Advances in Water Resources 166 (2022) 104252

Contents lists available at ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Hydrogeological multiple-point statistics inversion by adaptive sequential Monte Carlo

Macarena Amaya[a,*], Niklas Linde[a], Eric Laloy[b]

[a] Institute of Earth Sciences, University of Lausanne, Lausanne, 1015, Switzerland
[b] Institute for Environment, Health and Safety, Belgian Nuclear Research Center, Mol, 2400, Belgium

ARTICLE INFO

Keywords:
Adaptive sequential Monte Carlo
Particle methods
Sequential geostatistical resampling
Multiple-point statistics
Bayesian inference
Evidence computation

ABSTRACT

For strongly non-linear and high-dimensional inverse problems, Markov chain Monte Carlo (MCMC) methods may fail to properly explore the posterior probability density function (PDF) given a realistic computational budget and are generally poorly amenable to parallelization. Particle methods approximate the posterior PDF using the states and weights of a population of evolving particles and they are very well suited to parallelization. We focus on adaptive sequential Monte Carlo (ASMC), an extension of annealed importance sampling (AIS). In AIS and ASMC, importance sampling is performed over a sequence of intermediate distributions, known as power posteriors, linking the prior to the posterior PDF. The AIS and ASMC algorithms also provide estimates of the evidence (marginal likelihood) as needed for Bayesian model selection, at basically no additional cost. ASMC performs better than AIS as it adaptively tunes the tempering schedule and performs resampling of particles when the variance of the particle weights becomes too large. We consider a challenging synthetic groundwater transport inverse problem with a categorical channelized 2D hydraulic conductivity field defined such that the posterior facies distribution includes two distinct modes. The model proposals are obtained by iteratively re-simulating a fraction of the current model using conditional multiple-point statistics (MPS) simulations. We examine how ASMC explores the posterior PDF and compare with results obtained with parallel tempering (PT), a state-of-the-art MCMC inversion approach that runs multiple interacting chains targeting different power posteriors. For a similar computational budget, ASMC outperforms PT as the ASMC-derived models fit the data better and recover the reference likelihood. Moreover, we show that ASMC partly retrieves both posterior modes, while none of them is recovered by PT. Lastly, we demonstrate how the power posteriors obtained by ASMC can be used to assess the influence of the assumed data errors on the posterior means and variances, as well as on the evidence. We suggest that ASMC can advantageously replace MCMC for solving many challenging inverse problems arising in the field of water resources.

1. Introduction

Markov chain Monte Carlo (MCMC) methods are widely used to tackle probabilistic inverse problems arising in hydrology. As the dimensionality of the parameter space and the non-linearity of the forward problem increases, standard MCMC methods often fail to explore the posterior probability density function (PDF) given realistic computational constraints. This happens as the Markov chains may be trapped in local minima for long times or they may be unable to move between modes of high posterior probability. To circumvent such issues, methods exploring a series of so-called power posteriors have been developed. In a power posterior, less weight is given to the likelihood function as it is raised to the inverse of a temperature that is larger than one, something that is typically referred to as tempering. Tempering-based methods take advantage of the enhanced freedom of exploration at higher temperatures (sampling closer to the prior PDF) as popularized by the widely-used simulated annealing method for global optimization (Kirkpatrick et al., 1983).

Parallel tempering (Earl and Deem, 2005) is a MCMC method in which many interacting chains target different power posteriors. Through proposed swaps of the states between chains, states sampled at higher temperatures can act as model proposals in the chains targeting the posterior distribution, also called unit temperature chains. Analogous to classical MCMC methods, PT approximates the posterior PDF by the states of the unit temperature chains sampled after burn-in. In the context of geophysical inversion, Sambridge (2014) demonstrated that PT can drastically improve sampling efficiency leading to an expanded exploration of the parameter space compared to standard MCMC. The

[*] Corresponding author.
E-mail address: macarena.amaya@unil.ch (M. Amaya).

https://doi.org/10.1016/j.advwatres.2022.104252
Received 4 April 2022; Received in revised form 1 June 2022; Accepted 8 June 2022
Available online 16 June 2022
0309-1708/© 2022 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

PT method has lately been applied in a range of geoscientific problems such as landscape evolution (Chandra et al., 2019), groundwater flow and transport (Laloy et al., 2016; Reuschen et al., 2020) and earthquake source inversion (Gallovič et al., 2019).

A highly parallelizable alternative to MCMC is offered by particle methods such as Annealed Importance Sampling (AIS, Neal (2001)) and Sequential Monte Carlo (SMC, Doucet and Johansen (2011)), where the posterior distribution is approximated using a weighted sample of particle states. In these methods, importance sampling steps are performed sequentially along a sequence of power posteriors. What differentiates these two methods from each other is that SMC performs resampling of the particle population when the variance of the importance weights becomes too high. One outstanding feature of both methods is that the evidence, the normalizing constant in Bayes' theorem and the key quantity in Bayesian model selection, is estimated as well. Compared with its extensive use in science and engineering, SMC appears poorly explored in the Earth sciences (Linde et al., 2017). Zhou et al. (2016) proposed an adaptive version of SMC (ASMC) that automatically tunes the temperature reduction between neighboring power posteriors. Recently, adaptive SMC algorithms were introduced and successfully implemented in geophysical applications for posterior PDF and evidence estimations (Amaya et al., 2021; Davies et al., 2021).

Realistic geological priors can often not be expressed by two-point geostatistical models (e.g., multivariate Gaussian), for example, when connectivity patterns play an essential role in determining the system response (Gómez-Hernández and Wen, 1998; Renard and Allard, 2013). Multiple-point statistics (MPS) is a sub-field of geostatistics aiming at generating conditional model realizations that honor higher-order statistics found in a so-called training image, a gridded 2-D or 3-D representation of the spatial field of interest that is built from generic or previous geological knowledge of the site (Mariethoz and Caers, 2014). To generate MPS-based candidate models within MCMC inversions, one popular approach is sequential geostatistical resampling (SGR) (Ruggeri et al., 2015), in which model proposals are generated by re-simulating a random fraction of the current model conditioned to the remaining grid values. The SGR framework embraces two end-member strategies: either a randomly located boxed-shaped area is resimulated as in sequential Gibbs sampling by Hansen et al. (2012) and in blocking MCMC by Fu and Gómez-Hernández (2009), or random points throughout the model domain are resimulated as in iterative spatial resampling by Mariethoz et al. (2010a). Recently, hybrid methods determining an optimal combination of these end-members have been proposed by Reuschen et al. (2021). Other approaches relying on much faster model proposals are offered, for instance, by graph cuts (Zahner et al., 2016) or by encoding the complex priors in a much lower-dimensional space using deep generative networks, thereby, reducing the number of inferred parameters from several thousands of unknowns to some tenths of latent variables (Laloy et al., 2017, 2018).

In this paper, we consider the challenging groundwater transport inverse problem introduced by Laloy et al. (2016). It consists of a 2-D synthetic tracer experiment in which concentration is monitored at pumping wells and the aim is to recover the hydraulic conductivity field assuming a binary geological media (their case study 2). This test case is particularly challenging for three reasons: (i) the underlying non-linearity caused by long-range connectivity of high-conductivity zones and a conductivity ratio of 100 between permeable channels and less permeable matrix material, (ii) a large number of observations with a high signal-to-noise ratio and (iii) a true field that is designed such that the targeted posterior distribution is bimodal with the modes being located far from each other. Laloy et al. (2016) demonstrated how PT clearly outperforms standard MCMC when used within a SGR framework. Nevertheless, even if PT offered important improvements it did not sample any of the posterior modes and the simulated data of the generated model realizations did not fit the true data to the level of the added noise. Compared to PT, ASMC presents the following advantages: (i) adaptive determination of the temperature schedule, (ii) the model proposal scale is tuned adaptively using the acceptance rate at the previously considered temperature and, (iii) the evidence is calculated along the run with updates being made every time the temperature changes. In PT, the temperature schedule and the proposal scale need to be pre-defined. The evidence estimation in PT is reduced to a one-dimensional integral over the inverse temperatures, which can imply large approximation errors if the temperatures are comparatively few or poorly chosen. We assess the performance of ASMC for this test case and compare the results with the PT results of Laloy et al. (2016). We further discuss the insights offered by analyzing the results at intermediate temperatures corresponding *de facto* to assumptions of larger measurement noise. With respect to the geophysical ASMC study by Amaya et al. (2021), the present work considers a hydrogeological problem that is much more non-linear and the model parameterizations and model proposal schemes are entirely different. In ASMC-SGR we need to consider as many model parameters as there are pixels (7500 in our example) while the deep generative network used by Amaya et al. (2021) only considered a few tenths of unknown latent variables.

2. Method

2.1. Bayes' theorem

It is often beneficial to pose inverse problems within a probabilistic framework using Bayes' theorem, in which the parameters to infer are treated as random variables. If we consider a conceptual model composed by parameters $\boldsymbol{\theta}$, the posterior pdf $\pi(\boldsymbol{\theta}|\mathbf{y})$ is given by:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{\pi(\mathbf{y})}. \tag{1}$$

The prior PDF $\pi(\theta)$ represents the a priori information concerning the model parameters. This information is then weighted by the likelihood function $p(\mathbf{y}|\theta)$ that expresses, for a given noise model, how probable it is that a particular set of parameter values have produced the observations $\mathbf{y}$. Assuming the noise on the data to be uncorrelated and normally distributed with a constant variance σ^2, the likelihood is expressed as:

$$p(\mathbf{y}|\boldsymbol{\theta}) = (\sqrt{2\pi\sigma^2})^{-m_d} \exp\left[-\frac{1}{2\sigma^2}\sum_{i}^{m_d}(y_i - F_i(\boldsymbol{\theta}))^2\right], \tag{2}$$

where m_d is the number of data points and $F(\boldsymbol{\theta})$ the simulated data given a set of model parameter values. It can be convenient to consider the variable component of the natural logarithm of the likelihood:

$$l(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\sum_{i}^{m_d}(y_i - F_i(\boldsymbol{\theta}))^2, \tag{3}$$

which we refer to as the reduced log-likelihood as it ignores the constant terms.

The evidence, also known as the marginal likelihood, is the normalizing constant in Bayes' theorem. This quantity can be used to rank alternative conceptual models, defined by different prior models, as it represents how consistent a conceptual model is with the set of observations under consideration (Kass and Raftery, 1995). The evidence is a multidimensional integral over the parameter space:

$$\pi(\mathbf{y}) = \int \pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}, \tag{4}$$

making it very challenging to calculate for high-dimensional models. Brunetti et al. (2019) focus particularly on how to compute the evidence to compare different conceptual models within a MPS inversion framework.

