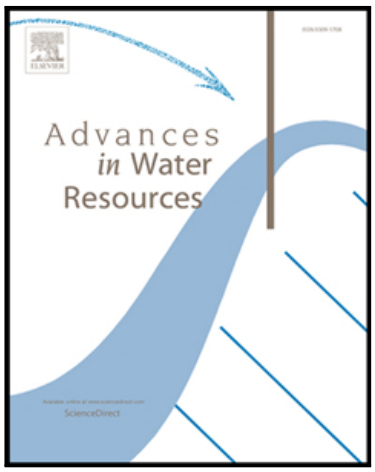

Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields

Eric Laloy, Niklas Linde, Diederik Jacques, Grégoire Mariethoz

PII: S0309-1708(16)30033-1
DOI: 10.1016/j.advwatres.2016.02.008
Reference: ADWR 2558

To appear in: *Advances in Water Resources*

Received date: 5 November 2015
Revised date: 16 February 2016
Accepted date: 16 February 2016

Please cite this article as: Eric Laloy, Niklas Linde, Diederik Jacques, Grégoire Mariethoz, Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields, *Advances in Water Resources* (2016), doi: 10.1016/j.advwatres.2016.02.008

Highlights

- Posterior sampling from a complex geological prior model is an important task
- Sequential Geostatistical Resampling cannot cope well with data-rich situations
- Merging parallel tempering with SGR is shown to substantially improve sampling
- A synthetic bimodal benchmark is used for testing sampling algorithms

Merging parallel tempering with sequential geostatistical resampling for improved posterior exploration of high-dimensional subsurface categorical fields

Eric Laloy[a,*], Niklas Linde[b], Diederik Jacques[a], Grégoire Mariethoz[c]

[a] *Institute for Environment, Health and Safety, Belgian Nuclear Research Centre*
[b] *Applied and Environmental Geophysics Group, Institute of Earth Sciences, University of Lausanne*
[c] *Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland.*

Abstract

The sequential geostatistical resampling (SGR) algorithm is a Markov chain Monte Carlo (MCMC) scheme for sampling from possibly non-Gaussian, complex spatially-distributed prior models such as geologic facies or categorical fields. In this work, we highlight the limits of standard SGR for posterior inference of high-dimensional categorical fields with realistically complex likelihood landscapes and benchmark a parallel tempering implementation (PT-SGR). Our proposed PT-SGR approach is demonstrated using synthetic (error corrupted) data from steady-state flow and transport experiments in categorical 7575- and 10,000-dimensional 2D conductivity fields. In both case studies, every SGR trial gets trapped in a local optima while PT-SGR maintains an higher diversity in the sampled model states. The advantage of PT-SGR is most apparent in a inverse transport problem where the posterior distribution is made bimodal by construction. PT-SGR then converges

*Corresponding author
Email addresses: elaloy@sckcen.be (Eric Laloy), niklas.linde@unil.ch (Niklas Linde), djacques@sckcen.be (Diederik Jacques), gregoire.mariethoz@unil.ch (Grégoire Mariethoz)

towards the appropriate data misfit much faster than SGR and partly recovers the two modes. In contrast, for the same computational resources SGR does not fit the data to the appropriate error level and hardly produces a locally optimal solution that looks visually similar to one of the two reference modes. Although PT-SGR clearly surpasses SGR in performance, our results also indicate that using a small number (16-24) of temperatures (and thus parallel cores) may not permit complete sampling of the posterior distribution by PT-SGR within a reasonable computational time (less than 1-2 weeks).

Keywords: Parallel tempering, sequential geostatistical resampling, training image, MCMC, multiple-point statistics

1. Introduction

A general Markov chain Monte Carlo (MCMC) simulation strategy based on sequential geostatistical resampling of spatially-distributed prior models has recently been introduced in the geosciences to infer the posterior distribution of subsurface property fields. The approach creates candidate fields by conditioning a geostatistical field realization from a possibly complex prior model to a randomly chosen fraction of the current state (and hence model/field) of the Markov chain. Both parametric (e.g., multi-Gaussian) and non-parametric prior models can be considered. The multi-Gaussian prior basically consists of a variogram model that encodes the 2-point statistics to be honored. As of non-Gaussian structures, they can be generated using a multiple-point statistics (MPS) simulation method. Such algorithms aim at reproducing not only the 2-point but also higher-order statistics found

in a so-called training image (TI). The TI is a gridded 2D or 3D conceptual representation of the target spatial field and can be either continuous or categorical (e.g., geologic facies image). It can either be built from a geologic model or from an observed structure (e.g., outcrop).

Various authors have independently introduced the probabilistic sequential geostatistical resampling (SGR) idea outlined above. The conference paper by Hansen et al. [13] was probably the first to describe the approach, considering so-called block updates where a box-shaped randomly located section of the current model is iteratively resimulated. Almost simultaneously, Fu and Gómez-Hernández [6] proposed a variant of the method that they termed blocking MCMC (BMCMC), which handles multi-Gaussian conditional simulation only. Shortly after, Mariethoz et al. [23] presented a SGR algorithm that resimulates a randomly chosen set of pixels/voxels rather than a contiguous block of pixels/voxels. This approach that was named iterative spatial resampling (ISR) was coupled with the direct sampling (DS) MPS algorithm of Mariethoz et al. [24] for resampling of both categorical and continuous priors. Finally, Hansen et al. [14] applied the approach by Hansen et al. [13] to more case studies and clarified the theoretical background of SGR, which they referred to as sequential Gibbs sampling (SGS). Even if all SGR variants presented above fall under the umbrella of Gibbs sampling theory [10], it is worth noting that the latter also forms a common framework for unconditional multi-Gaussian simulation [e.g., 7, 21].

The SGS [13, 14] and ISR [23] variants differ only in the geometry of the resimulated grid points, which is a box-shaped area for SGS and a set of points for ISR. For convenience, from here on we will follow Ruggeri et

al. [36] and use the generic name "SGR" for both SGS and ISR. To sample from a complex prior, SGR can in principle be implemented with any MPS algorithm. It is however very important that the considered MPS code can condition on a large fraction of grid data points (i.e., resimulating only a small fraction of the model). This is currently achieved only by pixel-based MPS techniques, for example, the DS and SNESIM [39] algorithms. We use DS in this study as it possesses good conditioning capabilities and is memory-efficient and relatively fast.

Ruggeri et al. [36] performed a systematic evaluation of SGR within a multi-Gaussian framework. They compared a gradual deformation [16] proposal mechanism with point and block SGR updates for a synthetic linear geophysical inverse problem using a multi-Gaussian prior and different numbers of measurements and noise levels. Results by Ruggeri et al. [36] suggest that the computational cost of producing one independent realization of the posterior by SGR is often prohibitively large even for relatively simple inverse problems. Ruggeri et al. [36] conclude that this finding warrants further research into model parameter reduction techniques that reduce parameter dimensionality and thus complexity of the inverse problem. This is in line with the work by Laloy et al. [20] who proposed a new reduced multi-Gaussian model parameterization, that is easily coupled with advanced MCMC sampling techniques [e.g., 19, 40, 42]. For training-image based inference of non-Gaussian/categorical structures, however, reducing the dimensionality of the parameter field is arguably more difficult. Even though a few model reduction methods have recently been proposed [17, 22, 41], the conceptual simplicity and flexibility of SGR remain attractive. To the best of our knowl-

edge, no critical analysis of SGR performance has been proposed so far for non-multi-Gaussian cases. Only rather simple problems involving either only 9 data points [23] or unrealistically large measurement errors [14, 15] have been considered. Using a very limited number of measurement data and/or large measurement errors makes the likelihood function rather flat and the posterior target is thus easy to sample. It is unclear at this stage whether training-image based SGR can handle more complicated problems with more realistically peaked likelihood functions. Furthermore, the posterior distribution might be multi-modal which, as shown herein, is not easily dealt with by standard SGR.

Parallel tempering (PT) [4, 5, 11], also called Metropolis-coupled MCMC, consists of parallel Markov chains that sample unnormalized target posterior density functions (pdf) raised to different powers, the inverse of which are called temperatures. The different chains regularly swap their temperatures, with the hot chains sampling a flattened posterior density landscape while the unit temperature(s) chain(s) explore(s) the desired distribution. The hot chains can more easily jump from one basin of attraction of the posterior to another, and this information is shared through swapping with the cold chain(s) that more intensively explore individual modes. This process can dramatically improve exploration of multi-modal posterior distributions while preserving a theoretically consistent sampling [4, 5, 11].

Up to now, application of PT to geosciences problems remains limited. In the area of reservoir simulation, Mohamed et al. [27] applied PT to the inversion of the Imperial College fault (ICF) model, considering 3 unknown model parameters and 10 parallel Markov chains. For this application, PT

was shown to explore much more efficiently the posterior parameter space than two stochastic optimization algorithms which got stuck within local optima. The study by Carter and White [3] is also focused on posterior exploration of the ICF model, considering from 1 to 13 unknown parameters and using 48 to 64 parallel chains. Carter and White [3] compared a simple random walk Metropolis (RWM) [26] sampler against the same algorithm equipped with PT for an ICF model with one unknown. This clearly demonstrated the superiority of PT for sampling the associated multi-modal posterior parameter distribution. Lastly, Sambridge [37] used as many as 380 parallel chains (and temperatures) to solve a synthetic trans-dimensional (the number of parameters is unknown) geophysical inverse problem for which the true model has 13 unknowns. Results by Sambridge [37] show a spectacular performance improvement by PT in terms of mixing and convergence towards the target data misfit. The use of PT thus appears to be beneficial not only for recovering multi-modal posterior distributions, but also for finding the maximum a posteriori estimate (MAP) of complex unimodal distributions. To date, PT applications in the geosciences have been concerned with rather low-dimensional parameter spaces. We hypothesize that PT may be advantageous for posterior inference in high parameter dimensions as well, such as spatially-distributed subsurface properties. We further suggest that this is possible even when considering a number of levels in the temperature ladder that is very small compared to the dimensionality of the parameter space. For completeness, we note that independently of our work, the idea of coupling PT with SGR also recently appeared as an outlook in the study by Ruggeri et al. [36].

In this paper, we illustrate the limits of the standard SGR method for posterior sampling of categorical fields, and benchmark a PT implementation with respect to both data fitting and diversity of the sampled posterior distribution. We refer to the proposed algorithm as PT-SGR for parallel tempering SGR. In contrast to previous work with PT in the geosciences, the inverse problems considered herein are quite high-dimensional (7575 to 10,000 sampled parameters). Moreover, only a relatively limited amount of parallel chains are used: from 16 to 24. This allows for parallel implementation on workstation computers or small clusters. Besides PT, we also investigate which settings of the SGR algorithm achieve the best performance for categorical field inference. Our proposed PT-SGR approach is demonstrated using synthetic (error corrupted) data from two flow and transport experiments in categorical 10,000- and 7575-dimensional 2D hydraulic conductivity fields that represent a channelized aquifer. These inverse problems involve realistically complex likelihood landscapes, one of which is made bimodal by construction.

This paper is organized as follows. Section 2 presents the different elements of our inversion approach. This is followed in section 3 with numerical experiments which include a performance analysis of SGR for different algorithmic settings, and a benchmarking of PT-SGR against SGR for the same multi-core computational resources. Section 4 then provides further discussion of the performance and limitations of our method and discusses possible future developments. Finally, section 5 concludes this paper with a summary of the most important findings.

2. Methods

2.1. Bayesian inference

A common stochastic representation of the forward problem is

$$F(\boldsymbol{\theta}) = \mathbf{d} + \mathbf{e}, \tag{1}$$

where $F(\boldsymbol{\theta})$ is a deterministic, error-free forward model that expresses the relation between the unknown parameters $\boldsymbol{\theta}$ and the measurement data $\mathbf{d} = (d_1, \ldots, d_N) \in \mathbb{R}^N, N \geq 1$, and the noise term $\mathbf{e}$ lumps all sources of errors.

