Chapter 5
Bayesian Modelling of Dependence Between Experts: Some Comparisons with Cooke's Classical Model

David Hartley and Simon French

Abstract A Bayesian model for analysing and aggregating structured expert judgement (SEJ) data of the form used by Cooke's classical model has been developed. The model has been built to create predictions over a common dataset, thereby allowing direct comparison between approaches. It deals with correlations between experts through clustering and also seeks to recalibrate judgements using the seed variables, in order to form an unbiased aggregated distribution over the target variables. Using the Delft database of SEJ studies, compiled by Roger Cooke, performance comparisons with the classical model demonstrate that this Bayesian approach provides similar median estimates but broader uncertainty bounds on the variables of interest. Cross-validation shows that these dynamics lead to the Bayesian model exhibiting higher statistical accuracy but lower information scores than the classical model. Comparisons of the combination scoring rule add further evidence to the robustness of the classical approach yet demonstrate outperformance of the Bayesian model in select cases.

5.1 Introduction

Algorithmic approaches for combining judgements from several experts have evolved over the years. Initially, techniques were either simple averaging, known as opinion polling, or in essence Bayesian (French 1985, 2011). However, the Bayesian approach did not prove practical and fell by the wayside, whilst the opinion polling techniques gained traction. In practice, Cooke's development of a performance-weighted opinion polling approach, known as the classical model (Cooke 1991, 2007), dominated among the mathematical approaches to eliciting and aggregating expert judgement data and remains the exemplar in this field. Non-mathematical

D. Hartley (✉) · S. French
Department of Statistics, University of Warwick, Coventry CV4 7AL, UK
e-mail: d.s.hartley@warwick.ac.uk

S. French
e-mail: simon.french.50@gmail.com

A. M. Hanea et al. (eds.), *Expert Judgement in Risk and Decision Analysis*,
International Series in Operations Research & Management Science 293,
https://doi.org/10.1007/978-3-030-46474-5_5

approaches (designated "behavioural" approaches) to combining experts' assessments have also been applied in many contexts. Here, typically, a group of experts discuss and agree on some form of consensus probability distribution within a structured framework (Garthwaite et al. 2005). There are benefits and risks to either behavioural or mathematical aggregation techniques, both practically and philosophically; however, both are possible and the choice in practice on which to use is context-dependent (EFSA 2014).

Bayesian approaches (Hartley and French 2021) for SEJ start with the formalisation of a prior probability representing the decision-maker's belief ahead of hearing from the experts. Experts' judgements are then treated as *data*, and appropriate likelihood functions are created to represent the information inferred from their stated judgements. Bayes' theorem is applied to combine the prior with the elicited judgements on the uncertainty, to give the decision-maker's posterior perspective given the experts' statements. Calculation of the, potentially very complex, likelihood function was one of the key challenges that made early Bayesian models intractable.

Bayesian methods are starting to become more tractable with the advent of more effective computational approaches, particularly Markov Chain Monte Carlo (MCMC) (Wiper and French 1995; Clemen & Lichtendahl 2002; Lichtendahl 2005; Albert et al. 2012; Billari et al. 2014). At the same time, many of the principles early Bayesian models sought to highlight, e.g. expert to expert correlation, have not been explicitly tackled within existing non-Bayesian models. Thus the time is right to more formally assess these new Bayesian frameworks versus current approaches in an aim to build their credibility with decision-makers.

One of the key characteristics of Bayesian models is that they can utilise a parametric structure and thus infer a final posterior parametric distribution to represent the decision-maker's belief given the experts' judgements. This is a motivating factor for considering Bayesian frameworks for mathematical aggregation in SEJ. Opinion pooling techniques result in non-parametric representations of the consensus output. SEJ outputs are often used as inputs to broader parametric models and thus having the consensus in a parametric form can be very powerful. Another motivating factor for considering Bayesian models, in addition to the ability to encode more complex dynamics such as expert to expert correlation, is the ability to specifically incorporate prior knowledge into the process. If we are deploying SEJ in contexts where there is a well-defined decision-maker, the Bayesian approach can help understand how her unique position changes given the experts' inputs. Note, in some cases, SEJ will be used to act as a "rational scientist". In these cases, informative priors may be inappropriate and the model can be adjusted to take "naive" priors, whereby nearly all the information encapsulated in the posterior comes from the experts' judgements as the priors have been selected to be intentionally uninformative.

Bayesian models are often structurally context-dependent, and many models have been utilised in only a small number of settings, eliminating the possibility of a broad meta-analysis. Building on the work of (Lichtendahl 2005; Albert et al. 2012) and (Billari et al. 2014), within this chapter, we have considered a Bayesian framework applied retrospectively to existing SEJ studies. This allows us to generate predictions against a common data set and compare existing models accordingly. We recognise

this does necessitate some compromises within the Bayesian paradigm which may limit efficacy (such as no input during expert elicitation on parametrisation or prior selection), however, does set a benchmark for the use of generalised Bayesian models within SEJ.

It is important to note that any Bayesian approach may suggest a different procedure to the elicitation and documentation of SEJ studies (Hartley and French 2021; EFSA 2014; Cooke et al. 2000). Our aim in this chapter is to demonstrate some practical applications of the Bayesian framework utilising the database[1] of studies compiled by Cooke and Goossens (2008) and provide some performance comparisons with the classical model. Performance assessments are used to build the case for the feasibility of generalised Bayesian frameworks and to provide evidence that such a framework could be a credible choice for a decision-maker. Whilst not a primary focus within this chapter, we shall to a lesser extent note some of the more procedural elements that are important when considering Bayesian approaches.

5.2 Overview of the Bayesian Model

Bayesian approaches treat experts judgements as *data* and then create appropriate likelihood functions to represent the information implicit in their statements. The main complexity in applying Bayesian methods relates to:

- the experts' ability to encode their knowledge probabilistically and their potential for overconfidence (Clemen & Lichtendahl 2002; O'Hagan et al. 2006; Hora 2007; Lin and Bier 2008);
- shared knowledge and common professional backgrounds which drives correlation between expert' judgements (Shanteau 1995; Mumpower and Stewart 1996; Wilson 2016; Hartley and French 2021);
- correlation that may exist between the experts judgements and the decision-makers own judgements (French 1980);
- the effects of other pressures which may drive bias. These may arise from conflicts of interests, fear of being an outlier, concern about future accountabilities, competition among the experts themselves, more general psychological biases, and emotional and cultural responses to context (Hockey et al. 2000; Skjong and Wentworth 2001; Lichtendahl and Winkler 2007; French et al. 2009; Kahneman 2011).

The Bayesian perspective makes it clear that one needs to think about correlation between experts' judgements due to shared knowledge; other approaches to aggregating expert judgements do not. As any statistician knows, ignoring dependencies

[1]This database is constantly growing as studies are completed. To give an indication of scale, when the Eggstaff OOS validation analysis (Eggstaff et al. 2014) was conducted, 62 datasets were evaluated. These sets included 593 experts and 754 seed variables which resulted in 6,633,508 combinations and 67,452,126 probability judgements. A subset of these are considered within this chapter for cross-validation of the Bayesian approach.

between data leads to overconfidence in estimates. The same is true here, although we have noted that allowing for correlations between experts has been a considerable hurdle to the development of practical Bayesian methods.

The Bayesian framework we have employed simplifies some of the inherent complexity by breaking the post-processing into four distinct steps:

- Expert clustering
- Distribution fitting
- Recalibration
- Aggregation

The method is applied to judgements in the form used by the classical model. Here, estimates are elicited for both the *target* variables of interest and for *seed* variables, for which the analyst conducting the study knows the values a priori but the experts do not. These variables can be used as a calibration dataset within the Bayesian paradigm but are used by the classical model in order to calculate the performance weighting scores. All elicitations are made against a standard set of quantiles (typically three–0.05, 0.5, 0.95 or five–0.05, 0.25, 0.5, 0.75, 0.95).

We will not give a full mathematical exposition of our Bayesian framework here; however, we shall outline some of the key components behind each of the above steps to help with the analysis later on. For a full mathematical background, please consult Hartley and French (2021).

5.2.1 Expert Clustering

One of the risks leading to overconfidence in a final posterior comes from the shared knowledge or common professional backgrounds that experts may have which drives correlation. As we outlined before, finding such an underlying correlation and correcting for it is often a challenge for Bayesian models. One approach to bypass the issue of directly calculating complex correlation matrices would be to identify the sources of the underlying similarity in estimation and with this knowledge cluster experts into homogeneity groups in which all experts with similar historic knowledge are grouped together. As part of the aggregation exercise, this knowledge could be utilised to reduce the risk of overconfidence (Albert et al. 2012; Billari et al. 2014). One approach to forming these groups would be to attempt to elicit information about potential sources of common knowledge, in addition to the quantiles, from the experts. This approach is appealingly simple and would require only a procedural update. In practice, however, this elicitation is likely to be challenging as sources of this correlation may be opaque, even to the experts themselves. Thus, algorithmic approaches, which attempt to infer these groupings, could be considered.

The framework we have employed, similar to (Billari et al. 2014), utilises algorithmic clustering techniques in order to group and re-weight experts. Given the classical model data structure, there is a choice of data set to use for the clustering exercise, the target variables, the seed variables, or a combination thereof. We have chosen here to

use the seed variables. If there is underlying correlation between experts, driven by their shared knowledge, then this correlation should be apparent in their seed variable estimations. If there is no such link on the seed variables, then we would argue that there is limited risk of overconfidence on the target variables. This is clearly true only if the seed variables are within the same domain as that of the target variables, i.e. shared knowledge of experts in rare genetic conditions within hamsters does not imply shared knowledge in the risk of a bolt breaking within a suspension bridge. Representativeness of seed variables is similarly a core tenet underlying the use of these variables within the classical model. Please note that definition of meaningful seed variables is not easy, and there are those that would question the use of these variables altogether, although extensive cross-validation literature on Cooke's model does demonstrate their value (Colson and Cooke 2017; Eggstaff et al. 2014; Lin and Cheng 2009; Flandoli et al. 2011). We will leave this aside for now, however, and note that similar to Billari et al. (2014), the target variables could have been used in their place. Given seed variable estimations, it is easy to apply any number of clustering algorithms to the seed variable space (in which each expert is a point) in order to generate the expert groupings. We recommend utilising either hierarchical clustering, due to its efficacy over sparse datasets (and easy comprehension by decision-makers) or mixture models, specifically, Dirichlet process mixture models, due to their limited assumptions about the number of groupings a priori and their ability to integrate easily with the broader Bayesian framework (Billari et al. 2014).