2.2. Sequential geostatistical resampling

Prior models are often represented by mathematical functions allowing any prior model realization to be evaluated in terms of its probability. Examples include uniform priors, multivariate Gaussian priors and latent space distributions learned by deep generative neural networks. However, such explicit prior model parameterizations are not always suitable, or possible, when seeking to encode realistic geological spatial heterogeneity (Linde et al., 2015). As an alternative, one can instead consider realizations of MPS simulation tools (Strebelle, 2002) as samples drawn from the prior model. These realizations honor the higher-order statistics of training images that can be built based on expected geological structures, outcrops, geophysical or borehole data. The downside of such prior sampling-based approaches is that one cannot calculate the prior probabilities of model realizations as needed in most MCMC algorithms.

Sequential geostatistical resampling is a mechanism allowing MCMC inference when model proposals are drawn using MPS algorithms that sample proposals proportionally to the prior density. It builds on the foundational paper by Mosegaard and Tarantola (1995) in geophysics. However, it is noteworthy that the underlying philosophy of such a prior sampling-based algorithm has more recently received strong theoretical backing in the context of infinite-dimensional inversion problems (e.g., Cotter et al. (2013)). At each MCMC iteration, a new model proposal is generated by re-simulating a random fraction of the current model realization using an MPS algorithm, conditioned to the remaining pixel values. There are two end-member approaches to determine the locations of the pixels that are to be resimulated: either a randomly located box-shaped area (Alcolea and Renard, 2010; Hansen et al., 2012) or randomly located points (Mariethoz et al., 2010a). In this study, we use boxes as it provided the best results in Laloy et al. (2016). We further rely on the DeeSse MPS algorithm (http://www.randlab.org/research/deesse/) that is, in turn, based on the direct sampling method by Mariethoz et al. (2010b). To re-simulate the value of a certain uninformed pixel, the algorithm scans the training image searching for patterns that agree with those found in the vicinity of this pixel. If a similar-enough pattern is found, it assigns the value of the pixel under consideration in the training image to the one in the new proposed model. This procedure is repeated for all the pixels that are to be re-simulated.

In MCMC algorithms, the Metropolis rule is used to accept or reject model proposals obtained from symmetric proposal distributions. The acceptance probability Γ to move from a current state $\boldsymbol{\theta}_c$ to a proposed state $\boldsymbol{\theta}_p$ is:

$$\Gamma(\boldsymbol{\theta}_p, \boldsymbol{\theta}_c) = min\left(1, \frac{\pi(\boldsymbol{\theta}_p)p(\mathbf{y}|\boldsymbol{\theta}_p)}{\pi(\boldsymbol{\theta}_c)p(\mathbf{y}|\boldsymbol{\theta}_c)}\right). \tag{5}$$

As mentioned above, this rule cannot be used with MPS algorithms such as DeeSse as $\pi(\boldsymbol{\theta})$ is unknown. Instead, MPS-based inversions often rely on the extended Metropolis (Mosegaard and Tarantola, 1995) method that is applicable if the model proposal mechanism generates samples drawn proportionally to the prior PDF. The acceptance probability is then reduced to:

$$\Gamma(\boldsymbol{\theta}_p, \boldsymbol{\theta}_c) = min\left(1, \frac{p(\mathbf{y}|\boldsymbol{\theta}_p)}{p(\mathbf{y}|\boldsymbol{\theta}_c)}\right), \tag{6}$$

which involves only likelihood ratios.

2.3. Adaptive sequential Monte Carlo (ASMC)

2.3.1. Power posteriors

Tempering consists in introducing a temperature variable flattening the likelihood function in Eq. (1). The corresponding tempered posterior PDFs are called power posteriors and can, in their unnormalized form, be expressed as:

$$\gamma_t(\boldsymbol{\theta}_t|\mathbf{y}) \equiv \pi(\boldsymbol{\theta}_t)p(\mathbf{y}|\boldsymbol{\theta}_t)^{\alpha_t}, \tag{7}$$

where the likelihood is raised to an inverse temperature $\alpha_t \in [0,1]$. The effect of increasing the temperature (decreasing α_t) is that the likelihood function becomes less peaky, that is, with less pronounced modes. Targeting these power posteriors, instead of only targeting the posterior PDF at unit temperature as in standard MCMC, increases the exploration capacity because the tempering process decreases the probability of getting trapped in local minima. A graphical explanation regarding the advantages of tempered exploration can be found in Sambridge (2014). Fig. 1 illustrates the main structural differences between standard MCMC, and the methods of PT and AIS that both rely on tempering.

2.3.2. Annealed importance sampling (AIS)

Importance sampling is a Monte Carlo method used to estimate properties of a distribution that it is not possible to sample from (Hammersley and Handscomb, 1964). It relies on an auxiliary distribution $q(\boldsymbol{\theta})$ for drawing the samples, that must include and should ideally be slightly inflated with respect to the target distribution. For most applications, sampling from the prior distribution in order to estimate properties of the posterior PDF suffers from the curse of dimensionality, meaning that the computational effort needed to draw enough samples with a significant likelihood, as needed to enable reliable estimates, is unfeasible. In contrast, sampling from a well-chosen importance distribution allows focusing the sampling in regions of high posterior probability. The samples drawn are then used to compute the desired property while correcting for the bias resulting from the chosen importance distribution. If the target distribution is the unnormalized posterior PDF $\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})$, the importance weights are given by:

$$w = \frac{\pi(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta})}{q(\boldsymbol{\theta})}. \tag{8}$$

Neal (2001) combined tempering and importance sampling to produce the AIS method. It uses N chains, each of them representing evolving particles that target sequentially a sequence of power posteriors at different temperatures ranging from the prior to the unnormalized posterior PDF of interest. The sequence is given by $\left\{\gamma_t(\boldsymbol{\theta}_t|\mathbf{y})\right\}_{t=0}^{T}$, and it contains unnormalized power posteriors given by Eq. (7) with α_t ranging from $\alpha_{t=0} = 0$ (the prior) to $\alpha_{t=T} = 1$ (the unnormalized posterior PDF). The normalized power posteriors are given by:

$$\pi_t(\boldsymbol{\theta}_t|\mathbf{y}) = \frac{\gamma_t(\boldsymbol{\theta}_t|\mathbf{y})}{Z_t}, \tag{9}$$

where Z_t is the normalizing constant of the distribution.

In AIS, importance sampling steps are performed sequentially between each pair of consecutive power posteriors. A subsequent power posterior $\gamma_t(\boldsymbol{\theta}_t|\mathbf{y})$ is approximated by using the estimation of the previous power posterior $\gamma_{t-1}(\boldsymbol{\theta}_t|\mathbf{y})$ as the importance sampling distribution. In contrast to standard importance sampling were the samples are drawn directly from the importance distribution, in AIS the $\gamma_{t-1}(\boldsymbol{\theta}_t|\mathbf{y})$ samples are obtained by performing K MCMC iterations targeting this power posterior starting from the approximation $\gamma_{t-2}(\boldsymbol{\theta}_t|\mathbf{y})$. By performing multiple intermediate importance sampling steps between the prior and the posterior PDF, it is possible to ensure that each importance distribution is of high quality (slightly inflated with respect to the target) leading to estimates with low uncertainty (variance). After the importance sampling step (represented by the longer arrows in between different colored circles in Fig. 2), again each of the N chains perform K MCMC steps targeting now $\gamma_t(\boldsymbol{\theta}_t|\mathbf{y})$. This process is repeated until $\alpha_t = 1$.

We refer to the importance weights (Eq. (8)) resulting from each intermediate importance sampling step as the incremental weights. For a particle i at state $\boldsymbol{\theta}^i_{t-1}$, the incremental weight w^i_t that result from using $\gamma_{t-1}(\boldsymbol{\theta}_{t-1}|\mathbf{y})$ as an importance distribution for $\gamma_t(\boldsymbol{\theta}_t|\mathbf{y})$ is:

$$w^i_t = \frac{\gamma_t(\boldsymbol{\theta}^i_{t-1}|\mathbf{y})}{\gamma_{t-1}(\boldsymbol{\theta}^i_{t-1}|\mathbf{y})}. \tag{10}$$

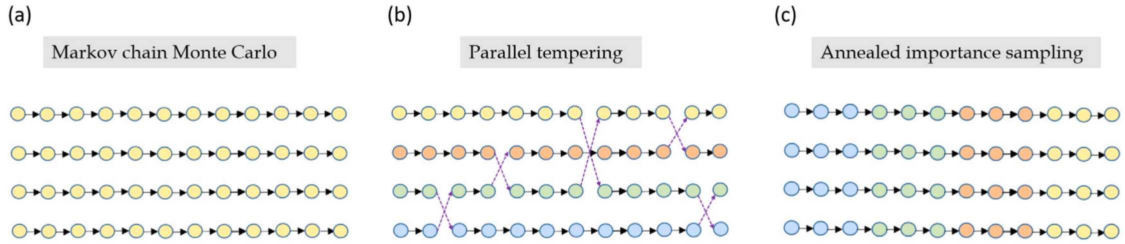

Fig. 1. Diagram illustrating structural differences between probabilistic inversion methods with the circles representing the evolving states at each iteration and the colors denoting the temperatures of the targeted power posteriors. In (a) standard MCMC, all the chains target the posterior PDF at unit temperature (shown in yellow). (b) Parallel tempering uses a number of chains targeting different power posteriors while allowing eventual swaps between them (shown as purple dashed lines), whereas in (c) annealed importance sampling, the targeted power posteriors change during the run in response to a gradually cooling sequence.

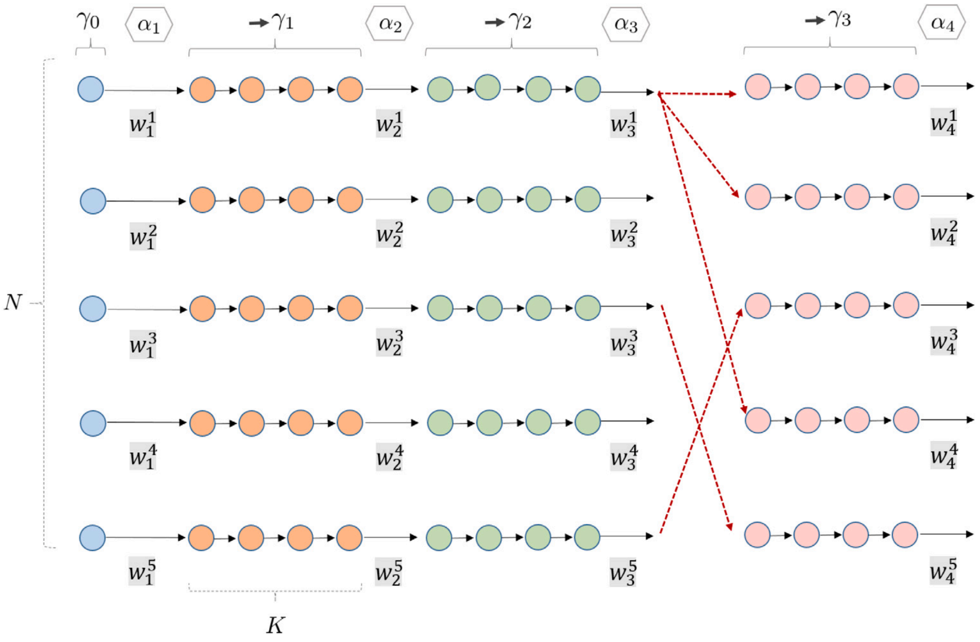

Fig. 2. Schematic representation of Sequential Monte Carlo (SMC) using $N = 5$ particles evolving in parallel. After the initial sampling from the prior PDF (blue circles), $K = 4$ Markov steps are performed to approximate power posteriors γ_t. In these power posteriors, the likelihood is raised to an inverse temperature α_t that increases gradually. At the end of each approximation, an importance sampling step is performed to calculate an incremental weight w_t. Adaptive Sequential Monte Carlo (ASMC) incorporates two modifications with respect to Annealed Importance Sampling (AIS): (i) adaptive determination of the α_t-sequence defining the sequence of power posteriors and (ii) resampling when the variance of the particle weights becomes too large (indicated by red dashed lines).

