ORIGINAL PAPER

Flexible iterative ensemble smoother for calibration of perfect and imperfect models

Muzammil Hussain Rammay[1] · Ahmed H. Elsheikh[1] · Yan Chen[2]

Abstract

Iterative ensemble smoothers have been widely used for calibrating simulators of various physical systems due to the relatively low computational cost and the parallel nature of the algorithm. However, iterative ensemble smoothers have been designed for perfect models under the main assumption that the specified physical models and subsequent discretized mathematical models have the capability to model the reality accurately. While significant efforts are usually made to ensure the accuracy of the mathematical model, it is widely known that the physical models are only an approximation of reality. These approximations commonly introduce some type of model error which is generally unknown and when the models are calibrated, the effects of the model errors could be smeared by adjusting the model parameters to match historical observations. This results in a bias estimated parameters and as a consequence might result in predictions with questionable quality. In this paper, we formulate a flexible iterative ensemble smoother, which can be used to calibrate imperfect models where model errors cannot be neglected. We base our method on the ensemble smoother with multiple data assimilation (ES-MDA) as it is one of the most widely used iterative ensemble smoothing techniques. In the proposed algorithm, the residual (data mismatch) is split into two parts. One part is used to derive the parameter update and the second part is used to represent the model error. The proposed method is quite general and relaxes many of the assumptions commonly introduced in the literature. We observe that the proposed algorithm has the capability to reduce the effect of model bias by capturing the unknown model errors, thus improving the quality of the estimated parameters and prediction capacity of imperfect physical models.

Keywords Ensemble smoother · Calibration of imperfect model · Model error

1 Introduction

Bayesian inversion is a generic inference framework that is widely adopted for calibration of mathematical models while accounting for different types/sources of uncertainties. In the Bayesian framework, the prior model parameters belief or probability is updated by integrating the observed data to obtain the posterior probability. Several algorithms could be used to generate samples from the posterior distribution of the model parameters. Among those,

Markov chain Monte Carlo (MCMC) is an exact method for sampling. However, MCMC can be computationally expensive due to the large number of iterations (number of sampling steps) needed to reach convergence and the sequential nature of the method. Ensemble-based methods have been gaining popularity in the last two decades for data assimilation and calibration of simulation models of various physical systems [5]. The main advantage of ensemble-based methods is the low computational cost for higher dimensional data assimilation and inverse problems with some trade-off between the convergence speed/dimensionality and Gaussian approximation of the posterior distribution. Ensemble smoother (ES) is one of the ensemble-based techniques similar to the ensemble Kalman filter [6]. ES updates the model parameters by simultaneous assimilation of all available data. However, ES fails to solve highly non-linear inverse problems. A number of iterative ensemble-based methods have been proposed for parameter estimation of non-linear mathematical models. The ensemble smoother with multiple data assimilation

✉ Muzammil Hussain Rammay
mhr3@hw.ac.uk; muzammil.h.rammay@uis.no

Ahmed H. Elsheikh
a.elsheikh@hw.ac.uk

[1] Heriot–Watt University, Edinburgh, UK

[2] Total Geoscience Research Centre, Aberdeen, UK

(ES-MDA) [5], Levenberg-Marquardt ensemble randomized maximum likelihood method (LM-EnRML) [3] and iterative ensemble smoother (IES) [18, 27] are some of these techniques. In ensemble-based methods, the prior ensemble members of the model parameters are computed from the initial statistical distribution of the unknown model parameters and the objective is to find an approximate posterior distribution (i.e., posterior ensemble) of the model parameters conditioned to observation data. After calibration, the posterior ensemble of the model parameters is used for making predictions. Ensemble-based methods are designed with the assumption that utilized mathematical model provides a complete representation of real physical systems and that the model errors are small enough that it could be neglected during the calibration process. This assumption might introduce bias in the estimated parameter distribution [1, 26] and as a consequence results in bad quality predictions using the calibrated models.

A large number of studies have been conducted to explore the possible ways to account for the model error during the calibration process. In a Bayesian inversion context, three broad lines of research have emerged in the published literature. In the first line of research, the prior model error statistics were computed using pairs of high-fidelity and low-fidelity models. These error statistics were utilized during the calibration by using different types of algorithmic frameworks [10, 16, 17, 20, 22]. These frameworks vary according to the behavior of different physical systems and could perform poorly in high-dimensional problems with variable boundary conditions and when the distributions of the model error statistics are complex or multi-modal. The problem of complex statistics of model error was addressed by Köpke et al. [13] and Köpke et al. [14], where the authors accounted for the model error component using orthonormal basis generated from the difference between pairs of high-fidelity and low-fidelity models. These bases were evaluated locally at each update iteration as well as the model pairs (high and low fidelity) were re-run during the calibration process. More recently, an input/output independent formulation of model error was introduced to handle high-dimensional parameter estimation problems as well as handling problems with time-varying boundary conditions [23, 24]. However, this line of research relies on the availability of high-fidelity models such that we could learn the statistical properties of the model errors by evaluating both the high- and low-fidelity models at the same set of model parameters. This assumption might not be valid for a wide range of applied problems where we only have access to one model.

The second line of research for addressing model error/bias during calibration is related to joint calibration of physical models with a second model that is assumed to represent the model error. Different parameterization

for the model bias/error have been proposed, for example Gaussian process regression [12] or autoregressive error models [8, 9]. However, without any prior knowledge of the error model (unknown model error), the joint calibration may be prone to break the physical constraints of the systems and as a consequence might fail to improve the predictive capacity of the calibrated physical model [25]. Evin et al. [7] compared the joint calibration approach to a post-processing approach of accounting for the model error and concluded that the joint calibration approach was found to be less robust. This leads us to the third line of research, where model error is generally estimated from the residual (data mismatch obtained from difference between reality and simulation models). For example, Evin et al. [7] estimated the model error by using normalized residuals and autoregressive model with linear heteroscedasticity. However, this formulation has limitation in the scenarios where model error exhibits strong structural features and non-linear heteroscedasticity. Along the same line, Sargsyan et al. [25] utilized approximate Bayesian computation (ABC) to capture the model error uncertainty using residuals. However, the use of ABC in ensemble-based methods is not clearly understood. Recently, Oliver and Alfonzo [21] estimated the correlated structure of an approximate model error by computing a covariance matrix of the total residual obtained after one-round of calibration. This covariance matrix was utilized to estimate the model error statistics and then a re-calibration step is introduced in order to compensate the model error effects until a termination criterion is satisfied. This approach while novel requires multiple re-calibration iterations and has the limitation to handle non-Gaussian behavior of residuals. Furthermore, if the imperfect model is flexible enough, the total residual might vanish after one calibration step and then the model error/bias will be underestimated.

In this paper, a flexible iterative ensemble smoother is introduced to calibrate both perfect and imperfect models. The algorithm is simple to implement and has negligible computational overhead over the standard ES-MDA algorithm. ES-MDA is reformulated by splitting of data-residual (mismatch between model output and observation) into two parts by estimating a split parameter (scalar value) during the calibration process. The first part of the residual is used to update the model parameters and the second part is assumed to represent the model error. The objective of the proposed algorithm is to reduce the model bias by capturing the unknown model error uncertainty during the calibration of imperfect models in order to improve the prediction capability of the calibrated physical model. Furthermore, the proposed algorithm could be used as a diagnostic tool to check the reliability of the physical models and the need for a model refinement step. In this work, three test cases have been used to observe

the performance of the proposed algorithm. These test cases are related to the calibration of polynomial functions, simple machine and an imperfect reservoir model. In the first test case, cubic function is considered as the perfect model and imperfect models are represented by quadratic and linear functions. The second test case is related to estimation of efficiency of the simple machine model which lacks physics in terms of friction component. The third test case is related to calibration of imperfect reservoir model, which has blurred channelized geological patterns. The imperfect reservoir model has two sources of modeling errors, i.e., simplified geological representation and up-scaling errors.

The outline of this paper is as follows. In Section 2 the formulation of flexible iterative ensemble smoother is described. Following that, three test cases related to calibration of polynomial functions, simple machine model and imperfect reservoir model are described with results in Section 3. Section 4 is related to the conclusions of the paper.

2 Formulation of flexible iterative ensemble smoother

Standard data assimilation and parameter estimation procedure often rely on an implicit assumption that the model errors are generally small and could be neglected (i.e., the mathematical model is perfect). Mathematically, if a perfect model is utilized, the observed data is formulated as,

$$\mathbf{d}_{obs} = g(\mathbf{m}_{true}) + \boldsymbol{\epsilon}_d, \tag{1}$$

where $g(.)$ is a function representing the perfect model, $\mathbf{m}_{true}$ is the true model parameters, $\boldsymbol{\epsilon}_d$ is the measurement errors which is usually assumed to follow a normal distribution $\mathcal{N}(0, \mathbf{C}_D)$ and $\mathbf{C}_D$ is the covariance matrix of the measurement errors.

Generally, physical models are treated as if they are perfect during the calibration process and subsequently parameter mismatch/errors are assumed as the main source of the observation/data mismatch. However, it is widely known that physical models are imperfect and are only an approximation of reality in terms of description, scale, assumption or complexity. These approximations commonly introduce model errors that are often neglected during calibration. The mismatch between historical data (i.e., the true system response) $g(m_{true})$ and simulated data from an imperfect model $\tilde{g}(m)$ is a result of both parameter error and model error. Mathematically, if an imperfect model is utilized, the observed data could be formulated as,

$$\mathbf{d}_{obs} = \tilde{g}(\mathbf{m}_{true}) + \boldsymbol{\epsilon}_d + \boldsymbol{\epsilon}_m, \tag{2}$$

where $\boldsymbol{\epsilon}_m$ is the model error and $\tilde{g}(.)$ is a function representing the imperfect mathematical model. Failing to account for the model error effects during the model calibration process can result in significant bias in the calibrated model parameters and consequently might result in inaccurate predictions. Generally, the magnitude of the model error is unknown before calibration and data assimilation (i.e., we do not know the extent of the imperfection of the mathematical model). In this section, an iterative ensemble smoother is formulated with an approximate method to quantify the model errors during the calibration process in order to reduce and in some cases eliminate any bias in the estimated model parameters. Furthermore, we aim to develop an iterative ensemble smoother that is flexible enough to match the observed data to the level of measurement noise in the case of utilizing a perfect mathematical model for inversion. In order to achieve this objective, we first present the standard ES-MDA, followed by the proposed flexible formulation that could handle the cases where the model errors cannot be neglected. In standard settings, ES-MDA update equation can be written as [5],

$$\Delta \mathbf{M} = \mathbf{C}_{MD} (\mathbf{C}_{DD} + \alpha \, \mathbf{C}_D)^{-1} (\mathbf{D}_{uc} - \mathbf{D}), \tag{3}$$

where $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \mathbf{m}_3, \ldots, \mathbf{m}_{N_e}] \in \mathbb{R}^{N_m \times N_e}$ is an ensemble of model parameters of size N_e, $\mathbf{m}_r$ is a realization r of the model parameters of size N_m, $\mathbf{D} \in \mathbb{R}^{N_d \times N_e}$ is an ensemble of model outputs of size N_d generated by a perfect model denoted by an operator g (i.e., $\mathbf{D} = g(\mathbf{M})$), $\mathbf{D}_{uc} \in \mathbb{R}^{N_d \times N_e}$ is the ensemble of perturbed observations ($\mathbf{d}_{obs} \in \mathbb{R}^{N_d \times 1}$), α is the noise inflation parameter and $\mathbf{C}_D$ is the covariance matrix of the measurement error/noise. The covariances $\mathbf{C}_{MD}$ and $\mathbf{C}_{DD}$, representing approximate sensitiveness of the model response to changes in the model parameters, are defined using the following equations:

$$\mathbf{C}_{MD} = \frac{1}{N_e - 1} (\mathbf{M} - \mathbf{M}_{mean} \overrightarrow{1_{N_e}})(\mathbf{D} - \mathbf{D}_{mean} \overrightarrow{1_{N_e}})^{\mathsf{T}}, \tag{4}$$

$$\mathbf{C}_{DD} = \frac{1}{N_e - 1} (\mathbf{D} - \mathbf{D}_{mean} \overrightarrow{1_{N_e}})(\mathbf{D} - \mathbf{D}_{mean} \overrightarrow{1_{N_e}})^{\mathsf{T}}. \tag{5}$$

where $\mathbf{M}_{mean} \in \mathbb{R}^{N_m \times 1}$ is the ensemble mean of model parameters, $\mathbf{D}_{mean}$ is the ensemble mean of model outputs and $\overrightarrow{1_{N_e}} \in \mathbb{R}^{1 \times N_e}$ is a row vector of ones. In ES-MDA, non-linear inverse problem is solved iteratively with an inflated noise covariance matrix and the inflation factor α is normally set to the total number of data assimilations/iterations N_a.