5.2.2 Distribution Fitting

Within Bayesian frameworks, it is common for distribution fitting to be utilised in order to apply parametric models in the post-analysis of experts' assessments. This both makes the computation simpler and aligns with the assumption, in many practical applications, that underlying phenomena are parametric in nature. One of the benefits of a Bayesian approach is this parametric form. Often the outputs of an SEJ study can feed further analysis and having a fully parametrised posterior distribution can make calculations of future models much simpler. Opinion pooling method outputs are typically non-parametric.

Due to the complexity of eliciting experts' expectations on parameters, it is often preferable to elicit on observables first and then parametrise *post hoc*. Ideally, this would be done in conjunction with the experts (similar to behavioural SEJ approaches) to ensure that they are comfortable with the final statement about their beliefs; however, as we are applying this analysis retrospectively, this is not feasible. Thus, a choice must be made of which parametrisation to use. The aim of any fitting process must be to select a distribution which minimises the discrepancy to quantiles elicited from the experts in order to ensure that the fitted distribution reasonably reflects their underlying beliefs. Commonly used distributions within SEJ are the Gaussian distribution (Albert et al. 2012; Billari et al. 2014), the log-normal distribution (de Vries and van de Wal 2015) or a piecewise distribution which is uniform on the interior

quantiles and exponential on the tails (Clemen & Lichtendahl 2002). All of these distributions have advantages and disadvantages, e.g. in fitting the log-normal distribution, assumptions must be made on the un-elicited minimum and on the exponential behaviour post the top quantile, or, in the Gaussian, assumptions of symmetry about the mean. To this extent, we have chosen to utilise a two-piece Gaussian distribution (a Gaussian distribution with different variances above and below the median). This choice allows exact fitting to the expert quantiles, with minimal points of discontinuity and no assumptions on the extremities. It is, however, admittedly an *ad hoc* choice and our framework is generic, and thus could be applied to many parametrisations. The impact of different parametrisations on the final decision-maker posterior is an area for further research. If this Bayesian framework were to be applied to a study from the offset (rather than utilising data *post hoc* as we are doing here), then discussions about the appropriate distributions to use should be had with the experts.

5.2.3 Recalibration

Bayesian models typically consider the topic of recalibration differently to frequentist approaches. In the Bayesian model, as probability is subjective and thus a property of the observer (typically the decision-maker) of the system, it appears reasonable, for any such observer to consider all the information at hand in forming their final posterior distribution. An example of such information may be any bias which the experts have exhibited in historic judgements. Many potential drivers of bias, such as anchoring (Kahneman et al. 1982; Kahneman 2011), can be minimised through elicitation procedure (Cooke et al. 2000). Others, such as consistent over/under confidence, are often still visible (Burgman 2015). Thus if expert A, from a pool of experts, has historically been systematically overconfident, a Bayesian decision-maker may choose to broaden the tails in expert A's elicited judgement distributions, before aggregating with other experts, in order to truly reflect the decision-maker's belief of the uncertainty. Please note that there is significant resistance to this form of recalibration in certain areas with the argument that you should not adjust the experts forecasts as this creates an ownership problem (effectively the forecasts are no longer the experts' once you have adapted them, they belong to the analyst) and an accountability issue accordingly.[2] We would argue that the use of recalibration is context-dependent. In expert judgement problems with a single decision-maker, it would potentially be remiss to ignore any such information about potential additional uncertainty. Regardless, the model we have used is modular in design and recalibration could be included or excluded as appropriate given the context of the

[2] If it is assumed that experts are operating as coherent subjective Bayesians (Finetti 1974; De Finetti 1957), then there are mathematical inconsistencies with certain forms of recalibration (Kadane and Fischhoff 2013; Lichtenstein et al. 1982). However, there is evidence of incoherence among expert judgements within the Delft data, even on the small number of elicited quantiles. The exact form of calibration we are employing is also explicitly excluded from the mathematical analysis in Kadane and Fischhoff (2013).

problem at hand. Significant overconfidence is apparent in many studies within the Delft database; thus, this analysis has included recalibration. Further work should be conducted to empirically assess the impact of this recalibration. One approach to this would be to conduct the same analysis outlined in this chapter using both calibrated and un-calibrated Bayesian approaches. This is left for further research.

Seed variables, used for the performance weighting calculation within the classical model framework, can be used for the quantification of bias adjustments. These variables, elicited from the experts with true realisations known by the facilitator a priori, allow an analyst to identify if there are any systemic biases between prediction and realisation which need to be eliminated. The approach we use for this, taken originally from Clemen & Lichtendahl (2002), is to identify "inflation" factors which are multiplicative parameters inferred from the seed variable estimates and their realisations. In the case of a study in which three quantiles are elicited, there are three multiplicative parameters. The first of these is a positioning inflation parameter that assesses if there is consistent over or under forecasting of the *median* assessment. The other two parameters are then multiplicative dispersion parameters. These are calculated on the distance (or in the case of the two-piece Gaussian, the standard deviation), defined by the gap between the median and the upper/lower estimates, respectively. In this way, the dispersion inflation parameters control for any systemic bias in over- or underestimating the *uncertainty* in the judgements the experts give. Posterior estimates for these multiplicative[3] inflation factors can be inferred for each expert by starting with the assumption that they are well calibrated and then utilising the seed variables provided in the study as data and passing each through Bayes' rule. In practice, we also strengthen the analysis by allowing an inflation factor dependency structure between experts. We infer this through hierarchical models and MCMC as per Clemen & Lichtendahl (2002).

5.2.4 Aggregation

Once the set of expert homogeneity groups (H) and a final set of individual experts' judgements on the target variables (which have been recalibrated and fit to appropriate parametric distributions) have been confirmed, we can combine these to create a final posterior through aggregation. In order to do this, we utilise a hierarchical model, first proposed in Albert et al. (2012), which includes a novel approach to capturing the dependencies between experts. The aggregation model assumes that each expert's parameterised beliefs, derived from the elicited quantiles, are linked to that of the other experts in their group through a common shared group distribution. Each group will have a parametrised distribution, with parameters defined by the

[3]Utilising multiplicative inflation factors in this way does put constraints on the scales of variables (both seed and target) as it assumes that all variables are of a similar order of magnitude. If we imagine some variables are logged, then this form of recalibration would not work. This is currently a restriction with this framework and more research is required into potential solutions, although one possible approach is outlined in Wiper and French (1995).

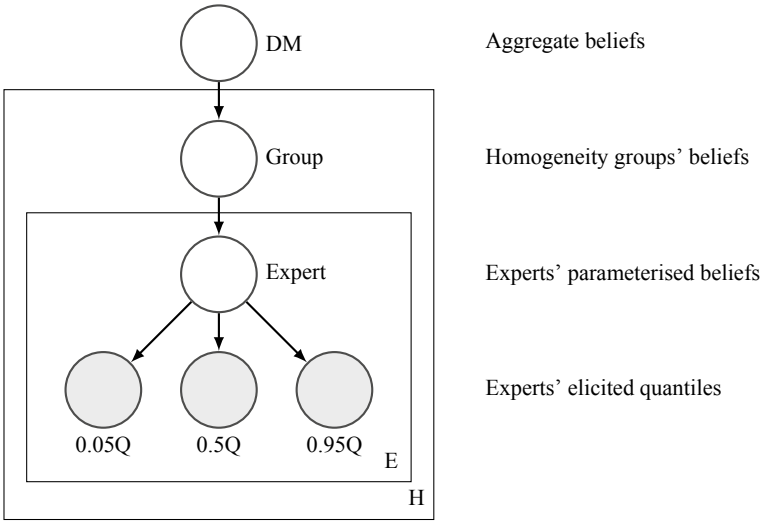

Fig. 5.1 A belief hierarchy for aggregation of expert judgement with homogeneity groups (DM - decision-maker)

combined beliefs of its expert members. The groups likewise are linked to each other via a common shared universal distribution that of the Supra-Bayesian. The final combined posterior distribution represents the updated decision-maker judgement and is calculated through MCMC. A simple diagram of this model is shown in Fig. 5.1.

The motivation for this expert partition is that rather than explicitly calculating the correlation matrix, the grouping approach is used to appropriately weigh the impact of each expert in the final model, offsetting overconfidence effects driven by correlation.

One of the advantages of this approach is that the hierarchical model can capture both the underlying consensus and diversity between experts. Opinion pooling methods do not attempt to assess consensus of opinion. Additionally, hierarchical Bayesian models of this nature allow inference not only the posterior distributions of the target variables but also all of the other latent elements within the model, such as inter-group dependencies. To this extent, it is possible to recover after the analysis has been completed, all of the homogeneity groups' beliefs as well as the parametrisations used for the individual experts. This gives the analyst a diagnostic tool to help understand how uncertainty has propagated through the model.

In order to combine each of the above four steps within our framework, we utilise MCMC. We have chosen to do our grouping utilising agglomerative hierarchical clustering (Charrad et al. 2014) within this chapter. This is to ensure deterministic group definitions given the significant number of predictions made. This means we have a two-step process, one step to create the necessary clusters and the second to do the parametrisation, recalibration and aggregation, which are all done within a single piece of MCMC code. If we utilise a Bayesian hierarchical clustering algorithm (such

as Dirichlet process mixture models), then this allows us to do all four steps within each iteration of the MCMC algorithm. This is philosophically appealing as it means we only use each piece of data once (seed variable information is used twice in the two-step model) but proves less stable with very small datasets and less intuitive to the decision-maker. To this extent, the choice of whether to use a one-step or two-step process is context-dependent. For a full mathematical exposition of the Bayesian framework, and to understand how the four elements are combined together, a review of Hartley and French (2021) is recommended.

Given this framework, it is interesting to understand how results compare to those of Cooke's classical model. Thanks to the open availability of historic Expert Judgement Studies, and the underlying data that Roger Cooke has kindly provided, we see how results differ from classical model outputs. We will start by doing a deeper dive into a couple of specific examples within geology and environmental resource management before looking more generically across the breadth of studies within the Delft database.

5.3 Effusive Eruption

Following the eruption of the Icelandic volcano, Eyjafjallajökull, in 2010 a scientific emergency group (SAGE) was appointed by the UK government. One of the tasks of this group was to consider the potential of future eruption scenarios that may impact the UK, and volcanic eruptions were subsequently added to the UK National Risk Register. One of the key scenarios adopted by the UK National Risk Register was considering the eruption of the Grimsvötn volcano (commonly known as the Laki Eruption due to its presence within the Laki crater) which occurred in 1783–84. This volcano had a huge impact on Europe, particularly in Iceland where 60% of the grazing livestock died (predominantly by Fluorosis) and 20% of the Icelandic population were also killed as a result of illness, famine and environmental stress. This eruption was considered to represent a "reasonable worst-case scenario" for future eruptions.