To calculate the total weight of a particle, one needs to account for all the intermediate importance sampling steps. To achieve this, the incremental weight w_t^i is used to update the normalized weight of particle i by:

$$W_t^i = \frac{W_{t-1}^i w_t^i}{\sum_{j=1}^N W_{t-1}^j w_t^j}, \tag{11}$$

where W_{t-1}^i is the normalized weight (that is, $\sum_{i=1}^N W_{t-1}^i = 1$) of the previous importance sampling step. The posterior PDF is then approximated through a particle approximation, in which the relative probabilities of the last N states of the particles are determined by the final normalized weights W_T^i. By saving intermediate normalized weights and corresponding particle states, the method allows also to approximate intermediate power posteriors that represent the solutions to the equivalent tempered problems.

2.3.3. Resampling

The variance of the particle weights influences strongly the quality of the importance sampling estimator (Neal, 2001). When using AIS, this variance may grow exponentially, resulting in poor estimations of the posterior PDF and the evidence. Sequential Monte Carlo (SMC) is a family of particle approaches that, as AIS, rely on sequential importance sampling. However, SMC incorporates also resampling (Del Moral et al., 2006; Doucet and Johansen, 2011). In a resampling step, the states of the particles are replicated according to a probability that is proportional to their current normalized weights, and all the weights are re-set to $1/N$. The replacement of particles with lower weights and increasing those with higher weights results in two advantages: (i) it avoids the variance of the weights to grow indefinitely and (ii) it orients the exploration towards regions of higher posterior probability. Nevertheless, since the resampling process increases the variance of the estimates (Douc and Cappe, 2005), it is often better to perform resampling only when needed. The effective sample size (ESS) (Kong et al., 1994) is expressed as:

$$ESS_t = \frac{(\sum_{i=1}^N W_{t-1}^i w_t^i)^2}{\sum_{j=1}^N (W_{t-1}^j)^2 (w_t^j)^2}. \tag{12}$$

It quantifies the number of effective samples in the particle approximation. The common approach is to monitor the ESS along the run, and perform resampling when it is lower than a specified threshold. In this paper, we rely on systematic resampling due to its good performance and easy implementation (Doucet and Johansen, 2011). Fig. 2 shows a graphical example of SMC with $N = 5$ particles, in which the resampling step is indicated with red dashed lines.

2.3.4. Adaptive tempering schedule

One complication of the AIS and SMC methods is the difficulty to pre-define a suitable tempering schedule (Fig. 2). Zhou et al. (2016) propose an adaptive SMC method (ASMC) (their algorithm 4) in which an appropriate α-step-size increment is determined before each importance sampling step. To do so, they rely on the conditional effective sample size ($CESS$) quantifying the quality of using the particle approximation $\gamma_{t-1}(\boldsymbol{\theta}_{t-1}|\mathbf{y})$ as an importance distribution to estimate expectations for the $\gamma_t(\boldsymbol{\theta}_{t-1}|\mathbf{y})$ arising for different choices of α_t. The $CESS$ is given by:

$$CESS = N\frac{(\sum_{i=1}^{N} W_{t-1}^{i} w_t^{i})^2}{\sum_{j=1}^{N} W_{t-1}^{j}(w_t^{j})^2}. \tag{13}$$

The ESS and $CESS$ are both obtained by a sample approximation of a Taylor expansion of the relative variance of the estimator (Kong et al., 1994). The difference between them is that the ESS embraces the accumulated mismatch between the importance and target distributions, whereas the $CESS$ focuses on the quality of the current importance sampling step. If resampling was to be performed at every iteration, then the ESS and $CESS$ quantities would be equal. A detailed derivation of the $CESS$ can be found in the supplementary material of Zhou et al. (2016).

The $CESS$ depends on the incremental weights w_t that in turn depend on α_t. The strategy consists in finding the α-increment between consecutive power posteriors, that is, the $\Delta\alpha_t$ such that $\alpha_t = \alpha_{t-1} + \Delta\alpha_t$, giving the $CESS$ that is the closest to a pre-defined quality expressed by $CESS_{op}$. To find $\Delta\alpha_t$, we rely on a binary search within a sequence of possible $\Delta\alpha$ values. First, the $CESS$ is computed using the middle value of the $\Delta\alpha$ sequence and it is compared with $CESS_{op}$. Depending on if it is higher or lower, one of the two $\Delta\alpha$ half-intervals is kept. This procedure is repeated until the $\Delta\alpha$ that gives the $CESS$ that is the closest to $CESS_{op}$ is found.

If we increase $CESS_{op}$, we obtain higher-quality estimates as the number L of intermediate power posteriors increases, but at the expense of a longer ASMC run. The total number of iterations per particle is $L \times K$, with K the number of MCMC steps per intermediate power posterior. In practice, the ratio $CESS_{op}/N$ is chosen close to 1 in order to ensure high quality estimates. It has been suggested that it should be at least 0.99 to build a smooth α-sequence (Amaya et al., 2021), but the optimal value is highly problem-dependent. The impact of the $CESS_{op}/N$ value on the resulting L is non-linear and not easy to predict.

2.3.5. ASMC-based evidence estimation

Evidence estimation is essential for Bayesian model selection. Considering two neighboring distributions $\gamma_{t-1}(\boldsymbol{\theta}_{t-1}|\mathbf{y})$ and $\gamma_t(\boldsymbol{\theta}_t|\mathbf{y})$, we can express the ratio of their normalizing constants as:

$$\frac{Z_t}{Z_{t-1}} = \frac{\int \gamma_t(\boldsymbol{\theta}_t|\mathbf{y})d\boldsymbol{\theta}_t}{\int \gamma_{t-1}(\boldsymbol{\theta}_{t-1}|\mathbf{y})d\boldsymbol{\theta}_{t-1}}. \tag{14}$$

Del Moral et al. (2006) propose an approximation of this ratio as:

$$\frac{Z_t}{Z_{t-1}} \approx \sum_{i=1}^{N} W_{t-1}^{i} w_t^{i}. \tag{15}$$

The evidence $\pi(\mathbf{y})$ is the normalizing constant Z_T of the unnormalized posterior PDF, that is, the last distribution of the sequence when $\alpha_{t=T} = 1$. Considering that the prior PDF integrates to one, $Z_0 = 1$, we can express the evidence as the product of the normalizing constant ratios:

$$\pi(\mathbf{y}) = Z_T = \frac{Z_T}{Z_0} = \prod_{t=1}^{T}\frac{Z_t}{Z_{t-1}} \approx \prod_{t=1}^{T}\sum_{i=1}^{N} W_{t-1}^{i} w_t^{\,i}. \tag{16}$$

Consequently, the evidence can be updated along the run by accounting for the evolving particle weights.

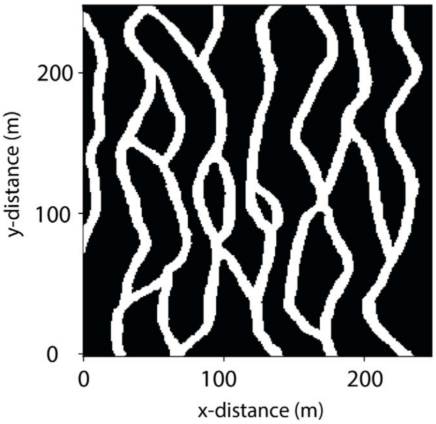

Fig. 3. Channelized binary training image from Strebelle (2002) with the spatial dimensions used in the present study.

2.4. Full ASMC-SGR algorithm

Our algorithm combining the SGR method for model proposals with ASMC for posterior PDF and evidence estimation is given in Algorithm 1. We denote this algorithm as ASMC-SGR (following the nomenclature in Laloy et al. (2016)). In this study, the proposal scale ϕ indicates half of the side-length in meters of the box that is being re-simulated at each iteration. In addition to the previously mentioned advantages of adaptive tempering and evidence estimation, the algorithm also has the attractive feature that the proposal scale ϕ can be tuned on-the-go without violating detailed balance conditions as would be the case for MCMC or PT applications. This is simply achieved by keeping track of the acceptance rate for the K MCMC steps at the previous α_{t-1} and then to use this information to adapt the proposal scale for the next number of K MCMC steps, such that the acceptance rate remains within a pre-defined range. This saves a lot of time compared with standard MCMC and PT algorithms that often necessitate tuning using multiple time-consuming trial runs.

3. Results

3.1. Test case

We consider the second test case from Laloy et al. (2016), in which the concentration of an injected tracer is measured at regular time intervals. The conceptual model is represented by a 250 × 250 categorical binary training image from Strebelle (2002) (Fig. 3). The 2-D reference model is located in the x-y plane and has a dimension of 75 m × 100 m with a discretization cell size of 1 m. The hydraulic conductivity K is 0.01 m/s for the channels and 0.0001 m/s for the matrix. A conservative tracer with a concentration of 1 kg/m^3 is injected at 8 locations on the top and bottom of the model (Fig. 4). The concentration is measured every 8 h during 10 days at 11 pumping wells (a total of 330 observations) that extract 0.0005 m^2/s of water, and the facies at these points are assumed to be known. This test exhibits symmetry with respect to the x-axis, such that any model and its mirrored image produce the same simulated concentration data and, therefore, the same likelihood. Consequently, the posterior PDF is bimodal with two distinct modes. Mode 1 of the reference model was obtained as a random realization from the DeeSse algorithm (Fig. 4a) and then mirrored to obtain the mode 2 reference model (Fig. 4b).