For the proposed Flexible ES-MDA, the output of the imperfect model is related to the perfect model output using the following equation:

$$\mathbf{D} = \widetilde{\mathbf{D}} + \mathbf{E}, \tag{6}$$

where $\widetilde{\mathbf{D}}$ is an ensemble of model output generated from the imperfect model denoted by the operator $\widetilde{g}$ using $\widetilde{\mathbf{D}} = \widetilde{g}(\mathbf{M})$ and $\mathbf{E}$ is an ensemble of the model error. Similarly, the mean of the previous equation over the ensemble outputs resulted the following equation:

$$\mathbf{D}_{mean} = \widetilde{\mathbf{D}}_{mean} + \mathbf{E}_{mean}. \tag{7}$$

For the approximate sensitivities, substituting Eqs. 6 and 7 into Eqs. 4 and 5 resulted in the following modified covariances:

$$\mathbf{C}_{MD} = \mathbf{C}_{M\widetilde{D}} + \mathbf{C}_{ME}, \tag{8}$$

$$\mathbf{C}_{DD} = \mathbf{C}_{\widetilde{D}\widetilde{D}} + \mathbf{C}_{\widetilde{D}E} + \mathbf{C}_{E\widetilde{D}} + \mathbf{C}_{EE}. \tag{9}$$

where,

$$\mathbf{C}_{M\widetilde{D}} = \frac{1}{N_e - 1}(\mathbf{M} - \mathbf{M}_{mean}\overrightarrow{1_{N_e}})(\widetilde{\mathbf{D}} - \widetilde{\mathbf{D}}_{mean}\overrightarrow{1_{N_e}})^{\mathsf{T}} \tag{10}$$

$$\mathbf{C}_{ME} = \frac{1}{N_e - 1}(\mathbf{M} - \mathbf{M}_{mean}\overrightarrow{1_{N_e}})(\mathbf{E} - \mathbf{E}_{mean}\overrightarrow{1_{N_e}})^{\mathsf{T}} \tag{11}$$

$$\mathbf{C}_{\widetilde{D}\widetilde{D}} = \frac{1}{N_e - 1}(\widetilde{\mathbf{D}} - \widetilde{\mathbf{D}}_{mean}\overrightarrow{1_{N_e}})(\widetilde{\mathbf{D}} - \widetilde{\mathbf{D}}_{mean}\overrightarrow{1_{N_e}})^{\mathsf{T}} \tag{12}$$

$$\mathbf{C}_{\widetilde{D}E} = \frac{1}{N_e - 1}(\widetilde{\mathbf{D}} - \widetilde{\mathbf{D}}_{mean}\overrightarrow{1_{N_e}})(\mathbf{E} - \mathbf{E}_{mean}\overrightarrow{1_{N_e}})^{\mathsf{T}} \tag{13}$$

$$\mathbf{C}_{E\widetilde{D}} = \frac{1}{N_e - 1}(\mathbf{E} - \mathbf{E}_{mean}\overrightarrow{1_{N_e}})(\widetilde{\mathbf{D}} - \widetilde{\mathbf{D}}_{mean}\overrightarrow{1_{N_e}})^{\mathsf{T}} \tag{14}$$

$$\mathbf{C}_{EE} = \frac{1}{N_e - 1}(\mathbf{E} - \mathbf{E}_{mean}\overrightarrow{1_{N_e}})(\mathbf{E} - \mathbf{E}_{mean}\overrightarrow{1_{N_e}})^{\mathsf{T}} \tag{15}$$

Using these sensitivities and substitution of Eqs. 6, 8 and 9 into Eq. 3, the Flexible ES-MDA update equation while accounting for the model error can be written as:

$$\Delta\mathbf{M} = (\mathbf{C}_{M\widetilde{D}} + \mathbf{C}_{ME})(\mathbf{C}_{\widetilde{D}\widetilde{D}} + \mathbf{C}_{\widetilde{D}E} + \mathbf{C}_{E\widetilde{D}} + \mathbf{C}_{EE} + \alpha\,\mathbf{C}_D)^{-1}(\mathbf{D}_{uc} - \widetilde{\mathbf{D}} - \mathbf{E}). \tag{16}$$

In general, the ensemble of modeling errors $\mathbf{E}$ is unknown. However, the residual of data mismatch includes the model error effects and we postulate that this data-residual could be split between a parameter update component and a modeling error component. In this study, we approximate the ensemble of model errors $\widetilde{\mathbf{E}}$ using a fraction (percentage) of the data-residual ensemble as shown in the following equation:

$$\widetilde{\mathbf{E}} = s_p\mathbf{R}, \tag{17}$$

where $\mathbf{R} = \mathbf{d}_{obs}\overrightarrow{1_{N_e}} - \widetilde{\mathbf{D}} \in \mathbb{R}^{N_d \times N_e}$ is the ensemble of residuals obtained from the differences between the observation and the ensemble of imperfect model outputs and s_p is a scaler split parameter. The prior (initial) split parameter is computed based on the ratio of the Euclidean norm of mean residual (mean deviation from observed data) and Euclidean norm of maximum residual (maximum

absolute deviation from observed data).

$$s_p{}^{(1)} = \frac{\|\boldsymbol{\sigma}_m\|}{\|\boldsymbol{\sigma}_{max}\|}, \tag{18}$$

where $\boldsymbol{\sigma}_m = \text{mean}(\mathbf{R}) \in \mathbb{R}^{N_d \times 1}$ is the mean residual (mean deviation from observed data) and $\boldsymbol{\sigma}_{max} = \text{max}(\text{abs}(\mathbf{R})) \in \mathbb{R}^{N_d \times 1}$ is the maximum residual (maximum absolute deviation from observed data). During calibration (data assimilation) this split parameter is updated based on the ratio of the Euclidean norm of mean residuals obtained in current iteration i and previous iteration $i - 1$.

$$s_p{}^{(i)} = \frac{\|\boldsymbol{\sigma}_m{}^{(i)}\|}{\|\boldsymbol{\sigma}_m{}^{(i-1)}\|}. \tag{19}$$

The statistical interpretation of the split parameter (Eqs. 18, 19) can be attributed to the correspondence of the mean residual to the model error/bias. Equation 18 is used to normalize the initial mean residual in the range of [0, 1] and Eq. 19 is used to represent the change in the mean residual during the calibration/data assimilation process. We note that if model response/output consists of multiple time series, the split parameter should be calculated individually for each time series.

Following the introduction of the ensemble of approximate model errors as defined in Eq. 17, we can rewrite Eq. 16 as the following:

$$\Delta\mathbf{M} = (\mathbf{C}_{M\widetilde{D}} + \mathbf{C}_{M\widetilde{E}})(\mathbf{C}_{\widetilde{D}\widetilde{D}} + \mathbf{C}_{\widetilde{D}\widetilde{E}} + \mathbf{C}_{\widetilde{E}\widetilde{D}} + \mathbf{C}_{\widetilde{E}\widetilde{E}} + \alpha\,\mathbf{C}_D)^{-1}(\mathbf{D}_{uc} - \widetilde{\mathbf{D}} - \widetilde{\mathbf{E}}). \tag{20}$$

We note that the proposed algorithm introduces an assumption about the model error form as defined in Eq. 17. This form is based on several numerical observations and is validated by extensive numerical testing. However, there is no guarantee that this form is optimal for all imperfect models. For those case where Eq. 17 fails to approximate the true model errors, utilizing the full update equation (Eq. 20) might result in invalid updates (and subsequently divergence) because of the corrupted sensitivity (cross-covariance) terms (i.e., $\mathbf{C}_{M\widetilde{E}}$, $\mathbf{C}_{D\widetilde{E}}$ and $\mathbf{C}_{\widetilde{E}D}$). Therefore, we adopted a relatively robust approach by retaining only the model error components with regularization effects ($\mathbf{C}_{\widetilde{E}\widetilde{E}}$). This results in a flexible ES-MDA formulation that is able to fit high-quality models with negligible model errors (a.k.a. perfect models) as well as suitable for calibration of imperfect models where model errors cannot be neglected.

Therefore, Eq. 20 can be simplified to the following form:

$$\Delta\mathbf{M} = \mathbf{C}_{M\widetilde{D}}(\mathbf{C}_{\widetilde{D}\widetilde{D}} + \mathbf{C}_{\widetilde{E}\widetilde{E}} + \alpha\,\mathbf{C}_D)^{-1}(\mathbf{D}_{uc} - \widetilde{\mathbf{D}} - \widetilde{\mathbf{E}}). \tag{21}$$

Equation 21 represents the final update form for flexible iterative ensemble smoother. We also note that after

---

Algorithm 1 Flexible ensemble-based algorithm for perfect and imperfect models.

1: Choose N_a ▷ number of data assimilation/iterations
2: $i \leftarrow 1$
3: $\alpha = N_a$
4: **while** $i <= N_a$ **do**
5:     $\mathbf{D}_{uc} = \mathbf{d}_{obs}\overrightarrow{1_{N_e}} + \sqrt{\alpha}\, \mathbf{C}_D^{1/2}\, \mathbf{Z}_d$,   ▷ Observation perturbations, $\mathbf{Z}_d = [\mathbf{z}_{d1}\ \mathbf{z}_{d2}\ \mathbf{z}_{d3}\ ......\ \mathbf{z}_{dN_e}] \in \mathbb{R}^{N_d \times N_e}$,
    $\mathbf{z}_d \sim \mathcal{N}(0, \mathbf{I}_{N_d}) \in \mathbb{R}^{N_d \times 1}$,
6:     $\widetilde{\mathbf{D}} = \tilde{g}(\mathbf{M})$ ▷ Generate ensemble of model outputs $\widetilde{\mathbf{D}} \in \mathbb{R}^{N_d \times N_e}$ from ensemble $\mathbf{M}$
7:     $\mathbf{R} = \mathbf{d}_{obs}\overrightarrow{1_{N_e}} - \widetilde{\mathbf{D}}$
8:     $\widetilde{\mathbf{E}} = s_p^{(i)}\mathbf{R}$ ▷ Compute ensemble of approximate model error, $\widetilde{\mathbf{E}} \in \mathbb{R}^{N_d \times N_e}$
9:     $\mathbf{M} \leftarrow \mathbf{M} + \mathbf{C}_{M\widetilde{D}}\, (\mathbf{C}_{\widetilde{D}\widetilde{D}} + \mathbf{C}_{\widetilde{E}\widetilde{E}} + \alpha\, \mathbf{C}_D)^{-1}(\mathbf{D}_{uc} - \widetilde{\mathbf{D}} - \widetilde{\mathbf{E}})$ ▷ Update ensemble $\mathbf{M}$
10:     $i \leftarrow i + 1$
11: **end while**
12: $\mathbf{M}_{post} = \mathbf{M}$ ▷ $\mathbf{M}_{post} \in \mathbb{R}^{N_m \times N_e}$ is the posterior ensemble of model parameters
13: $\widetilde{\mathbf{D}}_{post} = \tilde{g}(\mathbf{M}_{post})$ ▷ Generate posterior ensemble of model predictions, $\widetilde{\mathbf{D}}_{post}$

---

data assimilation, the magnitude of $\widetilde{\mathbf{E}}$ and $\mathbf{C}_{\widetilde{E}\widetilde{E}}$ could be used an indicator for the model accuracy. Higher magnitude refers to high model-inadequacy in terms of limited physics, reservoir geology, grid coarsening/up-scaling, parameterization or prior realizations descriptions. Ideally, after data assimilation for the perfect model cases, $\widetilde{\mathbf{E}}$ and $\mathbf{C}_{\widetilde{E}\widetilde{E}}$ would approach zero (i.e., $\mathbf{D} = \widetilde{\mathbf{D}}$). The details of the Flexible ES-MDA are shown in Algorithm 1. We note that, if the model response/output consists of multiple time series with different range of values, the rescaling and SVD thresholding of the inverse matrix $(\mathbf{C}_{\widetilde{D}\widetilde{D}} + \mathbf{C}_{\widetilde{E}\widetilde{E}} + \alpha\, \mathbf{C}_D)^{-1}$ should be done in the same way as shown in the paper [4].

3 Test cases

In this work, three test cases have been used to observe the performance of the proposed algorithm. These test cases are related to the calibration of polynomial functions, simple machine and imperfect reservoir model. For comparison purpose, calibration is performed using both the standard ES-MDA [5] and the proposed Flexible ES-MDA algorithm. An ensemble of 100 members is used with 8 number of iterations for the calibration of all test cases. In the first test case, cubic polynomial function is considered as perfect model and imperfect models are represented by quadratic and linear functions. The objective is to test the performance of the proposed algorithm for calibration of perfect and imperfect models. The second test case is related to the estimation of efficiency of the simple machine model which lacks physics in terms of friction component and used [1] as a test case for a joint calibration framework with the model error parameterized by Gaussian process regression. In this work, the objective is to test the calibration and prediction

improvement of simple machine model using the proposed algorithm without joint calibration. The third test case is related to calibration of imperfect reservoir model, which has blurred channelized geological patterns. The imperfect reservoir model has two sources of modeling errors, simplified geological representation and up-scaling errors.