In the Bayesian paradigm, parameters in $\boldsymbol{\theta}$ are viewed as random variables with a posterior pdf, $p(\boldsymbol{\theta}|\mathbf{d})$, given by

$$p(\boldsymbol{\theta}|\mathbf{d}) = \frac{p(\boldsymbol{\theta})\, p(\mathbf{d}|\boldsymbol{\theta})}{p(\mathbf{d})} \propto p(\boldsymbol{\theta})\, L(\boldsymbol{\theta}|\mathbf{d}), \tag{2}$$

where $p(\boldsymbol{\theta})$ denotes the prior distribution of $\boldsymbol{\theta}$ and $L(\boldsymbol{\theta}|\mathbf{d}) \equiv p(\mathbf{d}|\boldsymbol{\theta})$ signifies the likelihood function of $\boldsymbol{\theta}$. The normalization factor $p(\mathbf{d}) = \int p(\boldsymbol{\theta})\, p(\mathbf{d}|\boldsymbol{\theta})\, d\boldsymbol{\theta}$ can be obtained from numerical integration over the parameter space so that $p(\boldsymbol{\theta}|\mathbf{d})$ is a proper pdf that integrates to unity. The quantity $p(\mathbf{d})$ is generally difficult to estimate in practice but is not required for parameter inference when the parameter dimensionality is fixed. In the remainder of this paper, we will thus focus on the unnormalized density $p(\boldsymbol{\theta}|\mathbf{d}) \propto p(\boldsymbol{\theta})\, L(\boldsymbol{\theta}|\mathbf{d})$.

To avoid numerical over- or underflow, it is convenient to work with the logarithm of $L(\boldsymbol{\theta}|\mathbf{d})$ (log-likelihood), $\ell(\boldsymbol{\theta}|\mathbf{d})$, instead of $L(\boldsymbol{\theta}|\mathbf{d})$. If we assume $\mathbf{e}$ to be normally distributed, uncorrelated and with known constant variance, σ_e^2, the component of $\ell(\boldsymbol{\theta}|\mathbf{d})$ that depends on $\boldsymbol{\theta}$ can be written as

$$\ell(\boldsymbol{\theta}|\mathbf{d}) = -\frac{1}{2}\sigma_e^{-2} \sum_{i=1}^{N} \left[d_i - F_i(\boldsymbol{\theta})\right]^2, \tag{3}$$

where the $F_i(\boldsymbol{\theta})$ are the simulated equivalents to the $i = 1, \cdots, N$ measurement data, d_i.

An exact analytical solution of $p(\boldsymbol{\theta}|\mathbf{d})$ is not available for the type of inverse problems considered herein. We therefore resort to MCMC simulation to generate samples from the posterior pdf [see, e.g., 32]. The SGR algorithm independently developed by Hansen et al. [13, 14] and Mariethoz et al. [23] is used to approximate the posterior distribution. A detailed description of this sampling scheme can be found in the cited references and a convergence proof is given by Hansen et al. [14]. A brief summary of SGR is given in section 2.2.

2.2. Sequential geostatistical resampling from a training image

For a symmetric proposal distribution, the classical Metropolis acceptance probability, $\alpha(i, j)$ is given by

$$\alpha(i, j) = 1 \wedge \left(\frac{p(\boldsymbol{\theta}_j) L(\boldsymbol{\theta}_j|\mathbf{d})}{p(\boldsymbol{\theta}_i) L(\boldsymbol{\theta}_i|\mathbf{d})} \right), \tag{4}$$

where the function $\wedge$ takes the minimum of the left and right hand side numbers. For complex prior models, however, computing $p(\boldsymbol{\theta})$ might be difficult if not impossible. To overcome this limitation, Mosegaard and Tarantola [28] introduced a different version of the Metropolis algorithm in which the prior probabilities, $p(\boldsymbol{\theta})$, need not to be computed. The approach of Mosegaard and Tarantola assumes that a generating algorithm, $G = q(i, j)$, exists that is able to sample from $p(\boldsymbol{\theta})$ directly, in such a way that any proposal, $\boldsymbol{\theta}_j$, created by perturbation of the current state, $\boldsymbol{\theta}_i$, is itself a draw from $p(\boldsymbol{\theta})$. The Metropolis acceptance probability of a move from $\boldsymbol{\theta}_i$ to $\boldsymbol{\theta}_j$ can then be

reduced to

$$\alpha(i,j) = 1 \wedge \left(\frac{L(\boldsymbol{\theta}_j|\mathbf{d})}{L(\boldsymbol{\theta}_i|\mathbf{d})} \right). \tag{5}$$

Later called "extended Metropolis" sampler [14], this approach forms the basis of SGR. After initializing the chain with $\boldsymbol{\theta}_i$ drawn from $p(\boldsymbol{\theta})$, the latter proceeds in the three following steps to generate a Markov chain. First a candidate model, $\boldsymbol{\theta}_j$, is generated by resimulating a random fraction of the current state, $\boldsymbol{\theta}_i$, according to the prior model distribution. Since the conditioning points are chosen at random, this mechanism corresponds to a symmetric proposal distribution, $q(i,j)$, thus honoring the detailed balance condition: $q(i,j) = q(j,i)$ [see, e.g. 32, for theoretical details about MCMC]. Next, $\boldsymbol{\theta}_j$ is either accepted or rejected using equation (5). Finally, the chain either moves to $\boldsymbol{\theta}_j$ ($\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_j$) if the proposal is accepted, or remains at its current location ($\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i$) otherwise. Upon convergence of the chain, the generated model states constitute a set of representative draws from the posterior pdf.

Priors with complex structures can be handled by using a MPS algorithm that samples from a prescribed training image, which acts as prior model. Similarly as Mariethoz et al. [23], we use herein the DS method by Mariethoz et al. [24] as generating and conditioning algorithm. The selected TI is the most classical 250 × 250 binary image representing a channelized aquifer (not shown) that was introduced by Strebelle [39] [see also Figure 4a in 24].

2.3. Parallel tempering sequential geostatistical resampling

In parallel tempering [11, 5, 4], a temperature ladder, $\mathbf{T} = [T_1, \cdots, T_n]$ with $T_1 = 1 < T_2 < \cdots < T_n$, is used to increasingly flatten either the

posterior density

$$p\left(\boldsymbol{\theta}, T|\mathbf{d}\right) \propto \left[p\left(\boldsymbol{\theta}\right) L\left(\boldsymbol{\theta}|\mathbf{d}\right)\right]^{1/T}, \tag{6}$$

or the likelihood function

$$p\left(\boldsymbol{\theta}, T|\mathbf{d}\right) \propto p\left(\boldsymbol{\theta}\right) L\left(\boldsymbol{\theta}|\mathbf{d}\right)^{1/T}, \tag{7}$$

using a so called temperature T [12]. When $T \to \infty$ in equation (7), the tempered distribution $p\left(\boldsymbol{\theta}, T|\mathbf{d}\right)$ becomes the prior distribution $p\left(\boldsymbol{\theta}\right)$. Conversely, $p\left(\boldsymbol{\theta}, T|\mathbf{d}\right)$ becomes the posterior pdf, $p\left(\boldsymbol{\theta}|\mathbf{d}\right)$, when $T = 1$. In this work, $p\left(\boldsymbol{\theta}\right)$ does not need to be calculated explicitly as the MPS algorithm generates proposals directly from $p\left(\boldsymbol{\theta}\right)$ which is formed by the TI. In this case, equations (6) and (7) reduce to $p\left(\boldsymbol{\theta}, T|\mathbf{d}\right) \propto L\left(\boldsymbol{\theta}|\mathbf{d}\right)^{1/T}$.

Each tempered chain undergoes two possible moves throughout sampling: within-chain and between-chain proposals. In our proposed PT-SGR implementation, the within-chain proposal consists of a standard SGR update where a random fraction of the current state of the chain is resimulated according to the TI. The between-chain proposal consists of a swap of model states at two temperature levels i and j

$$\left[\left(\boldsymbol{\theta}_i, T_i\right), \left(\boldsymbol{\theta}_j, T_j\right)\right] \to \left[\left(\boldsymbol{\theta}_i, T_j\right), \left(\boldsymbol{\theta}_j, T_i\right)\right], \tag{8}$$

where $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ are the model parameter vectors in chains i and j immediately before the proposed swap. Exchange swap proposals improve the sampling at two levels. At the beginning of the search, they make it easier for the unit temperature ($T = 1$) chain to access regions of the model space with high posterior probabilities that are well removed from its current position. After burn-in, they allow for the unit temperature chain to jump between multiple peaks of the posterior density landscape.

Using equation (7), the Metropolis acceptance probability, $\alpha_s(i,j)$, of an exchange swap between models $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ with temperatures T_i and T_j is given by [12]

$$\alpha_s(i,j) = 1 \wedge \frac{L(\boldsymbol{\theta}_j|\mathbf{d})^{1/T_i}\, p(\boldsymbol{\theta}_j)\, L(\boldsymbol{\theta}_i|\mathbf{d})^{1/T_j}\, p(\boldsymbol{\theta}_i)}{L(\boldsymbol{\theta}_j|\mathbf{d})^{1/T_j}\, p(\boldsymbol{\theta}_j)\, L(\boldsymbol{\theta}_i|\mathbf{d})^{1/T_i}\, p(\boldsymbol{\theta}_i)}. \tag{9}$$

Canceling the equivalent terms in the numerator and denominator and performing some reorganization leads to

$$\alpha_s(i,j) = 1 \wedge \left[\frac{L(\boldsymbol{\theta}_j|\mathbf{d})}{L(\boldsymbol{\theta}_i|\mathbf{d})}\right]^{(1/T_i - 1/T_j)}, \tag{10}$$

where the prior distributions $p(\boldsymbol{\theta}_i)$ and $p(\boldsymbol{\theta}_j)$ do not appear, thereby allowing us to couple PT with SGR.

The temperature swapping (equation (8)) is often restricted to neighboring temperatures [4], either by considering all pairs of neighbors at once [e.g., 30] or only one pair at the time [e.g., 38]. Other authors [e.g., 37] instead proposed to randomly swap models independently of their temperature levels. All these schemes are valid in the sense that the unit temperature chain(s) will ultimately converge to the proper stationary distribution, provided that any given temperature is involved in possible exchange with no more than one other temperature [30] at each proposal step.

A general guideline is that exchange swaps must not happen too frequently such that (1) swapping-induced correlation in the tempered chains is reduced and (2) the risk for the unit temperature chain(s) to get trapped by cycling locally within a certain temperature interval is minimized [29, 30, 31]. Nonetheless, it has also been shown that optimal sampling performance is attained with a relatively high frequency of the exchange swap proposals

[29]. The mean acceptance probability, α_s, of a swap move for a given pair (equation (10)) is recognized to be another important diagnostic of parallel tempering performance. Obviously, an excessively small α_s will hamper exploration by the unit temperature chain. As α_s increases, however, the tempered chains will tend to keep exchanging each other's models without creating new configurations, thereby slowing down posterior sampling [29]. Optimal α_s values of about 20% [31] and 39% [30] have been proposed under certain conditions, whereas good sampling performance was found with a α_s value as low as 8% [29].

A pseudo-code of the proposed PT-SGR algorithm is as follows.

Algorithm 1 Parallel tempering sequential geostatistical resampling

1: **procedure** PT-SGR($\mathbf{T}$,m,n,p_{SW},$meth_{SW}$,$\mathbf{P}$) ▷

$\mathbf{T}$ is the temperature ladder of size n (with n even), m is the number of MCMC iterations, p_{SW} is the probability of performing a swap update with temperature selection procedure $meth_{SW}$, and $\mathbf{P}$ encapsulates the SGR algorithmic settings (e.g., ϕ, ...).

