; x < L^*_{Xeh} \\ \frac{P_M - P_L}{M^*_{Xeh} - L^*_{Xeh}} & if \;\; L^*_{Xeh} \leqslant x < M^*_{Xeh} \\ \frac{P_U - P_M}{U^*_{Xeh} - M^*_{Xeh}} & if \;\; M^*_{Xeh} \leqslant x < U^*_{Xeh} \\ (1 - P_U) \lambda_U e^{-\lambda_U (x_i - U^*_{Xeh})} & if \;\; x > U^*_{Xeh} \end{cases} \tag{10}$$

Here parameters λ_L and λ_U are given by:

$$\lambda_L = (\frac{P_M - P_L}{M^*_{Xeh} - L^*_{Xeh}}) \frac{1}{P_L} \tag{11}$$

and

$$\lambda_U = (\frac{P_U - P_M}{U^*_{Xeh} - M^*_{Xeh}}) \frac{1}{1 - P_U} \tag{12}$$

This approach has a distinct advantage over the basic Gaussian parameterisation as the distribution will exactly fit the quantiles given by the expert and thus there is no loss of data. However, the uniform component puts very little mass near the central quantile suggesting that the expert gives us very little information except the range of probable outcomes (0.05 - 0.95 quantiles).

The approach that we have taken is to utilise the natural shape of the Gaussian, in which we suggest that in practice experts have a strong belief about the mid-quantile with diminishing probabilities from here, without the associated loss of information. In this way we will model utilising a split normal:

$$g_e(x|L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}) \sim \begin{cases} \frac{1}{\sigma^*_{Xleh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x - M^*_{Xeh}}{\sigma^*_{Xleh}})^2} & if x < M^*_{Xeh} \\ \frac{1}{\sigma^*_{Xueh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x - M^*_{Xeh}}{\sigma^*_{Xueh}})^2} & if x \geqslant M^*_{Xeh} \end{cases} \tag{13}$$

where, the unbiased standard deviations σ^*_{Xleh} and σ^*_{Xueh} are calculated by:

$$\sigma^*_{Xleh} = \frac{M^*_{Xeh} - L^*_{Xeh}}{\delta_1} \;\; and \;\; \sigma^*_{Xueh} = \frac{U^*_{Xeh} - M^*_{Xeh}}{\delta_2} \tag{14}$$

Here δ_i represents the number of standard deviations between the elicited quantiles. Clearly if the probabilities for L and U are symmetric around M then $\delta_1 = \delta_2$. In the case 0.05, 0.5 and 0.95 then $\delta_1 = \delta_2 \cong 1.64$. τ^*_{Xleh} and τ^*_{Xueh} follow directly from these assignments and are then hierarchically calculated as per (8). The final parameterisations are given by $\gamma^*_{Xeh} = (M^*_{Xeh}, \tau^*_{Xleh}, \tau^*_{Xueh})$. With this model the decision maker location prior will be M_{DM} rather than μ_{DM}.

The formulation of the split normal in this way will not result in a fully continuous distribution as at the mid-quantile there is a step. It would be trivial to adjust for this simply by factoring each half of the distribution, however, the result of this would be a shift in the median point, which is an unappealing result given the way we have defined calibration. Given that it does not largely affect the complexity of the modelling we leave this point of slight discontinuity.

Remark. this model is only one of many approaches which could be taken, it would be interesting, although not covered within this paper, to assess the impact of non-Gaussian parameterisations on the final output.

To build our model we have used the JAGs package embedded within R. JAGs, built on the BUGs language, works using a Gibbs Sampling approach to MCMC, and in this combination allows relatively efficient calculations of results. Complete model runs typically take circa 10 minutes to complete 100,000 iterations. Modelling has been run in RStudio with R version 3.6.1, JAGs version 4-10 on an AMD Ryzen 7 PRO 3700U processor, with 4 cores, 8 logical processors and 16.4GBs of virtual memory. These specifications are for a standard laptop. Timing was provided based on modelling only on a single core. Modelling was implemented in this way to mimic hardware available to study analysts whilst conducting an SEJ study live. If run-time is a concern, significant improvements can be made utilising a multi-threaded version of the code and deploying in a virtual environment with many cores.

4 Analysis of results

In this section we assess the effectiveness of our combined model by running it against empirical studies from Cooke's database and comparing the resultant target variable forecasts to the more commonly utilised opinion pooling techniques. The results on two studies will be outlined. Studies have been chosen at random from a subset in which all forecasts are on a uniform scale.

The first study considered (Arkansas) formed part of a broader study conducted by the Centre for Disease Dynamics, Economics and Policy looking at grant effectiveness and child health insurance enrolment for the Robert Wood Johnson Foundation. The second case study we will review (CWD) was conducted for the University of Ottowa to assess infection transmission risks for chronic wasting disease (cwd) from deer to humans ([46]; [47]).

Whilst in principle we may wish to encode DM knowledge into the analysis, these studies have been conducted in the rational scientist context outlined in EFSA [18].

In this context, relatively uninformative priors are appropriate, outlined below, and a consistent set of priors can be used for both studies.

The rational scientist context is a common one in SEJ. Here, rather than modelling the impact of elicited expert judgements to an individual DM's belief, the aim is to combine the judgements in a way that represents what a rational scientist would believe given the experts' inputs. Typically, all of the knowledge for that hypothetical rational scientist should be encapsulated within the experts' judgements and hence relatively flat (and thus uninformative) priors are desired.

One of the advantages of modelling expert judgement in a Bayesian way is that this can easily be done, but, if DM's do have prior belief they wish to embed in the model the mechanisms inherently exist to do this. In many deployments of Bayesian models the modelling process can be quite distinct from DMs and model priors are defined by the analysts performing the work. It is strongly recommended for SEJ that this approach, for target variable priors, unless in a rational scientist context, is avoided.

If any prior knowledge on the target variables needs to be incorporated into the model, this should be elicited directly from the DM. This can be difficult, particularly as studies often run at a distance from decision makers, but the authors contend that it is an essential step in deploying a Bayesian study correctly. The process of eliciting DM prior belief can operate as a mechanism to increase traction with stakeholders, help facilitators understand the sensitivity of the decision to judgements given and ultimately increase the chance of study outputs being utilised. Whilst admittedly challenging, the authors have successfully elicited prior beliefs from many decision makers in the past.

Incorporation of decision maker belief, often happens when SEJ is deployed in private enterprise. Public sector SEJ studies often take the rational scientist view outlined. As such, publicly available data-sets are often in this context, which, given the desire to compare to existing models, is the reason for the two case studies outlined. SEJ is more commonly found in the public sector than the private sector, although the authors would like to see greater traction for these methods in private enterprises.

In a typical study there are often very few experts and elicited quantiles. The full set of hyperparameters, $(a_l,b_l,a_u,b_u,M_{DM},\rho_0,\rho,\rho_h,\xi,\xi_h,\tau_0,a)$ consequently must be modelled carefully as their influence will be important.

The calibration component parameters $(A_{lh},B_{lh},A_{uh},B_{uh})$ will be set, as per (6), such that $A_{lh} \sim Pois(a_l)$ and $B_{lh} \sim Exp(b_l)$ (and equivalent for the upper bounds) and $a_l = b_l = a_u = b_u = 2\ \forall h \in \mathbf{H}$. This provides a suitably diffuse prior, centred around 1, for the dispersion parameters of the calibration model. This is identical to the parameter values first proposed by Clemen and Lichtendahl [9].

The aggregation component hyperparameters $(M_{DM},\rho_0,\rho,\rho_h,\xi,\xi_h,\tau_0,a)$ are set with weakly informative priors.

Remark. given our split normal parameterisation there are actually hyperparameters for both the upper and lower model dispersion parameters, thus, there is effectively a ξ_u and ξ_l, and equivalent. In practice we set these hyperparameters to be identical, therefore for simplicity this distinction is omitted below.

Similar to Albert et al. [2], we shall apply truncated Normal priors for the dispersion components as this is a useful simplification of the folded noncentral-t distribution which is conjugate for these parameters, as shown by [24] :

$$\sqrt{\rho_0} \sim \mathscr{N}_+(0,\varphi_0); \quad \sqrt{\rho} \sim \mathscr{N}_+(0,\varphi); \quad \sqrt{\rho_h} \sim \mathscr{N}_+(0,\varphi_h); \tag{15}$$

where $\varphi_0, \varphi, \varphi_h$ are selected to be weakly informative. Given the uniform scales of the examples that we are looking at, and the fact that in practice to ease modelling we normalise everything onto [0,1] (and then readjust back to the original scale at the end), setting $\varphi_0 = \varphi = \varphi_h = 1000$, provides a sufficiently diffuse prior for these hyperparameters. Similarly, we can switch the prior on the variance component τ^{-1}, from a gamma ($\tau_0^{-1}\Gamma(a,a)$) to a truncated normal prior $\sqrt{\tau^{-1}} \sim \mathscr{N}_+(0,\psi_0)$ where ψ_0 is selected in order to be reasonably diffuse, here we also select $\psi_0 = 1000$. For the intermediary hyperparameters, ξ, ξ_h, we stick with a gamma prior, as outlined in the orignal model, however, we do not use completely diffuse priors and set $\xi = \xi_h = 1.5$. As Gelman [24] highlighted, utilising very small components for the terms in the gamma distribution puts a lot of the mass at zero, which for this model is unfavourable. Utilising the above structure, allows us to set a prior which has more of the mass centred at 1, building the assumption that there is similarity in intra homogeneity group dispersion parameters, whilst being sufficiently diffuse enough to learn the true nature of these intra group dispersion relationships from the data.

With these hyperparameters set, we review the results for each of the studies outlined earlier and compare these to opinion pooling methods. It is important to note here that the comparison is not aimed at showing superiority of our method compared to the other methods. We merely wish to better understand how the hierarchical modelling, and the focus on the consensus of experts, drives a differing perspective to the opinion pooling methods. Specifically, the two methods which we will compare to are an equal weighted linear opinion pool of the form (1) where $\omega_e = 1/|\mathbf{E}|$ (with a DM referred to as EWDM) and a performance weighted linear opinion pool where ω_e is defined by performance over the seed variables and is determined by Cooke's Classical model, outlined previously (PWDM).

4.1 Arkansas Example

The Arkansas study, originally conducted in 2012, had 4 experts who were required to assess 10 seed variables and 20 target variables. An example of the seed questions utilised (with the values known *a priori*) is "What is the ratio between the number of children without health insurance in Arkansas / number of children without health insurance in Louisiana?" with a true realisation of 0.66. The target questions were of a similar nature; e.g. "What would the participation rate for public insurance be in 2020 if CHIPRA were not renewed in 2013?". Here CHIPRA refers to the Children's Health Insurance Program Reauthorization Act, which was signed into action by President Obama February 4th 2009. All of the data was elicited against 5 quantiles

(0.05, 0.25, 0.5, 0.75, 0.95), although given the above parameterisation we will only utilise 3 of these within our model.

The first component to analyse is the clustering component (the outputs of which are then embedded within the broader calibration and aggregation model). Running our agglomerative hierarchical clustering approach over the seed variable space results in a proposal to split the experts in to three homogeneity groups. The dendogram for the hierarchical clustering can be seen in Fig. 14 in the appendix. Expert 1 and expert 4 sit within their own groups and experts 2 and 3 are clustered within a single homogeneity group.

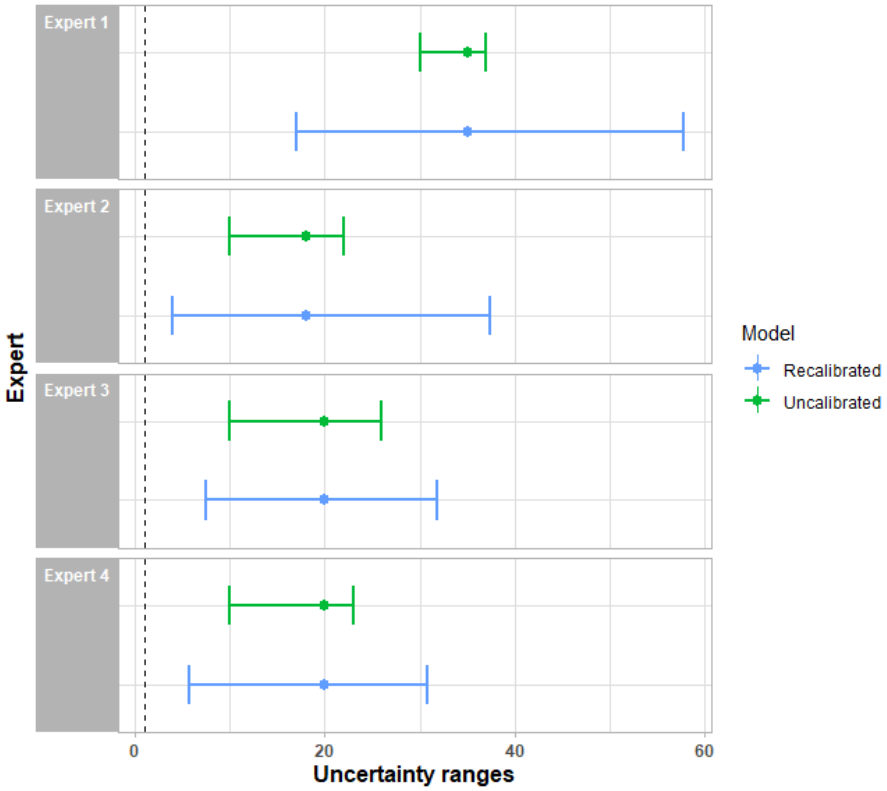

Fig. 3 Effect of recalibration on experts' estimates within the Arkansas study on the question: "What would the adolescent well-care visit rate be in 2020 without RWJF Covering Kids and if CHIPRA were not renewed in 2013?" All experts show a significant broadening of their quantiles as a result of the recalibration, particularly expert 1.

Following homogeneity group definition, we assess the impact of calibration. A very simple pre-analysis of the data suggests the experts are not well calibrated, with a significant bias towards overconfidence. If all of the experts were statistically accurate we would expect that circa 4 of the 40 seed variable estimations across the

experts (10%) would be outside of the experts' elicited quantiles. In practice 58% of the realisations fell above the experts' 0.95 quantile or below the 0.05 quantile. This ranged from 80% for expert 1 to 30% of assessments for expert 3. To this extent, when we review the calibration parameters in the Bayesian model we would expect these to compensate for this behaviour and increase the expected uncertainty versus what was outlined by the experts. The medians for the experts' miscalibration parameters range from 1.3-3.6 for a_{le}, and from 2.0-11.4 for a_{ue}. All of these values are greater than 1, indicating systemic over-confidence rather than under-confidence, aligned to the expectations from the pre-analysis. Expert 1, sits at the upper end of this range for both variables. The scale of these calibration parameters suggests the experts' forecasts are significantly over confident. The DM should consequently be very careful when assessing whether the originally elicited judgements from these experts give a true picture of the uncertainty present.

