Bayesian Analysis (2020) **15**, Number 4, pp. 1251–1280

Calibrating Expert Assessments Using Hierarchical Gaussian Process Models

Tommi Perälä[*], Jarno Vanhatalo[†], and Anna Chrysafi[‡]

Abstract. Expert assessments are routinely used to inform management and other decision making. However, often these assessments contain considerable biases and uncertainties for which reason they should be calibrated if possible. Moreover, coherently combining multiple expert assessments into one estimate poses a long-standing problem in statistics since modeling expert knowledge is often difficult. Here, we present a hierarchical Bayesian model for expert calibration in a task of estimating a continuous univariate parameter. The model allows experts' biases to vary as a function of the true value of the parameter and according to the expert's background. We follow the fully Bayesian approach (the so-called supra-Bayesian approach) and model experts' bias functions explicitly using hierarchical Gaussian processes. We show how to use calibration data to infer the experts' observation models with the use of bias functions and to calculate the bias corrected posterior distributions for an unknown system parameter of interest. We demonstrate and test our model and methods with simulated data and a real case study on data-limited fisheries stock assessment. The case study results show that experts' biases vary with respect to the true system parameter value and that the calibration of the expert assessments improves the inference compared to using uncalibrated expert assessments or a vague uniform guess. Moreover, the bias functions in the real case study show important differences between the reliability of alternative experts. The model and methods presented here can be also straightforwardly applied to other applications than our case study.

Keywords: expert elicitation, bias correction, Gaussian process, Supra Bayes, fisheries science, environmental management.

MSC2020 subject classifications: Primary 62F15, 62P12; secondary 60G15.

1 Introduction

Expert elicitation is an important part of statistical analyses in various fields of research and decision making (O'Hagan et al., 2006; Dias et al., 2018; Albert et al., 2012). In a typical situation where expert knowledge is used, data are lacking or the time and resources to collect the data are limited (Burgman et al., 2011; Morgan, 2014) and thus, the available information is insufficient to make meaningful inference about the phenomenon of interest (Burgman, 2005; Roman et al., 2008; Zickfeld et al., 2010; Wilson et al., 2018). Expert opinions can be valuable, for example, within the context

[*] Department of Biological and Environmental Science, University of Jyväskylä, Finland, tommi.a.perala@jyu.fi

[†] Department of Mathematics and Statistics and Organismal and Evolutionary Biology Research Program, University of Helsinki, Finland, jarno.vanhatalo@helsinki.fi

[‡] Water & Development Research Group, Aalto University, Finland, anna.chrysafi@aalto.fi

https://doi.org/10.1214/19-BA1180

of Bayesian inference (Garthwaite et al., 2005; Uusitalo et al., 2005; Low-Choy et al., 2009) or in decision and risk analysis (Usher and Strachan, 2013; Landquiste et al., 2017; Nevalainen et al., 2018) as a means for specifying informative prior distributions for unknown parameters. Although expert knowledge is often a valuable, and sometimes even the only available source of information, its successful utilization in decision making immediately raises at least two practical questions. The first question is related to the optimal design of the expert elicitation process itself. What kind of a procedure would best utilize the expertise and capture the possibly informal knowledge of the expert? The second question concerns the proper usage of the assessments in statistical decision making. How the assessments of multiple experts should be utilized in statistical inference and decision making so that the uncertainties and possible systematic errors or biases in the assessments are properly accounted for (Tversky and Kahneman, 1974; Lindley et al., 1979; O'Hagan et al., 2006; Burgman et al., 2011; Dias et al., 2018)? Even though these issues are intertwined, here we focus on the latter paying special attention to considering the biases, or in other words, to the calibration of experts' assessments.

More specifically, we want to infer an unknown system parameter using experts' assessments of it in a setting where the experts' assessments may be contaminated by systematic errors or biases. We call this assessed parameter a system parameter to help distinguish it from the other parameters in our model. We assume that instead of being constant, the biases in the expert assessments may vary depending on the true value of the system parameter. We also assume that we have access to calibration data containing the experts' previous assessments of system parameters whose true value is known. We build a formal statistical model which utilizes this information about the experts' past performance to learn about the experts' biases, and, more importantly, to correct for them in their future assessments in similar situations. In order to establish the utility of our work, to elucidate the applicability of the proposed method, and to further clarify our approach, we next discuss some potential example applications for it.

In ecology, plant coverage data is commonly used, for example, in species distribution models. The plant coverage data is often expressed as a percentage describing the fraction of a survey plot the plant inhabits. Typically, these data are based on expert assessments where an expert (researcher working on the field) visually estimates the plant coverage of the survey plot. In practice, the accuracy of the expert's assessment varies depending on the true coverage since assessing very small or large percentage coverages is (relatively) harder than assessing intermediate values. In environmental management applications, expert assessment is used to estimate, for example, the vulnerability of a species to contaminants such as oil (Nevalainen et al., 2018). The assessment, and thus its accuracy and value in decision making, may depend on the true value of the vulnerability parameter because of the experts' (unconsciously) precautionary attitude, which encourages the experts to overestimate the vulnerability. Moreover, in general, experts tend to underestimate probabilities (Lindley and Singpurwalla, 1986). Often the word expert is also used to refer to a (deterministic) computer model. In their seminal work (Kennedy and O'Hagan, 2001) modelled the bias of a computer model as a function of covariates (computer model inputs). However, these covariates may not be available, or the computer model can be biased within certain output range regardless of the covariates. For example, due to numerical approximations, environmental simulators may overestimate concentrations near zero. The case study for this work is motivated by the dire need for improved methods for data-limited fisheries stock assessment. Here, the unknown system parameter is *stock status*, which is defined as the ratio of the current and virgin biomasses of a fish stock (Section 2 and Chrysafi et al., 2019). We use calibration data and fisheries expert assessments to infer the stock status for a set of fish stocks. In all these examples we could, at least in principle, collect calibration data to enable learning of the experts' biases. For example, in the plant coverage estimation, the calibration data would be easy to collect as a small number of plots could be accurately measured and the measurements compared with the experts' assessments. The potential biases could then be accounted and corrected for when using the experts' assessments in plant coverage estimation in the future.

Tversky and Kahneman (1974) demonstrated that experts (or humans in general) are sensitive to a host of psychological idiosyncrasies and subjective biases. There are four main heuristics that experts unconsciously use when making their assessments. The first one is called *representativeness*, and it often occurs when assessing the probability of events such as "A belongs to B" or "C originates from D". The second one is called *availability*, and it occurs when the expert ("she" hereafter) assesses the probability of an event by trying to recall the number of past occurrences of that event. The third one is called *adjustment and anchoring*, and it affects the expert assessments when the expert starts from an initial value (the anchor) and then makes incremental adjustments to it in order to arrive at her final assessment. Incremental adjustments around the anchor are typically inefficient (Tversky and Kahneman, 1974). The fourth heuristic is called *overconfidence* (Kynn, 2008; Speris-Bridge et al., 2010), which means the expert systematically overestimates the accuracy of or underestimates the uncertainty in her assessment. These heuristics can lead to very narrow and biased probability distributions (Griffiths et al., 2007; Kuhnert et al., 2009, 2010). If the biases are systematic, they can be inferred from calibration data, and corrected for; that is, the experts can be calibrated (Lindley, 1982, 1983; Burgman et al., 2011; Morgan, 2014; Hartley and French, 2018).

Methods for expert elicitation have been extensively studied, and comprehensive reviews and detailed treatments of the subject are provided by O'Hagan et al. (2006) and Dias et al. (2018). Here, we assume that during the elicitation process, each expert's knowledge has been formulated as a probability distribution. There are two main approaches for conducting statistical inference using expert assessments: a) the fully Bayesian approach (also called "supra-Bayesian") (French, 1980) and b) opinion pooling. In the former approach, expert assessments are used as "observations" to update the analyst's beliefs about the phenomena under study using the Bayes' theorem (Morris, 1974; Lindley and Singpurwalla, 1986; Gelfand et al., 1995; French, 2011; Albert et al., 2012; Hartley and French, 2018). The information provided by the experts is linked to the unknown system parameters through a conditional probability distribution (likelihood function). However, it can be challenging to formulate the likelihood function (Genest and Schervish, 1985; O'Hagan et al., 2006), which is probably why the latter approach, the opinion pooling, has gained more popularity (Dias et al., 2018). Opinion pooling, which also forms the basis of the classical models (Cooke and Goossens, 2008), is based on combining the probability distributions provided by the experts either by weighted arithmetic or logarithmic averaging (McConway, 1981; O'Hagan et al., 2006;

Dietrich and List, 2014; Farr et al., 2018). The weights can be used to give more influence on those experts that have performed better in a validation test (Cooke and Goossens, 2008; de Little et al., 2018).

In this work, we are interested in three questions: 1) how to combine several experts' assessments in a theoretically coherent manner, 2) how to conduct statistical inference on the biases in these assessments, and 3) how to account and correct for the biases when using the same experts as a source of information later in similar situations to estimate an unknown system parameter. All these questions can be answered within the framework of hierarchical Bayesian statistics (Hartley and French, 2018). Hence, we follow the fully Bayesian approach where we act as the analyst who builds an explicit statistical model for the experts' biases and their relationships to the unknown system parameters and uses the model to update his beliefs about the system under study using the Bayes' rule. Lindley (1982) presents an early approach for calibrating experts by comparing the experts' assessments of certain events with the realizations of those events (e.g. weather forecasting). In his application, the experts' assessments consisted of estimates of probabilities for binary events. Lindley (1983) and Lindley and Singpurwalla (1986) apply and extend the approach to continuous variables. Later applications of Bayesian calibration and updating are provided, for example, by Clemen and Lichtendahl (2002) and Albert et al. (2012). Here, we further extend these approaches making them suitable for problems where the experts' biases are not necessarily constant, but instead may vary depending on the true value of the unknown system parameter being assessed. We use hierarchical Gaussian processes to model the experts' biases as continuous functions of the unknown system parameter value. We infer the bias functions from the calibration data and use them to correct for the bias in experts' assessments in new situations. We evaluate and demonstrate our model performance with simulations and apply it to a real case study.

The rest of the article is organized as follows. In Section 2, we describe the motivating case study that will be analyzed in the experiments. In Section 3, we present our statistical model and inference methods, and in Section 4, we present the simulation and case study results. We end with discussion and conclusions in Section 5.

2 Expert assessments in data-limited fisheries management

Currently, approximately 80% of the world's exploited fish stocks are unassessed (Costello et al., 2012), which is a major concern both to ocean sustainability and food security since appropriate management decisions should be based on the status of the fish stock (Food and Agriculture Organization of the United Nations, 1995). However, assessing data-limited stocks can be challenging as the traditional stock assessment methods require large amounts of fishery dependent and independent data, (Magnusson and Hilborn, 2007; Methot and Wetzel, 2013) which are typically lacking for small-scale fisheries and in developing countries (Salas et al., 2007; Meissa et al., 2013). The list of data-limited stock assessment methods is, thus, extensive (Geromont and Butterworth, 2015; Chrysafi and Kuparinen, 2015) and the use of expert knowledge is often recommended in the data-limited fisheries literature (Berkson and Thorson, 2014; Newman et al., 2015). Expert assessment is motivated by the fact that fisheries data are not straightforward to link to the underlying stock (Daan et al., 2011). Experts possess experiential knowledge on fishing methods, data collection, reporting behavior of fishermen as well as on biology and behavior of fish stocks that help them interpret data. Such tacit information is especially important in data-limited stock assessment.