2: **for** $i = 1, \cdots, m$ **do** ▷ Loop over MCMC iterations

3: **for** $j = 1, \cdots, n$ **do in parallel** ▷ Loop over temperature ladder

4: $p\left(\boldsymbol{\theta}, T_j|\mathbf{d}\right)_i \leftarrow \mathbf{SGR}_{\mathrm{MCMC}}(T_j, \mathbf{P})$ ▷ classical MCMC move with SGR. The $j = 1, \cdots, n$ updates are done in parallel.

5: **end for**

6: **if** $p_{SW} > U\left(0, 1\right)$ **then**▷ Perform a swap update with probability p_{SW}, with $U\left(0, 1\right)$ indicating an uniform random draw between 0 and 1.

7: $\mathbf{r} = \mathrm{SelectPairs}(n, meth_{SW}, i)$ ▷ Select pairs of temperatures

8: **for** $j = 1, \cdots, n/2$ **do**

9: $v = \mathbf{r}(2 \times (j-1) + 1) \quad w = \mathbf{r}(2 \times j)$ ▷ Propose swapping of selected pairs of chains, one possible exchange swap per temperature

10: $\alpha\left(v, w\right) \leftarrow 1 \wedge \left[\dfrac{L\left(\boldsymbol{\theta}_w|\mathbf{d}\right)}{L\left(\boldsymbol{\theta}_v|\mathbf{d}\right)}\right]^{(1/T_v - 1/T_w)}$

11: **if** $\alpha < U\left(0, 1\right)$ **then** ▷ Swap chain temperatures

12: $T_v \leftarrow T_w, \quad T_w \leftarrow T_v$

13: **end if**

14: **end for**

15: **end if**

16: **end for**

17: **end procedure**

As discussed above, there are several options for selecting temperature pairs at swapping time. Two such options are given below: considering either randomly located or adjacent temperatures, with each temperature level involved in one exchange swap. In the case of adjacent temperatures, the same pairs can of course not be selected every time. When swapping occurs, the chosen pairs are thus either $(1,2),(3,4),\cdots$ or $(2,3),(4,5),\cdots$ depending on whether the iteration number is even or odd [e.g., 30].

Algorithm 2 Selection of temperature indices at swapping time

procedure $\mathbf{r}$ = SelectPairs(n,$meth_{SW}$,i)
  if $meth_{SW}$ = "*random*" **then** ▷ Select temperatures randomly
    $\mathbf{r} = permute(n)$ ▷ Create a random permutation of the temperature indices
  elseIf $meth_{SW}$ = "*adjacent*" ▷ Consider all adjacent pairs
    if $mod(i) = 1$ **then** ▷ MCMC iteration number is odd
      $\mathbf{r} = [1, \cdots, n]$
    else ▷ MCMC iteration number is even
      $\mathbf{r} = [2, \cdots, n-1, 1, n]$
    end if
  end if
end procedure

3. Case studies

3.1. Case study 1: steady-state flow

Our first synthetic case study considers steady-state head data collected at various locations within a channelized 2D aquifer (Figure 1a). The 100

$\times$ 100 modeling domain lies in the $x - y$ plane with a grid cell size of 1 m. Channels and matrix are assigned hydraulic conductivity values of 1×10^{-2} m/s and 1×10^{-4} m/s, respectively. Steady state groundwater flow is simulated using MaFloT [18] which is a finite-volume algorithm for 2D flow and transport in porous media. We assume no flow boundaries at the upper and lower sides and fixed head boundaries on the left and right sides of the domain so that a lateral head gradient of 0.025 (-) is imposed, with water flow in the x-direction. A pumping well extracting 0.003 m^2/s is located at the center of the domain. Simulated heads are collected at 49 locations that are regularly spread over the domain (Figure 1a-b). These data were then corrupted with a Gaussian white noise using a standard deviation of 0.01 m, leading to a root-mean-square-error (RMSE) of 0.01 m for the measurement data (Figure 1c). This translates into a reference log-likelihood (component that depends on $\boldsymbol{\theta}$), $\ell(\boldsymbol{\theta}|\mathbf{d})$, of -24.5.

3.2. DS settings for case study 1

The parameters of the DS simulation used for case study 1 are a neighborhood made of 50 nodes and a distance threshold set to 0.05. This means that for any simulated node, the data event (pattern) made of the 50 closest neighbors is considered, and up to only 2 mismatching nodes are allowed [see 24, for details]. The maximum scanned fraction of the TI is set to 0.9.

3.3. SGR settings

Apart from the employed MPS algorithm, important SGR algorithmic settings are (1) the type of conditioning, that is, whether the pixels to resimulate are defined by a set of points that are distributed throughout the

model domain or if they all belong to a box-shaped area, and (2) the size of the randomly located model fraction that is resimulated, ϕ. The latter can be fixed beforehand, adapted during burn-in of the MCMC sampling or drawn randomly from a certain probability distribution. All of these three options are explored in this study. In this section, we study the sampling performance achieved by conditioning on points (S1), or on all the points outside a square block (S2) for different sizes of ϕ. For S1, the six following fractions were considered: 0.995, 0.99, 0.95, 0.75, 0.5 and 0.25. As of S2, ϕ was set to 0.5, 0.25, 0.1 and 0.05. For each combination of settings, the setup of case study 1 (section 3.1) was used to perform 4 different MCMC trials for a total of 5000 forward model runs. In this and all other MCMC experiments conducted in this study, we initialized each Markov chain by randomly sampling $p(\boldsymbol{\theta})$.

Figure 2 displays the resulting sampled root-mean-square error (RMSE) trajectories and mean acceptance rate (AR) of the MCMC. Averages of the 4 trials are presented. Clearly, resimulating a box-shaped area (S2) shows a superior performance with respect to data fitting. As expected, the AR decreases with ϕ for both S1 and S2. Large resimulated fractions induce low AR values, below 1 or 2% ($\phi_{\mathrm{S1}} = 0.995$ to $\phi_{\mathrm{S1}} = 0.5$ and $\phi_{\mathrm{S2}} = 0.5$ and $\phi_{\mathrm{S2}} = 0.25$). Such small AR values characterize a prohibitively slow evolution of the MCMC chain.

Based on the above findings, we decided to use the resimulation strategy S2 in all of the following tests. Since the optimal value of ϕ is likely to depend on the problem at hand, in the remainder of this paper and unless stated otherwise ϕ_{S2} is tuned online to try to reach an AR value of 20%

during the first 10% of the MCMC iterations. The motivation for this target value is based on the fact that an AR of about 23% is considered optimal for Gaussian proposal and target distributions whereas an AR in the range 10% - 50% is generally recommended [33, 34].

3.4. Parallel tempering settings

For the case studies considered in this paper, limited testing with the different swapping strategies described in section 2.3 showed no overwhelming advantage of any specific strategy. Nevertheless, randomly proposing to swap model states after every regular within-chain MCMC update appeared to be the most robust and efficient approach. We therefore do so for all of our numerical experiments. With respect to the α_s values of the individual tempered chains, we use a common loglinear temperature ladder [3, 35] with maximum level such that α_s is (almost) always comprised between 5% and 30%. The rationale behind a loglinear scale is that a pair of neighboring cool chains ($T = 1$ or slightly higher) likely needs a smaller temperature difference for an exchange swap to be accepted compared to a pair of hotter chains. Other settings that perform better may very well exist, the quest for which is beyond the scope of this study.

3.5. Convergence of the Markov chain Monte Carlo simulation

The use of multiple (unit temperature) Markov chains makes it possible to use of the potential scale reduction factor, $\hat{R}$ [9], for monitoring convergence of the MCMC sampling. The $\hat{R}$ statistic compares for each parameter of interest the average within-chain variance to the variance of all the chains mixed together. The closer the values of these two variances, the closer to

unity the value of $\hat{R}$. Values of $\hat{R}$ smaller than 1.2 are commonly deemed to indicate convergence to a limiting distribution. In principle, $\hat{R}$ offers a stronger convergence assessment than merely considering the moment when the sampled (log-)likelihood (and thus RMSE) values reach an equilibrium. The latter indeed signifies only that the posterior distribution has been located, whereas the former aims at evaluating whether it has been adequately explored. For example, in the study by Laloy et al. [see 20] it was found that 25 times more MCMC steps where needed to appropriately explore a 1000-dimensional posterior than to start sampling it. For the considered case studies and computational budgets, our simulation results indicate that $\hat{R}$-convergence is never achieved by SGR, no matter whether computed on the basis of 3 (randomly chosen) chains or all of the 16 (case study 1) or 24 (case study 2) independent trials. Unfortunately, the $\hat{R}$ statistic is not well suited for monitoring convergence of several unit temperature chains within a given tempered ensemble. Indeed, running PT-SGR for case study 1 with 3 out of the 16 temperatures set to 1 results in an exaggeratedly fast $\hat{R}$-convergence (not shown). The 10,000 individual $\hat{R}$ values may even jointly fall below 1.2 after less iterations than required for the unit chains to sample the appropriate likelihood values. The reason for this is that the swapping dynamics causes large correlations between states/models of neighboring cool chains. The within-chain variances thus become similar enough for $\hat{R}$ to be satisfied prematurely. We therefore refrain from using the $\hat{R}$ diagnostic. Instead we simply resort to the point in time when (log-)likelihood values start to fluctuate around a constant level to define burn-in. From this moment on our algorithm starts drawing posterior samples. One must bear in mind,

however, that given the large dimensionality of the considered problems, it is evident that the posterior target is not fully explored within our limited computational budget (10,000 to 25,000 MCMC iterations) and we do not claim to do so. By a convenient abuse of terminology, we nevertheless refer to the resulting set of posterior samples as the "posterior distribution".

3.6. Inversion results for case study 1

For this case study, a total of 10,000 MCMC iterations is allowed for both SGR and PT-SGR. For a classical single-chain SGR trial, this translates into 10,000 forward model evaluations. The PT-SGR algorithm with n temperature levels is best run with parallel calculation of both the n DS simulations and n forward model evaluations performed per MCMC iteration. This roughly leads to a similar CPU-time per MCMC iteration (from 5 to 10 s herein) between SGR and PT-SGR. One must note, however, that some minor additional computational time is needed for PT-SGR due to communication overhead, the extent of which depends on hardware- and software-specific details. Here PT-SGR is ran on a multi-core platform, with $n = 16$. A loglinear temperature ladder was selected between unity and a maximum temperature of 6, together with a single unit temperature chain.

Figure 3a depicts the sampled negative log-likelihood ($-\ell(\boldsymbol{\theta}|\mathbf{d})$) trajectories for the unit temperature PT-SGR chain and the 16 independent SGR trials. It is observed that the PT-SGR trial samples appropriate mean RMSE values after some 1250 iterations. In contrast, the basic SGR algorithm shows a large spread of trajectories. Overall, the PT-SGR chain converges towards the reference data misfit at least as fast as the fastest of the 16 SGR chains. It also takes about 8000 MCMC iterations for the mean of the 16 SGR tri-

als to reach the target RMSE value (not shown). For this particular run, this leads to a 6 times speed-up of PT-SGR. Limited additional testing with PT-SGR confirmed (I) a similar data fitting efficiency of PT-SGR to that of the best performing SGR trial and (II) a 5-8 times speedup of PT-SGR for locating the posterior compared to the mean SGR behavior. This speedup is accomplished by the (random) mixing across the whole temperature ladder (Figure 3b). Across the tempered PT-SGR chains, the AR associated with the regular and swap moves, α and α_s, are 24% (with range of 19% - 26%) and 19% (with range of 6% - 24%), respectively.

