

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

4 Eric Laloy<sup>a,\*</sup>, Niklas Linde<sup>b</sup>, Diederik Jacques<sup>a</sup>, Grégoire Mariethoz<sup>c</sup>

5 <sup>a</sup>*Institute for Environment, Health and Safety, Belgian Nuclear Research Centre*

6 <sup>b</sup>*Applied and Environmental Geophysics Group, Institute of Earth Sciences, University  
7 of Lausanne*

8 <sup>c</sup>*Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland.*

---

9 **Abstract**

10 The sequential geostatistical resampling (SGR) algorithm is a Markov  
11 chain Monte Carlo (MCMC) scheme for sampling from possibly non-Gaussian,  
12 complex spatially-distributed prior models such as geologic facies or categor-  
13 ical fields. In this work, we highlight the limits of standard SGR for poste-  
14 rior inference of high-dimensional categorical fields with realistically complex  
15 likelihood landscapes and benchmark a parallel tempering implementation  
16 (PT-SGR). Our proposed PT-SGR approach is demonstrated using synthetic  
17 (error corrupted) data from steady-state flow and transport experiments in  
18 categorical 7575- and 10,000-dimensional 2D conductivity fields. In both  
19 case studies, every SGR trial gets trapped in a local optima while PT-SGR  
20 maintains an higher diversity in the sampled model states. The advantage of  
21 PT-SGR is most apparent in a inverse transport problem where the posterior  
22 distribution is made bimodal by construction. PT-SGR then converges to-  
23 wards the appropriate data misfit much faster than SGR and partly recovers  
24 the two modes. In contrast, for the same computational resources SGR does

---

\*Corresponding author

*Preprint submitted to Advances in Water Resources, February 15, 2016*  
Email addresses: [elaloy@sckcen.be](mailto:elaloy@sckcen.be) (Eric Laloy), [niklas.linde@unil.ch](mailto:niklas.linde@unil.ch) (Niklas  
Linde), [djacques@sckcen.be](mailto:djacques@sckcen.be) (Diederik Jacques), [gregoire.mariethoz@unil.ch](mailto:gregoire.mariethoz@unil.ch)  
(Grégoire Mariethoz)

25 not fit the data to the appropriate error level and hardly produces a locally  
26 optimal solution that looks visually similar to one of the two reference modes.  
27 Although PT-SGR clearly surpasses SGR in performance, our results also in-  
28 dicate that using a small number (16-24) of temperatures (and thus parallel  
29 cores) may not permit complete sampling of the posterior distribution by  
30 PT-SGR within a reasonable computational time (less than 1-2 weeks).

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

---

## 33 **1. Introduction**

34 A general Markov chain Monte Carlo (MCMC) simulation strategy based  
35 on sequential geostatistical resampling of spatially-distributed prior mod-  
36 els has recently been introduced in the geosciences to infer the posterior  
37 distribution of subsurface property fields. The approach creates candidate  
38 fields by conditioning a geostatistical field realization from a possibly complex  
39 prior model to a randomly chosen fraction of the current state (and hence  
40 model/field) of the Markov chain. Both parametric (e.g., multi-Gaussian)  
41 and non-parametric prior models can be considered. The multi-Gaussian  
42 prior basically consists of a variogram model that encodes the 2-point statis-  
43 tics to be honored. As of non-Gaussian structures, they can be generated  
44 using a multiple-point statistics (MPS) simulation method. Such algorithms  
45 aim at reproducing not only the 2-point but also higher-order statistics found  
46 in a so-called training image (TI). The TI is a gridded 2D or 3D conceptual  
47 representation of the target spatial field and can be either continuous or cat-  
48 egorical (e.g., geologic facies image). It can either be built from a geologic

49 model or from an observed structure (e.g., outcrop).

50 Various authors have independently introduced the probabilistic sequen-  
51 tial geostatistical resampling (SGR) idea outlined above. The conference  
52 paper by Hansen et al. [13] was probably the first to describe the approach,  
53 considering so-called block updates where a box-shaped randomly located  
54 section of the current model is iteratively resimulated. Almost simultane-  
55 ously, Fu and Gómez-Hernández [6] proposed a variant of the method that  
56 they termed blocking MCMC (BMCMC), which handles multi-Gaussian con-  
57 ditional simulation only. Shortly after, Mariethoz et al. [23] presented a SGR  
58 algorithm that resimulates a randomly chosen set of pixels/voxels rather than  
59 a contiguous block of pixels/voxels. This approach that was named iterative  
60 spatial resampling (ISR) was coupled with the direct sampling (DS) MPS  
61 algorithm of Mariethoz et al. [24] for resampling of both categorical and con-  
62 tinuous priors. Finally, Hansen et al. [14] applied the approach by Hansen et  
63 al. [13] to more case studies and clarified the theoretical background of SGR,  
64 which they referred to as sequential Gibbs sampling (SGS). Even if all SGR  
65 variants presented above fall under the umbrella of Gibbs sampling theory  
66 [10], it is worth noting that the latter also forms a common framework for  
67 unconditional multi-Gaussian simulation [e.g., 7, 21].

68 The SGS [13, 14] and ISR [23] variants differ only in the geometry of  
69 the resimulated grid points, which is a box-shaped area for SGS and a set  
70 of points for ISR. For convenience, from here on we will follow Ruggeri et  
71 al. [36] and use the generic name “SGR” for both SGS and ISR. To sample  
72 from a complex prior, SGR can in principle be implemented with any MPS  
73 algorithm. It is however very important that the considered MPS code can

74 condition on a large fraction of grid data points (i.e., resimulating only a  
75 small fraction of the model). This is currently achieved only by pixel-based  
76 MPS techniques, for example, the DS and SNESIM [39] algorithms. We  
77 use DS in this study as it possesses good conditioning capabilities and is  
78 memory-efficient and relatively fast.

79 Ruggieri et al. [36] performed a systematic evaluation of SGR within a  
80 multi-Gaussian framework. They compared a gradual deformation [16] pro-  
81 posal mechanism with point and block SGR updates for a synthetic linear  
82 geophysical inverse problem using a multi-Gaussian prior and different num-  
83 bers of measurements and noise levels. Results by Ruggieri et al. [36] suggest  
84 that the computational cost of producing one independent realization of the  
85 posterior by SGR is often prohibitively large even for relatively simple in-  
86 verse problems. Ruggieri et al. [36] conclude that this finding warrants further  
87 research into model parameter reduction techniques that reduce parameter  
88 dimensionality and thus complexity of the inverse problem. This is in line  
89 with the work by Laloy et al. [20] who proposed a new reduced multi-Gaussian  
90 model parameterization, that is easily coupled with advanced MCMC sam-  
91 pling techniques [e.g., 19, 40, 42]. For training-image based inference of  
92 non-Gaussian/categorical structures, however, reducing the dimensionality  
93 of the parameter field is arguably more difficult. Even though a few model  
94 reduction methods have recently been proposed [17, 22, 41], the conceptual  
95 simplicity and flexibility of SGR remain attractive. To the best of our knowl-  
96 edge, no critical analysis of SGR performance has been proposed so far for  
97 non-multi-Gaussian cases. Only rather simple problems involving either only  
98 9 data points [23] or unrealistically large measurement errors [14, 15] have

99 been considered. Using a very limited number of measurement data and/or  
100 large measurement errors makes the likelihood function rather flat and the  
101 posterior target is thus easy to sample. It is unclear at this stage whether  
102 training-image based SGR can handle more complicated problems with more  
103 realistically peaked likelihood functions. Furthermore, the posterior distri-  
104 bution might be multi-modal which, as shown herein, is not easily dealt with  
105 by standard SGR.