Focussing on experts' forecasts for a single target variable, Fig. 3 outlines the impact this recalibration has to the estimates used in the model. Variable 10 ("What would the adolescent well-care visit rate be in 2020 without Robin Wood Johnson Foundation (RWJF) Covering Kids and if CHIPRA were not renewed in 2013?"), is chosen here to demonstrate the difference in output for the three different modelling types. Of particular note, is expert 1 whom despite being very confident in their response initially, once recalibrated, as expected, display a much broader distribution.

It is interesting to also note the underlying structure of the responses to this variable. In the original forecasts expert 1 estimated with high certainty the true value will be greater than 30 whereas the other three experts estimate the median of this variable at circa 20 (and strictly less than 30). The recalibration exercise has shrunk this discrepancy. The recalibrated judgement for expert 1 now overlaps considerably with the other three experts. This suggests that there may not be such a stark underlying difference, aligning with our aim to identify consistency between judgements. We will return to this later when we assess the output distributions.

Finally, we focus on the aggregation component and the posterior distribution for the DM. One element to be considered when building this final posterior distribution is how to combine the posteriors of the components (M, τ_l and τ_u) into a single output. Within the initial MCMC these components are modelled separately; in each run we have a posterior distribution for each but no combined distribution. We create this combination by applying a secondary Monte-Carlo analysis drawing triplets from each distribution (here we actually use samples from the original model, post convergence) which we fit back to our split normal structure. We sample from this to give our combined posterior.

Reviewing all target variables in the study; Table 1 outlines the resultant posterior and how this compares to the estimates for the EWDM and the PWDM. (Please note. Variables 14 and 15 have been removed from this list as not all experts predicted these two target variables). It is clear from the data that there are similarities between the approaches in the mid-quantile assessment with the Bayesian decision maker (BDM), having a mean difference of 1.5% from the EWDM and 2.2% from the PWDM, (with mean absolute differences of 4.6% and 14.8% respectively). For the outer quantiles (0.05,0.95) the BDM model produces much wider final distributions

than either of the other two models. Here the BDM suggests there is a significant probability that the true value of these realisations will lie significantly below or above the estimates of either the PWDM or the EWDM. This should not be surprising, given the calibration data reviewed earlier, however, it is important to review the full distributional forms rather than just the quantiles to understand the impact of these fluctuations.

Rather than review all of the distributions in detail, we will examine the distribution for the tenth target variable[5], which was outlined earlier. Variable 10 was selected as there was notable discrepancy between the EWDM and the PWDM distributions. This gives us an opportunity to understand how the BDM model compares to other models in cases where there is more complication in the underlying data structure. Distributions for all remaining target variables have been included in Fig. 15 in the appendix.

The EWDM distribution in Fig. 4, is multimodal, aligned to the calibration point plot shared earlier. There is more mass under the first peak reflecting the lower assessments provided by three of the experts. For the PWDM we see a unimodal dis-

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Variable 1	51.9	74	97.5	56.2	95.4	97.9	46.2	75.6	95.3
Variable 2	62.1	92.5	99.5	85.5	96	99.6	58.7	87.9	98.7
Variable 3	54	76.5	98.6	56.2	94.3	97.9	51.2	79.1	96.8
Variable 4	35.4	84.4	95.6	54.3	92.9	96	39.2	73.7	95.6
Variable 5	32.1	70.1	95.4	36.6	92.8	95.9	33.8	67.2	92.1
Variable 6	10.2	17.5	25.2	10.3	20	25.9	6.6	18.5	30
Variable 7	15.9	28	38.1	16.8	26.2	37.5	12.9	29	44.1
Variable 8	11.9	26.2	38.1	10.2	21.1	37.3	10.3	28.1	45.6
Variable 9	11	23.2	36.4	12.5	25	30	10	25.7	41.9
Variable 10	10.4	20.8	36.4	10	19.1	26	7.8	23.8	41.6
Variable 11	5.9	16.1	30.4	5	18.1	32.6	-0.5	16.7	37.9
Variable 12	24.7	51.9	67.9	20.6	52.1	67.8	20.9	52.5	88.9
Variable 13	15.2	49	66.3	10.6	35.6	58.9	17.8	49.2	84.3
Variable 16	39.5	71	84.4	35.5	65.8	84.6	45	72.7	93.7
Variable 17	80.9	90.5	98.6	80.3	92.5	99	77.9	90.5	98.3
Variable 18	50.6	87.3	94.5	45.7	72.8	92.4	60.8	85.9	97.5
Variable 19	67.6	89.1	96.2	75.1	87.6	96.5	66.2	85.8	97.2
Variable 20	45.9	81.2	90.8	40.7	70.8	86.9	52.9	80.9	96

Table 1 Comparison of DM quantiles for different modelling approaches to the Arkansas Study.

[5] From the nature of the questions asked within the study, some of the variables are bounded, i.e. the output is a % that must be between 0 and 1. Without intervention, the BDM in these cases may produce a posterior distribution that sits outside of these bounds as we do not constrain the model in formulation. Thanks to the constant in the Bayesian formula, we can simply do this by applying a further prior which is uniform on the unit interval (and zero outside of this) and then rescaling the posterior as necessary. All of the values in this study, when comparing to the other modelling types, have been adjusted accordingly. Another mechanism for imposing bounds, considering the generic model outlined, would be to select a parameterisation, such as a beta distribution, which constrains the bounds by default.

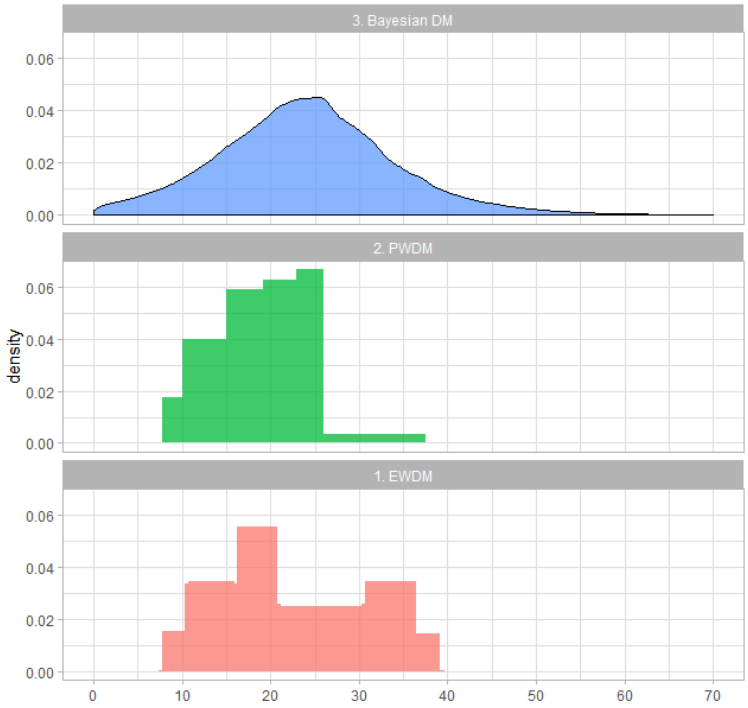

Fig. 4 Comparison of final distributions to the question "What would the adolescent well-care visit rate be in 2020 without RWJF Covering Kids and if CHIPRA were not renewed in 2013?" The Bayesian model (blue) demonstrates a larger support, aligned to the overconfidence demonstrated by experts in the seed variables.

tribution aligned to the lower estimations. This is driven by the relative weighting of Cooke's model which determined that 23% of the performance weighting should be assigned to expert 2 and 77% of the weight assigned to expert 3. Expert 1 and expert 4 were eliminated. Thus, the final distribution is dominated by experts who had a lower estimate of the variable as expert 1's judgements are not included. By design, as we are assessing the consistency in opinion, the Bayesian model also has a unimodal posterior. Evident from Fig. 4 is also the heavier BDM tails relative to the other two models. The mode of the distribution, unsurprisingly, sits between the two peaks of the EWDM, however is skewed slightly to the lower peak. This is aligned to the calibration behaviour shown earlier and the relative homogeneity group weightings in the model. In this way, the Bayesian model smooths the variabilities between the experts, whilst still modelling the underlying differences in the estimations.

The beauty of the Bayesian model is that even though the homogeneity group structure has been compressed into this final posterior distribution, it is still possible to learn about the underlying model behaviour. We can examine the posterior homogeneity group parameters which can easily be recovered in addition to the

complete posterior DM distribution. Fig. 5 outlines the homogeneity group distributions. Group 1 comprises of just expert 1, group 2 includes expert 2 and expert 3, and group 3 is just expert 4. The rationale for the skew in the Bayesian model towards the lower peak of the EWDM is evident here. Group 1 has a distinctly different distribution to groups 2 and 3, driven by expert 1's differentiated belief compared to their peers. The final Bayesian DM distribution is weighted more to the common belief demonstrated in group 1 and group 2, but less so than if the experts had just been aggregated directly. In this way, the differentiated perspective of expert 1 has had increased weight. This is one of the advantages of the Bayesian model, the model design and software implementation facilitates a deepdive of the results, beyond just the final distributions, to support the DM decision making.

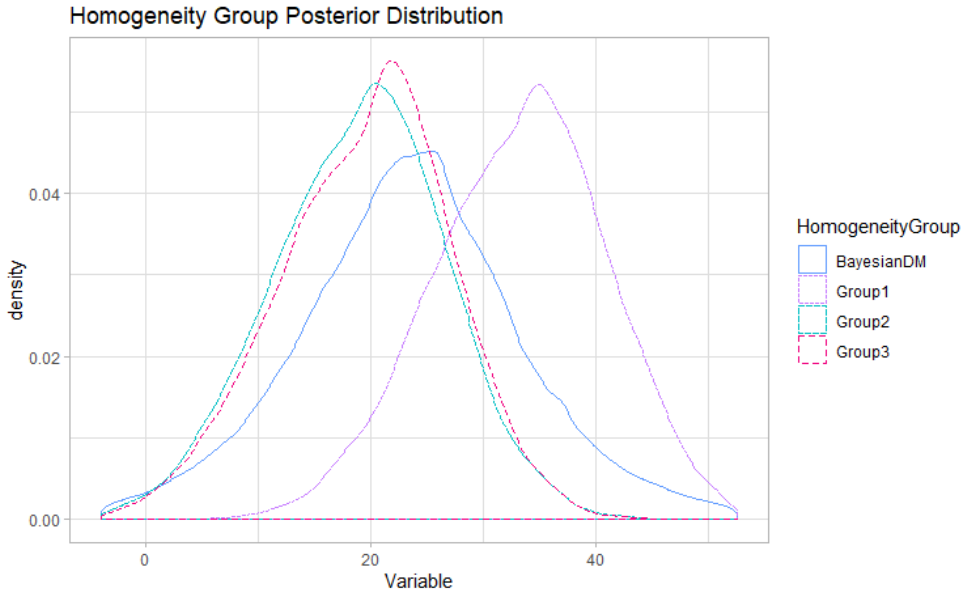

Fig. 5 Comparison of homogeneity group distributions to the question "What would the adolescent well-care visit rate be in 2020 without RWJF Covering Kids and if CHIPRA were not renewed in 2013?" Group 1 = {expert 1}, Group 2 ={expert 2, expert 3} and Group 3 = {expert 4}.

4.2 The impact of our elicited quantile choice

The Arkansas study originally elicited 5 quantiles (0.05, 0.25, 0.5, 0.75 and 0.95) for both the target and seed variables. Up to now, given the parameterisation choice outlined earlier, we have only been considering three of these (0.05, 0.5, 0.95) thereby effectively discarding some of the elicited information. Utilising the remaining data

points and different modelling approaches allows us to determine the impact of our quantile choice and inform more efficient elicitation in the future. There are many possible combinations of quantiles which we could use to either inform the calibration inflation factors or the target variable aggregations. We have prioritised four combinations for further analysis and compare these to our original model structure:

- ***OuterQuantiles***: Our original parameterisation: quantiles 0.05, 0.5 and 0.95 are used to determine two inflation factors α_{le} and α_{ue}. The same quantiles are used for aggregation on the target variables.
- ***InnerQuantiles***: We still use three quantiles for determining α_{le} and α_{ue} but now consider the elicited quartiles 0.25, 0.5 and 0.75. The same quantiles are used for aggregation on the target variables.
- ***AllFiveQuantiles***: Use all five elicited quantiles for determining α_{le} and α_{ue}, effectively giving more power to the calibration model. Only the OuterQuantiles are used for aggregation on the target variables.
- ***BetaOuterQuantiles***: All five elicited quantiles are used for calibration. α_{le} and α_{ue} are calculated as before but we extend the model to allow positional uncertainty in the median by creating an inflation factor β_e such that $M_e^* = \beta_e M_e$. Outer quantiles (0.05, 0.5, 0.95) are used for aggregation on the target variables.
- ***BetaInnerQuantiles***:Identical to the BetaOuterQuantiles except the inner quantiles (0.25, 0.5, 0.75) are used for aggregation of target variables.

Applying these five model parameterisation types, to the target variable outlined earlier, results in different posterior distributions for the DM as demonstrated in Fig. 6. What is striking in Fig. 6 is that whilst there are some minor differences between the posteriors of each approach (particularly in the tails) these are relatively insubstantial. It is reasonable to assert that a DM is unlikely to make a substantively different decision regardless of the parameterisation choice used. We apply a Kolmogorov–Smirnov test to 10000 samples from each distribution and consider relative to the original model (OuterQuantiles) to assess the mathematical difference between the cumulative distribution functions (c.d.fs), Table 2. The p-value in this test demonstrates that in all cases the posterior distributions are with very high likelihood not the same (or strictly speaking, they are not both samples from an identical underlying distribution). Utilising the D statistic from the test does demonstrate however, that whilst they are not identical distributions they are very similar. The D-statistic can be interpreted as the maximum distance between the two tested c.d.fs. Thus if we consider the two parameterisations which do not have a beta term (InnerQuantiles and AllFiveQuantiles), the maximum distance between either of these distributions and our original parameterisation is circa 4%. Even if we assess calibration by placing an inflation factor on the median estimate (BetaOuterQuantiles and BetaMidQuantiles) we still do not generate massively different distributions. The maximum distance here relative to the original model is circa. 9%. If we discard the tails and apply a Kolmogorov-Smirnov (KS) test to the bulk of the distributions (samples in the 25th-75th percentiles), these numbers drop further to $< 2\%$ and $< 7\%$ respectively.

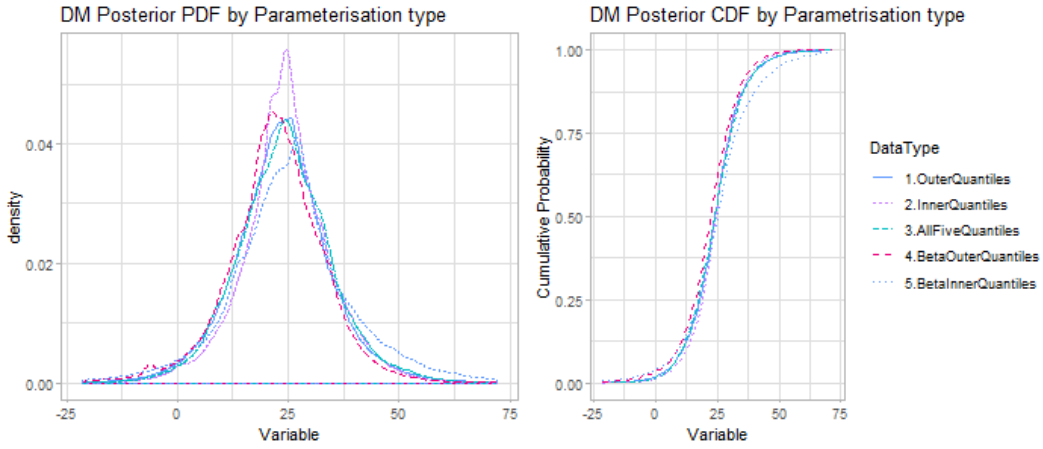

Fig. 6 Comparison of the posterior DM distributions by parameterisation type for target variable 10 in the Arkansas study. The choice of parameterisation has had little impact to the overall uncertainty.