Our case study, is motivated by data-limited stock assessment methods that require information of the so-called stock status system parameter, expressed as $x_t = B_t/B_0 \in [0, 1]$, where B_t is the fish stock's biomass at time t and B_0 is the virgin biomass (Dick and MacCall, 2011; Cope, 2013; Froese et al., 2017). Recently, Chrysafi et al. (2019) conducted an expert elicitation experiment where simulated and data-rich stocks with known stock statuses were used to construct a data-limited stock assessment test. For each stock, six fisheries experts with varying levels of experience in stock assessment were provided with data that imitated the typical data available in data-limited stock assessment. The experts were then asked to provide their estimates for the stock statuses together with estimates of their own uncertainty in their ability to estimate the system parameters as Beta distributions. The main objectives of Chrysafi et al. (2019) were to quantify the degree of the bias in the experts' assessments for the stock status x_t and to explore the ways in which the fisheries expert knowledge may help to inform management decisions in a data-limited case. Their main findings, which also inspired the methods developed in this work, can be summarized as follows. Firstly, the experts' biases varied as a function of the true stock status. Experts tend to overestimate low stock statuses and to underestimate high stock statuses. Secondly, the experts' amount of experience in stock assessment affected both their degree of bias as well as their confidence in their estimates. Experienced experts' opinions were better calibrated, and they understood the concept of uncertainty better than the inexperienced ones. However, Chrysafi et al. (2019) did not consider pooling the expert assessments nor correcting for their bias, developing methodology that would be extensible to other expert elicitation applications. These issues are treated in this work.

3 Materials and methods

3.1 Expert assessment model

We consider ourselves as an analyst poised with the task of estimating an unknown system parameter $\tilde{x} \in [0, 1]$[1] using expert assessments of the system parameter. We assume that each expert $j \in \{1, \ldots, J\}$ has previously assessed a similar system parameter $x_i, i \in \{1, \ldots, n\}$ for n different systems. The vector containing the previously assessed system parameters is denoted by $\mathbf{x} = [x_1, \ldots, x_n]^T$. We assume that the system parameter has the same interpretation in each system, and that the systems and the experts' knowledge about them are similar enough so that we can anticipate consistency in the experts' assessments of x_i for each $i \in \{1, \ldots, n\}$. In our case study, the system parameter represents the ratio of the current biomass and the virgin biomass of a fish stock

[1] Since any interval can be mapped to the unit interval, we focus here without loss of generality on the unit interval pointing out that the approach can be generalized to any interval.

and the different systems represent different fish stocks. Earlier similar expert elicitation examples are provided by, e.g., Lindley and Singpurwalla (1986) who considered the failure rates of machine components, Gelfand et al. (1995) who considered the number of points NBA teams would score in a series game and Albert et al. (2012) who consider dose-response for Listeria contamination and assessment of length of PhD studies.

In general, it is advisable to ask the experts, in addition to their best guess, to also report their own uncertainty about their assessment in the form of a probability density function (O'Hagan et al., 2006). Here, we restrict our treatment to parametric distributions and thus our first two assumptions are:

(A1) Each expert $j \in \{1, \ldots, J\}$ has expressed her beliefs about the system parameters $x_i, i \in \{1, \ldots, n\}$ and the uncertainty about her beliefs as parametric probability distributions and provided us with the summary statistics $\mathcal{D}_{ji}$ that fully describe the distributions.

(A1)* These distributions are Beta distributions,[2] $\text{Beta}(m_{ji}s_{ji}, (1-m_{ji})s_{ji})$, and the summary statistics, $\mathcal{D}_{ji} = (m_{ji}, s_{ji})$, consist of the mean (m_{ji}) and dispersion (s_{ji}) parameters of the Beta distribution. The mean parameter is interpreted as the expert's point estimate (or measurement) of x_i, whereas the dispersion parameter (more commonly known as *sample size*) quantifies the expert's own perception of the uncertainty in her point estimate (accuracy of the measurement).

For notational simplicity, we assume without loss of generality that each expert assesses the same set of systems[3] and denote by $\mathcal{D}_j = \{\mathcal{D}_{ji}, i \in \{1, \ldots, n\}\}$ the set of assessments for all systems by expert $j \in \{1, \ldots, J\}$.

We formulate our prior beliefs about the system parameters $\mathbf{x} = [x_1, \ldots, x_n]^T$ in the study systems as a probability distribution $p(\mathbf{x})$. In order to update our prior beliefs with the expert assessments, we need a model describing what kind of assessments the experts produce for given system parameter values. More formally, we have to specify the observation model as a conditional probability distribution for all expert assessments conditioned on the unknown system parameters $p(\mathcal{D}|\,\mathbf{x})$, where $\mathcal{D} = \{\mathcal{D}_j, j = 1, \ldots, J\}$. Our updated beliefs are then represented by the posterior distribution obtained using the Bayes' theorem

$$p(\mathbf{x}|\mathcal{D}) \propto p(\mathcal{D}|\,\mathbf{x})p(\mathbf{x}). \tag{3.1}$$

This update rule provides a theoretically coherent method for updating analyst's beliefs about the systems under study and it is sometimes called the supra-Bayesian approach to distinguish it from various (non-coherent) pooling and averaging methods (Hartley and French, 2018). However, the challenge with the fully Bayesian approach is the formulation of the conditional model for the expert assessments (Lindley and Singpurwalla, 1986; Gelfand et al., 1995; Hartley and French, 2018).

[2] Naturally, other distributions could be used as well (see e.g., Lindley and Singpurwalla, 1986). However, the Beta distribution is a reasonable choice here since the system parameter is defined in the unit interval.

[3] Our model can be generalized to situations where each expert assesses a different set of systems.

Because of the possibility of bias and overconfidence in the experts' assessment due to the psychological reasons discussed earlier and further supported by the findings of the expert elicitation experiment of Chrysafi et al. (2019), we do not fully trust the experts' ability to estimate the unknown system parameters. However, we do believe that the experts' assessments contain valuable information that we want to utilize to estimate the unknown system parameter. We treat the experts as faulty or uncalibrated measurement devices that produce measurements that can be biased, and the reported measurement accuracy may be wrong. We model this using an observation model, where the possible bias is explicitly accounted for. We do not have, however, a priori knowledge of the bias, and thus we use calibration data to learn about the biases.

We build the conditional model for the expert assessments hierarchically and first define a parametric joint observation model for the expert assessments, $p(\mathcal{D}|\,\mathbf{x},\theta)$. Here θ denotes the parameters of the observation model. The conditional model for the expert assessments can be now written as $p(\mathcal{D}|\,\mathbf{x}) = \int p(\mathcal{D}|\,\mathbf{x},\theta)p(\theta|\,\mathbf{x})\mathrm{d}\theta$ where $p(\theta|\,\mathbf{x})$ is the conditional probability density function of the observation model parameters given the system parameters. The joint observation model can be further expanded as

$$\begin{aligned} p(\mathcal{D}|\,\mathbf{x},\theta) &= \prod_{j=1}^{J} p(\mathcal{D}_j|\,\mathbf{x},\theta_j) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{J} p(m_{ji},s_{ji}|x_i,\theta_{ji}), \end{aligned} \tag{3.2}$$

where $\theta_j = \{\theta_{ji}\}_{i=1}^{n}$ denotes the set of the jth expert's observation model parameters for each system and $\theta = \{\theta_j\}_{j=1}^{J}$ denotes the set of all experts' observation model parameter sets. Next, we write the jth expert's observation model for the system i as $p(m_{ji},s_{ji}|x_i,\theta_{ji}) = p(m_{ji}|s_{ji},x_i,\theta_{ji})p(s_{ji}|x_i,\theta_{ji})$, and make the following assumptions.

(A2) The expert's assessment of her own uncertainty, s_{ji}, does not contain information about x_i in itself, implying that $p(s_{ji}|x_i,\theta_{ji}) = p(s_{ji}|\theta_{ji})$.

(A3) The observation model for the jth expert's point estimate m_{ji} for the ith system parameter x_i is a Beta distribution $p(m_{ji}|s_{ji},x_i,\theta_{ji}) = \mathrm{Beta}(m_{ji}|\mu_{ji}\eta_{ji}, (1-\mu_{ji})\eta_{ji})$, parameterized using the mean $\mu_{ji} = \mu_j(x_i) \in (0,1)$ and the dispersion parameter $\eta_{ji} = \eta_j(s_{ji}) \in \Re_+$, where and $\mu_j(\cdot)$ and $\eta_j(\cdot)$ are the jth expert's mean and dispersion functions. Hence, the parameters in (3.2) are $\theta_{ji} = \{\mu_{ji},\eta_{ji}\}$. These parameters are related to the natural parameters of the beta-distribution as $\alpha_{ji} = \mu_{ji}\eta_{ji}$ and $\beta_{ji} = (1-\mu_{ji})\eta_{ji}$, and the variance of the beta-distribution is $\mu_{ji}(1-\mu_{ji})/(\eta_{ji}+1)$ (Gelman et al., 2013).

Assumption *(A2)* is the same that was used already by Lindley (1983) and Lindley and Singpurwalla (1986) with log-Gaussian observation model. Similarly, assuming independence between μ_{ji} and s_{ji} means that the expert's uncertainty estimate does not contain information about her point estimate or the “best guess”, which is encoded by the mean parameter. These assumptions are convenient at minimum but may not be realistic in all applications. We make them here, since in our case study, there is no reason to assume that the experts' estimates of their uncertainty about their assessment would depend on the true value of the system parameter x_i being assessed. The Beta distribution in Assumption *(A3)* is a natural choice here since the support of the Beta distribution is the unit interval which also happens to be the interval containing all the possible values of the expert's point estimate m_{ji}.

By letting the dispersion parameter η_{ji} depend on s_{ji}, we assume that the experts can provide us with useful information on how much we should trust their assessment. If we fixed $\eta_i(s_{ji}) = s_{ji}$ we would fully trust the expert's estimate of the uncertainty about her assessment. In this case, at the limit $s_{ji} = 0$ expert's assessment would not be used to update our prior belief $p(x_i)$. Similarly, at the limit $s_{ji} = \infty$ our prior would be replaced by point mass one at $\mu_j(x_i)$. In order to learn about the goodness of expert's uncertainty estimate, we let $\eta_j(s_{ji}) = s_{ji}\rho_j$, where ρ_j is a reliability parameter. It corrects for possible errors in the expert's estimate of the accuracy of her assessment. For example, expert's overconfidence would be described by a reliability parameter value less than one. The reliability parameter is an unknown parameter of the observation model and must be inferred from the calibration data (see Section 3.2). We formulate our prior beliefs about the reliability parameter as $\rho_j \sim \text{Gamma}(2, 0.5)$. This prior distribution is centered around one and assigns 60% probability for reliability parameter values that are between 0.5 and 2, thus encoding a weak preference for moderate deviations from the expert's own uncertainty assessment. We also consider an alternative more restrictive assumption

(A4)* The expert's estimate of the accuracy in her assessment, s_{ji}, does not contain information about the assessed system parameter x_i implying that $p(m_{ji}|s_{ji}, x_i, \theta_j) = p(m_{ji}|x_i, \theta_j)$, and the dispersion parameter in the observation model is $\eta_{ji} = \eta_j(s_{ji}) = \eta_j$.