Risk to the UK from such a scenario recurring would be in the form of volcanic gases, aerosols, acid rain and deposition of acids. These factors can have significant environmental impact (due to deposition on vegetation, buildings and potential impact to groundwater), or impact on transport, particularly aviation (as we saw with Eyjafjallajkull), where sulphur dioxide and sulphuric acid can cause damage to airframes and turbines, engine corrosion, or put crew and passengers at risk of exposure. To model this complexity, meteorological (weather and atmospheric transport) models, in addition to chemistry models, are considered. In order to support this modelling and determine a set of prior values for some of the source characteristics, an expert judgement study was conducted in 2012 (Loughlin et al. 2012) (note: this study followed an earlier study conducted on the same topic in 2010 (EFSA 2010)).

Structurally, the elicitation was conducted with 14 multidisciplinary experts. Experts were from academia, research institutes and other institutes with opera-

tional responsibilities. These experts were able to cover all of the modelling fields described earlier (meteorology, atmospheric dispersion, chemistry) in addition to specific volcanology expertise. Quantitative responses were captured for 8 seed variables, alongside 28 target variables (22 volcanological in nature and 6 related to plume chemical processes). Not all questions were answered by all experts, with number of responses for each variable coming from between 5 and all 14 experts. For comparisons between the Bayesian framework and the classical performance-weighted model, we shall consider only the 10 target variables which were responded to by all experts and for which details are captured within the Delft database (Cooke and Goossens 2008).

Seed variables experts were asked to quantify were related to the historic Laki eruption, e.g.

- What was the area of the Laki Lava flows in km^2?
- What was the estimated production of Laki in CO_2 megatonnes?

With true realisations of 500 km^2 and 349 megatonnes, respectively, an example target variable question was:

- What is the likelihood that in the next Laki-like eruption there is an episode which releases 10 times more SO_2 on the same timescale as the peak eruption episode during Laki?

With other questions similarly linked back to the Laki eruption, this link is important as it helps ensure that the seed variables are truly representative of the target variables and are thus suitable for use within the recalibration exercise inherent within the Bayesian model (and likewise for appropriate performance weighting in the classical model).

Across the total 112 seed variable estimations, if we were to a priori assume that experts were well calibrated/statistically accurate, we should expect to see 11–12 of the seed variable realisations sitting outside the range given by the 0.05 and 0.95 quantiles provided by the experts. Individual experts would expect to have no more than one judgement where the true realisation sits outside of these bounds. In practice, actually 64% (72) of the true realisations fell outside of the 90th percentile bounds given by the experts. For individual experts, between 37.5% and 100% of realisations fell outside of the confidence bounds given (Fig. 5.2). These results are potentially shocking to the uninitiated and may appear to point to a lack of true "expertise" of the experts in the panel, in practice, however, these types of numbers are not uncommon for judgements within SEJ (Burgman 2015) and reflect the complexity of the underlying dynamics within the contexts in which SEJ operates (hence, the need for judgement in the first place). What it does point to, however, is a cautionary note for decision-makers that experts can often be, and in this case are demonstrated to be, systemically overconfident in their judgements. This furthers the case for recalibration, without which, further uncertainty driven by this overconfidence will be ignored.

Running the classical model over this dataset results in 3 experts getting a weighting (Expert 10–53%, Expert 14–31%, Expert 12–16%) and 11 experts being removed

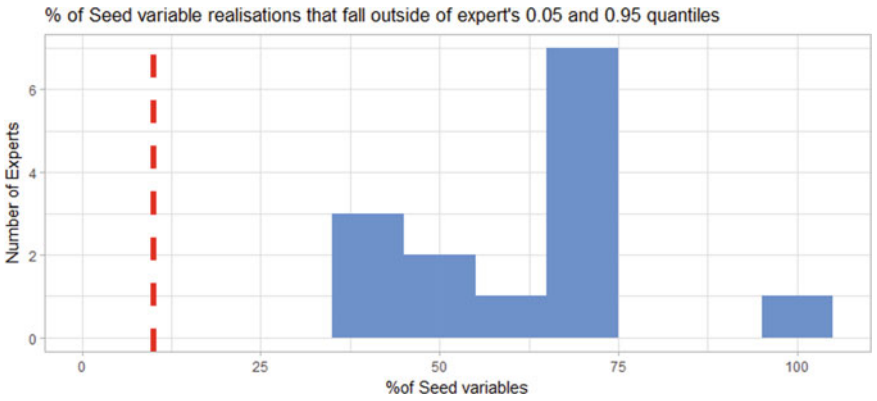

Fig. 5.2 Across seed variables within the Laki effusive eruption scenario study, experts demonstrate significant overconfidence in judgements. All experts have more variables outside of the 0.05 and 0.95 quantiles than would be expected for high statistical accuracy. Red line indicates the expected % of variables for a perfectly calibrated expert

from the final CM optimised decision-maker quantile calculation altogether.[4] To compare the impact of this to the Bayesian framework, we can first consider the homogeneity groups that are created as a result of the first step within the model, the clustering exercise. Running this process identifies five core homogeneity groups within the expert pool (Fig. 5.3), of which two are formed of a single expert and three

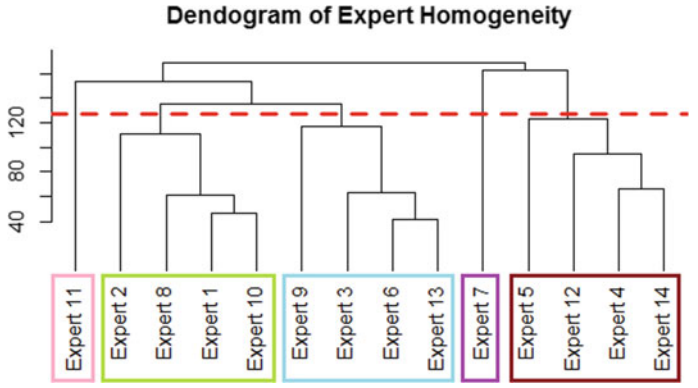

Fig. 5.3 Experts are clustered into five homogeneity groups (coloured boxes) demonstrated by a horizontal cut on the dendogram (red line). These five groupings, based on seed variable responses, cluster the experts into three groups of four individuals and two outliers (Experts 11 and 7) who sit within their own homogeneity groups as they regularly offer differing opinions to the other experts

[4]Please note that, whilst not included in the final calculation of the quantiles within Cooke's methods optimised volcanology, these experts' assessments are still involved in determining the intrinsic range of the random variables.

are each of four experts. The two experts who are grouped within their own pools have done so as a direct result of a significant divergence in judgement between themselves and the remaining groups as it pertains to the seed variables. Thus, the Bayesian model identifies that there is the potential for discrepancy in opinion on the target variables that should be considered and upweights these individuals relative to their peers. In this way, the Bayesian model is capturing the diversity of thinking across the experts. Please note that, at this stage, no judgements have been recalibrated; thus, we do not yet know whether this diversity is a result of different mental models by these experts or due to miscalibration. The recalibration exercise ensures that experts are well calibrated before aggregation, and thus we minimise the risk of simply up-weighting a "poor" forecaster.

Expert judgements are subsequently passed through the distribution fitting, recalibration and aggregation processes described earlier to create a single decision-maker posterior distribution. It is important here to reflect on the context of the analysis that we are conducting, and hence the decision-maker we are trying to model. In the case of this volcanology study, there is not an individual decision-maker whose belief is being updated by the experts. The study is being conducted in order to arrive at a consensus distribution which reflects that of a rational scientist. As a result of this, we need to be thoughtful about the choice of priors that we use in our model. As outlined in Hartley and French (2021), in rational scientist scenarios, there is no individual decision-maker with a significant a priori belief thus we recommend using diffuse priors to minimise the impact of the analyst on the output. The only exception to this is on the median inflation factor, for which we set a tight prior, centred around 1. This assumes that experts are well calibrated on their median and ensures that there are only minor changes feasible to the centrally elicited quantile. However, the model is given freedom to adjust the upper and lower tails to mitigate the overconfidence seen earlier. If extensive changes to the median were allowed in the model, it could be argued that the judgements no longer reflect that of the expert, and thus the aim of achieving a rational scientific consensus would be compromised (note, in the context of an individual updating their beliefs, further recalibration of the median may be appropriate). Numerically, the set of priors considered for our model here are as outlined in Hartley and French (2021), where rational scientist consensus is also the goal. These priors were consistent across all of the analysis within this chapter. More extensive exposition of the considerations of priors within Bayesian SEJ models is outlined in Hartley and French (2021).

Before getting on to discussions regarding the uncertainty bounds provided by the Bayesian/Cooke's models, it is first interesting to assess differences between the posterior median for the Bayesian model and Cooke's optimised decision-maker's 0.5 quantile. In many contexts, final decision-makers will look to a point estimate from which to base their next best action. As the Bayesian model is trying to consider both the consensus and diversity in opinion, the hierarchical nature enforces unimodality in the posterior distribution. This posterior mode (which, due to the parametrisation used, will be located at the median) reflects the most likely single value a decision-maker would use to represent a point estimate. Whilst we recognise that ignoring uncertainty in this way is counter to the goals of risk management for which SEJ is

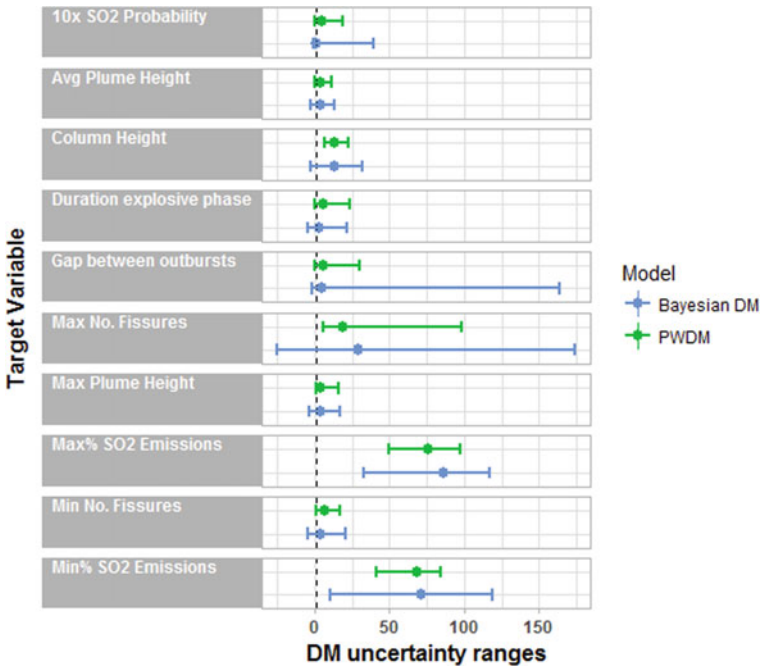

Fig. 5.4 The Bayesian model produces median estimates similar to that of the PWDM and always within the PWDM uncertainty bounds. The Bayesian decision-maker, however, suggests a higher level of underlying uncertainty

typically employed, the use of point estimates is a common decision-making reality and thus worth assessing. For brevity, outputs from Cooke's classical model will henceforth be referred to as the PWDM (Performance-Weighted Decision-Maker with optimisation).