	InnerQuantiles	AllFiveQuantiles	BetaOuterQuantiles	BetaMidQuantiles
D Statistic	0.0395	0.0248	0.0852	0.0791
p-value	$< 1*10^{-9}$	$<1*10^{-9}$	$<1*10^{-9}$	$<1*10^{-9}$

Table 2 Kolmogorov-Smirnov test on the impact of quantile parameterisation for target variable 10 in the Arkansas Study.

Extending this further to consider all of the target variables within this study (Table 3 and Fig. 16-32 in the appendix), we can see that this distributional similarity is very consistent across variables. Indeed the maximum difference between any different parameterisation choice and our original model in any target variable at any point in the c.d.f is circa 15%. This is an incredibly small difference given the substantively different data sets, and calibration parameterisation choices used to generate these distributions. Thus, for the Arkansas study, we have seen posterior consistency when considering different parameterisations of the calibration model and number of elicited quantiles being modelled.

This is only a single study, and significant empirical analysis would be required to demonstrate that this behaviour applies consistently. However, it is a very appealing result nonetheless. Firstly, this suggests the modelling approach we have outlined is robust and relatively consistent regardless of the quantile choice made. This builds confidence that the model is identifying underlying behaviour without being overly sensitive to our input choices. Secondly, if our choice of quantiles does not impact the robustness of the output, DMs can make a choice prior to elicitation on which quantiles they would like their experts to inform on. The cost of an elicitation exercise increases as the number of elicited data points increases. Eliciting more invariably takes more time from both facilitators and experts. Thus, if this model gives you consistently similar results regardless of whether you elicit three

Target Variable	InnerQuantiles	AllFiveQuantiles	BetaOuterQuantiles	BetaMidQuantiles
Variable 1	0.1004	0.0305	0.0456	0.0907
Variable 2	0.079	0.046	0.0631	0.0392
Variable 3	0.0586	0.0364	0.0473	0.0966
Variable 4	0.0378	0.0695	0.0767	0.067
Variable 5	0.0386	0.0311	0.0384	0.0838
Variable 6	0.0927	0.0308	0.0572	0.0963
Variable 7	0.1088	0.038	0.0142	0.0476
Variable 8	0.0553	0.019	0.0496	0.0759
Variable 9	0.0841	0.024	0.0336	0.0698
Variable 10	0.0395	0.0248	0.0852	0.0791
Variable 11	0.1305	0.0291	0.0142	0.0305
Variable 12	0.1589	0.0148	0.0167	0.0936
Variable 13	0.0839	0.0493	0.0466	0.1195
Variable 16	0.1595	0.0404	0.0395	0.1027
Variable 17	0.1258	0.0214	0.018	0.0799
Variable 18	0.0892	0.0406	0.0422	0.0889
Variable 19	0.0915	0.0348	0.0631	0.0945
Variable 20	0.0801	0.0259	0.0117	0.0953

Table 3 D Statistic from a Kolmogorov-Smirnov test on the impact of quantile parameterisation across all Target Variables in the Arkansas Study.

or five quantiles, you can make your elicitation exercise more effective by eliciting less, without loss on the output. Another way this behaviour could be beneficial is that experts may favour elicitation in different ways, i.e. some experts wish to give their judgements on outer quantiles and others on inner quantiles. Given the model provides consistent outputs, an elicitor could tailor the elicitation exercise to each individual expert and then place these mixed quantiles into the model. The model could then adjust and standardise internally as required without loss of precision.

4.2.1 All-in-one-method

The full modelling method outlined so far relies on two steps, a homogeneity group assignment step, followed by a combined calibration/aggregation step. In the interest of moving to a fully Bayesian method, an all-in-one method which links these three elements completely is desired. As described earlier, one approach to this is to consider mixture models for clustering over the seed variable space. We outline such a model and apply it to the Arkansas study here for two reasons. Firstly to assess how reasonable an approximation the two-step method is and secondly to address the suitability and challenges of modelling in this way.

Rather than use a standard mixture model to do our clustering, we utilise a Dirichlet process mixture model. A Dirichlet process is used here as not only the expert assignments but also the number of homogeneity groups may be unknown to the DM. The non-parametric nature of the Dirichlet process mixture model allows the

DM to simply define a tuning parameter, o, based on an expected maximum number of homogeneity groups. The identified number of groups is then an output of the model.

The Dirichlet process mixture can be described by the following generative process, [6].

1. Generate a set of mixing weights, ν where $\nu = \{\nu_k\}_{k\in 1:|E|}$ according to a stick breaking process dependent on tuning parameter o.
2. Generate a set of parameters θ where $\theta = \{\theta_k\}_{k\in 1:|E|}$ for each potential cluster k, according to a prior distribution with parameters θ_0.
3. For each observation in the seed variable space $\mathbf{Y}_e$, assign a component label, h, according to the mixing proportion ν.
4. Generate $\mathbf{Y}_e$ according to the h^{th} component of θ.

When the model is a mixture of Gaussians, θ is defined as a mean and a precision matrix and the prior distribution on θ is modelled as a normal-Wishart. The model can be written algebraically as:

$$\mathbf{Y}_e \sim \mathcal{N}(\theta_h) \tag{16}$$
$$h \sim Cat(\nu) \tag{17}$$
$$\theta_k \sim Normal - Wishart(\theta_0) \tag{18}$$
$$\nu \sim GEM(\mathrm{o}) \tag{19}$$

where GEM denotes the stick break process, [43]. The DAG of this process is outlined in Fig. 13 in the appendix.

The DPMM can be integrated into the linked calibration/aggregation model by utilising the homogeneity group assignment h for expert e in the calibration and aggregation element, when $\mathbf{Y}_e$ is assigned to cluster h. Rather than being calculated *a priori*, clusterings are then defined at each step within the global MCMC.

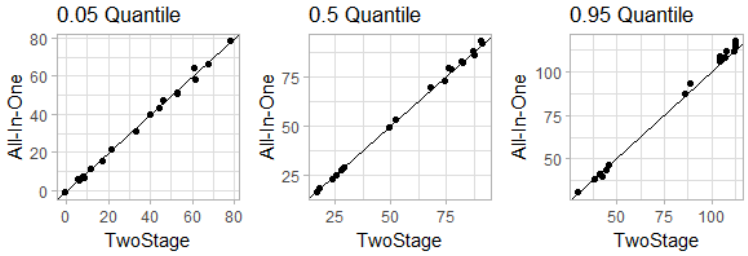

Fig. 7 Comparison of the posterior DM Quantiles between the one-stage and two-stage method

We compare this all-in-one method to the two step method, with hierarchical clustering, outlined earlier by generating posterior distributions for all target variables in the Arkansas data. Running the all-in-one method takes considerably longer. On the same machine, modelling takes approximately 10 times as long.

Final results obtained using the two methods were extremely similar. Fig. 7 outlines the relationship between the quantiles of the final posterior distribution for all of the target variables in both the all-in-one method and the two-stage method. In each case the line $x = y$ is plotted. Across all three variables we see very tight clustering of the results around these lines. The DM posterior has been largely invariant to the change in methodology. This is reassuring as it further builds the case for the robustness of the method, whilst providing options for analysts on how to perform the modelling based on both practicality (time needed to model, availability of data) and theory (modelling all-in-one allows us to avoid having to reuse the data).

The mixture model for clustering will work best when we have studies with a significant number of experts and seed/target variables in them, as we now have a very tangible increase in the number of model parameters. Methods for operating on a large scale are becoming more relevant due to mass participation expert judgement studies conducted over the internet, [22], however, these are still the minority. As such, for most studies today, often the clustering space will be very sparse and potentially high dimensional. In such cases, principal component analysis (PCA) could be run, here as part of the clustering process itself, to reduce dimensionality ahead of cluster identification if an all-in-one method is considered and convergence is an issue. The analysis here however, suggests that the two-step method is a reasonable approximation and will likely be more appropriate in the short term. The issues with an all-in-one method overall are significant increase in model complexity (and consequently stability), the need for more data and risk that final posterior distributions become exceptionally diffuse with the integration of more areas of uncertainty, thereby reducing the value for the decision maker.

Whilst costs of this all-in-one method may outweigh the theoretical benefits today, as bigger studies are undertaken this balance is likely to shift. Application of this method to a mass participation expert judgement study is likely to test the enhanced efficacy it can bring. Given the scale/cost required to implement mass-participation studies, it would probably be best to integrate this test into a study conducted for other purposes, rather than designing a test study explicitly. This is left to further research.

4.3 CWD

The CWD study was outside of the health insurance domain, instead looking at the transmission risks for chronic wasting disease from deer to humans. The study was comprised of 10 seed variables and 13 target variables. With 14 participants, this study had significantly more experts present than the Arkansas study. These experts were separated into 5 homogeneity groups by the model. Three experts (1,4 and 10) were placed in individual groups as they demonstrated consistently different behaviour over the seed variables than their counterparts. The model breaks the remaining experts into the following two sets, $\{2,6,8,9,11,12\}$ and $\{3,5,7,13,14\}$. The separation of the individual groups is evident in a simple PCA plot of the first

two components of the seed variable space, demonstrated in Fig. 8. Even within just these two components over half (∼54%) of the variability in the seed variable space can be explained.

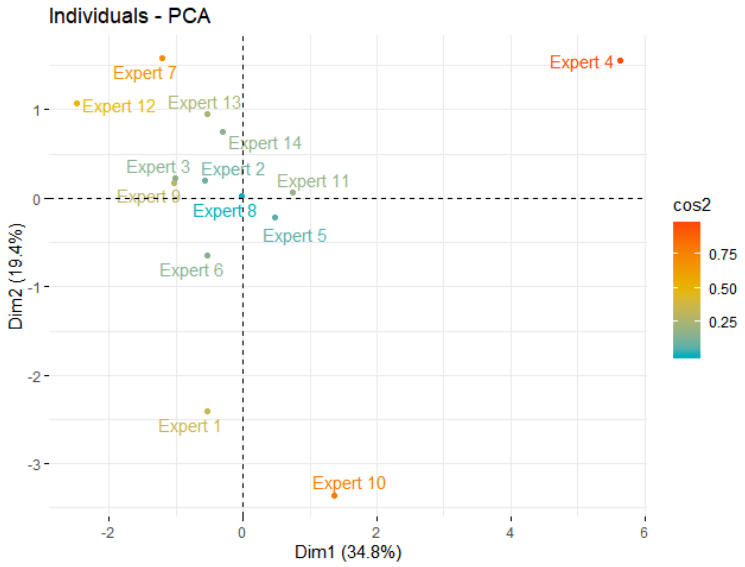

Fig. 8 PCA plot of first two components for the seed variable space. Note the separation of experts 1,4 and 10 from the bulk of the remainder. These experts resultantly sit in their own homogeneity groups.

Whilst the three groups of individual experts are visible from the PCA plot of the first two principal components, there is no apparent logical seperation of the remaining experts. A split into two further groups is not readily visible. The third principal component of the PCA captures a further 15.4% of the variance within the seed variable space. As outlined in Fig. 9, plotting the third principal component pairwise versus the first, the groups produced by the model become readily apparent.

Whilst the number of experts, the number of questions and the domain of the study in this case are very different to the Arkansas case, the statistical accuracy of the experts is similar. Here over 51% of judgements sit outside of the 5th to 95th quantile bounds. In the most extreme cases, 2 experts had 80% of their judgements outside of these bounds. This level of miscalibration suggests systemic overconfidence. Thus as anticipated in the majority of cases the Bayesian model has significantly broader tails across the aggregate target variables than the two opinion pooling methods. This behaviour can be seen in a plot of the distributions for each target variable, captured in Fig. 10. Table 4 in the appendix outlines the standard quantiles, by target variable, associated with each of these distributions.

One case in this study where we see different behaviour to any of the variables in the Arkansas study, is where experts predict significant imbalance in the upper

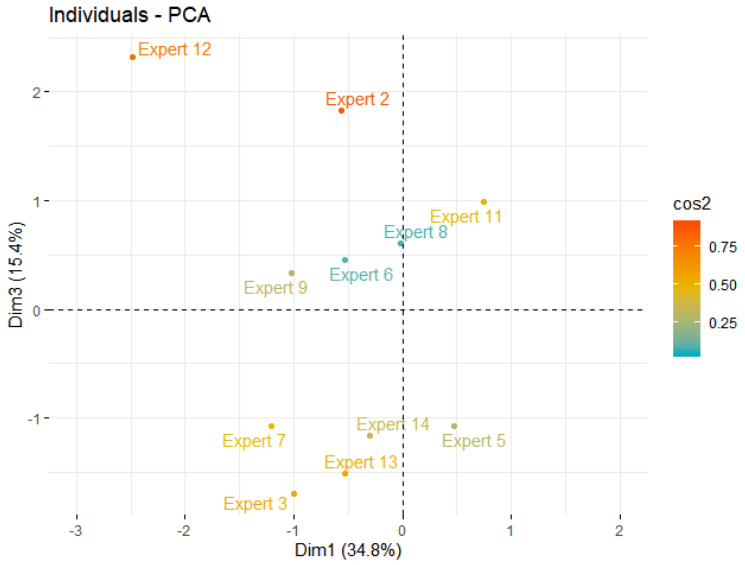

Fig. 9 PCA plot of first and third components for the seed variable space. Two further expert clusterings are visible. One cluster consisting of experts, 2,6,8,9,11 and 12, all of whom sit above the line y=0. One further cluster with experts 3,5,7,13 and 14 is visible with members sitting below the line y=0. Experts 1,4 and 10 have been removed from this plot to reduce clutter.

and lower uncertainty. Variable 10 in CWD study, outlined in Fig. 11, which looks at the link between wild and farmed deer, is a good example. The distance between the 50th percentile and the 95 percentile in the final DM estimation is 2%-3% across all three model types (EWDM, PWDM and Bayesian DM). This compares to up to 57.5% in the lower half of the distribution (between the 5th percentile and the 50th percentile). For this type of variable, the EWDM will often have uniform probability all the way out to the limits because there is a single expert who has had this extreme judgement. The PWDM may not include this expert in the final aggregation (or significantly down weight them) and therefore does not recognise this tail. The Bayesian p.d.f however maintains the potential for the low value, but decays much more rapidly than the EWDM. This is due both to the chosen parameterisation and because the expert with the extreme perspective (expert 12) sits within a broader homogeneity group, this will effectively down-weight the effect of this expert in the model. This is an interesting counterpoint to the Arkansas study, in which the grouping up-weighted the differentiated view as that individual was grouped alone, here the view is included but down-weighted as their perspective is grouped with many others that are less extreme. This leads to a lower risk profile for the Bayesian DM which sits between the EWDM and the PWDM. The final result, in Fig. 11 is that the Bayesian model acknowledges the additional lower uncertainty identified by a small subset of the experts (and highlighted in the EWDM) whilst maintaining the mass closer to the bulk of the estimations, similar to the PWDM. In these cases this