3.1 Test case 1: the polynomial functions

In this test case, the data is generated from a cubic polynomial functions and perturbed with an additive measurement noise $\epsilon_d \sim \mathcal{N}(0, \mathbf{C}_D)$ using Eq. 22. The complexity of the calibrated models vary from a first order polynomial to a third order polynomials as shown in Eqs. 23, 24 and 25. The objective is to test the performance of the proposed algorithm on both scenarios where the calibrated model complexity matches the data generating process (i.e., perfect model) and when the calibrated model is less parameterized than the data generating process (i.e., imperfect models). The domain of x lies within the interval $[-1, 1]$. We calibrate three models (cubic, quadratic and linear models) to obtain the posterior distributions of each model parameters vector $\boldsymbol{\lambda}$ and the corresponding models outputs. The objective is to evaluate the flexibility of the proposed algorithm in quantifying the model error uncertainty for imperfect models as well as the ability to match the data for the perfect model case. Prior parameters are sampled from a standard normal distribution with zero mean ($\mu = 0$) and standard deviation $\sigma = 10$.
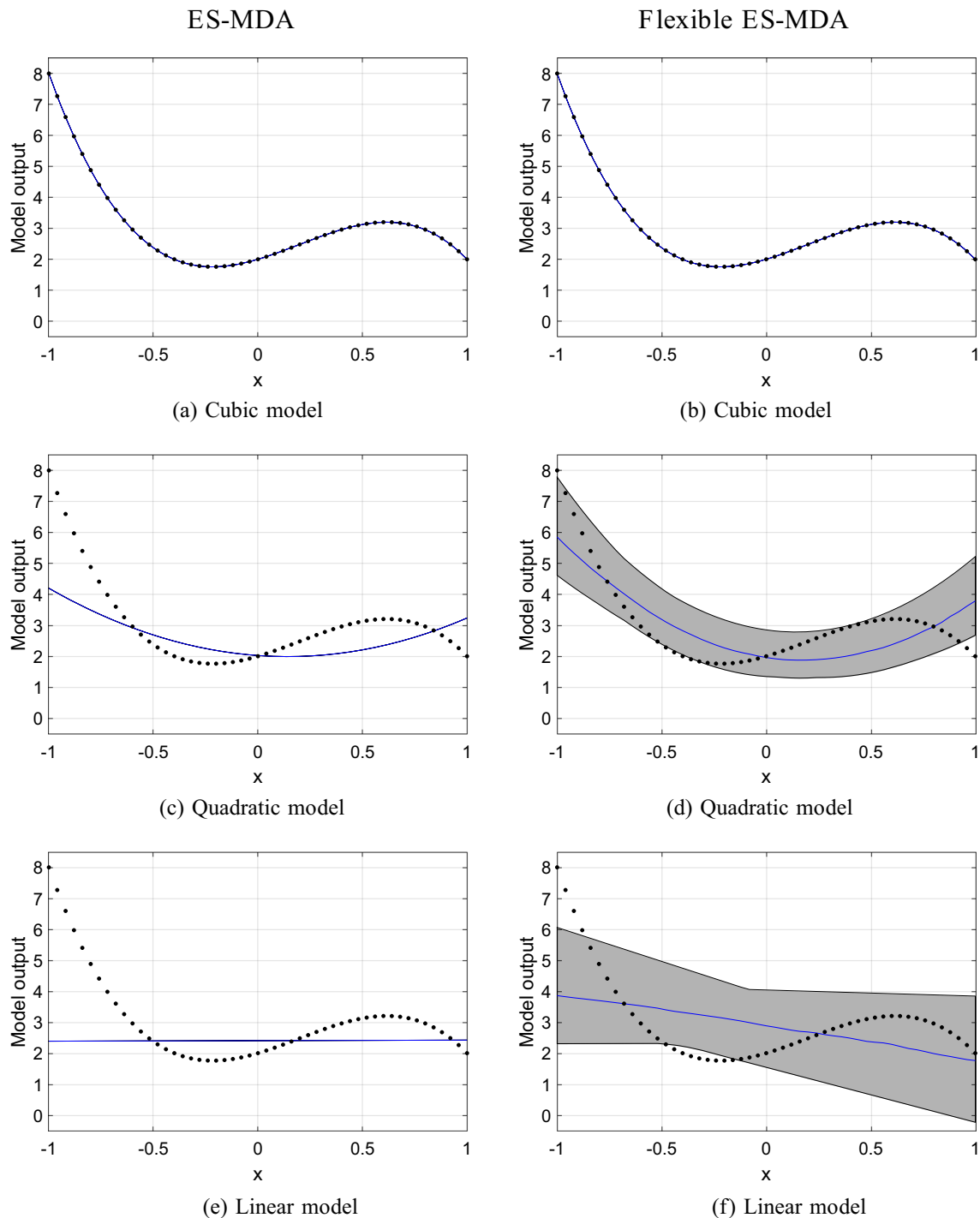
$$d_{obs} = 2 + 2x + 3x^2 - 5x^3 + \epsilon_d, \qquad (22)$$

$$f_1(x) = \lambda_0 + \lambda_1 x, \qquad (23)$$

$$f_2(x) = \lambda_0 + \lambda_1 x + \lambda_2 x^2, \tag{24}$$

$$f_3(x) = \lambda_0 + \lambda_1 x + \lambda_2 x^2 + \lambda_3 x^3. \tag{25}$$

The calibration of polynomial functions are performed using two levels of measurement noise. In the first level, negligible measurement error (i.e., order of magnitude 10^{-12}) is considered. Figure 1 shows the posterior distribution of cubic, quadratic and linear models outputs obtained by ES-MDA and Flexible ES-MDA algorithms. We observe that both algorithms manage to match the observations for cubic model (perfect model case), without any bias as shown in Fig. 1a and b. This shows that the

ES-MDA   Flexible ES-MDA

(a) Cubic model   (b) Cubic model

(c) Quadratic model   (d) Quadratic model

(e) Linear model   (f) Linear model

Fig. 1 Posterior distribution of models (cubic, quadratic and linear) outputs for test case 1 (negligible measurement error). Blue lines show $p50$ percentiles, gray shaded areas show 99% confidence intervals and solid black dots show observation data

proposed algorithm did not introduce any bias and has the flexibility to match the data for the perfect model case with negligible measurement errors. The posterior distributions of cubic model output (Fig. 1a and b) are appeared as the point estimate (i.e., exact data match) due to the effect of negligible measurement error (i.e., order of magnitude 10^{-12}).

Figure 1c and d show the posterior distribution of the quadratic model output. It is clear that the standard ES-MDA fails to match the data, while the calibrated model using the Flexible ES-MDA results in model outputs that account for model error uncertainty and provides a good coverage for the observations as shown in Fig. 1d. Similar results are observed for the linear model case in Fig. 1e and f, where a better coverage is obtained by the proposed Flexible ES-MDA. However, we note that the coverage is not as good for the quadratic model case. This shows that the capacity as well as the limitation of the proposed algorithm in accounting for the modeling errors during the calibration.

For further validation of the proposed algorithm, we compare the best split factor (computed using the true model error ensemble $\mathbf{E}_{actual}$) with the approximated split factor s_p estimated using the formulation proposed in this manuscript. The best split is computed by minimizing the Frobenius norm of the difference between true model error ensemble and the residual ensemble, i.e., $\min \|\mathbf{E}_{actual} - s_p\mathbf{R}\|$. This simple minimization problem is solved using differential evolution algorithm [28].

Figure 2 shows both the best split factor s_p when the true model error is known and the approximated split versus the number of ES-MDA iterations. At iteration 0 (i.e., prior), both the best and proposed split factors show very low values, because the prior residual is relatively larger than the model error. For subsequent iterations, the split parameter is increasing with respect to the iteration number, which is attributed to the convergence of the proposed flexible algorithm to the level of model error uncertainty. At iteration 2, the proposed split is very close to best split value for both linear and quadric imperfect models, which show the robustness of the proposed equation in approximating the split factor. After iteration 2, the proposed split parameter approached 1, which show that all the data mismatch (residual ensemble) is treated as a model error ensemble and further reduction of uncertainty is not possible due to the limited capacity of both the linear and quadratic models in matching the data. These results demonstrate the ability of the proposed algorithm in capturing the unknown model error uncertainties during the calibration of imperfect models.

Using the same problem setup, we add a measurement error of 5% of true function value (Eq. 22). The objective of this experiment is to evaluate the proposed algorithm when measurement errors are present while calibrating perfect and imperfect models. Figure 3 shows the posterior outputs of the calibrated cubic, quadratic and linear models. For the cubic models case, both the standard ES-MDA and the Flexible ES-MDA algorithms managed to match the structural feature of the data as shown in Fig. 3a and b. However, the proposed Flexible ES-MDA algorithm performs significantly better in terms of capturing the noisy features of the data (Fig. 3b). This effect shows that the split formulation of proposed algorithm act as a adaptive regularizer for the perfect model scenario. Similar results (w.r.t negligible measurement noise case) are observed for both the quadratic and the linear models in the presence of both measurement and model errors (Fig. 3c, d, e
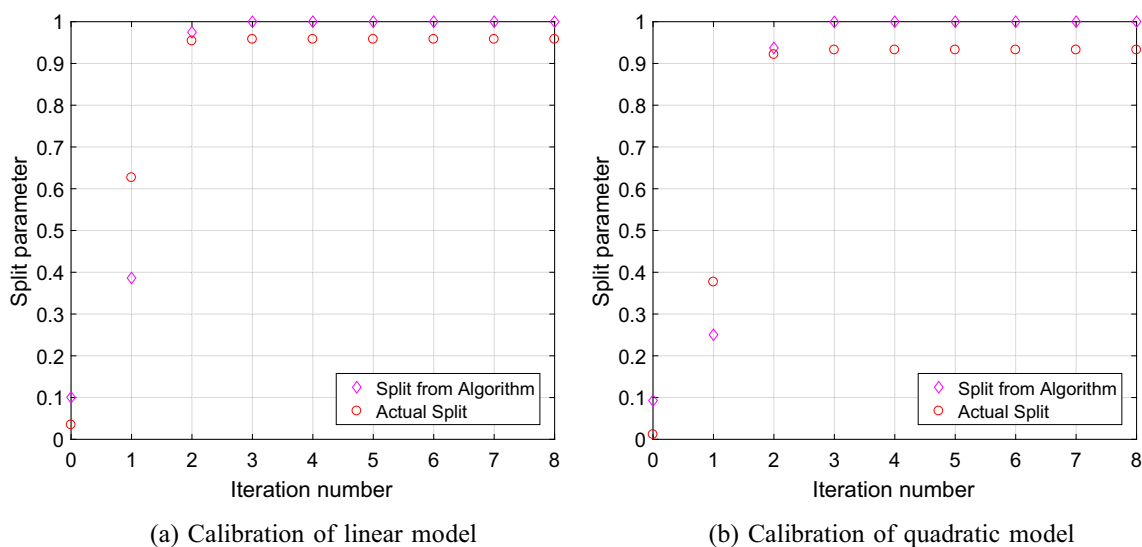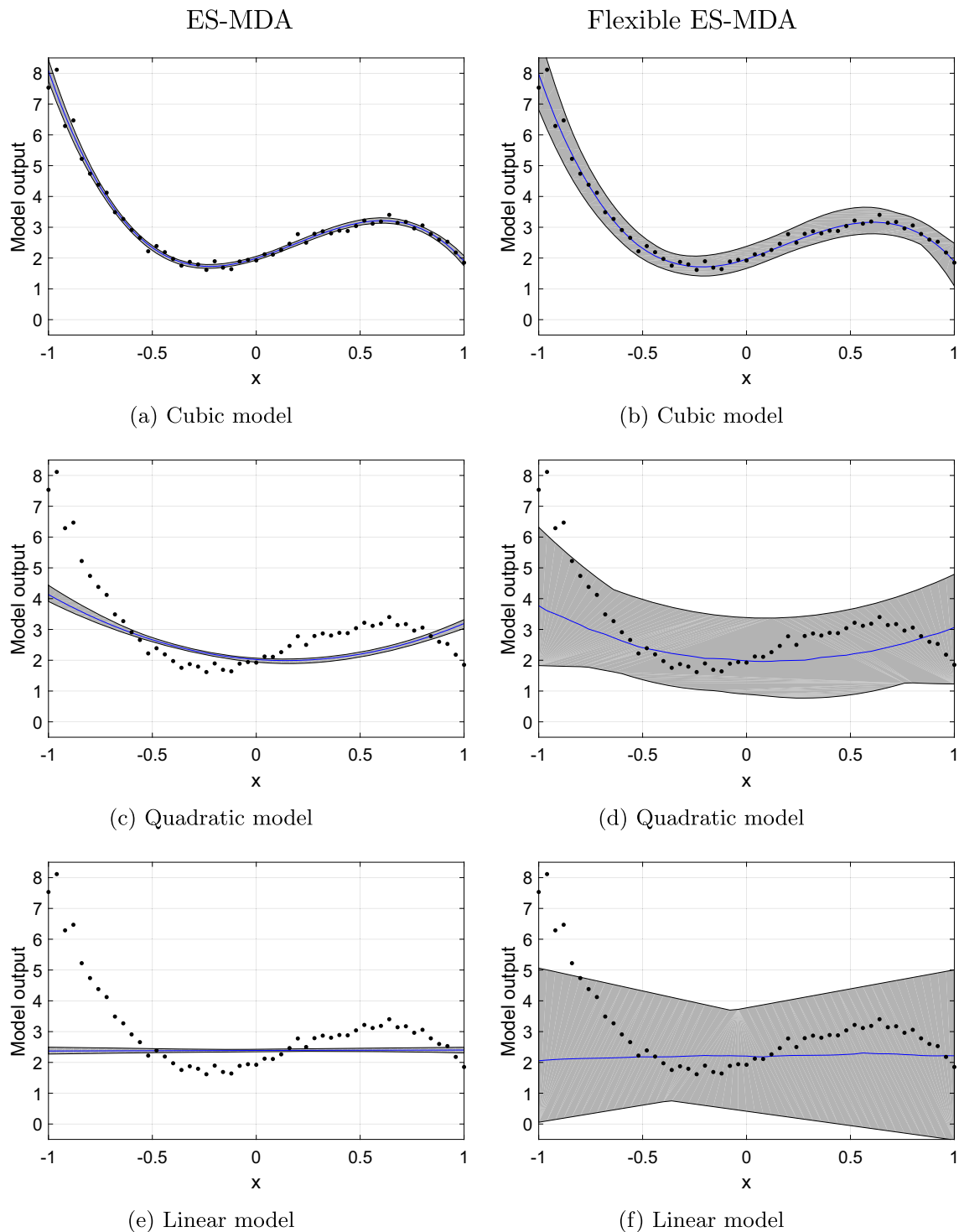

(a) Calibration of linear model   (b) Calibration of quadratic model

Fig. 2 Comparison between splits obtained from actual model error and proposed algorithm during calibration

ES-MDA

Flexible ES-MDA

(a) Cubic model

(b) Cubic model

(c) Quadratic model

(d) Quadratic model
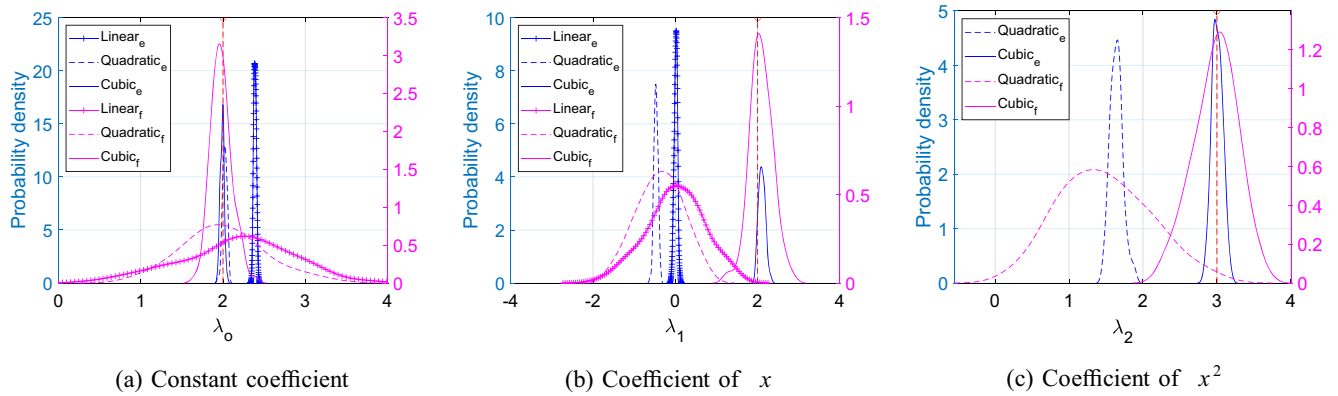
(e) Linear model

(f) Linear model

Fig. 3 Posterior distribution of models (cubic, quadratic and linear) outputs for test case 1 (with measurement error of 5%). The descriptions of lines and colors are same as in Fig. 1

and f), where the results from the Flexible ES-MDA are clearly better than the standard ES-MDA in quantifying the unknown model error uncertainty along in the presence of measurement errors.