Figure 5.4 outlines the final uncertainty ranges for each of the 10 target variables and each of the ascribed models.[5] It is important to note that across all of the distributions the median for the Bayesian model sits within the uncertainty bounds of the PWDM. This is reassuring. Given the extent to which the PWDM has been utilised in practical studies, if there were fundamental concerns on this number these are likely to have been surfaced before. This suggests that a decision-maker considering either model is not likely to make a significantly different decision based on the expected value alone. There is a noticeable difference, however, in the ranges given by the two models. The PWDM has consistently narrower bounds than the Bayesian decision-

[5]Table 5.3 and Fig. 5.13 in the Supplementary material outline the same responses but also include the results from considering an equal-weighted linear opinion pool (EWDM), omitted initially for brevity and clarity. Comparisons versus the EWDM are considered later when assessing the distributional forms as this provides greater clarity on the difference in modelling approaches than simply the uncertainty bounds alone. All EWDM results have been taken directly from the tool EXCALIBUR used to calculate the PWDM optimised DM.

maker. This is as we would expect and is caused by two predominant factors. Firstly, the PWDM selection criteria, by design, are optimising for statistical accuracy *and* information, and will often trade minor reductions in statistical accuracy for significantly improved information. This occurs due to the fact that information is a slower responding function than statistical accuracy. Secondly, the Bayesian decision-maker is recalibrating the experts' judgements. Given that experts have demonstrated overconfidence, the decision-maker has correspondingly increased uncertainty ranges.

Please note that some of the posterior uncertainty ranges for the target variables within Fig. 5.4 and Table 5.3 modelled using the Bayesian framework are demonstrating infeasible values (e.g. for variable 6, the maximum number of fissures, as a result of the recalibration exercise, negative values have appeared. For obvious reasons, it is not possible to have negative values for such a variable). The Bayesian framework does allow variable constraints to be considered, and the key is the constant of proportionality within Bayes' rule. This constant allows us to apply such bounds *post hoc*, by removing the infeasible area and rescaling accordingly. The other approach would be to consider the framework utilising distributions which are inherently constrained, for example, utilising a beta distribution if the target variable is a percentage as this is naturally constrained to the interval [0, 1]. Neither adjustment has been performed here as the focus is on highlighting the impact of different modelling approaches at a macro-level although these considerations are very important when applying the framework in practice.

Whilst the median and the uncertainty bounds themselves are critical, it is also important to understand the shape of the final decision-maker distribution for each model. Figures 5.5 and 5.6 outline two such distributions, selected as these show different behaviours of the models. The equal-weighted decision-maker (EWDM) distribution has also been added to these slides for comparison. The equal-weighted decision-maker is the result of a linear opinion pool with identical weighting given to each expert.

Target variable 3 (Fig. 5.5) demonstrates common behaviour of the Bayesian model versus Cooke's performance-weighted approach and the equal-weighted decision-maker, notably, a single modal point with a Gaussian decay in either direction (rather than multi-modality), narrower shoulders, and a broader support. One of the outlined aims of the Bayesian framework is to identify underlying consensus in opinion from the experts, and thus this distributional shape is by design and reflects a starting assumption that the Supra-Bayesian decision-maker's belief is of this form. Note that this is a decision to be made in setting up the model, and the framework is generic in nature to support many other possible parametrisations. Broader support is driven by the recalibration portion of the model and the overconfidence displayed by the experts on the seed variables. If experts were systemically under confident, we would expect to see narrower tails on the Bayesian model. In practice, overconfidence is much more common (Burgman 2015).

Target variable 10 (Fig. 5.6) demonstrates a slightly different picture; here, once again the modal point of the Bayesian model is similar to that of the PWDM, and the unimodal shape is maintained. However, in this instance, the EWDM is demonstrating a slightly different picture of the uncertainty. Both the PWDM and the Bayesian model

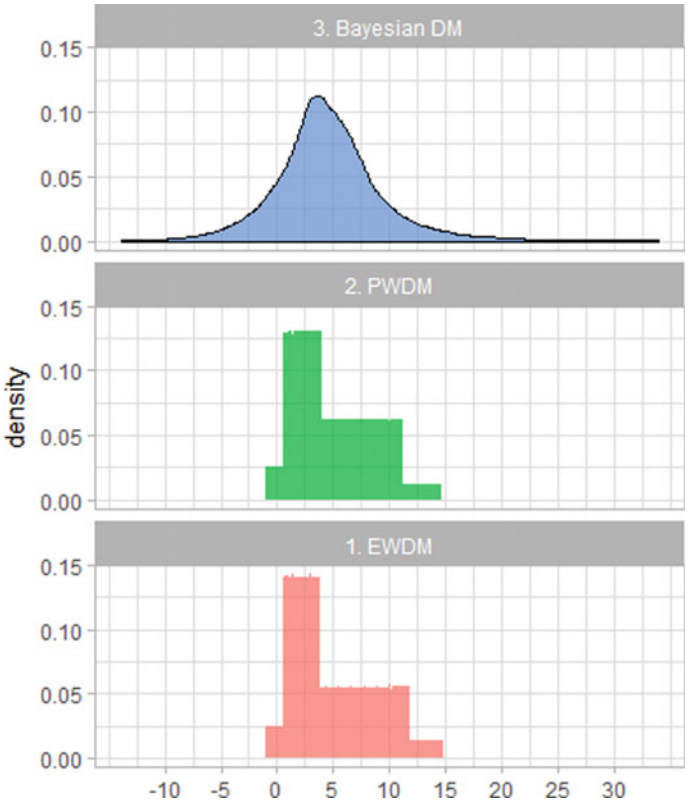

Fig. 5.5 Final distributions for target variable 3: 'After the initial explosive phase (i.e. first days), what is the likely average sustained plume height for gases above the vent for the remainder of the active episode?' Note the similar shapes between the PWDM and EWDM distributions. The Bayesian model demonstrates a slightly higher modal point, and more uniform shape as it is focussed on the underlying consensus in opinion. Note also the larger support of the Bayesian decision-maker as this recalibrates for overconfidence

put a very significant amount of the density at the modal point with little probability to a value below this and a limited but positive probability of more extreme values. The EWDM, however, has significantly less mass around the modal point, a larger probability of a lower realisation and more significant density in the upper tail. It is a positive sign that similar distributional shapes are visible here for the Bayesian and PWDM as there is no a priori reason that this should be the case and suggests that they may both be pointing to similar underlying consensus between experts.

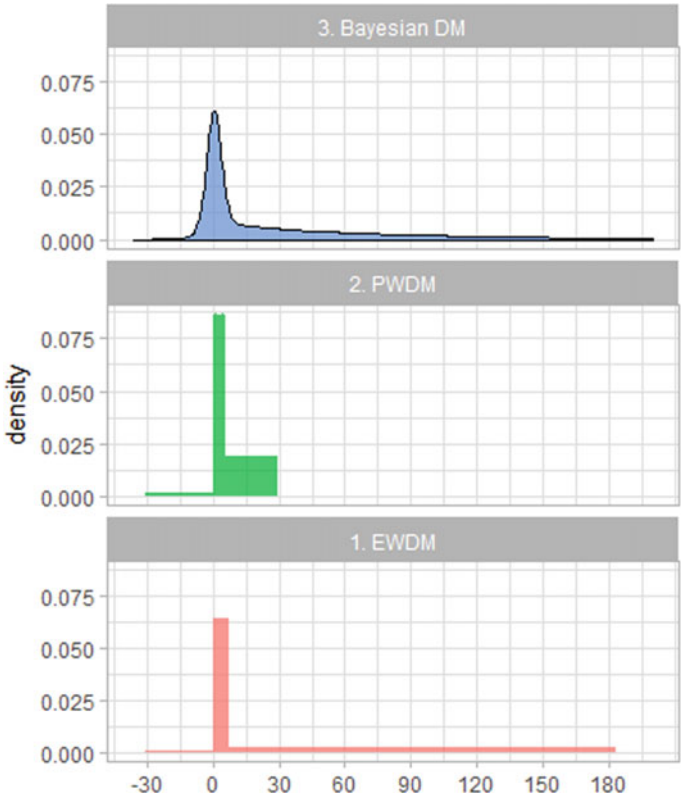

Fig. 5.6 Final distributions for target variable 10: 'What is the typical gap between major gas outburst episodes?' Note, similar to the PWDM (although not the same extent) the Bayesian decision-maker puts more of the distributional density in the region just greater than the modal point. Whilst the Bayesian model includes the greater range suggested by the EWDM, it tapers off much faster

5.4 Invasions of Bighead and Silver Carp in Lake Erie

Forecasting the likelihood of, and damage caused by, invasive non-indigenous species within many natural environments is difficult and poses a problem for those responsible for natural resource management. Similar to the context outlined before, often, the data necessary to build comprehensive decision models is incomplete, and thus expert judgement can be used to supplement what data is available. A recent study, (Wittmann et al. 2015, 2014; Zhang et al. 2016), utilised expert judgement through the classical model to forecast the impacts of Asian carp in Lake Erie. Asian carp is non-indigenous and currently believed not to be established within the lake. Assessments were made to quantify potential aspects of the Asian carp population (biomass, production and consumption) as well as impacts to existing fish species, in the instance that these carp become established within the lake. Establishment could occur as a

result of contamination of bait, release by humans or through waterway connections linked to currently established populations.

Structurally, the study comprised of 11 experts, each of whom was asked to assess 84 variables (20 seeds and 64 targets) within the elicitation questionnaire. In practice, for 5 of the seed variables, actual realisations did not become available and for 1 expert, only 11 of the seed questions were responded to. Hence, within this analysis, to ensure consistency across modelling approaches, these 5 seed variables and this 1 expert have been removed, to leave 15 seed variables and 10 experts. Please note that this selection choice differs from the original paper in which the expert was left in the study but a further four seed variables were removed. Elicitations were made against the standard three quantiles (0.05, 0.5, 0.95).