Under Assumption (A4*), the Beta distribution's dispersion parameter, η_{ji}, is independent from the experts' own uncertainty assessment, which means that the analyst does not trust the expert's ability to estimate her own accuracy, and rather infers the accuracy from the calibration data. In this case, we give the dispersion parameter a wide Gamma prior distribution, $\eta_j \sim \text{Gamma}(1, 10)$.

We expect that the experts can give systematically biased assessments, that is, their point estimates m_{ji} deviate from the true system parameter values x_i in a consistent and predictable manner. Hence, we explicitly model this systemic deviation or bias for each expert defining a bias function $\beta_j(x) = \mu_j(x) - x$ as the difference between the expert's point estimate and the true system parameter value. We do not want to impose strong prior assumptions about the functional form of the bias and give weakly informative priors for $\mu_j(x)$ defined directly in the function space. This is achieved using Gaussian processes (GP, Williams and Rasmussen, 2006). We define the prior distributions for the mean functions in the expert's observation models by first introducing a latent GP $b_j(x)$ with a mean function $h_j(x) = \text{E}[b_j(x)]$ and a covariance function $k_j(x, x') = \text{Cov}\left(b_j(x), b_j(x'); l_j, \sigma_j^2\right)$ where l_j is the correlation length-scale and σ_j^2 is the variance parameter of the covariance function of the GP. In this work, we use the exponentiated square (i.e., Gaussian) covariance function $k_j(x, x') = \sigma_j^2 \exp\left(-(x-x')^2/l_j^2\right)$ and a zero mean function $h_j(x) = 0$. We will consider three different models for the mean of the observation model, $\mu_j(x)$ in $x \in [0, 1]$:

Model 1 (additive): $$\mu_j = [x + b_j(x)]_{(0,1)} \tag{3.3}$$

Model 2 (logit additive): $$\text{logit}(\mu_j) = \text{logit}(x) + b_j(x) \tag{3.4}$$

Model 3 (marginally uniform predictive prior): $$\mu_j = \Phi\left(b_j(x)/\sqrt{\sigma_j^2}\right) \tag{3.5}$$

The shrinkage operator $[\cdot]_{(0,1)}$ in (3.3) is required to assure that the mean is in the unit interval. In (3.5), $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Gaussian random variable. The prior distributions induced for the mean function and the rationale behind these models are discussed next.

Model 1, the additive bias model, assumes a priori that the experts are biased towards the center of the interval (Figure 1). In other words, the experts are reluctant to believe that the system parameter value is close to either end of the interval but instead tend to favor intermediate values. Even though $b_j(x)$ is assigned a zero-mean GP prior, the shrinkage operator in (3.3) causes the prior for the bias function, $\beta_j(x)$, to be a "truncated" GP resulting in an asymmetrical distribution of the probability mass around zero everywhere except in the middle of the interval. The asymmetry of the prior is most apparent near the endpoints of the interval (Figure 1).

Model 2, the logit additive bias model, aims to encode a prior assumption that the bias is close to zero (Figure 1). This could be justified if we assume that the experts can provide unbiased estimates and have no prior information to suggest otherwise. The logit transformation in (3.4) forces the prior expectation of the bias function to be close to zero even at the endpoints of the interval. Moreover, even though the prior expectation of the bias is zero only at the center of the interval, its prior median is zero everywhere.

Model 3, the marginally uniform prior predictive model, encodes analyst's total prior ignorance by making no assumptions about the relationship between the true system parameter value and the expert's point estimate, stating that the expert's assessment can be anything ranging from a random guess to a very informed and accurate assessment of any value of the system parameter x (Figure 1). This assumption could be justified in a situation where expert assessment has never been used before. A zero-mean GP prior for $b_j(x)$ implies a Gaussian marginal distribution for each $b_j(x) \sim N(0, \sigma_j^2)$. Hence, the transformation of b_j/σ_j^2 through the cumulative distribution function of the standard normal random variable in (3.5) induces a uniform prior predictive distribution for the expert's point estimate m_j for all system parameter values.

We end the construction of the models for the expert means and biases with two alternative assumptions for the joint distribution of the experts' biases $b_j(x), j = 1, \ldots, J$

(A5) The biases of the individual experts are mutually independent. Hence, the processes $b_j(\cdot)$ are mutually independent zero-mean GP.

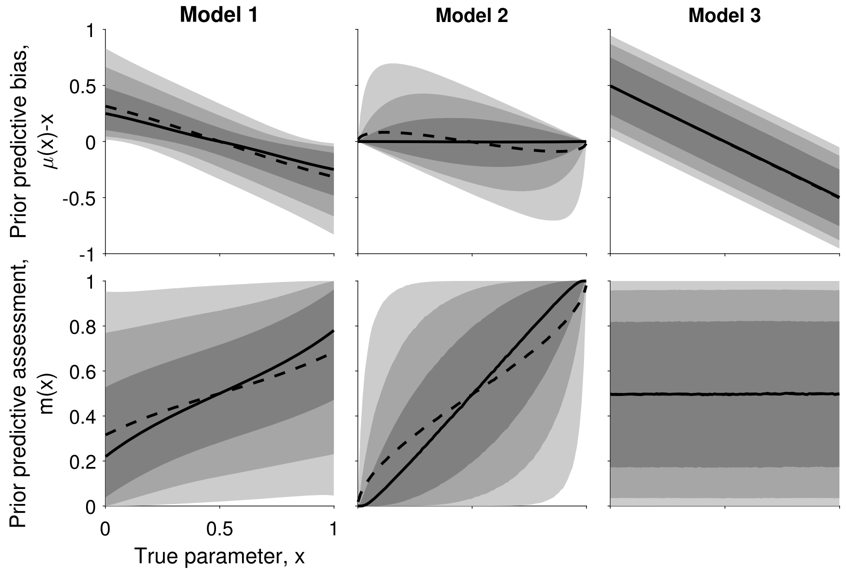


Figure 1: The prior predictive density of the bias $\beta(x)$ and the prior predictive density of the expert's mean estimate as a function of the true parameter. The solid black line is the median, the dashed black line is the mean, and the shaded areas represent the 90%, 75% and 50% central probability intervals.

(A5)* The experts are assigned to K groups based on their similarity determined for example by their educational backgrounds. The expert biases are assumed mutually dependent of the biases of experts in the same group and independent of the biases of experts in a different group. This hierarchical dependence structure is encoded by setting $h_j(x) = \bar{b}_k(x)$ if expert j belongs to group k, where $\bar{b}_k(\cdot)$ is the groupwise mean function for all $b_j(x)$ in group k. The groupwise mean functions are modeled by mutually independent zero-mean GPs with covariance functions $k_k(x, x'; l_k, \sigma_k^2), k = 1, \ldots, K$.

Assumption *(A5*)* is an extension of the hierarchical model of Albert et al. (2012) where experts were assigned to homogeneity groups that shared common hyperparameters. Here, instead of the groupwise hyperparameters, we use GPs that are shared by the experts in the same homogeneity group. These homogeneity groups may result, for example, from similar education and historic frames of reference among the experts (Albert et al., 2012; Hartley and French, 2018). For example, in our case study, the experts form three groups according to their education and experience and we can assume the experts to be exchangeable within each group. For the covariance function parameters, we used prior distributions that give more weight for slowly varying bias functions with small magnitude: $p(\sigma_j^2) \propto \text{Cauchy}(\sigma_j^2; 0, 5)\text{I}_{[0,50]}(\sigma_j^2)$ and $p(l_j^{-1}) \propto \text{Cauchy}(l_j^{-1}; 0, 10)\text{I}_{[0.2,3]}(l_j^{-1})$ where $\text{I}_{[\cdot,\cdot]}(\cdot)$ is the indicator function used to

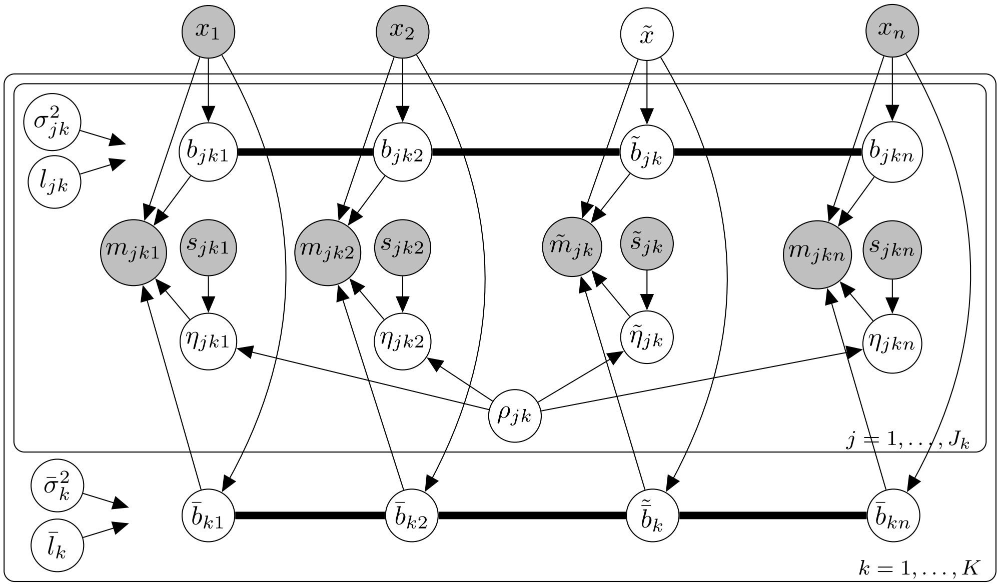


Figure 2: Graphical representation of the model. Gray circles denote the observed nodes, and white circles denote the unknown variables and functions. The variable $\tilde{x}$ denotes the parameter of interest in system for which we do not know the true system parameter value and variables x with gray background correspond to the calibration systems. The thick black lines denote GP with undirected links between all pairs of latent variables where $\bar{b}_{ki}$ denotes the ith groupwise latent variable of group k and b_{jki} is the ith latent variable of expert j in group k. The inner panel includes the expert wise GP and the outer panel includes the groupwise GPs (Assumption *A5*). Under assumption *A5** (prior independence between experts) the outer panel is removed from the model.

truncate the probability distributions thus restricting the parameter values to a closed interval. A graphical representation of the models is shown in Figure 2.