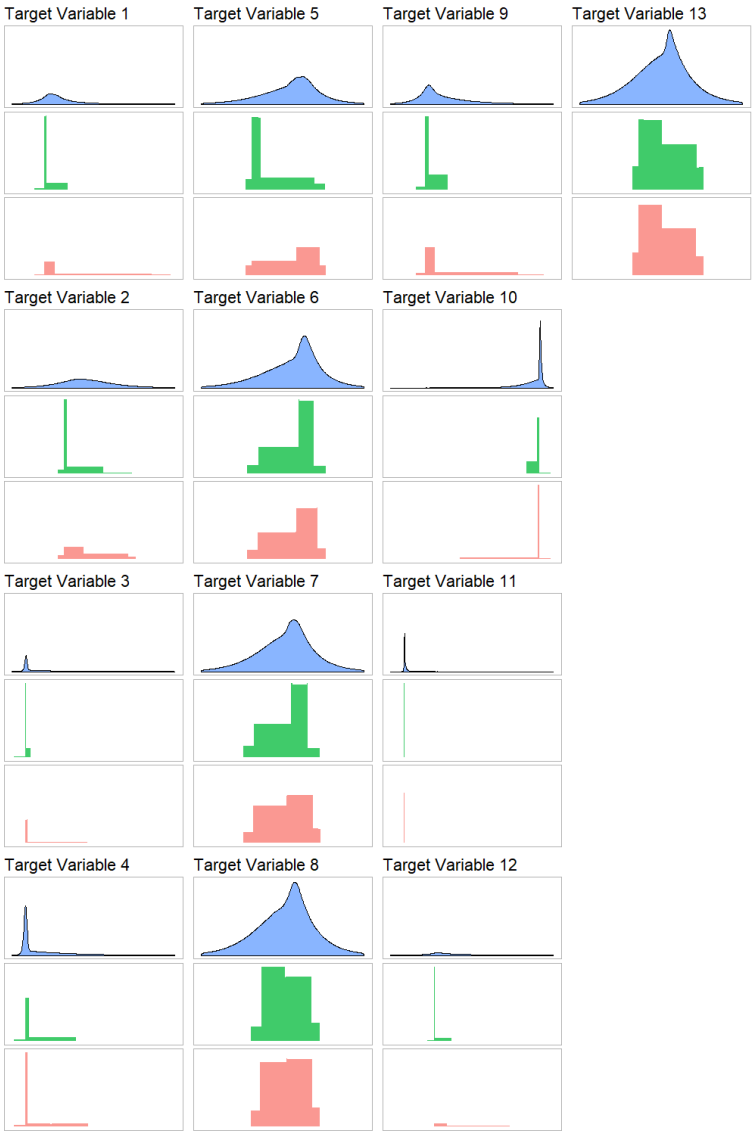

Fig. 10 Final distributions for all target variables in the CWD study. The Bayesian decision maker (blue) consistently demonstrates a wider support than the PWDM (green) or the EWDM (pink). The Bayesian model is also always unimodal, emphasising the underlying consistency in the estimations.

behaviour can outweigh the impact of recalibration due to overconfidence and lead to tighter distributions than the EWDM in one tail.

As is evident across the studies analysed so far, it is very common to see overconfidence (as defined by low statistical accuracy denoting too narrow bounds) in

expert judgement studies. Outside of simulated data the authors are yet to see a study with experts consistently demonstrating under-confidence in their judgement. This cross-domain tendency for experts to be overconfident should give DMs pause for thought. Recalibration of expert opinions is a controversial subject, and there are certainly contexts when it is unwise; we would argue however that a DM should not neglect this critical information when assessing their belief in light of the elicited judgements. This would lead such a DM to a Bayesian model similar to the one outlined here.

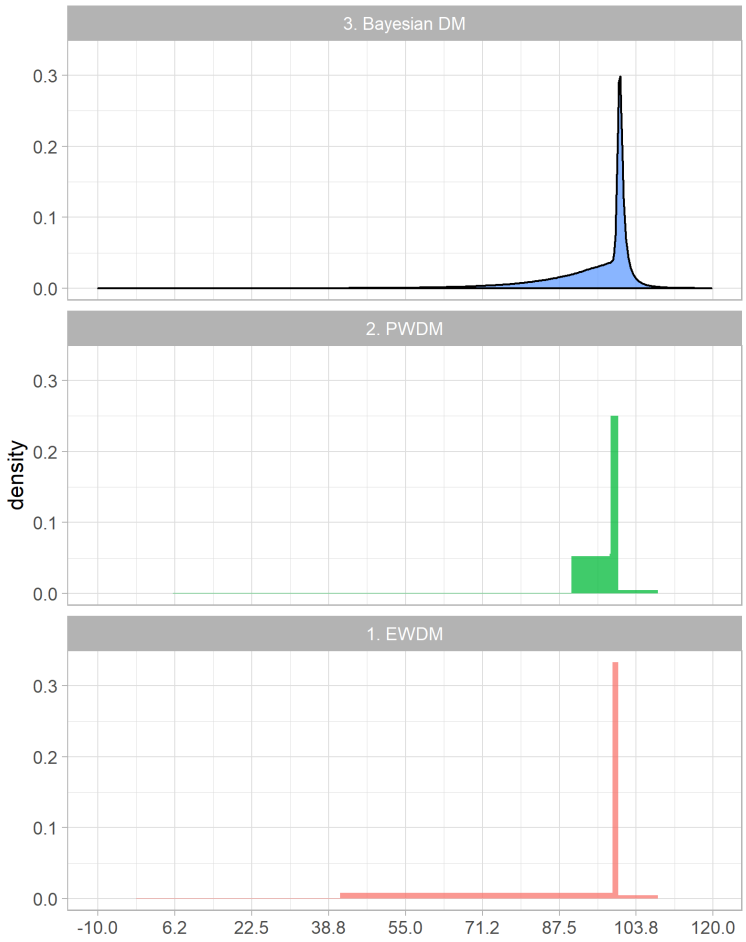

Fig. 11 Final distributions for target variable 10 in the CWD study. The Bayesian decision maker (blue) lower portion of the pdf decays much quicker than the EWDM (pink). This leads to a posterior expectation on the tracked quantiles between the EWDM and the PWDM (green)

5 Discussion

The above examples demonstrate that a Bayesian framework for SEJ can generate comparable results to those of the current exemplars in this space (Cooke's Classical model and EWDM models) whilst emphasising different components of the information provided by the experts. Across the target variables shown we have seen consistently that the Bayesian model demonstrates broader support in the posterior than the output of both the EWDM and the PWDM. This suggests the overshoot percentages utilised in pooling methods should be very carefully considered.

Posterior consistency has been evident when considering different parameterisations of the calibration component of the model (i.e. assuming two or three recalibration inflation factors) and utilising varying numbers of elicited quantiles. This could provide an opportunity for study facilitators to minimise the number of elicited items and thereby reduce time and cost.

One of the advantages of the Bayesian approach is that visualisations can be produced throughout its application which gives an exploratory tool for the DM to understand drivers of uncertainty captured within the posterior. These visualisations can be on core model elements, such as PCA plots of the clustering or homogeneity group posteriors, or, on other latent variables within the model. The ability to see the uncertainty represented visually can give confidence to the DM in their ability to move forward with understanding.

The underlying model we have described in the paper is quite generic, although we have given a more specific formulation for our examples. Not withstanding this, we note two potential areas of future research to allow the model to be used in broader contexts. The first of these is scale invariance. When we have variables on both a linear and a log scale, appropriate calibration adjustments must be considered differently. The current method of utilising multiplicative inflation factors is not suitable for such an endeavour. One option would be to transform the experts' judgements prior to modelling in order to allow them be dealt with consistently. One method would be to pass them through the DM's prior, this approach was first outlined by French and Wiper, [50]. Another option would be to recalibrate the percentiles themselves rather than the variables, i.e. apply the inflation factor to the quantile probabilities, thus moving the problem all into the single scaled probability domain.

The second area that warrants further investigation is how we handle DM/expert correlation. Currently we have bypassed this with the use of diffuse priors however with a knowledgeable DM this would not necessarily be appropriate [32]. A suitable model adjustment would need to be found.

To demonstrate this model is truly a feasible alternative for a DM, it is important to test the model across a number of studies and decisions to demonstrate the specific contexts in which this model will provide advantages. Luckily, given the wealth of studies that have been conducted historically and reside within Cooke's Database [14], there is a number of conditions within which to test this approach. One method is to apply the model to all of the studies within this database and compare versus the current models utilising out-of-sample (OOS) validation. Unfortunately as reali-

sations of events assessed in SEJ studies are, by definition, uncommon; true OOS validation is rarely feasible. Typical approaches ([8], [19], [12]) utilise cross-validation techniques, removing one or more of the seed variable, setting these as target variables and training the model with the remaining seeds. Forecast precision is then measured in standard ways as the true values for these targets is known *a priori*. Further comparisons with the performance of our model with the Classical Model provide additional support for the conclusions in this paper on the relative merits of the models [33].

Finally, it is hypothesised that as the number of experts increases we would expect to see the results of utilising a Bayesian method and the PWDM approach to further align. In such situations the PWDM will typically have a broader number of experts with non-trivial weights, thereby representing a mixture of many well-calibrated individuals, which is conceptually very similar to the Bayesian model. The over shoot percentages defined as part of Cooke's Classical PWDM are analagous to the tails on the Bayesian model. It is conjectured if these were relaxed further than the 10% commonly used today, as the number of experts were to increase, there is potential for convergence between the posterior distributions for the PWDM and the Bayesian model. Many studies today are not conducted on a significant enough scale to demonstrate this behaviour. As mass participation subject judgement events become more common through virtual elicitation over the internet, we would expect to see enough data to become available to test this hypothesis in the future.

References

1. Abdallah, N.B., Mouhous-Voyneau, N. and Denoeux, T. (2014). "Combining statistical and expert evidence using belief functions: Application to centennial sea level estimation taking into account climate change." *International Journal of Approximate Reasoning*, 55(1): 341–354. doi: http://dx.doi.org/10.1016/j.ijar.2013.03.008
2. Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., and Rousseau, J. (2012). "Combining expert opinions in prior elicitation." *Bayesian Analysis*, 7(3): 503–532. doi: http://dx.doi.org/10.1214/12-BA717
3. Aspinall, W. P. (2006). "Structured elicitation of expert judgement for probabilistic hazard and risk assessment in volcanic eruptions." *Statistics in volcanology*, 1: 15–30. doi: http://dx.doi.org/10.1144/IAVCEI001.2
4. Billari, F. C., Graziani, R., and Melilli, E. (2014). "Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm." *Demography*, 51(5), 1933-1954. doi: http://dx.doi.org/10.1007/s13524-014-0318-5
5. Boujelben, M. A., De Smet, Y., Frikha, A., and Chabchoub, H. (2011). "A ranking model in uncertain, imprecise and multi-experts contexts: The application of evidence theory" *International journal of approximate reasoning*, 52(8): 1171–1194. doi: http://dx.doi.org/10.1016/j.ijar.2011.06.008
6. Chen, H., Leung, C.-C., Xie, L., Ma, B., Li, H., (2015). "Parallel inference of dirichlet process Gaussian mixture models for unsupervised acoustic modeling: a feasibility study" *INTERSPEECH-2015*, 3189–3193
7. Chesley, G. R. (1975). "Elicitation of subjective probabilities: a review" *The Accounting Review*, 50(2): 325–337.

8. Clemen, R. T. (2008). "Comment on Cooke's classical method" *Reliability Engineering and System Safety*, 93(5): 760–765. doi: http://dx.doi.org/10.1016/j.ress.2008.02.003
9. Clemen, R. T. and Lichtendahl, K. C. (2002). "Debiasing expert overconfidence: A Bayesian calibration model" *Sixth International Conference on Probablistic Safety Assessment and Management (PSAM6)*.
10. Clemen, R. T. and Winkler, R. L. (1999). "Combining probability distributions from experts in risk analysis" *Risk analysis*, 19(2): 187–203. doi: http://dx.doi.org/10.1111/j.1539-6924.1999.tb00399.x
11. Colson, A. and Cooke, R. M. (2018). "Expert Elicitation: Using the Classical Model to Validate Experts' Judgments" *Review of Environmental Economics and Policy*, 12(1): 113–132. doi: http://dx.doi.org/10.1093/reep/rex022
12. Colson, A. R. and Cooke, R. M. (2017). "Cross validation for the classical model of structured expert judgment" *Reliability Engineering & System Safety*, 163: 109–120. doi: http://dx.doi.org/10.1016/j.ress.2017.02.003
13. Cooke, R. M. (1991). "Experts in uncertainty: opinion and subjective probability in science" *Oxford University Press on Demand*
14. Cooke, R. M. (2014). "Validating Expert Judgment with the Classical Model" *Experts and Consensus in Social Science*, 191–212. doi: http://dx.doi.org/10.1007/978-3-319-08551-7_10
15. Cooke, R. M. and Goossens, L. H. J. (2000). "Procedures guide for structural expert judgement in accident consequence modelling" *Radiation Protection Dosimetry*, 90(3): 303–309. doi: http://dx.doi.org/10.1093/oxfordjournals.rpd.a033152
16. Cooke, R. M. and Goossens, L. H. J. (2004). "Expert judgement elicitation for risk assessments of critical infrastructures" *Journal of Risk Research*, 7(6): 643–656. doi: http://dx.doi.org/10.1080/1366987042000192237
17. Cox, D. R. (1958). "Two further applications of a model for binary regression" *Biometrika*, 562–565. doi: http://dx.doi.org/10.2307/2333203
18. EFSA (2014). "Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment" *EFSA Journal*.
19. Eggstaff, J. W., Mazzuchi, T. A., and 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. doi: http://dx.doi.org/10.1016/j.ress.2013.07.015
20. French, S. (1980). "Updating of Belief in the Light of Someone Else's Opinion" *Journal of the Royal Statistical Society. Series A (General)*, 143(1): 43–48. doi: http://dx.doi.org/10.2307/2981768
21. French, S. (1985). "Group consensus probability distributions: A critical survey" *Bayesian statistics*, 2: 183–202.
22. French, S. (2011). "Aggregating expert judgement" *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 105(1): 181–206. doi: http://dx.doi.org/10.1007/s13398-011-0018-6
23. Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005). "Statistical methods for eliciting probability distributions" *Journal of the American Statistical Association*, 100(470): 680–701. doi: http://dx.doi.org/10.1198/016214505000000105
24. Gelman, A. (2004). "Prior distributions for variance parameters in hierarchical models" Report, EERI Research Paper Series.
25. Genest, C. and McConway, K. J. (1990). "Allocating the weights in the linear opinion pool" *Journal of Forecasting*, 9(1): 53–73. doi: http://dx.doi.org/10.1002/for.3980090106
26. Genest, C. and Zidek, J. V. (1986). "Combining probability distributions: A critique and an annotated bibliography" *Statistical Science*, 114–135. doi: http://dx.doi.org/10.1214/ss/1177013825
27. Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007). "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2): 243–268. doi: http://dx.doi.org/10.1111/j.1467-9868.2007.00587.x
28. Gosling, J. P., Hart, A., Mouat, D. C., Sabirovic, M., Scanlan, S., and Simmons, A. (2012). "Quantifying experts' uncertainty about the future cost of exotic diseases" *Risk Analysis*, 32(5): 881–893. doi: http://dx.doi.org/10.1111/j.1539-6924.2011.01704.x