The posterior distribution sampled by PT-SGR is illustrated in Figure 4 that shows the reference field together with the posterior mean and 7 successive posterior realizations from the unit temperature chain. The posterior mean in Figure 4 resembles the true model (Figure 1a) relatively well and the derived posterior uncertainty is rather small (compare realizations c - i in Figure 4). With respect to SGR, each of the 16 sampling runs turns out to remain in a specific region of the model space, as depicted by Figure 5. Indeed, the variability within the individual Markov chains is quite limited: models sampled more than 8500 iterations apart look very similar both to each other and to the chain average. Though also limited, the variability sampled by the (unit temperature) PT-SGR chain is nevertheless larger than for any given SGR trial. This is confirmed by the mean autocorrelation functions (ACF) calculated for the two tested algorithms (Figure 6): the ACF of PT-SGR drops much more rapidly than that of SGR, and stabilizes around a 1.6 times smaller value: 0.33 against 0.54 for SGR. The stabilization around a value larger than zero is caused by the fact that some specific

binary grid elements never change of value throughout the considered set of MCMC draws. Label switching for these grid elements is proposed but the resulting models have always too low likelihood to be accepted by the Markov chain. An ACF value of 1 is thus assigned to these grid elements which influences the mean ACF. Herein 33% of the 10,000 grid elements have not been updated for PT-SGR, against up to 54% for SGR.

The above results show that for case study 1, running PT-SGR with 16 parallel chains is a better option than running 16 independent SGR chains. For the CPU budget needed by PT-SGR to start sampling the posterior, most of the SGR chains are still exploring parts of the prior that do not belong to the posterior (Figure 3a). Indeed, it requires about 6 times more computational time for the 16 SGR chains to jointly sample the posterior. As of posterior diversity, each SGR trial gets trapped in a small region of the posterior model space (Figure 5). The situation is arguably better for PT-SGR (see Figures 4 and 6) even though it is far from having explored the full posterior range.

3.7. Case study 2: tracer experiment

Our second case study uses simulated tracer breakthrough curves at different wells as measurement data. The modeling domain is 75 $\times$ 101 and is located in the $x - y$ plane with a grid cell size of 1 m. Channels and matrix are again assigned hydraulic conductivity values of 1×10^{-2} m/s and 1×10^{-4} m/s, respectively. Steady state groundwater flow and conservative transport are both simulated using MaFloT. No-flow boundaries are assumed at the left and right sides, and fixed head boundaries on the upper and lower sides of the domain. These fixed heads are set to 0 m at both sides, and 11

pumping wells individually extracting 0.0005 m^2/s of water are spaced 7 m apart along the horizontal line located at equal distance from the top and bottom sides (Figure 7a). The facies values at the 11 wells are assumed to be known exactly and serve for direct conditioning. A conservative tracer with concentration of 1 kg m^{-3} is applied within 8 model cells of the top and bottom boundaries using a step function. The x-y coordinates of these cells are (14,1), (30,1), (46,1), (62,1), (14,101), (30,101), (46,101), and (62,101) (Figure 7a). The background solute concentration is assumed to be 0.01 kg m^{-3}. Ignoring density effects, conservative transport of the tracer through the subsurface is simulated using open boundaries on all sides, and longitudinal and transverse dispersivities both set to 0.1 m. Solute transport was monitored during a period of 10 days with concentration measurements made every 8 hours in the 11 extraction wells, resulting into a total of 330 observations. These simulated data were then corrupted with a Gaussian white noise using a standard deviation equivalent to 3% of the mean observed concentration. This led to root-mean-square-error (RMSE) and log-likelihood ($\ell(\boldsymbol{\theta}|\mathbf{d})$) of 0.0030 kg m^{-3} and -165, respectively, for the measurement data (Figure 7c).

This setup has the attractive feature of causing the posterior facies distribution to include two distinct modes with equal probability. Indeed, the reference field of Figure 7a and its mirrored image shown in Figure 7b both lead to the exact same simulated concentration data and thus likelihood function value. We thus consider this rather challenging case study to be especially instructive as the posterior target is known to present (at least) two separate modes.

3.8. DS settings for case study 2

The parameters of the DS simulation used for case study 2 are a neighborhood made of 75 nodes, a distance threshold of 0.01 and a maximum scanned fraction of the TI of 0.9.

3.9. Inversion results for case study 2

A total of 25,000 MCMC iterations is used for this case study. The PT-SGR algortihm is again run on a multi-core machine with $n = 24$ and the computational cost incurred by 1 MCMC iteration is in the range of 20-30 s for this workstation. A loglinear temperature ladder between unity and a maximum temperature of 2 is selected together with a single unit temperature chain. Using such a small maximum temperature was needed given the trade-off between the number of available parallel cores and the complexity of the (log-)likelihood landscape. For instance, the peakier the likelihood function, the smaller the temperature intervals need to be for α_s to be significantly larger than zero. With 24 temperature levels, using a larger maximum temperature than 2 essentially results in a frequency of accepted swaps that is impractically low. In addition, the update mechanism of ϕ that is described in section 3.3 was slightly modified in an attempt to generate more diverse proposals while keeping the acceptance rate of regular MCMC moves reasonably high. Rather than a tuned constant resimulation block size, ϕ was taken as the (adapted) mean of a triangular pdf bounded between 0.01 and 0.25 from which the actual size of the block to be resimulated was drawn. This was similarly done for both SGR and PT-SGR.

The sampled $-\ell(\boldsymbol{\theta}|\mathbf{d})$ trajectories for the unit temperature PT-SGR chain and the 24 independent SGR trials are presented in Figure 8a. The

PT-SGR unit chain evolves towards the reference $-\ell(\boldsymbol{\theta}|\mathbf{d})$ value much faster than the fastest of the SGR trials. Furthermore, the spread of the SGR trajectories is rather large. After 25,000 iterations, only 4 trials are sampling $-\ell(\boldsymbol{\theta}|\mathbf{d})$ values in the range sampled by PT-SGR and 5 trials are still exploring areas associated with twice as large $-\ell(\boldsymbol{\theta}|\mathbf{d})$ values. Equation (5) can be used to calculate the probability of a direct jump from the reference model to the most likely model found by PT-SGR on the one hand, and the most likely model among the 24 SGR chains on the other hand. Doing so reveals that the most likely model sampled by PT-SGR is more than 1×10^7 times more likely than that of SGR. The PT-SGR algorithm thus clearly outperforms SGR for this case study.

Even if PT-SGR surpasses SGR, it is evident that the unit temperature PT-SGR chain fluctuates around a slightly larger $-\ell(\boldsymbol{\theta}|\mathbf{d})$ value than the reference value of -165. In fact, for iterations 10,000-25,000 the mean sampled $-\ell(\boldsymbol{\theta}|\mathbf{d})$ value by PT-SGR exceeds the reference value of 165 by 9% and the sampled range actually never contains it (Figure 8a). This means that the samples produced by PT-SGR are not representative of the posterior distribution. That said, this inverse problem is much more difficult to solve than for case study 1 (see section 3.6). This is because (I) the large amount of good quality (moderately corrupted) measurements (330) causes the two well separated (log-) likelihood modes to be more peaky, and (II) using transport data induces a more nonlinear relationship between model (parameters) and (log-) likelihood than using steady-state head data.

The AR associated with the PT-SGR chains are lower than for case study 1 but still acceptable: across the whole temperature ladder, α and α_s, are

8% (with range of 7% - 9%) and 14% (with range of 6% - 20%), respectively. The corresponding swap exchange dynamics looks visually good from iteration 10,000 onwards (Figure 8b). The PT-SGR unit temperature chain does however not mix well. Indeed, the chain basically cycles over the (nearly) same 6-7 models during the last 15,000 MCMC iterations (Figure 9). The reason for this is likely twofold. First, the maximum temperature of 2 does not flatten the likelihood enough for sufficient exploration by the hot chains. Second and most important, for this rather complex likelihood landscape the exchange swaps appear to be mostly performed in cycles between a few neighboring temperature levels with colder (hotter) samples almost never traveling to the highest (lowest) levels (not shown). The only way to solve this problem would therefore be to use a ladder with a much larger number of levels and a wider temperature range.

As depicted in Figure 9, the two reference modes (Figures 7ab) are very roughly recovered by the PT-SGR trial. Furthermore, the "left" reference mode (Figure 7a) appears to be better identified (compare Figure 2a with Figures 9c,d,f, and i). This finding is fairly positive given the relatively limited computational budget and the use of a small temperature ladder in regard to the problem dimensionality. Perhaps not surprisingly, the SGR performance is substantially worse. Here every independent chain is clearly trapped in one local optimum, which always has a larger data misfit than the reference RMSE of 0.0030 kg/m^{-3} (Figure 10). Visual inspection of the final states of the 24 SGR trials also shows that almost none of the model realizations looks similar to any of the two reference modes (see Figure 10 for three such examples). In average over the 24 trials, only 9% of the simulated

pixels present a different facies between iterations 10,000 and 25,000. As a result, the mean ACF of SGR takes a value as high as 0.82 at lag 5000 (Figure 11). With an ACF value of 0.13, PT-SGR produces 6 times less autocorrelated samples at lag 5000 (Figure 11).

4. Discussion

Our results demonstrate that the standard SGR approach cannot cope efficiently with situations where the measurement data are collected at a relatively high spatial and/or temporal density. Standard SGR has however been shown to work for a data-poor situation, where the information content of the data does not constrain much the facies distribution and the posterior uncertainty is thus quite large [e.g., 23].

Parallel tempering improves the SGR performance. Sampling of the complicated bimodal posterior distribution of case study 2 is hence much improved by parallel tempering, but not to the point of drawing samples from the correct stationary distribution within the allowed computing time and when using 24 temperatures (and thus parallel cores). Significantly increasing the number of temperatures, say by a factor 10, is expected to strongly enhance posterior sampling. Our future work will focus on two alternatives to simply increasing the available computing power. First, parallel tempering could be coupled with Wang-Landau (WL) sampling [2] for better exploration capabilities. The main principle of WL sampling is to adaptively bias the Metropolis acceptance probability in order to sample a flat histogram of posterior density values with pre-defined bins. The derived histogram thus contains an approximately equal amount of samples for each class of density

value, and these samples can then be reused to approximate the posterior distribution (e.g., via importance sampling or by seeding a new MCMC run). The method might however not help with the observed problem of sampling slightly too large (log-)likelihoods, and thus a wrong stationary distribution. Second and perhaps more promising is the use of a more informal ensemble-based multiscale approach. The latter would consist in sequentially solving the inference problem from an upscaled coarser scale to the finer scale of interest, in the spirit of the work by Gardet et al. [8] for the multi-Gaussian case. The underlying idea is that the (upscaled) parameter space can be scanned quickly using a coarse resolution, thereby allowing for the subsequent finer scale searches to concentrate on the most productive parts of the prior distribution. Starting from different random points would then eventually provide an ensemble of solutions that (informally) approximate the posterior target.

On a more practical level, the rationale for our DS settings deserves special attention. A neighborhood of 50 (case study 1) or 75 (case study 2) nodes may seem large [25] as the DS simulation time increases with the number of neighboring nodes. Nevertheless, using such large values was necessary to minimize the occurrence of artifacts in the generated models, which is caused by repeatedly conditioning on a large amount of grid points throughout the MCMC sampling. Combined with a large fraction of conditioning data (say $> 50\%$ of the image), a small neighborhood can sometimes result in model proposals that are somewhat degraded compared to the TI. It is indeed the restricted neighborhood size that gives freedom to the DS to produce structures that are different than those found in the TI. Also, all it takes for a

slightly degraded model to appear in the Markov chain is for the Metropolis rule (equation 5) to accept it. In other words, even if a model proposal with some artifacts is only rarely proposed, this model can persist in the Markov chain if the associated simulated data fit the observations sufficiently well. Even with the employed neighborhood of 50 nodes for case study 1 (see section 3.2), trials $\phi_{S1} = 0.5$ and $\phi_{S1} = 0.75$ of section 3.3 nevertheless showed artifacts in the proposed models, typically manifested by overly broad channels (not shown). To a lesser extent, other artifacts such as isolated patches and broken channels also occurred for trials $\phi_{S2} = 0.9$ and $\phi_{S2} = 0.95$ (not shown). The used distance threshold of 0.01 for case study 2 (see section 3.8) also incurs a larger computational cost than that of using the more common value of 0.05 [e.g., 23]. The value of 0.01 turned out to be required to (almost) systematically honor the point measurement data (see section 3.7).