The simulations are performed using MaFloT, a finite-volume open-source code for transport simulations in porous media (Künze and Lunati, 2012). Fixed head boundaries of 0 m on the top and bottom of the domain and no-flow boundaries on the sides are assumed to simulate steady-state groundwater flow. For the tracer transport, we

Algorithm 1: ASMC-SGR

The SGR section of the algorithm is adapted from Laloy et al. (2016) and the ASMC section from Zhou et al. (2016) (their algorithm 4)

```
Variables to pre-define:
    Number of particles ($N$), optimal CESS ($CESS_{op}$), ESS threshold ($ESS^*$), number of MCMC iterations at each intermediate distribution ($K$),
    minimal and maximal acceptance rate ($AR_{min}$, $AR_{max}$), minimal ($\phi_{min}$) and maximal ($\phi_{max}$) proposal scale and its percentage of change ($f$)
Initialization: Set $t = 0$
    Set $\alpha = 0$
    Sample $\theta_0$ from the prior $\pi(\theta)$ $N$ times
    Set the $N$-dimensional vector of normalized weights $\mathbf{W}_0 = [\frac{1}{N}; \frac{1}{N}; ...; \frac{1}{N}]$
    Set evidence $\pi(\mathbf{y}) = 1$
Iteration : Set $t = t + 1$
    Search for incremental distribution
        Do binary search for the increment $\Delta\alpha$ that gives the CESS (Eq. (13)) that is the closest to $CESS_{op}$
        Update $\alpha = min(1, \alpha + \Delta\alpha)$ and define the following intermediate distribution $\gamma_t(\theta_t|\mathbf{y}) = \pi(\theta_t)p(\mathbf{y}|\theta_t)^\alpha$
        Perform the importance sampling step: compute the weight increments $w_t^i$ (Eq. (10)), update and save the normalized weights $W_t^i$ (Eq. (11))
        and the evidence $\pi(\mathbf{y}) = \pi(\mathbf{y}) \sum_{i=1}^{N} W_{t-1}^i w_t^i$ (Eq. (16))
    Resampling
        Calculate ESS (Eq. (12)), if $ESS < ESS^*$ do resampling: re-organize $\theta_t$ states and update $\mathbf{W}_t = [\frac{1}{N}; \frac{1}{N}; ...; \frac{1}{N}]$
    Do $K$ MCMC iterations for each of the $N$ particles (chains):
        Propose moves $\theta_p$: randomly select the location of a box with dimensions $2\phi \times 2\phi$, and run the MPS simulation using the points outside the box as
        conditioning points and accept or reject based on the extended Metropolis rule, with an acceptance probability given by: $\Gamma(p, c) = min\left(1, \frac{p(\mathbf{y}|\theta_p)^\alpha}{p(\mathbf{y}|\theta_c)^\alpha}\right)$ (Eq. (6))
        Save the $N$ models $\theta$ and their likelihoods
        Set last state as $\theta_{t+1}$
    Tune proposal scale
        If acceptance rate $AR < AR_{min}$ then decrease proposal scale factor: $\phi = \phi * (1 - \frac{f}{100})$
        If acceptance rate $AR > AR_{max}$ then increase proposal scale factor: $\phi = \phi * (1 + \frac{f}{100})$
        If $\phi < \phi_{min}$ then $\phi = \phi_{min}$, if $\phi > \phi_{max}$ then $\phi = \phi_{max}$
Repeat until $\alpha$=1
```

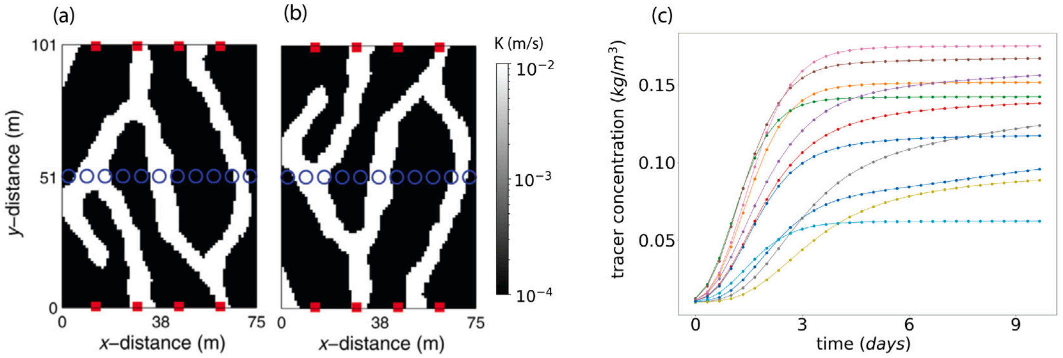

Fig. 4. (a) Reference model [mode 1] and (b) mirrored reference model [mode 2], the red squares represent the points where the tracer is injected and the blue circles represent the pumping wells where measurements are made. Both models result in the same (c) simulated concentration over time (shown before being contaminated with uncorrelated Gaussian noise), where each color represents the observations at one pumping well.

assume open boundaries, an hydraulic dispersivity of 0.1 m and a background concentration of 0.01 kg/m^3. The simulated data were corrupted with uncorrelated Gaussian noise with a standard deviation of $\sigma = 0.003$ kg/m^3, approximately 3% of the mean concentration.

3.1.1. ASMC-SGR settings

The proposal scale ϕ used to create candidate models is tuned along the run (see Algorithm 1) by increasing or decreasing it by $f = 20\%$ to ensure that the acceptance rate stays within the range of $AR_{min} = 15\%$ and $AR_{max} = 35\%$. It is further constrained to be between $\phi_{max} = 50$ m and $\phi_{min} = 5$ m. For the DeeSse simulations, we follow Laloy et al. (2016) and use 75 neighbors, which implies that the patterns that are searched by the algorithm are composed of the 75 informed nodes that are the closest to the one being re-simulated. The fraction of the training image that is scanned is 0.9 and the distance threshold to accept a pattern is 0.01 (Mariethoz et al., 2010b).

3.2. ASMC-SGR results

3.2.1. Test 1: ASMC-SGR with 24 particles

We first compare the ASMC-SGR results with those obtained by Laloy et al. (2016) for a similar computational budget: 24 chains and 25,000 iterations per chain. To achieve this, we chose $N = 24$ particles and $CESS_{op}/N = 0.9997$ combined with $K = 18$, which resulted in 25,956 iterations per particle. The resampling threshold ESS^*/N was set to 0.3 (Del Moral et al., 2006). The user-defined parameters and length of the run are summarized in Table 1 (ASMC-SGR 24p).

We first consider the evolution of the tempered log-likelihood, that is, the likelihood raised to the inverse temperature in the natural log-scale (Fig. 5a). The tempered log-likelihood of each particle is seen to evolve according to the reference tempered log-likelihood curve (calculated using the assumed noise standard deviation $\sigma = 0.003$ kg/m^3). If $CESS_{op}/N$ or K would be too low, then the particles would have considerably lower tempered likelihoods than the reference curve, thereby, indicating that the sampled log-likelihoods are too low and that the associated computational budget is insufficient for the problem at hand. Consequently, this type of curve is a useful diagnostic plot allowing the user to terminate an ASMC run at an early stage if the tempered log-likelihoods fall below the reference curve.

The automatically tuned proposal scale ϕ (Fig. 5e) enables the acceptance rate to stay within the pre-defined range (Fig. 5c). The resulting α_t-sequence (Fig. 5b) demonstrates that roughly half of the forward simulations are carried out with α_t-values less than 0.01, corresponding to temperatures above 100. The plot showing the evolution

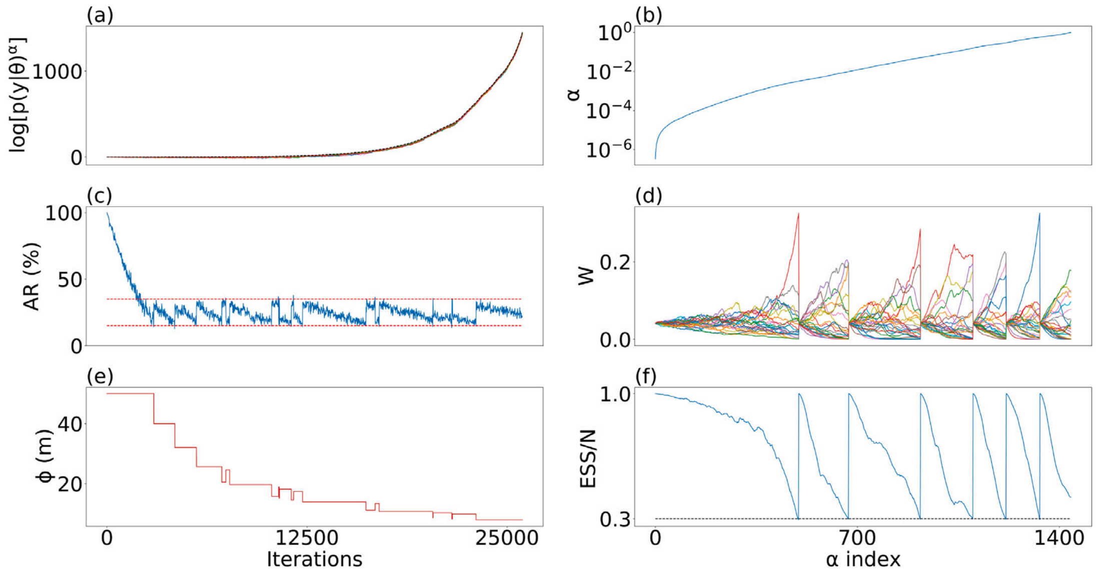

Fig. 5. ASMC-SGR results using 24 particles: (a) tempered log-likelihood vs. iterations per particle, the colors represent different particles and the black dashed line indicates the reference tempered log-likelihood; (b) α-sequence vs. α index; (c) acceptance rate vs. iterations per particle, the dashed line indicates the pre-defined minimum and maximum range; (d) normalized weights vs. α index, the colors represents different particles; the (e) proposal scale vs. iterations per particle; (f) ESS/N vs. α index, the black dashed line indicates the 0.3 threshold below which resampling is made.

Table 1
User-defined parameters, resulting sequence length and data fitting for ASMC-SGR using 24 and 72 particles. The reduced reference log-likelihood $l(\mathbf{y}|\theta_{ref})$ (Eq. (3)) for this test case is −165. Using PT-SGR, Laloy et al. (2016) obtained a $\Delta l(\mathbf{y}|\theta_t)$ = 9% for a numerical demand of 600,000 forward simulations.

	ASMC-SGR 24p	ASMC-SGR 72p
Particles (N)	24	72
$CESS_{op}/N$	0.9997	0.9997
ESS^*/N	0.3	0.3
AR_{min}	15%	15%
AR_{max}	35%	35%
K iterations	18	18
L power posteriors	1442	1533
Iterations per particle	25,956	27,594
Resampling times	6	6
Total number of forward simulations	622,944	1,986,768
$\Delta l(\mathbf{y}|\theta_t)$	3.76%	1.5%
$l(\mathbf{y}|\theta_T)$ range	[−196, −164]	[−188, −160]

of the normalized weights (Fig. 5d) illustrates the divergence of the weights between resampling steps and the re-alignment of the weights when the normalized effective sample size ESS/N (Fig. 5f) reaches below the 0.3 threshold.

