Computational Geosciences
https://doi.org/10.1007/s10596-023-10241-2

ORIGINAL PAPER

Bayesian model evaluation for multiple scenarios

Sigurd Ivar Aanonsen[1] · **Kristian Fossum**[1] · **Trond Mannseth**[1]

Received: 19 September 2022 / Accepted: 18 July 2023

Abstract
Traditional uncertainty analysis for subsurface models is typically based on a single dynamic model with a number of uncertain parameters. Improved and more robust forecasting can be obtained by combining several models in a Bayesian setting using model averaging. The traditional Bayesian Model Averaging (BMA), however, suffers from several drawbacks, such as too large sensitivity to prior model assumptions and instability with respect to measurement perturbations, especially when the number of measurements is large. We suggest a modified version of BMA (MBMA) where the calculations are stabilized using an ensemble of measurements. Bayesian stacking (BS) is a method that is directly focused on the performance of the combined predictive distribution of several models. The original version of BS (BSLOO) is based on leave-one-out cross-validation and requires a Bayesian inversion for each data point which may be very time consuming. We suggest a modified version of stacking (MBS) that requires only a single history match and uses an ensemble of measurements. MBS may be used with either prior (MBS-pri) or posterior (MBS-post) predictive distributions. The behavior of the methods is illustrated using three synthetic, linear examples. One is a simple mixture model. The other two are inspired by 4D seismic data. The results with MBS-pri are very similar to the results with MBMA. The results with MBS-post are similar to those of BSLOO when the data are uncorrelated. MBS can take into account correlated data or measurement errors, while correlations are neglected in the BSLOO weight calculations.

Keywords Uncertainty quantification · Bayesian model averaging · Bayesian stacking · Model combination

1 Introduction

Dynamic models of the underground typically depend on a large number of unknown and/or uncertain parameters and should also be conditioned to a set of, typically time-dependent, data. While uncertainty in model parameters is commonly taken into account when solving subsurface-related inverse problems, the uncertainty related to the model itself is most often ignored. This is not because the involved models are certain to be correct, but reflects that model uncertainty is challenging to handle properly. Most approaches to uncertainty analysis are still based on perturbations around a single "base case" model, and the uncertainty related to the model itself is ignored even if several alternative scenarios may often be possible *á priori*.[1] Our goal is to make robust predictions and improve the uncertainty quantification for the predictions by keeping, in principle, all viable models and average these in a Bayesian setting. That is, given a set of viable models, $\mathcal{M} = \{M_1, ..., M_K\}$, which have been fitted to the data separately, we want to make predictions by combining the posterior distributions for the individual models using a posterior model weight.

Bayesian model probability (BMP) or model evidence (BME) to discriminate between alternative models have been applied within a number of fields during the last few decades, including medicine [1], weather forecasting [2] and climate research [3]. The posterior model probabilities are also typically used as the weighting factors in Bayesian model selection and averaging (BMS/BMA) [4]. A drawback of

Kristian Fossum and Trond Mannseth contributed equally to this work.

✉ Sigurd Ivar Aanonsen
siaanons@gmail.com

Kristian Fossum
krfo@norceresearch.no

Trond Mannseth
trma@norceresearch.no

[1] NORCE Norwegian Research Centre, P.O. Box 22 Nygårdstangen, 5838 Bergen, Norway

[1] Notice that in this paper we will use both "models" and "scenarios" and not distinguish between these.

Published online: 29 August 2023

BMA in the context of reservoir modelling is that it will asymptotically select the candidate model that is closest to the true model in Kullback-Leibler divergence [5]. That is, for a large amount of data, such as seismic data, the BMP will often be 1.0 for one of the models, even if several of the models could explain the data within the given uncertainty. It will also be very sensitive to the prior models as well as uncertainty in the measurements.

How the true data-generating model relates to the candidate models have been classified as $\mathcal{M}$-closed, $\mathcal{M}$-complete, and $\mathcal{M}$-open, see e.g., Yao et al. [5]. $\mathcal{M}$-closed means that one of the candidate models is actually the data-generating model. Thus, when the amount of data increases, the BMP for this model *should* converge to 1.0, and other values only "reflect a statistical inability to distinguish the hypotheses based on limited data" [6]. In the $\mathcal{M}$-complete and $\mathcal{M}$-open settings, the true model is not any of the candidate models. In $\mathcal{M}$-complete, the true model is known, but it is not practical to include it in the list of candidate models, while $\mathcal{M}$-open refers to a situation where it is not possible to specify the true model because it is too difficult conceptually or computationally [5]. Höge et al. [7] introduces even a 4th setting, Quasi-$\mathcal{M}$-closed, where the true model is not in the candidate model list, but very close to one of them. Within this context, BMA is appropriate only for the $\mathcal{M}$-closed situation, but not in the $\mathcal{M}$-complete and $\mathcal{M}$-open settings. $\mathcal{M}$-open would be the normal case in reservoir modelling where the true reservoir typically is far more complex than any of the models.

To avoid some of the problems with BMA, Gelman [8] recommends to expand the discrete set of models into one continuous model family if possible. Alternatively, one could use other methods for Bayesian model combination than BMA. Yao et al. [5] discuss alternative model weighting approaches, including Pseudo-BMA, Akaike weights and Bayesian Stacking (BS), and generally recommend stacking (of predictive distributions) for the task of combining separately-fit Bayesian posterior predictive distributions. With BS one tries to find an optimal combination in the space spanned by all candidate models, which in a Bayesian context means finding a predictive distribution that is close to the true data generating distribution. Höge et al. [7] compare various Bayesian model combination methods for finding the best model for the spatial distribution of hydraulic conductivity from a sandbox lab experiment. The methods considered are BMS/BMA, pseudo BMS/BMA, BS and Bayesian bootstrapping (BB). Höge et al. [7] recommend using BS in $\mathcal{M}$-open situations if averaging of distributions for broad coverage of predictive uncertainty is the goal. The version of BS introduced in [5] (which we will denote BSLOO) is based on the principle of "Leave-One-Out Cross-Validation" (LOOCV).

We aim at applying Bayesian model combination for large subsurface models—with 4D seismic data in particular—and the amount of data may be huge. Also, there is typically only one set of measurements, which may consist of correlated or uncorrelated data. When applying Bayesian inversion to such problems, the actual measurement vector, or sometimes a smoothed version, is assumed to be the mean of the random data variable, and the statistical properties of this variable is represented by the measurement error PDF. Also these measurements will often not be independent and identically distributed (iid), and the validity of the LOOCV approaches may be questionable [7]. Thus, we suggest a modified version of BS (MBS) and replace the LOOCV-replications with an ensemble of data realizations. MBS can be used with both prior and posterior predictive distributions and take into account correlated measurement errors. We also suggest to stabilize the BMA calculations by averaging BMP over many data realizations along the lines used by Hong et al. [9] to integrate model uncertainty in a probabilistic decline curve analysis. This modified version of BMA will be denoted MBMA.

In the next section we define the concepts of model likelihood or model evidence, model probability and model averaging and introduce the modified versions of stacking. Then the alternative methods for calculating model weights are compared on some simple, linear Gaussian examples where all calculations can be performed analytically. We demonstrate that MBS based on posterior predictive distributions give similar results as BSLOO when measurements are iid. We also demonstrate MBMA may greatly improve the results from BMA by reducing the sensitivity to the measurements and prior assumptions as well as the tendency that the probability of one model approaches 100% when the number of measurements is large.

2 Methods for Bayesian model combination

Let h denote a quantity of interest. h may for instance be Net Present Value of a given development strategy for a petroleum reservoir. It could also be an unknown model parameter. We assume that the calculation of h is based on subsurface models, which depend on a number of unknown parameters, $\theta \in R^{N_\theta}$. We further assume that the models are constrained by a set of, typically dynamic, data, $d \in R^{N_d}$. All quantities are assumed to be random variables. We will for simplicity use small letters both for the random variables and their realizations or variates.

The Bayesian average of multiple models is the weighted average of the model-wise posterior predictive distributions, $p(h|d, M_k)$, i.e.,

$$p\,(h|d) = \sum_k p\,(h|d, M_k)\, w\,(M_k|d). \tag{1}$$

$w\,(M_k|d)$ is the posterior model weight given the data, and the estimation of this quantity is the main focus of this paper. In the following we will denote this by just w_k.

2.1 Traditional Bayesian model average and selection, BMA/BMS

As mentioned in the introduction, BMA and BMS are based on a discrete version of Bayes rule, i.e., an assumption that one of the models is the true one—an $\mathcal{M}$-closed setting. The model weight can then be interpreted as the posterior probability for model k being the true model (the BMP) and is given by,

$$w_k = p\,(M_k|d) = \frac{p(d|M_k)p(M_k)}{\sum_l p(d|M_l)p(M_l)} = \frac{1}{\sum_l \frac{p(d|M_l)p(M_l)}{p(d|M_k)p(M_k)}}, \tag{2}$$

where,

$$p\,(d|M_k) = \int p(d|\theta_k, M_k)\,p(\theta_k|M_k)\,\mathrm{d}\theta_k \tag{3}$$

is the marginal distribution or *prior predictive distribution* for d under model M_k, also called model likelihood or model evidence. The ratio of two model likelihoods, $p(d|M_l)/p(d|M_k)$, is commonly called the Bayes factor for model M_l compared to model M_k [8].

It is well known that the calculation of BMP is very sensitive and unstable with respect to e.g., small uncertainties in the measurements, especially in high (data) dimensions. It was demonstrated in Aanonsen et al. [10] how the BMP may quickly approach 0 or 1 when the number of measurements increases, and it can be shown that BMA will asymptotically select the model in the list that is closest in Kullback-Leibler (KL) divergence [5]. However, if the BMP is considered as a function of the random variable, d, a sample of the BMP may be obtained by calculating it for an ensemble of N_e data realizations generated from the data PDF, and a more stable estimate of the posterior probabilities may be obtained by taking the mean of this sample. One model may then receive 100% probability for each $d^{(j)}$, but which model that receives 100% probability will typically vary when j varies. Normally, none of the models will therefore receive 100% probability when averaging over the ensemble. An example of this is shown in Fig. 9, right plot. BMP for model k is then calculated as,

$$w_k = \frac{1}{N_e}\sum_{j=1}^{N_e} p\left(M_k|d^{(j)}\right), \tag{4}$$

where $d^{(j)}$ is a realization of the full data set obtained by adding a realization from the measurement error distribution to the actual measurements. That is, assuming measurement errors being normally distributed with zero mean,

$$d^{(j)} = d + e^{(j)}, \text{ where } e^{(j)} \in R^{N_d} \text{ is a realization from } \mathcal{N}(0, C_d). \tag{5}$$

A similar approach was used by e.g., Hong et al. [9] to integrate model uncertainty in a probabilistic decline curve analysis. It is also related to the modified bootstrap method used by Cheng et al. [11] for probabilistic estimation of oil reserves from production data. However, if the measurement realizations are generated by adding uncorrelated error realizations, variations between the individual measurements will typically cancel out, and one model typically becomes the one closest to the data (measured by the Mahalanobis distance) for all the realizations. Using correlated error realizations, this effect is avoided, and a much more stable BMP is obtained.

Thus, instead of generating measurement realizations using Eq. 5, we will use 100% correlated perturbations, i.e., define $d_i^{(j)}$ by,

$$d_i^{(j)} = d_i + \sigma_i e^{(j)}, \quad i = 1, ..., N_d, \quad j = 1, ..., N_e \tag{6}$$

where σ_i is the standard deviation of measurement i, and $e^{(j)}$ is a realization from the one-dimensional standard, normal distribution, $\mathcal{N}(0, 1)$. The difference between using Eqs. 5 and 6 will be discussed further in the examples section below.