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

117 Up to now, application of PT to geosciences problems remains limited.  
118 In the area of reservoir simulation, Mohamed et al. [27] applied PT to the  
119 inversion of the Imperial College fault (ICF) model, considering 3 unknown  
120 model parameters and 10 parallel Markov chains. For this application, PT  
121 was shown to explore much more efficiently the posterior parameter space  
122 than two stochastic optimization algorithms which got stuck within local  
123 optima. The study by Carter and White [3] is also focused on posterior

124 exploration of the ICF model, considering from 1 to 13 unknown parame-  
125 ters and using 48 to 64 parallel chains. Carter and White [3] compared a  
126 simple random walk Metropolis (RWM) [26] sampler against the same algo-  
127 rithm equipped with PT for an ICF model with one unknown. This clearly  
128 demonstrated the superiority of PT for sampling the associated multi-modal  
129 posterior parameter distribution. Lastly, Sambridge [37] used as many as 380  
130 parallel chains (and temperatures) to solve a synthetic trans-dimensional (the  
131 number of parameters is unknown) geophysical inverse problem for which the  
132 true model has 13 unknowns. Results by Sambridge [37] show a spectacular  
133 performance improvement by PT in terms of mixing and convergence towards  
134 the target data misfit. The use of PT thus appears to be beneficial not only  
135 for recovering multi-modal posterior distributions, but also for finding the  
136 maximum a posteriori estimate (MAP) of complex unimodal distributions.  
137 To date, PT applications in the geosciences have been concerned with rather  
138 low-dimensional parameter spaces. We hypothesize that PT may be advan-  
139 tageous for posterior inference in high parameter dimensions as well, such as  
140 spatially-distributed subsurface properties. We further suggest that this is  
141 possible even when considering a number of levels in the temperature ladder  
142 that is very small compared to the dimensionality of the parameter space.  
143 For completeness, we note that independently of our work, the idea of cou-  
144 pling PT with SGR also recently appeared as an outlook in the study by  
145 Ruggeri et al. [36].

146 In this paper, we illustrate the limits of the standard SGR method for  
147 posterior sampling of categorical fields, and benchmark a PT implementa-  
148 tion with respect to both data fitting and diversity of the sampled posterior

149 distribution. We refer to the proposed algorithm as PT-SGR for parallel  
150 tempering SGR. In contrast to previous work with PT in the geosciences,  
151 the inverse problems considered herein are quite high-dimensional (7575 to  
152 10,000 sampled parameters). Moreover, only a relatively limited amount of  
153 parallel chains are used: from 16 to 24. This allows for parallel implementa-  
154 tion on workstation computers or small clusters. Besides PT, we also inves-  
155 tigate which settings of the SGR algorithm achieve the best performance for  
156 categorical field inference. Our proposed PT-SGR approach is demonstrated  
157 using synthetic (error corrupted) data from two flow and transport experi-  
158 ments in categorical 10,000- and 7575-dimensional 2D hydraulic conductivity  
159 fields that represent a channelized aquifer. These inverse problems involve  
160 realistically complex likelihood landscapes, one of which is made bimodal by  
161 construction.

162 This paper is organized as follows. Section 2 presents the different ele-  
163 ments of our inversion approach. This is followed in section 3 with numerical  
164 experiments which include a performance analysis of SGR for different algo-  
165 rithmic settings, and a benchmarking of PT-SGR against SGR for the same  
166 multi-core computational resources. Section 4 then provides further discus-  
167 sion of the performance and limitations of our method and discusses possible  
168 future developments. Finally, section 5 concludes this paper with a summary  
169 of the most important findings.

170 **2. Methods**

171 *2.1. Bayesian inference*

172 A common stochastic representation of the forward problem is

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

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

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

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

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

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

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

191 where the  $F_i(\boldsymbol{\theta})$  are the simulated equivalents to the  $i = 1, \dots, N$  measure-  
 192 ment data,  $d_i$ .

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

### 201 *2.2. Sequential geostatistical resampling from a training image*

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

$$204 \quad \alpha(i, j) = 1 \wedge \left( \frac{p(\boldsymbol{\theta}_j) L(\boldsymbol{\theta}_j|\mathbf{d})}{p(\boldsymbol{\theta}_i) L(\boldsymbol{\theta}_i|\mathbf{d})} \right), \quad (4)$$

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

214 reduced to

$$215 \quad \alpha(i, j) = 1 \wedge \left( \frac{L(\boldsymbol{\theta}_j | \mathbf{d})}{L(\boldsymbol{\theta}_i | \mathbf{d})} \right). \quad (5)$$

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

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

### 235 *2.3. Parallel tempering sequential geostatistical resampling*

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

238 posterior density

$$239 \quad p(\boldsymbol{\theta}, T|\mathbf{d}) \propto [p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{d})]^{1/T}, \quad (6)$$

240 or the likelihood function

$$241 \quad p(\boldsymbol{\theta}, T|\mathbf{d}) \propto p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|\mathbf{d})^{1/T}, \quad (7)$$

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

248 Each tempered chain undergoes two possible moves throughout sampling:  
249 within-chain and between-chain proposals. In our proposed PT-SGR imple-  
250 mentation, the within-chain proposal consists of a standard SGR update  
251 where a random fraction of the current state of the chain is resimulated ac-  
252 cording to the TI. The between-chain proposal consists of a swap of model  
253 states at two temperature levels  $i$  and  $j$

$$254 \quad [(\boldsymbol{\theta}_i, T_i), (\boldsymbol{\theta}_j, T_j)] \rightarrow [(\boldsymbol{\theta}_i, T_j), (\boldsymbol{\theta}_j, T_i)], \quad (8)$$

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

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

$$265 \quad \alpha_s(i, j) = 1 \wedge \frac{L(\boldsymbol{\theta}_j|\mathbf{d})^{1/T_i} p(\boldsymbol{\theta}_j) L(\boldsymbol{\theta}_i|\mathbf{d})^{1/T_j} p(\boldsymbol{\theta}_i)}{L(\boldsymbol{\theta}_j|\mathbf{d})^{1/T_j} p(\boldsymbol{\theta}_j) L(\boldsymbol{\theta}_i|\mathbf{d})^{1/T_i} p(\boldsymbol{\theta}_i)}. \quad (9)$$

266 Canceling the equivalent terms in the numerator and denominator and per-  
 267 forming some reorganization leads to

$$268 \quad \alpha_s(i, j) = 1 \wedge \left[ \frac{L(\boldsymbol{\theta}_j|\mathbf{d})}{L(\boldsymbol{\theta}_i|\mathbf{d})} \right]^{(1/T_i - 1/T_j)}, \quad (10)$$

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

271 The temperature swapping (equation (8)) is often restricted to neighbor-  
 272 ing temperatures [4], either by considering all pairs of neighbors at once [e.g.,  
 273 30] or only one pair at the time [e.g., 38]. Other authors [e.g., 37] instead pro-  
 274 posed to randomly swap models independently of their temperature levels.  
 275 All these schemes are valid in the sense that the unit temperature chain(s)  
 276 will ultimately converge to the proper stationary distribution, provided that  
 277 any given temperature is involved in possible exchange with no more than  
 278 one other temperature [30] at each proposal step.

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

285 [29]. The mean acceptance probability,  $\alpha_s$ , of a swap move for a given pair  
286 (equation (10)) is recognized to be another important diagnostic of paral-  
287 lel tempering performance. Obviously, an excessively small  $\alpha_s$  will hamper  
288 exploration by the unit temperature chain. As  $\alpha_s$  increases, however, the  
289 tempered chains will tend to keep exchanging each other's models without  
290 creating new configurations, thereby slowing down posterior sampling [29].  
291 Optimal  $\alpha_s$  values of about 20% [31] and 39% [30] have been proposed under  
292 certain conditions, whereas good sampling performance was found with a  $\alpha_s$   
293 value as low as 8% [29].