To compare these ASMC-SGR 24p results with those obtained by Laloy et al. (2016), we first consider the measure used in their study as an indicator of data fitting:

$$\Delta l(\mathbf{y}|\theta)[\%] = \frac{\bar{l}(\mathbf{y}|\theta_T) - l(\mathbf{y}|\theta_{ref})}{l(\mathbf{y}|\theta_{ref})} \times 100, \quad (17)$$

where $l(\mathbf{y}|\theta_{ref})$ is the reduced reference log-likelihood (Eq. (3)) and $\bar{l}(\mathbf{y}|\theta_T)$ is the mean sampled reduced log-likelihood. For MCMC and PT, this mean is simply the arithmetic average of the reduced log-likelihoods after burn-in (only considering unit temperature chains for PT), whereas for ASMC it is the weighted average of the N final reduced log-likelihoods. Laloy et al. (2016) demonstrated a drastic improvement when using PT-SGR compared with MCMC-SGR following Hansen et al. (2012). The indicator $\Delta l(\mathbf{y}|\theta_t)$ was 9%; an important improvement of 70% on average compared with MCMC-SGR. Still, the reference reduced log-likelihood was actually not contained in the range of sampled reduced log-likelihoods with PT-SGR, indicating that these samples are not representative of the posterior PDF. For our ASMC-SGR 24p run, the indicator $\Delta l(\mathbf{y}|\theta_t)$ is 3.76% and the log-likelihood range contains the reference value (Table 1).

Fig. 6a–d show exemplary PT-SGR posterior samples from Laloy et al. (2016). These samples do not resemble either mode 1 or mode 2, even if Fig. 6b has some structural similarities with mode 1 (Fig. 4a). In contrast, the final states obtained by ASMC-SGR 24p (Fig. 6e–h) recover models that resemble both reference modes: The realizations in Fig. 6e–g resemble mode 2 (Fig. 4b) and the one in Fig. 6h resembles mode 1 (Fig. 4a).

The reference mean (Fig. 7a) is the mean of mode 1 (Fig. 4a) and mode 2 (Fig. 4b) of the reference model. The true posterior mean is unknown and it is likely to be slightly biased towards models resembling one of the modes. The reason for this is that even if mode 1 and 2 have the same likelihood, they do not have the same prior probability. This is a consequence of using the training image in Fig. 3 that is likely to favor certain orientations of structures when generating prior samples. Nevertheless, Fig. 7a provides a sensible point of comparison.

The ASMC-SGR 24p weights (Fig. 6e–h) and the posterior mean corresponding to the weighted arithmetic mean of the samples (Fig. 7b) suggest that the total weights given to the two “modes” is unbalanced with “mode 2” having a higher total weight than “mode 1”. Still, these results show that ASMC-SGR can sample the two modes of this very challenging inverse problem and that the structures of the reference mean are partly recovered (unlike for the PT-SGR mean, see Fig. 9b in Laloy et al. (2016)).

3.2.2. Test 2: ASMC-SGR with 72 particles

ASMC provides an approximation not only of the posterior PDF but also of every tempered intermediate power posterior. In this section, we focus on the evolution of the (unnormalized) power posteriors as α increases from the prior ($\alpha = 0$) to the posterior PDF ($\alpha = 1$). One way of interpreting these power posteriors is to consider them as posterior PDFs for different assumptions on the data error level. Indeed, decreasing the α-exponent has the same impact on the likelihood variable component as increasing the assumed standard deviation σ of the data noise ($\alpha \propto \frac{1}{\sigma^2}$ in Eqs. (3) and (7)). Thus, the effect of tempering with a given α could also be achieved by considering an assumed standard deviation of $\sigma_\alpha = \sigma/\sqrt{\alpha}$, where σ is the original standard deviation of 0.003 kg/m^3 (for example, $\alpha = 0.25$ is analogous to assuming a standard deviation that is twice as large $\sigma_\alpha = 0.006$ kg/m^3).

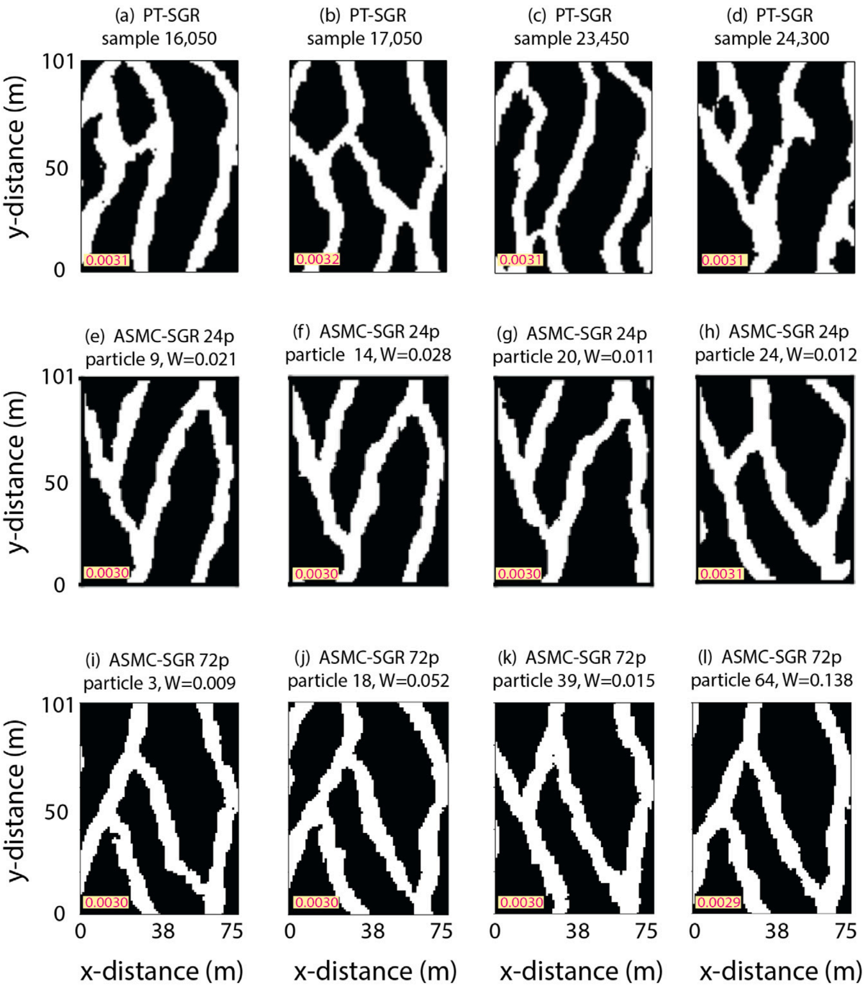

Fig. 6. Samples from the posterior PDF obtained with: (a)–(d) PT-SGR by Laloy et al. (2016); (e)–(f) ASMC-SGR using 24 particles and (i)–(l) ASMC-SGR using 72 particles (the corresponding weights W of the particles are shown). The root mean square error (RMSE) without units is indicated for each sample; the corresponding value for the true model (modes 1 and 2) is 0.0030.

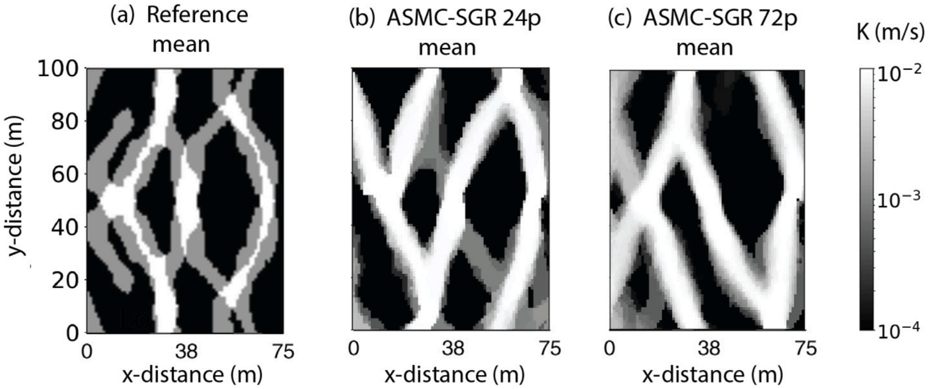

Fig. 7. (a) Mean of the reference model's two modes; (b) ASMC-SGR 24p posterior mean and (c) ASMC-SGR 72p posterior mean obtained as a weighted mean of the final states of the particles.

Since the objective in this section is no longer to compare the results with Laloy et al. (2016) for a similar computational budget, we now consider more particles. We increase the number of particles running in parallel from 24 to 72, thereby, aiming for improved approximations of the intermediate power posteriors, while keeping fixed the other user-defined parameters (ASMC-SGR 72p in Table 1). The number of power posteriors needed to honor the targeted $CESS_{op}$ are slightly higher compared to the ASMC-SGR 24p test. The indicator $\Delta l(\mathbf{y}|\boldsymbol{\theta})[\%]$ (Eq. (17)) for ASMC-SGR 72p is 1.5%, that is, 60% less than for the 24 particles test. Furthermore, the likelihood range of the final particles is also reduced. The posterior mean for ASMC-SGR 72p (Fig. 7c) and four samples from the posterior PDF (Fig. 6i–l) indicate that most of the samples resemble mode 1 instead of mode 2 of the reference model, that is, the opposite behavior compared with the ASMC-SGR 24p run.

The structural similarity index measure (SSIM) (Wang et al., 2004) can be used to quantify the similarity between two images. It varies

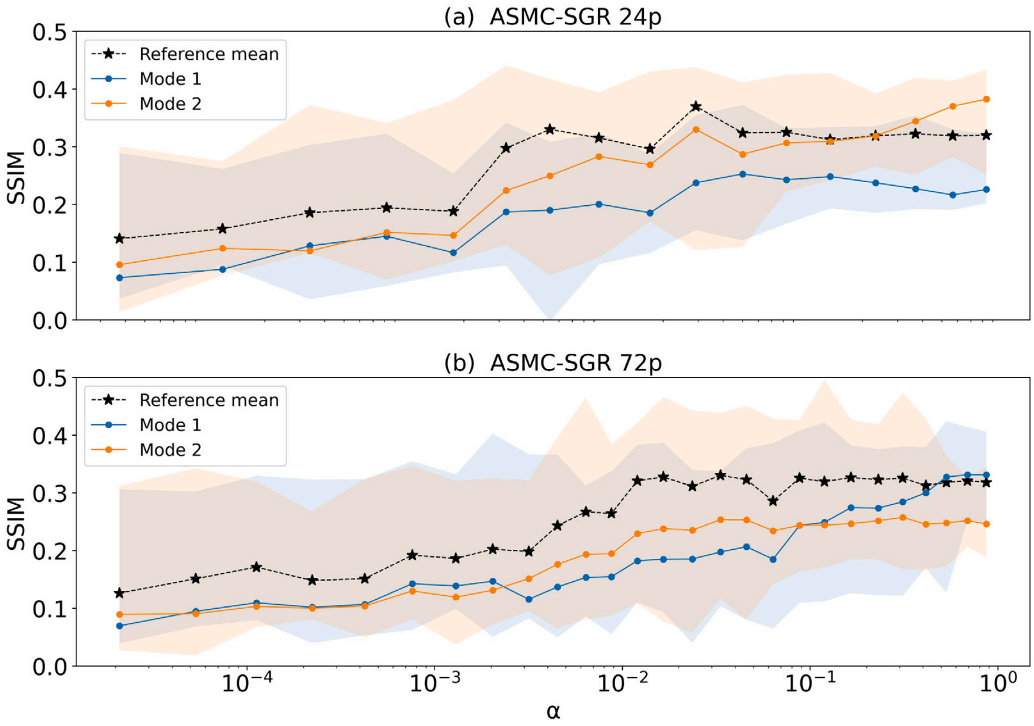

Fig. 8. Structural similarity index measure (SSIM) of the weighted mean models for a subset of the estimated power posteriors with respect to the reference mean model (Fig. 7a) and reference models (modes 1 and 2) (Fig. 4a–b) vs. α. For modes 1 and 2, the range for all samples is indicated with shading. Results are shown for (a) ASMC-SGR 24p and (b) ASMC-SGR 72p.