Before proceeding to the description of posterior inference in Section 3.2, we first elaborate some of the model assumptions. Naturally, the bias function $\beta(x)$ does not need to depend on the true system parameter value, x, directly. The plant coverage estimation is an example where the direct dependence on x is a justified assumption. An expert can, in principle, observe the value of the system parameter (plant coverage), and thus the observation model $p(\mathcal{D}_{ji}|x_i)$ describes how well she is able to translate her visual observation into quantitative estimate for x_i. When an expert cannot observe the true value but bases her assessment on indirect information on x_i we can denote by I_{ji} the background information expert j has about system i and by $p_j(x_i|I_{ji})$ her belief on x_i conditional on I_{ji}. She then summarizes this belief with $\mathcal{D}_{ji}$ so that we can denote by $p(\mathcal{D}_{ji}|I_{ji})$ the model for her belief on the parameter value conditional on her background information. If $p(I_{ji}|x_i)$ denotes the dependence between an expert's background information and the true parameter value, the marginal distribution for expert assessment can be written as $p(\mathcal{D}_{ji}|x_i) = \int p(\mathcal{D}_{ji}|I_{ji})p(I_{ji}|x_i)dI_{ji}$. Hence, if we do not know I_{ji} and $p(I_{ji}|x_i)$ the model reduces to the original one. An example where an expert's bias could be directly related to her (true) belief is an application where an expert is (implicitly) biased through her precautionary attitude. Let $\mathcal{D}_{ji}$ again denote the summary statistics of her true belief and let $\mathcal{D}^{\mathrm{r}}_{ji}$ denote the values she reports to us. A model $p(\mathcal{D}^{\mathrm{r}}_{ji}|\mathcal{D}_{ji})$ then summarizes how she reports (either consciously or unconsciously) her belief with different $\mathcal{D}_{ji}$. The marginal distribution for expert assessment can now be written as $p(\mathcal{D}^{\mathrm{r}}_{ji}|x_i) = \int p(\mathcal{D}^{\mathrm{r}}_{ji}|\mathcal{D}_{ji})p(\mathcal{D}_{ji}|I_{ji})p(I_{ji}|x_i)d\mathcal{D}_{ji}dI_{ji}$ so if we do not know $\mathcal{D}_{ji}$ or $p(\mathcal{D}_{ji}|x_i)$ the model reduces again to the original one. The difficulty of eliciting the expert's true belief is discussed in more detail by, e.g., O'Hagan and Oakley (2004) and O'Hagan et al. (2006).

Naturally, the bias can also be independent of the true value x. In that case, if an expert can provide useful information about x, the bias function would be a constant whereas if the expert cannot provide any information about x the bias function would be such that $p(\mathcal{D}_{ji}|x_i) \propto 1$. Note that in the absence of direct dependence between x and the bias function, the smoothness and continuity assumptions for the bias function in models 1-3 imply that $p(I_{ji}|x_i)$ should also vary smoothly and continuously with respect to x. That is, the experts should have qualitatively "similar" background information from all systems. If this cannot be assumed, for example due to qualitatively different systems, it would be natural to define own bias function for these systems leading to hierarchical bias over systems similarly as the hierarchical structure over experts in assumption (*A5**). Alternatively, we could add into the model covariates that distinguish the systems. In an extreme case, when the systems are so different that expert assessments for them should be treated independent, we could not learn between the systems.

If available, we should add into model covariates, $\mathbf{z}$, that describe experts and the systems. In this case the update rule (3.1) is revised to

$$p(\mathbf{x}|\mathcal{D}, \mathbf{z}) \propto p(\mathcal{D}|\, \mathbf{x}, \mathbf{z})p(\mathbf{x}\,|\mathbf{z}). \tag{3.6}$$

The observation model $p(\mathcal{D}|\, \mathbf{x}, \mathbf{z})$ is as in *(A4)* with $\mu_j(\mathbf{x})$ replaced by $\mu_j(\mathbf{x}, \mathbf{z})$ and the bias function $\beta(x, \mathbf{z})$ would be a function of the covariates as well. For example, in our case study, the covariates could describe experts experience (e.g., years in fisheries stock assessment work) or quality and amount of data on an individual fish stock available to the experts. In plant percentage cover estimates, an informative covariate could be the size of the plot which governs whether an expert can survey it thoroughly or only partially. As done in the case study (Section 4.2), in the absence of covariates we can use random effects to explain heterogeneity in the bias function that cannot be explained by x alone. In (3.6), the original prior $p(\mathbf{x})$ is updated to $p(\mathbf{x}\,|\mathbf{z}) \propto p(\mathbf{z}|\,\mathbf{x})p(\mathbf{x})$, that is, our *posterior* distribution for $\mathbf{x}$ in the light of the covariates. Hence, if covariates themselves are informative on $\mathbf{x}$, that is $p(\mathbf{x}\,|\mathbf{z}) \neq p(\mathbf{x})$ and not only on bias function, we can first update our understanding based on the covariates and after that use the expert assessment to provide us more information through $p(\mathcal{D}|\,\mathbf{x}, \mathbf{z})$. If we assumed that the covariates contain all the background information of experts, $\mathbf{z} = I$, the model for expert assessment would be $p(\mathcal{D}|\,\mathbf{x}, \mathbf{z}) = p(\mathcal{D}|\mathbf{z})$ and our posterior distribution would reduce to

$p(\mathbf{x}|\mathcal{D},\mathbf{z}) \propto p(\mathbf{x}\,|\mathbf{z})$. That is, we would not gain anything from experts' opinions. This illustrates also that expert elicitation is reasonable only if experts are assumed to have information about $\mathbf{x}$ that can be obtained only through their assessment for it.

3.2 Posterior inference

In order to learn about the parameters of the experts' observation models, we collect calibration data

$$\mathcal{C} = \{(m_{ji}, s_{ji}, x_i), i = 1, \dots, n, j = 1, \dots, J\}\,,$$

which consists of 3-tuples of expert's means, measures of uncertainty and known parameter values. Each calibration system is chosen so that we know the real parameter value corresponding to that system. Furthermore, we will denote by $\tilde{x}$ the unknown system parameter value which we want to estimate based on the expert assessments. Similarly, $\tilde{\mathcal{D}} = \{(\tilde{m}_j, \tilde{s}_j), j = 1, \dots, J\}$ denotes the J expert assessments for that system.

In our first task, we are interested in studying the expert bias. In order to visualize the bias, we define a vector $\tilde{\mathbf{x}} = [\tilde{x}_1, \dots, \tilde{x}_p]$ of fixed parameter values at p regularly spaced intervals and denote by $\tilde{\mathbf{b}}_j = [b_j(\tilde{x}_1), \dots, b_j(\tilde{x}_p)]$ the corresponding vector of latent bias variables of jth expert at those values. Now we can calculate the posterior distribution

$$p(\tilde{\mathbf{b}}_j|\mathcal{C},\tilde{\mathbf{x}}) \propto \int p(\tilde{\mathbf{b}}_j,\mathbf{b}|\vartheta,\mathbf{x},\tilde{\mathbf{x}})p(\eta|\vartheta)p(\vartheta)\prod_{j=1}^{J}\prod_{i=1}^{n} p(m_{ji},s_{ji}|x_i,b_{ji},\eta_{ji})d\eta d\vartheta d\mathbf{b}, \quad (3.7)$$

where $\mathbf{b} = [b_{ji}]_{i,j=1}^{n,J}$ is a vector of latent bias parameters $b_{ji} = b_j(x_i) + \bar{b}_{k_j}(x_i)$ for each expert at calibration data points and ϑ collects all the covariance function parameters and the scaling parameters. Due to GP priors $p(\tilde{\mathbf{b}}_j, \mathbf{b}|\vartheta)$ is a multivariate Gaussian which allows the information flow between the calibration data and $\tilde{\mathbf{b}}_j$ (See also Figure 2). Once we have solved $p(\tilde{\mathbf{b}}_j|\mathcal{C}, \tilde{\mathbf{x}})$ we can calculate the posterior distribution of expert biases $\tilde{\beta}_j = [\beta_j(\tilde{x}_1), \dots, \beta_j(\tilde{x}_p)]$. The (3.7) generalizes also to joint distribution of all experts' bias functions.

In our second task, we want to calculate the posterior distribution for $\tilde{x}$ in a new system for which we do not know $\tilde{x}$. In this case, we calculate the posterior distribution

$$p(\tilde{x}|\mathcal{C},\tilde{\mathcal{D}}) \propto \int p(\tilde{x})p(\tilde{b}_1,\dots\tilde{b}_J,\mathbf{b}|\vartheta,\mathbf{x},\tilde{x})p(\eta|\vartheta)p(\vartheta)$$
$$\prod_{j=1}^{J}\left(p(\tilde{m}_j,\tilde{s}_j|\tilde{x},\tilde{b}_j,\tilde{\eta}_j)\prod_{i=1}^{n}p(m_{ji},s_{ji}|x_i,b_{ji},\eta_{ji})\right)d\eta d\vartheta d\mathbf{b} d\tilde{b}_1,\dots d\tilde{b}_J, \quad (3.8)$$

where $p(\tilde{x})$ is our prior distribution for $\tilde{x}$. In this work, we used $\tilde{x} \sim \mathrm{Uniform}(0,1)$.

The posterior inference was conducted using Markov chain Monte Carlo sampling with probabilistic programming language Stan version 2.9.0 (Stan Development Team, 2016; Hoffman and Gelman, 2014). We used a warmup period of 1000 samples after which the next 10000 samples were recorded. To speed up the inference, we sampled standardized latent variables $\dot{\mathbf{b}} \sim N(0, \mathbf{I})$ as a model parameter in the code. The original latent variables were then obtained by $\mathbf{b} = L\dot{\mathbf{b}}$ where L is the Cholesky decomposition of the covariance matrix of $\mathbf{b}$ (Vanhatalo and Vehtari, 2007). With this reparameterization of the model there were no significant problems with the convergence of the Markov chain Monte Carlo (MCMC) chains. We fitted approximately 300 models (see experiments) and only 5 of them had not converged nor mixed well at the first attempt. The MATLAB and Stan codes for the posterior inference are available in GitHub (`https://github.com/Tommipe/expert_calibration`)

3.3 Model assessment and comparison

We examined our models' performance in estimating the experts' bias functions $\beta(x)$ using simulated data with varying levels of bias and uncertainty. The models' performance in inferring an unknown system parameter $\tilde{x}$ was tested using both the simulated data and a real case study. In the real case study, we used leave-one-out cross validation, whereas in the simulation studies, the tests were done at 10 equally spaced values for $\tilde{x}$ in the interval $[0.001, 0.999]$. The simulations are described in Section 4 and the real case study in Section 4.2.

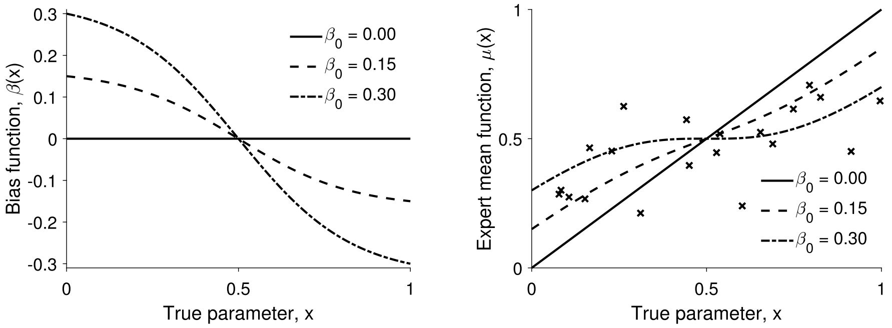
In both the simulated and the real case study, we compared the models' performance using the log point-wise posterior density statistics (LPD, Vehtari and Ojanen (2012)), i.e., we calculate the value of the log posterior density function at the true parameter values. Since the posterior inference was conducted with MCMC, we approximated the posterior densities for $\beta(x)$ and $\tilde{x}$ with a kernel density estimator. In the simulation studies, we also calculated the root mean squared error (RMSE) between the median of the posterior distribution of the parameter and the true parameter value. Log posterior density statistics is a proper scoring rule for comparing competing models (Vehtari and Ojanen, 2012). However, it is often useful to examine in more detail the model's posterior distributions from the calibration perspective. Calibration refers to statistical consistency between the posterior distributions and observations; that is in the long term, the frequency of events with probability p should be p (Gneiting et al., 2007). We examined the calibration of the posterior distributions by calculating the coverage of the 50%, 75% and 90% central probability intervals (CPI). That is, we calculated the frequency of how often the true value was inside these posterior probability intervals. In the simulation study, we calculated this for both the posterior distributions of the bias function and the unknown $\tilde{x}$. In the real case study, only the true parameter value was known, and thus we calculated this only for the latter case.