2.2 Bayesian stacking

Yao et al. [5] defines the model weights, $w = \{w_k\}$, for stacking of predictive distributions by:

$$w^* = \arg\max_w \frac{1}{N_d}\sum_{i=1}^{N_d} \log \sum_k w_k\, p\left(d_i|d_{-i}, M_k\right), \quad w_k \geq 0, \quad \sum_k w_k = 1, \tag{7}$$

where $p\,(d_i|d_{-i}, M_k)$ is the posterior predictive distribution of d_i conditioned to all the other data, d_{-i}, given by

$$p\,(d_i|d_{-i}, M_k) = \int p(d_i|\theta_k, M_k)\,p(\theta_k|d_{-i}, M_k)\,\mathrm{d}\theta_k. \tag{8}$$

This approach (BSLOO) is based on LOOCV which in principle requires data to be iid (see e.g., [8] p. 176). We suggest using a modified version BS along the lines of the

modified version of BMA described above and define the weights by

$$w^* = \arg\max_w \frac{1}{N_e}\sum_{j=1}^{N_e}\log\sum_k w_k\, p\left(d^{(j)}|d, M_k\right), \quad w_k \in [0,1], \quad \sum_k w_k = 1, \tag{9}$$

where,

$$p\left(d^{(j)}|d, M_k\right) = \int p(d^{(j)}|\theta_k, M_k)\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k. \tag{10}$$

In BSLOO, the full predictive distribution, $p\left(\tilde{d}|d, M_k\right)$, evaluated at a new dataset $\tilde{d}$, is replaced with the corresponding LOO predictive distribution. The final equation for the stacking is derived by using a logarithmic score Eq. 7. In our modified approach we instead evaluate the full predictive distribution by drawing a new random dataset. These new datasets can be considered replications, and this method is described in Chap. 6.3 of [8]. Eq. 10 then corresponds to Eq. (6.1) in [8]. As a summarizing statistic we have selected to calculate the mean of the replications.

A natural choice for $d^{(j)}$ would be to use measurement realizations as given by Eq. 5. However, the issue discussed in the previous section will also apply here, and thus we will use 100% correlated perturbations as defined in Eq. 6 also for the modified stacking.

We will also consider an alternative expression based on prior predictive distributions:

$$w^* = \arg\max_w \frac{1}{N_e}\sum_{j=1}^{N_e}\log\sum_k w_k\, p\left(d^{(j)}|M_k\right), \quad w_k \in [0,1], \quad \sum_k w_k = 1. \tag{11}$$

This approach will be directly comparable with MBMA. The two alternative formulations of modified stacking will be denoted as MBS-post and MBS-pri, respectively.

2.3 Computational issues

All the alternative methods for calculating model weight are based on predictive distributions, i.e., integrals of the same type as Eq. 3. It is well known that these are challenging to estimate in the general case, and especially in high dimensions. Assuming that the measurements represent a particular realization of the random vector, d, the integral represents the value of the marginal distribution of d evaluated at this particular realization, and it is evident that this will be very sensitive to uncertainties in this distribution. The various stacking methods, as well as Eq. 4 aims at stabilizing this calculation by averaging over data realizations or the individual measurements. The goal of this paper is to evaluate the success of the various methods to produce stable, reliable and "reasonable" estimates for model weights without having to consider uncertainties in the PDF's and the integrals. Thus, in the examples, we use only linear, Gaussian models where the predictive distributions are Gaussian and can be calculated exactly using the formulas in Appendix A. For large amounts of data, inversion of the relevant covariance matrices may also be a challenge. Here, we do not consider this issue, and in the examples exact inversion has been performed for all matrices.

With respect to the computational time, the methods based on (full) measurement realizations requires that the predictive distributions are calculated in N_d-space for each data realization. In addition, MBS-post requires one Bayesian inversion given the full dataset. BSLOO in its basic form requires N_d Bayesian inversions, which will not be feasible for most real-life problems. Yao et al. [5] suggest importance sampling to calculate the integral in Eq. 8 using the posterior distribution as the importance sampler. The required posterior predictive distribution is then approximated by,

$$p\left(d_i|d_{-i}, M_k\right) \approx \left(\frac{1}{N_s}\sum_s\left(\frac{1}{p(d_i|\theta_k^s, M_k)}\right)\right)^{-1}, \tag{12}$$

where θ_k^s are simulation draws from the full posterior $p(\theta_k|d, M_k)$. It is seen that the distribution is given by the harmonic averages of likelihoods, a calculation which may potentially be unstable due to very small tail densities. Yao et al. suggest resolving this using a Pareto smoothing [12, 13]. A derivation of Eq. 12 is presented in Appendix B, and we notice that this requires iid data.

An alternative approach for calculating Eq. 8 also utilizing the iid assumption is used by Höge et al. [7]:

$$\begin{aligned} p\left(d_i|d_{-i}, M_k\right) &= \int p(d_i|\theta_k, M_k)\, p(\theta_k|d_{-i}, M_k)\, \mathrm{d}\theta_k \\ &= \int p(d_i|\theta_k, M_k)\frac{p(d_{-i}|\theta_k, M_k)\, p(\theta_k|M_k)}{p(d_{-i}|M_k)}\, \mathrm{d}\theta_k \\ &\overset{\text{iid}}{=} \frac{1}{p(d_{-i}|M_k)}\int p(d_i, d_{-i}|\theta_k, M_k)\, p(\theta_k|M_k)\, \mathrm{d}\theta_k \\ &= \frac{p(d|M_k)}{p(d_{-i}|M_k)}. \end{aligned} \tag{13}$$

That is, the prior predictive distribution based on all data divided by the prior predictive distribution based on the remaining data. With this approach, inversion is not necessary, but the prior predictive distribution must be calculated in $(N_d - 1)$-space for each measurement.

To summarize, neither the basic version on BSLOO Eq. 7, nor the Höge et al. approach Eq. 13 will be feasible for cases with large amounts of data, like 4D seismic. The approximate version of BSLOO Eq. 12, requires just one Bayesian inversion and N_d likelihood calculations in 1D. However, the amount of additional effort and accuracy involved when applying the Pareto smoothing suggested by Yao et al. [5] in the general, high-dimensional case is not quite clear and needs to be evaluated. In our BSLOO calculations, we have used a subset of the full dataset, d, and performed an analytic Bayesian inversion for each measurement in the subset. Notice that the prior predictive distributions Eq. 8 used in BSLOO will be univariate versions of Eq. A.2, and consequently, the non-diagonal elements of the covariance matrix, C_k are not used. However, the full matrices will be used in the Bayesian inversions when conditioning on d_{-i}. On the other hand, the predictive distributions to be calculated for the methods based on measurement realizations (Eqs. 4,9,11) are PDF's of the full, possibly very high-dimensional random vector d. Although challenging, these calculations should be feasible as long as the number of measurement realizations required is limited. For spatially distributed data, efficient calculation of the predictive distributions for larger and more realistic, non-linear models may be performed using multilevel techniques as shown in a separate paper [14]. However, the methods based on high-dimensional PDF's may be more exposed to the so-called curse of dimensionality, since the computations typically will involve distances between vectors in high dimensions.

3 Examples

In this section we will compare model weights based on the various methods described above using three synthetic, linear examples. The first example is an $\mathcal{M}$-open problem based on a Gaussian mixture model taken from Yao et al. [5]. The other two are inspired by the interpretation of data from repeated (4D) seismic surveys. The first of these is a type of Quasi-$\mathcal{M}$-closed setting, with two distinct scenarios which both could explain the data: pressure depletion or water flooding. The true data is based on one of these scenarios, but the corresponding model is an approximation to the true data-generating model. The second is an $\mathcal{M}$-open problem inspired by the case considered by Aanonsen et al. [10]. In [10], the probability of alternative seismic interpretations of the top reservoir surface was estimated from 4D seismic measurements of gas-cap thickness. Here, we simplify this by assuming that the unknown surface is observed directly with uncertainty. This is an example of a problem which could have been expanded to a hierarchical problem, defining e.g., the mean of the unknown surface as a hyperparameter. For the last example, we also repeat the weight calculations using several data realizations to get some information about the uncertainty in the estimated stacking weights.

In example 2 and 3 θ and d are defined on spatial grids. The parameter and measurement grids may be different. The measurements are given on a 2D, regular grid with $50 \times 50 = 2500$ cells. This is small enough to allow for exact calculations, while still being large enough to illustrate the issues related to the calculation of Bayesian model weights for large amounts of data. Thus, the number of measurements $N_d = N_x \times N_y = 2500$. Parameters and measurements are Gaussian, and the required predictive distributions can be calculated by the formulas given in Appendix A using the appropriate expected values and covariances.

In the discussion of example 2 and 3 the *expected value* of a quantity will denote the vector of expected values defined on the corresponding grid, while we will use the term *mean value* to denote the average of a quantity over the grid cells. Thus, the expected value will be a vector with dimension equal to the number of grid cells, while the mean will be a single, scalar value.

3.1 Example 1

In the first example we assume N_d independent observations of a single quantity coming from a normal distribution $\mathcal{N}(3.4, 1)$, not known to the data analyst. That is, $d = (d_i, i = 1, ..., N_d)$. We assume 8 candidate models, $\mathcal{N}(\mu_k, 1)$, with $\mu_k = k$ for $1 \leq k \leq 8$, each with equal prior model probability, $P(M_k) = 1/8$. This is then an $\mathcal{M}$-open problem where none of the candidate models is the true model.

To be consistent with the formulation in Section 2, we define the 8 models as follows:

$$d_k = G\theta_k + \epsilon, \qquad k = 1, ..., 8, \tag{14}$$

where,

$$G = \{1, ..., 1\}^T \in R^{N_d \times 1}, \quad \text{and} \quad \epsilon \sim \mathcal{N}(0, 1), \tag{15}$$

and

$$p(\theta_k|M_k) = \delta(\theta_k - \mu_k), \quad \mu_k = k. \tag{16}$$

The prior predictive distribution for the data (BME) becomes:

$$\begin{aligned} p(d|M_k) &= \int p(d|\theta_k, M_k)\, p(\theta_k|M_k)\, \mathrm{d}\theta_k \\ &= p(d|\mu_k) \propto \exp\{-\frac{1}{2}\sum_{i=1}^{N_d}(d_i - \mu_k)^2\}. \end{aligned} \tag{17}$$

Furthermore, since there are no parameters to estimate in this case, the posterior predictive distribution for an

unobserved quantity, $\tilde{d}$, becomes identical to the prior predictive distribution, i.e., MBS-pri and MBS-post coincide:

$$\begin{aligned} p(\tilde{d}|d, M_k) &= \int p(\tilde{d}|\theta_k, M_k)\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k \\ &= \int p(\tilde{d}|\theta_k, M_k) \frac{p(d|\theta_k, M_k)\, p(\theta_k|M_k)}{p(d|M_k)}\, \mathrm{d}\theta_k \\ &= p(\tilde{d}|\mu_k) \frac{p(d|\mu_k)}{p(d|M_k)} \\ &= p(\tilde{d}|\mu_k). \end{aligned} \tag{18}$$

In the following we derive the expressions for posterior model weights for BMA, MBMA, BSLOO and MBS.

$$w_{k\text{-BMA}} = p(M_k|d) = \frac{p(d|M_k)}{\sum_l p(d|M_l)} = \frac{\exp\{-\frac{1}{2}\sum_{i=1}^{N_d}(d_i-\mu_k)^2\}}{\sum_{l=1}^{8}\exp\{-\frac{1}{2}\sum_{i=1}^{N_d}(d_i-\mu_l)^2\}}, \tag{19}$$

$$w_{k\text{-MBMA}} = \frac{1}{N_e}\sum_{j=1}^{N_e} p(M_k|d^{(j)}), \tag{20}$$

$$w_{k\text{-BSLOO}} = \arg\max_{w} \frac{1}{N_d}\sum_{i=1}^{N_d} \log \sum_{k=1}^{8} w_k\, p\,(d_i|d_{-i}, M_k)\,, \quad w_k \in [0,1]\,, \quad \sum_k w_k = 1, \tag{21}$$

where now,

$$p\,(d_i|d_{-i}, M_k) = p(d_i, M_k) = \frac{1}{\sqrt{2\pi}}\exp\{-\frac{1}{2}(d_i-\mu_k)^2\}, \tag{22}$$

and finally,

$$w_{k\text{-MBS}} = \arg\max_{w} \frac{1}{N_e}\sum_{j=1}^{N_e} \log \sum_{k} w_k\, p\left(d^{(j)}|d, M_k\right), \quad w_k \in [0,1]\,, \quad \sum_k w_k = 1, \tag{23}$$

where,

$$p\left(d^{(j)}|d, M_k\right) = p\left(d^{(j)}|M_k\right) = \frac{1}{(2\pi)^{N_d/2}}\exp\{-\frac{1}{2}\sum_{i=1}^{N_d}\left(d^{(j)}-\mu_k\right)^2\}. \tag{24}$$

The expressions for BMA and BSLOO are now the same as given in [5].