Finally, it would be interesting to investigate the performance of parallel tempering when used in conjunction with the patch-based geostatistical resimulation algorithm by Zahner et al. [43] which uses graph cuts. This method has been shown to be 40 times faster than DS for generating a 2D model, with a resulting posterior distribution that is (at least) of similar quality as that obtained by using DS.

5. Conclusion

This study is concerned with the application of sequential geostatistical resampling (SGR) to high-dimensional categorical field inference problems that present realistically complex likelihood functions. We highlight the limitations of the classical SGR approach and propose a parallel tempering im-

plementation that, for a similar multi-core computing budget, provides much improved results with respect to both convergence towards the appropriate data misfit and sampling diversity. Two synthetic case studies are considered: a steady-state flow and a transport inverse problem, involving from 7501 to 10,000 unknowns. For the transport problem, the corresponding likelihood function is made bimodal with two well separated modes. In both case studies, every SGR MCMC chain gets trapped in a local optima while parallel tempering within sequential geostatistical resampling (PT-SGR) does not. The advantage of PT-SGR becomes more apparent for the bimodal inverse transport problem, for which PT-SGR is found to converge towards the reference data misfit much faster that SGR and to indicate the existence of two posterior modes. In contrast, for the same computational resources SGR appears to be barely able to appropriately fit the data and does almost not produce any single solution that looks visually similar to one of the two reference modes. Although PT-SGR outperforms SGR, our results also demonstrate that using a reasonably small number of temperatures (and thus parallel cores) in the range 16-24 may not allow sampling of the posterior distribution by PT-SGR within an affordable computational time. As an alternative to significantly increasing the number of temperatures and thus computational needs, coupling PT-SGR with Wang-Landau sampling and (2) reframing SGR within an ensemble-based multiscale optimization framework are two potentially useful approaches that will be investigated in future work. More generally, PT could also prove useful when used in conjunction with dimensionality reduction approaches.

6. Acknowledgments

A MATLAB code of the proposed PT-SGR approach is available from the first author. The patented DeeSee (DS) multiple-point statistics code is available for academic use upon request to one of its developers (Grégoire Mariethoz, Philippe Renard, Julien Straubhaar). Also, the synthetic measurements and forward modeling setup of our two case studies can be downloaded from `http://www.minds.ch/gm/downloads.htm`. We would like to thank the associated editor and three anonymous referees for their positive feedback and useful comments.

References

[1] Alcolea A, Renard P. Blocking Moving Window algorithm: Conditioning multiple-point simulations to hydrogeological data. Water Resour. Res. 2010;46:W08511. http://dx.doi.org/10.1029/2009WR007943.

[2] Bornn L, Jacob PE, Del Moral P, Doucet A. An adaptive interacting WangLandau algorithm for automatic density exploration. J. Comput. Graph. Stat. 2013; 22(3):749-773. http://dx.doi.org/10.1080/10618600.2012.723569.

[3] Carter JN, White DA. History matching on the Imperial College fault model using parallel tempering. Comput. Geosci. 2013;17:43–65. http://dx.doi.org/10.1007/s10596-012-9313-3.

[4] Earl DJ, Deem, MW. Parallel tempering: theory, applications, and new perspectives. Phys. Chem. Chem. Phys. 2005;7: 3910–3916.

[5] Falcioni M, Deem MW. A biased Monte Carlo scheme for zeolite structure solution. J. Chem. Phys. 1999;110: 1754–1766.

[6] Fu J, Gómez-Hernández JJ. A blocking Markov chain Monte Carlo method for inverse stochastic hydrogeological modeling. Math. Geosci. 2009;41:105–128. http://dx.doi.org/10.1007/s11004-008-9206-0.

[7] Galli A, Gao H. Rate of convergence of the Gibbs sampler in the Gaussian case. Math. Geol. 2001;33(6):653–677.

[8] Gardet C, Le Ravalec M, Gloaguen E. Multiscale parameterization of petrophysical properties for efficient history-matching. Math. Geosci. 2014;46(3):315-336. http://dx.doi.org/10.1007/s11004-013-9480-3.

[9] Gelman AG, Rubin DN. Inference from iterative simulation using multiple sequences. Stat. Sci. 1992;7:457–472.

[10] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. 1984;6(6):721–741.

[11] Geyer, CJ. Markov Chain Monte Carlo maximum likelihood. 1991. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, American Statistical Association, New York, pp. 156–163.

[12] Geyer, CJ. Chapter 11: Importance sampling, simulated tempering and umbrella sampling. 2011. In Brooks S, Gelman A, Jones GL, Meng XL (eds.) *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC, Boca Raton, pp. 295–306.

[13] Hansen TM, Mosegaard K, Cordua KC. Using geostatistics to describe complex a priori information for inverse problems. In: Ortiz, JM, Emery, X (eds.) VIII International Geostatistics Congress. 2008;vol. 1:329–338. Mining Engineering Department, University of Chile, Santiago.

[14] Hansen TM, Cordua KC, Mosegaard K. Inverse problems with non-trivial priors: efficient solution through sequential Gibbs sampling. Comput. Geosci. 2012;16:593–611. http://dx.doi.org/10.1007/s10596-011-9271-1.

[15] Hansen TM, Cordua KC, Looms MC, Mosegaard K. SIPPI: a Matlab toolbox for sampling the solution to inverse problems with complex prior information: Part 2 - Application to crosshole GPR tomography. Comput. Geosci. 2013;52: 481–492. http://dx.doi.org/10.1016/j.cageo.2012.10.001.

[16] Hu, LY. Gradual deformation and iterative calibration of Gaussian related stochastic models. Math. Geol. 2000;32(1):87–108.

[17] Khaninezhad MM, Jafarpour B, Li L. Sparse geologic dictionaries for subsurface flow model calibration: Part I. Inversion formulation. Adv. Water Resour. 2012;39(0):106121. http://dx.doi.org/10.1016/j.advwatres.2011.09.002.

[18] Künze R, Lunati I. An adaptive multiscale method for density-driven instabilities. J. Comput. Phys. 2012;231:5557–5570.

[19] Laloy E, Vrugt JA. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and

high-performance computing. Water Resour. Res. 2012:48(1). http://dx.doi.org/10.1029/2011WR010608.

[20] Laloy E, Linde N, Jacques D, Vrugt JA. Probabilistic inference of multi-Gaussian fields from indirect hydrological data using circulant embedding and dimensionality reduction. Water Resour. Res. 2015:51. http://dx.doi.org/10.1002/2014WR016395.

[21] Lantuéjoul C. Geostatistical simulation: models and algorithms, Springer; 2002.

[22] Lochbühler T, Vrugt JA, Sadegh M, Linde N. Summary statistics from training images as prior information in probabilistic inversion. Geophys. J. Int. 2015;201:157–171. http://dx.doi:10.1093/gji/ggv008.

[23] Mariethoz G, Renard P, Caers J. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resour. Res. 2010a; 46:W11530. http://dx.doi.org/10.1029/2010WR009274.

[24] Mariethoz G, Renard P, Straubhaar J. The Direct Sampling method to perform multiple-point geostatistical simulations. Water Resour. Res. 2010b;46:W11536. http://dx.doi.org/10.1029/2008WR007621.

[25] Meerschman E, Pirot G, Mariethoz G, Straubhaar J, Van Meirvenne M. A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. Comput. Geosci. 2013;52:307–324.

[26] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. J. Chem. Phys. 1953;21:1087–1092.

[27] Mohamed L, Calderhead B, Filippone M, Christie M, Girolami M. Population MCMC methods for history matching and uncertainty quantification. Comput. Geosci. 2012;16(2):423–436.

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

[29] Opps SB, Schoffield, J. Extended state-space Monte Carlo methods. Phys. Rev. E. 2001;60:056701. http://dx.doi.org/10.1103/PhysRevE.63.056701.

[30] Predescu C, Predescu M, Ciobanu CV. On the efficiency of exchange in parallel tempering Monte Carlo simulations. J. Phys. Chem. B. 2005;109:4189–4196.

[31] Rathore N, Chopra M, de Pabloa JJ. Optimal allocation of replicas in parallel tempering simulations. J. Chem. Phys. 2005;122:024111. http://dx.doi.org/0.1063/1.1831273.

[32] Robert, CP, Casella G. Monte Carlo statistical methods, second edition. Springer; 2004.

[33] Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. Appl. Prob. 1997;7:110–120.

[34] Roberts GO, Rosenthal JS. Optimal scaling for various Metropolis-Hastings algorithms. Stat. Sci. 2001;16:351–367.

[35] Romary T. Integrating production data under uncertainty by parallel interacting Markov chains on a reduced dimensional space. Comput. Geosci. 2009;13:103–122. http://dx.doi.org/10.1007/s10596-008-9108-8.

[36] Ruggeri P, Irving J, Holliger K. Systematic evaluation of sequential geostatistical resampling within MCMC for posterior sampling of near-surface geophysical inverse problems. Geophys. J. Int. 2015;202:961–975. http://dx.doi.org/10.1093/gji/ggv196.

[37] Sambridge M. A Parallel tempering algorithm for probabilistic sampling and multimodal optimization. Geophys. J. Int. 2014;196:357–374. http://dx.doi.org/10.1093/gji/ggt342.

[38] Schrek A, Fort G, Moulines E. Adaptive Equi-Energy Sampler: Convergence and Illustration. ACM Trans. Model. Comput. Simul.2013;23(1):5:1–5:27. http://doi.acm.org/10.1145/2414416.2414421.

[39] Strebelle S. Conditional simulation of complex geological structures using multiple point statistics. Math. Geol. 2002;34(1):1–22.

[40] ter Braak, CJ, Vrugt JA. Differential evolution Markov chain with snooker updater and fewer chains. Stat. Comput. 2008;18(4):435–446.

[41] Vo HX, Durlofsky LJ. A new differentiable parameterization based on Principal Component Analysis for the low-dimensional representation of complex geological models, Math. Geosci. 2014;46:775-813.

[42] Vrugt JA, Ter Braak, C, Diks C, Robinson BA, Hyman, JM, Higdon, D. Accelerating Markov chain Monte Carlo simulation by differential evolu-

tion with self-adaptive randomized subspace sampling. Int. J. Nonlin. Sci. Numer. Simul. 2009;10(3):273–290.

[43] Zahner T, Lochbühler T, Mariethoz G,Linde N. Image synthesis with graph cuts: a fast model proposal mechanism in probabilistic inversion. Geophys. J. Int. 2016;204(2):1179–1190. http://dx.doi.org/10.1093/gji/ggv517.

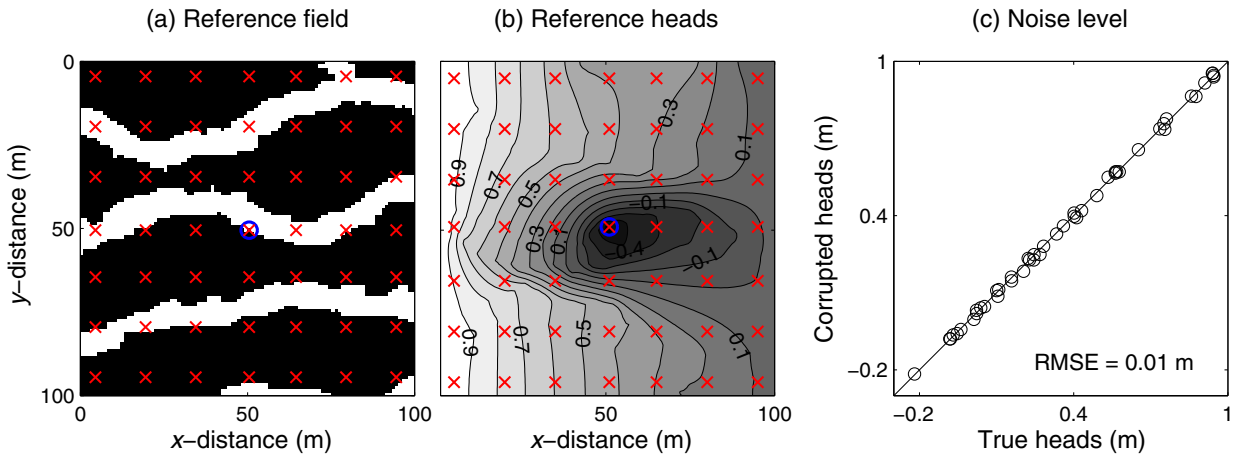

Figure 1: (a) Reference categorical field, (b) associated heads and (c) noise-corrupted measurement data used for case study 1. In subfigures a and b, the blue circle marks the location of the pumping well and the red crosses indicate piezometers.