294 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.,  $\phi$ , ...).
2:   for  $i = 1, \dots, m$  do ▷ Loop over MCMC iterations
3:     for  $j = 1, \dots, n$  do in parallel ▷ Loop over temperature ladder
4:        $p(\boldsymbol{\theta}, T_j | \mathbf{d})_i \leftarrow \mathbf{SGR}_{\text{MCMC}}(T_j, \mathbf{P})$  ▷ classical MCMC move
       with SGR. The  $j = 1, \dots, n$  updates are done in parallel.
5:     end for
6:     if  $p_{SW} > U(0, 1)$  then ▷ Perform a swap update with probability
        $p_{SW}$ , with  $U(0, 1)$  indicating an uniform random draw between 0 and 1.
7:        $\mathbf{r} = \text{SelectPairs}(n, meth_{SW}, i)$  ▷ Select pairs of temperatures
8:       for  $j = 1, \dots, n/2$  do
9:          $v = \mathbf{r}(2 \times (j - 1) + 1)$   $w = \mathbf{r}(2 \times j)$  ▷ Propose swapping
       of selected pairs of chains, one possible exchange swap per temperature
10:         $\alpha(v, w) \leftarrow 1 \wedge \left[ \frac{L(\boldsymbol{\theta}_w | \mathbf{d})}{L(\boldsymbol{\theta}_v | \mathbf{d})} \right]^{(1/T_v - 1/T_w)}$ 
11:        if  $\alpha < U(0, 1)$  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
```

---

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

---

**Algorithm 2** Selection of temperature indices at swapping time

---