4 Experiments

4.1 Simulation experiments

Simulation set up

We generated simulation data sets for scenarios with one expert who has different levels of bias and uncertainty in her assessments. We used a sigmoid function to model the



(a) The simulated bias function, $\beta(x)$.

(b) Mean of the Beta-distribution.

Figure 3: The simulated bias function (4.1) with $q = 1.1$, and the corresponding mean of the Beta distribution, $\mu(x) = x + b(x)$. On the right hand also an example of simulated data, shown by crosses, with $\eta = 15$, $n = 20$ and $\beta_0 = 0.30$.

bias function

$$\beta(x) = \beta_0 q \left(\frac{1 - \left(\frac{q+1}{q-1}\right)^{2x-1}}{1 + \left(\frac{q+1}{q-1}\right)^{2x-1}} \right), \tag{4.1}$$

where the parameter $\beta_0 = \beta(0)$ is the maximum bias obtained at the lower bound of the interval at $x = 0$, and the parameter $q > 1$ controls how close to the asymptotic values the sigmoid function gets at the boundaries $x = 0$ and $x = 1$. The function $\beta(x)$ is a monotonically decreasing function which is motivated by the fact that the bias cannot be negative at zero nor positive at one. See Figure 3.

For the calibration data, we first generated true parameter values $x_i \sim \text{Uniform}(0, 1)$ for $i = 1, \ldots, n$ systems. Then we drew the corresponding expert mean estimates m_i from a Beta-distribution $m_i \sim \text{Beta}\left(\mu(x_i)\eta, (1 - \mu(x_i))\,\eta\right)$ where $\mu(x_i) = x + \beta(x_i)$ is the biased expert mean. We tested three different levels of maximum bias $\beta_0 \in \{0, 0.15, 0.3\}$ and generated data sets with different number of expert assessments $n \in \{5, 10, 20\}$. Furthermore, to test the effect of the level of noise in the data, we tested three different values of the dispersion parameter, $\eta \in \{5, 15, 80\}$. We generated 10 data sets with each combination of n, β_0 and η resulting with 270 data sets. For each calibration data set we generated also expert assessments for 100 test values of x whose real values were not included into inference. We then evaluated the models' ability to infer these test values.

As a second set of test data we generated otherwise similar data sets but drew the dispersion parameter η from $\text{Gamma}(\text{shape} = 1.3437, \text{scale} = 11.1208)$, for which the 90% CPI is $[1.4463, 40.3983]$. Each simulated dispersion parameter was then used as an expert's uncertainty assessment, that is $s_{ji} = \eta_{ji}$. The parameters of the Gamma distribution were obtained by fitting a Gamma distribution to the experts' uncertainty estimates in the case study data (see Section 4.2) using maximum likelihood estimation. Here, we also varied the number of data points $n \in \{5, 10, 20\}$ and the value of maximum

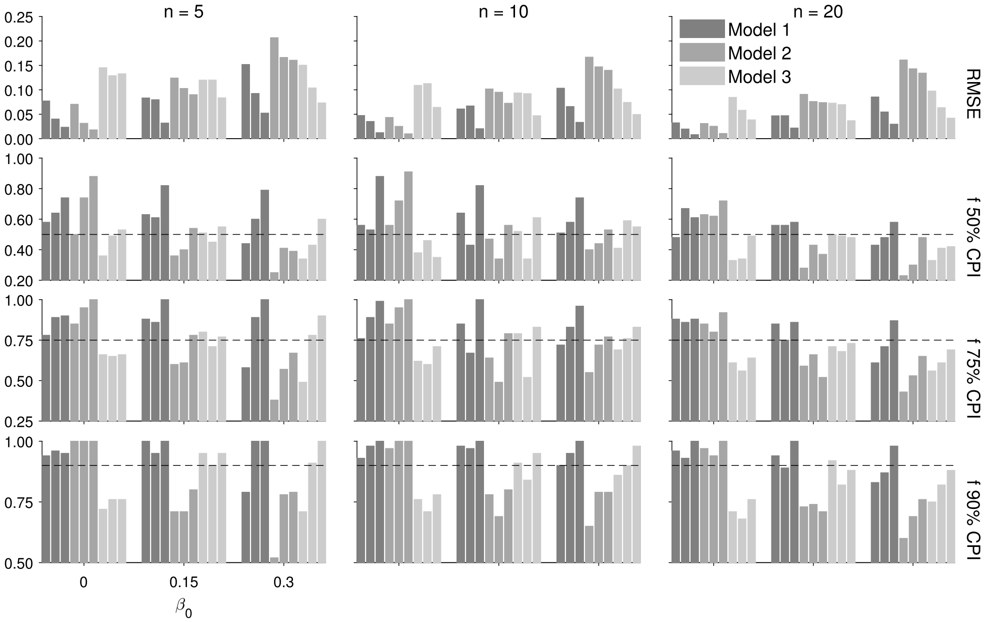


Figure 4: Results of simulation experiment on the bias inference when using only expert's mean estimates as data (assumption $A4^*$). Colors denote models as shown in the legend. The maximum bias used in the simulations, β_0 is shown on the x-axis. Each group of three bars includes three dispersion parameters increasing from left to right. The number of calibration systems, n, increases from the left column to the right.

bias $\beta_0 = \{0, 0.15, 0.3\}$. Again, we generated 10 data sets with each combination of n and β_0 ending up with total of 90 data sets. For each calibration data set we also generated expert assessment for 100 test values of x whose real values were not included in calibration data. We then calculated the models' ability to infer these test values.

Results

Figure 4 summarizes the results for the bias function inference when experts' uncertainty estimates were not used as data. The figure shows only RMSE and posterior probability interval tests. The log posterior density results were qualitatively similar to the RMSE results. In terms of RMSE, all models perform better as the number of calibration systems and the dispersion parameter used in simulations increase. This is reasonable since both lead to more informative data. Hence, the more calibration systems we have, and the more accurate experts are, in terms of η_{ji}, the better we can infer the bias function. However, there are clear differences between models' relative performances with different levels of bias. The performances of Model 1 and Model 2 decrease as the bias increases whereas the performance of Model 3 remains rather stable. Moreover, Model 1 and Model 2 are practically equally good when there is no bias ($b_0 = 0$);

with the intermediate bias ($b_0 = 0.15$), Model 1 is the best and Model 2 and Model 3 have approximately equal RMSE; and when bias is largest ($b_0 = 0.3$), Model 2 is the worst and Model 1 and Model 3 perform practically equally well. These differences are reasonable since models 1 and 2 have most of the prior probability mass near zero bias whereas Model 3 has most of the prior probability mass for non-zero biases. Hence, if there is strong *a priori* information for bias models 1 and 3 should be preferred. We also examined visually individual bias function estimates (not shown here). In all cases the posterior of the bias function was able to track the simulated bias and the match was the better the more informative the data was.

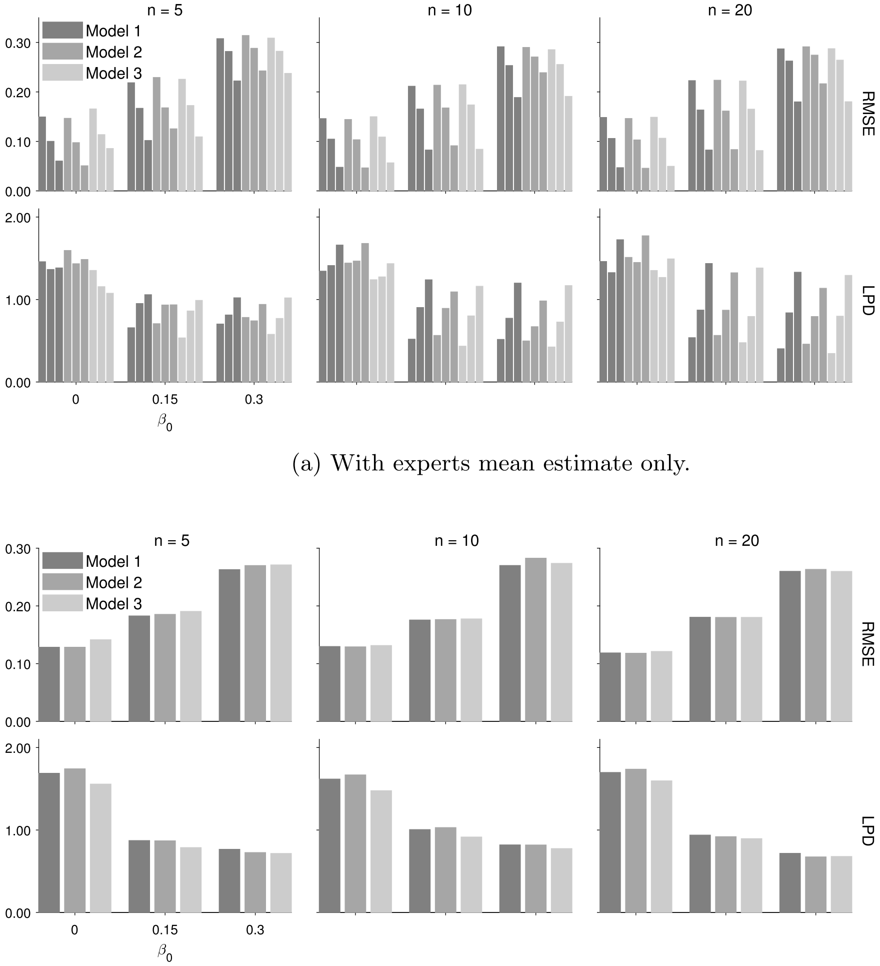
All models performed the better the more data and the less uncertainty in the expert assessments there was also in terms of coverage of the central posterior probability intervals. In the absence of bias, Model 1 and Model 2 have on average too wide 50% and 75% CPIs and very accurate 90% CPIs whereas Model 3 has on average too narrow CPIs in all these cases. This is again reasonable since Model 3 has the smallest prior probability for small biases. With larger biases Model 3 has the most accurate 50% CPI whereas the other intervals are approximately equally good among the models. However, the statistics in Figure 4 are averages over ten equally spaced values in $(0, 1)$ and only over 10 replicates so there is considerable amount of noise in these values. Thus, this test mostly shows that the models work as envisaged but their long-term performance would need to be further tested for the specific problem at hand.