29. Gosling, J. P., Oakley, J. E., and O'Hagan, A. (2007). "Nonparametric elicitation for heavy-tailed prior distributions" *Bayesian Analysis*, 2(4): 693–718. doi: http://dx.doi.org/10.1214/07-BA228
30. Ha-Duong, M. (2008). "Combining Experts' Judgments: Comparison of Algorithmic Methods Using Synthetic Data" *International Journal of Approximate Reasoning*, 49(3): 555–574.
31. Hammitt, J. K. and Zhang, Y. (2013). "Hierarchical fusion of expert opinions in the Transferable Belief Model, application to climate sensitivity" *Risk Analysis: An International Journal*, 33(1): 109–120.
32. Hartley, D. and French, S. (2018). "Elicitation and calibration: a Bayesian perspective" In L. Dias, A. Morton, & J. Quigley (Eds.) *Elicitation*, Springer, Cham, 119–140. doi: http://dx.doi.org/10.1007/978-3-319-65052-4_6
33. Hartley, D. and French, S. (2020). "Bayesian modelling of dependence between experts: some comparisons with Cooke's Classical Model" In Hanea, A, Nane, G.F., Bedford, T & French, S (Eds.) *Expert Judgement in Risk and Decision Analysis*, Springer, Cham, *In Press*. doi: 10.1007/978-3-030-46474-5
34. Kadane, J. B. and Fischhoff, B. (2013). "A cautionary note on global recalibration." *Judgment and Decision Making* 8(1), 25–27.
35. Kahneman, D. and Tversky, A. (1979). "Prospect theory: An analysis of decision under risk" *Econometrica: Journal of the Econometric Society*, 47(2) 263–291. doi: http://dx.doi.org/10.2307/1914185
36. Kaplan, S. (2000). "'Combining Probability Distributions from Experts in Risk Analysis'" *Risk Analysis*, 20(2): 155–156. doi: http://dx.doi.org/10.1111/0272-4332.202015
37. Kinnersley, N. and Day, S. (2013). "Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study" *Pharmaceutical statistics*, 12(2): 104–113. doi: http://dx.doi.org/10.1002/pst.1552
38. Lindley, D. (1983). "Reconciliation of probability distributions" *Operations Research*, 31(5), 866–880. doi: http://dx.doi.org/10.1287/opre.31.5.866
39. Morris, P. A. (1974). "Decision analysis expert use" *Management Science*, 20(9): 1233–1241. doi: http://dx.doi.org/10.1287/mnsc.20.9.1233
40. Morris, P. A. (1977). "Combining expert judgments: A Bayesian approach" *Management Science*, 23(7): 679–693. doi: http://dx.doi.org/10.1287/mnsc.23.7.679
41. Mumpower, J. L. and Stewart, T. R. (1996). "Expert Judgement and Expert Disagreement" *Thinking and Reasoning*, 2(2/3): 191–212. doi: http://dx.doi.org/10.1080/135467896394500
42. Perälä, T., Vanhatalo, J. and Chrysafi, A. (2019). "Calibrating expert assessments using hierarchical Gaussian process models" *Bayesian Analysis*, Advance Publication doi: 10.1214/19-BA1180
43. Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica sinica*, 639–650 doi: 10.1214/19-BA1180
44. Shrestha, G. and Rahman, S. (1996). "A statistical representation of imprecision in expert judgments" *International journal of approximate reasoning*, 5(1): 1–25. doi: http://dx.doi.org/10.1016/0888-613X(91)90004-6
45. Smith, J. Q. (2010). *Bayesian decision analysis: principles and practice*. Cambridge University Press. doi: http://dx.doi.org/10.1017/CBO9780511779237
46. Tyshenko, M. G., ElSaadany, S., Oraby, T., Darshan, S., Aspinall, W., Cooke, R., Catford, A., and Krewski, D. (2011). "Expert elicitation for the judgment of prion disease risk uncertainties" *Journal of Toxicology and Environmental Health, Part A*, 74(2-4): 261–285. doi: http://dx.doi.org/10.1080/15287394.2011.529783
47. Tyshenko, M. G., ElSaadany, S., Oraby, T., Darshan, S., Catford, A., Aspinall, W., Cooke, R., and Krewski, D. (2012). "Expert judgement and re-elicitation for prion disease risk uncertainties" *International Journal of Risk Assessment and Management*, 16(1-3): 48–77. doi: http://dx.doi.org/10.1504/IJRAM.2012.047552
48. Wilson, K. J. (2017) "An investigation of dependence in expert judgement studies with multiple experts." *International Journal of Forecasting*, 33(1): 325–336. doi: http://dx.doi.org/10.1016/j.ijforecast.2015.11.014

49. Wilson, K. J. and Farrow, M. (2018) "Combining judgements from correlated experts" *Elicitation*, Springer, Cham, 211–240 doi: http://dx.doi.org/10.1007/978-3-319-65052-4_9
50. Wiper, M. P. and French, S. (1995). "Combining experts' opinions using a normal-wishart model" *Journal of Forecasting*, 14(1): 25–34. doi: http://dx.doi.org/10.1002/for.3980140103

Acknowledgements

The authors wish to thank the three anonymous reviewers whose comments have significantly improved the content and structure of this paper.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

A Appendix

A.1 Full method outline

Here we provide the mathematical outline for the full method combining the hierarchical model homogeneity group calculation, aggregation and calibration processes.

A.1.1 Data going into the model:

- A set of experts **E**, a set of seed variables **Y** and a set of target variables **X**
- A set of probabilities associated with quantiles that were elicited from each expert for each variable - P_L, P_M, P_U
- Elicited quantiles from each expert for the target variables - $L_{Xeh}, M_{Xeh}, U_{Xeh}$ $\forall e \in \mathbf{E}, X \in \mathbf{X}$.
- Elicited quantiles from each expert for the seed variables - $L_{Yeh}, M_{Yeh}, U_{Yeh}$ $\forall e \in \mathbf{E}, Y \in \mathbf{Y}$.
- A parameterised distribution to be fit to the elicited data for each expert - g_e $\forall e \in \mathbf{E}$ with cumulative distribution function G_e
- The decision maker's prior for each target variable - π_{DM_X}

A.1.2 Step one - Homogeneity group calculation

For each elicited seed variable mid-quantile, M_{Yeh}, rescale onto the unit interval. Term the new rescaled value rM_{Yeh}.

$$rM_{Yeh} := \frac{M_{Yeh} - min(\{M_{Yeh} : e \in \mathbf{E}\})}{max(\{M_{Yeh} : e \in \mathbf{E}\}) - min(\{M_{Yeh} : e \in \mathbf{E}\})} \tag{20}$$

Define $\mathbf{Y}_e$ to be the $|\mathbf{Y}|$ dimensional tuple:

$$\mathbf{Y}_e = (rM_{1eh}, rM_{2eh}, ..., rM_{|\mathbf{Y}|eh}) \tag{21}$$

Run an agglomerative hierarchical clustering process. Set each item $\mathbf{Y}_e$ to be its own cluster $\mathbf{C}_e$. Create a dendogram by merging clusters one at a time based on the Euclidean distance between them in $\mathbb{R}^{|\mathbf{Y}|}$.

i.e. Merge the two clusters $\mathbf{C}_i$ and $\mathbf{C}_j$ that minimise D($\mathbf{C}_i$,$\mathbf{C}_j$). Where

$$D(\mathbf{C}_i, \mathbf{C}_j) = max\, d(\mathbf{Y}_i, \mathbf{Y}_j) \quad \mathbf{Y}_i \in \mathbf{C}_i, \mathbf{Y}_j \in \mathbf{C}_j, \tag{22}$$

and

$$d(\mathbf{Y}_i, \mathbf{Y}_j) = \sqrt{\sum_{k \in 1:|\mathbf{Y}|} (\mathbf{Y}_i(k) - \mathbf{Y}_j(k))^2} \tag{23}$$

(This defines an agglomerative process with a Euclidean distance metric and a complete linkage criterion. These are standard metrics to use for this form of clustering but many others are available.)

This merging process is repeated using the same criteria until all elements form a single cluster.

Final homogeneity groupings, **H**, are then defined by selecting a cut of this dendogram which can be done either manually based on visual inspection or utilising a dynamic tree cutting approach such as in the R package NbClust.

The cluster groupings that sit along the cut are assignments of experts to homogeneity groups. Suppose along this cut ten experts were clustered in the following three groups. Group 1 - experts $\{1,3,5,7,9\}$, Group 2 - experts$\{2,4,6\}$, Group 3 - experts$\{8,10\}$. **H** would then be the array $\{1,2,1,2,1,2,1,3,1,3\}$.

Validate the homogeneity group choices by running a principal component analysis (PCA) over the seed variable space. Visualise the first two or three principal components pairwise and consider a scree plot of the PCA to understand the level of variance captured within these components. Determine whether there is agreement with the choices made algorithmically and then finalise homogeneity group assignments.

A.1.3 Step two - calibration and aggregation

Calibration

For each Y in **Y** and e in **E**, assume the true realisation of Y (y_Y) are random draws from a distribution of structure g_e defined by the unbiased quantile estimates of the expert e.

$$y_Y \sim g_e(\cdot|L^*_{Yeh}, U^*_{Yeh}, M^*_{Yeh}) \quad e \in \mathbf{E}, Y \in \mathbf{Y} \tag{24}$$

where the unbiased quantile estimates are defined by:

$$\begin{aligned} L^*_{Yeh} &:= (1-\alpha_{le})M_{Ye} + \alpha_{le}L_{Ye} && e \in \mathbf{E} \\ U^*_{Yeh} &:= (1-\alpha_{ue})M_{Ye} + \alpha_{ue}U_{Ye} && h = \mathbf{H}(e) \\ M^*_{Yeh} &:= M_{Ye} && X \in \mathbf{X} \end{aligned} \tag{25}$$

and the inflation factors for each expert are random draws from a distribution which is consistent across experts within a single homogeneity group:

$$\begin{aligned} \alpha_{le}|A_{lh}, B_{lh} &\sim \Gamma(A_{lh}+1, B_{lh}) && e \in \mathbf{E} \\ \alpha_{ue}|A_{uh}, B_{uh} &\sim \Gamma(A_{uh}+1, B_{uh}) && h = \mathbf{H}(e) \end{aligned} \tag{26}$$

where A_{lh}, A_{uh}, B_{lh} and B_{uh} are defined by:

$$\begin{aligned} A_{lh} \sim Pois(a_l) \;\; and \;\; B_{lh} \sim Exp(b_l) \\ A_{uh} \sim Pois(a_u) \;\; and \;\; B_{uh} \sim Exp(b_u) \end{aligned} \tag{27}$$

Hyperparameters a_l, a_u, b_l and b_u are consistent across all experts and homogeneity groups.

Aggregation

Assume that the elicited quantile for each expert target variable pair is a function of the underlying unbiased quantiles and the inflation factors inferred.

$$\begin{aligned} L_{Xeh} &= (L^*_{Xeh} - (1-\alpha_{le})M_{Xe})/\alpha_{le} & e \in \mathbf{E} \\ U_{Xeh} &= (U^*_{Xe} - (1-\alpha_{ue})M_{Xe})/\alpha_{ue} & h = \mathbf{H}(e) \\ M_{Xeh} &= M^*_{Xe} & X \in \mathbf{X} \end{aligned} \tag{28}$$

where the unbiased parameters $L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}$ are those such that:

$$G_e(L^*_{Xeh}|\gamma^*_{Xeh}) = P_L \tag{29}$$

$$G_e(M^*_{Xeh}|\gamma^*_{Xeh}) = P_M \tag{30}$$

$$G_e(U^*_{Xeh}|\gamma^*_{Xeh}) = P_U \tag{31}$$

where each experts' unbiased parameterised values γ^*_{Xeh}, for a given target variable $X \in \mathbf{X}$ and expert e are random draws from a distribution, f, defined by the homogeneity group $h \in \mathbf{H}$ within which e sits. The appropriate functional form of f is determined by the functional form of g_e. Homogeneity group parameters are random draws from a global distribution, which have the decision maker's prior.

$$\begin{aligned} \gamma^*_{Xeh} &\sim f(\cdot|\gamma_{Xh}, \rho_{Xh}) & \forall e \in \mathbf{E} \\ \gamma_{Xh} &\sim f(\cdot|\gamma_X, \rho_X) & \forall h \in \mathbf{H} \\ \gamma_X &\sim \pi_{DM_X} \end{aligned} \tag{32}$$

The parameters γ_X are then used to infer the target posterior given by $g_{DM}(\cdot|\gamma_X)$.