Figure 1 shows the posterior predictive distribution $p(\tilde{d}|d) = \sum_k w_k\, p(\tilde{d}|d, M_k)$ for the 4 methods from one simulation with sample size varying from 3 to 200 compared to the data distribution. The distributions jumps back and forth somewhat following the average of the measurements, which varies a lot for small sample sizes. For large sample sizes, the mean approaches 3.4 for all methods, except BMA. The behavior of the methods is also illustrated in Fig. 2 showing the mean and standard deviation of $p(\tilde{d}|d)$ vs sample size. Here, the data mean and standard deviation are calculated from the actual samples. For most of the N_d-values, the results with MBMA and MBS are almost identical. Except for small sample sizes, the mean of MBMA, BSLOO and MBS follow the variations in data mean, while BMA picks model 3 if data mean is closer to 3.0, or model 4 if data mean is closer to 4.0. Standard deviation for BMA and BSLOO is close to data standard deviation, while the predictive distributions estimated by MBMA and MBS have a larger spread than the data. The reason is that MBMA and MBS are based on the full ensemble, $d^{(j)}$, and not just the data vector d.

The estimated model weights are plotted versus sample size in Fig. 3. Again, we see that BMA picks model 3 or model 4 with 100% probability more and more often as the sample size increases. BSLOO picks model 3 and 4 with a slightly larger weight for model 3 for most of the sample sizes. The weights estimated with MBMA and MBS are highest for model 3 and 4. However, also model 2 and 5 get significant weights explaining the larger variances seen in Figs. 1 and 2. Thus, if the objective is to retain all possible models with some probability, MBMA and MBS may be better than BSLOO. Notice also that the spread in weights is considerably larger for BSLOO than for MBS.

In this case the results with MBS are almost the same as with MBMA. This will normally not be the case for a parameter estimation problem, where the posterior predictive distributions are based on the posterior parameter distributions. Figure 4 shows model weights estimated with MBMA and MBS for the same example, but extended to a parameter estimation problem by defining the prior densities as,

$$p\,(\theta_k|M_k) = \mathcal{N}\,(\mu_k, 1)\,. \tag{25}$$

With the given prior and measurement uncertainties, the expected posterior parameter values are almost equal for all models. Thus, the estimated MBS weights now oscillate around 0.125 corresponding to almost equal weight for all the models, while the results with MBMA are very similar to those without parameter estimation.

3.2 Example 2

3.2.1 Example specification

In this example we consider the problem of discriminating between pressure and saturation response, which is a common problem within 4D seismic analysis. We assume that seismic data is available in a limited area of an oil reservoir

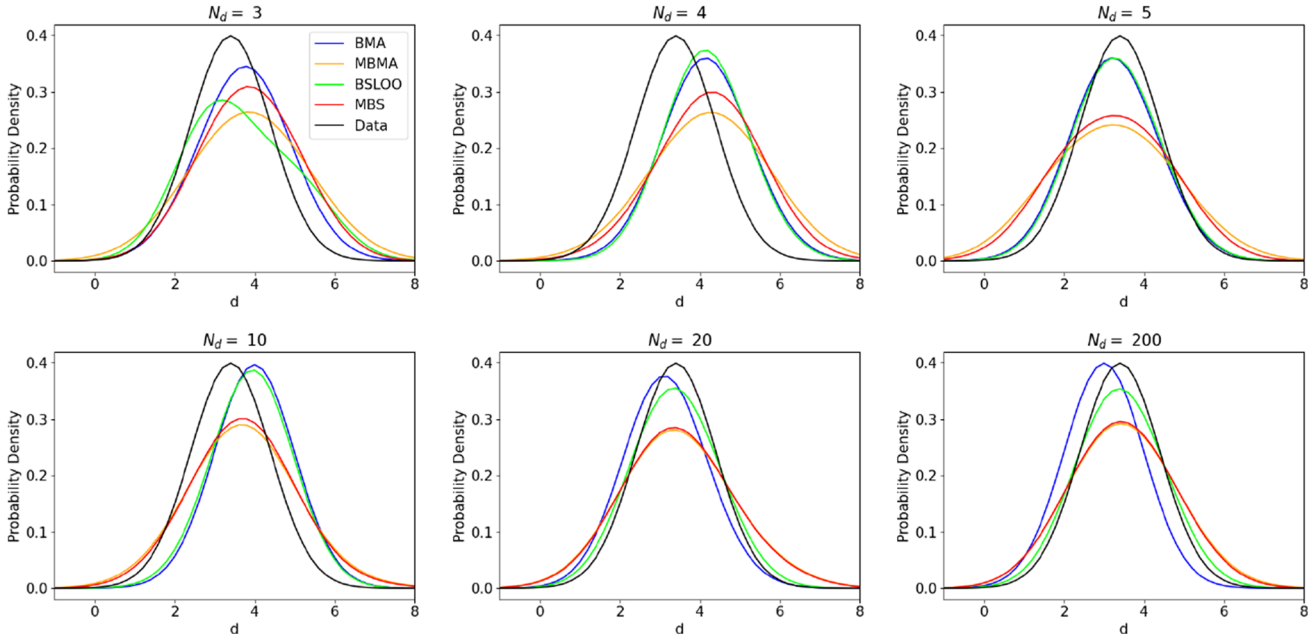

Fig. 1 Predictive distributions for selected numbers of measurements

and consider two scenarios: 1) the area is water flooded to residual oil saturation while retaining the initial reservoir pressure, and 2) the area is a part of a segment which has been pressure depleted without changing the oil saturation. We further assume that the seismic data is given as change in acoustic impedance induced by the water flooding or depletion, respectively. This difference in acoustic impedance will be denoted DAI. The data is thus DAI given on a 2D, horizontal grid lying inside a larger reservoir. Measurement errors are assumed to be additive, Gaussian with zero mean, and can be correlated or uncorrelated. Scenario 1 (water flooding) is assumed to be the true scenario in all cases, and the data is generated using a quite standard rock physics model calculating acoustic impedance as a function

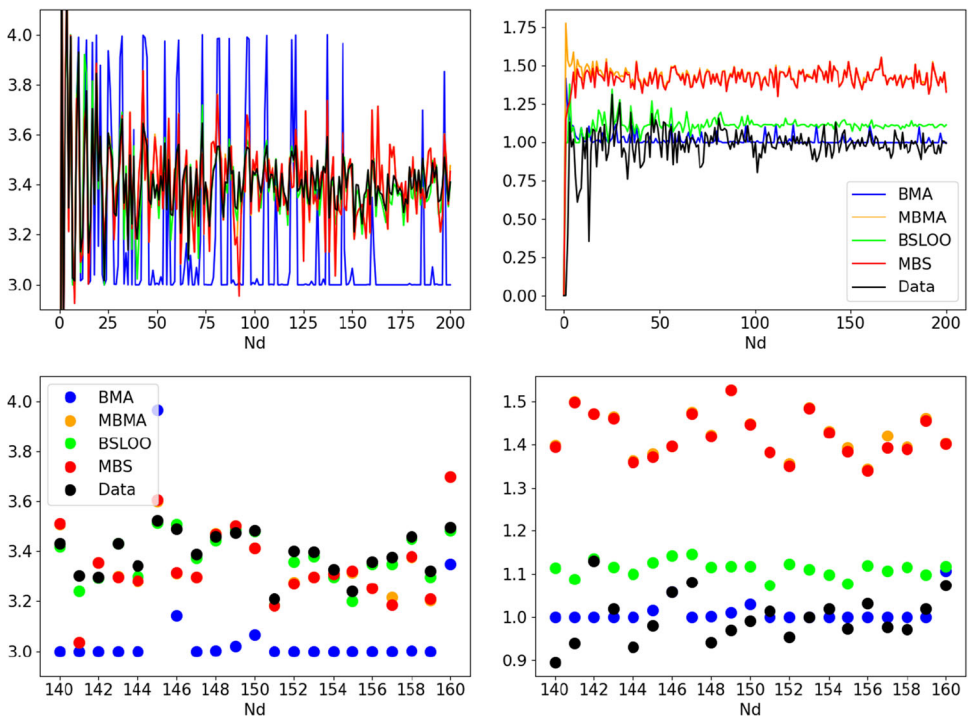

Fig. 2 Mean (left) and standard deviation (right) of the predictive distributions vs number of measurements. To further emphasize the behavior, the bottom 2 plots show the results within a limited N_d interval

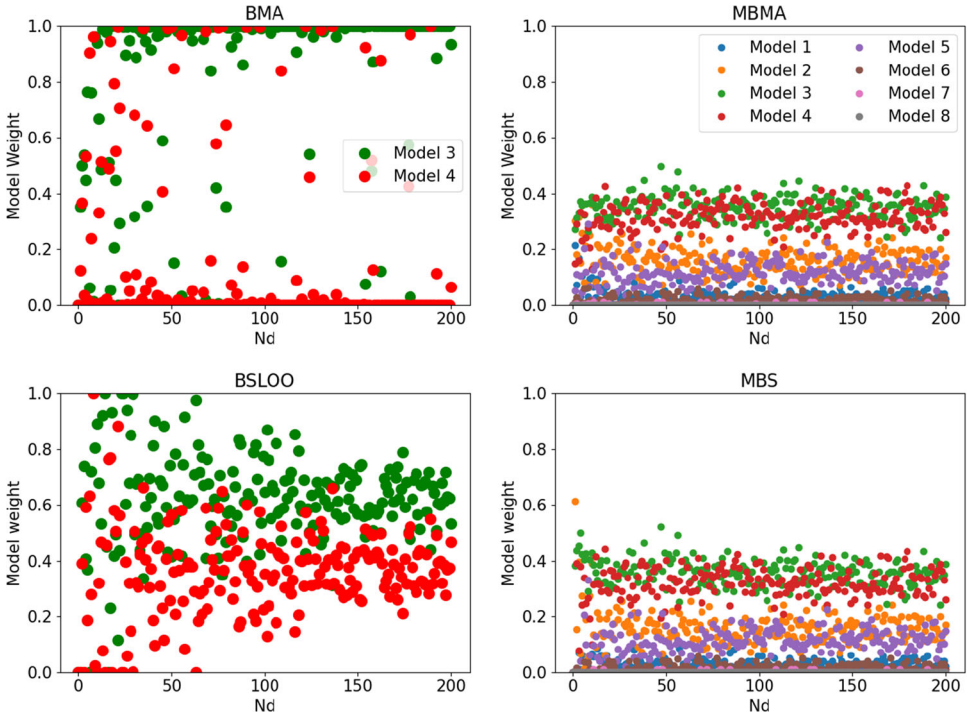

Fig. 3 Model weights vs number of measurements. For BMA and BSLOO only weights for model 3 and model 4 are shown, since the weights for the other models are mostly equal to zero

of pressure, saturation and porosity. For details on the rock physics model, see [15]. The synthetic, true pressures and saturations are assumed constant over the 4D region, while the porosity is assumed to be heterogeneous and vary between cells. The porosity field is taken as a realization of a Gaussian random field with long correlation length, and a reference acoustic impedance difference, $\widetilde{\text{DAI}}$, is calculated from this porosity field and the homogeneous pressures and saturations. Notice that because of nonlinearities in the rock physics model, $\widetilde{\text{DAI}}$ will vary between cells even if the same porosity field is used for the initial and final calculations. Three synthetic datasets are generated from $\widetilde{\text{DAI}}$. In Case 1, the data is equal to $\widetilde{\text{DAI}}$, while in Case 2 and 3 the data is generated by adding a realization, $\widetilde{E}$, from a Gaussian random field with zero mean and covariance matrix $C_{\widetilde{E}}$. $C_{\widetilde{E}}$ is defined by a constant standard deviation, $\sigma_{\widetilde{E}} = 0.05$ MPa-s/m and a spherical variogram with range $R_{\widetilde{E}}$. $R_{\widetilde{E}} = 1$ cell in Case 2 and 0.75 times the length of the area (i.e., 37.5 cells) in Case 3. The measurement error distribution is defined by a measurement error standard deviation, σ_d, and a spherical variogram model with range R_d. In Case 2 the measurement error covariance matrix, C_d, is diagonal ($R_d = 1$ cell). In Case 3 it is non-diagonal with $R_d = 37.5$ cells. For Case 1 C_d is either diagonal (Case 1a) or nondiagonal (Case 1b). All parameters used to generate the true models are listed in Table 1. The 3 different sets of measurements are shown in Fig. 5. Distributions of the measurements over the 2500 grid cells are shown in Fig. 6. Here, we have also plotted the distributions of 400 realizations of the data obtained by adding realizations from the error distributions. It is seen that in Case 1a and 1b the spread in DAI over the grid cells is much less than the spread over all measurement realizations, while in Case 2 and 3, the spread is similar. Thus, it could be expected that BSLOO, which utilizes the spread in DAI over grid cells,