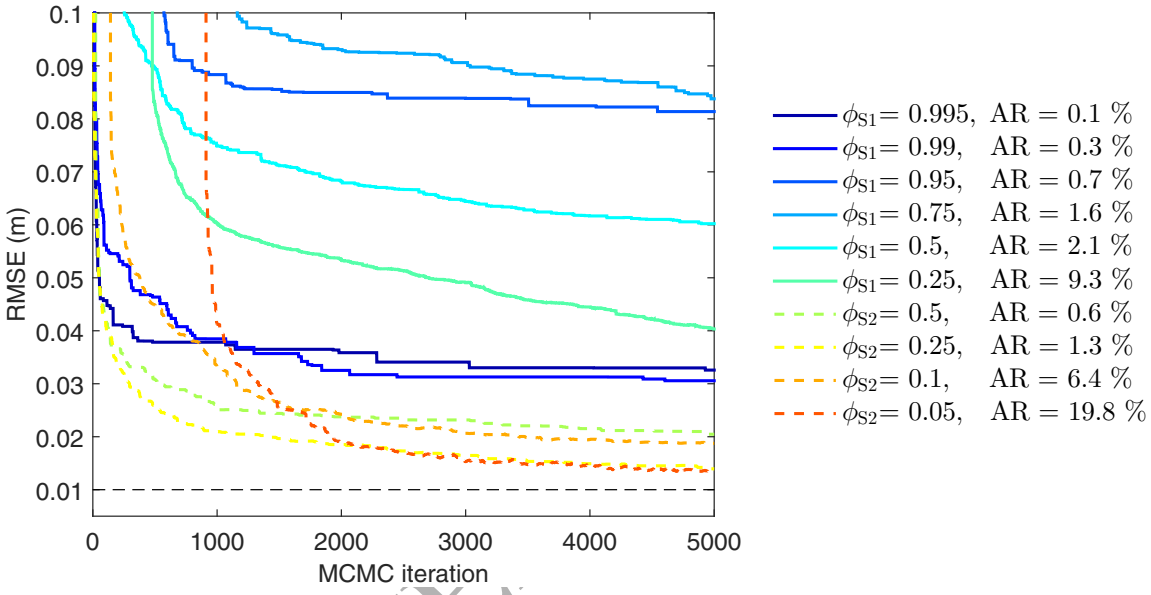

Figure 2: Trace plot of the mean sampled RMSE values across 4 repetitions for the tested conditioning strategies. Solid and dashed colored lines denote resimulating a set of points (S1) and a box-shaped area (S2), respectively. Each color represents a given size of the (randomly located) model fraction that is resimulated. The dashed black line signifies the true RMSE.

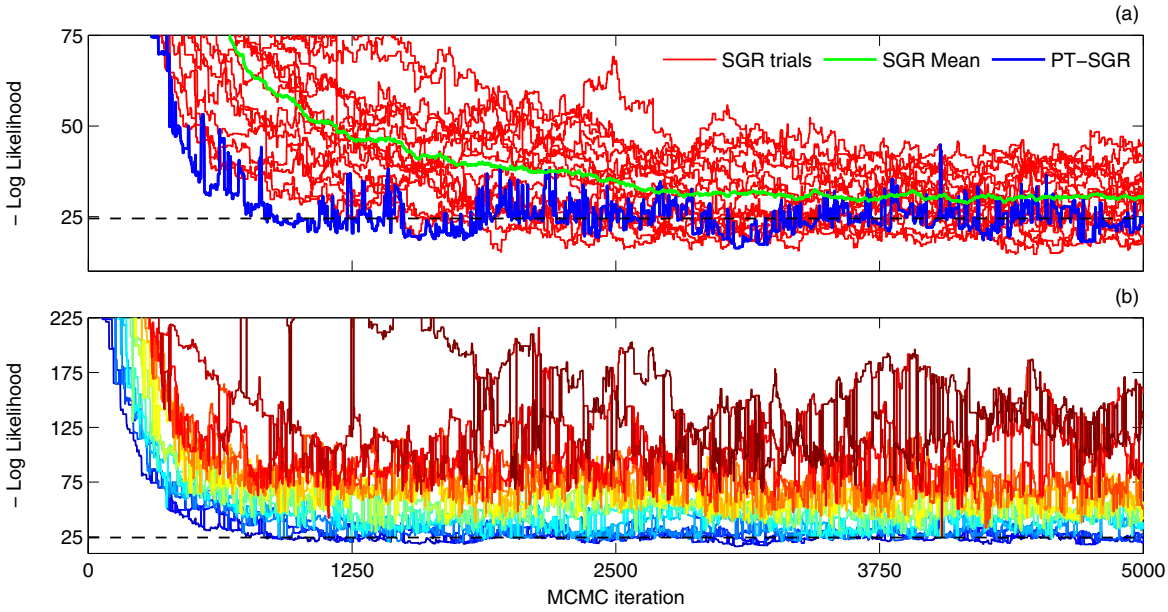

Figure 3: (a) Trace plot of the sampled negative log-likelihood values by the unit temperature chain evolved by PT-SGR (blue line) and the 16 independent SGR trials (red lines) for case study 1. The green line denotes the mean trajectory of the 16 SGR trials. (b) Trace plot of the sampled negative log-likelihood values by the 16 PT-SGR chains with each temperature coded with a different color. The temperature increases as the color varies from dark blue (temperature index of 1) to dark red (temperature index of 6). In both subfigures, the horizontal dashed black line denotes the true negative log-likelihood of 24.5, corresponding to a RMSE of 0.01 m.

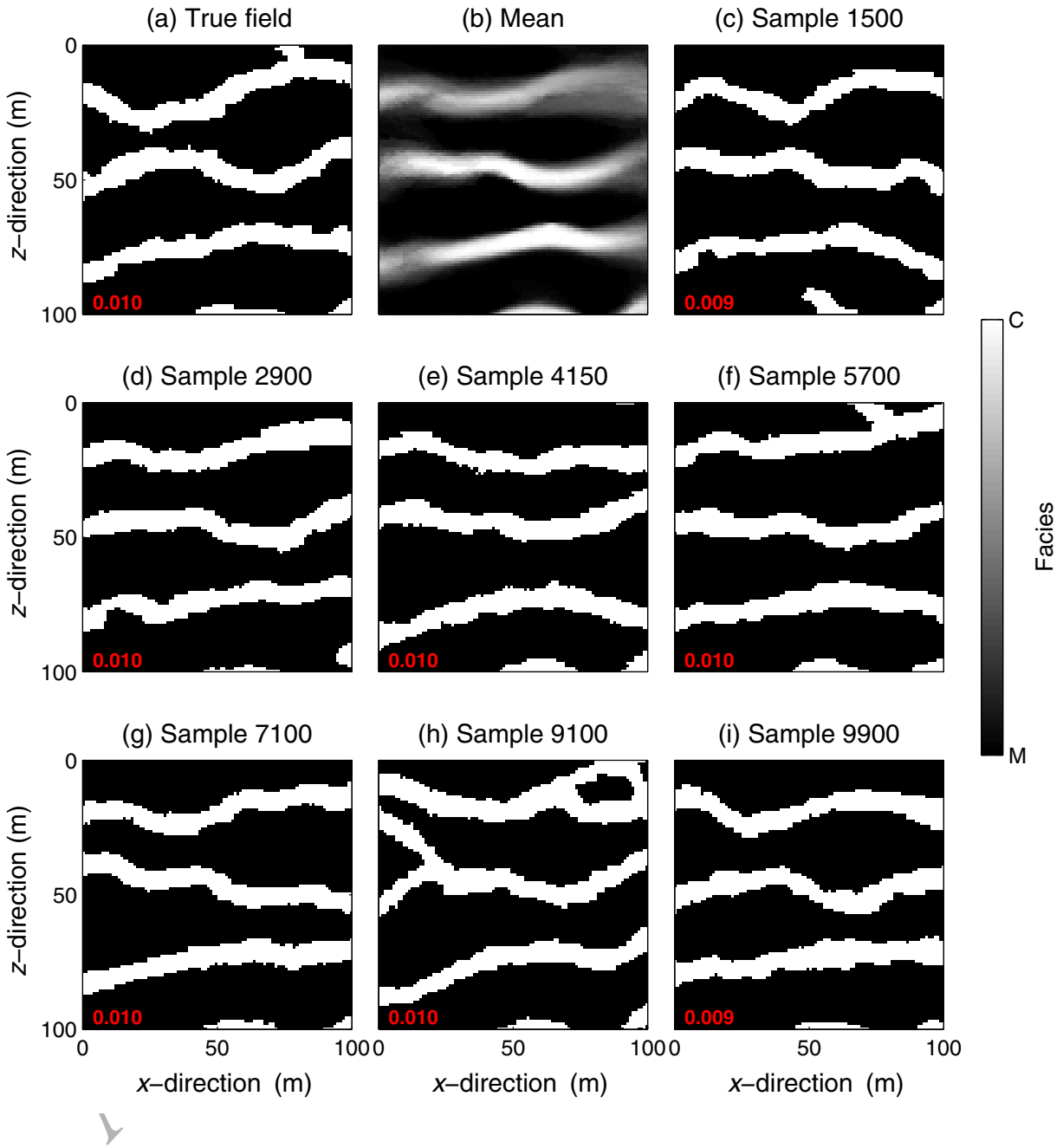

Figure 4: Posterior mean and 8 successive posterior realizations taken at regular intervals throughout sampling for the PT-SGR trial of case study 1. The posterior mean is computed on the basis of the samples produced by the unit temperature chains after a burn-in of 1500 MCMC iterations and using a thinning factor of 50, thus leading to a total of 170 posterior samples. The red number in the lower left corner of each plot is the corresponding RMSE (m).