Please note: In practice, when encoding in a language such as BUGS*, the logical determination in equations (28)-(31) is embedded within the first line of (32). Thus the data is encoded as a random draw. To this extent, this is modelled as* $L_{Xeh} \sim f(\cdot|\gamma_{Xh}, \rho_{Xh}, \alpha_{le})$ *and the functional forms of* g_e *and equation (28) determine the structure of this draw. Similar for* U_{Xeh} *and* M_{Xeh}*.*

A.2 Full method outline - split normal parameterisation

When the distributions g_e are all defined to be a split normal then they can be represented by a single function such that:

$$g_e(x|L^*_{Xeh}, M^*_{Xeh}, U^*_{Xeh}) \sim \begin{cases} \frac{1}{\sigma^*_{Xleh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-M^*_{Xeh}}{\sigma^*_{Xleh}})^2} & if x < M^*_{Xeh} \\ \frac{1}{\sigma^*_{Xueh}\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-M^*_{Xeh}}{\sigma^*_{Xueh}})^2} & if x \geqslant M^*_{Xeh} \end{cases} \tag{33}$$

where, the unbiased standard deviations σ^*_{Xleh} and σ^*_{Xueh} are calculated by:

$$\sigma^*_{Xleh} = \frac{M^*_{Xeh} - L^*_{Xeh}}{\delta_1} \quad and \quad \sigma^*_{Xueh} = \frac{U^*_{Xeh} - M^*_{Xeh}}{\delta_2} \tag{34}$$

and $\tau^*_{Xleh} := 1/\sigma^{2*}_{Xleh}$ and $\tau^*_{Xueh} := 1/\sigma^{2*}_{Xueh}$. Here δ_i represents the number of standard deviations between the elicited quantiles. The equation for g_e is identically defined for the seed variables and used in equation (24). The parameters γ^*_{Xeh} in equation (32) are then given by the triple, $(M^*_{Xeh}, \tau^*_{Xleh}, \tau^*_{Xueh})$. In this instance the generic aggregation component (equation (32)) is now replaced by:

$$\begin{aligned} L_{Xeh}|M_{Xh}, \alpha_{le}, \rho_{Xh} &\sim \mathcal{N}(M_{Xh} - \frac{\delta_1}{\sqrt{\tau^*_{Xleh}}}\alpha_{le}), \rho_{Xh}) \\ U_{Xeh}|M_{Xh}, \alpha_{ue}, \rho_{Xh} &\sim \mathcal{N}(M_{Xh} + \frac{\delta_2}{\sqrt{\tau^*_{Xueh}}}\alpha_{ue}), \rho_{Xh}) \\ M_{Xeh}|M_{Xh}, \rho_{Xh} &\sim \mathcal{N}(M_{Xh}, \rho_{Xh}) \end{aligned} \tag{35}$$

with **Location parameter aggregation:**

$$\begin{aligned} M_{Xh}|M_X, \rho_X &\sim \mathcal{N}(M_X, \rho_X) \\ M_X &\sim \mathcal{N}(M_{DM_X}, \rho_{X0}) \end{aligned} \tag{36}$$

and **Dispersion parameter aggregation:**

$$\begin{aligned} \frac{\tau_{Xlh}}{\tau^*_{Xleh}}|\tau_{Xlh}, \xi_{Xlh} &\sim \Gamma(\xi_{Xlh}, \xi_{Xlh}) & \frac{\tau_{Xuh}}{\tau^*_{Xueh}}|\tau_{Xuh}, \xi_{Xuh} &\sim \Gamma(\xi_{Xuh}, \xi_{Xuh}) \\ \frac{\tau_{Xl}}{\tau_{Xlh}}|\tau_{Xl}, \xi_{Xl} &\sim \Gamma(\xi_{Xl}, \xi_{Xl}) & \frac{\tau_{Xu}}{\tau_{Xuh}}|\tau_{Xu}, \xi_{Xu} &\sim \Gamma(\xi_{Xu}, \xi_{Xu}) \\ \tau^{-1}_{Xl} &\sim \tau^{-1}_{Xl0}\Gamma(a,a) & \tau^{-1}_{Xu} &\sim \tau^{-1}_{Xu0}\Gamma(a,a) \end{aligned} \tag{37}$$

The parameters now used to infer the target posterior are given by the triple $(M_X, \tau_{Xl}, \tau_{Xu})$. These are used as inputs into equation (33) to create the full aggregate distribution.

A.3 DAG for connected calibration and aggregation models

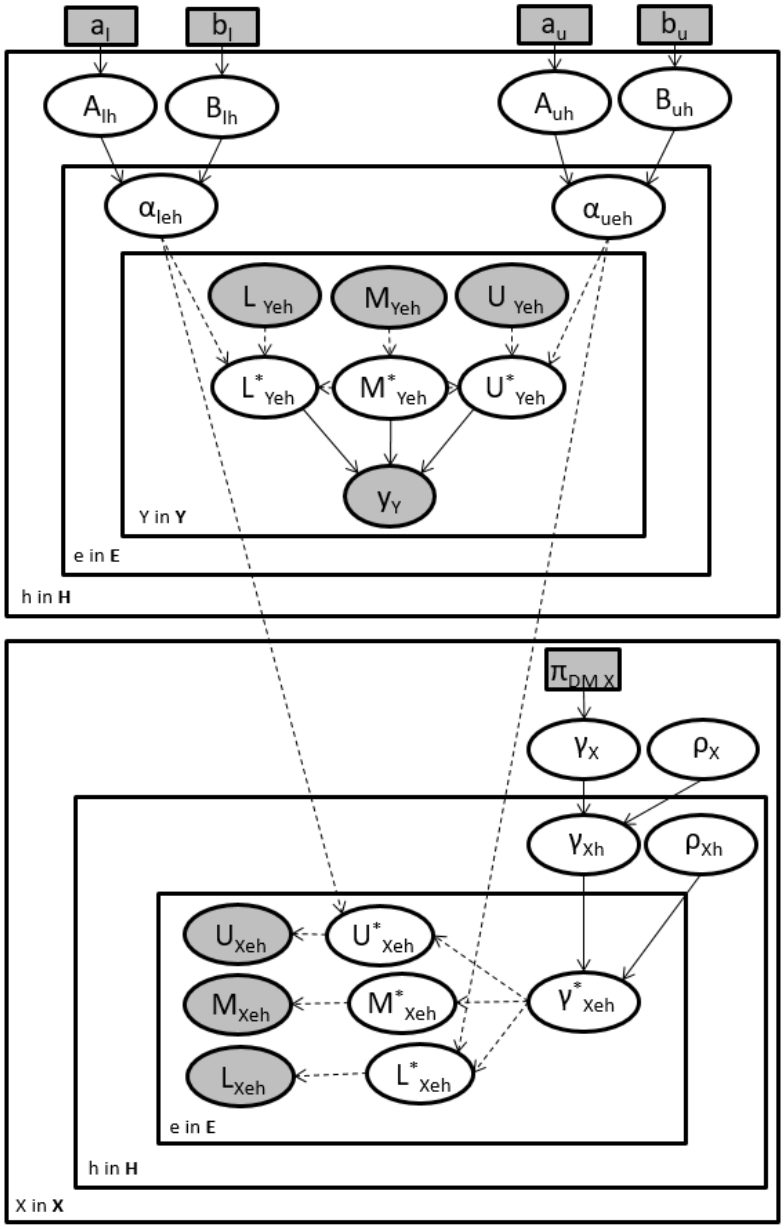

Fig. 12 Directed acyclic graph for the linked aggregation and calibration models. The inflation factors from the calibration model are used to logically determine unbiased estimators in the aggregation model.

A.4 DAG for a Dirichlet process mixture model

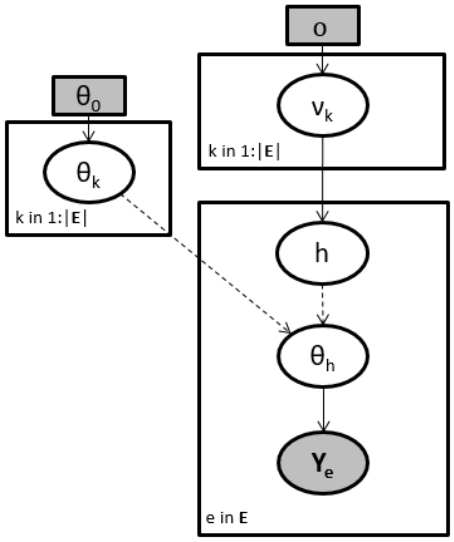

Fig. 13 Dirichlet process mixture model for homogeneity group definition. Experts are points in the space $\in \mathbb{R}^{|\mathbf{Y}|}$ and are clustered utilising a mixture model (typically Gaussian).

A.5 Additional Arkansas study analysis and figures

A.5.1 Dendogram of expert homogeneity groups

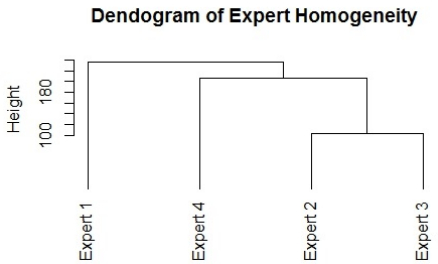

Fig. 14 Hierarchical clustering dendogram for the identification of expert homogeneity groups within the Arkansas Study. Expert 2 and 3 form a single homogeneity group.

A.5.2 Distributions for all target variables

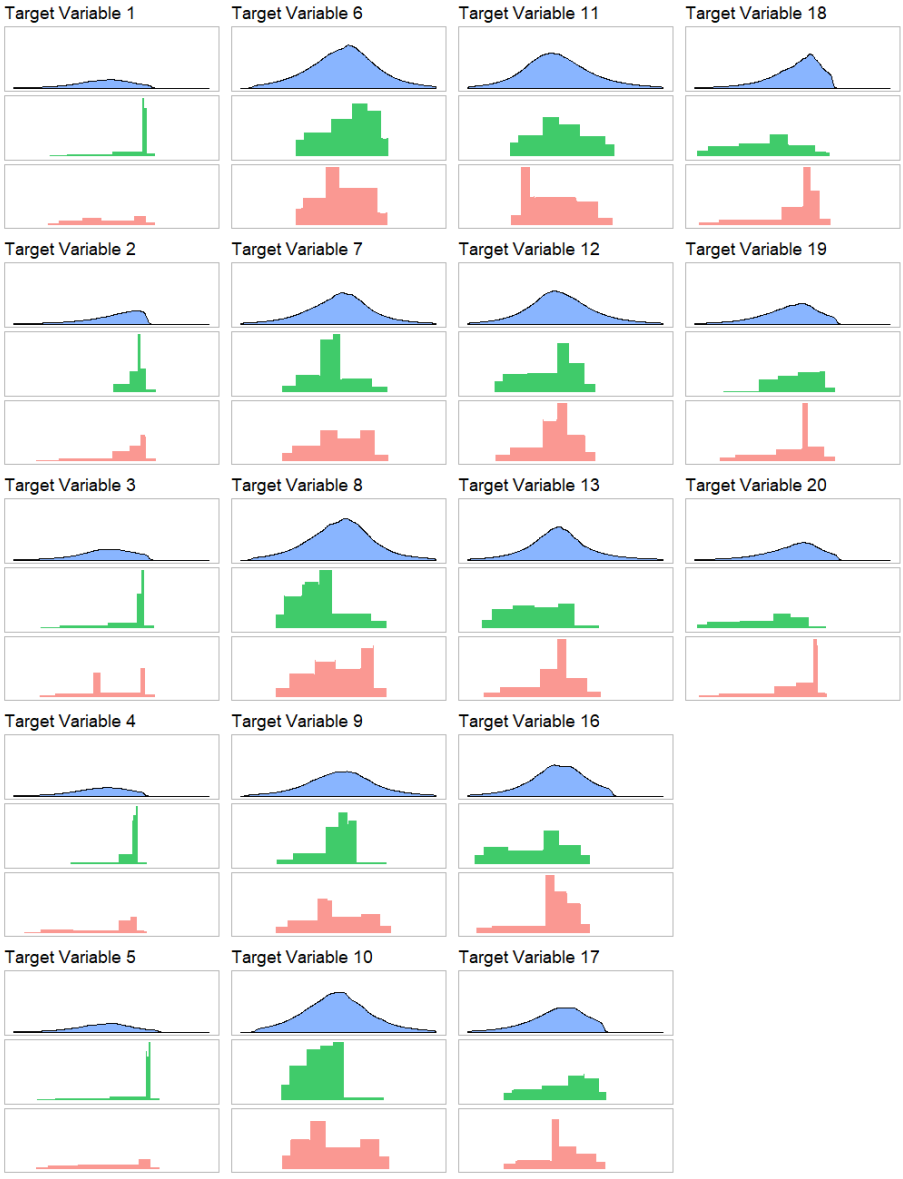

Fig. 15 Comparison of final distributions across all target variables within the Arkansas study. The Bayesian model (blue) demonstrates a larger support, aligned to the overconfidence demonstrated by experts in the seed variables.

A.5.3 Cumulative density functions for different parameterisations of the calibration and aggregation model

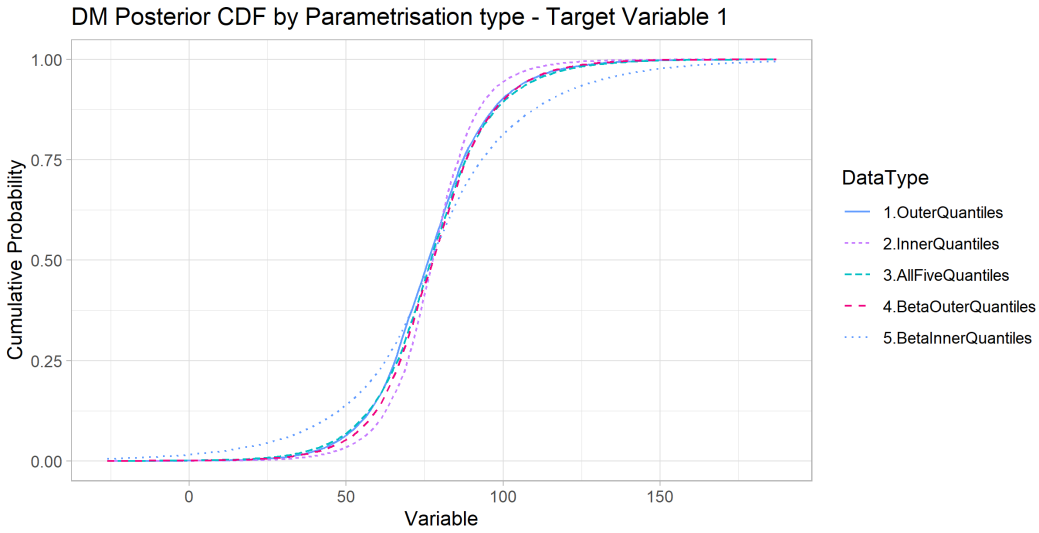

Fig. 16 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 1 in the Arkansas study.

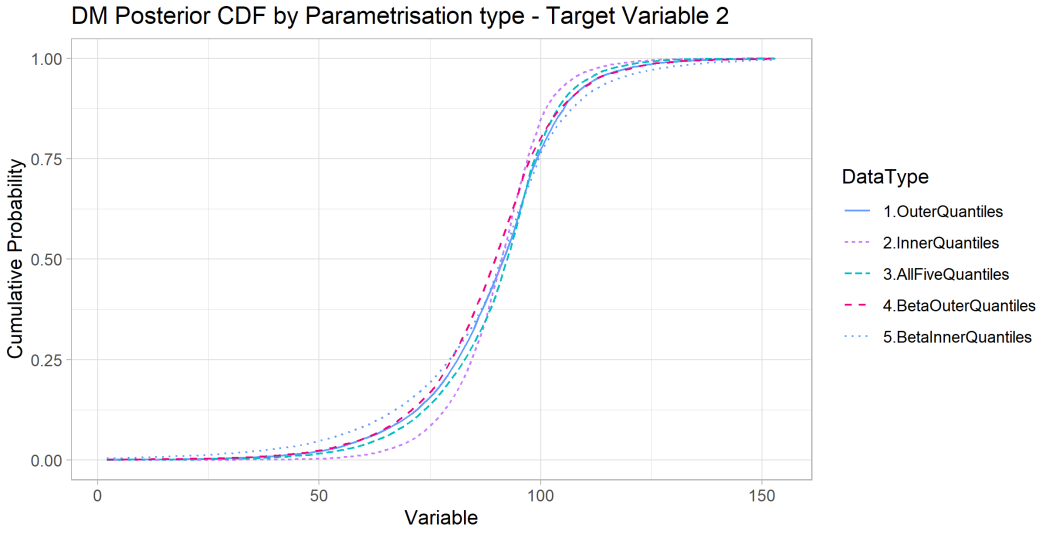

Fig. 17 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 2 in the Arkansas study.