```

1: procedure  $\mathbf{r} = \text{SelectPairs}(n, \text{meth}_{SW}, i)$ 
2:   if  $\text{meth}_{SW} = \text{"random"}$  then           ▷ Select temperatures randomly
3:      $\mathbf{r} = \text{permute}(n)$                    ▷ Create a random permutation of the
      temperature indices
4:   elseif  $\text{meth}_{SW} = \text{"adjacent"}$            ▷ Consider all adjacent pairs
5:     if  $\text{mod}(i) = 1$  then                   ▷ MCMC iteration number is odd
6:        $\mathbf{r} = [1, \dots, n]$ 
7:     else                                   ▷ MCMC iteration number is even
8:        $\mathbf{r} = [2, \dots, n - 1, 1, n]$ 
9:     end if
10:  end if
11: end procedure

```

---

302 **3. Case studies**

303 *3.1. Case study 1: steady-state flow*

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

306  $\times 100$  modeling domain lies in the  $x - y$  plane with a grid cell size of 1  
 307 m. Channels and matrix are assigned hydraulic conductivity values of  $1 \times$   
 308  $10^{-2}$  m/s and  $1 \times 10^{-4}$  m/s, respectively. Steady state groundwater flow is  
 309 simulated using MaFloT [18] which is a finite-volume algorithm for 2D flow  
 310 and transport in porous media. We assume no flow boundaries at the upper  
 311 and lower sides and fixed head boundaries on the left and right sides of the  
 312 domain so that a lateral head gradient of 0.025 (-) is imposed, with water  
 313 flow in the  $x$ -direction. A pumping well extracting  $0.003 \text{ m}^2/\text{s}$  is located at  
 314 the center of the domain. Simulated heads are collected at 49 locations that  
 315 are regularly spread over the domain (Figure 1a-b). These data were then  
 316 corrupted with a Gaussian white noise using a standard deviation of 0.01 m,  
 317 leading to a root-mean-square-error (RMSE) of 0.01 m for the measurement  
 318 data (Figure 1c). This translates into a reference log-likelihood (component  
 319 that depends on  $\boldsymbol{\theta}$ ),  $\ell(\boldsymbol{\theta}|\mathbf{d})$ , of -24.5.

### 320 *3.2. DS settings for case study 1*

321 The parameters of the DS simulation used for case study 1 are a neigh-  
 322 borhood made of 50 nodes and a distance threshold set to 0.05. This means  
 323 that for any simulated node, the data event (pattern) made of the 50 closest  
 324 neighbors is considered, and up to only 2 mismatching nodes are allowed [see  
 325 24, for details]. The maximum scanned fraction of the TI is set to 0.9.

### 326 *3.3. SGR settings*

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

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

343 Figure 2 displays the resulting sampled root-mean-square error (RMSE)  
 344 trajectories and mean acceptance rate (AR) of the MCMC. Averages of the  
 345 4 trials are presented. Clearly, resimulating a box-shaped area (S2) shows  
 346 a superior performance with respect to data fitting. As expected, the AR  
 347 decreases with  $\phi$  for both S1 and S2. Large resimulated fractions induce  
 348 low AR values, below 1 or 2% ( $\phi_{S1} = 0.995$  to  $\phi_{S1} = 0.5$  and  $\phi_{S2} = 0.5$  and  
 349  $\phi_{S2} = 0.25$ ). Such small AR values characterize a prohibitively slow evolution  
 350 of the MCMC chain.

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

355 during the first 10% of the MCMC iterations. The motivation for this target  
356 value is based on the fact that an AR of about 23% is considered optimal for  
357 Gaussian proposal and target distributions whereas an AR in the range 10%  
358 - 50% is generally recommended [33, 34].

### 359 *3.4. Parallel tempering settings*

360 For the case studies considered in this paper, limited testing with the  
361 different swapping strategies described in section 2.3 showed no overwhelming  
362 advantage of any specific strategy. Nevertheless, randomly proposing to swap  
363 model states after every regular within-chain MCMC update appeared to  
364 be the most robust and efficient approach. We therefore do so for all of  
365 our numerical experiments. With respect to the  $\alpha_s$  values of the individual  
366 tempered chains, we use a common loglinear temperature ladder [3, 35] with  
367 maximum level such that  $\alpha_s$  is (almost) always comprised between 5% and  
368 30%. The rationale behind a loglinear scale is that a pair of neighboring cool  
369 chains ( $T = 1$  or slightly higher) likely needs a smaller temperature difference  
370 for an exchange swap to be accepted compared to a pair of hotter chains.  
371 Other settings that perform better may very well exist, the quest for which  
372 is beyond the scope of this study.

### 373 *3.5. Convergence of the Markov chain Monte Carlo simulation*

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

379 unity the value of  $\hat{R}$ . Values of  $\hat{R}$  smaller than 1.2 are commonly deemed  
380 to indicate convergence to a limiting distribution. In principle,  $\hat{R}$  offers a  
381 stronger convergence assessment than merely considering the moment when  
382 the sampled (log-)likelihood (and thus RMSE) values reach an equilibrium.  
383 The latter indeed signifies only that the posterior distribution has been lo-  
384 cated, whereas the former aims at evaluating whether it has been adequately  
385 explored. For example, in the study by Laloy et al. [see 20] it was found  
386 that 25 times more MCMC steps were needed to appropriately explore a  
387 1000-dimensional posterior than to start sampling it. For the considered case  
388 studies and computational budgets, our simulation results indicate that  $\hat{R}$ -  
389 convergence is never achieved by SGR, no matter whether computed on the  
390 basis of 3 (randomly chosen) chains or all of the 16 (case study 1) or 24 (case  
391 study 2) independent trials. Unfortunately, the  $\hat{R}$  statistic is not well suited  