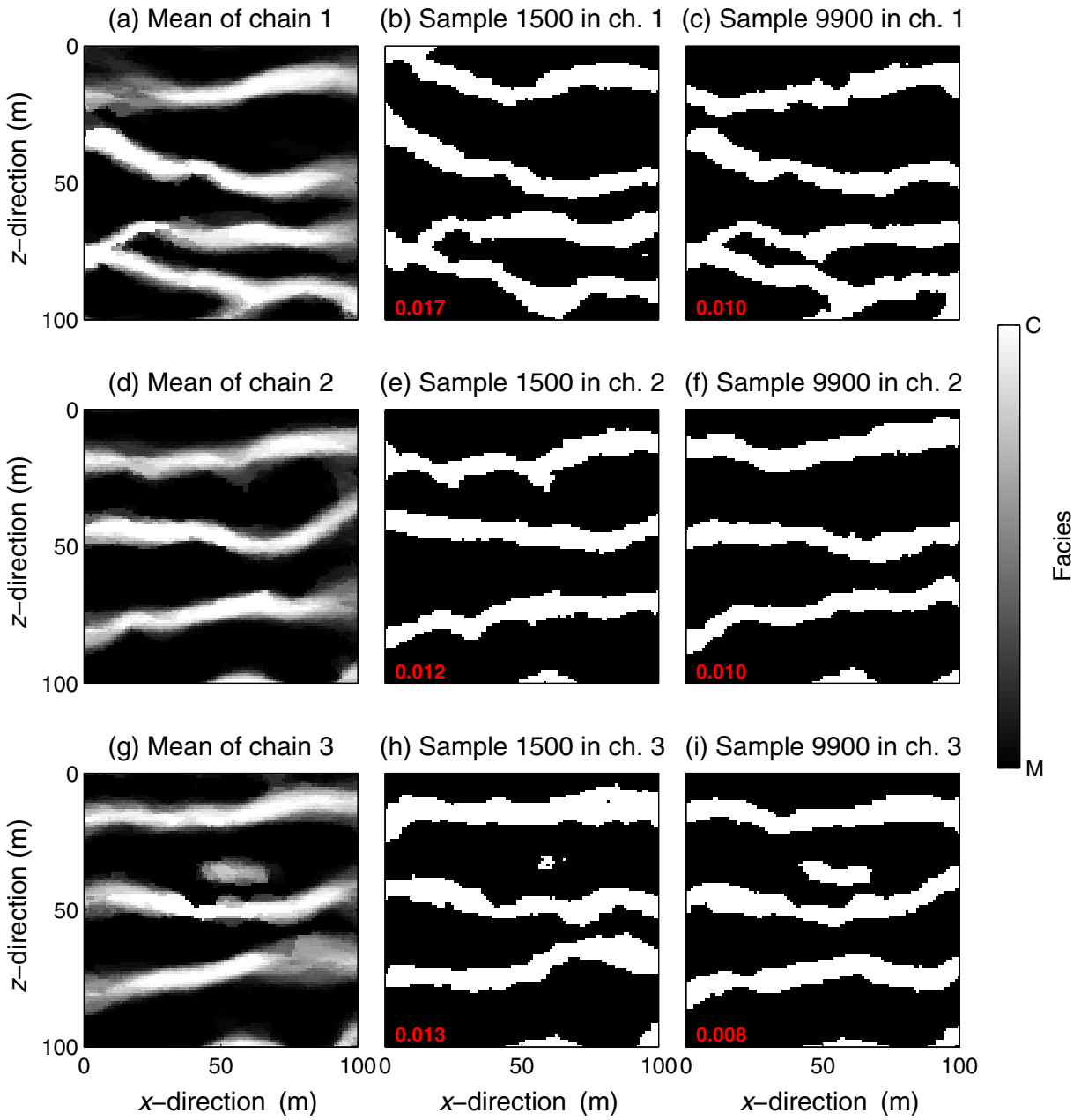

Figure 5: Mean sampled model over MCMC iterations 1500-10,000 (using a thinning factor of 50), and sampled models after 1500 and 9900 MCMC iterations for 3 out of the 16 independent SGR chains and case study 1. The red number in the lower left corner of each plot is the corresponding RMSE (m).

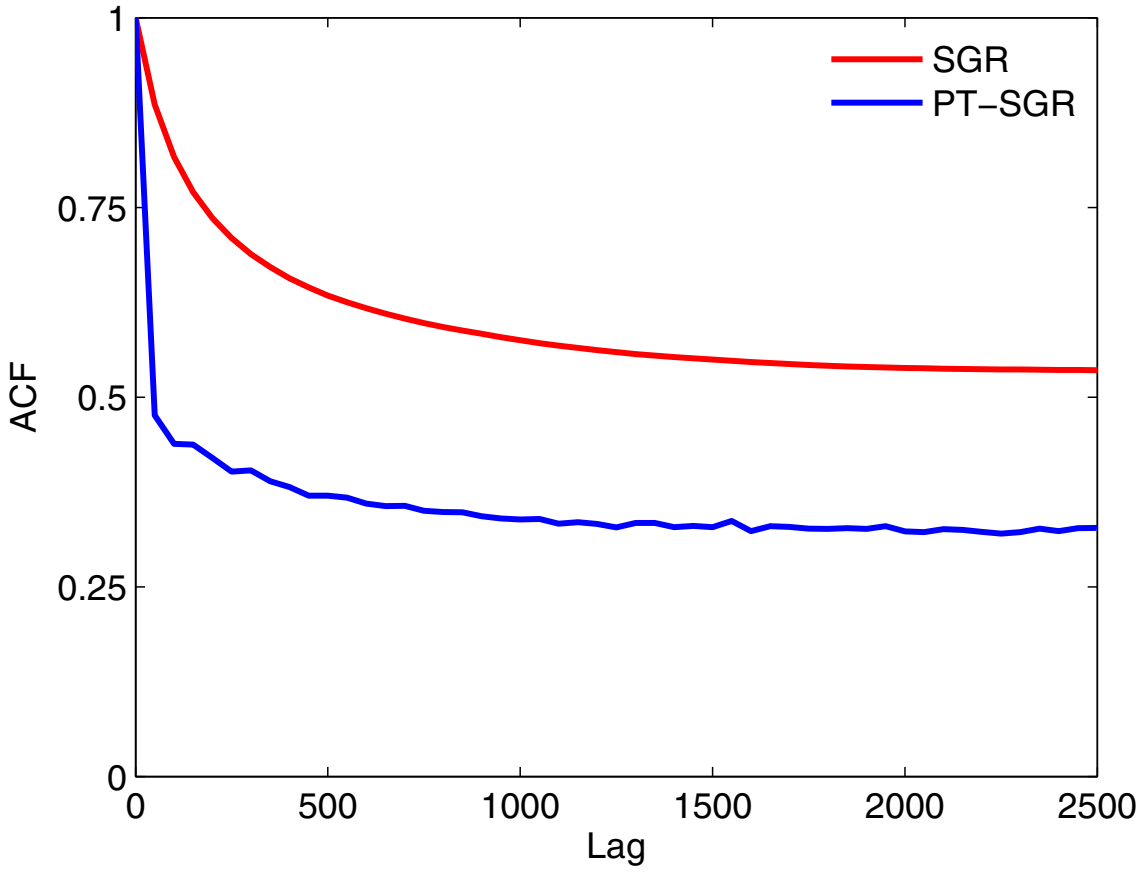

Figure 6: Mean autocorrelation function (ACF) of the 10,000 conductivity grid values derived from PT-SGR (blue line) or SGR (red line) for lags 0-2500 and case study 1. The lag-k autocorrelation is defined as the correlation between draws k lags apart. Listed statistics are computed for the last 8500 iterations of the unit temperature chain of PT-SGR or the 16 independent SGR chains, using a thinning factor of 50 thereby leading to a set of 170 sampled models for each chain. For SGR, the average of the 16 chains is presented.

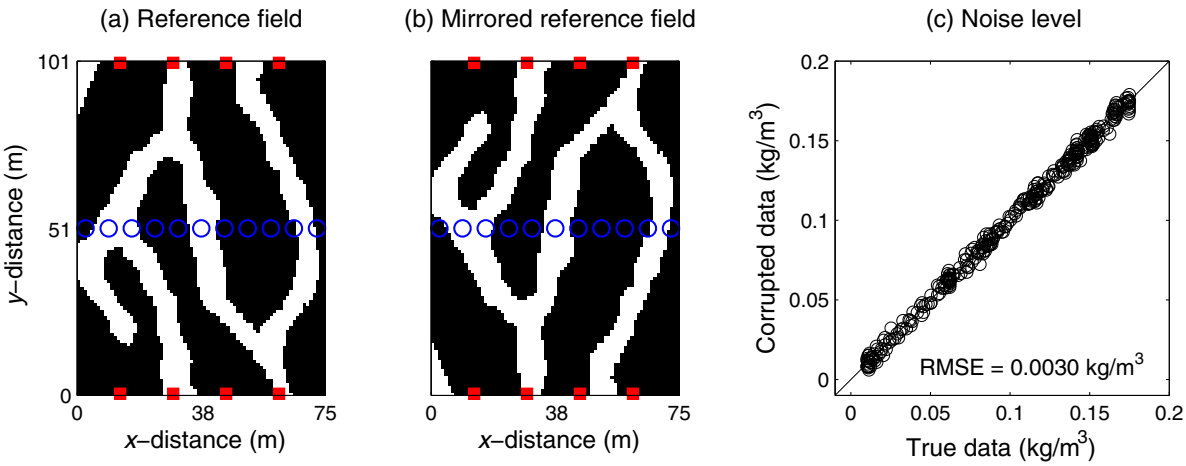

Figure 7: (a) Reference categorical field, (b) associated symmetric (mirrored) field and (c) noise-corrupted measurement data used for our second synthetic case study. In subfigures a and b, the red squares denote the application points of the tracer and the blue circles mark the locations of the pumping wells where concentrations are monitored.

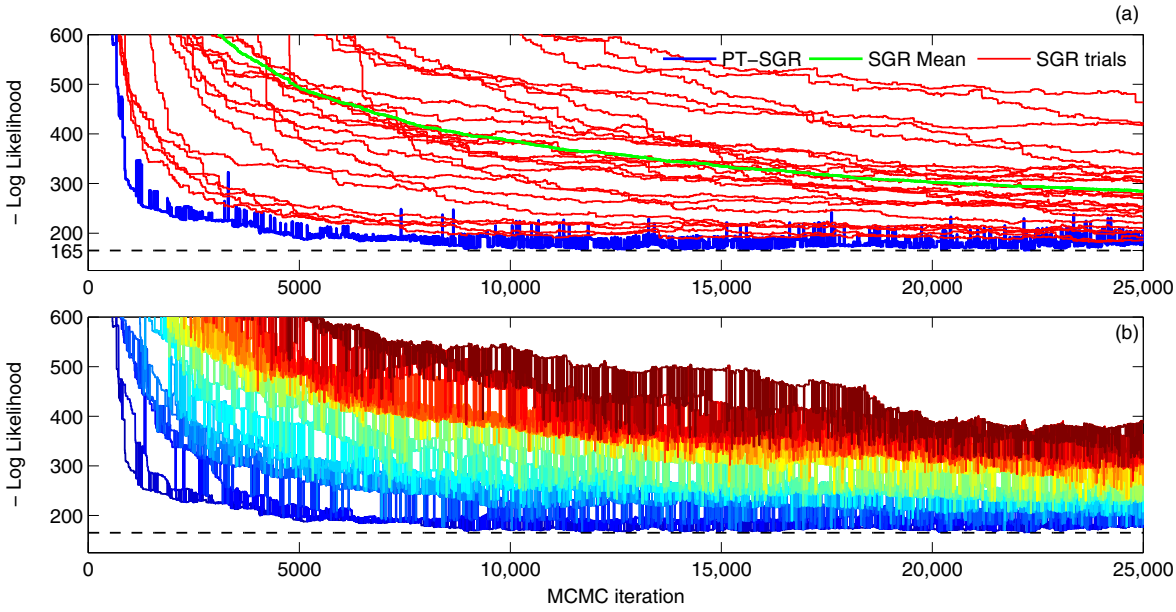

Figure 8: (a) Trace plot of the sampled negative log-likelihood values by the unit temperature chain evolved by PT-SGR (blue line) and the 24 independent SGR trials (red lines) for case study 2. The green line denotes the mean trajectory of the 24 SGR trials. (b) Trace plot of the sampled negative log-likelihood values by the 24 PT-SGR chains with each temperature coded with a different color. The temperature increases as the color varies from dark blue (temperature index of 1) to dark red (temperature index of 2). In both subfigures, the horizontal dashed black line denotes the true negative log-likelihood of 165, corresponding to a RMSE of 0.003 kg/m^{-3}.