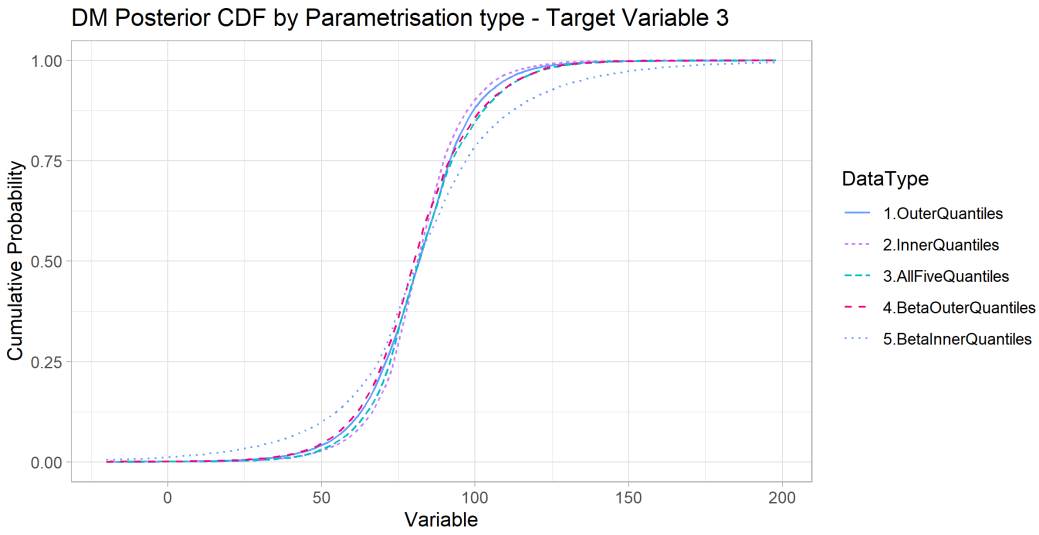

Fig. 18 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 3 in the Arkansas study.

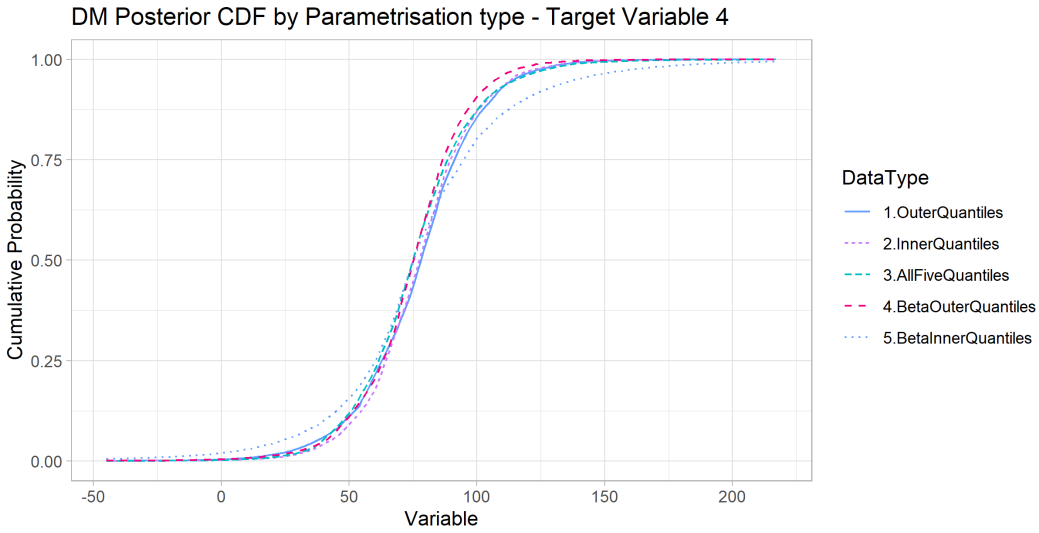

Fig. 19 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 4 in the Arkansas study.

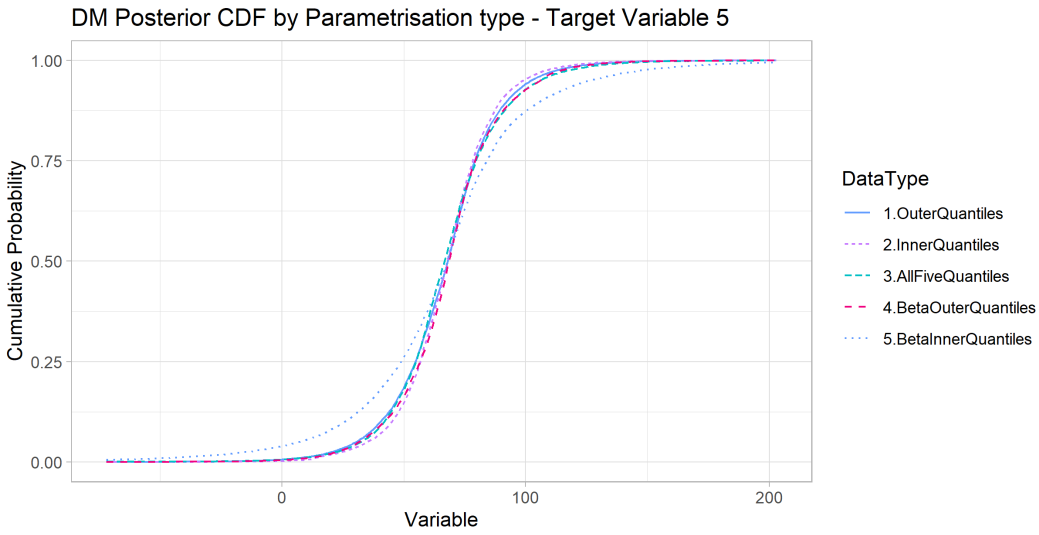

Fig. 20 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 5 in the Arkansas study.

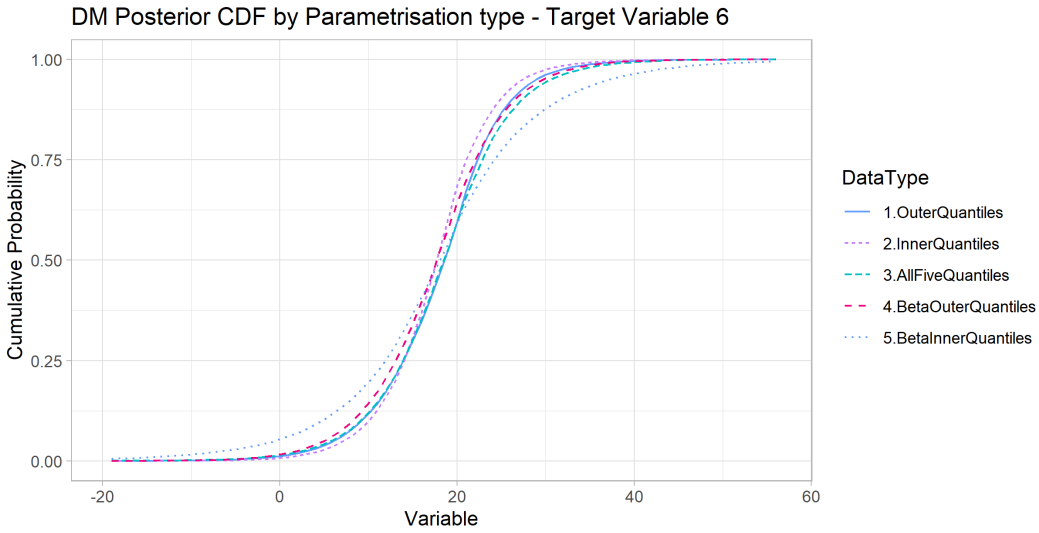

Fig. 21 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 6 in the Arkansas study.

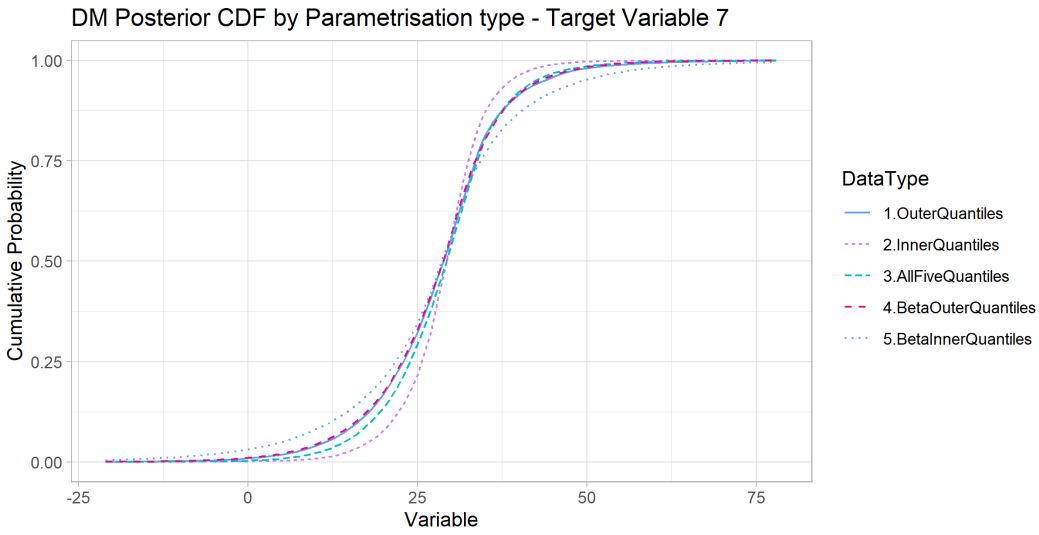

Fig. 22 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 7 in the Arkansas study.

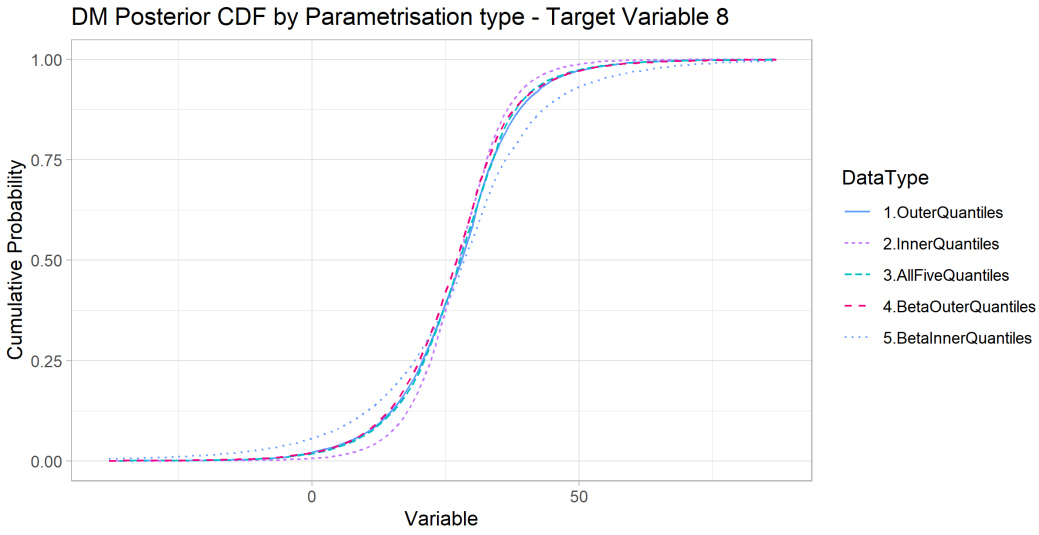

Fig. 23 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 8 in the Arkansas study.

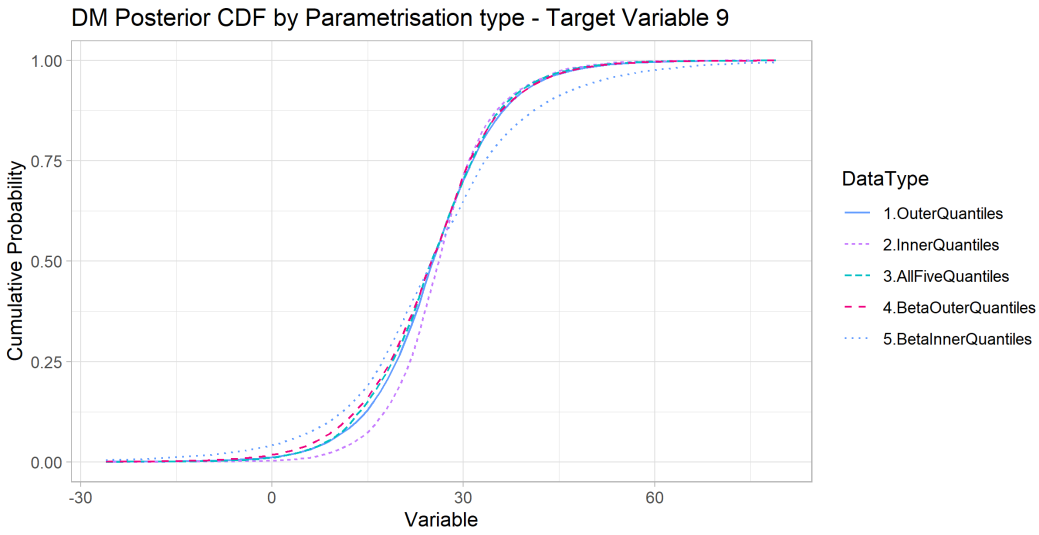

Fig. 24 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 9 in the Arkansas study.

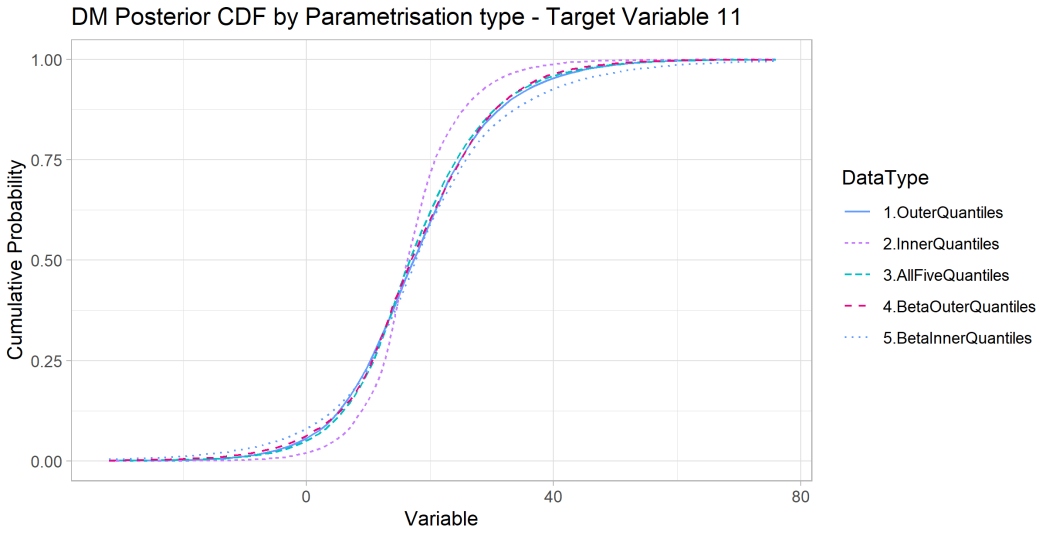

Fig. 25 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 11 in the Arkansas study.