392 for monitoring convergence of several unit temperature chains within a given  
393 tempered ensemble. Indeed, running PT-SGR for case study 1 with 3 out of  
394 the 16 temperatures set to 1 results in an exaggeratedly fast  $\hat{R}$ -convergence  
395 (not shown). The 10,000 individual  $\hat{R}$  values may even jointly fall below 1.2  
396 after less iterations than required for the unit chains to sample the appro-  
397 priate likelihood values. The reason for this is that the swapping dynamics  
398 causes large correlations between states/models of neighboring cool chains.  
399 The within-chain variances thus become similar enough for  $\hat{R}$  to be satis-  
400 fied prematurely. We therefore refrain from using the  $\hat{R}$  diagnostic. Instead  
401 we simply resort to the point in time when (log-)likelihood values start to  
402 fluctuate around a constant level to define burn-in. From this moment on  
403 our algorithm starts drawing posterior samples. One must bear in mind,

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

### 409 *3.6. Inversion results for case study 1*

410 For this case study, a total of 10,000 MCMC iterations is allowed for  
411 both SGR and PT-SGR. For a classical single-chain SGR trial, this trans-  
412 lates into 10,000 forward model evaluations. The PT-SGR algorithm with  
413  $n$  temperature levels is best run with parallel calculation of both the  $n$  DS  
414 simulations and  $n$  forward model evaluations performed per MCMC itera-  
415 tion. This roughly leads to a similar CPU-time per MCMC iteration (from  
416 5 to 10 s herein) between SGR and PT-SGR. One must note, however, that  
417 some minor additional computational time is needed for PT-SGR due to  
418 communication overhead, the extent of which depends on hardware- and  
419 software-specific details. Here PT-SGR is ran on a multi-core platform, with  
420  $n = 16$ . A loglinear temperature ladder was selected between unity and a  
421 maximum temperature of 6, together with a single unit temperature chain.

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

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

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

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

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

### 470 *3.7. Case study 2: tracer experiment*

471 Our second case study uses simulated tracer breakthrough curves at dif-  
472 ferent wells as measurement data. The modeling domain is  $75 \times 101$  and  
473 is located in the  $x - y$  plane with a grid cell size of 1 m. Channels and  
474 matrix are again assigned hydraulic conductivity values of  $1 \times 10^{-2}$  m/s and  
475  $1 \times 10^{-4}$  m/s, respectively. Steady state groundwater flow and conservative  
476 transport are both simulated using MaFloT. No-flow boundaries are assumed  
477 at the left and right sides, and fixed head boundaries on the upper and lower  
478 sides of the domain. These fixed heads are set to 0 m at both sides, and 11

479 pumping wells individually extracting  $0.0005 \text{ m}^2/\text{s}$  of water are spaced 7 m  
 480 apart along the horizontal line located at equal distance from the top and  
 481 bottom sides (Figure 7a). The facies values at the 11 wells are assumed to be  
 482 known exactly and serve for direct conditioning. A conservative tracer with  
 483 concentration of  $1 \text{ kg m}^{-3}$  is applied within 8 model cells of the top and bot-  
 484 tom boundaries using a step function. The  $x$ - $y$  coordinates of these cells are  
 485  $(14,1)$ ,  $(30,1)$ ,  $(46,1)$ ,  $(62,1)$ ,  $(14,101)$ ,  $(30,101)$ ,  $(46,101)$ , and  $(62,101)$  (Fig-  
 486 ure 7a). The background solute concentration is assumed to be  $0.01 \text{ kg m}^{-3}$ .  
 487 Ignoring density effects, conservative transport of the tracer through the sub-  
 488 surface is simulated using open boundaries on all sides, and longitudinal and  
 489 transverse dispersivities both set to 0.1 m. Solute transport was monitored  
 490 during a period of 10 days with concentration measurements made every 8  
 491 hours in the 11 extraction wells, resulting into a total of 330 observations.  
 492 These simulated data were then corrupted with a Gaussian white noise using  
 493 a standard deviation equivalent to 3% of the mean observed concentration.  
 494 This led to root-mean-square-error (RMSE) and log-likelihood ( $\ell(\boldsymbol{\theta}|\mathbf{d})$ ) of  
 495  $0.0030 \text{ kg m}^{-3}$  and -165, respectively, for the measurement data (Figure 7c).

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

503 *3.8. DS settings for case study 2*

504 The parameters of the DS simulation used for case study 2 are a neigh-  
505 borhood made of 75 nodes, a distance threshold of 0.01 and a maximum  
506 scanned fraction of the TI of 0.9.

507 *3.9. Inversion results for case study 2*

508 A total of 25,000 MCMC iterations is used for this case study. The  
509 PT-SGR algorithm is again run on a multi-core machine with  $n = 24$  and  
510 the computational cost incurred by 1 MCMC iteration is in the range of  
511 20-30 s for this workstation. A loglinear temperature ladder between unity  
512 and a maximum temperature of 2 is selected together with a single unit  
513 temperature chain. Using such a small maximum temperature was needed  
514 given the trade-off between the number of available parallel cores and the  
515 complexity of the (log-)likelihood landscape. For instance, the peakier the  
516 likelihood function, the smaller the temperature intervals need to be for  $\alpha_s$  to  
517 be significantly larger than zero. With 24 temperature levels, using a larger  
518 maximum temperature than 2 essentially results in a frequency of accepted  