Figure 5a shows the RMSE and LPD for simulation experiment for inferring x when the expert's uncertainty estimates were not used as data. As the bias function inference, the system parameter inference performs the better the more calibration data there is and the smaller the expert's uncertainties, s_{ji}, are. Again, Models 1 and 2 performed better than Model 3 in absence of bias, and once bias increases the difference between Model 1 and Model 3 vanishes whereas Model 2 has slightly worse RMSE and LPD statistics. However, the differences between models are smaller than in the bias function inference. The reason for this is that the posterior distributions for $\tilde{x}$ are much more skewed than those for the bias functions. Similarly, the posterior densities evaluated at true values of $\tilde{x}$ are smaller than those for the biases for which reason the differences between LPDs are also smaller.

The between model differences are also similar when the experts' uncertainty estimates are included in the calibration data but the absolute differences between the models are smaller than when using only the mean estimate. This is illustrated in Figure 5b, which shows the RMSE and LPD tests for the task of inferring the unknown system parameter $\tilde{x}$. In the bias function inference the differences in LPD were similar to the differences between the models when only the experts' mean estimates were used. However, CPIs better reflected the true (simulated) distribution of the biases.

4.2 Real data case study on data-limited fisheries

As shortly described in Section 2, Chrysafi et al. (2019) tested fisheries experts' performance in a simulated elicitation experiment. The authors used 18 data-rich stocks and two simulated stocks (total of $n = 20$ systems) with known (model derived) stock



(a) With experts mean estimate only.

(b) With experts mean and uncertainty estimate.

Figure 5: Results of the simulation experiments on inferring the unknown system parameter x. Colors denote different models. The groups of three same colored bars have the same maximum bias and the level of noise decreases (dispersion parameter increases) from left to right within the groups. The maximum bias is shown on the x-axis. The number of calibration points increases from the left column to the right.

status system parameter, x_i, to build the study. The elicitation included six experts (two inexperienced, two novice and two experienced) to account for the differences in the degree of bias and uncertainty associated with the different levels of experience in stock assessment. The authors provided the experts with data on catch history, limited entries of commercial length compositions, life history and the starting year of management actions, aiming to imitate the amount of information available to experts in real data-limited situations. With the data provided and their prior knowledge attributed to their education and experience, the experts formulated their beliefs about the stock status system parameters as beta distributions with the aid of quantiles, which are typically easier for experts to estimate than means or variances (Garthwaite et al., 2005; O'Hagan et al., 2006; Dias et al., 2018). Here, we will compare the elicited distributions to (point) estimates of the stock status from the corresponding data-rich fisheries stock assessment models or simulated stocks. Even though the data-rich stock assessment estimates are not the true stock status, they represent the best available estimate and for the purposes of this work they are treated as the true values of the system parameter, x_i. In the case of simulated stocks, the true stock status is known.

It is important to note that in this application, the experts make inferences about the stock status, utilizing not only the provided data, but also their experiential knowledge about the fish population dynamics and the fisheries targeting those populations. Hence, they possess relevant information about the stock status that the fisheries manager (the analyst in this work) does not. With this tacit information, the experts provide additional relevant information that would not otherwise be available to the analyst.

When analyzing the data, we noticed that there was a significant amount of variability between the experts' assessments. Hence, we added a system and expert specific random effect, $\epsilon_{ji} \sim N(0, \sigma_\epsilon^2), \sigma_\epsilon^2 \sim \text{Cauchy}(0, 0.1)\text{I}_{[0,1]}$ to the latent bias variables $b_j(x)$. This random effect accounts for the occasional non-typical errors in the expert assessments. We validated the performance of this extended model with similar simulation studies as detailed in Section 4.1. The model validation showed that these extended models performed otherwise similarly to the original models (Section 4.1) but, as expected, the posterior distributions were just wider.

We compared the models presented above to a simple "random guess" and to each expert's own assessments. In the former, the stock status was given a uniform distribution, $p(x) \propto \delta_{[0,1]}(x)$, which represents total ignorance. In our framework, the random guess corresponds to the analyst's prior and, hence, to a situation with no expert assessments. For the expert elicitation to be useful the performance of our models should improve when conditioned on the experts' assessments as opposed to only using the analyst's prior distribution. Moreover, since various pooling methods are the most common approach to combine probabilistic expert assessments, we compared our method also to simple linear and logarithmic pooling with equal weights (O'Hagan et al., 2006).

Results

Tables 1 and 2 show the model comparison with leave-one-out cross-validation LPD and frequency of the true system parameter being captured by 50%, 75% and 90% posterior CPIs. Model 3 performed the best with each combination of the calibration data and both the hierarchical and the non-hierarchical prior structure for $b(x)$ in terms of both LPDs and the posterior CPIs. In general, when only the experts' mean estimates were used, the hierarchical models performed better than their non-hierarchical counterparts. When both the mean and the uncertainty estimates were used the non-hierarchical models performed better. In terms of LPD, none of the individual experts performed better

	Data, $\mathcal{D}_{ji}$	Model 1	Model 2	Model 3
Non-hierarchical	m_{ji}	0.0303	-0.0327	0.0592
Non-hierarchical	(m_{ji}, s_{ji})	-1.3607	0.1036	0.1375
Hierarchical	m_{ji}	0.0552	0.0165	0.1683
Hierarchical	(m_{ji}, s_{ji})	0.0483	0.0711	0.0946
Individual expert assessment		-0.8050	-2.4819	-5.4086
only		-0.0067	-0.0966	-6.8062
Uniform prior only		0		
Linear pooling, equal weights		0.0345		
Logarithmic pooling, equal weights		-1.3274		

Table 1: The average leave-one-out cross-validation LPDs of $\tilde{x}$ evaluated at the true system parameter value. The first four rows summarize the models presented in this work. The fifth and sixth rows summarize the performance of the individual experts, and the seventh row corresponds to the random guess. The last two rows show the performance of the traditional linear and logarithmic pooling with equal weights.

		50% interval			75% interval			90% interval		
	Data, $\mathcal{D}_{ji}$	M 1	M 2	M 3	M 1	M 2	M 3	M 1	M 2	M 3
Non-hierarchical	m_{ji}	0.50	0.35	0.45	0.65	0.55	0.65	0.70	0.70	0.80
Non-hierarchical	(m_{ji}, s_{ji})	0.50	0.45	0.50	0.70	0.70	0.65	0.75	0.75	0.85
Hierarchical	m_{ji}	0.35	0.35	0.50	0.65	0.60	0.75	0.70	0.70	0.80
Hierarchical	(m_{ji}, s_{ji})	0.40	0.45	0.50	0.65	0.60	0.65	0.75	0.75	0.85
Individual expert assessment		0.15	0.25	0.10	0.35	0.40	0.25	0.50	0.50	0.35
only		0.70	0.20	0.20	0.90	0.65	0.30	1.00	0.75	0.35
Uniform prior only		0.60			0.80			0.95		
Linear pooling, eq. weights		0.40			0.70			0.95		
Logarithmic pooling, eq. weights		0.05			0.25			0.40		

Table 2: The average leave-one-out cross-validation frequency of true parameter being inside p% posterior CPIs. The first four rows summarize the models presented in this work. The fifth and sixth rows summarize the performance of the individual experts, and the seventh row corresponds to the random guess. The last two rows show the performance of the traditional linear and logarithmic pooling with equal weights.

than the uniform prior. Linear pooling with equal weights outperformed the uniform prior whereas the logarithmic pooling with equal weights performed worse than the uniform prior. However, linear pooling outperformed two versions of Model 1 and Model 2 and the uniform prior outperformed one version of Model 1 and Model 2. When looking at the individual expert assessments the best uncalibrated expert performed better than the worst performing calibration model. When we examined the distribution of the cross-validation LPDs there was one system where these two models performed considerably worse than at the other systems. This poor performance in this system resulted from a single expert giving a very bad assessment. A single expert assessment does not have as big of an effect in the hierarchical models for which reason the hierarchical prior for $b(x)$ seems to offer more robustness against “outlying” assessments.

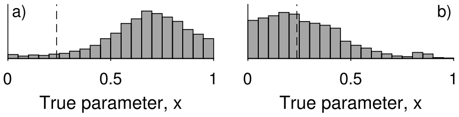


Figure 6: Examples of leave-one-out posterior densities when the true system parameter value lies far from the mode (a) and near the model (b) of the posterior distribution.

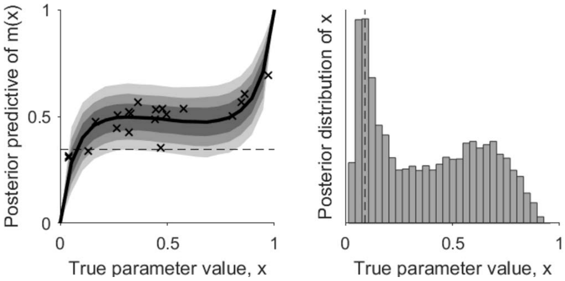


Figure 7: A simulated example of a posterior predictive distribution of the expert's mean function (black line denoting the posterior median and gray shadings 50%, 75% and 90% CPIs) and a bimodal posterior distribution of $\tilde{x}$. Here, the model does not include the expert and system specific random effect for which reason an outlying observation near $x = 0.5$ pulls the posterior predictive distribution of $m(x)$ down so that when the expert gives an assessment $m = 0.3455$ (shown by the dashed horizontal line) the resulting posterior has two modes: a higher one near the true value (shown by the dashed vertical line) and a lower one far from the true value.

The LPDs are rather close to zero. As an example, the value of the posterior density function evaluated at the true system parameter value of the best model is on average only $\exp(0.17) = 1.18$ times that of the uniform prior. However, this is in the same order of magnitude as the LPDs in the simulation studies with the largest bias, $\beta_0 = 0.3$ and $n = 20$ and approximately equal to the case with the largest simulated expert uncertainty, $\eta = 5$. Hence, the performance of the best model is comparable to simulation studies. Moreover, in most cases the expert assessments inform the posterior distribution of $\tilde{x}$ considerably well as illustrated in Figure 6b, and the minor on average improvement in the performance compared to the uniform prior results from complete failure in one or two cases as illustrated in Figure 6a. These "random outlier assessments" decrease the average performance of all models. This emphasizes also the importance of having the system and expert specific random effects in the model. These random effects increase the uncertainty related to estimates of $\tilde{x}$ which reflects the fact that there is significantly non-zero chance that experts might give occasionally very bad assessments even if they are skillful in general. Without the random effects the occasional large errors in the expert's assessment could lead to multimodal posterior distributions as illustrated in Figure 7.

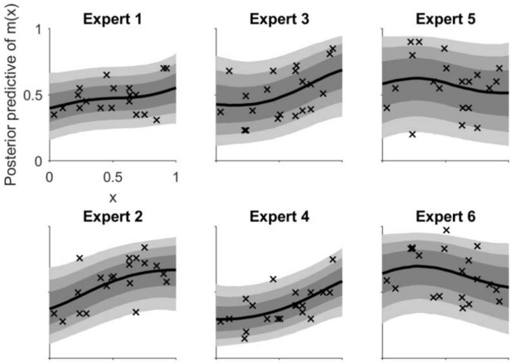


Figure 8: Posterior predictive distribution of the experts' mean estimates $m(x)$ as a function of x together with their actual assessments. The experts' point estimates are shown by crosses and the posterior predictive distributions by black line solid lines (posterior median) and gray shading (50%, 75% and 90% CPIs).