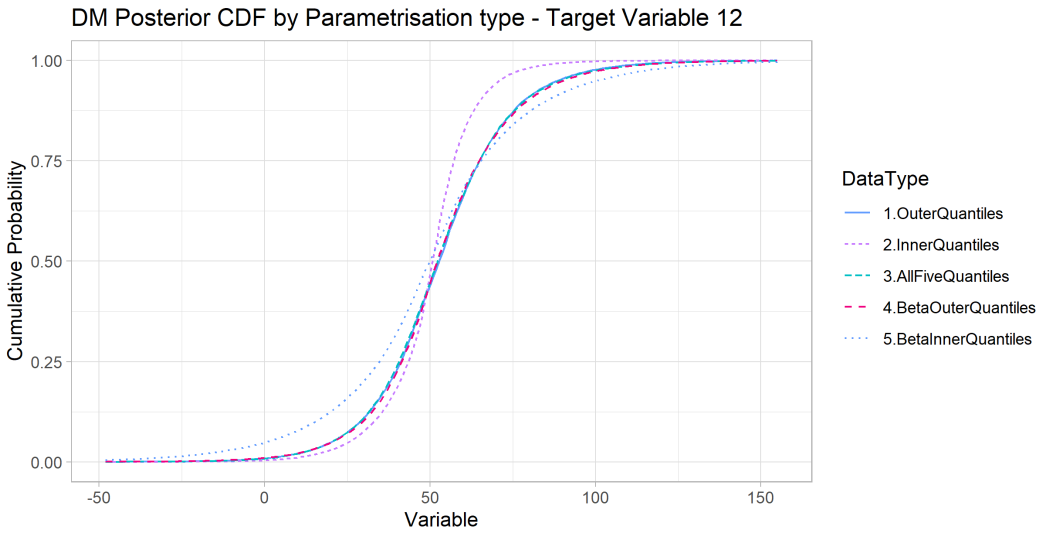

Fig. 26 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 12 in the Arkansas study.

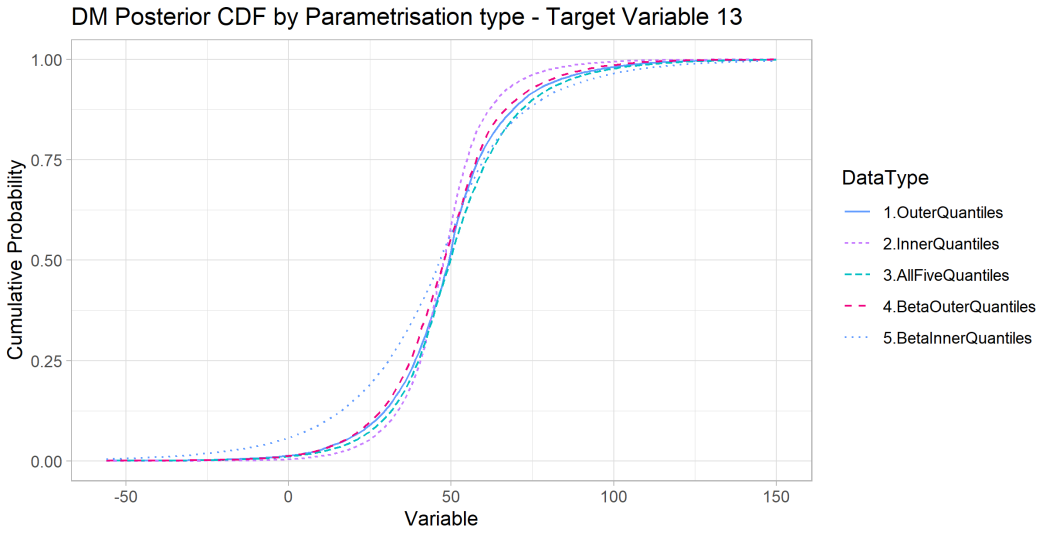

Fig. 27 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 13 in the Arkansas study.

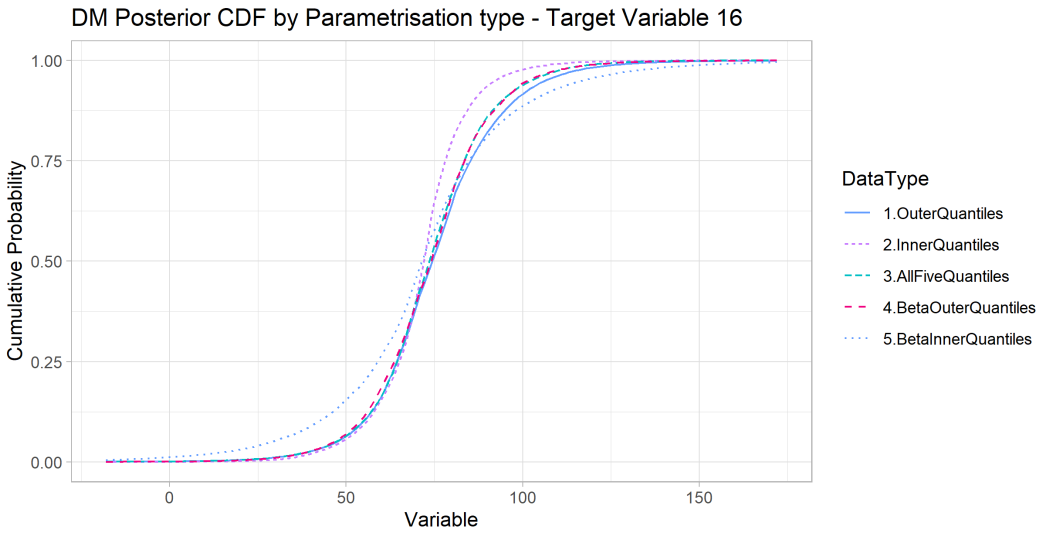

Fig. 28 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 16 in the Arkansas study.

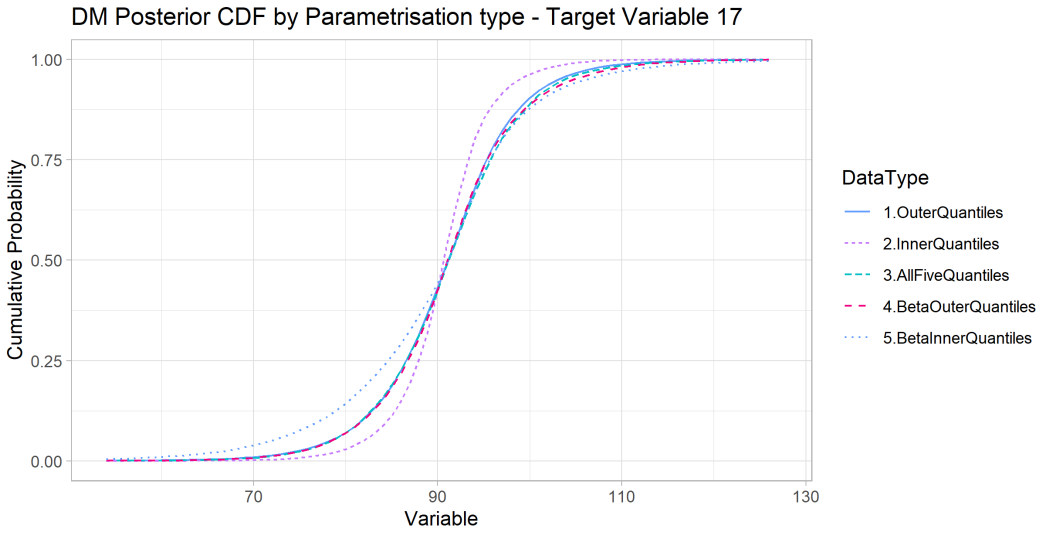

Fig. 29 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 17 in the Arkansas study.

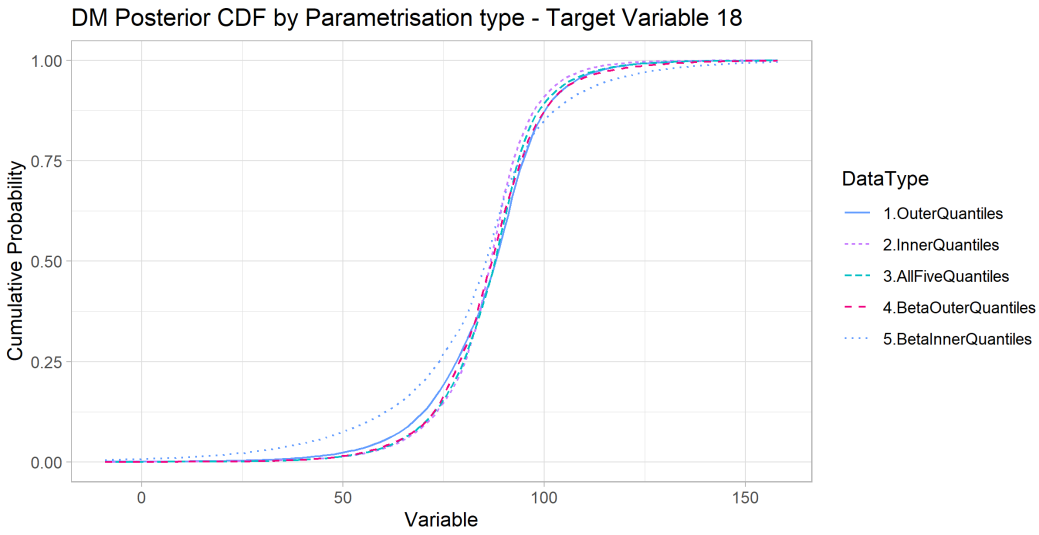

Fig. 30 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 18 in the Arkansas study.

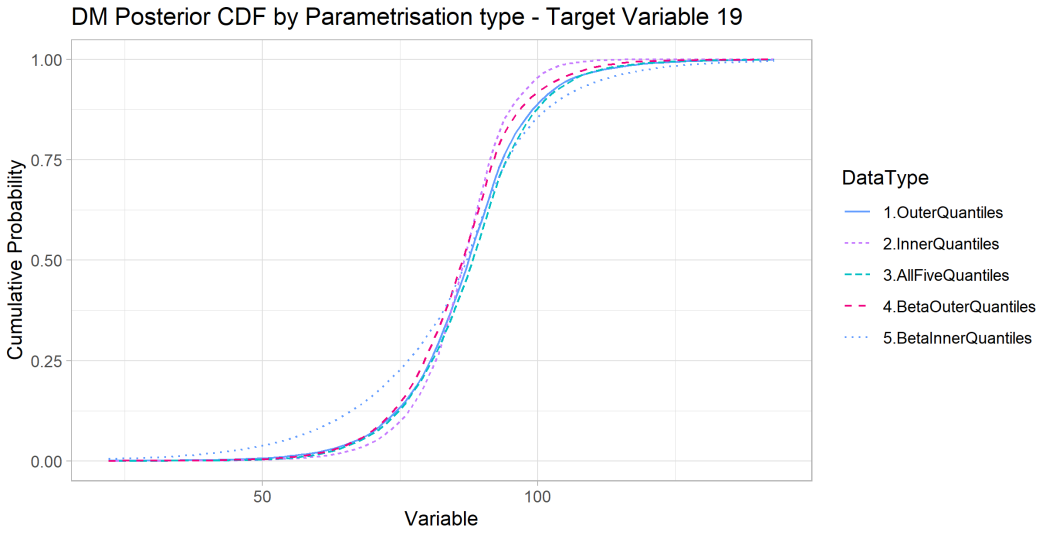

Fig. 31 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 19 in the Arkansas study.

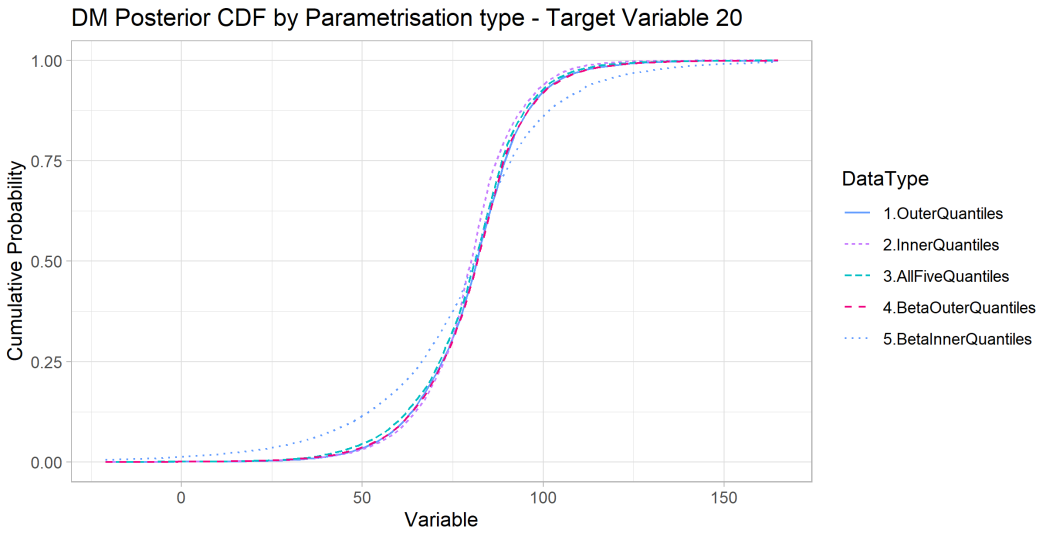

Fig. 32 Comparison of the posterior DM cumulative density functions by parameterisation type for target variable 20 in the Arkansas study.

A.6 Additional CWD study analysis

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Variable 1	0.0	9	90.5	0.0	1.7	19.8	-12.3	8.3	47.3
Variable 2	0.2	29.7	95.8	0.0	5	58.7	-37.9	30.8	114.8
Variable 3	0.6	14.3	454.4	0.5	4.8	40	-13.4	11	457
Variable 4	0.0	9.1	238.7	0.0	15.6	193.4	-8	7.1	345.6
Variable 5	0.1	65.3	100	0.0	13.3	92.1	-24.7	62.9	122.4
Variable 6	12.1	66.4	95.1	13.9	69.5	89.7	-18.9	68.6	121.8
Variable 7	11.7	56.5	91.8	12	62.1	84.8	-17	59.2	119.1
Variable 8	7.8	49.7	89.6	10	46.6	89	-36.6	52.8	123.4
Variable 9	0.0	8.7	83	0.0	3.4	19.8	-14.6	8.4	69.7
Variable 10	41.2	98.7	100	90.1	98.4	100	75.3	99.8	102.7
Variable 11	0.0	1.3	461.7	0.1	0.9	2	-2	1.1	38.2
Variable 12	0.0	14.8	87.8	0.0	0.7	19.7	-24.6	12.8	87.5
Variable 13	2.1	39.5	95.3	1.9	39.6	97.4	-46	44.7	120.2

Table 4 Comparison of DM quantiles for the 13 target variables, considering different modelling approaches to the CWD Study. The Bayesian decision maker (BDM) demonstrates a broader range of uncertainty than the equal weighted (EWDM) or performance weighted (PWDM) decision makers.