519 swaps that is impractically low. In addition, the update mechanism of  $\phi$  that  
520 is described in section 3.3 was slightly modified in an attempt to generate  
521 more diverse proposals while keeping the acceptance rate of regular MCMC  
522 moves reasonably high. Rather than a tuned constant resimulation block size,  
523  $\phi$  was taken as the (adapted) mean of a triangular pdf bounded between 0.01  
524 and 0.25 from which the actual size of the block to be resimulated was drawn.  
525 This was similarly done for both SGR and PT-SGR.

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

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

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

551 The AR associated with the PT-SGR chains are lower than for case study  
 552 1 but still acceptable: across the whole temperature ladder,  $\alpha$  and  $\alpha_s$ , are

553 8% (with range of 7% - 9%) and 14% (with range of 6% - 20%), respectively.  
554 The corresponding swap exchange dynamics looks visually good from itera-  
555 tion 10,000 onwards (Figure 8b). The PT-SGR unit temperature chain does  
556 however not mix well. Indeed, the chain basically cycles over the (nearly)  
557 same 6-7 models during the last 15,000 MCMC iterations (Figure 9). The  
558 reason for this is likely twofold. First, the maximum temperature of 2 does  
559 not flatten the likelihood enough for sufficient exploration by the hot chains.  
560 Second and most important, for this rather complex likelihood landscape  
561 the exchange swaps appear to be mostly performed in cycles between a few  
562 neighboring temperature levels with colder (hotter) samples almost never  
563 traveling to the highest (lowest) levels (not shown). The only way to solve  
564 this problem would therefore be to use a ladder with a much larger number  
565 of levels and a wider temperature range.

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

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

#### 582 **4. Discussion**

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

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

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

616 On a more practical level, the rationale for our DS settings deserves spe-  
617 cial attention. A neighborhood of 50 (case study 1) or 75 (case study 2) nodes  
618 may seem large [25] as the DS simulation time increases with the number  
619 of neighboring nodes. Nevertheless, using such large values was necessary to  
620 minimize the occurrence of artifacts in the generated models, which is caused  
621 by repeatedly conditioning on a large amount of grid points throughout the  
622 MCMC sampling. Combined with a large fraction of conditioning data (say  
623  $> 50\%$  of the image), a small neighborhood can sometimes result in model  
624 proposals that are somewhat degraded compared to the TI. It is indeed the  
625 restricted neighborhood size that gives freedom to the DS to produce struc-  
626 tures that are different than those found in the TI. Also, all it takes for a

627 slightly degraded model to appear in the Markov chain is for the Metropolis  
628 rule (equation 5) to accept it. In other words, even if a model proposal with  
629 some artifacts is only rarely proposed, this model can persist in the Markov  
630 chain if the associated simulated data fit the observations sufficiently well.  
631 Even with the employed neighborhood of 50 nodes for case study 1 (see sec-  
632 tion 3.2), trials  $\phi_{S1} = 0.5$  and  $\phi_{S1} = 0.75$  of section 3.3 nevertheless showed  
633 artifacts in the proposed models, typically manifested by overly broad chan-  
634 nels (not shown). To a lesser extent, other artifacts such as isolated patches  
635 and broken channels also occurred for trials  $\phi_{S2} = 0.9$  and  $\phi_{S2} = 0.95$  (not  
636 shown). The used distance threshold of 0.01 for case study 2 (see section  
637 3.8) also incurs a larger computational cost than that of using the more com-  
638 mon value of 0.05 [e.g., 23]. The value of 0.01 turned out to be required to  
639 (almost) systematically honor the point measurement data (see section 3.7).

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

## 646 5. Conclusion

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

651 plementation that, for a similar multi-core computing budget, provides much  
652 improved results with respect to both convergence towards the appropriate  
653 data misfit and sampling diversity. Two synthetic case studies are consid-  
654 ered: a steady-state flow and a transport inverse problem, involving from  
655 7501 to 10,000 unknowns. For the transport problem, the corresponding like-  
656 lihood function is made bimodal with two well separated modes. In both case  
657 studies, every SGR MCMC chain gets trapped in a local optima while par-  
658 allel tempering within sequential geostatistical resampling (PT-SGR) does  
659 not. The advantage of PT-SGR becomes more apparent for the bimodal  
660 inverse transport problem, for which PT-SGR is found to converge towards  
661 the reference data misfit much faster than SGR and to indicate the existence  
662 of two posterior modes. In contrast, for the same computational resources  
663 SGR appears to be barely able to appropriately fit the data and does al-  
664 most not produce any single solution that looks visually similar to one of the  
665 two reference modes. Although PT-SGR outperforms SGR, our results also  
666 demonstrate that using a reasonably small number of temperatures (and thus  
667 parallel cores) in the range 16-24 may not allow sampling of the posterior  
668 distribution by PT-SGR within an affordable computational time. As an  
669 alternative to significantly increasing the number of temperatures and thus  
670 computational needs, coupling PT-SGR with Wang-Landau sampling and  