Figure 4 shows the posterior distribution of parameters for the case of 5% measurements errors. In this figure, blue and magenta colors show the posterior distribution obtained from ES-MDA and Flexible ES-MDA, respectively.

(a) Constant coefficient (b) Coefficient of x (c) Coefficient of x^2

Fig. 4 Posterior distribution of models (cubic, quadratic and linear) parameters for test case 1 (with measurement error of 5%). Red dashed lines with small circles on ends show the reference model parameters, which are used to generate observations. In figure legends, subscripts

"e" and "f" show the calibration using ES-MDA and Flexible ES-MDA respectively. Blue and magenta show the y-axes correspond to PDF values of posterior distribution obtained using ES-MDA and Flexible ES-MDA respectively

ES-MDA results in a biased posterior distribution of the model parameters as shown in Fig. 4a, b and c except for the perfect model case and constant coefficient of the quadratic model. Flexible ES-MDA reduces model bias in parameter estimation of linear and quadratic models by capturing the model error effects and covers the reference solution as shown in Fig. 4a, b and c except for coefficient of x, (i.e., λ_1) of the quadratic model.

We use the prediction interval coverage probability (PICP) of posterior distributions as one of the calibration metric. PICP is estimated by counting the number of observations in the confidence intervals (10% to 99%) of the posterior distribution normalized by the total number of observations [29]. PICP is a very useful metric for quantifying both under-estimation and/or over-estimation of uncertainties in the obtained posterior distributions [29]. The mathematical description of PICP is shown in Appendix 1. PICP values close to 45° line (dashed red line in Fig. 5) indicates a perfect posterior distribution. Figure 5 shows the PICP of the prior and posterior distribution of the linear, quadratic and cubic models outputs. We observe

that the posterior distribution of cubic (perfect) model output obtained by the standard ES-MDA underestimates the uncertainty. However, the proposed Flexible ES-MDA shows more robust uncertainty quantification as shown in Fig. 5a. This effect is due to the efficient coverage of the noisy features of the observation data by the proposed algorithm (Fig. 3b). PICP of posterior distribution obtained from ES-MDA for the imperfect linear and quadratic models show severe under-estimation of uncertainties (Fig. 5b and c) as the model error uncertainties are missing in the formulation. Further, we observe a more robust PICP values obtained from the proposed algorithm for calibrating the imperfect linear and quadratic models as shown in Fig. 5b and c.

Mean continuous ranked probability score (CRPS) is one of the most useful calibration and prediction metrics to evaluate the precision and accuracy of the posterior ensemble. Less accurate or poor quality results are indicated by the higher values of CRPS. The complete detail of CRPS for ensemble prediction system can be found in [11]. The mathematical descriptions of mean CRPS along with

(a) PICP of Cubic model (b) PICP of Quadratic model (c) PICP of Linear model
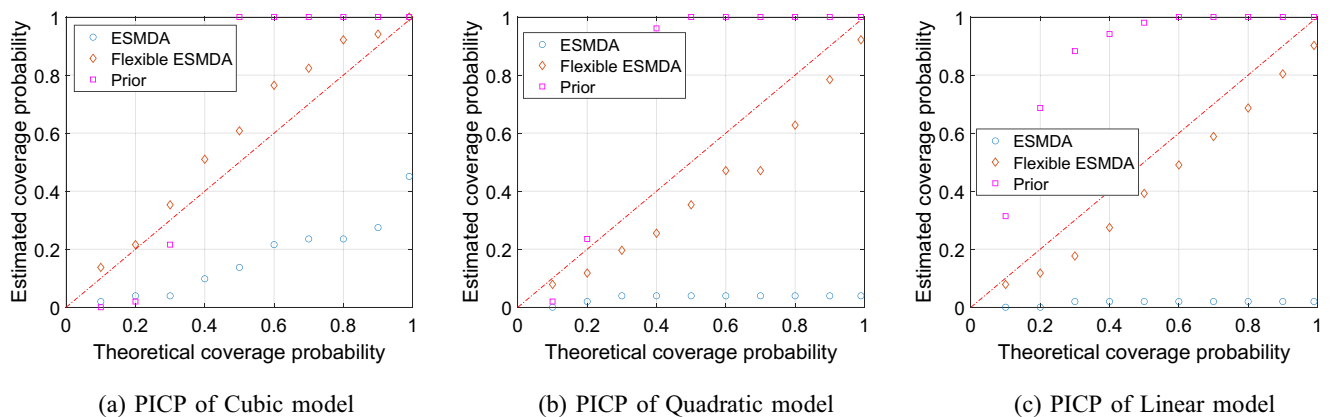
Fig. 5 Prior and posterior PICP of linear, quadratic and cubic models (with measurement error of 5%)

Mean square error (MSE) are presented in Appendix 1. Figure 6 shows the comparison of mean CRPS and MSE of the posterior ensemble of models outputs obtained by the ES-MDA and the proposed algorithm. Calibration results from the proposed algorithm show lower CRPS values for linear, quadratic and cubic models (Fig. 6a), which indicates more reliable results in terms of precision and accuracy. In Fig. 6b, MSE shows more wide distribution of posterior ensemble due to the coverage of unknown model error uncertainty using Flexible ES-MDA; however, lower values of MSE is also observed using the proposed algorithm due to significant reduction in model bias.

3.2 Test case 2: the simple machine

In this test case, the simple machine model is calibrated with the observation data obtained from the unknown complex machine. The main objective is to test the flexibility of proposed algorithm for the cases, where the model lacks some physics. The imperfect model for simple machine is shown in Eq. 26.

$$y = \theta x, \tag{26}$$

where y is the work obtained from machine, x is the effort on machine and θ is the efficiency of machine. True complex machine $T(x)$ includes the friction effect, which is unknown. The observed data d_{obs} is generated using the function of true complex machine using:

$$T(x) = \frac{\theta x}{1 + x/a}, \tag{27}$$

$$d_{obs} = T(x) + \epsilon_d, \tag{28}$$

where $a = 20$ and ϵ_d is the measurement error, which is taken as 5% of true function. In this test case, the task is to estimate the efficiency of the machine θ and the corresponding output of the calibrated model. The prior distribution for the model parameters is assumed to follow a standard normal distribution, i.e., mean $\mu = 0$ and standard deviation $\sigma = 1$. The domain of effort variable x lies within the interval [0, 6]. Total number of points are 61, i.e., x contains values from 0 to 6 with the difference of 0.1 between two consecutive points, where 40% are used for calibration/parameter estimation purpose and 60% are used for predictions.

In this test case, an ensemble of 100 members is used with 8 number of iterations for the calibration of simple machine model. Figure 7 shows the posterior distribution of efficiency (parameter) of the machine and model output obtained by ES-MDA and Flexible ES-MDA algorithms. Both algorithms match the historical/training data; however, predictions obtained from the calibrated model using ES-MDA are overconfident, inaccurate and generally unreliable (Fig. 7a). This is due to the biased posterior distribution of efficiency (parameter) of machine as shown in Fig. 7d. This bias in the estimated efficiency parameter of the machine is reduced significantly and posterior distribution covers the true efficiency of the machine using proposed Flexible ES-MDA (Fig. 7d). Due to this effect more reliable predictions are obtained from the calibrated simple machine model using the proposed Flexible ES-MDA algorithm (Fig. 7b).

Figure 8 shows the PICP of training data and predictions. The proposed algorithm shows more robust uncertainty quantification of training data as shown in Fig. 8a. The prediction PICP is also improved by the Flexible ES-MDA


| (a) Mean CRPS Comparison | (b) MSE Comparison |
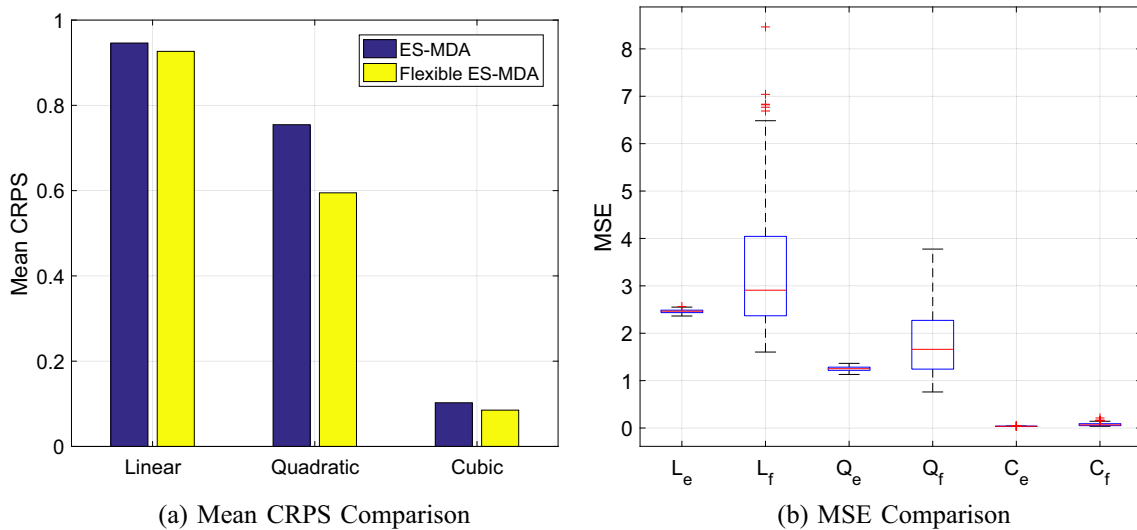
Fig. 6 Mean CRPS and MSE of posterior ensemble for test case 1 (with measurement error of 5%). In part (**b**), "L," "Q" and "C" show MSE from linear, quadratic and cubic models respectively. Subscripts "e" and "f" show the calibration using ES-MDA and Flexible ES-MDA respectively

(a) ES-MDA

(b) Flexible ES-MDA

(c) Prior distribution

(d) Posterior distribution
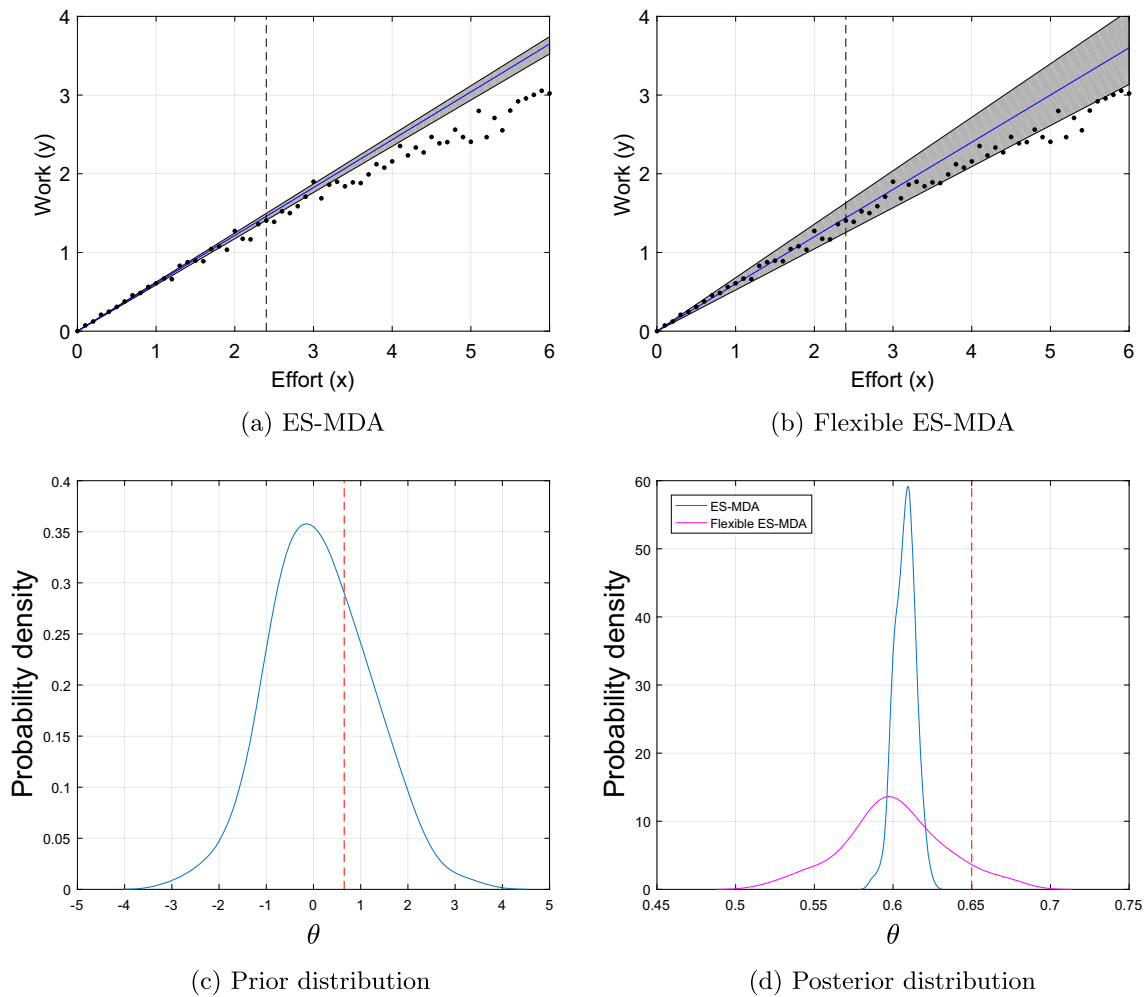
Fig. 7 Parameter estimation, calibration and prediction results for test case 2. Parts **a** and **b** show the posterior distribution of model output. In parts **a** and **b** blue lines show $p50$ percentiles, gray shaded areas show 99% confidence intervals, dashed black line show the end of historical/training (i.e., use for parameter estimation) data and solid black dots show observation data. Parts **c** and **d** show prior and posterior distribution of model parameter. In parts **c** and **d** red dashed line show the true efficiency of the machine.