	η			l			σ^2		
	5%	mean	95%	5%,	mean	95%	5%	mean	95%
Expert 1	10.7	24.2	44.9	0.35	1.15	3.52	0.26	7.1	27.1
Expert 2	8.3	19.2	37.9	0.36	0.98	2.85	0.31	6.6	24.2
Expert 3	6.2	15.0	31.9	0.36	1.08	3.12	0.24	6.2	22.8
Expert 4	10.7	23.6	44.7	0.37	1.13	3.15	0.23	6.1	22.5
Expert 5	4.1	11.2	25.3	0.36	1.19	3.55	0.22	6.3	23.3
Expert 6	5.1	13.8	31.0	0.36	1.12	3.31	0.22	6.0	22.5

Table 3: Summary of posterior marginal distributions of the expert observation model parameters.

There was considerable variation between the quality of the expert assessments as illustrated in Figure 8 and summarized in Table 1. Four of the experts (Experts 1-4) gave consistently good assessments and their biases were "coherent" in the sense that the posterior predictive distribution of $m(x)$ was monotonically increasing. The remaining two experts (Experts 5 and 6) were not able to give useful assessments which is reflected by large uncertainty in their posterior predictive distributions of $m(x)$. Moreover, the posterior distributions of η_{ji} for Experts 5 and 6 are concentrated to smaller values than the corresponding posterior distributions of experts 1-4 (Table 3) implying partial down-weighting of them. It should be also noted, that in Figure 8 expert's mean assessment corresponds to the expected stock status which in case of the true data-rich stocks is compared to the estimate from a "golden standard" model that uses all available data. Hence, some of the variability between expert assessments and true system parameter value in Figure 8 might result from the stochasticity and uncertainty in this "golden standard" stock status. With simulated data the true stock status was known exactly.

5 Discussion and conclusions

We have presented a hierarchical Bayesian model for calibrating and combining multiple experts' assessments. We proposed three different Gaussian process priors for the expert bias functions and showed in simulation studies that these models can be used to infer the expert biases and thus to calibrate the expert assessments in a real case study.

As the size of the calibration data and the accuracy of the experts' assessments increase the models' performances improve. Moreover, the prior assumptions in the different bias models are also reflected in our models' performances. Model 1 and Model 3 seem to be the most appropriate choices, especially if the expert assessments are assumed to be biased *a priori*. Model 2 seems to contain too little prior probability for the bias at the ends of the parameter interval, which decreases its performance. When designing our models, we began with the assumption that the expert's true beliefs can be expressed as a parametric probability density function as in Albert et al. (2012), in the form of a beta distribution. Unlike Albert et al. (2012), however, we assumed that the expert data $\mathcal{D}$ fully described their beliefs about x although eliciting experts' true beliefs is challenging (O'Hagan and Oakley, 2004). We used the mean and dispersion parameters of the elicited distribution for building the experts' observation models.

In the case study, all the alternative models improved the inference of the stock status compared to the uncalibrated expert assessments. Overall, the Bayesian approach taken here performed better than the most common pooling methods (the linear and logarithmic with equal weights) and the uncalibrated individual expert assessments. Moreover, our models allowed inference about experts' reliability in terms of the bias functions and η_{ji} parameters as expected. If an expert consistently gives accurate assessments the model allows her assessments to improve estimates concerning the unknown system parameter $\tilde{x}$. On the other hand, if an expert never gives useful assessments, the model effectively ignores her assessments. The confusion resulting from occasional outlier assessments are modeled by the expert and system specific random effects in the description of $b(x)$ which increase the uncertainty on posterior of system parameter. Our assumption that the mean of the expert's observation model, μ_{ji}, does not depend on s_{ji} could be relaxed if the experts were assumed to be risk averse in the sense that they would prefer to give smaller mean estimates m_{ji} the more uncertain they were. In real fisheries stock assessment this could be reasonable. Fisheries scientists that work with stock assessment are trained to follow the precautionary approach recommended in fisheries management (Hilborn et al., 2001; Consalez-Laxe, 2005), and hence they could prefer giving assessments leading to yield loss over assessments that might lead to overfishing and even stock collapse (Chrysafi et al., 2019).

One could argue that the expert bias in the fisheries case study could be eliminated by posing the problem as a regression analysis. However, interpreting and making statistical inference from fisheries data requires knowledge and experience on fishing and data collection methods, reporting behavior of fishermen as well as on biology and behavior of fish stock. There is strong evidence that naive regression analysis based on, for example, catch and effort data is not typically applicable to fisheries stock assessment as such (Kuparinen et al., 2012). Through past experiences, fisheries experts have such knowledge which they transfer to their assessment on the system parameter when analyzing the available data in light of their experience. However, the rationale behind expert judgments should be documented as carefully as possible during the expert elicitation. As discussed around equation (3.6) expert's background knowledge could be encoded into covariates $\mathbf{z}$ that could then be used to inform the analysis through $p(\mathbf{x}\,|\mathbf{z})$ and $p(\mathcal{D}|\,\mathbf{x},\mathbf{z})$. If experts' background knowledge could be elicited fully, we would not expect expert assessment $\mathcal{D}$ to contain any extra information compared to $\mathbf{z}$. In that case $p(\mathbf{x}\,|\mathcal{D},\mathbf{z}) = p(\mathbf{x}\,|\mathbf{z})$. In fisheries studies, experts' background knowledge would typically correspond to different hypotheses behind the data generating process in which case an alternative modeling option would be Bayesian model averaging over the alternative models suggested by experts (Mäntyniemi et al., 2013). However, typically in cases where expert judgments are combined it is infeasible to elicit experts' background knowledge fully (the situation assumed here). Moreover, even if it was feasible, the problem of constructing model for $p(\mathcal{D}|\,\mathbf{x})$ would just be redefined to problem of modeling $p(\mathcal{D}_z|\mathbf{z})$ where $\mathcal{D}_z$ denotes experts assessment of the true underlying covariates; that is we would need to model our trust on how well experts' are able to describe their true background knowledge (O'Hagan and Oakley, 2004; Albert et al., 2012) and how well that resembles the true state of world.

By far the most commonly used method of combining elicited expert knowledge has been the classical model with linear pooling (Cooke and Goossens, 2008; Dias et al., 2018), whereas Bayesian methods have rarely been used. Here, we decided to keep the comparison to linear pooling simple and to exclude more sophisticated weighting schemes for two reasons. Firstly, pooling with equal weights is probably the most commonly used method (O'Hagan et al., 2006). Secondly, the classical model with unequal weights would have required many choices concerning the scoring function and its parameterization, which is out of the scope of this work. Nevertheless, the included pooling methods were used to verify that our method is at least comparable to the most standard expert assessment combination methods. Pooling methods have been argued to be easier and more straightforward to apply than the Bayesian approach (O'Hagan et al., 2006; Hartley and French, 2018). However, pooling does not represent the actual beliefs of any individual and hence, does not behave as one would expect a probability distribution to behave. For example, when expert assessment is used as a complementary source of information to available observational data, pooling can be performed either for experts' prior or posterior distributions. However, the result will be different depending on the process stage where the pooling occurs, leading to incoherent inference (O'Hagan et al., 2006; Farr et al., 2018). Moreover, none of the pooling methods allows for bias correction as such, but the bias needs to be corrected with a statistical model, whereas our model explicitly models the bias thus providing the bias correction naturally.

Our case study was based on simulated expert assessment study where the true system parameter value was known. However, an evident challenge in the expert bias inference and correction in general is how to collect the calibration data. In some cases, such as in the plant coverage estimation, calibration data could be collected from a subset of systems. In some applications, for example in assessments of failure rates of machine components, it would be possible to first collect expert assessments and then calibrate them sequentially as data from the assessed systems is collected. This is also similar to the traditional use of the classical model (Cooke and Goossens, 2008). In addition to the challenge of collecting calibration data common challenges with any (human) expert assessment calibration method are also that experts may themselves already know some of the calibration data, experts may themselves apply corrections to future judgments if they observed errors in their past judgments (O'Hagan et al., 2006) and that experts may not answer truthfully if they know their responses are going to be adjusted (O'Hagan et al., 2006, Section 4.5.4). In the first two cases we could, at least in principle, build the model to account for whether or not an expert has seen calibration data or assessment of her performance. The challenge of getting experts to answer truthfully should be communicated with the experts. Also with pooling methods an expert's assessment gets different weight depending how well she is believed to perform. One interesting future research direction could be also to extend the methods presented here for multivariate variables. This could in principle be done by extending the GP models to multinormal or multivariate Gaussian processes.

Even when there is available observational data, ecological questions pertinent to formal decision-making are characterized by uncertainty and paucity of empirical data (Kuhnert et al., 2010). At the same time, as also shown by our case study example, the degree of bias in the expert opinions can vary between experts and even within the assessments of single expert (Burgman et al., 2011; Cooke and Goossens, 2008; Kynn, 2008; de Little et al., 2018). Hence, our work is not limited to fisheries stock assessment, but it can also have applications in other fields. The plant coverage estimation was described as another example in this work and we mentioned few other examples in the Introduction. Other applications can be also found from medicine; e.g. Wilson et al. (2018) elicited expert opinions in the probability of disease progression in patients with undetected melanoma and Albert et al. (2012) assessed dose response in a contamination study. Moreover, even though our main objective was in unknown system parameter inference, our model could be used to specifically study the experts' biases.

References

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. MR2981623. doi: https://doi.org/10.1214/12-BA717. 1251, 1253, 1254, 1256, 1260, 1273, 1274, 1275

Berkson, J. and Thorson, J. T. (2014). "The determination of data-poor catch limits in the United States: is there a better way?" *ICES Journal of Marine Science*, 72(1): 237–242. 1255

Burgman, M. (2005). *Risks and Decisions for Conservation and Environmental Management*. Cambridge University Press. 1251

Burgman, M., Carr, A., Godden, L., Gregory, R., McBride, M., Flander, L., and Maguire, L. (2011). "Redefining expertise and improving ecological judgment." *Conservation Letters*, 4(2): 81–87. 1251, 1252, 1253, 1275

Chrysafi, A., Cope, J., and Kuparinen, A. (2019). "Eliciting expert knowledge to inform stock status for data-limited stock assessments." *Marine Policy*, (101): 167–176. 1253, 1255, 1257, 1267, 1273

Chrysafi, A. and Kuparinen, A. (2015). "Assessing abundance of populations with limited data: Lessons learned from data-poor fisheries stock assessment." *Environmental Reviews*, 24(1): 25–38. 1254

Clemen, R. T. and Lichtendahl, K. C. (2002). "Debiasing expert overconfidence: A Bayesian calibration model." Working paper, Duke University. 1254

Consalez-Laxe, F. (2005). "The precautionary principle in fisheries management." *Marine Policy*, 29: 495–505. 1273

Cooke, R. M. and Goossens, L. L. (2008). "TU Delft expert judgment data base." *Reliability Engineering & System Safety*, 93(5): 657–674. Expert Judgement. 1253, 1254, 1274, 1275

Cope, J. M. (2013). "Implementing a statistical catch-at-age model (Stock Synthesis) as a tool for deriving overfishing limits in data-limited situations." *Fisheries Research*, 142: 3–14. 1255

Costello, C., Ovando, D., Hilborn, R., Gaines, S. D., Deschenes, O., and Lester, S. E. (2012). "Status and solutions for the world's unassessed fisheries." *Science*, 338: 517–520. 1254