The clustering of experts defined by seed variable responses suggested three core homogeneity groups within the expert pool. The largest group consisted of six members (experts 3, 4, 7, 8, 9 and 10), the second group three members (experts 1, 2 and 5) and finally expert six sat within their own group as their responses consistently differed to those of the remaining groups, suggesting that they may be using a different set of reference data or mental models through which to base their judgements. Supporting a decision-maker in identifying why a particular homogeneity grouping may have arisen could be difficult as the space in which the clustering is performed may be high dimensional (in this case 15 dimensions). Principal component analysis can be used to reduce dimensionality and create a lower dimensional visual representation of the variation in responses between experts. Figure 5.7 outlines some key PCA outputs for this study and identifies the emergent groups in a visual way.

Similar to the prior study outlined, overconfidence was common in the experts across the Lake Erie study. 47 of the 150 (31.3%), seed variable/judgement combinations had realisations sitting outside of the bounds given by the experts. Significantly

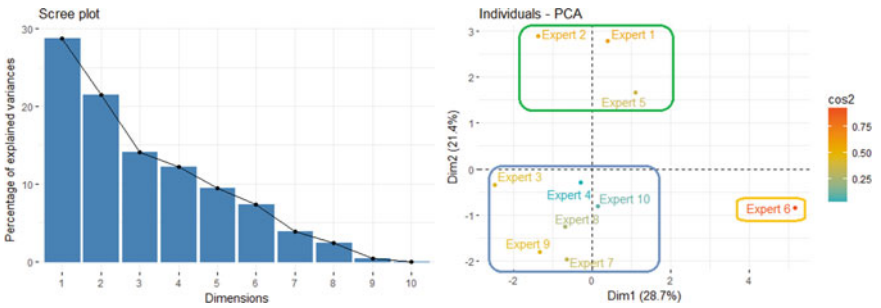

Fig. 5.7 A scree plot of the principal component analysis demonstrates that the first two identified components explain 50.1% of the variance across the original 15 dimensions within the seed variable space. When we isolate these two components and look at where the individual experts sit, the homogeneity groups identified by the model (coloured boxes) emerge. Expert 6 is separated from the remainder and thus is within their own homogeneity group, as they systemically give differentiated responses to the other experts. Note that groupings here are for visualisation purposes only; actual clustering occurs over the full set of seed variable dimensions

Table 5.1 Biomass levels (t/km^2) predicted for the Lake Erie study sole invader scenario

	EWDM			PWDM			BDM		
Target variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Peak biomass									
Bighead carp	0.0	2.4	17.2	1.6	8.9	25.9	0.7	4.2	13.0
Silver carp	0.0	2.3	17.0	1.6	8.8	25.9	0.7	4.1	11.9
Equilibrium biomass									
Bighead carp	0.0	1.2	9.1	0.4	3.0	12.2	0.3	2.0	6.2
Silver carp	0.0	1.1	8.0	0.4	3.0	12.2	0.3	2.3	6.8

more than the ~15 (10%), we would have expected assuming the experts were all well calibrated. Unlike the effusive eruption example given earlier, however, the range of calibration across the experts was broad. Expert 4 demonstrated strong statistical accuracy, only 1 of the realisations (7%) fell outside of the judgement bounds they gave. As expected this translates into significantly less recalibration within the Bayesian model. Expert 4 had the lowest recalibration parameters within the group. Classical model analysis of the study put all weighting to Expert 4, thereby effectively removing all other experts' judgements from the quantile aggregation within the PWDM optimised decision-maker. Note that, as before, all experts are still included in the calculation of the intrinsic ranges.

The key finding of the original elicitation was that given the right starting condition, there is significant potential for the establishment of Asian carp within Lake Erie. In particular, they have the potential to achieve a biomass level similar to some already established fish species currently harvested commercially or recreationally (yellow perch, walleye, rainbow smelt and gizzard shad). These findings remain when considering the final posteriors proposed by the Bayesian model. The Bayesian model estimations of peak biomass levels for bighead and silver carp, in scenarios where they are the sole invader, suggest higher medians than those predicted by the EWDM but lower than those predicted by the PWDM for both bighead and silver carp (Table 5.1).

Uncertainty ranges within the Bayesian model are slightly narrower than either of the other two. Equilibrium biomass estimates in the same scenarios were lower than peak biomass levels but displayed consistent behaviour between models. The PWDM estimates that the median equilibrium values were approximately 1/3 of the peak value compared to approximately 1/2 in the Bayesian or EWDM models. In the joint invasion scenario (where both bighead and silver carp are established), the Bayesian estimation of the equilibrium biomass was marginally higher than both the PWDM and the EWDM (Table 5.2). Final quantile estimations of the proportion of the total biomass that is bighead carp in the joint invasion scenario were identical in the PWDM and Bayesian models and marginally higher than that of the EWDM (Table 5.2 and Fig. 5.8).

Table 5.2 Estimates for the Lake Erie joint invasion scenario

	EWDM			PWDM			BDM		
Target Variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
Equilibrium biomass	0.0	2.2	12.3	0.4	3.0	12.2	0.6	3.6	10.4
Proportion Bighead carp	0.0	0.3	0.9	0.1	0.5	0.9	0.1	0.5	1

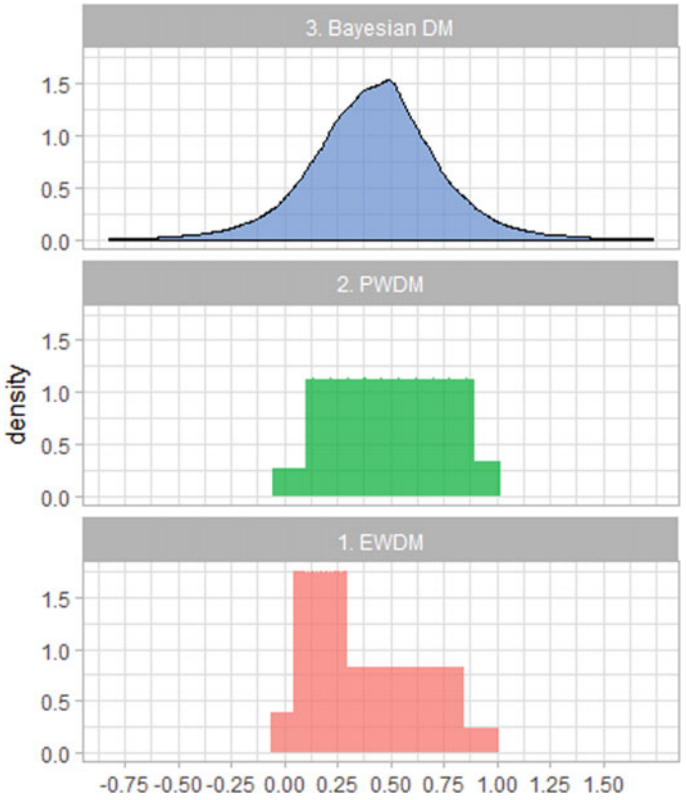

Fig. 5.8 Final distributions for Target variable 10 in the Lake Erie study: 'What is the proportion of the total biomass that is bighead within the joint invasion scenario?' The Bayesian model and the PWDM suggest a marginally higher proportion of the biomass will be the non-indigenous bighead carp than EWDM predictions. Note the narrower shoulders and broader tails of the Bayesian model, consistent with other target variable estimations

Overall, similar to what was seen in the effusive eruption case, the quantities of interest resulting from the Bayesian model do not vary significantly (where significance is defined as implying a radically different conclusion from the judgement data) from those of the PWDM, and in this case the EWDM. All models have suggested that there is significant potential for the establishment of tangible biomass of these carp, in relation to existing fish populations, although each model has demonstrated a slightly different posterior distribution of the uncertainty as they emphasise different underlying elements of the judgements.

This is reassuring for a new model, such as the Bayesian framework, as existing models have been used and tested extensively. If radically different values had been found, significant justification would be required.

In both this and the earlier effusive eruption example, we have seen that the median estimate was similar in the performance-weighted and Bayesian approaches. If we look at a broader subset of the Delft studies, specifically a subset where all variables are on a uniform scale to ensure the recalibration algorithm is applicable, we can assess the final decision-maker medians for each of the target variables. In total, there are 20 considered studies with 548 forecasted target variables. In Fig. 5.9, we can see that this similarity between median estimates is true more broadly as there is a strong correlation (0.82) between the final median estimates of the two approaches (0.99, removing outliers).

A log scale is used in Fig. 5.9 to allow us to compare across studies. Whilst variables are on a consistent scale within a single study, they may be on very different

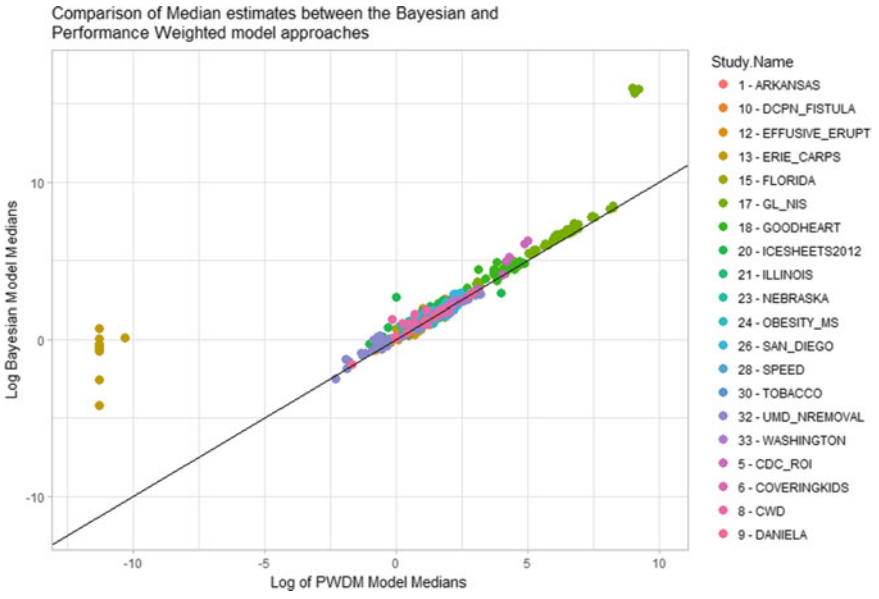

Fig. 5.9 Median estimates from the final decision-maker distributions are highly correlated between the Bayesian and performance-weighted approaches across studies within the Delft database

scales in different studies. There are a small number of outliers at either end of the plot, from the Lake Erie study and the GL_NIS study, where the Bayesian model has a final median of a significantly different order of magnitude to the PWDM. Those at the lower end, from Lake Erie, have been driven by the fact that the performance weighting approach selected a single expert who had a value for these variables many orders of magnitude lower than some of their compatriots who were also included in the Bayesian aggregation. On the upper end, the discrepancy is driven by a small number of target variables within the GL_NIS study whose estimated values were many orders of magnitude higher than other target variables. This will have broken the constraint of scale uniformity from the recalibration process within the Bayesian model and potentially projected higher than realistic values here. Aside from this small number of outliers, however, there is broadly good consistency between approaches on this median value. Whilst important for decision-makers, the median is not the only element that is being considered by those utilising the output of expert judgement studies, with the way that uncertainty is being expressed also of critical importance. A recent study on the impact of melting ice sheets, and the subsequent commentary papers, emphasised some of these considerations.