(a) PICP of training/historical data

(b) PICP of prediction data
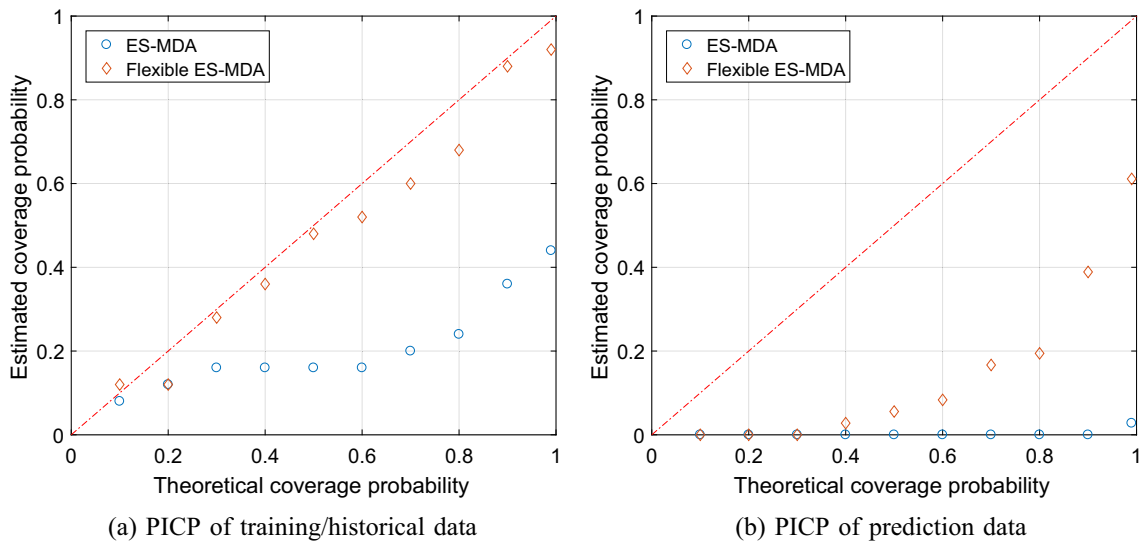
Fig. 8 Posterior PICP of training/historical data and prediction from simple machine

(Fig. 8b), which shows the increase in predictions reliability from calibrated imperfect models (which lacks some physics). Similar results are observed in terms of mean CRPS and MSE metrics for both historical/training and prediction data as shown in Fig. 9.

3.3 Test case 3: the imperfect reservoir model

This test case is related to the calibration of imperfect model of subsurface oil reservoir with channelized geological patterns of log-permeability field. The subsurface reservoir has dimensions of 7500 ft $\times$ 7500 ft $\times$ 20 ft in the x, y and z directions, respectively. The reservoir contains oil and water phases with in-compressible flow in porous media. The reservoir has uniform porosity of 20% and the initial reservoir pressure is 5000 psi. Figure 10 shows the true model, which consists of 75 $\times$ 75 grid blocks, along with different wells open/shut schedule (Fig. 10b) and controls (Fig. 10c). The true model is used to generate the observed data and the observations are perturbed by an additive measurement noise, which is taken as 5% of true model response. The reservoir contains one injector well (I1) and three production wells (P1, P2, P3). The production wells are operated under constant bottom hole pressure constraint of 4500 psi and the injector well is operated with time-varying constraint of constant injection rate as shown in Fig. 10c. The relative permeability is represented by Corey's power law model, which is described along with parameter values and fluid properties in Appendix 2. The capillary pressure and gravitational effects are neglected.

The reservoir is simulated using a 2-D grid with the MATLAB Reservoir Simulation Toolbox (MRST) [15]. Well P3 is used in the prediction period in order to assess

the prediction capabilities of the calibrated reservoir model for future/new wells. The reference and prior fine scale channelized log-permeability fields are generated using a two facies training image as an input to a direct sampling version of MPS algorithm [19]. The calibrated reservoir model is an up-scaled version of true model with a size of 15 $\times$ 15 grid block with no parameterization of the geological features (parameters corresponds to grid block values). The up-scaled model (a.k.a. imperfect reservoir model) contains two sources of modeling errors, simplified geological representation and up-scaling errors. Figure 10a shows the log-permeability field of true model with the channelized features and Fig. 11a shows the corresponding up-scaled log-permeability field where the channelized features are blurred due to harmonic up-scaling of the grid properties. Except for permeability, all other inputs to the reservoir simulator take the same value for both the reference fine model and the up-scaled models used in history matching, for example relative permeability and porosity.

In this test case, an ensemble of 100 members (i.e., one hundred realizations of log-permeability field) is used with 8 number of iterations for the calibration of imperfect reservoir model. Figure 11 shows the mean of the posterior distribution of the log-permeability field obtained by standard ES-MDA and the proposed Flexible ES-MDA algorithms. For the standard ES-MDA, we observe over-shooting (red and blue values) in the mean log-permeability field at a large number of grid blocks. In this case, ES-MDA aggressively tried to adjust the model parameters to match the observation while neglecting the model errors. For the Flexible ES-MDA, we obtained relatively smooth log-permeability fields as shown in Fig. 11c, as the algorithm

(a) Mean CRPS comparison

(b) MSE comparison
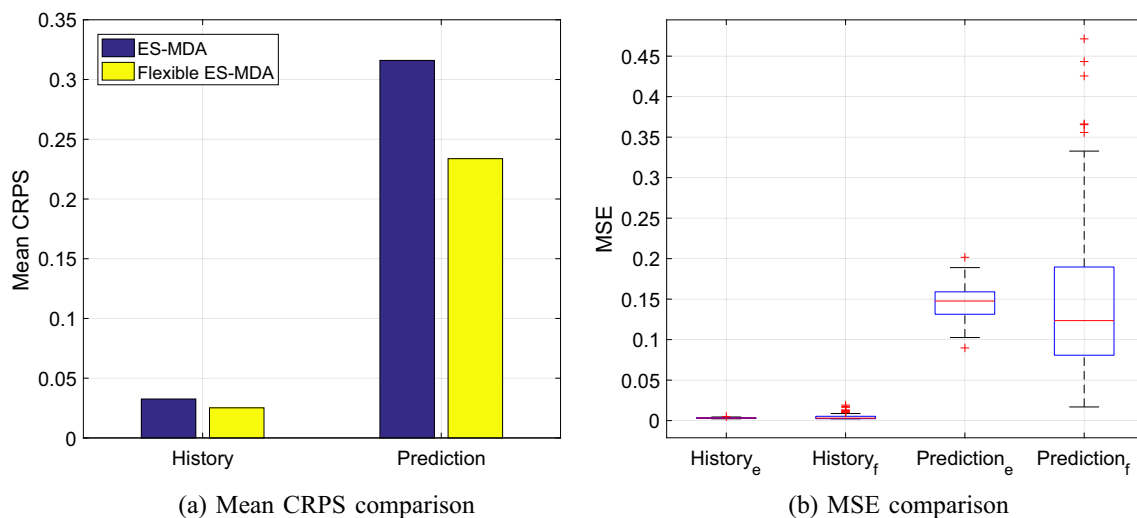
Fig. 9 Mean CRPS and MSE of posterior ensemble for historical and prediction data of test case 2. In part (b) subscripts "e" and "f" show the calibration from ES-MDA and Flexible ES-MDA respectively

(a) Reference fine   (b) Well schedule   (c) Injection well control
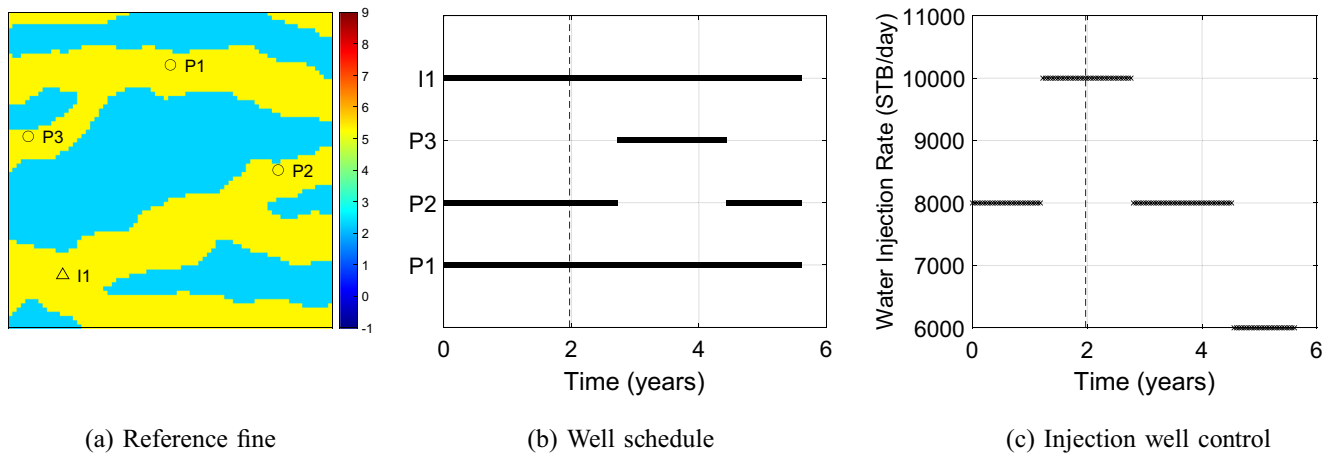
Fig. 10 The true reservoir model (use to generate observations). Part **a** shows the fine scale (75 × 75) reference log-permeability with channelized features. Part **b** shows wells open/shut schedule. In part **b** solid back lines indicate the time periods when a well is open to flow. Part **c** shows water injection rate of injector well I1. Vertical dashed black lines show the end of the historical period in part (**b**) and (**c**)

only updates the model parameters using a percentage of the data mismatch defined by the split parameter s_p to account for unknown model error effects. Figure 12 shows the standard deviation of log-permeability field obtained using both the standard ES-MDA and the proposed Flexible ES-MDA algorithms. We observe relatively high values of standard deviation from Flexible ES-MDA (Fig. 12b) as compared to ES-MDA (Fig. 12a), due to the effect of additional uncertainty of unknown model error.

Figure 13 shows the prior and posterior distribution of oil rates from the production wells. ES-MDA performs well in terms of matching the observations as shown in Fig. 13a; however, poor-quality predictions are observed for wells P2 and P3. For well P1, the predictions start to deviate after 4.2 years. These overconfident and inaccurate predictions are due to the over-shooting of log-permeability values around

production wells (Fig. 11b). This is a common problem in the petroleum industry, where the historical data is usually matched and often the calibrated models suffers from severe predictability problem [2]. One of the primary reasons of this problem is failing to account for model error effects during the history matching/calibration process as evident in these results. In contrast, the proposed Flexible ES-MDA algorithm performs relatively better in terms of predictions quality for production wells P1, P2 and P3 as shown in Fig. 13b. This could be easily attributed to the smooth mean log-permeability fields of posterior ensemble as shown in Fig. 11c.