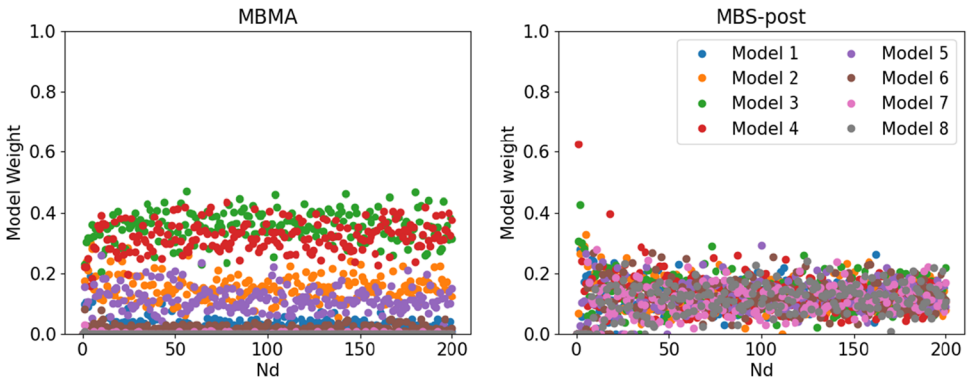

Fig. 4 MBMA and MBS model weights vs number of measurements. With parameter estimation

Table 1 Example 2. Parameter settings for the true model

Reference model, $\widetilde{\text{DAI}}$:	
Reservoir pressure (MPa)	24 .0
Initial oil saturation, S_{oi}	0.8
Residual oil saturation, S_{or}	0.3
Porosity, ϕ	Realization from $\mathcal{N}(\overline{\phi}, C_\phi)$
$\overline{\phi}$	0.3
σ_ϕ	0.02
C_ϕ	Spherical variogram; range 37.5 $(0.75N_x)$ cells
Measurements:	
d = DAI	Case 1: DAI = $\widetilde{\text{DAI}}$
	Case 2 and 3: DAI = $\widetilde{\text{DAI}} + \widetilde{E}$,
	$\widetilde{E}$ being one realization from $\mathcal{N}(0, C_{\widetilde{E}})$
$\sigma_{\widetilde{E}}$ (MPa-s/m)	0.05
$C_{\widetilde{E}}$	Case 2: Diagonal.
	Case 3: Spherical variogram; range 37.5 $(0.75N_x)$ cells
Measurement error:	$\mathcal{N}(0, C_d)$
σ_d (MPa-s/m)	0.05
C_d	Case 1a and 2: Diagonal.
	Case 1b and 3: Spherical variogram; range 37.5 $(0.75N_x)$ cells

will behave similarly as MBS-post in Case 2 and 3, but not necessarily in Case 1a and 1b.

In both the two alternative model scenarios the porosity is assumed to be constant. To speed up the calculations, linear approximations are applied to the rock physics model, and the example then also reflects that models may not be exact. The linear approximations are relatively accurate in the relevant parameter intervals as shown in Fig. 7. In the models the residual oil saturation, S_{or}, may be constant over the 4D area, or vary between the cells due to e.g., a heterogeneous permeability. Thus, model 1 has either 1 parameter or 2500 parameters. Constant reservoir pressure would normally be assumed over the 4D area, i.e., model 2 has one unknown parameter, the reservoir pressure after depletion, p. However, we will in this investigation also allow the pressure to vary between grid cells. In the cases where a parameter is equal in several grid cells, the predictions will be correlated between these cells. For instance, if the number of parameters for model k, $N_{\theta k} = 1, G_k = \{1, 1, ..., 1\}^T$ (cf. Appendix A). In this case, the covariance of the predicted data will be proportional to a matrix with all entries equal to 1. Input to the prior models is listed in Table 2.

The main objective of this example is to evaluate the sensitivity to prior model assumptions, which are known to be large for BMA. Thus, we will estimate model weights for varying parameter prior mean and variance for Scenario 1, while keeping everything constant for Scenario 2. Figure 8 shows the distribution of prior predictions for Scenario 1 and

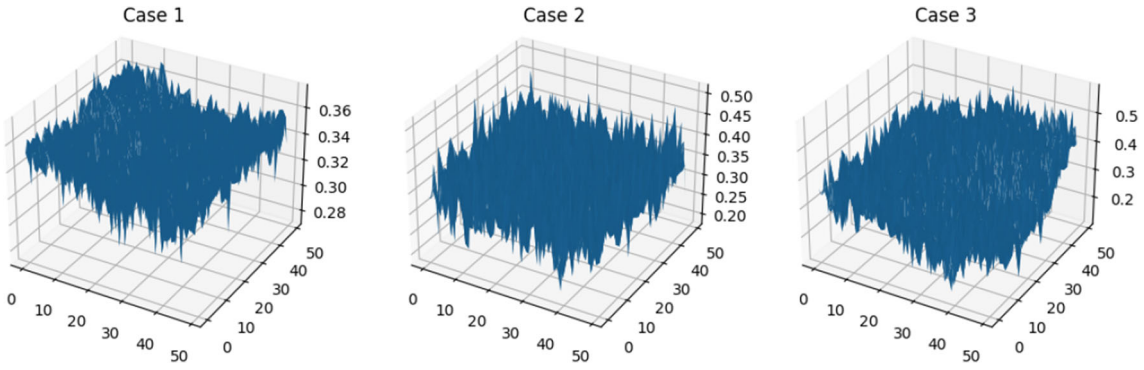

Fig. 5 Synthetic measurements, DAI (Mpa-s/m)

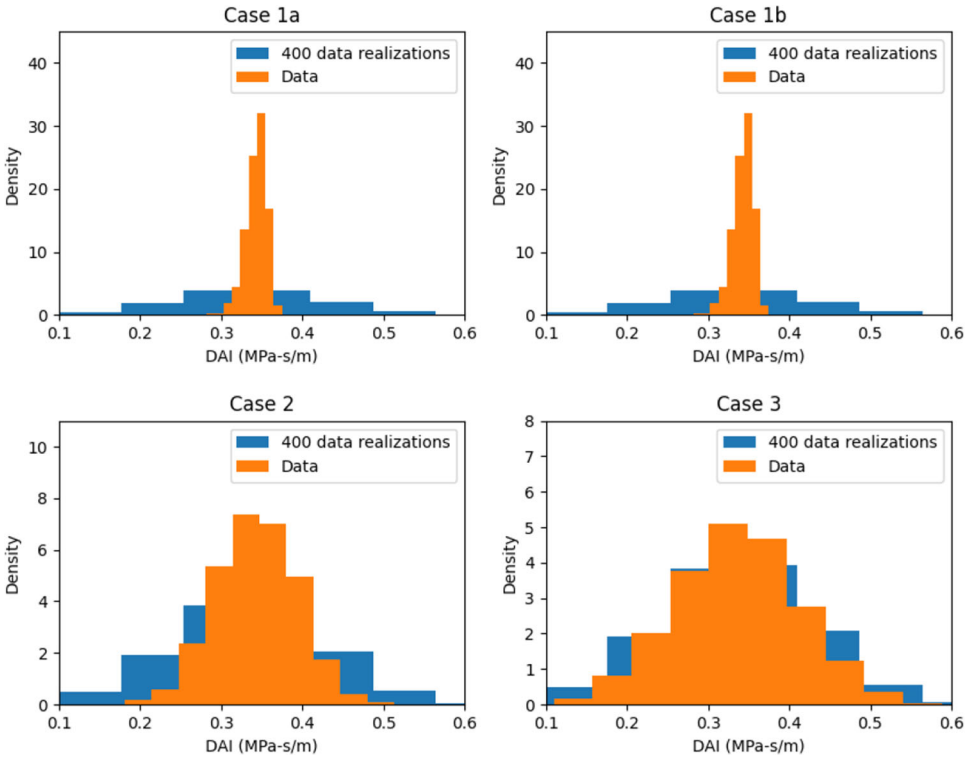

Fig. 6 Measurement (data) distribution over the 2500 grid cells, and distribution of the measurements plus 400 realizations from the error distributions

Scenario 2 compared to Case 2 data distributions for prior mean, $S_{or\text{-}pri} = 0.2$ and 0.3. Prior mean pressure for Scenario 2 is equal to 21.9 MPa in both cases. As can be seen, the prior model predictions are almost overlapping with the data distributions when prior mean $S_{or\text{-}pri} = 0.3$, and both scenarios would then be expected to get almost the same weight. When $S_{or\text{-}pri} = 0.2$, there is still a significant overlap between the data distribution and the distribution of prior predictions for Scenario 1, so although Scenario 2 should get a higher weight, we would normally not want to discard Scenario 1 completely for predictions. Thus, a good method should produce a non-zero weight for Scenario 1 also in this case. The plot shows distributions for $N_{\theta 1} = N_{\theta 2} = 1$. However, the behavior is similar in the other cases.

3.2.2 Evaluation of method performance

In this section we first demonstrate how the traditional BMA may be improved by averaging BMP over many data realizations. Then we illustrate how the original Bayesian Stacking introduced by Yao et al. [5] may fail for correlated measurements. Finally, we compare the sensitivity with respect to prior assumptions for the different Bayesian model combination methods presented in Section 2. There will, of course, always be some dependency on prior probabilities, and the strength of that dependency will vary from method to method. Our aim is just to illustrate how the methods we have considered behave in this respect.

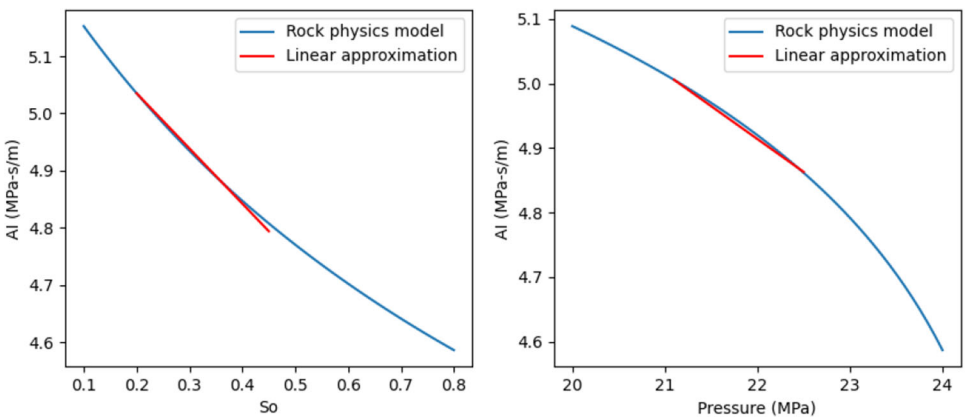

Fig. 7 Rock physics model with linear approximations plotted in the relevant intervals. Acoustic Impedance (AI) vs pressure and saturation for porosity = 0.30

Table 2 Example 2. Parameter setting for prior model scenarios

Model 1 (flooded):		
Reservoir pressure (MPa)	24.0	
Initial oil saturation, S_{oi}	0.8	
Residual oil saturation, S_{or}	$\mathcal{N}(S_{or\text{-}pri}, C_{pri\text{-}1})$	
$S_{or\text{-}pri}$	Varying	
$\sigma_{pri\text{-}1}$	0.03	
$C_{pri\text{-}1}$	$\{\sigma_{pri\text{-}1}{}^2\}$,	$N_{\theta 1} = 1$
	$diag\{\sigma_{pri\text{-}1}{}^2\}$,	$N_{\theta 1} > 1$
Porosity, ϕ	0.3 (constant)	
DAI (MPa-s/m)	$0.4487 - 0.964(S_{or} - 0.2)$	
Model 2 (depleted):		
Initial reservoir pressure (MPa)	24 .0	
Final reservoir pressure	$\mathcal{N}(p_{pri}, C_{pri\text{-}2})$	
p_{pri} (MPa)	21.9	
$\sigma_{pri\text{-}2}$ (MPa)	0.3	
$C_{pri\text{-}2}$	$\{\sigma_{pri\text{-}2}{}^2\}$,	$N_{\theta 2} = 1$
	$diag\{\sigma_{pri\text{-}2}{}^2\}$,	$N_{\theta 2} > 1$
Oil saturation, S_{oi}	0.8	
Porosity, ϕ	0.3 (constant)	
DAI (MPa-s/m)	$0.4192 - 0.102(p - 21.1)$	

Figure 9 shows two examples of BMP calculated from Eqs. 4 and 6 for increasing values of N_e compared to the BMP calculated from the individual data realizations (i.e., Eq. 2). The plot illustrates the behavior seen in all cases tested: although the individual calculations are unstable and very sensitive to the measurements, a relatively small number of measurement realizations is required to get a stable value for the average. Notice that although C_d is diagonal, C_k in Eq. A.3 is non-diagonal because of the correlations in the prior predictions when $N_\theta = 1$.