between −1 and 1, the higher the SSIM the more similar the two compared images are (SSIM = 1 indicates identical images). The SSIM of the power posterior mean models with respect to the reference mean model (Fig. 7a) initially increases before stagnating when α reaches 0.01 for both ASMC-SGR 24p (Fig. 8a) and ASMC-SGR 72p (Fig. 8b). For ASMC-SGR 24p, the SSIM values with respect to mode 2 continue to increase while the SSIM values with respect to mode 1 is even decreasing at the end of the run (Fig. 8a). For ASMC-SGR 72p, the situation is the opposite with the SSIM values with respect to mode 1 being those that continue to increase for larger α-values (Fig. 8b). For ASMC-SGR 24p, the SSIM remains the highest for mode 2 for all α-values above 0.001, while the SSIM values with respect to mode 2 for ASMC-SGR 72p only start to dominate for α-values above 0.1. This is a consequence of the larger number of particles and the corresponding increased ability to approximate the power posterior. The range of SSIM values between the particle realizations and the reference models are shown to decrease as the run progresses.

Fig. 9 shows the posterior means and standard deviations at five stages of the ASMC-SGR 72p run. The mean of the prior models (Fig. 9a) is computed from the initial DeeSse simulations using the facies at the pumping wells as conditioning data. At $\alpha = 2.0e{-}3$ (Fig. 9b), $\alpha = 1.7e{-}2$ (Fig. 9c) and $\alpha = 8.8e{-}2$ (Fig. 9d), the power posterior mean models already resembles patterns of the reference mean (Fig. 7a). When $\alpha = 1$, the posterior mean model is dominated by mode 1 (Fig. 9e). The standard deviations are initially high except in the vicinity of the conditioning points (Fig. 9f) and they decrease as expected with increasing α-values (Fig. 9g–j) as the run evolves towards the posterior PDF. Four samples from the power posteriors corresponding to $\alpha = 2.0e{-}3$ (Fig. 10e–h), $\alpha = 1.7e{-}2$ (Fig. 10i–l) and $\alpha = 8.8e{-}2$ (Fig. 10m–p) indicate that the variability among the realizations are high at the beginning with large corresponding RMSE values. As α increases, the variability among the realizations and the corresponding RMSE values decrease as the samples start resembling the modes and fit the data better.

3.2.3. Resampling and Eve indices

Resampling has the advantage of reducing the variance of the particle weights and focusing the sampling in regions of high posterior probability. However, the corresponding decrease in the variability of the sample realizations has also an adverse impact on the ASMC estimations. A conservative way of estimating the number of independent particles remaining in a run is to trace back the origin of the particles using the Eve indices. Before any resampling is performed, the Eve indices of the particles are $1 : N$. As resampling implies re-organization and replication of particles, the Eve indices change along the run. At time t, each particle i has an Eve index E_t^i that denotes the original index of the particle that moved there (see Lee and Whiteley (2018) for a detailed and illustrative explanation).

The evolution of the Eve indices are shown for tests ASMC-SGR 24p (Fig. 11a) and ASMC-SGR 72p (Fig. 11b). The Eve indices are modified after each resampling step: particles with higher weights are more likely to be replicated, and as they bring their Eve indices (their origin) with them, these Eve indices are replicated as well, while other Eve indices corresponding to particle states with low weights are lost on the way. Consequently, the number of distinct Eve indices is reduced along the run due to resampling; the more resampling there is, the fewer surviving Eve indices at the end of the run. In each of our two example runs there is six resampling steps; this led to two surviving Eve indices out of 24 for ASMC-SGR 24p and only one surviving Eve index out of 72 for ASMC-SGR 72p. Of course, the particles with the same Eve indices are generally not identical as they develop independently after resampling in response to the MCMC proposal steps. Despite inherent randomness, a larger number of Eve indices are expected when reducing the number of resampling steps or increasing the number of particles. For our two test cases, the few surviving Eve indices indicate that a higher number of particles N, intermediate power posteriors or K steps would be beneficial.

3.2.4. Evidence estimation

The evidence $\pi(\mathbf{y})$ (Eq. (4)), which can be used for Bayesian model selection and ranking, is obtained as a byproduct of the ASMC algorithm (Eq. (16)). The log-evidence is shown to evolve similarly for the

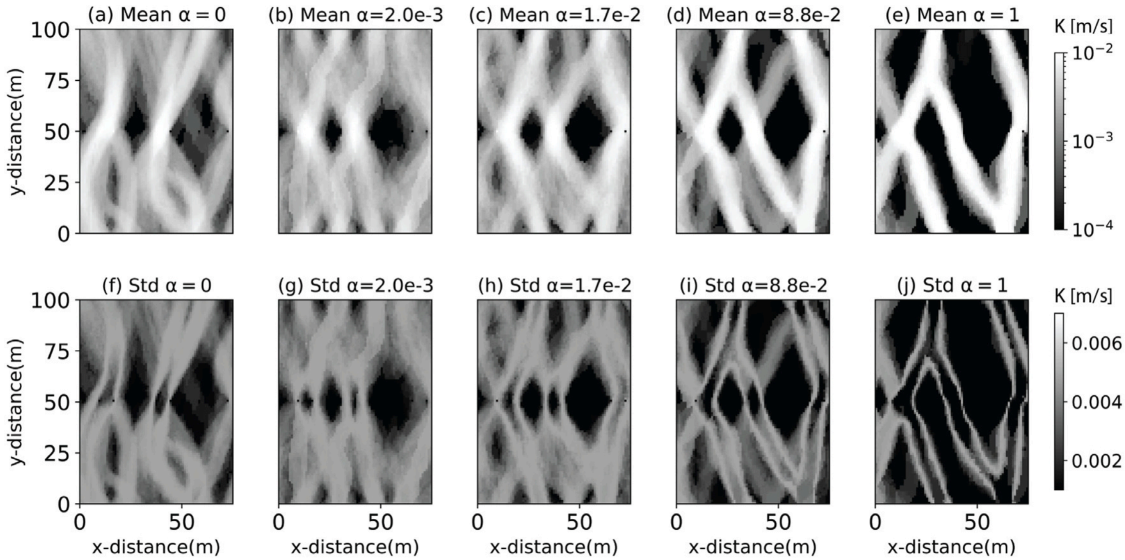

Fig. 9. Posterior (a–e) means and (f–j) standard deviations of five different power posteriors for the ASMC-SGR 72p run: (a), (f) $\alpha_t = 0$; (b), (g) $\alpha_t = 0.002$; (c), (h) $\alpha_t = 0.017$; (d), (i) $\alpha_t = 0.088$ and (e), (j) $\alpha_t = 1$.

ASMC-SGR 24p and ASMC-SGR 72p (Fig. 12a) runs. Both evidence curves have the same shape as α increases, and the final evidence estimates are close: $\pi(\mathbf{y}) = 1374.14$ for ASMC-SGR 72p and $\pi(\mathbf{y}) = 1371.06$ for ASMC-SGR 24p. For many model selection studies focused on conceptual model comparison, the differences in the evidence between conceptual models are often much larger (Amaya et al., 2021; Brunetti et al., 2017, 2019) than this discrepancy (Fig. 12b), thereby, suggesting that only 24 particles would probably provide sufficiently accurate results. Analogous to the power posteriors, it is also possible to interpret the intermediate evidences as those corresponding to larger assumed σ-values. This necessitates a correction nevertheless, as the multiplicative term $((\sqrt{2\pi\sigma^2})^{-m_d}$ in Eq. (2)) does not follow the proportionality $\alpha \propto \frac{1}{\sigma^2}$. The intermediate log-evidences $log[\pi(\mathbf{y},\alpha)]$ can be corrected to $log[\pi(\mathbf{y},\alpha)]_{corr}$ following:

$$log[\pi(\mathbf{y}),\alpha]_{corr} = log[\pi(\mathbf{y},\alpha)] + \alpha\, m_d\, log(\sqrt{2\pi}\sigma) - \, m_d\, log(\sqrt{2\pi}\sigma_\alpha), \tag{18}$$

where σ is the originally assumed standard deviation of 0.003 kg/m^3 and $\sigma_\alpha = \sigma/\sqrt{\alpha}$ is the standard deviation corresponding to that particular α. The results highlight that the estimated evidences depend very strongly on the assumed error level (Fig. 12c).

4. Discussion

For a similar computational budget, ASMC-SGR has been shown to outperform PT-SGR in terms of data fitting (Table 1). Moreover, ASMC-SGR recovers particle states (Fig. 6e-h) that resemble both of the reference modes (Fig. 4a–b), while none of them are recovered when using PT-SGR (Fig. 6a–d). The ASMC algorithm adaptively tunes both the proposal scale and the α-sequence (inverse temperatures) along the run, which implies much less user effort compared to the tedious testing needed to make PT-SGR perform well. If conceptual model comparison is intended, ASMC becomes even more attractive as it provides evidence estimations (Zhou et al., 2016) that are reliable and in agreement with unbiased estimations obtained using brute force Monte Carlo (Amaya et al., 2021).

The intermediate power posterior approximations offered by the ASMC algorithm are highly instructive (Figs. 9–10). When the ASMC-SGR algorithm starts considering α-values above a given threshold (lower for 24 particles than for 72 particles), the sampling tends to become unbalanced in our two example runs and there is one mode that ends up having a higher posterior probability than the other. This could be addressed by increasing the computational budget: either by considering a much larger number of particles (one could imagine using hundreds or thousands of particles), or by increasing $CESS_{op}$ or K that would reduce the number of resampling steps. Resampling plays the important role in particle methods of focusing the sampling towards high-probability regions by controlling the variance of the particle weights. Unfortunately, this advantage comes at the expense of losing the independence between the particles leading, in our case, to over-prediction of one of the posterior modes. In our test example, it would be straightforward to facilitate sampling of both modes simply by allowing for model proposals that would mirror the present state. However, this would not be possible in most realistic settings.