Daan, N., Gislason, H., Pope, J. G., and Rice, J. C. (2011). "Apocalypse in world fisheries? The reports of their death are greatly exaggerated." *ICES Journal of Marine Science*, 68(7): 1375–1378. 1255

de Little, S. C., Casas-Mulet, R., Patulny, L., Wand, J., Miller, K. A., Fidler, F., Stewardson, M. J., and Webb, J. A. (2018). "Minimising biases in expert elicitations to inform environmental management: Case studies from environmental flows in Australia." *Environmental Modelling and Software*, 100: 146–158. 1254, 1275

Dias, L. C., Morton, A., and Quigley, J. (2018). *Elicitation*. Springer International Publishing. MR3700912. doi: https://doi.org/10.1007/978-3-319-65052-4. 1251, 1252, 1253, 1269, 1274

Dick, E. J. and MacCall, A. D. (2011). "Depletion-Based Stock Reduction Analysis: A catch-based method for determining sustainable yields for data-poor fish stocks." *Fisheries Research*, 110(2): 331–341. 1255

Dietrich, F. and List, C. (2014). *Probabilistic opinion pooling*. Oxford University Press. MR3643906. 1253

Farr, C., Ruggeri, F., and Mengersen, K. (2018). "Prior and Posterior Linear Pooling for Combining Expert Opinions: Uses and impact on Bayesian networks." *Entropy*, 20(3): 209. 1253, 1274

Food and Agriculture Organization of the United Nations (1995). "Code of Conduct for responsible Fisheries." 1254

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. MR0567946. doi: https://doi.org/10.2307/2981768. 1253

French, S. (2011). "Aggregating expert judgement." *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, 105(1): 181–206. MR2783806. doi: https://doi.org/10.1007/s13398-011-0018-6. 1253

Froese, R., Demirel, N., Coro, G., Kleisner, K. M., and Winker, H. (2017). "Estimating fisheries reference points from catch and resilience." *Fish and Fisheries*, 18(3): 506–526. 1255

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. MR2170464. doi: https://doi.org/10.1198/016214505000000105. 1252, 1269

Gelfand, A. E., Mallick, B. K., and Dey, D. K. (1995). "Modeling Expert Opinion Arising As a Partial Probabilistic Specification." *Journal of the American Statistical Association*, 90(430): 598–604. MR1340512. 1253, 1256

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press. MR3235677. 1257

Genest, C. and Schervish, M. J. (1985). "Modeling Expert Judgements for Bayesian Updating." *The Annals of Statistics*, 13(3): 1198–1212. MR0803766. doi: https://doi.org/10.1214/aos/1176349664. 1253

Geromont, H. F. and Butterworth, D. S. (2015). "A Review of assessment methods and the development of management procedures for data-poor fisheries." *FAO report, FAO*. 1254

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. MR2325275. doi: https://doi.org/10.1111/j.1467-9868.2007.00587.x. 1264

Griffiths, S. P., Kuhnert, P. M., Venables, W. N., and Blaber, S. J. (2007). "Estimating abundance of pelagic fishes using illnet catch data in data-limited fisheries: A Bayesian approach." *Canadian Journal for Fisheries and Aquatic Sciences*, 64(7): 1019–1033. 1253

Hartley, D. and French, S. (2018). "Elicitation and Calibration: A Bayesian Perspective." In Dias, L. C., Morton, A., and Quigley, J. (eds.), *Elicitation The science and Art of Structuring Judgement*, 119–140. Springer International Publishing. MR3700917. 1253, 1254, 1256, 1260, 1274

Hilborn, R., Maquire, J., Parma, A. M., and Rosenberg, A. A. (2001). "The precautionary approach and risk management: can they increase the probability of success in fisheries." *Canadian Journal of Fisheries and Aquatic Sciences*, 58: 99–107. 1273

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research*, 1593–1623. MR3214779. 1263

Kennedy, M. C. and O'Hagan, A. (2001). "Bayesian calibration of computer models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3): 425–464. MR1858398. doi: https://doi.org/10.1111/1467-9868.00294. 1252

Kuhnert, P. M., Hayes, K., Martin, T. G., and McBride, M. F. (2009). "Expert opinion in statistical models." *18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation.*, 4264–4268. 1253

Kuhnert, P. M., Martin, T. G., and Griffiths, S. P. (2010). "A guide to eliciting and using expert knowledge in Bayesian ecological models." *Ecology letters*, 13(7): 900–914. 1253, 1275

Kuparinen, A., Mäntyniemi, S., Hutchings, J., and Kuikka, S. (2012). "Increasing biological realism of fisheries stock assessment: towards hierarchical Bayesian methods." *Environmental Reviews*, 20: 135–151. 1273

Kynn, M. (2008). "The 'heuristics and biases' bias in expert elicitation." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(1): 239–264. MR2412655. doi: https://doi.org/10.1111/j.1467-985X.2007.00499.x. 1253, 1275

Landquiste, H., Norman, J., Lindhe, A., Norberg, T., Hassellöv, I., Lindgren, J. F., and Rosen, L. (2017). "Expert elicitation for deriving input data for probabilistic risk assessment of shipwrecks." *Marine Pollution Bulletin*, 125: 399–415. 1252

Lindley, D. V. (1982). "The Improvement of Probability Judgements." *Journal of Royal Statistical Society. Series A (General)*, 145(1): 117–126. MR0662171. doi: https://doi.org/10.2307/2981425. 1253, 1254

Lindley, D. V. (1983). "Reconciliation of Probability Distributions." *Operations Research*, 31(5): 866–880. MR0726369. doi: https://doi.org/10.1287/opre.31.5.866. 1253, 1254, 1257

Lindley, D. V. and Singpurwalla, N. D. (1986). "Reliability (and fault tree) analysis using expert opinions." *Journal of the American Statistical Association*, 81(393): 87–90. MR0830568. 1252, 1253, 1254, 1256, 1257

Lindley, D. V., Tversky, A., and Brown, R. V. (1979). "On the Reconciliation of Probability Assessments." *Journal of the Royal Statistical Society. Series A (General)*, 142(2): 146–180. MR0547236. doi: https://doi.org/10.2307/2345078. 1252

Low-Choy, S., O'Leary, R., and Mengersen, K. (2009). "Elicitation by design in ecology: using expert opinion to inform priors for Bayesian statistical models." *Ecology*, 90(1): 265–277. 1252

Magnusson, A. and Hilborn, R. (2007). "What makes fisheries data informative?" *Fish and Fisheries*, 8(4): 337–358. 1254

Mäntyniemi, S., Haapasaari, P., Kuikka, S., Parmanne, R., Lehtiniemi, M., and Kaitaranta, J. (2013). "Incorporating stakeholders' knowledge to stock assessment: Central Baltic herring." *Canadian Journal of Fisheries and Aquatic Sciences*, 70(4): 591–599. 1274

McConway, K. (1981). "Marginalization and Linear Opinion Pools." *Journal of the American Statistical Association*, 71: 410–414. MR0624342. 1253

Meissa, B., Gascuel, D., and Rivot, E. (2013). "Assessing stocks in data-poor African fisheries: a case study on the white grouper Epinephelus aeneus of Mauritania." *African Journal of Marine Science*, 35: 253–267. 1254

Methot, R. D. and Wetzel, C. R. (2013). "Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management." *Fisheries Research*, 142: 86–99. 1254

Morgan, M. G. (2014). "Use (and abuse) of expert elicitation in support of decision making for public policy." *Proceedings of the National Academy of Sciences*, 111(20): 7176–7184. 1251, 1253

Morris, P. A. (1974). "Decision Analysis Expert Use." *Management Science*, 20(9): 1233–1241. 1253

Nevalainen, M., Helle, I., and Vanhatalo, J. (2018). "Estimating the acute impacts of Arctic marine oil spills using expert elicitation." *Marine Pollution Bulletin*, 131: 782–792. 1252

Newman, D., Berkson, J., and Suatoni, L. (2015). "Current methods for setting catch limits for data-limited fish stocks in the United States." *Fisheries Research*, 164: 86–93. 1255

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006). *Uncertain Judgements: Eliciting Experts' Probabilities*. John Wiley & Sons. 1251, 1252, 1253, 1256, 1262, 1269, 1274, 1275

O'Hagan, A. and Oakley, J. E. (2004). "Probability is perfect, but we can't elicit it perfectly." *Reliability Engineering & System Safety*, 85(1): 239–248. 1262, 1273, 1274

Roman, H. A., Walker, K. D., Walsh, T. L., Conner, L., Richmond, H. M., Hubbel, B. J., and Kinnery, P. L. (2008). "Expert Judgment Assessment of the Mortality Impact of Changes in Ambient Fine Particulate Matter in the U.S." *Environmental Science & Technology*, 42(7): 2268–2274. 1251

Salas, S., Chuenpagdee, R., Seij, J., and Charles, A. (2007). "Challenges in the assessment and management of small-scale fisheries in Latin America and Caribbean." *Fisheries Research*, 87: 5–16. 1254

Speris-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., and Burgman, M. (2010). "Reducing Overconfidence in the Interval Judgements of Experts." *Risk Analysis*, 30: 512–523. 1253

Stan Development Team (2016). "Stan: A C++ Library for Probability and Sampling, Version 2.9.0." URL `http://mc-stan.org/`. 1263

Tversky, A. and Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and Biases." *Science*, 185(4157): 1124–1131. 1252, 1253

Usher, W. and Strachan, N. (2013). "An expert elicitation of climate, energy and economic uncertainties." *Energy Policy*, 61: 811–821. 1252

Uusitalo, L., Kuikka, S., and Romakkaniemi, A. (2005). "Estimation of Atlantic salmon smolt carrying capacity of rivers using expert knowledge." *ICES Journal of Marine Science: Journal du Conseil*, 62(4): 708–722. 1252

Vanhatalo, J. and Vehtari, A. (2007). "Sparse Log Gaussian Processes via MCMC for Spatial Epidemiology." *JMLR Workshop and Conference Proceedings*, 1: 73–89. 1264

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6: 141–228. MR3011074. doi: `https://doi.org/10.1214/12-SS102`. 1264

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT Press. MR2514435. 1258

Wilson, E. C., Usher-Smith, J. A., Emery, J., Corrie, P. G., and Walter, F. M. (2018). "Expert elicitation of multinomial probabilities for decision-analytic modelling: An application to rates of disease progression in undiagnosed and untreated melanoma." *Value in Health*, in press. 1251, 1275

Zickfeld, K., Morgan, M. G., Frame, D. J., and Keith, D. W. (2010). "Expert judgments about transient climate response to alternative future trajectories of radiative forcing." *Proceedings of the National Academy of Sciences*, 107(28): 12451–12456. 1251

Acknowledgments

TP was funded by the Academy of Finland (through Academy Research Fellow Grant to Anna Kuparinen), by the University of Jyväskylä, and by the European Research Council (COMPLEX-FISH 770884 to Anna Kuparinen). AC was funded by the University of Helsinki through DENVI Doctoral Program Fellowship. JV has additionally been funded by Academy of Finland [Grants 304531 and 317255] and the research funds of University of Helsinki [decision No. 465/51/2014]. We thank the editor and the anonymous reviewers for their thorough work and constructive comments that helped significantly improve the paper.