5.5 Ice Sheet Example

Climate change is one of the major issues of the current age and as such is an area where strong scientific insight is fundamental to building the case for necessary political decision-making and public behavioural change. Unsurprisingly, despite the wealth of geological datasets and sophisticated models, understanding the complexity in and predicting the outcome of many climate change problems relies heavily on expert judgement. Willy Aspinall, who performed the effusive eruptions elicitation study, alongside Jonathan Bamber, conducted a glaciological study to predict the impact of melting ice sheets, due to global warming, on rising sea levels (Bamber and Aspinall 2013).

This research has been extensively cited and has positively contributed to the ongoing debate regarding the appropriate use of expert judgement within the geological community. One commentary by de Vries and van de Wal (2015), and subsequent discussion papers (Bamber et al. 2016; de Vries and van de Wal 2016), assessed and questioned a number of key elements of applying the classical model in this context. In particular:

- The correct way of assessing the lack of consensus in the interpretation of post-processing the experts' answers.
- The reduction in the "effective" number of experts caused by the classical model weighting process.
- The choice of underlying distributions.

Please note. One of the other topics raised in the commentary paper was regarding the choice of variables to elicit from the experts. In this case, the primary elicited

variables reflected the experts' predictions on the impact to sea level rise from three separate ice sheets (East Antarctic, West Antarctic and Greenland) and were then combined *post hoc* to create a total sea level rise estimate utilising a Monte Carlo model. Questions were raised whether this would accurately reflect the experts' underlying belief of the final target variable as the choice of model can impact the total uncertainty bounds. Hence, it was suggested that the total sea level rise estimates should have been elicited explicitly. This is a question of study design and we will not tackle it here, except to comment that it is very common for expert judgement to be used both to make judgements on final decision-making variables, or variables which are then inputs into a broader model. The choice selected here may be largely context-dependent. One of the design elements of the Bayesian model (a fully parameterised posterior) is a support tool for decision-makers and analysts utilising the output of expert judgement studies as priors in other models.

Many of these topics are not unique to glaciology and have been commented on elsewhere in the literature with respect to the classical model. The Bayesian hierarchical model, by design, takes a philosophically different approach to each of these areas than the classical model. Thus, whilst it will not address all of the comments posed in de Vries and van de Wal (2016), it would be interesting to consider how the application of the Bayesian aggregation model to the same data performs relative to the performance-weighted approach.

In the effusive eruption and Lake Erie example, we compared some of the forecasts for target variables, however, made no comment as to the validity of the final estimates nor how this varies between models. To be confident in any forecast a decision-maker should have prior validation of a model's results. To this extent, we shall not assess the two models over the target variables within the ice sheet studies as we have done before but will look for ways of assessing how well the models perform (in this context and the previous studies) using some cross-validation techniques.

5.6 Cross-Validation

Measurable target variable realisations are uncommon within expert judgement studies due to the inherent rarity of events assessed or the lack of ethical means of collecting data. These are the same drivers which lead to the studies in the first instance. Consequently, standard models of assessing forecast accuracy, e.g. Out-of-sample validation, are rarely feasible. There are two primary concerns when designing a validation framework for expert judgement studies. Firstly, how to generate a significant sample of testing data for which there are both modelled aggregate judgements and realisations, and secondly, what testing methodology to use to assess the validity of the final distributions on this testing set.

Validation within SEJ models is relatively new, and there is much work to do in order to formally define an agreed-upon approach. Several different methods of building the testable set have been proposed, all of which fall into the context of cross-validation. Cross-validation involves taking only judgements about seed variables

(considered as these are the variables against which both judgements have been made and true realisations are known) and permuting through certain subsets of these, using each subset as a training set and then modelling the remaining subset. Clemen (2008) proposed such a technique using a method known as ROAT (Remove One At A Time). Here, each seed variable is removed from the training set one at a time, and all remaining variables are used to train the model, i.e. if there are S seed variables in the data set, each training set will be of size, S-1, and there will be S final forecasts. ROAT is a fast method of cross-validation as relatively few judgements need to be made; however, it was demonstrated that this method could have an inherent bias against a performance-weighted decision-maker (Cooke 2008). Other methods of cross-validation considered have utilised bigger training subsets (Colson and Cooke 2017; Lin and Cheng 2009; Flandoli et al. 2011) and (Cooke et al. 2014), although questions arose over the implementation of a couple of these studies as the numbers quoted did not align with those from Cooke's modelling platform EXCALIBUR (Cooke 2016; Cooke and Solomatine 1992).

Arguably the most comprehensive cross-validation of PWDM was outlined in Eggstaff et al. (2014). Within this cross-validation model, the authors considered every permutation of the seed variable partitions from 1 to S-1 (the code was also vetted against EXCALIBUR). Modelling this level of data showed strong support for the advantages of a PWDM model over an EWDM model but relied on an extremely large number of forecasts which would be a struggle to replicate at scale for other modelling approaches. Given that the number of subsets of a set of size n is 2^n, for a single study of 10 variables, this would create 1022 forecasted subsets (both the empty set and the complete set are removed). Given that each subset forecast within the Bayesian model can take a few minutes to complete, computation cost of this number of forecasts is very high (speed is definitely a distinct advantage of the PWDM over Bayesian MCMC approaches). Colson and Cooke (2017), whilst building on the work of Eggstaff et al. (2014), have recently recommended considering all permutations of training subsets 80% of the size of the original set of seed variables. This creates a manageable sized set of forecasts to perform whilst overcoming some of the biases in the ROAT methodology. For all subsequent analysis within this chapter, we have utilised this 80% methodology. For a study of size 10, there are 45 training subsets of size 8 and 90 resultant forecasts (two for each model run). If 80% was non-integer, we have shrunk the training set size to the nearest integer, and where necessary the minimum number of variables removed was set to 2 to ensure the methodology was not applying a ROAT process.

Methodologies for assessing the accuracy of the forecasts on the given testable sets also vary, and there is further opportunity for research and consolidation on an agreed approach here. One simple method that is considered in many studies is to ignore the uncertainty bounds within the model and simply assess the median within the distribution, utilising a metric such as the mean average percentage error (MAPE), assuming that this represents the most likely value a decision-maker would use in practice. This gives an indicative value on the discrepancy between the forecasts and the actual realisations for these point estimates but does not assess the full richness of the analysis conducted.

The aim of other methods is to quantitatively assess how representative the full distribution is compared to the observed phenomena, within the test sets. One such method is to consider a reapplication of the classical model itself (Eggstaff et al. 2014; Colson and Cooke 2017). Here, each modelling type is considered an "expert", the testable set is the set of seed variables and target variables are omitted. The performance measures (statistical accuracy and information) are then calculated for each model across all of the forecasted variables considered. Typically, for a given study, each subset is assessed in this way and then aggregate statistical accuracy and information scores are calculated for the total study by taking the mean or median of the scores for each subset. Geometric means are also calculated but the arithmetic mean is the value most commonly utilised.

Applying this cross-validation technique to the three studies, we have discussed within this chapter and using the 80% subset rule results in 527 separate subsets and 1529 individual forecasts. Statistical accuracy and information scores for these forecasts are then calculated within R. The R code was validated by taking a sample of these forecasts, rebuilding it from scratch within EXCALIBUR and ensuring consistency of the output. All numbers for Cooke's classical model (PWDM, and EWDM when relevant) have been drawn directly from Eggstaff et al. (2014) supplementary material, kindly provided by Roger Cooke, in order to ensure consistency. Mean statistical accuracy scores, Fig. 5.10, show that the Bayesian model (0.53 effusive eruption, 0.54 Lake Erie, 0.57 ice sheets) scored higher in each study than the PWDM (0.29, 0.45, 0.31). Conversely, Cooke's model (1.6, 0.85, 1.01) performed better than the Bayesian model (0.98, 0.38, 0.63) according to the information criteria outlined. This highlights exactly the behaviour we might expect to see, given the distributions we saw earlier, the fatter tails of the Bayesian model as a result of calibration, and the inherent trade-off made within Cooke's model.

Perhaps more surprising is the performance relative to the EWDM, which has also been included in Fig. 5.10 to provide another reference point. Across the studies

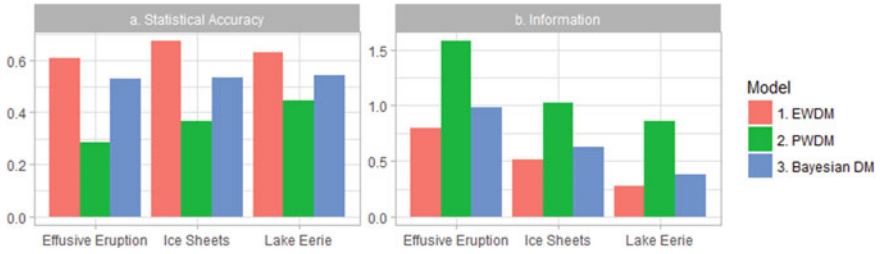

Fig. 5.10 Arithmetic mean of the statistical accuracy and information scores for each tested model across the three studies previously discussed. The Bayesian model typically demonstrates better statistical accuracy but lower information than the PWDM, as we would expect. Perhaps surprisingly, given the broader support, in these studies, the Bayesian model demonstrates higher informativeness than the EWDM. This is due to the Bayesian model having narrower shoulders. Please note that information is a relative measure and absolute informativeness numbers are not relevant cross studies and should only be considered across models within a single study

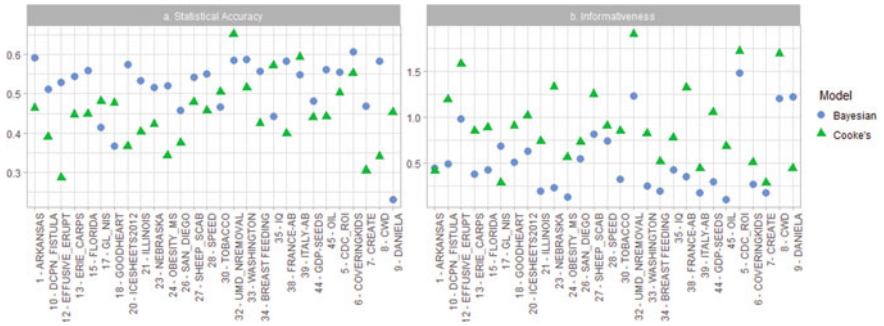

Fig. 5.11 Statistical accuracy and information plots for each analysed study within the Delft database. The Bayesian model demonstrates consistently higher statistical accuracy than the PWDM but lower information scores

outlined, the Bayesian model had higher information scores than the EWDM (0.80, 0.28, 0.52) but lower statistical accuracy scores (EWDM; 0.61, 0.63, 0.35). This may seem counterintuitive, as the Bayesian model has been specifically recalibrated, whereas the EWDM has not. The reason for this behaviour is the consensus focus that the Bayesian model has, rather than diversity which is emphasised in the EWDM approach. Despite the fatter tails, by looking for a consensus view, the Bayesian model typically has narrower shoulders than the EWDM, as we have seen in some of the earlier distributions, e.g. Fig. 5.5. Narrower shoulders are likely to reduce the statistical accuracy score but increase information. In this way, the Bayesian model is also trading off between statistical accuracy and information. If we were to only apply the recalibration component of the Bayesian framework and then aggregate utilising the EWDM, we should expect to see the highest statistical accuracy scores but the lowest information of any of the models discussed so far.