The test example was primarily designed to ensure that the posterior had two posterior modes located far from each other, thereby, enabling comparison of different probabilistic methods for a very challenging inverse problem. In order to allow a fair comparison between the previously published PT results and the new ASMC results, the training image and the DeeSse simulation parameters were the same as in Laloy et al. (2016). However, this implies that the prior probability of sampling modes 1 and 2 are different, and consequently that the two posterior modes have unequal posterior probabilities despite that the likelihoods are equivalent. To ensure that the true posterior has two modes of equal posterior probability, one could use a training image with two layers. The first layer would be the original training image and the second layer would be obtained by mirroring the training image similarly to how mode 2 was created. At each SGR step, the MPS algorithm would scan from either layer 1 or 2. Nevertheless, the fact that ASMC-SGR 24p primarily sampled mode 2 and ASMC-SGR 72p primarily sampled mode 1 suggests that the main limitation in the presented runs are the limited computational budgets that prohibit sampling the two posterior modes well during one ASMC run.

One option to reduce the computational time and, thereby, allow for longer runs would be to use faster algorithms for generating the candidate models: either newer versions of DeeSse, quick sampling (Gravey and Mariethoz, 2020), graph cuts (Zahner et al., 2016), or by replacing MPS-based algorithms with deep learning-based generators as in the study by Amaya et al. (2021). Also, a computational gain could be achieved by replacing the expensive forward solver with a surrogate (e.g., by polynomial chaos expansion (Laloy et al., 2013; Meles et al., 2022)). This should not bias the results if the surrogate is only applied in the intermediate K Markov steps, while still using the expensive forward solver for the importance sampling steps.

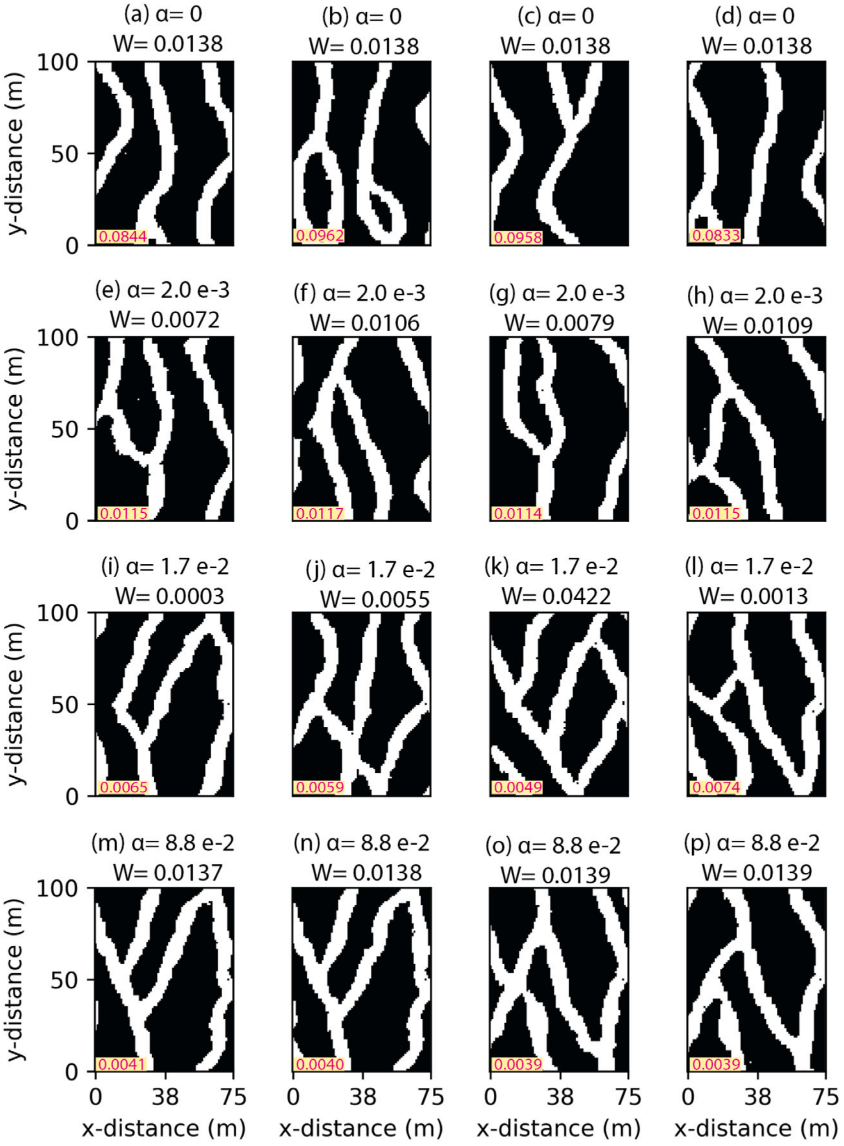

Fig. 10. Four samples from different power posteriors sampled with the ASMC-SGR 72p run: (a)–(d) $\alpha_t = 0$; (e)–(h) $\alpha_t = 0.002$; (i)–(l) $\alpha_t = 0.017$ and (m)–(p) $\alpha_t = 0.088$. The corresponding weights W are shown and the root mean square error (RMSE) without units is indicated for each sample with the corresponding value of the reference model (modes 1 and 2) being 0.0030.

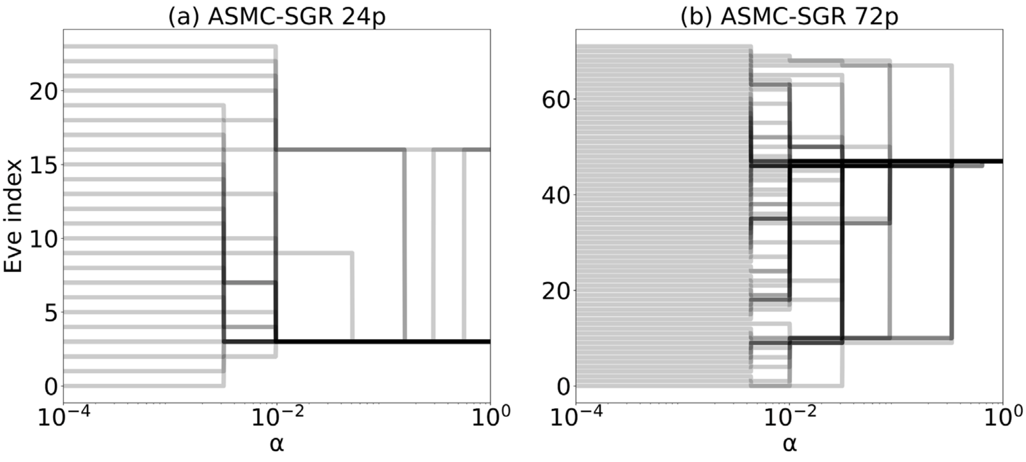

Fig. 11. Evolution of the Eve indices for the (a) ASMC-SGR 24p and (b) ASMC-SGR 72p in the range from $\alpha = 10^{-4}$ until $\alpha = 1$. The opacity of the lines is proportional to the number of particles that have the same Eve index (same origin) at a given α.

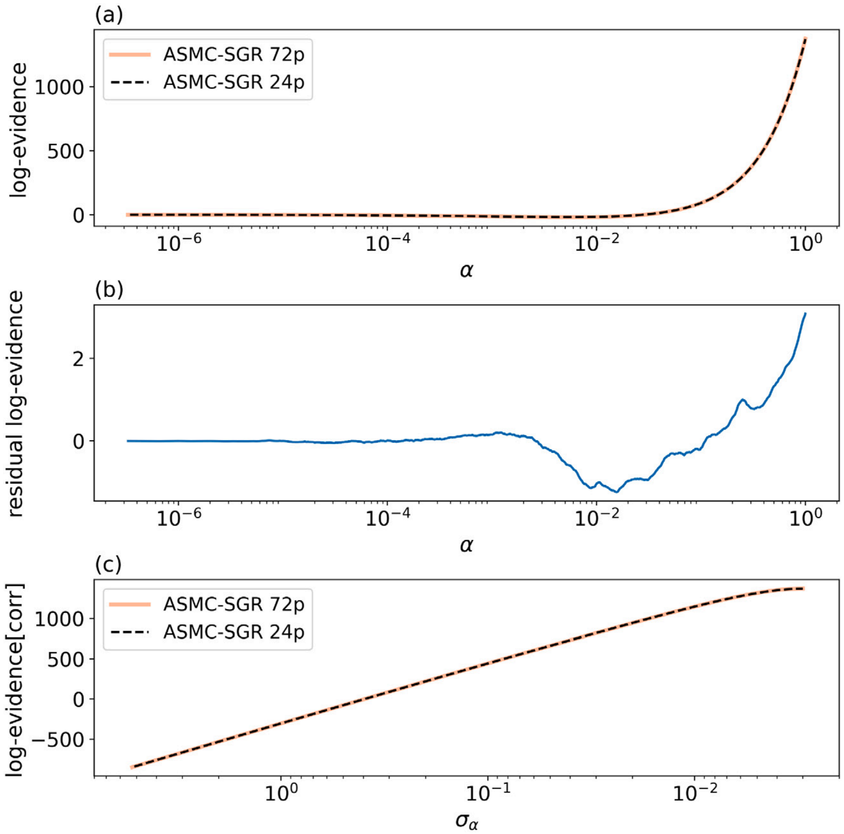

Fig. 12. (a) Natural log-evidence evolution vs. α for ASMC-SGR using 24 and 72 particles; (b) difference of the log-evidence estimates for the two test cases vs. α; (c) corrected natural log-evidence (Eq. (18)) vs. different assumptions of σ_α.

The power posterior approximations can also be interpreted as posterior PDF approximations for different assumed data error levels (Fig. 9). By raising the likelihood function to an inverse temperature α that is less that 1, the impact on the reduced log-likelihood is the same as if increasing the assumed error level. That is, flattening the likelihood and, thereby, enhancing the freedom of the exploration. A similar effect is obtained by decreasing the number of data points considered m_d ($\alpha \propto m_d$ in Eq. (3)): keeping a subset of the original observations will have a similar impact as reducing α or increasing σ^2. Tempering, assuming artificially high data errors or reducing the number of data are not uncommon in the literature when addressing challenging Bayesian inversions (e.g., Juda and Renard (2021)). This results in an easier to solve, but different, inverse problem that is conservative in the sense that the posterior mean is less informative and the posterior variance is larger than for the original problem. One important advantage of ASMC is that it explores all these intermediate problems, but also use the information gained to sample the original posterior PDF that is unfeasible for many other methods. Similarly, the evidence computations can be re-scaled to correspond to different assumptions of data error levels (Fig. 12).

In field applications, the data error level is typically poorly known. ASMC can then be very helpful, as one could assume a noise level that is likely too low and then obtain approximations of several power posterior corresponding to different (larger) error assumptions. One could then consider choosing an optimal error level based on the ASMC intermediate results using the relationship between α and σ. For instance, one could perhaps choose the error level and the corresponding posterior (and evidence) approximations by considering the divergence between the reference target log-likelihood and the tempered log-likelihoods with increasing α. In Fig. 2a there is no such divergence as the true data error level is assumed. This would be much more efficient than running multiple MCMC runs with different assumptions concerning σ.