Similar results are observed in Fig. 14 for the water rates of production wells and injector well I1. However, we observe that the historical data for well I1 is not fully covered by the posterior distribution obtained by

(a) Reference (15x15)   (b) ESMDA   (c) Flexible ESMDA
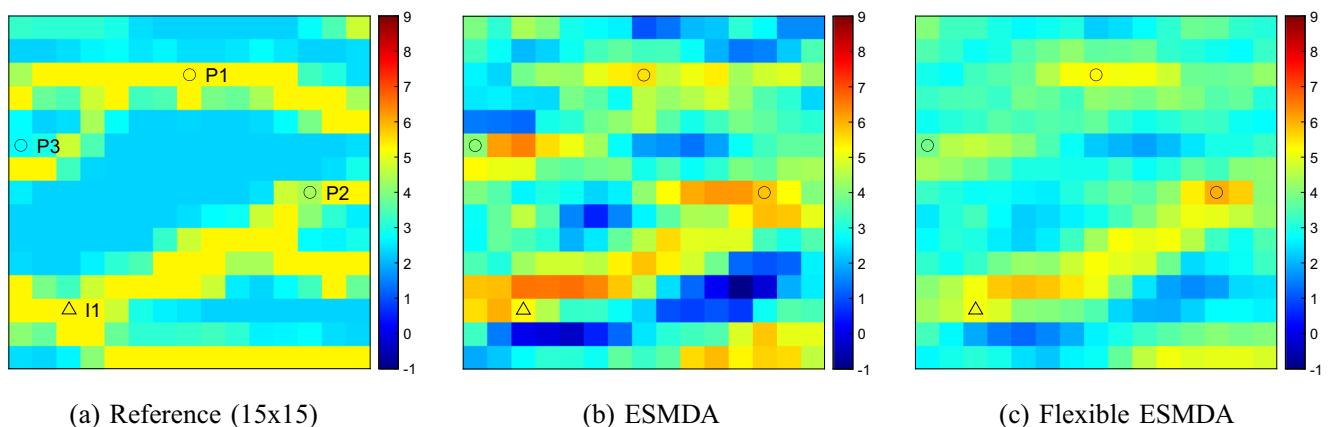
Fig. 11 Reference log-permeability field of the imperfect reservoir model with parameter estimation results. Part **a** shows the reference log-permeability field. Part **b** shows the posterior mean obtained from ES-MDA calibration. Part **c** shows the posterior mean obtained from the proposed Flexible ES-MDA

(a) ESMDA
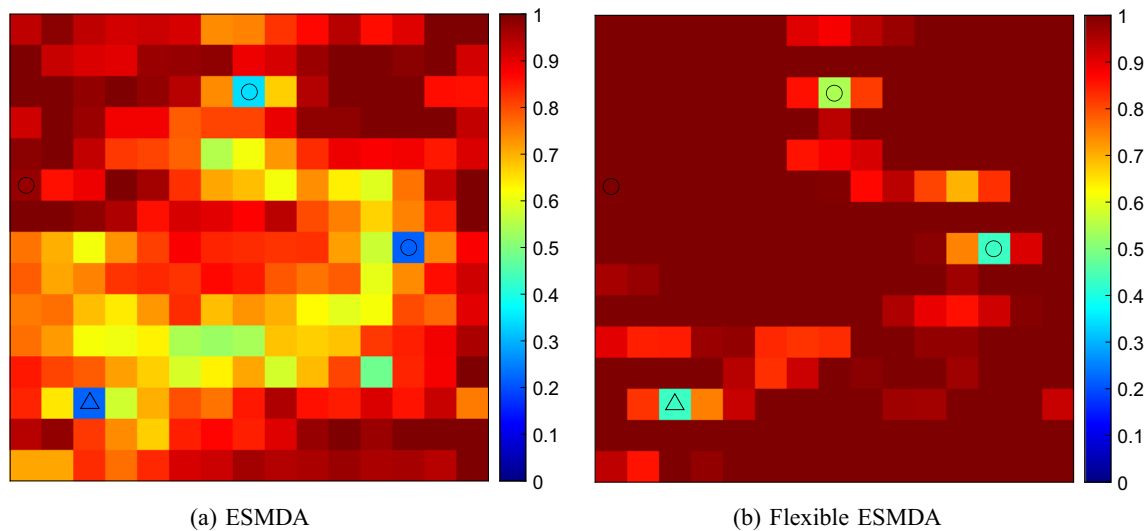
(b) Flexible ESMDA

Fig. 12 Standard deviation of log-permeability field obtained from posterior ensemble

the Flexible ES-MDA as shown in the bottom row of Fig. 14b. This is due to the bad prior distribution (shown by brown lines) of the injection pressure for well I1. We note that this low quality prior is a consequence of modeling errors effect (up-scaling) and this can be avoided by using model/grid refinement around wells that are difficult to match. Figure 15 shows the PICP of parameter estimation, historical data and predictions. The Flexible ES-MDA shows very robust uncertainty estimates for the model parameters (log-permeability field) as the PICP values lie close to the reference line as shown in Fig. 15a. Additionally, the proposed algorithm shows improved uncertainty estimate of historical production data as shown in Fig. 15b, despite failing to match the historical data for well I1. The prediction PICP is also improved by the Flexible ES-MDA (Fig. 15c), which shows an increase in predictions reliability from calibrated imperfect reservoir models.

Figure 16 shows the mean CRPS and MSE metrics for both historical and prediction data. We observe that mean CRPS and MSE of the historical data show lower values for ES-MDA due to relatively better matching and higher values for Flexible ES-MDA due to the uncovered historical data of well I1. However, the Flexible ES-MDA shows significant improvement for mean CRPS and MSE of prediction data. The good matching using ES-MDA can mislead us to overconfident and inaccurate predictions in case of imperfect models and this over-confidence in inaccurate predictions could be avoided using the proposed Flexible ES-MDA.

In cases where model refinement/correction is not feasible, we propose a variant of the Flexible algorithm where the maximum split factor s_p is bounded by the coverage probability of the 99.99% confidence interval.

This results in an adaptive algorithm where the split parameter is adaptively adjusted for each time series using the following simple equation,

$$\text{If: } s_p > CP : \quad \text{then } s_p = CP, \tag{29}$$

where CP is the coverage probability of the 99.99% confidence interval. The only additional modification is the computation of the CP parameter for each time series at every iteration of Flexible ES-MDA algorithm. The mathematical description of CP is presented in Appendix 1. The idea behind the comparison of split parameter with coverage probability that both have same range from [0, 1], and in the case of a time series showing bad prior (failed to cover the data, i.e., $CP = 0$) then the Flexible ES-MDA performs like standard ES-MDA for this output time series during the initial few iterations of the ES-MDA. During later iterations it is expected that the CP becomes non-zero and the split parameter would be adjusted according to the value of CP.

Figure 17 shows the posterior results of log-permeability field using the proposed Flexible ES-MDA with adaptive adjustment of the split parameter with respect to coverage probability. We note that the quality of the mean log-permeability field is slightly decreased (Fig. 17a) as compared to earlier results shown in Fig. 11c without any adaptive adjustment of the split parameter. Figure 18 shows the calibration and prediction results of well outputs obtained from Flexible ES-MDA with adaptive adjustment of the split parameter. In this figure, we only present the results that are affected by the adaptive adjustment of the split parameter. We observe a better data match of the historical data for well I1 (Fig. 18) due to the adaptive adjustment of the split parameter in Flexible ES-MDA as compared to without any adjustment of split parameter
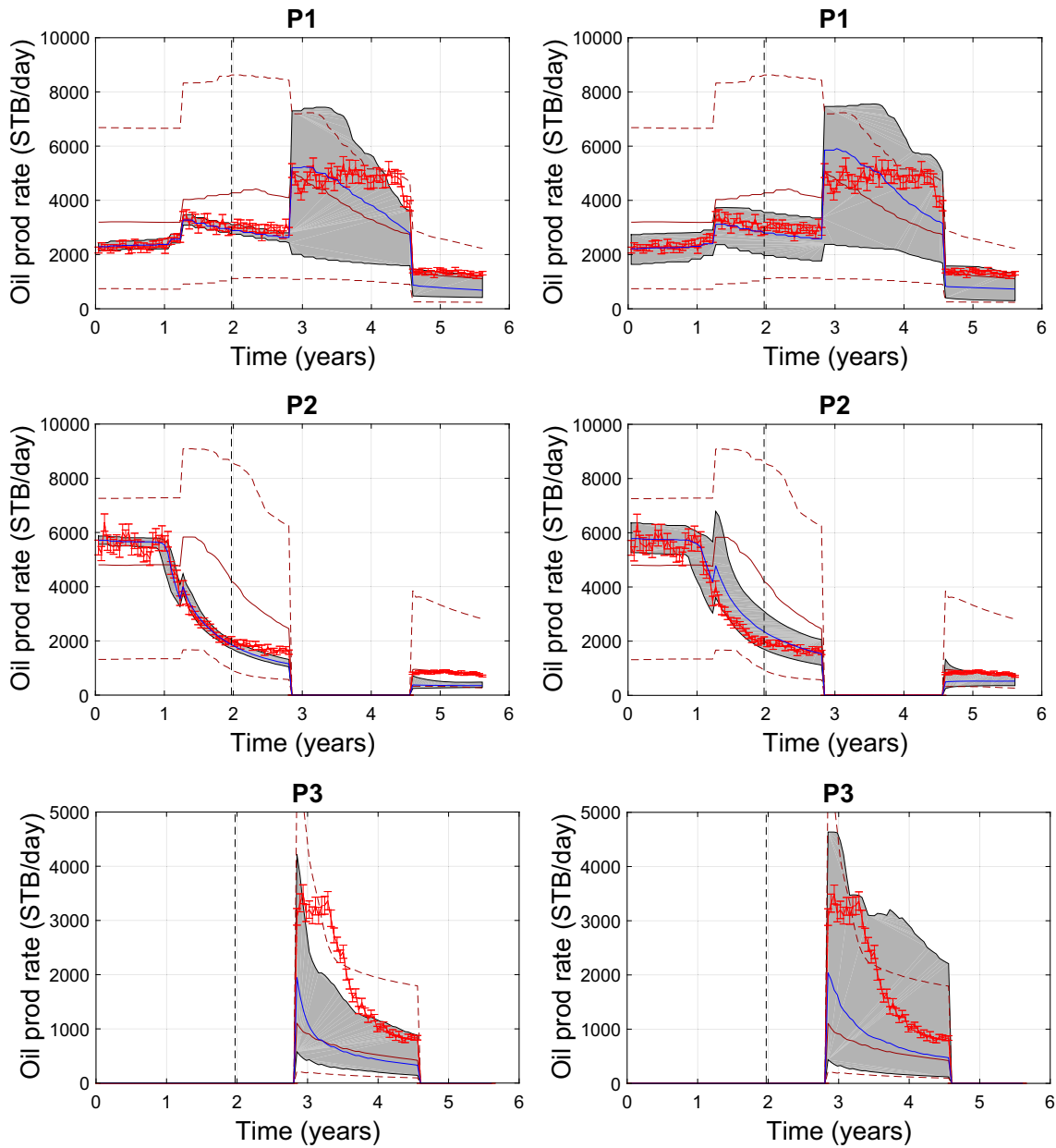
Fig. 13 Prior and posterior of oil production data for production wells P1, P2 and P3. Red lines show observed data and bars on red lines show measurement error. Vertical dashed black lines show the end of the historical period. Solid and dashed brown lines show 50th percentile *p*50 and 99% confidence interval of prior distribution respectively. Solid blue lines and gray shaded area show *p*50 and 99% confidence interval of posterior distribution

(Fig. 14b). In addition, we note that the confidence intervals of production wells P2 and P3 are reduced (Fig. 18) due to the relatively lower values of the standard deviation of the posterior ensemble (Fig. 17b) obtained from Flexible ES-MDA with adaptive split parameter as compared to Fig. 12b.

Figure 19 shows the PICP of model parameters, historical data and predictions. We observe that PICP of model parameters is slightly decreased as compared to the results in Fig. 15a. This is due to the adjustment of split parameter in order to match bad prior time series of well I1. However, the PICP of historical data and predictions are improved as compared to Fig. 15b and c using Flexible ES-MDA. This effect shows some trade-off between quality of estimated parameters and corresponding outputs (specially bad prior) of imperfect reservoir model using
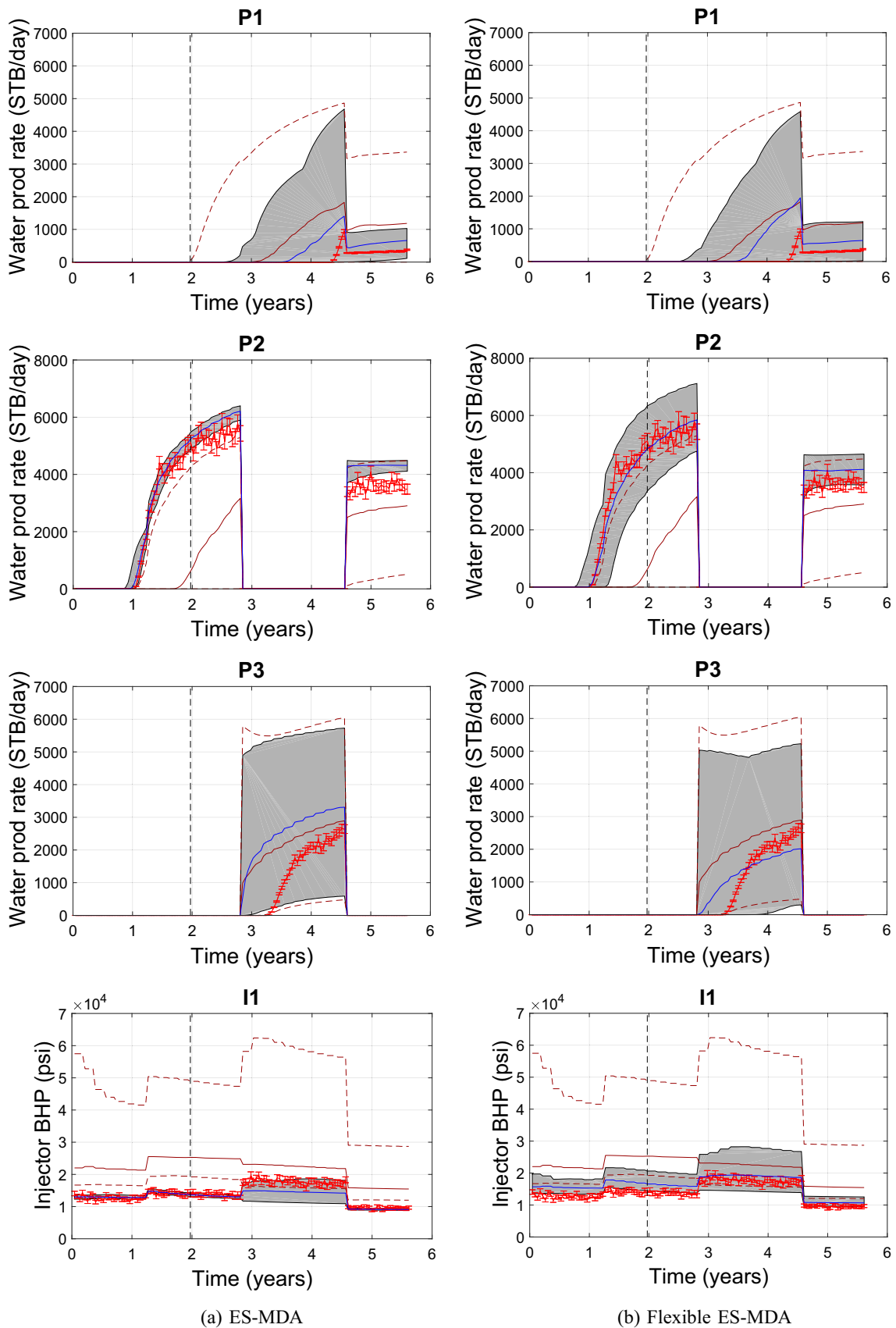
Fig. 14 Prior and posterior of water production and injection pressure data. The descriptions of lines and colors are same as in Fig. 1
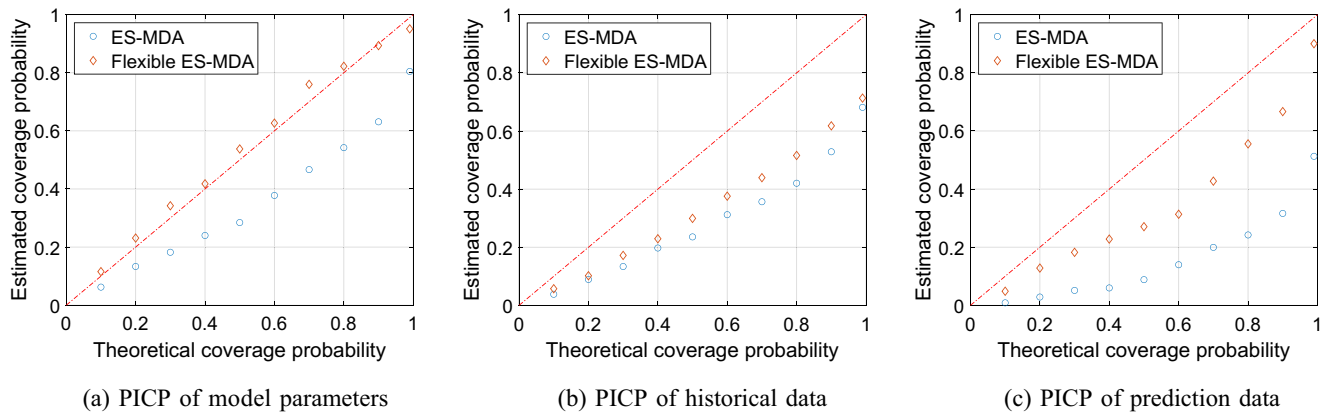
(a) PICP of model parameters   (b) PICP of historical data   (c) PICP of prediction data