671 (2) reframing SGR within an ensemble-based multiscale optimization frame-  
672 work are two potentially useful approaches that will be investigated in future  
673 work. More generally, PT could also prove useful when used in conjunction  
674 with dimensionality reduction approaches.

## 675 **6. Acknowledgments**

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

## 684 **References**

- 685 [1] Alcolea A, Renard P. Blocking Moving Window algorithm: Conditioning  
686 multiple-point simulations to hydrogeological data. *Water Resour. Res.*  
687 2010;46:W08511. <http://dx.doi.org/10.1029/2009WR007943>.
- 688 [2] Bornn L, Jacob PE, Del Moral P, Doucet A. An adap-  
689 tive interacting WangLandau algorithm for automatic den-  
690 sity exploration. *J. Comput. Graph. Stat.* 2013; 22(3):749-773.  
691 <http://dx.doi.org/10.1080/10618600.2012.723569>.
- 692 [3] Carter JN, White DA. History matching on the Imperial College  
693 fault model using parallel tempering. *Comput. Geosci.* 2013;17:43-65.  
694 <http://dx.doi.org/10.1007/s10596-012-9313-3>.
- 695 [4] Earl DJ, Deem, MW. Parallel tempering: theory, applications, and new  
696 perspectives. *Phys. Chem. Chem. Phys.* 2005;7: 3910-3916.

- 697 [5] Falcioni M, Deem MW. A biased Monte Carlo scheme for zeolite structure  
698 solution. *J. Chem. Phys.* 1999;110: 1754–1766.
- 699 [6] Fu J, Gómez-Hernández JJ. A blocking Markov chain Monte Carlo  
700 method for inverse stochastic hydrogeological modeling. *Math. Geosci.*  
701 2009;41:105–128. <http://dx.doi.org/10.1007/s11004-008-9206-0>.
- 702 [7] Galli A, Gao H. Rate of convergence of the Gibbs sampler in the Gaussian  
703 case. *Math. Geol.* 2001;33(6):653–677.
- 704 [8] Gardet C, Le Ravalec M, Gloaguen E. Multiscale parameterization  
705 of petrophysical properties for efficient history-matching. *Math. Geosci.*  
706 2014;46(3):315–336. <http://dx.doi.org/10.1007/s11004-013-9480-3>.
- 707 [9] Gelman AG, Rubin DN. Inference from iterative simulation using multiple  
708 sequences. *Stat. Sci.* 1992;7:457–472.
- 709 [10] Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the  
710 Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.*  
711 1984;6(6):721–741.
- 712 [11] Geyer, CJ. Markov Chain Monte Carlo maximum likelihood. 1991. In  
713 *Computing Science and Statistics: Proceedings of the 23rd Symposium on*  
714 *the Interface*, American Statistical Association, New York, pp. 156–163.
- 715 [12] Geyer, CJ. Chapter 11: Importance sampling, simulated tempering and  
716 umbrella sampling. 2011. In Brooks S, Gelman A, Jones GL, Meng XL  
717 (eds.) *Handbook of Markov Chain Monte Carlo*, Chapman & Hall/CRC,  
718 Boca Raton, pp. 295–306.

- 719 [13] Hansen TM, Mosegaard K, Cordua KC. Using geostatistics to describe  
720 complex a priori information for inverse problems. In: Ortiz, JM, Emery,  
721 X (eds.) VIII International Geostatistics Congress. 2008;vol. 1:329–338.  
722 Mining Engineering Department, University of Chile, Santiago.
- 723 [14] Hansen TM, Cordua KC, Mosegaard K. Inverse problems with non-  
724 trivial priors: efficient solution through sequential Gibbs sampling. *Com-  
725 put. Geosci.* 2012;16:593–611. [http://dx.doi.org/10.1007/s10596-011-9271-](http://dx.doi.org/10.1007/s10596-011-9271-1)  
726 1.
- 727 [15] Hansen TM, Cordua KC, Looms MC, Mosegaard K. SIPPI: a Matlab  
728 toolbox for sampling the solution to inverse problems with complex prior  
729 information: Part 2 - Application to crosshole GPR tomography. *Comput.  
730 Geosci.* 2013;52: 481–492. <http://dx.doi.org/10.1016/j.cageo.2012.10.001>.
- 731 [16] Hu, LY. Gradual deformation and iterative calibration of Gaussian re-  
732 lated stochastic models. *Math. Geol.* 2000;32(1):87–108.
- 733 [17] Khaninezhad MM, Jafarpour B, Li L. Sparse geologic dic-  
734 tionaries for subsurface flow model calibration: Part I. In-  
735 version formulation. *Adv. Water Resour.* 2012;39(0):106121.  
736 <http://dx.doi.org/10.1016/j.advwatres.2011.09.002>.
- 737 [18] Künze R, Lunati I. An adaptive multiscale method for density-driven  
738 instabilities. *J. Comput. Phys.* 2012;231:5557–5570.
- 739 [19] Laloy E, Vrugt JA. High-dimensional posterior exploration  
740 of hydrologic models using multiple-try DREAM<sub>(ZS)</sub> and

- 741 high-performance computing. *Water Resour. Res.* 2012;48(1).  
742 <http://dx.doi.org/10.1029/2011WR010608>.
- 743 [20] Laloy E, Linde N, Jacques D, Vrugt JA. Probabilistic inference of  
744 multi-Gaussian fields from indirect hydrological data using circulant  
745 embedding and dimensionality reduction. *Water Resour. Res.* 2015;51.  
746 <http://dx.doi.org/10.1002/2014WR016395>.
- 747 [21] Lantuéjoul C. *Geostatistical simulation: models and algorithms*,  
748 Springer; 2002.
- 749 [22] Lochbühler T, Vrugt JA, Sadegh M, Linde N. Summary statistics from  
750 training images as prior information in probabilistic inversion. *Geophys.*  
751 *J. Int.* 2015;201:157–171. <http://dx.doi:10.1093/gji/ggv008>.
- 752 [23] Mariethoz G, Renard P, Caers J. Bayesian inverse problem and op-  
753 timization with iterative spatial resampling. *Water Resour. Res.* 2010a;  
754 46:W11530. <http://dx.doi.org/10.1029/2010WR009274>.
- 755 [24] Mariethoz G, Renard P, Straubhaar J. The Direct Sampling method  
756 to perform multiple-point geostatistical simulations. *Water Resour. Res.*  
757 2010b;46:W11536. <http://dx.doi.org/10.1029/2008WR007621>.
- 758 [25] Meerschman E, Pirot G, Mariethoz G, Straubhaar J, Van Meirvenne M.  
759 A practical guide to performing multiple-point statistical simulations with  
760 the direct sampling algorithm. *Comput. Geosci.* 2013;52:307–324.
- 761 [26] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E.  
762 Equation of state calculations by fast computing machines. *J. Chem. Phys.*  
763 1953;21:1087–1092.