BMP calculations using Eq. 4 and the MBS , Eq. 9 or Eq. 11, require that the predictive distributions are calculated for N_e data realizations. In Fig. 10 we have plotted estimated weight of model 2 versus N_e for a typical example. The plot also shows the weight estimated with BSLOO using subsets of the full dataset. Here N_d^* denotes the number of datapoints used, and every 2nd datapoint is used if $N_d^* = N_d/2$, etc. Using only a subset of the data will greatly speed up the calculations with BSLOO as defined in Eqs. 7 and 8, since a Bayesian inversion is required for each individual measurement. However, it will still be much slower than MBS. We never ran BSLOO with all 2500 measurements. It would have taken an impractically long time, and the results seemed to converge with less measurements in all cases tested. It should be noted that, when $N_\theta = 1$, the data realizations (measurements for BSLOO) need to be split between the models for all the methods. This is because the parameters, and thus the predicted data, are the same in all cells, and the predictive

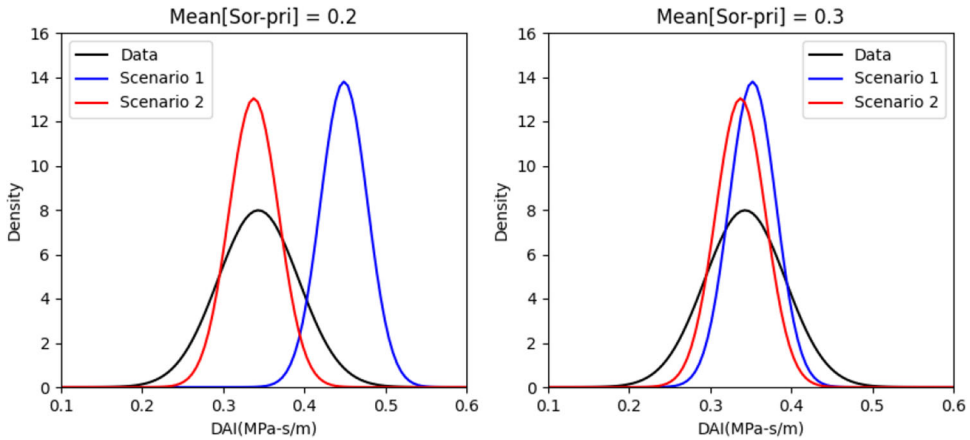

Fig. 8 Distribution of prior predictions for Scenario 1 and 2 compared to data distributions. $N_{\theta 1} = N_{\theta 2} = 1$. Case 2 data

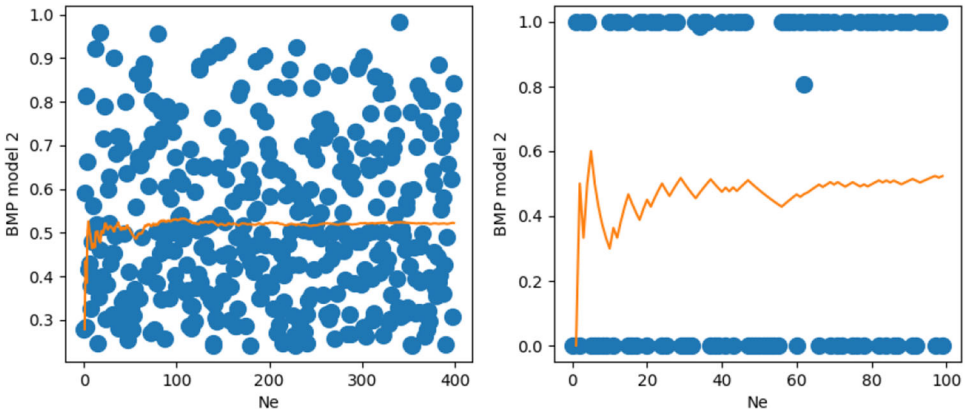

Fig. 9 Stabilizing effect of BMP averaging. Individual realizations (blue dots) and cumulative average over realizations (orange). Data Case 2. $N_{\theta 1} = N_{\theta 2} = 1$. Left: Based on full matrix Eq. A.3. Right: Based on diagonal approximation to Eq. A.3

distributions will be 100% correlated if the same data realizations, or datapoints for BSLOO, are used for both models. Thus, when e.g., N_e(or N_d^*) $= 1000$, 500 data realizations (or measurements) are used to estimate weights for each of the two models. We will in the following use 400 realizations for BMA and MBS, and 500 datapoints for BSLOO to ensure that the results of the method evaluations are not influenced by the number of data realizations or datapoints used.

The effect of BMP-averaging on prior model sensitivity is illustrated in Fig. 11. Here we plot BMP for model 2 vs Mean$[S_{or\text{-}pri}]$ for 3 different values of $\sigma_{pri\text{-}1}$. The traditional BMP-calculation yields either zero or one and is very sensitive to the prior assumptions. The results using Eqs. 4 and 5 are slightly better, but still not satisfactory. The results using Eqs. 4 and 6, however, looks much better considering that a BMP around 0.5 is to be expected for Mean$[S_{or\text{-}pri}]$ around 0.3. The sensitivity to prior model assumptions is also much lower.

Consider again Fig. 5. Although the data fields look quite similar in the 3 cases, the data in Case 1 will be highly correlated. The data in Case 2 are truly iid, while the data in Case 3 are Gaussian with a relatively long correlation length. For all the tests performed, BSLOO performs well for the cases 1a, 2 and 3. However, it fails for Case 1b data. While the other methods (MBMA, MBS-pri and MBS-post) give results which are consistent with the model parameters used, the weights calculated with BSLOO then typically become either 0 or 1. Remember that the measurements are the same in Case 1a and 1b. However, C_d is diagonal in Case 1a and non-diagonal in Case 1b. In general we have experienced that BSLOO is very stable also for correlated data when a realization from the error distribution is added to the data before doing the analysis, while it is less stable if not.

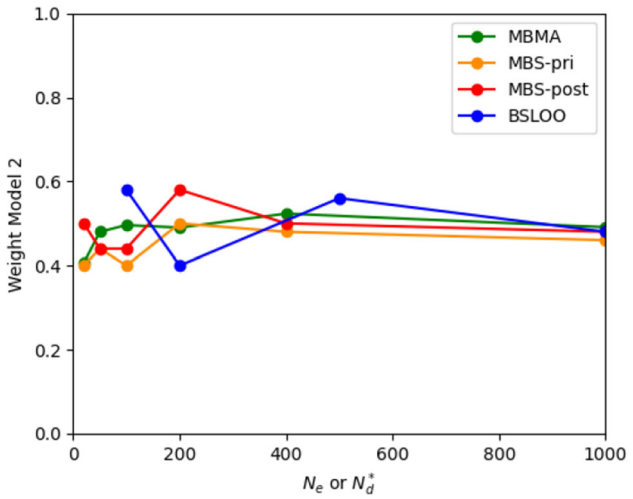

Fig. 10 Estimated weight model 2 vs N_e or, for BSLOO, N_d^*. Data Case 2. $N_{\theta 1} = N_{\theta 2} = 2500$

In the following we will consider only Case 2 and 3.

To investigate the performance of the methods, we have estimated model weights while varying the parameter prior mean and standard deviation of model 1, i.e., Mean$[S_{or\text{-}pri}]$ and $\sigma_{pri\text{-}1}$; with and without correlated data, and with a varying number of parameters and measurement uncertainty. The measurement error realizations, $d^{(j)}$, required for MBMA Eq. 4 and MBS Eqs. 9,11 are generated by adding 100% correlated error realizations as specified in Eq. 6.

In the following, the main results are summarized and illustrated with some selected plots of w_2 vs Mean$[S_{or\text{-}pri}]$ and $\sigma_{pri\text{-}1}$.

Consider first the sensitivity with respect to Mean$[S_{or\text{-}pri}]$ (Fig. 12).

When $N_{\theta 1} = N_{\theta 2} = 1$ (left column) the amount of data is very large compared to the number of parameters, and the posterior parameter estimates are almost independent of prior properties and with very small posterior uncertainty. Thus, the calculated weights are almost independent of Mean$[S_{or\text{-}pri}]$ for BSLOO and MBS-post. The results with MBMA are very similar to those obtained with MBS-pri, indicating that averaging BMP over many data realizations have a similar stabilizing effect as stacking over prior distributions—a result we have seen in all cases tested.

The middle column shows the case with $N_{\theta 1} = 2500$ and $N_{\theta 2} = 1$. We see that when the measurement errors are corre-

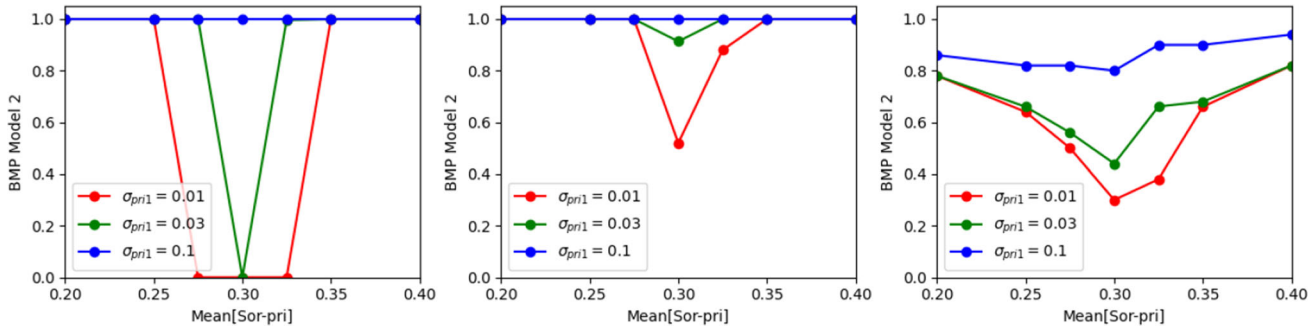

Fig. 11 Estimated BMP model 2 vs Mean[$S_{or\text{-}pri}$] and $\sigma_{pri\text{-}1}$. $N_{\theta 1} = N_{\theta 2} = 2500$. From Eq. 2 (left); from Eqs. 4 and 5 (middle); from Eqs. 4 and 6 (right)

lated, $w_2 = 0$ independently of Mean[$S_{or\text{-}pri}$] for all methods except BSLOO. This is because the heterogeneous fields will be closer to the data field when measured by the Mahalanobis distance, more or less independently of the mean. As commented by Stewart et al. [16] positive error correlations reduce weight given to the average of observations, but give more weight to differences between observed values. This illustrates a challenge when using methods which are based on the Mahalanobis distance in high dimensions: the estimated predictive densities become very sensitive to the shape of the surfaces, and also to small, long-range correlations, which normally are not well known. With Case 2 data, the measurement errors are uncorrelated. However, in this particular case, the calculations are dominated by correlations in the prior predictions for Scenario 2 (second term in Eq. A.3). Remember that since $N_{\theta 2} = 1$, the predicted data for model 2 will be 100% correlated, and the determinant, C_k, will be much smaller for model 2 than for model 1. The distance between predictions and data is not very different, and thus the determinant will dominate the BME calculation Eq. A.2 giving $w_2 = 1$ for MBMA and MBS-pri independently of Mean[$S_{or\text{-}pri}$]. On the other hand, the large number of parameters for model 1 allows for a much better posterior match to both the mean and shape of the measurements than model 2. Thus, w_2 is lower for MBS-post than for the other methods. Notice also that when $N_{\theta 1} = 2500$, and the data are uncorrelated, the matrices G, C_{pri} and C_d are all diagonal, and consequently, there will be no Bayesian update with