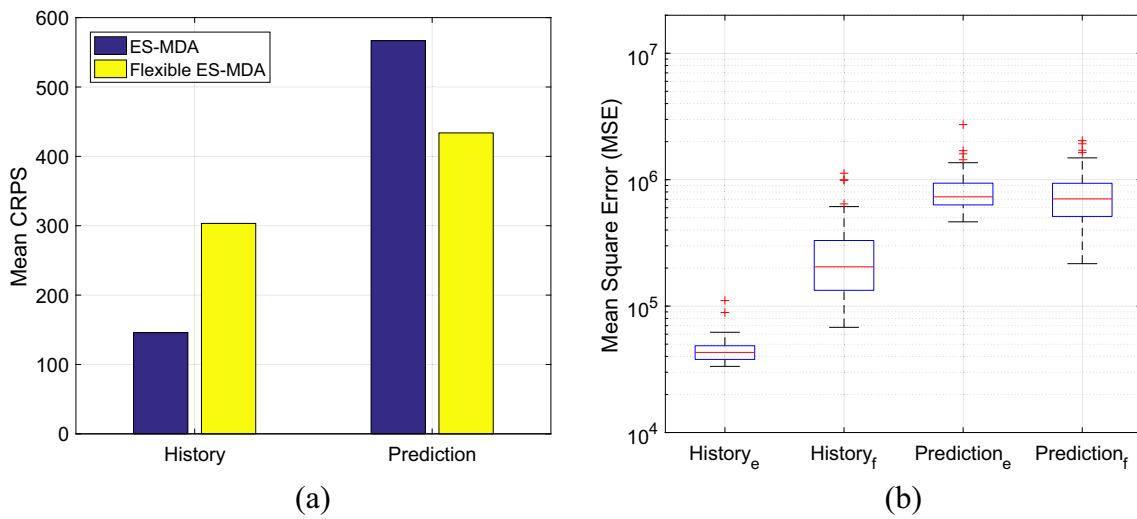**Fig. 15** Posterior PICP of model parameters, historical and prediction data of test case 3

Fig. 16 Mean CRPS and MSE of posterior ensemble for historical and prediction data of test case 3. In part (b) subscripts "e" and "f" show the MSE obtained from ES-MDA and Flexible ES-MDA respectively

(a) Mean   (b) Standard deviation
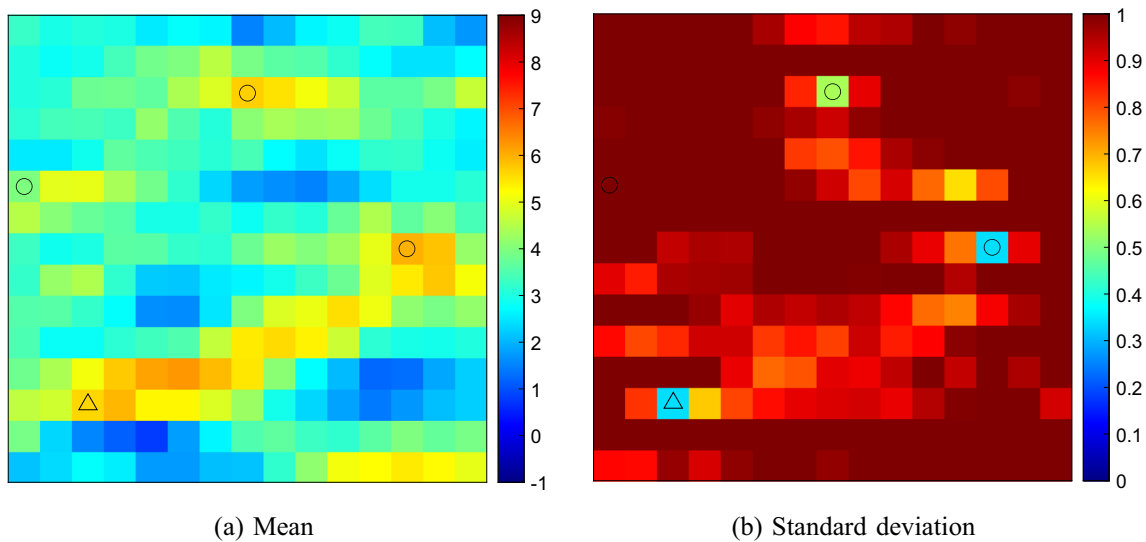
Fig. 17 Mean and standard deviation of log-permeability field of posterior ensemble obtained from Flexible ES-MDA with adaptive adjustment of split parameter
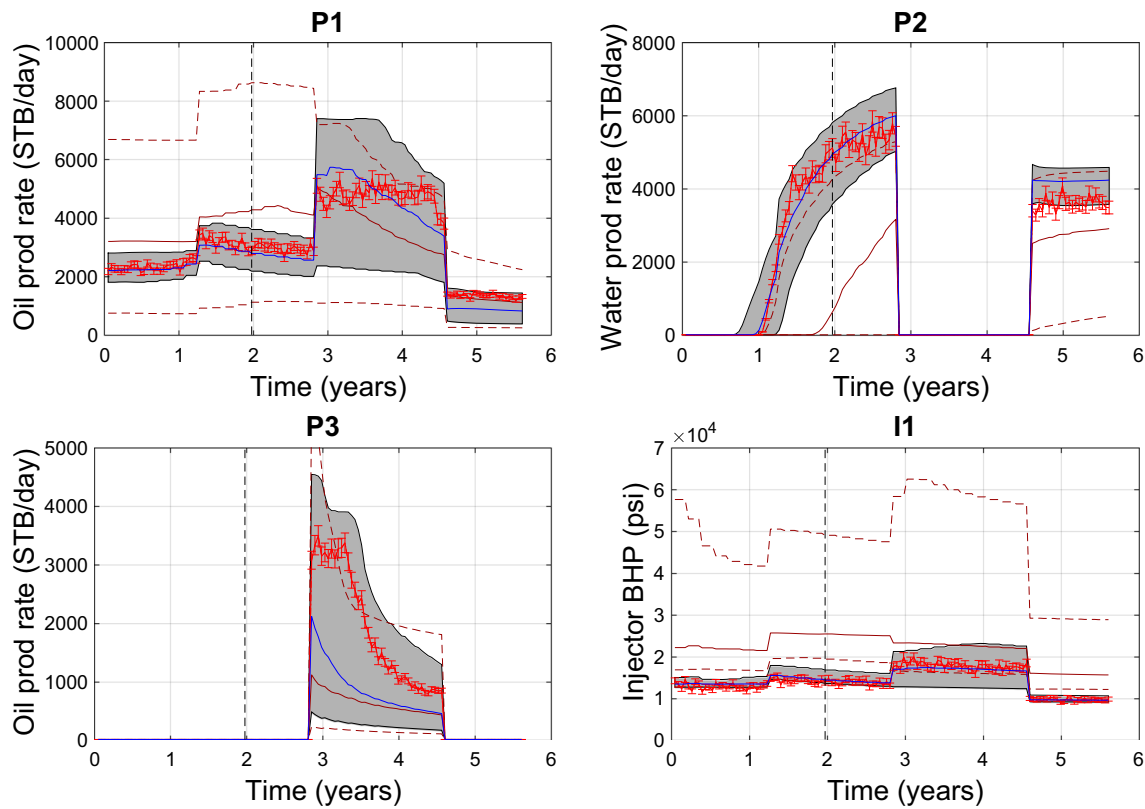
Fig. 18 Calibration and prediction results of production and injection data using Flexible ES-MDA with adaptive adjustment of split parameter. The descriptions of lines and colors are same as in Fig. 13
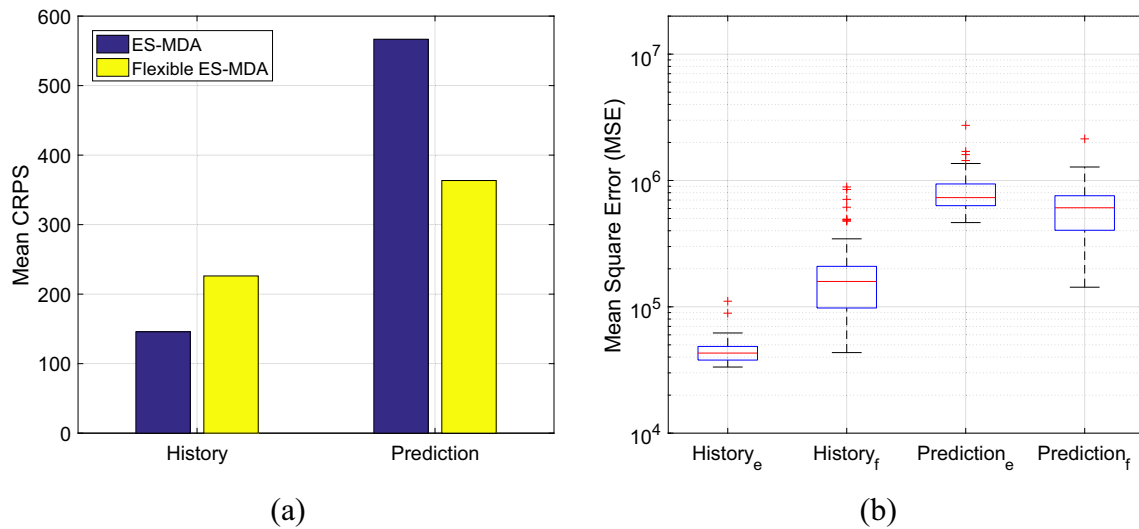
adaptive adjustment of the split parameter in the proposed algorithm.

Figure 20 shows the mean CRPS and MSE metrics for both historical and prediction data with adaptive adjustment of split parameter. We observe that mean CRPS and MSE of the historical and prediction data are improved using Flexible ES-MDA as compared to Fig. 16 due to matching

of historical data of well I1 and reduction in confidence intervals of water production of well P2 and oil production of well P3 by adaptive adjustment of split parameter. These results show that the adaptive adjustment of split parameter in Flexible ES-MDA can be used as an alternative option if model refinement is not feasible and output time series cannot be matched due to bad prior.