Expanding the cross-validation technique to the broader set of studies within the Delft database can help us ascertain whether we see the above behaviour consistently. As the recalibration within the Bayesian model cannot currently deal with variables on different scales, a subset of 28 studies which were utilised by Eggstaff and within the Delft database were considered. In each considered study, all of the variables were of similar order of magnitude. Study names align with those in the original paper. In total, this equated to 2706 forecasted subsets and 6882 individual variable forecasts (Fig. 5.11).

The Bayesian model outperformed the PWDM on statistical accuracy in 71.4% (20 out of the 28) studies based on the arithmetic mean. The PWDM outperformed the Bayesian model in mean informativeness in 93% (26 of the 28) of the cases. This is reassuring as it demonstrates that the model behaves consistently across studies relative to the PWDM and aligns with what we saw earlier. One of the studies (Study 9—DANIELA) is clearly an outlier with an extremely low statistical accuracy and high information for the Bayesian model. This is because there was a convergence issue with this model, believed to be due to the combination of a low number of

experts (4) and seed variables (7, given the holdout sample, only 5 of which would be included in each subset). Whilst more work is necessary to understand the impact of the number of seed variables on expert judgement models, performance weighting guidelines suggest that at least 10 seed variables are considered. Bayesian models with recalibration will similarly require minimum numbers to reach appropriate convergence which meaningfully reflects underlying expert bias.

The above analysis highlights that the Bayesian model and PWDM model are trading off between statistical accuracy and information to different degrees. We would argue the choice of which model to use in practice for a decision-maker may depend on the context in which the study is being performed and the sensitivity of the decision they are making to either information or statistical accuracy. To get a better sense, however, whether the trade-off that the Bayesian model is making is reasonable, we can consider the combination score, as per Cooke's performance weighting method. Here, the statistical accuracy and information scores are multiplied together to give a combined score. This metric for cross-validation is based on the same motivations that lie behind performance weighting. It is thus important to ensure that it is not biased towards a performance-weighted decision-maker. More research is needed to confirm this is the optimal unbiased cross-validation approach. We note this challenge and agree that more work should be done to define a set of cross-validation metrics and processes that are independently ratified, model agnostic and applied consistently to such studies. In the short term, however, this does remain the best available approach and gives us access to a body of knowledge built in the previously listed studies for comparison. Rather than considering the aggregate combined score, which may mask some of the underlying behaviour, we will consider the combined score of each forecasted subset for each study. Figure 5.12 plots the combined score of the PWDM versus the Bayesian decision-maker and an $x = y$ line to help identify relative performance.

This plot highlights a number of interesting elements about the performance of the two models across these subsets. Firstly, of keynote, is that across many of the studies outlined, a significant portion of the subset forecasts sit above the line $x = y$ (e.g. Study 23 or Study 35). This implies that for these studies the PWDM has outperformed the Bayesian model on aggregate whilst considering such a combination measure. Whilst this might appear disheartening for the Bayesian framework, it provides further evidence of the robustness of a performance-weighted approach, which should be admired for its consistent ability to stand up to scrutiny and can provide further reassurance for those who have relied on this model over the past 3 decades. On the positive side for the Bayesian framework, however, is that there are studies in which the mass of points have been more balanced (e.g. Study 1, Study 27 and Study 28) and a few studies in which the Bayesian model appears to be a better predictor across the given subsets (e.g. Study 1, Study 20, Study 8). In fact, there is only one study (Study 23) in which the Bayesian model did not outperform the PWDM on some subset of the seed variables when we consider a combination metric. In total across the 2706 subsets, the Bayesian model outperformed the performance-weighted model in approximately a third of cases (912) with the PWDM demonstrating higher combination scores in 1794 subsets. It is reassuring for the Bayesian approach that there

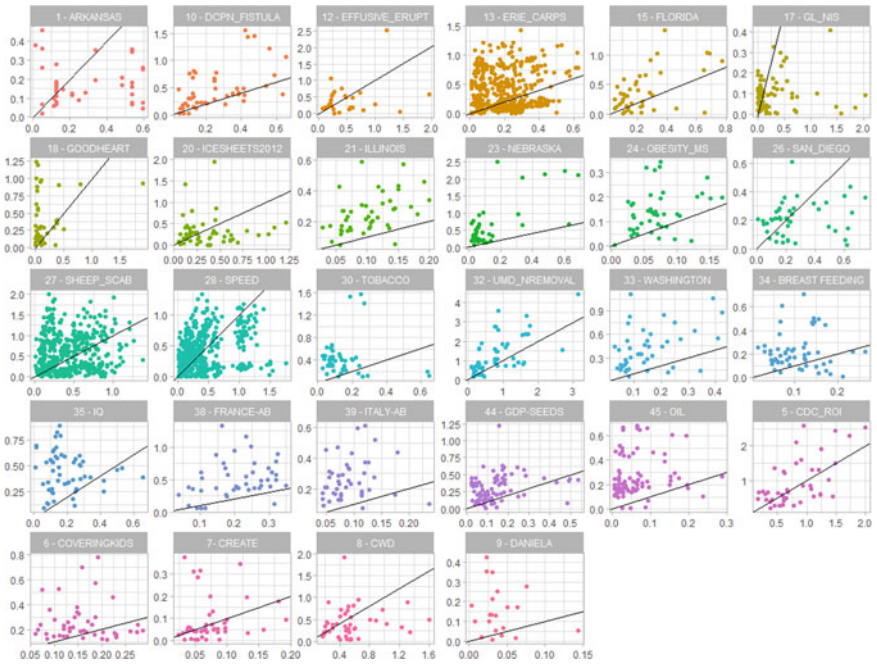

Fig. 5.12 A plot of the combination scores for each analysed study subset. The performance-weighted model (y-axis) demonstrates higher combination scores than the Bayesian model (x-axis) as a significant mass of the points are above the x = y line. There are studies, however, e.g. Study 20 (the ice sheets example), where the Bayesian model typically has higher combination scores

is a substantial number of cases where the model can meet the aims of providing a consensus distribution which is fully paramaterised, whilst performing well against the PWDM when considering a combined statistical accuracy and information score. To be a fully viable model, however, more research is required to understand the drivers of what causes certain combinations to perform better in the Bayesian context than others. One potential option, originally posited in Hartley and French (2021), is that performance here could be linked to the number of experts/seed variables present within the study. This assessment is left for future research.

5.7 Discussion

This chapter has outlined the application of the Bayesian approach to aggregating expert judgements, and its ability to supplement existing models, by:

- Assessing the extent to which experts display systematic over or under confidence.

- Minimising potential overconfidence for the decision-maker that arises from the impact of correlation between expert judgements driven by shared knowledge and common professional backgrounds
- Emphasising the underlying consensus between experts whilst reflecting the diversity of judgements.
- Providing a fully parametrised posterior distribution that is easy to integrate into further analysis.

The framework has been assessed in detail against a small number of studies and then at a macro-level across many studies within the Delft database.

This analysis has shown that such new Bayesian frameworks can be practical, unlike many preceding Bayesian approaches, and can be implemented without a significant overhead in defining complex priors. Utilising relatively diffuse priors (consistent across studies) has been shown to provide results on a similar order of magnitude to current approaches. This would also support the potential of applications of the Bayesian approach in contexts where the aggregate distribution is designed to emulate a rational scientist's perspective in addition to those where a specific decision-maker , potentially with significant a priori beliefs and consequently tighter priors, exists.

The outputs of a Bayesian model of expert judgement have been compared across studies to the performance weighting approach of Cooke's classical model. This comparison has shown that the resultant outputs of the Bayesian approach typically do not vary substantially from the performance-weighted approach when only the median point is considered, however, emphasise a different perspective of the uncertainty. Consistent with other analysis of the Bayesian approach (Hartley and French 2021), the Bayesian model displays a unimodal posterior, with narrower shoulders than an equal-weighted approach (as it emphasises underlying consensus) and has fatter tails than the performance-weighted approach (as it usually highlights systemic overconfidence of experts).

Through cross-validation, we have shown that, as we might expect a priori given its structure, the Bayesian model demonstrates higher statistical accuracy than the performance-weighted approach, but lower informativeness. This suggests that based on the decision-making context the potential sensitivity to each of these metrics may impact the choice of model considered.

Finally, by considering the single combined score metric (the product of the information and statistical accuracy), we have seen that the performance-weighted approach once again stands up to scrutiny and outperforms the Bayesian framework, when configured in this particular way, in the majority (circa 2/3) of cases. There are, however, a substantial number of cases (circa 1/3) for which the Bayesian model outperforms the performance-weighted approach lending credibility to the usage of the Bayesian model in general.

Overall, this chapter has demonstrated that the goal of a practical generic Bayesian framework for mathematical aggregation of expert judgement is feasible and can produce reasonable results when compared to current best in class approaches even

when considered broadly with a single set of parametrisations/priors. Much more work is required to assess:

- The impact of the number of seed variables/experts.
- Different parametrisations and priors within the generic framework.
- Approaches for dealing with variables on different scales.
- The drivers of out/underperformance relative to performance-weighted approaches.

However, we have now shown that there is sufficient evidence that the application of resources to assessing these areas is justified.

The performance-weighted approach outlined by Cooke clearly remains the exemplar in this space for many applications; however, we now have a Bayesian approach which can provide a different perspective, add value for decision-makers with specific needs and which we hope will continue to evolve and challenge the performance-weighted method.