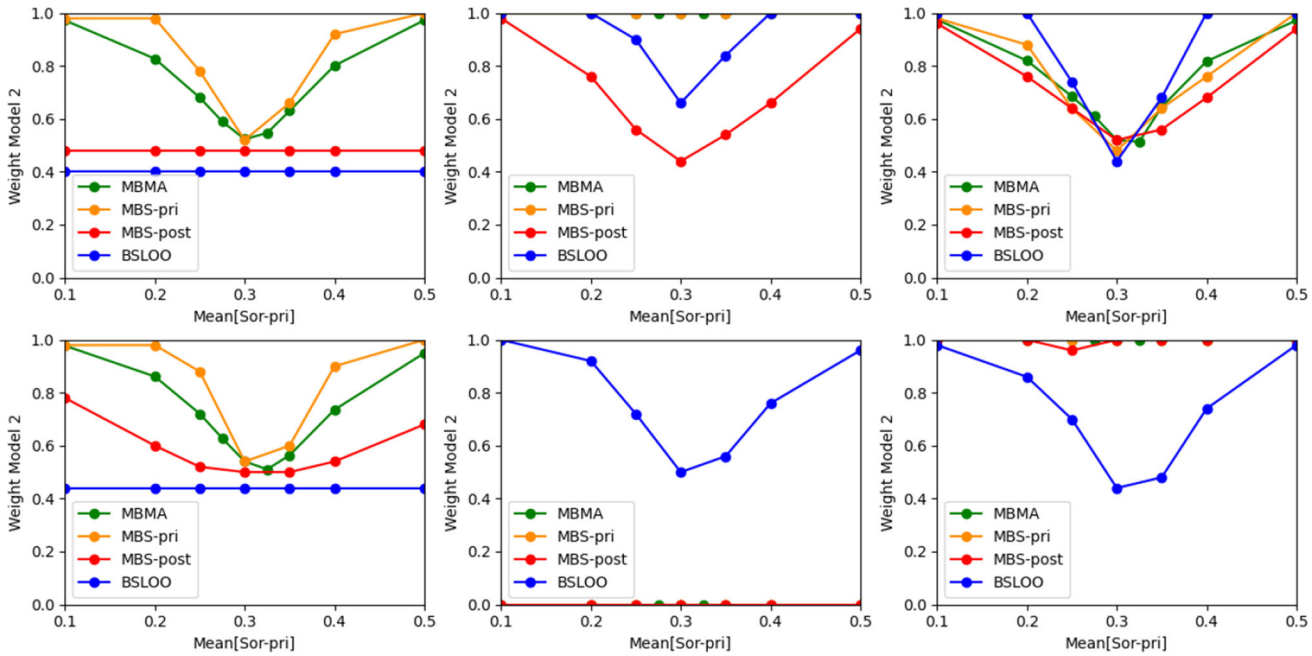

Fig. 12 Estimated weight model 2 vs Mean[$S_{or\text{-}pri}$]. Top row: Data Case 2 (uncorrelated). Bottom row: Data Case 3 (correlated; R_d = 37.5 cells). Left column:$N_{\theta 1} = N_{\theta 2} = 1$. Middle column: $N_{\theta 1} = 2500$, $N_{\theta 2} = 1$. Right column: $N_{\theta 1} = N_{\theta 2} = 2500$

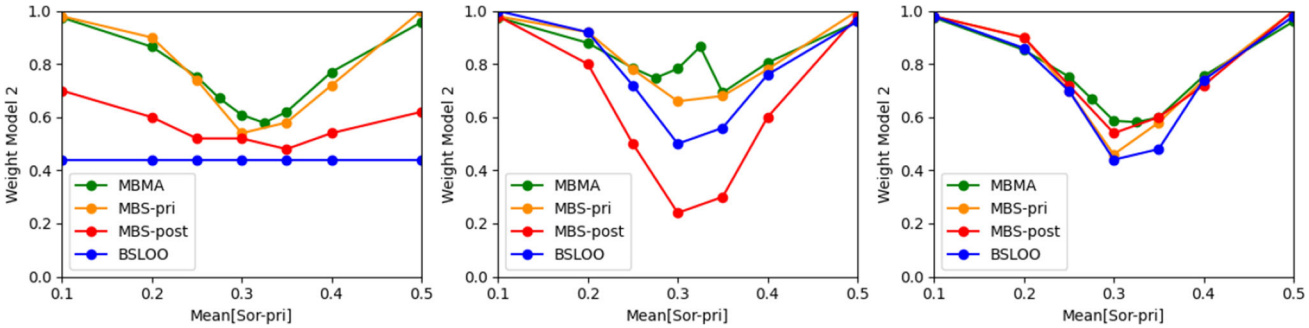

Fig. 13 Estimated weight model 2 vs Mean[$S_{or\text{-}pri}$]. Data Case 3. Weight calculations for MBMA and MBS based on diagonal approximation to Eq. A.3. Left column:$N_{\theta 1} = N_{\theta 2} = 1$. Middle column: $N_{\theta 1} = 2500$, $N_{\theta 2} = 1$. Right column: $N_{\theta 1} = N_{\theta 2} = 2500$

BSLOO. Here this applies to model 1, and the weights estimated with BSLOO become higher than for MBS-post, but not equal to 1.

When $N_{\theta 1} = N_{\theta 2} = 2500$ (right column), there is no update in any of the models with BSLOO when the data are uncorrelated, and the BSLOO results are very similar to the case with $N_{\theta 1} = 2500$ and $N_{\theta 2} = 1$. Also, since correlations are neglected with BSLOO, the results for Case 2 and Case 3 data will be very similar. For Case 2 data all methods give similar results since the Bayesian parameter updates are only minor with this relatively large measurement uncertainty. With Case 3 data, $w_2 = 1$ for all methods, except BSLOO. This is because the shape of prior model 2 matches the data slightly better than model 1 giving a high weight to model 2, and this again illustrates the challenge when using high-dimensional probability distributions.

Figure 13 illustrates the effect of neglecting off-diagonal terms in the matrix C_k in Eq. A.3. Comparing to the bottom row of Fig. 12 we see that when $N_{\theta 1} = N_{\theta 2} = 1$ the difference is small, since the prior and posterior surfaces are flat. However, the calculation of model weights using MBMA, MBS-pri and MBS-post is now dominated by the difference between mean surfaces and measurements and not the shape, and when $N_{\theta 1} = N_{\theta 2} = 2500$, the results become almost identical to those of BSLOO and also to those with Case 2 data. They are also similar to the BSLOO results when $N_{\theta 1} = 2500$ and $N_{\theta 2} = 1$, but now w_2 is lower with MBS-post since it is based on the Bayesian update.

Figure 14 shows estimated w_2 vs $\sigma_{pri\text{-}1}$ for the different methods. When $N_{\theta 1} = N_{\theta 2} = 1$ (left plot), w_2 will be independent of prior uncertainty for MBS-post and BSLOO because the posterior estimates are then almost independent of prior assumptions. For MBMA and MBS-pri, estimated weight for a given model will depend on its prior confidence, and thus w_1 decreases (and w_2 increases) with increasing $\sigma_{pri\text{-}1}$. When $N_{\theta 1} = 2500$ (middle and right plot), the sensitivity to prior uncertainty is even larger for MBMA and MBS, and because there is no Bayesian update of θ_1 with BSLOO in this case, also BSLOO show the same sensitivity. With MBS-post, however, w_2 decreases with $\sigma_{pri\text{-}1}$. This example illustrates the complexity involved in model weight calculations, and to understand the behavior, consider the expression for the predictive distributions, Eq. A.2. The main factors influencing the predictive distributions are the weighted mismatch between data and predictions and the determinant of the matrix, C_k. For the prior predictive distribution, the effect of an increase in det C_1 with increasing $\sigma_{pri\text{-}1}$ is larger than the effect of a slightly decreased mismatch (due to a changed weight). For the posterior predictive distribution, however, the effect of a decreased mismatch when $\sigma_{pri\text{-}1}$ increases is

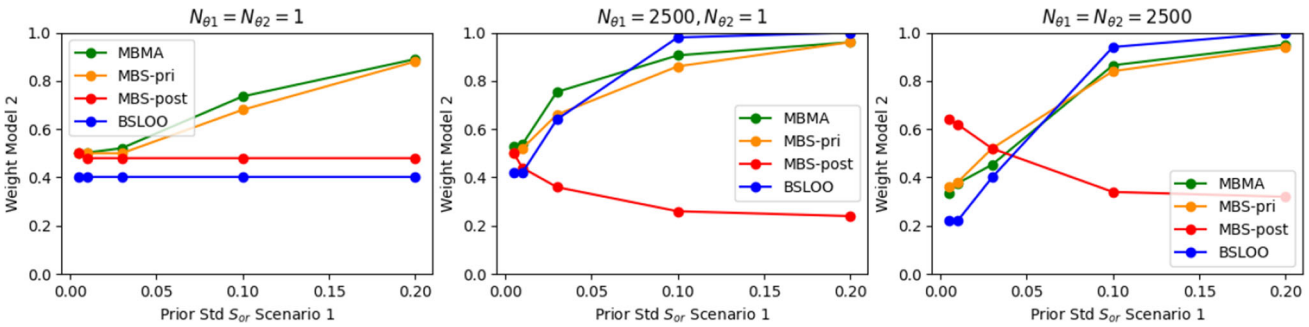

Fig. 14 Estimated weight model 2 vs $\sigma_{pri\text{-}1}$. Data Case 2

Table 3 Example 3. Properties of the prior surface interpretations, I_1 and I_2

	Model 1	Model 2
Mean, $\widetilde{M}$ (m)	1000	1010
Standard deviation (m)	5	5
Variogram model	Spherical	Spherical
Variogram range, $\widetilde{R}$ (grid cells)	5	50

larger than the effect of an increased det C_1, since increasing $\sigma_{pri\text{-}1}$ will increase the weight of the data, and thus give a significantly lower posterior mismatch.

3.3 Example 3

3.3.1 Example specification

In this example we assume that the unknowns are depths (in each grid cell) of an underground surface, and for simplicity, the data are assumed to be uncertain measurement of the same surface. That is, the forward model is given by a unit matrix, and the number of parameters and the number of measurements is both equal to 2500. We further assume that there exist two alternative prior seismic interpretations of this surface, I_1 and I_2, each defined by a single realization of two corresponding Gaussian random fields. Spatial correlations are defined by a spherical variogram model. Mean, variance and variogram range may be different for the two interpretations. Here we have one interpretation with a short correlation length reflecting an interpretation philosophy following small scale variations, and one with a long correlation length reflecting an interpretation philosophy favorizing smooth surfaces. Input data to generate the prior interpretations are listed in Table 3. The prior surface interpretations are plotted in Fig. 15. Cross sections of the surfaces along each row of the grid are plotted in Fig. 16 illustrating the effect of different variogram range. These two prior interpretations are then assumed to be expected values for two prior statistical models. These models are also assumed to be Gaussian random fields, where the uncertainty now represents uncertainty in the interpretations. Input data for the prior statistical models are listed in Table 4.