An alternative and somewhat related method to solve inverse problems with SGR is Population Expansion (PoPEx) introduced by Jäggli et al. (2017, 2018). This method is similar to ASMC in the sense that the proposal distribution progressively evolves along the run towards the posterior PDF. These evolving distributions provide information maps built to efficiently select conditioning data for new SGR model proposals based on previously sampled high-likelihood models. The posterior PDF is approximated by iteratively expanding the set of models along the run. The corrected PoPEx algorithm by Jäggli et al. (2018) can be interpreted as an adaptive importance sampling algorithm (Naylor and Smith, 1988), in which the evolving proposal distribution is the importance distribution and the posterior PDF is the target distribution. This is different from ASMC where the importance sampling relies on consecutive power posteriors. Compared with PoPEx, ASMC also includes resampling steps, thereby, avoiding the degeneracy that often seems to plague PoPEx. To address this problem, Jäggli et al. (2018) artificially reweigh the weights in order to achieve a lower variance and, hence, a richer representation of the approximated posterior.

5. Conclusions

Tempering of likelihood functions is used in a wide variety of Bayesian methods to enhance posterior exploration and for evidence computations, particularly when confronted with high-dimensional and multimodal posterior PDFs that standard MCMC methods often struggle with. We demonstrate that adaptive sequential Monte Carlo (ASMC) outperforms parallel tempering (PT) when using sequential geostatistical resampling (a multiple-point statistics approach) as model proposal scheme in the context of a challenging synthetic groundwater transport inverse problem involving 7500 model parameters with a bimodal posterior PDF. ASMC is found to be considerably more effective in locating the two posterior modes and to sample states with likelihoods that are in agreement with the data noise. The algorithm has a simple implementation and demands a minimal user effort in terms of tuning due to its adaptive features. Furthermore, it also estimates the evidence (marginal likelihood) at almost no additional computational cost. The intermediate results of the algorithm can be used to determine the posterior means, standard deviations and evidences corresponding to different assumptions of data errors. This can be very helpful as it avoids pre-defining one standard deviation on the noise (or doing many MCMC runs with different assumed errors) and it allows assessing how the posterior changes from the prior through a number of intermediate power posteriors to the targeted posterior PDF. The method is versatile, robust and very well suited for parallelization and could have wide applicability to solve inverse problems arising in the field of water resources using a wide range of model parameterizations, forward solvers and model proposal schemes. In the future, we will seek speed-ups through surrogate modeling to enable a larger number of particles or longer runs and, thereby, improve the posterior estimations further for a given computational cost. Indeed, our examples with 24 and 72 particles could locate the posterior modes, but the computational budgets were insufficient to robustly sample the two posterior modes during the same ASMC run.

CRediT authorship contribution statement

Macarena Amaya: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Niklas Linde:** Conceptualization, Methodology, Formal analysis, Writing – review & editing, Supervision, Funding acquisition. **Eric Laloy:** Software, Writing – review & editing.

Declaration of competing interest

Niklas Linde, director of the project number 184574 reports financial support was provided by Swiss National Science Foundation.

Acknowledgments

This work was supported by the Swiss National Science Foundation (project number: 184574). Our ASMC code and the test examples are available at the following GitHub repository: https://github.com/amaya-macarena/ASMC-SGR. We are grateful for the insightful comments provided by two anonymous reviewers. We would also like to thank Prof. Arnaud Doucet (University of Oxford) for his helpful suggestions.

References

Alcolea, A., Renard, P., 2010. Blocking moving window algorithm: Conditioning multiple-point simulations to hydrogeological data. Water Resour. Res. 46 (8), W08511.

Amaya, M., Linde, N., Laloy, E., 2021. Adaptive sequential Monte Carlo for posterior inference and model selection among complex geological priors. Geophys. J. Int. 226 (2), 1220–1238.

Brunetti, C., Bianchi, M., Pirot, G., Linde, N., 2019. Hydrogeological model selection among complex spatial priors. Water Resour. Res. 55 (8), 6729–6753.

Brunetti, C., Linde, N., Vrugt, J.A., 2017. Bayesian model selection in hydrogeophysics: Application to conceptual subsurface models of the South Oyster Bacterial Transport Site, Virginia, USA. Adv. Water Resour. 102, 127–141.

Chandra, R., Müller, R.D., Azam, D., Deo, R., Butterworth, N., Salles, T., Cripps, S., 2019. Multicore parallel tempering Bayeslands for basin and landscape evolution. Geochem. Geophys. Geosyst. 20 (11), 5082–5104.

Cotter, S.L., Roberts, G.O., Stuart, A.M., White, D., 2013. MCMC methods for functions: modifying old algorithms to make them faster. Statist. Sci. 28 (3), 424–446.

Davies, L., Ley-Cooper, A.Y., Sutton, M., Drovandi, C., 2021. Bayesian detectability of induced polarisation in airborne electromagnetic data using reversible jump sequential Monte Carlo. arXiv preprint arXiv:2109.00661.

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (3), 411–436.

Douc, R., Cappe, O., 2005. Comparison of resampling schemes for particle filtering. In: ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005. pp. 64–69.

Doucet, A., Johansen, A.M., 2011. A tutorial on particle filtering and smoothing: Fifteen years later. In: The Oxford Handbook of Nonlinear Filtering, Vol. 12. p. 3, (656–704).

Earl, D.J., Deem, M.W., 2005. Parallel tempering: Theory, applications, and new perspectives. Phys. Chem. Chem. Phys. 7 (23), 3910–3916.

Fu, J., Gómez-Hernández, J.J., 2009. A blocking Markov chain Monte Carlo method for inverse stochastic hydrogeological modeling. Math. Geosci. 41 (2), 105–128.

Gallovič, F., Valentová, L., Ampuero, J.-P., Gabriel, A.-A., 2019. Bayesian dynamic finite-fault inversion: 1. Method and synthetic test. J. Geophys. Res.: Solid Earth 124 (7), 6949–6969.

Gómez-Hernández, J.J., Wen, X.-H., 1998. To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. Adv. Water Resour. 21 (1), 47–61.

Gravey, M., Mariethoz, G., 2020. QuickSampling v1.0: a robust and simplified pixel-based multiple-point simulation approach. Geosci. Model Dev. 13 (6), 2611–2630.

Hammersley, J.M., Handscomb, D.C., 1964. General principles of the Monte Carlo method. In: Monte Carlo Methods. Springer Netherlands, Dordrecht, pp. 50–75.

Hansen, T.M., Cordua, K.S., Mosegaard, K., 2012. Inverse problems with non-trivial priors: efficient solution through sequential Gibbs sampling. Comput. Geosci. 16 (3), 593–611.

Jäggli, C., Straubhaar, J., Renard, P., 2017. Posterior population expansion for solving inverse problems. Water Resour. Res. 53 (4), 2902–2916.

Jäggli, C., Straubhaar, J., Renard, P., 2018. Parallelized adaptive importance sampling for solving inverse problems. Front. Earth Sci. 203.

Juda, P., Renard, P., 2021. An attempt to boost posterior population expansion using fast machine learning algorithms. Front. Artif. Intell. 4, 25.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. J. Amer. Statist. Assoc. 90 (430), 773–795.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.

Kong, A., Liu, J.S., Wong, W.H., 1994. Sequential imputations and Bayesian missing data problems. J. Amer. Statist. Assoc. 89 (425), 278–288.

Künze, R., Lunati, I., 2012. An adaptive multiscale method for density-driven instabilities. J. Comput. Phys. 231 (17), 5557–5570.

Laloy, E., Hérault, R., Jacques, D., Linde, N., 2018. Training-image based geostatistical inversion using a spatial generative adversarial neural network. Water Resour. Res. 54 (1), 381–406.

Laloy, E., Hérault, R., Lee, J., Jacques, D., Linde, N., 2017. Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. Adv. Water Resour. 110, 387–405.

Laloy, E., Linde, N., Jacques, D., Mariethoz, G., 2016. Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields. Adv. Water Resour. 90, 57–69.

Laloy, E., Rogiers, B., Vrugt, J.A., Mallants, D., Jacques, D., 2013. Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. Water Resour. Res. 49 (5), 2664–2682.

Lee, A., Whiteley, N., 2018. Variance estimation in the particle filter. Biometrika 105 (3), 609–625.

Linde, N., Ginsbourger, D., Irving, J., Nobile, F., Doucet, A., 2017. On uncertainty quantification in hydrogeology and hydrogeophysics. Adv. Water Resour. 110, 166–181.

Linde, N., Renard, P., Mukerji, T., Caers, J., 2015. Geological realism in hydrogeological and geophysical inverse modeling: A review. Adv. Water Resour. 86, 86–101.

Mariethoz, G., Caers, J., 2014. Multiple-Point Geostatistics: Stochastic Modeling with Training Images. John Wiley & Sons.

Mariethoz, G., Renard, P., Caers, J., 2010a. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resour. Res. 46 (11), W11530.

Mariethoz, G., Renard, P., Straubhaar, J., 2010b. The direct sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 46 (11), W11536.

Meles, G.A., Linde, N., Marelli, S., 2022. Bayesian tomography with prior-knowledge-based parametrization and surrogate modeling. arXiv preprint arXiv:2201.02444.

Mosegaard, K., Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. J. Geophys. Res.: Solid Earth 100 (B7), 12431–12447.

Naylor, J., Smith, A., 1988. Econometric illustrations of novel numerical integration strategies for Bayesian inference. J. Econometrics 38 (1–2), 103–125.

Neal, R.M., 2001. Annealed importance sampling. Stat. Comput. 11 (2), 125–139.

Renard, P., Allard, D., 2013. Connectivity metrics for subsurface flow and transport. Adv. Water Resour. 51, 168–196.

Reuschen, S., Jobst, F., Nowak, W., 2021. Efficient discretization-independent Bayesian inversion of high-dimensional multi-Gaussian priors using a hybrid MCMC. Water Resour. Res. 57 (8), e2021WR030051.

Reuschen, S., Xu, T., Nowak, W., 2020. Bayesian inversion of hierarchical geostatistical models using a parallel-tempering sequential Gibbs MCMC. Adv. Water Resour. 141, 103614.

Ruggeri, P., Irving, J., Holliger, K., 2015. Systematic evaluation of sequential geostatistical resampling within MCMC for posterior sampling of near-surface geophysical inverse problems. Geophys. J. Int. 202 (2), 961–975.

Sambridge, M., 2014. A parallel tempering algorithm for probabilistic sampling and multimodal optimization. Geophys. J. Int. 196 (1), 357–374.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Math. Geol. 34 (1), 1–21.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Zahner, T., Lochbühler, T., Mariethoz, G., Linde, N., 2016. Image synthesis with graph cuts: a fast model proposal mechanism in probabilistic inversion. Geophys. J. Int. 204 (2), 1179–1190.

Zhou, Y., Johansen, A.M., Aston, J.A., 2016. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. J. Comput. Graph. Statist. 25 (3), 701–726.