5.8 Supplementary Material

Table 5.3 Target variable predicted quantiles for the effusive eruption study across models

	EWDM			PWDM			BDM		
Target variable	0.05	0.5	0.95	0.05	0.5	0.95	0.05	0.5	0.95
10x SO2 probability	0.02	2.76	32.19	0.24	5.03	19.2	−0.41	0.99	39.44
Column height	5.24	13.61	22.68	6.52	12.87	22.22	−2.96	13.47	31.67
Avg plume height	0.6	3.79	11.92	0.58	4.05	11.26	−2.95	4.22	13.06
Max plume height	0.53	3.93	13.2	0.85	3.87	15.78	−3.63	3.98	16.74
Max% SO2 emissions	53.04	86.18	99.74	50	75.63	97.45	32.63	86.15	116.77
Min% SO2 emissions	25.09	70.89	89.96	40.81	68.2	84.61	10.27	71.21	118.77
Max no. fissures	3.88	26.15	472.4	6.08	18.45	98.43	−25.05	28.77	174.32
Min no. fissures	0.12	2.95	13.6	1.19	6.6	16.83	−4.31	3.55	20.27
Duration explosive phase	0.14	2.88	15.84	0.34	5.64	23.79	−4.3	2.86	21.44
Gap between outbursts	0.12	7.17	183.4	0.51	5.71	29.87	−2.17	4.37	163.99

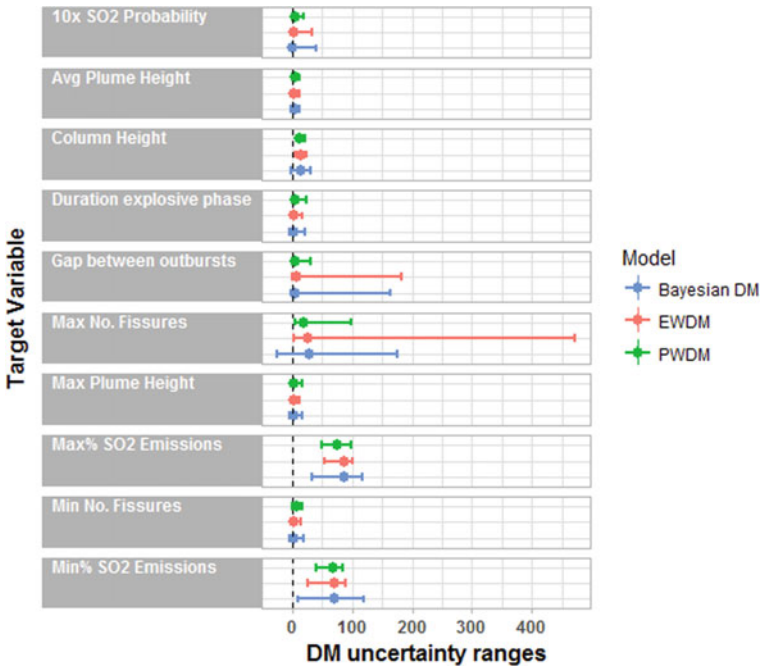

Fig. 5.13 Replication of Fig. 5.4 including the EWDM. The Bayesian model displays posterior uncertainty ranges consistently broader than the PWDM, however, displays uncertainty bounds both broader and narrower than the EWDM

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K., & Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Analysis* (pp. 503–532).

Bamber, J. L., & Aspinall, W. P. (2013). An expert judgement assessment of future sea level rise from the ice sheets. *Nature Climate Change, 3*(4), 424.

Bamber, J. L., Aspinall, W. P., & Cooke, R. M. (2016). A commentary on how to interpret expert judgment assessments of twenty-first century sea-level rise by Hylke de Vries and Roderik SW van de Wal" *Climatic Change, 137*(3–4), 321–328.

Billari, F. C., Graziani, R., Melilli, E. (2014). Stochastic population forecasting based on combinations of expert evaluations within the Bayesian paradigm. *Demography, 51*(5), 1933–1954.

Burgman, M. A. (2015). *Trusting judgements: How to get the best out of experts*. Cambridge University Press.

Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A. & Charrad, M.M. (2014). Package nbclust. *Journal of Statistical Software, 61*, 1–36

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering & System Safety, 93*(5), 760–765.

Clemen, R. T., & Lichtendahl, K. C. (2002). *Debiasing expert overconfidence: A Bayesian calibration model*. PSAM6: San Juan, Puerto Rico.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety, 163*, 109–120.

Cooke, R. M. (1991). *Experts in uncertainty*. Oxford: Oxford University Press.

Cooke, R. M. (Ed.). (2007). Expert judgement studies. *Reliability Engineering and System Safety*.
Cooke, R. M. (2008). Response to discussants. *Reliability Engineering & System Safety, 93*(5), 775–777.
Cooke, R. M. (2016). Supplementary Online Material for Cross Validation of Classical Model for Structured Expert Judgment.
Cooke, R. M., & Goossens, L. H. J. (2000). Procedures guide for structured expert judgement in accident consequence modelling. *Radiation Protection Dosimetry, 90*(3), 303–309.
Cooke, R. M., & Goossens, L. H. (2008). TU Delft expert judgment data base. *Reliability Engineering & System Safety, 93*(5), 657–674.
Cooke, R. M., & Solomatine, D. (1992). *EXCALIBUR integrated system for processing expert judgements version 3.0*. Delft: Delft University of Technology and SoLogic Delft.
Cooke, R. M., Wittmann, M. E., Lodge, D. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., & Mason, D. M. (2014). Out of sample validation for structured expert judgment of Asian carp establishment in Lake Erie. *Integrated Environmental Assessment and Management, 10*(4), 522–528.
De Finetti, B. (1974). *Theory of Probability*. Chichester: Wiley.
De Finetti, B. (1975). *Theory of Probability*. Chichester: Wiley.
de Vries, H., & van de Wal, R. S. W. (2015). How to interpret expert judgment assessments of 21st century sea-level rise. *Climatic Change, 130*(2), 87–100.
de Vries, H., & van de Wal, R. S. W. (2016). Response to commentary by JL Bamber, WP Aspinall and RM Cooke. *Climatic Change, 137*(3–4), 329–332.
Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cookes classical model. *Reliability Engineering & System Safety, 121*, 72–82.
EFSA. (2010). Statement of EFSA on the possible risks for public and animal health from the contamination of the feed and food chain due to possible ash fall following the eruption of the Eyjafjallaj kull volcano in Iceland. *EFSA Journal, 8*, 1593.
EFSA. (2014). Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*.
Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering & System Safety, 96*(10), 1292–1310.
French, S. (1980). Updating of belief in the light of someone else's opinion. *Journal of the Royal Statistical Society, A143*, 43–48.
French, S. (1985). Group consensus probability distributions: a critical survey (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley & A. F. M. Smith *Bayesian Statistics 2, North-Holland* (pp. 183–201).
French, S. (2011). Aggregating expert judgement. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales, 105*(1), 181–206.
French, S., Maule, A. J., & Papamichail, K. N. (2009). *Decision behaviour, analysis and support*. Cambridge: Cambridge University Press.
Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association, 100*(470), 680–701.
Hartley, D., & French, S. (2021). A Bayesian method for calibration and aggregation of expert judgement. *Journal of Approximate Reasoning, 130*, 192–225.
Hartley, D., French, S. (2018). Elicitation and calibration: A Bayesian perspective. *Elicitation: The science and art of structuring judgement* (pp. 119–140).
Hockey, G. R. J., Maule, A. J., Clough, P. J., & Bdzola, L. (2000). Effects of negative mood on risk in everyday decision making. *Cognition and Emotion, 14*, 823–856.
Hora, S. (2007). Eliciting probabilities from experts. In W. Edwards, R. F. Miles, & D. Von Winterfeldt, *Advances in decision analysis: From foundations to applications* (pp. 129–153). Cambridge: Cambridge University Press.

Kadane, J. B., & Fischhoff, B. (2013). A cautionary note on global recalibration. *Judgment and Decision Making, 8*(1), 25–27.

Kahneman, D. (2011). *Thinking, fast and slow*. London: Penguin, Allen Lane.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgement under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Lichtendahl, K. C. (2005). Bayesian models of expert forecasts. Ph.D. thesis.

Lichtendahl, K. C., & Winkler, R. L. (2007). Probability elicitation, scoring rules, and competition among forecasters. *Management Science, 53*(11), 1745–1755.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky *Judgement under uncertainty* (pp. 306–334). Cambridge: Cambridge University Press.

Lin, S.-W., & Bier, V. M. (2008). A study of expert overconfidence. *Reliability Engineering and System Safety, 93*, 711–721.

Lin, S.-W., & Cheng, C.-H. (2009). The reliability of aggregated probability judgments obtained through Cooke's classical model. *Journal of Modelling in Management, 4*(2), 149–161.

Loughlin, S. C., Aspinall, W. P., Vye-Brown, C., Baxter, P. J., Braban, C., Hort, M., et al. (2012). Large-magnitude fissure eruptions in Iceland: Source characterisation. *BGS Open File Report, OR/12/098*, 231pp. Retrieved from http://www.bgs.ac.uk/research/volcanoes/LakiEruptionScenarioPlanning.html.

Mumpower, J. L., & Stewart, T. R. (1996). Expert judgement and expert disagreement. *Thinking and Reasoning, 2*(2–3), 191–211.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, R., Garthwaite, P. H., Jenkinson, D., et al. (2006). *Uncertain judgements: Eliciting experts' probabilities*. Chichester: Wiley.

Shanteau, J. (1995). Expert judgment and financial decision making. *Risky business: Risk behavior and risk management*. B. Green: Stockholm, Stockholm University.

Skjong, R., & Wentworth, B. H. (2001). Expert judgement and risk perception. In *Proceedings of the eleventh (2001) international offshore and polar engineering conference*. Stavanger, Norway: The International Society of Offshore and Polar Engineers.

Wilson, K. J. (2016). An investigation of dependence in expert judgement studies with multiple experts. *International Journal of Forecasting*.

Wiper, M. W., & French, S. (1995). Combining experts' opinions using a normal-Wishart model. *Journal of Forecasting, 14*, 25–34.

Wittmann, M. E., Cooke, R. M., Rothlisberger, J. D., & Lodge, D. M. (2014). Using structured expert judgment to assess invasive species prevention: Asian Carp and the Mississippi-Great Lakes Hydrologic Connection. *Environmental Science & Technology, 48*(4), 2150–2156.

Wittmann, M. E., Cooke, R. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., Mason, D. M., et al. (2015). Use of structured expert judgment to forecast invasions by bighead and silver carp in Lake Erie. *Conservation Biology, 29*(1), 187–197.

Zhang, H., Rutherford, E. S., Mason, D. M., Breck, J. T., Wittmann, M. E., Cooke, R. M., et al. (2016). Forecasting the impacts of silver and bighead carp on the Lake Erie food web. *Transactions of the American Fisheries Society, 145*(1), 136–162.