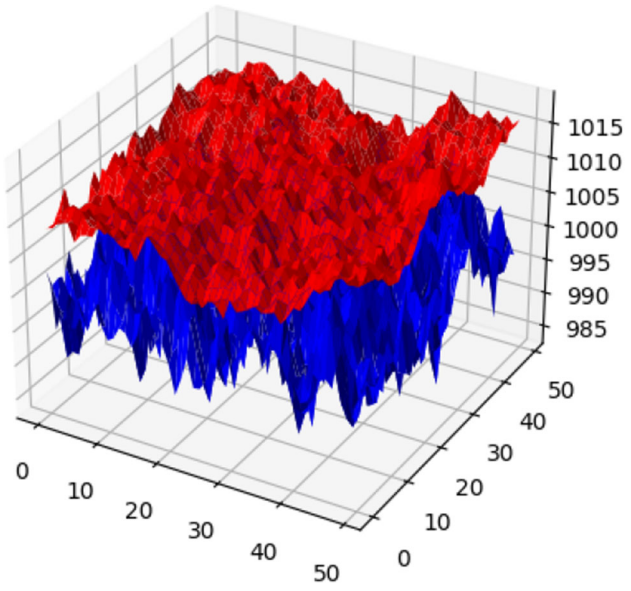

Fig. 15 Prior interpretations

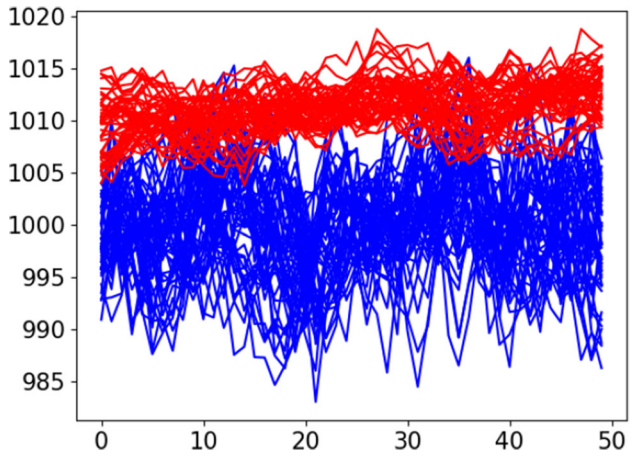

Fig. 16 Prior interpretations. Cross sections through all 50 rows

The main focus of this example is stability of the weight calculations with respect to data. Thus, we have generated a set of measurements varying between the two prior interpretations. This example also demonstrates how an inherently hierarchical model can be approximated by a small number of scenarios.

The measurements are generated in the same way as the prior interpretations, but several datasets are made with mean and correlation length being linear combinations of those of the prior interpretations. That is, for each value $\alpha = 0.0, 0.1, 0.2, ..., 1.0$ and $\beta = 0.0, 0.5, 1.0$ a surface, $\widetilde{d}(\alpha, \beta)$, is generated as a realization of a Gaussian random field with mean, $\widetilde{M}_d(\alpha)$, given by,

$$\widetilde{M}_d(\alpha) = \alpha \widetilde{M}_1 + (1 - \alpha)\widetilde{M}_2, \tag{26}$$

Table 4 Example 3. Properties of the prior models

	Model 1	Model 2
Expected value, $E[\theta]$	I_1	I_2
σ_{pri} (m)	5	5
Covariance matrix, C_{pri}	Diagonal	Diagonal

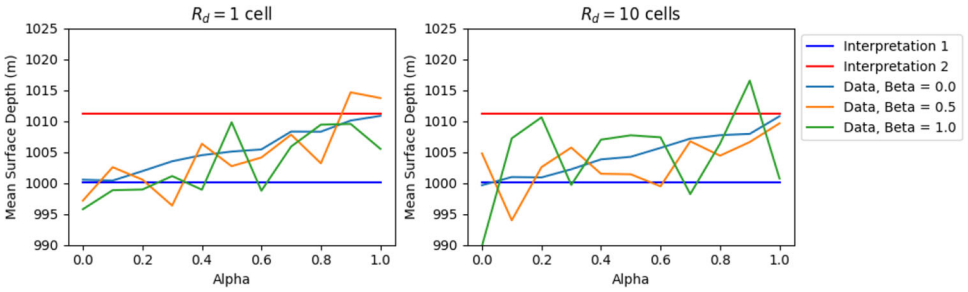

Fig. 17 Mean depths of prior interpretations and data

and a covariance matrix, $\widetilde{C}_d(\beta)$, with a variance $(10m)^2$ and a spherical variogram with variogram range, $\widetilde{R}_d(\beta)$, given by,

$$\widetilde{R}_d(\beta) = \beta\widetilde{R}_1 + (1-\beta)\widetilde{R}_2. \tag{27}$$

Again we assume that the measurement error is additive, Gaussian with zero mean. Given these surfaces, $\widetilde{d}(\alpha, \beta)$, measurements, $d(\alpha, \beta)$, are generated by adding a realization from the measurement error distribution. Notice that the measurement error covariance matrix, C_d, may be different from $\widetilde{C}_d$. Four datasets were made for each α, β corresponding to diagonal C_d and correlated C_d with range, $R_d = 10$ cells. $\sigma_d = 2$m or 10m.

Figure 17 shows mean of the prior surfaces and data vs α for different values of β, and it would be expected that the estimated weights should to some degree follow these variations in data mean when α and β are varied. The mean values are calculated from the generated realizations and may be slightly different from $\widetilde{M}_d$, $\widetilde{M}_1$ and $\widetilde{M}_2$. The variation with α is relatively smooth for the uncorrelated case ($\beta = 0$), but with more and more variations when β increases.

Using the generated prior interpretations and datasets, model weights were calculated using MBMA, MBS-pri, MBS-post and BSLOO. The models are given by Eq. A.1 with $G_k = I$, and the integrals Eqs. 3, 8 and 10 are calculated from Eqs. A.2 and A.3 using the appropriate prior or posterior properties.

For non-diagonal C_d we have used both the full matrix and a diagonal approximation when calculating the predictive distributions for MBMA and MBS. The full C_d is always used in the Bayesian inversion prior to MBS-post. Notice again that for BSLOO, since the matrices G and C_{pri} are diagonal, there will be no update of θ_i with d_{-i} in the cases where C_d is also diagonal. The weights calculated with BSLOO are based on 500 measurements, i.e., every 5th measurement is used.

Results with MBMA and MBS are based on using Eq. 6 with 400 realizations. If Eq. 5 is used, all measurement realizations will typically give a higher value for the predictive distribution to one of the models, and the estimated weight will be 0 or 1 as shown in the left plot in Fig. 18. On the other hand, using Eq. 6 gives a relatively smooth variation in weight increasing with increasing values of α (Fig. 18, right plot). The variations then mainly reflects the variations in Mean[d] (cf. Fig. 17). Like in example 2 above, the traditional BMA using only one dataset, always yields just 0 or 1 for BMP.

3.3.2 Evaluation of method performance

Some selected weight calculations illustrating the behavior of the alternative methods are shown in Fig. 19. Estimated weights for the second interpretation (model 2) are plotted vs α for different values of β, and since the variation with α and β is believed to follow the variation in the data mean as shown in Fig. 17, we have plotted the estimated w_2 together with the data mean. Much larger variations in the weight calculations with α and β than in the corresponding data mean may indicate a less stable method. To further quantify

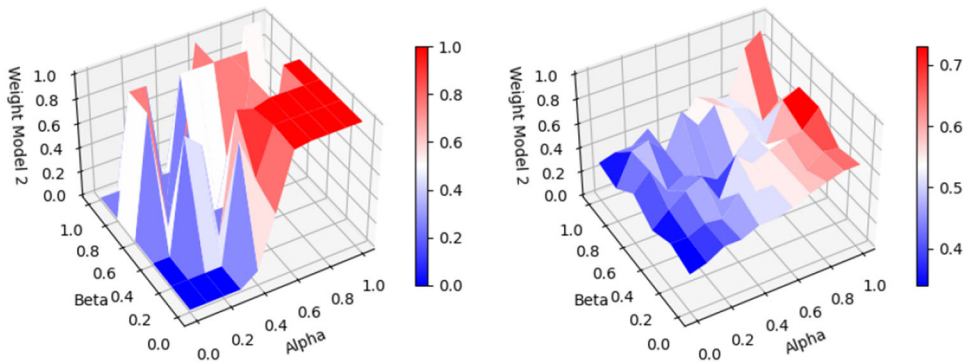

Fig. 18 Weight model 2 estimated with MBS-post vs α and β. $R_d = 1$ cell. Based on Eq. 5 (left). Based on Eq. 6 (right)

Fig. 19 Estimated weight model 2 (left axes) and mean depth of measurements (right axes) vs α for different values of β for each of the 4 datasets

Table 5 Distance between estimated weights vs α and shifted mean depth vs α for different values of β for each of the 4 datasets

β	σ_d (m) / R_d (cells)	2 / 1	10 / 1	2 / 10	10 / 10
0.0	MBMA	0.99	0.09	2.62	4.31
	MBS-pri	0.94	0.11	2.59	4.29
	MBS-post	0.07	0.09	0.89	2.75
	BSLOO	0.23	0.46	0.14	0.31
0.5	MBMA	0.76	0.06	2.45	4.20
	MBS-pri	0.74	0.08	2.43	4.20
	MBS-post	0.10	0.09	0.65	2.85
	BSLOO	0.64	0.56	0.42	1.32
1.0	MBMA	0.98	0.06	1.71	4.54
	MBS-pri	0.93	0.06	1.72	4.50
	MBS-post	0.05	0.07	1.04	2.77
	BSLOO	0.73	0.50	1.08	0.71

the results, the data mean depth vs α was shifted to the same axis as the weights (i.e., (990,1025) $\rightarrow$ (0.0,1.0)), and the distance between this depth and $w(\alpha)$ was calculated. The results are listed in Table 5.

Again, the results with MBS-pri and MBMA are almost identical in all cases. The general trend is that the degree of instability increases for all methods when β increases, i.e., the measurements get more correlated. MBS-post follows the data more closely than MBS-pri and MBMA, and the difference increases when σ_d decreases, as expected. Comparing MBS-post and BSLOO, MBS-post is better than BSLOO in the two cases where $R_d = 1$. For the case with $\sigma_d = 2$m and $R_d = 10$ cells, the results are quite similar. For $\sigma_d = 10$m and $R_d = 10$ cells, BSLOO apparently is better than MBS-post according to the measure in Table 5. However, when the data surface is highly correlated, one would expect a higher weight for model 2 which is based on a Gaussian realization with long correlation length, independently of α. This is also predicted with MBS-post as seen in Fig. 19. BSLOO, on the other hand does not take the correlations into account and also seems to be more unstable. In this case w_2 calculated with MBMA and MBS-pri becomes equal to 1 for all values of α and β.

Figure 20 and Table 6 show the same case using a diagonal approximation to Eq. A.3 in the analysis. Compared to the corresponding case based on the full matrix the weights calculated with MBMA, MBS-pri and MBS-post now follows the data mean closely because less weight is given to the shape of the data field. For BSLOO, there is no change as expected.

3.3.3 Uncertainty in stacking weights

To evaluate the uncertainty in the estimated stacking weights, we calculated weights with MBS-pri and MBS-post for 50 realizations of the data generated by adding realizations from the error distribution for $\alpha = 0.3$ and $\beta = 0.2$. Thus, these data realizations were generated using Eq. 5. The MBS weights calculated for each of these data realizations, however, where still based on Eq. 6. Mean and standard deviation of the 50 calculated weights for the 4 alternative measurement error covariance matrices are listed in Table 7. The standard deviation is relatively small in all cases. Diagonal approximations to C_d are used in the weight calculations. With a full C_d the standard deviation is similar or smaller.

4 Summary and conclusions

While uncertainty in model parameters is commonly taken into account when solving subsurface-related inverse problems, the uncertainty related to the model itself is most often ignored. More robust forecasting can be obtained by employing several models in the analysis and apply Bayesian model combination.