(a) PICP of model parameters   (b) PICP of historical data   (c) PICP of prediction data
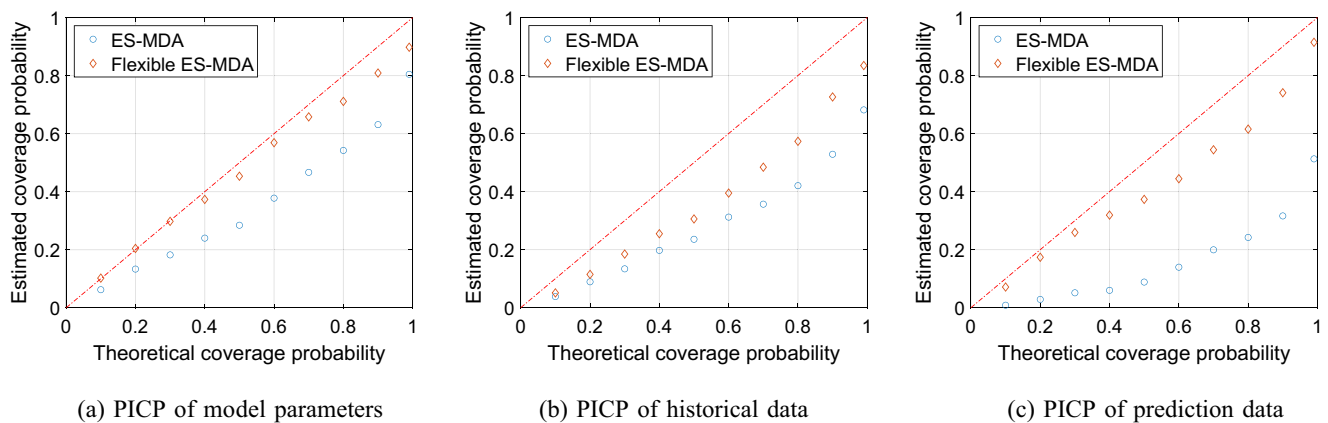
Fig. 19 Posterior PICP of model parameters, historical and prediction data of test case 3 with adaptive adjustment of split parameter in Flexible ES-MDA

Fig. 20 Mean CRPS and MSE of posterior ensemble for historical and prediction data of test case 3 with adaptive adjustment of split parameter in Flexible ES-MDA. In part (b) subscripts "e" and "f" show the MSE obtained from ES-MDA and Flexible ES-MDA respectively

4 Conclusions

In this paper, a flexible algorithm is proposed for calibration of perfect and imperfect models. This flexible algorithm builds on the ensemble smoother with multiple data assimilation, which has the assumption that models are perfect, i.e., accurate representation of the real systems. However, it is widely known that mathematical models are approximations of the real systems and some inherent errors always exist in computational models. Neglecting the model error during calibration causes bias in the estimated physical parameter and often results in unreliable predictions. In the proposed algorithm, the residual (data mismatch) is split into two parts. One part is used for parameter estimation/update and the second part is used to represent the model error. The initial split parameter is computed based on the ratio of norm of mean residual (mean deviation from observed data) and norm of maximum residual (maximum absolute deviation from observed data). During calibration (data assimilation) this split parameter is updated based on the ratio of norm of mean residuals obtained in current and previous iterations. In the proposed algorithm, we assumed that the data misfit/residual can be split into a parameter error and a model error according to the ratio of mean deviations between iterations. This split formulation shows very close correspondence to the best split value computed from true model error (see Test case 1).

In this work, three test cases have been used to observe the performance of the proposed algorithm. These test cases are related to the calibration of polynomial functions, simple machine and imperfect reservoir model.

In the first test case, cubic function is considered as perfect model and imperfect models are represented by quadratic and linear functions. We observe that if the model is perfect, the proposed algorithm exactly match the data and for imperfect models the algorithm has the flexibility to capture the unknown model error, which is very useful to avoid over-confidence and generally inaccurate predictions. The second test case is related to estimation the efficiency parameter of a simple machine model which lacks physics in terms of friction component. The calibration result from proposed algorithm shows more reliable estimation of the efficiency of the machine which provides significant improvement in prediction. The third test case is related to calibration of imperfect reservoir model, which has blurred channelized geological patterns. The imperfect reservoir model has two sources of modeling errors, simplified geological representation and up-scaling errors. We note that the proposed algorithm reduces the model bias in parameter estimation and as a consequence predictions are improved significantly from the calibrated imperfect models. We observe improvements in PICP, mean CRPS and MSE metrics after evaluating the parameter estimation, calibration and prediction performance of the proposed algorithm for three test cases of different physical systems. However, the well output which shows bad prior due to the model error effect can be difficult to match using the proposed algorithm. This can be avoided by adaptive adjustment of the split parameter in the proposed algorithm with some trade-off between quality of estimated parameters and corresponding model outputs.

We strongly recommend model improvement in terms of physics, assumptions, details and description, if a

large magnitude of model error is indicated by the proposed algorithm. However, we argue that the proposed algorithm provides good indicators about the reliability of the calibrated models especially for the cases with unknown model error. Sometimes for effective decision-making we need to reduce the uncertainty to threshold value (i.e., minimum value which effects decision-making process). If the model error uncertainty is greater than the threshold value, we strongly recommend to improve the model in order to reduce model error uncertainty. In cases where model error uncertainty is less than acceptable threshold value and there is a time-constraint, the proposed algorithm provides a robust method for calibrating imperfect models that can be used for decision support. In future, we would try to address different types of uncertain parameters and simulation models with different forms of model errors in terms of complex physics, structural geology, fluid dynamics and phase behavior.

Appendix 1: Calibration and prediction metrics

A.1 Prediction interval coverage probability (PICP)

The mathematical description of coverage probability (CP) is shown below,

$$CP = \frac{N_{CI}}{N_t}. \tag{A1.1}$$

N_{CI} = Number of observations or parameters in confidence interval

N_t = Total number of observations or parameters

Prediction interval coverage probability is obtained by computing coverage probability for confidence interval 10%, 20%, ..., 90%, 99%.

A.2 Mean square error (MSE)

The mathematical description of mean square error (MSE) is shown below,

$$MSE = \frac{1}{N_d} \sum_{n=1}^{N_d} (d_n - d_{obs,n})^2, \tag{A1.2}$$

where n is index of observation or model prediction at corresponding time.

A.3 Continuous ranked probability score (CRPS)

Mathematically, CRPS can be described as follows,

$$CRPS = \int_{-\infty}^{\infty} [p(x) - H(x - x_{obs})]^2 dx, \tag{A1.3}$$

where $p(x) = \int_{-\infty}^{x} \rho(y) dy$ Cumulative distribution of quantity of interest, $H(x - x_{obs})$ = Heaviside function (Step function), i.e.,

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases}$$

For an ensemble system with N_e realizations, the CRPS can be written as follows,

$$CRPS = \sum_{r=0}^{N_e} c_r. \tag{A1.4}$$

$$c_r = \alpha_r p_r^2 + \beta_r (1 - p_r)^2. \tag{A1.5}$$

where $p_r = P(x) = r/N_e$, for $x_r < x < x_{r+1}$ (cumulative distribution is a piece wise constant function).

$$\alpha_r = \begin{cases} 0 & \text{if } x_{obs} < x_r \\ x_{obs} - x_r & \text{if } x_r < x_{obs} < x_{r+1} \\ x_{r+1} - x_r & \text{if } x_{obs} > x_{r+1} \\ x_{obs} - x_{N_e} & \text{if } x_{obs} > x_{N_e} \\ 0 & \text{if } x_{obs} < x_1 \end{cases}$$

$$\beta_r = \begin{cases} x_{r+1} - x_r & \text{if } x_{obs} < x_r \\ x_{r+1} - x_{obs} & \text{if } x_r < x_{obs} < x_{r+1} \\ 0 & \text{if } x_{obs} > x_{r+1} \\ 0 & \text{if } x_{obs} > x_{N_e} \\ x_1 - x_{obs} & \text{if } x_{obs} < x_1 \end{cases}$$

Appendix 2: Reservoir properties

Corey model in form of power law is used to generate relative permeability data for the reservoir model. The mathematically description of Corey model in form of

Table 1 Reservoir fluid data and Corey relative permeability model parameters

Fluid properties		Corey relative permeability model parameters			
Water viscosity	0.5 cp	S_{or}	0.2	k_o^0	1
Oil viscosity	1 cp	S_{wc}	0.2	k_w^0	1
Water density	1000 kg/m^3	n_w	2		
Oil density	700 kg/m^3	n_o	2		

power law is shown below.

$$k_{rw} = (\hat{S_w})^{n_w} k_w^0. \tag{A2.1}$$

$$k_{ro} = (1 - \hat{S_w})^{n_o} k_o^0. \tag{A2.2}$$

$$\hat{S_w} = \frac{S_w - S_{wc}}{1 - S_{or} - S_{wc}}. \tag{A2.3}$$

The notations of above equations are described in MRST manual [15]. The fluid data and Corey relative permeability model parameters used in the reservoir model are shown in Table 1.

References

1. Brynjarsdottir, J., O' Hagan, A.: Learning about physical parameters: The importance of model discrepancy. Inverse Probl. **30**(11), 114007 (2014)
2. Carter, J.N., Ballester, P.J., Tavassoli, Z., King, P.R.: Our calibrated model has poor predictive value: An example from the petroleum industry. Reliab. Eng. Syst. Safety **91**(10), 1373–1381 (2006)
3. Chen, Y., Oliver, D.S.: Levenberg–Marquardt forms of the iterative ensemble smoother for efficient history matching and uncertainty quantification. Comput. Geosci. **17**(4), 689–703 (2013)
4. Emerick, A.A., Reynolds, A.C.: History matching time-lapse seismic data using the ensemble Kalman filter with multiple data assimilations. Comput. Geosci. **16**(3), 639–659 (2012)
5. Emerick, A.A., Reynolds, A.C.: Ensemble smoother with multiple data assimilation. Comput Geosci **55**, 3–15 (2013)
6. Evensen, G.: Data Assimilation: The Ensemble Kalman Filter. Springer Science & Business Media, New York (2009)
7. Evin, G., Thyer, M., Kavetski, D., McInerney, D., Kuczera, G.: Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour. Res. **50**(3), 2350–2375 (2014)
8. Giudice, D.D., Honti, M., Scheidegger, A., Albert, C., Reichert, P., Rieckermann, J.: Improving uncertainty estimation in urban hydrological modeling by statistically describing bias. Hydrol. Earth Syst. Sci. **17**(10), 4209–4225 (2013). https://doi.org/10.5194/hess-17-4209-2013. https://www.hydrol-earth-syst-sci.net/17/4209/2013/
9. Giudice, D.D., Löwe, R., Madsen, H., Mikkelsen, P.S., Rieckermann, J.: Comparison of two stochastic techniques for reliable urban runoff prediction by modeling systematic errors. Water Resour. Res. **51**(7), 5004–5022 (2015). https://doi.org/10.1002/2014WR016678
10. Hansen, T.M., Cordua, K.S., Jacobsen, B.H., Mosegaard, K.: Accounting for imperfect forward modeling in geophysical inverse problems—exemplified for crosshole tomography. Geophysics **79**(3), H1–H21 (2014). https://doi.org/10.1190/geo2013-0215.1
11. Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast. **15**(5), 559–570 (2000)
12. Kennedy, M.C., O'Hagan, A.: Bayesian calibration of computer models. J. R. Stat. Soc. Series B Stat. Methodol. **63**(3), 425–464 (2001). https://doi.org/10.1111/1467-9868.00294
13. Köpke, C., Irving, J., Elsheikh, A.H.: Accounting for model error in Bayesian solutions to hydrogeophysical inverse problems using a local basis approach. Adv. Water Resour. **116**, 195–207 (2018)
14. Köpke, C., Elsheikh, A.H., Irving, J.: Hydrogeophysical parameter estimation using iterative ensemble smoothing and approximate forward solvers. Front. Environ. Sci. **7**, 34 (2019). https://doi.org/10.3389/fenvs.2019.00034
15. Lie, K.-A.: An Introduction to Reservoir Simulation Using MATLAB - User Guide for the Matlab Reservoir Simulation Toolbox (MRST). Sintef ICT, Department of Applied Mathematics. https://www.sintef.no/Projectweb/MRST/Publications (2016)
16. Lødøen, O.P., Omre, H., Durlofsky, L.J., Chen, Y.: Assessment of uncertainty in reservoir production forecasts using upscaled flow models. In: Geostatistics Banff 2004, pp. 713–722. Springer (2005)
17. Lødøen, O.P., Tjelmeland, H.: Bayesian calibration of hydrocarbon reservoir models using an approximate reservoir simulator in the prior specification. Stat. Model. **10**(1), 89–111 (2010)
18. Luo, X., Stordal, A.S., Lorentzen, R.J., Naevdal, G., et al.: Iterative ensemble smoother as an approximate solution to a regularized minimum-average-cost problem: Theory and applications. SPE J. **20**(05), 962–982 (2015)
19. Mariethoz, G., Caers, J.: Multiple-Point Geostatistics: Stochastic Modeling with Training Images. Wiley, New York (2014). https://books.google.co.uk/books?id=Fe7sBQAAQBAJ
20. O' Sullivan, A., Christie, M.: Error models for reducing history match bias. Comput. Geosci. **9**(2-3), 125–153 (2005)
21. Oliver, D.S., Alfonzo, M.: Calibration of imperfect models to biased observations. Comput. Geosci. **22**(1), 145–161 (2018). https://doi.org/10.1007/s10596-017-9678-4
22. Omre, H., Lødøen, O.P., et al.: Improved production forecasts and history matching using approximate fluid-flow simulators. SPE J. **9**(03), 339–351 (2004)
23. Rammay, M.H., Elsheikh, A.H., Chen, Y.: Quantification of prediction uncertainty using imperfect subsurface models with model error estimation. J. Hydrol. **576**, 764–783 (2019). https://doi.org/10.1016/j.jhydrol.2019.02.056. http://www.sciencedirect.com/science/article/pii/S0022169419302136
24. Rammay, M.H., Elsheikh, A.H., Chen, Y.: Robust algorithms for history matching of imperfect subsurface models. SPE J SPE-193838-PA. https://doi.org/10.2118/193838-pa (2020)
25. Sargsyan, K., Najm, H., Ghanem, R.: On the statistical calibration of physical models. Int. J. Chem. Kinet. **47**(4), 246–276 (2015)
26. Schoups, G., Vrugt, J.A.: A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour. Res. 46(10) (2010)
27. Stordal, A.S., Elsheikh, A.H.: Iterative ensemble smoothers in the annealed importance sampling framework. Adv. Water Resour.

86, 231–239 (2015). https://doi.org/10.1016/j.advwatres.2015.09.030. http://www.sciencedirect.com/science/article/pii/S0309170815002377

28. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J. Global Optim. **11**(4), 341–359 (1997)

29. Xu, T., Valocchi, A.J.: Data-driven methods to improve baseflow prediction of a regional groundwater model. Comput. Geosci. **85**, 124–136 (2015)

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.