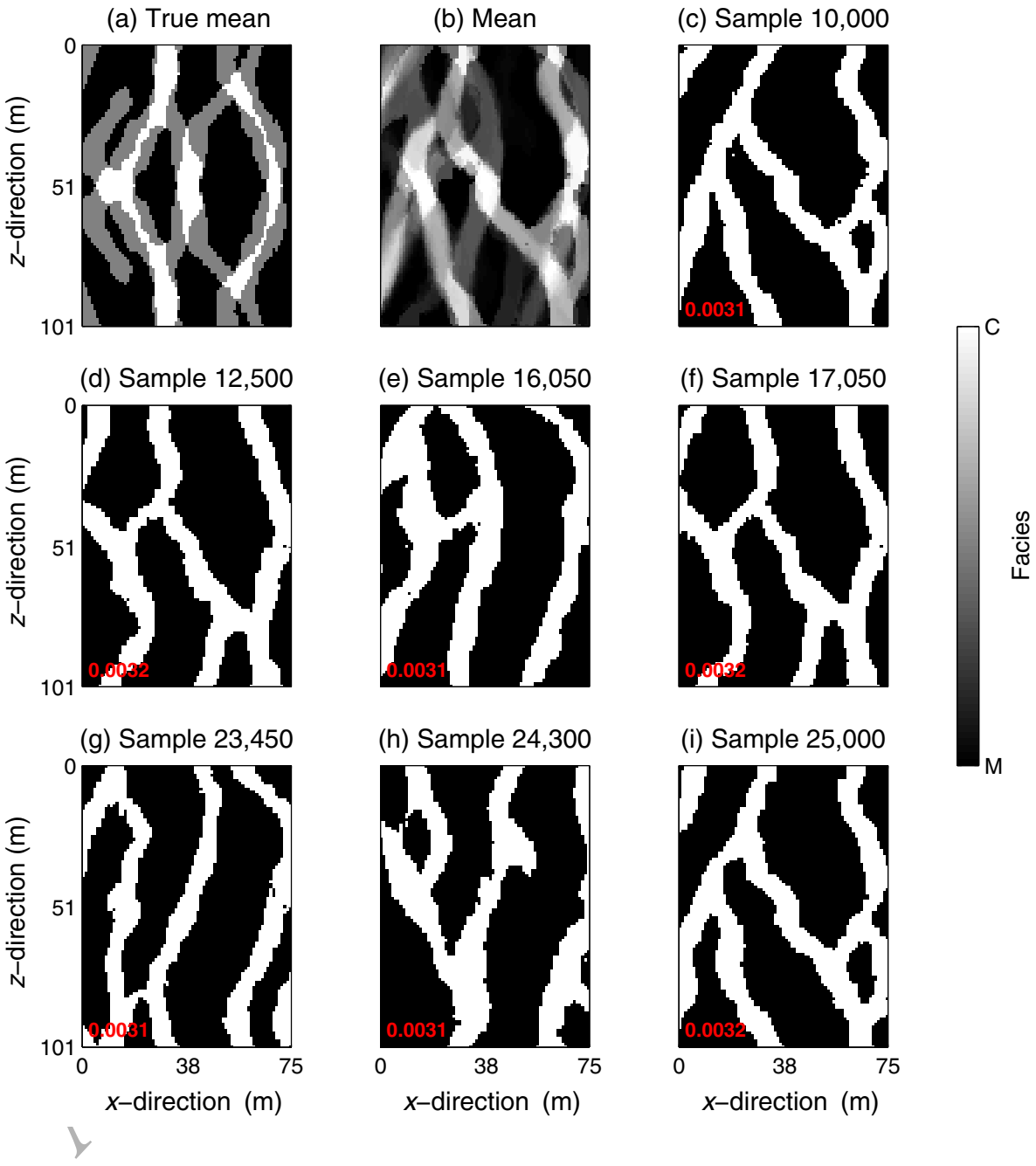

Figure 9: Mean and 8 successive model realizations taken at regular intervals throughout sampling for the PT-SGR trial of case study 2. The sample mean is computed on the basis of the samples produced by the unit temperature chains over iterations 10,000-25,000 and using a thinning factor of 50, thus leading to a total of 300 samples. The red number in the lower left corner of each plot is the corresponding RMSE (m).

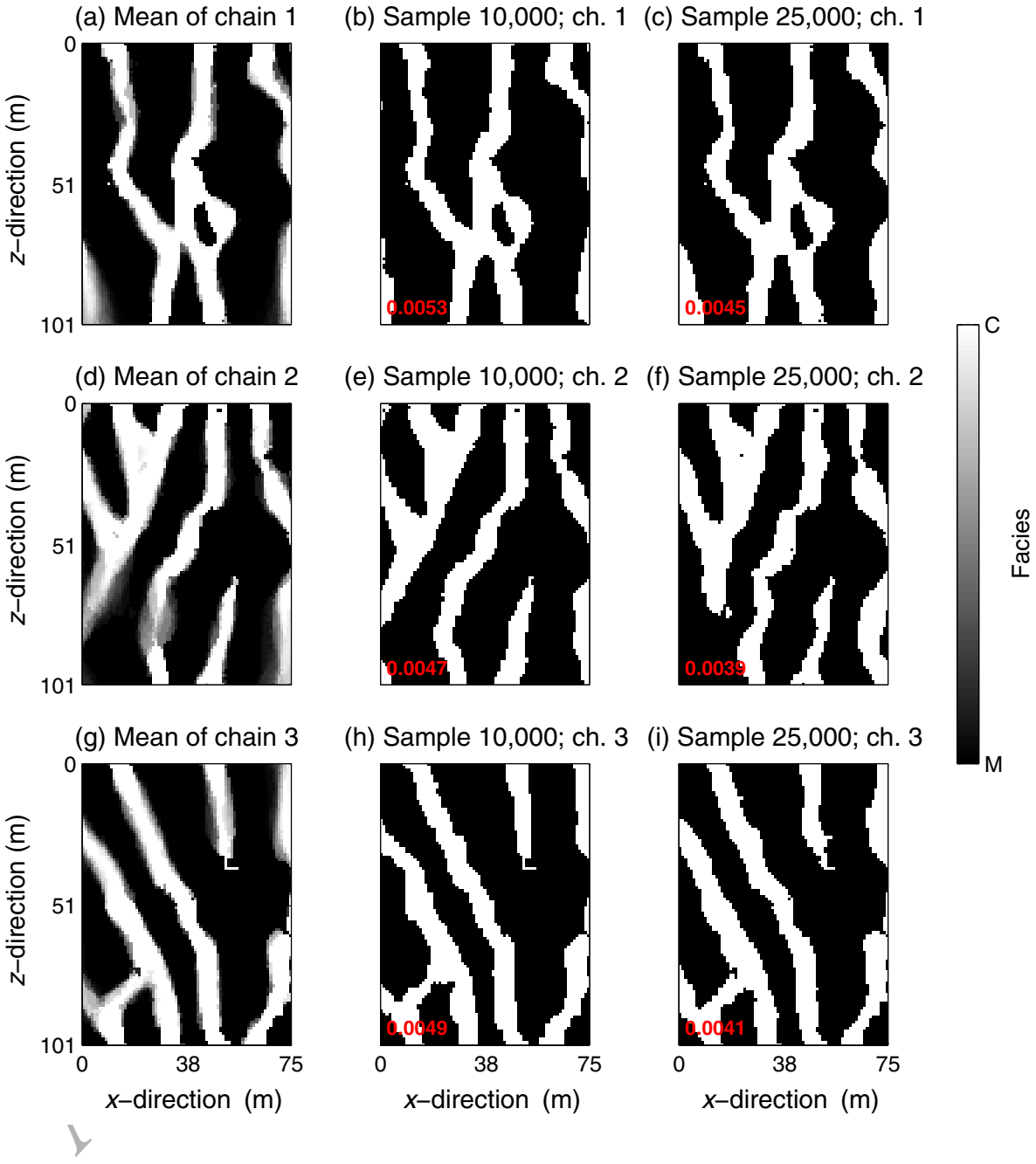

Figure 10: Mean sampled model over MCMC iterations 10,000-25,000 (using a thinning factor of 50), and sampled models after 10,000 and 25,000 MCMC iterations for 3 out of the 24 independent SGR chains and case study 2. The red number in the lower left corner of each plot is the corresponding RMSE (m).

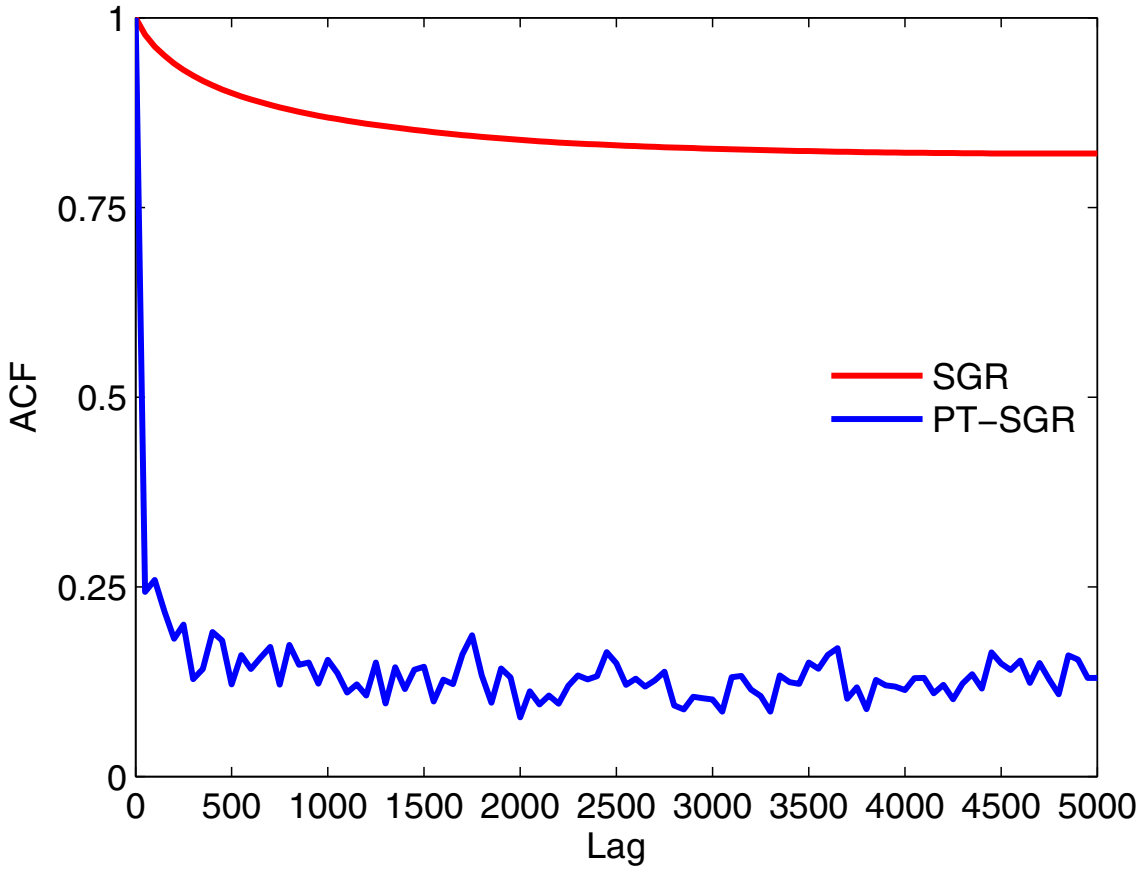

Figure 11: Mean autocorrelation function (ACF) of the 7575 conductivity grid values derived from PT-SGR (blue line) or SGR (red line) for lags 0-5000 and case study 1. The lag-k autocorrelation is defined as the correlation between draws k lags apart. Listed statistics are computed for the last 15,000 iterations of the unit temperature chain of PT-SGR or the 24 independent SGR chains, using a thinning factor of 50 thereby leading to a set of 300 sampled models for each chain. For SGR, the average of the 24 chains is presented.