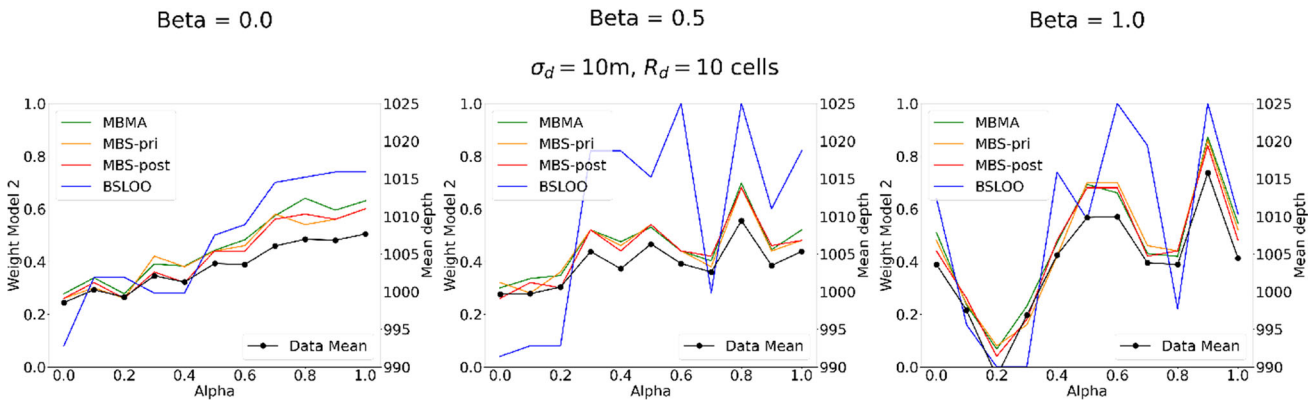

Fig. 20 Estimated weight model 2 (left axes) and mean depth of measurements (right axes) vs α for different values of β. Diagonal approximation to Eq. A.3 used in the weight calculations with MBMA and MBS

Table 6 Distance between estimated weights vs α and shifted mean depth vs α for different values of β diagonal approximation to Eq. A.3 used in the weight calculations with MBMA and MBS

β	σ_d (m) / R_d (cells)	10 / 10
0.0	MBMA	0.09
	MBS-pri	0.05
	MBS-post	0.04
	BSLOO	0.31
0.5	MBMA	0.06
	MBS-pri	0.05
	MBS-post	0.05
	BSLOO	1.32
1.0	MBMA	0.09
	MBS-pri	0.09
	MBS-post	0.06
	BSLOO	0.71

Formally, Bayesian Model Averaging, BMA, which is based on a discrete form of Bayes rule, is only appropriate for the $\mathcal{M}$-closed case where the true data is generated by one of the candidate models. Reservoir modelling problems will typically be $\mathcal{M}$-open with the true reservoir being far more complex than any of the models. With large amounts of data, like seismic data, BMA will typically select one of the candidate models with probability one and predicted uncertainty in parameters or future data may then be biased and too small. The calculation of model probabilities from the predictive distributions may also be unstable. We suggest to avoid these problems using a modified BMA (MBMA), where the model probabilities are averaged over a range of data realizations.

Model stacking is a method that is directly focused on the performance of the combined predictive distribution. The original version of Yao et al. [5] (BSLOO) is based on Leave-One-Out Cross Validation and requires a Bayesian inversion for each data point. We suggest a modified Bayesian stacking method (MBS), which is based on predictive distributions using an ensemble of measurement realizations. This modified stacking can be applied both with prior and posterior predictive distributions. With MBS-post, only a single Bayesian inversion is needed for each model. Thus, the computational effort with MBS-post is much less than for the basic BSLOO method. In addition, correlated measurement errors may be taken into account.

Table 7 Mean and standard deviation of estimated stacking weights model 2 from 50 measurement realizations $\alpha = 0.3$, $\beta = 0.2$

Error covariance		MBS-pri		MBS-post	
σ_d (m)	R_d (cells)	Mean$[w_2]$	Std$[w_2]$	Mean$[w_2]$	Std$[w_2]$
2	1	0.043	0.012	0.403	0.025
10	1	0.373	0.026	0.394	0.025
2	10	0.295	0.056	0.214	0.049
10	10	0.316	0.056	0.317	0.062

Using three synthetic, linear examples, MBMA, MBS-pri and MBS-post have been compared to the traditional BMA and BSLOO. The first example is a Gaussian mixture model taken from [5]. The other two are inspired by the interpretation of data from repeated (4D) seismic surveys. In one of these there are two distinct model scenarios, one of which is an approximation to the true data-generating model. The other is a case which could have been expanded to a hierarchical model.

We demonstrate that all the methods, MBMA, MBS-pri, MBS-post and BSLOO avoid the problem that the probability of one model approaches 100% when the number of measurements inceases. Also, the sensitivity to prior model assumptions are lower than with BMA. The MBS-pri results are very similar to those obtained with MBMA. The results using MBS-post are generally quite good, with lower sensitivity to the prior assumptions and more emphasis on the data than MBMA and MBS-pri. When measurement errors are correlated, MBS may be used with the full C_d or a diagonal approximation depending on whether the emphasis should be put on matching the shape of the data field or the mean. With respect to stability, MBS-post seems to perform equally good, or better, than BSLOO, especially with correlated data. The results with BSLOO become more unstable when correlations between measurements increase, even if data correlations are always neglected in the weight calculation. A disadvantage with MBS-pri, MBS-post and MBMA is that the predictive distributions are defined on the full data space, and for large dimensions the results with MBS-pri and MBS-post may be sensitive to small, and often very uncertain, correlations. These methods may be also more prone to the so-called curse of dimensionality than BSLOO, which is based on distributions defined in 1D. More investigations are needed to clarify this.

Appendix A: Predictive distribution for the linear, Gaussian case

In the linear, Gaussian case (i.e., the prior models are Gaussian, the forward model is linear and the measurement error is additive, Gaussian with zero mean), the posterior is also Gaussian for each model, and an analytical expression can

be derived for the predictive distribution or BME (see e.g. Bishop [17], Section 2.3.3). That is,

$$d = G_k\theta_k + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, C_d), \tag{A.1}$$

$$\begin{aligned} p(d|M_k) &= \mathcal{N}(G_k\bar{\theta}_k, C_k) \\ &= \left((2\pi)^{N_d} \det C_k\right)^{-1/2} \exp\left\{-\frac{1}{2}(d - G_k\bar{\theta}_k)^T C_k^{-1}(d - G_k\bar{\theta}_k)\right\}, \end{aligned} \tag{A.2}$$

$$C_k = C_d + G_k C_{\theta k} G_k^T. \tag{A.3}$$

Here $\bar{\theta}_k$ and $C_{\theta k}$ are the prior mean and covariance matrix for θ within model k, G_k is the forward model operator for model k and C_d is the measurement error covariance matrix. The weighted mismatch term appearing in the exponent is the *Mahalanobis distance* corresponding to the matrix C_k. This expression may also be applied to the other relevant predictive distributions by using the appropriate means and covariance matrices.

Appendix B: Predictive distribution for BSLOO

Yao et al. [5] suggests importance sampling to calculate the integral Eq. 8 required for BSLOO using the posterior distribution—given all the data—as the importance sampler. That is,

$$\begin{aligned} p(d_i|d_{-i}, M_k) &= \int p(d_i|\theta_k, M_k)\, p(\theta_k|d_{-i}, M_k)\, \mathrm{d}\theta_k \\ &= \frac{\int p(d_i|\theta_k, M_k)\, p(\theta_k|d_{-i}, M_k)\, \mathrm{d}\theta_k}{\int p(\theta_k|d_{-i}, M_k)\, \mathrm{d}\theta_k} \\ &= \frac{\int p(d_i|\theta_k, M_k)\, \frac{p(\theta_k|d_{-i}, M_k)}{p(\theta_k|d, M_k)}\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k}{\int \frac{p(\theta_k|d_{-i}, M_k)}{p(\theta_k|d, M_k)}\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k} \\ &= \frac{\int p(d_i|\theta_k, M_k)\, \frac{p(d_{-i}|\theta_k, M_k)/p(d_{-i})}{p(d|\theta_k, M_k, M_k)/p(d)}\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k}{\int \frac{p(d_{-i}|\theta_k, M_k)/p(d_{-i})}{p(d|\theta_k, M_k)/p(d)}\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k} \\ &= \frac{\int r_{i,k}^s\, p(d_i|\theta_k, M_k)\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k}{\int r_{i,k}^s\, p(\theta_k|d, M_k)\, \mathrm{d}\theta_k} \\ &\approx \frac{\sum_s r_{i,k}^s\, p(d_i|\theta_k^s, M_k)}{\sum_s r_{i,k}^s}, \end{aligned} \tag{B.1}$$

where θ_k^s are simulation draws from the full posterior $p(\theta_k|d, M_k)$, and

$$\begin{aligned} r_{i,k}^s &= \frac{p(d_{-i}|\theta_k^s, M_k)}{p(d|\theta_k^s, M_k)} \\ &\overset{\text{iid}}{=} \frac{1}{p(d_i|\theta_k^s, M_k)}. \end{aligned} \tag{B.2}$$

That is, for iid data

$$p(d_i|d_{-i}, M_k) \approx \left(\frac{1}{N_s}\sum_s \left(\frac{1}{p(d_i|\theta_k^s, M_k)}\right)\right)^{-1}. \tag{B.3}$$

Acknowledgements The authors acknowledge financial support from the NORCE research project "Assimilating 4D Seismic Data: Big Data Into Big Models", which is funded by industry partners, Equinor Energy AS, Aker BP ASA, Repsol Norge AS, Shell Global Solutions International B. V., TotalEnergies EP Norge AS, and Wintershall Dea Norge AS, as well as the Research Council of Norway (PETROMAKS2)

Data statement This manuscript has no associated data

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article

References

1. Genell, A., Nemes, S., Steineck, G., Dickman, P.W.: Model selection in medical research: a simulation study comparing bayesian model averaging and stepwise regression. BMC Med. Res. Methodol. **10**, 1317–1399 (2010)
2. Carrassi, A., Bocquet, M., Hannart, A., Ghil, M.: Estimating model evidence using data assimilation. Q. J. R. Meteorol. Soc. **143**, 866–880 (2017)
3. Carson, J., Crucifix, M., Preston, S., Wilkinson, R.D.: Bayesian model selection for the glacial-interglacial cycle. preprint arXiv:1511.03467
4. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian Model Averaging: A Tutorial. Statistical Sci. **14**(4), 382–417 (1999)
5. Yao, Y., Vehtari, A., Simpson, D., Gelman, A.: Using stacking to average bayesian predictive distributions (with discussion). Bayesian Anal. **13**(3), 917–1007 (2018)
6. Minka, T.P.: Bayesian model averaging is not model combination. Technical report, 2002. Available online: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.34.1359&rank=4
7. Höge, M., Guthke, A., Nowak, W.: Bayesian model weighting: The many faces of model averaging. Water 12(309) (2020)
8. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B: Bayesian Data Analysis. Chapman & Hall/CRS Press, third edition (2014)

9. Hong, A., Bratvold, R.B., Lake, L.W., Maraggi, L.M.R.: Integrating model uncertainty in probabilistic decline curve analysis for unconventional oil production forecasting. SPE Reservoir Eval. Eng. **22**(03), 861–876 (2019)
10. Aanonsen, S.I., Tveit, S., Alerini, M.: Using bayesian model probability for ranking different prior scenarios in reservoir history matching. SPE J. **24**(04), 1490–1507 (2019)
11. Cheng, Y., Wang, Y., McVay, D.A., Lee, W.J.: Practical application of a probabilistic approach to estimate reserves using production decline data. SPE Economics & Management **2**(01), 19–31 (2010)
12. Vehtari, A., Gelman, A., Gabry, J.: Pareto smoothed importance sampling. (2017). ArXiv e-print: arXiv:1507.02646
13. Vehtari, A., Gelman, A., Gabry, J.: Practical bayesian model evaluation using leave-one-out cross-validation and waic. Stat. Comput. **27**(5), 1413–1432 (2017)
14. Mannseth, T., Aanonsen, S.I., Fossum, K.: Calculating bayesian model evidence for porous-media flow using a multilevel estimator. Submitted (2022)
15. Fahimuddin, A.: 4D Seismic History Matching Using the Ensemble Kalman Filter (EnKF): Possibilities and Challenges. PhD thesis, Department of Mathematics, University of Bergen, Bergen, Norway, March (2010)
16. Stewart, A.M., Dance, S.L., Nichols, N.K.: Information content of spatially correlated observation errors. Technical report, Department of Mathematics, The University of Reading, Numerical analysis report 4/06 (2006)
17. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg (2006)