- 764 [27] Mohamed L, Calderhead B, Filippone M, Christie M, Girolami M. Pop-  
765 ulation MCMC methods for history matching and uncertainty quantifica-  
766 tion. *Comput. Geosci.* 2012;16(2):423–436.
- 767 [28] Mosegaard K, Tarantola, A. Monte Carlo sampling of solutions to inverse  
768 problems. *J. Geophys. Res.* 1995;100(B7):12431–12447.
- 769 [29] Opps SB, Schoffield, J. Extended state-space  
770 Monte Carlo methods. *Phys. Rev. E.* 2001;60:056701.  
771 <http://dx.doi.org/10.1103/PhysRevE.63.056701>.
- 772 [30] Predescu C, Predescu M, Ciobanu CV. On the efficiency of ex-  
773 change in parallel tempering Monte Carlo simulations. *J. Phys. Chem.*  
774 *B.* 2005;109:4189–4196.
- 775 [31] Rathore N, Chopra M, de Pabloa JJ. Optimal allocation of repli-  
776 cas in parallel tempering simulations. *J. Chem. Phys.* 2005;122:024111.  
777 <http://dx.doi.org/0.1063/1.1831273>.
- 778 [32] Robert, CP, Casella G. Monte Carlo statistical methods, second edition.  
779 Springer; 2004.
- 780 [33] Roberts GO, Gelman A, Gilks WR. Weak convergence and optimal scal-  
781 ing of random walk Metropolis algorithms. *Ann. Appl. Prob.* 1997;7:110–  
782 120.
- 783 [34] Roberts GO, Rosenthal JS. Optimal scaling for various Metropolis-  
784 Hastings algorithms. *Stat. Sci.* 2001;16:351–367.

- 785 [35] Romary T. Integrating production data under uncertainty by parallel in-  
786 teracting Markov chains on a reduced dimensional space. *Comput. Geosci.*  
787 2009;13:103–122. <http://dx.doi.org/10.1007/s10596-008-9108-8>.
- 788 [36] Ruggeri P, Irving J, Holliger K. Systematic evaluation of sequential  
789 geostatistical resampling within MCMC for posterior sampling of near-  
790 surface geophysical inverse problems. *Geophys. J. Int.* 2015;202:961–975.  
791 <http://dx.doi.org/10.1093/gji/ggv196>.
- 792 [37] Sambridge M. A Parallel tempering algorithm for probabilistic sam-  
793 pling and multimodal optimization. *Geophys. J. Int.* 2014;196:357–374.  
794 <http://dx.doi.org/10.1093/gji/ggt342>.
- 795 [38] Schrek A, Fort G, Moulines E. Adaptive Equi-Energy Sampler: Conver-  
796 gence and Illustration. *ACM Trans. Model. Comput. Simul.* 2013;23(1):5:1–  
797 5:27. <http://doi.acm.org/10.1145/2414416.2414421>.
- 798 [39] Strebelle S. Conditional simulation of complex geological structures us-  
799 ing multiple point statistics. *Math. Geol.* 2002;34(1):1–22.
- 800 [40] ter Braak, CJ, Vrugt JA. Differential evolution Markov chain with  
801 snooker updater and fewer chains. *Stat. Comput.* 2008;18(4):435–446.
- 802 [41] Vo HX, Durlafsky LJ. A new differentiable parameterization based on  
803 Principal Component Analysis for the low-dimensional representation of  
804 complex geological models, *Math. Geosci.* 2014;46:775–813.
- 805 [42] Vrugt JA, Ter Braak, C, Diks C, Robinson BA, Hyman, JM, Higdon, D.  
806 Accelerating Markov chain Monte Carlo simulation by differential evolu-

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

809 [43] Zahner T, Lochbühler T, Mariethoz G, Linde N. Image syn-  
810 thesis with graph cuts: a fast model proposal mechanism in  
811 probabilistic inversion. *Geophys. J. Int.* 2016;204(2):1179–1190.  
812 <http://dx.doi.org/10.1093/gji/ggv517>.

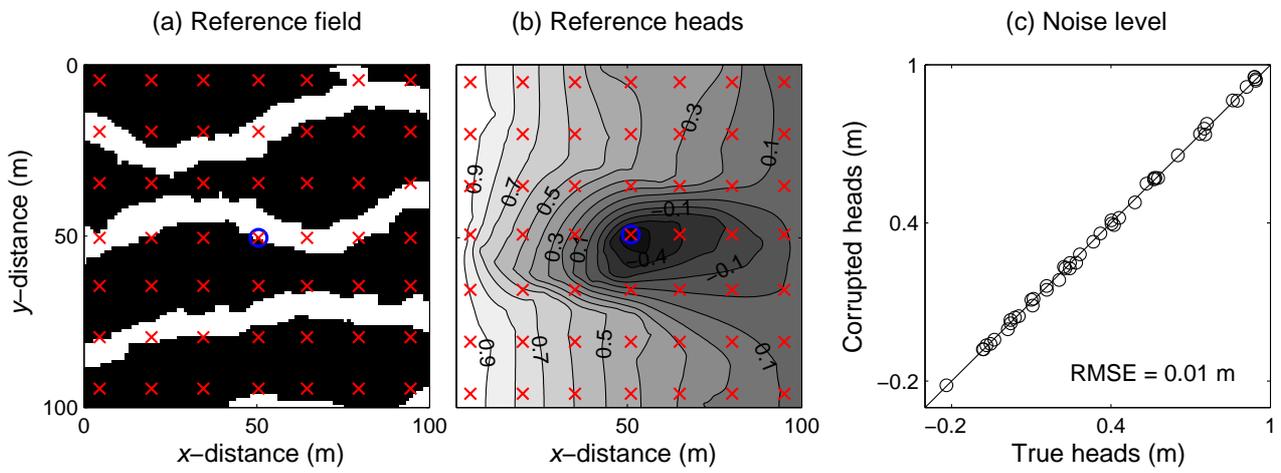


Figure 1: (a) Reference categorical field, (b) associated heads and (c) noise-corrupted measurement data used for case study 1. In subfigures a and b, the blue circle marks the location of the pumping well and the red crosses indicate piezometers.

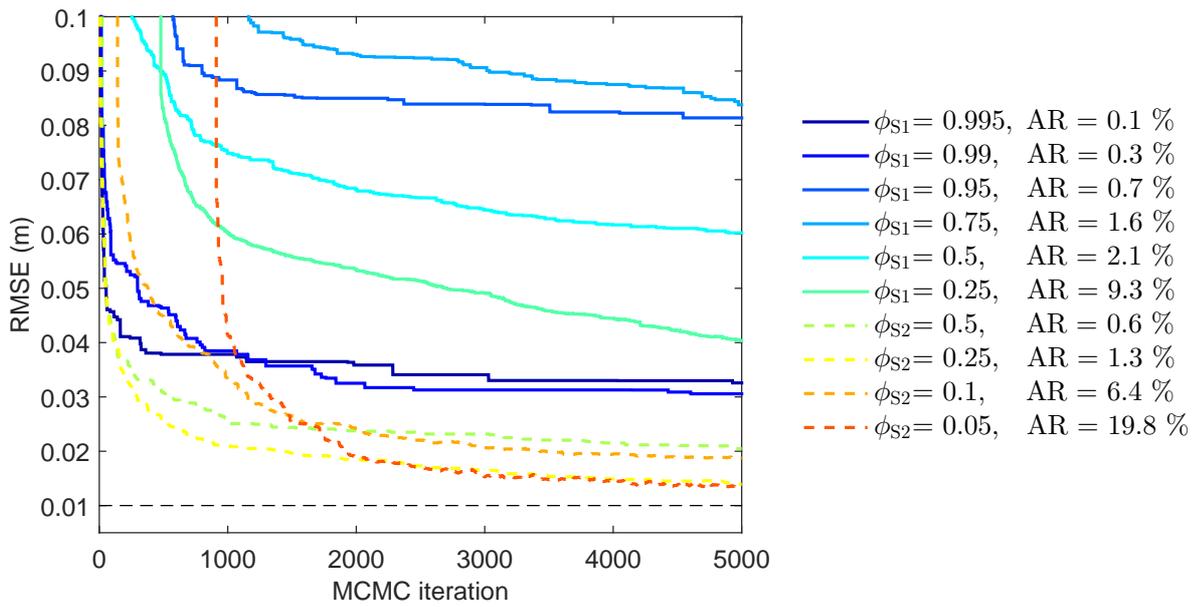


Figure 2: Trace plot of the mean sampled RMSE values across 4 repetitions for the tested conditioning strategies. Solid and dashed colored lines denote resimulating a set of points (S1) and a box-shaped area (S2), respectively. Each color represents a given size of the (randomly located) model fraction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

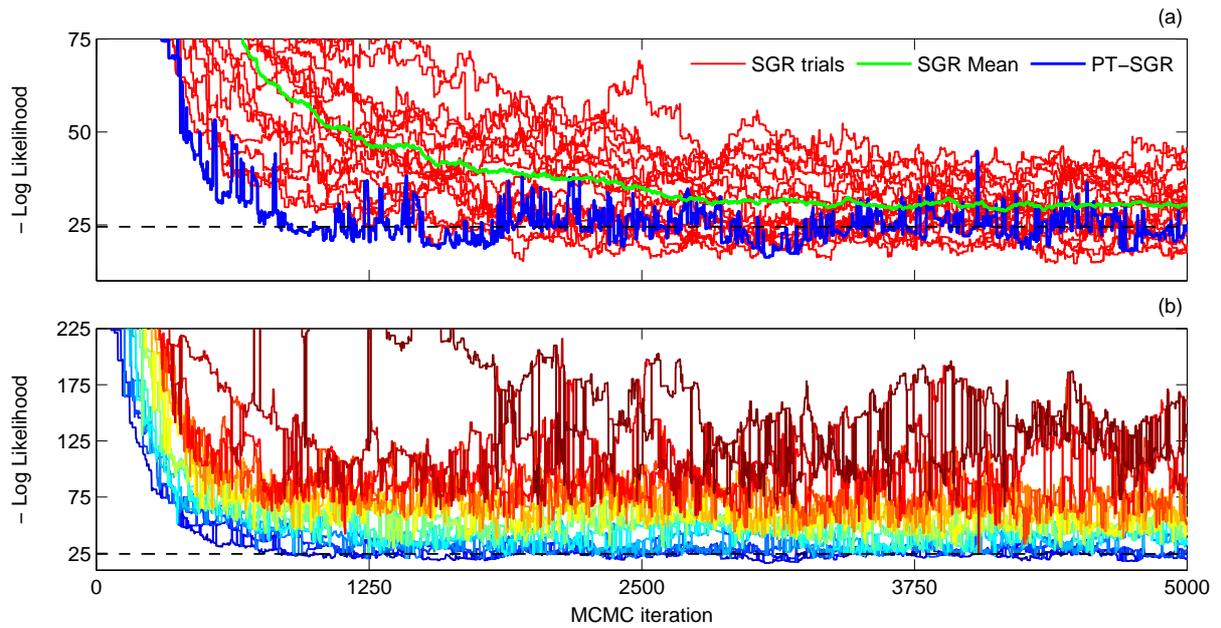


Figure 3: (a) Trace plot of the sampled negative log-likelihood values by the unit temperature chain evolved by PT-SGR (blue line) and the 16 independent SGR trials (red lines) for case study 1. The green line denotes the mean trajectory of the 16 SGR trials. (b) Trace plot of the sampled negative log-likelihood 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.

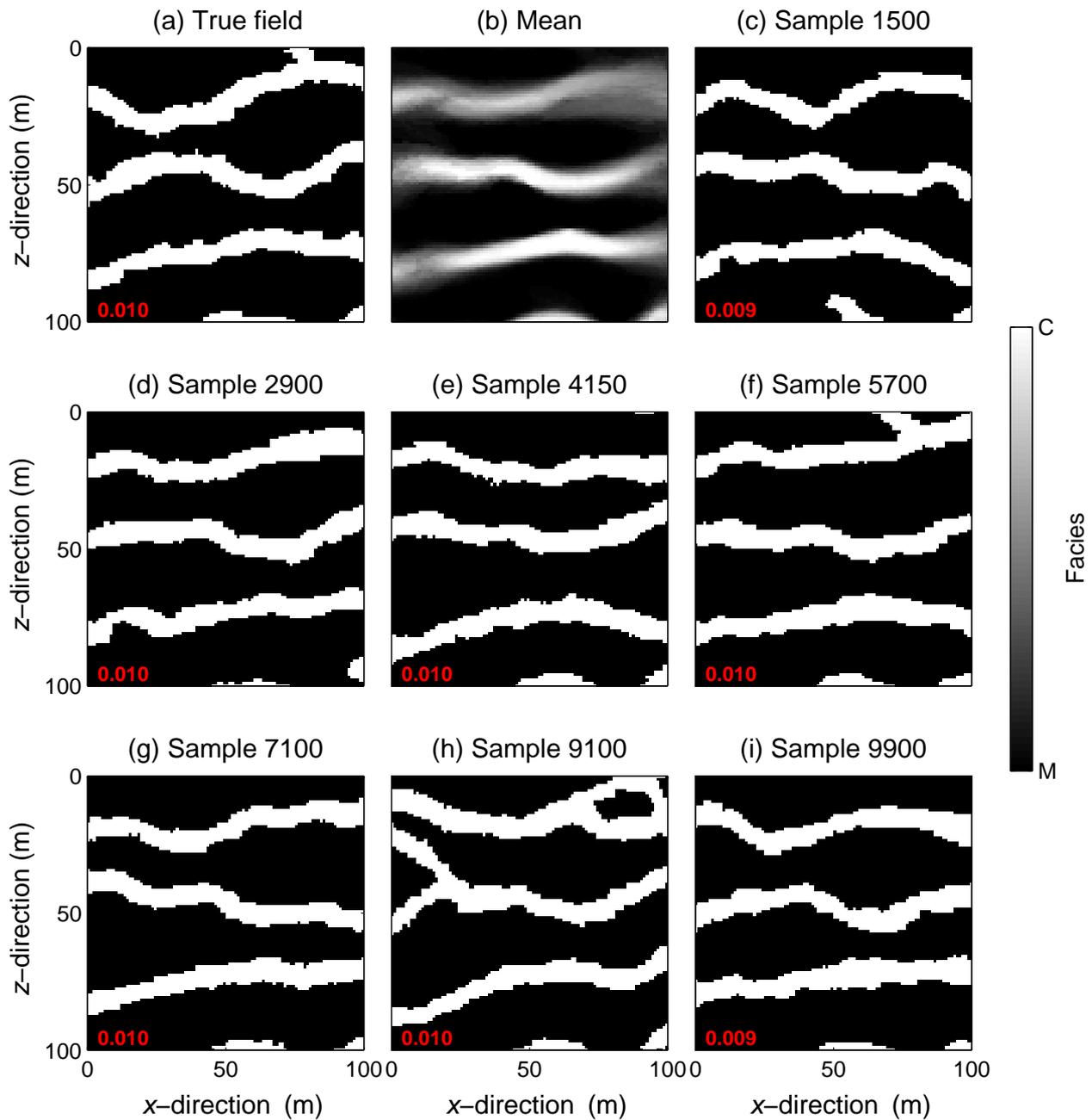


Figure 4: Posterior mean and 8 successive posterior realizations taken at regular intervals throughout sampling for the PT-SGR trial of case study 1. The posterior mean is computed on the basis of the samples produced by the unit temperature chains after a burn-in of 1500 MCMC iterations and using a 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).

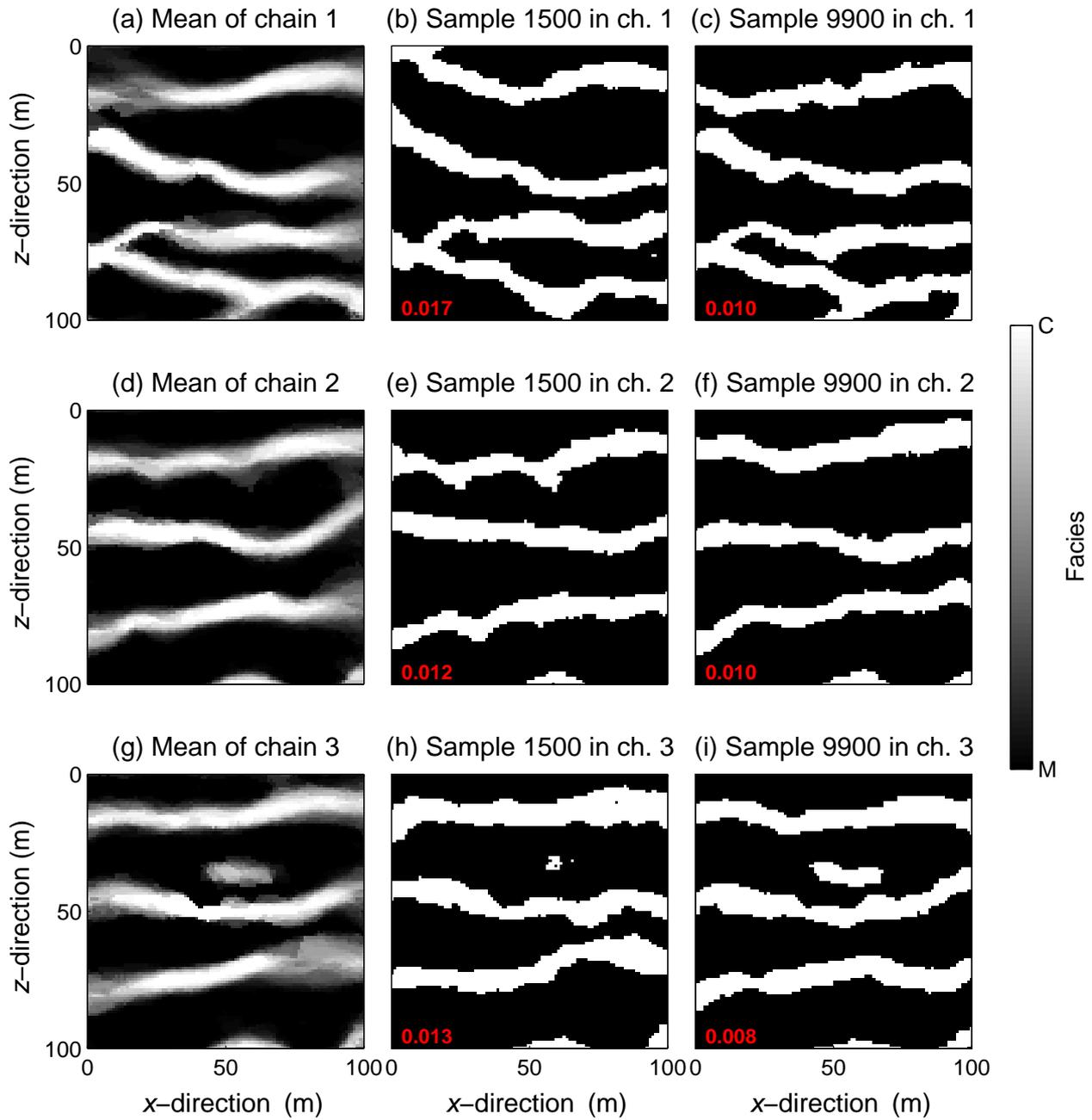


Figure 5: Mean sampled model over MCMC iterations 1500-10,000 (using a thinning factor of 50), and sampled models after 1500 and 9900 MCMC iterations for 3 out of the 16 independent SGR chains and case study 1. The red number in the lower left corner of each plot is the corresponding RMSE (m).<sup>42</sup>

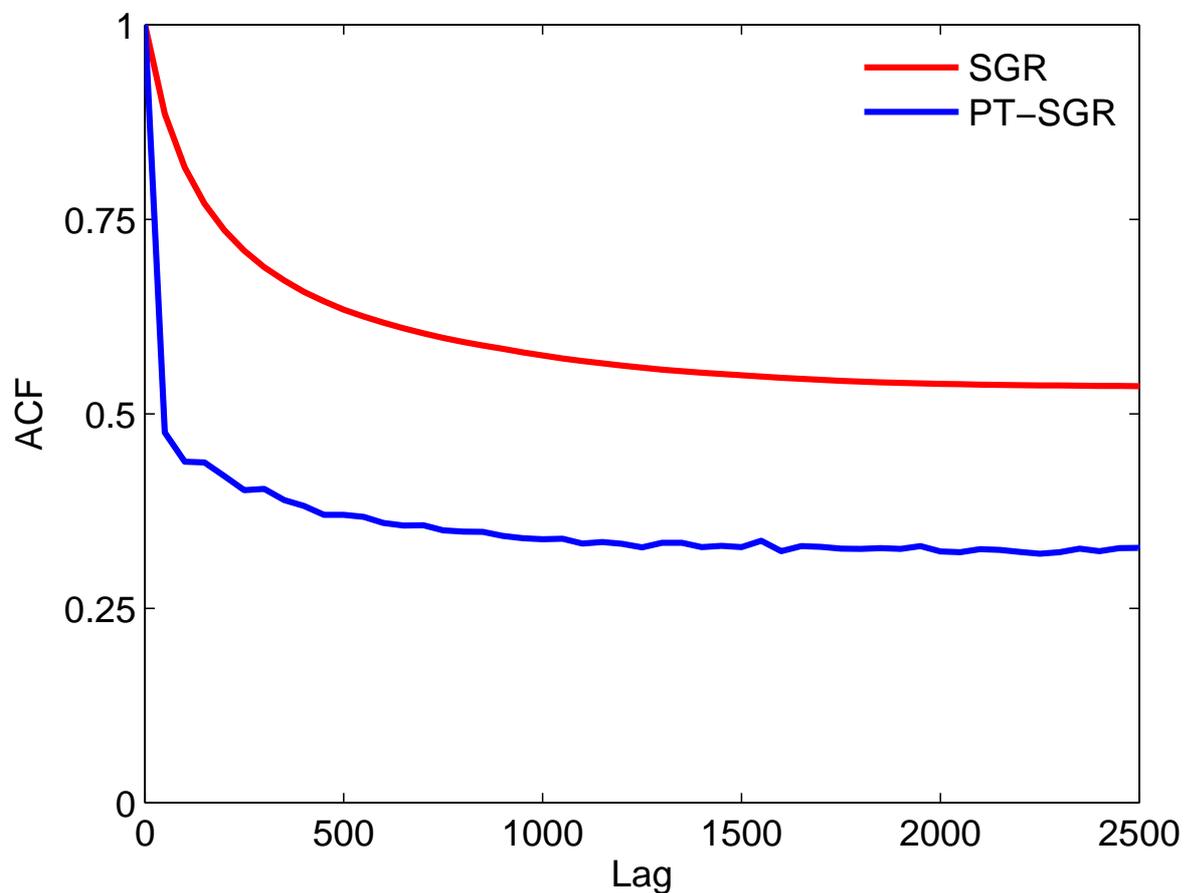


Figure 6: Mean autocorrelation function (ACF) of the 10,000 conductivity grid values derived from PT-SGR (blue line) or SGR (red line) for lags 0-2500 and case study 1. The lag-k autocorrelation is defined as the correlation between draws k lags apart. Listed statistics are computed for the last 8500 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.

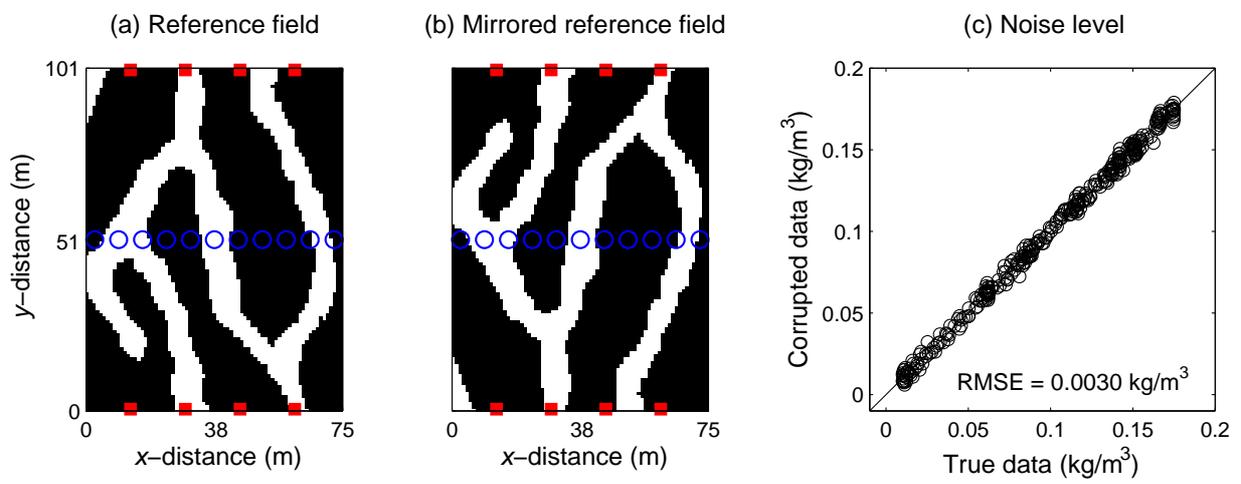


Figure 7: (a) Reference categorical field, (b) associated symmetric (mirrored) field and (c) noise-corrupted measurement data used for our second synthetic case study. In subfigures a and b, the red squares denote the application points of the tracer and the blue circles mark the locations of the pumping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

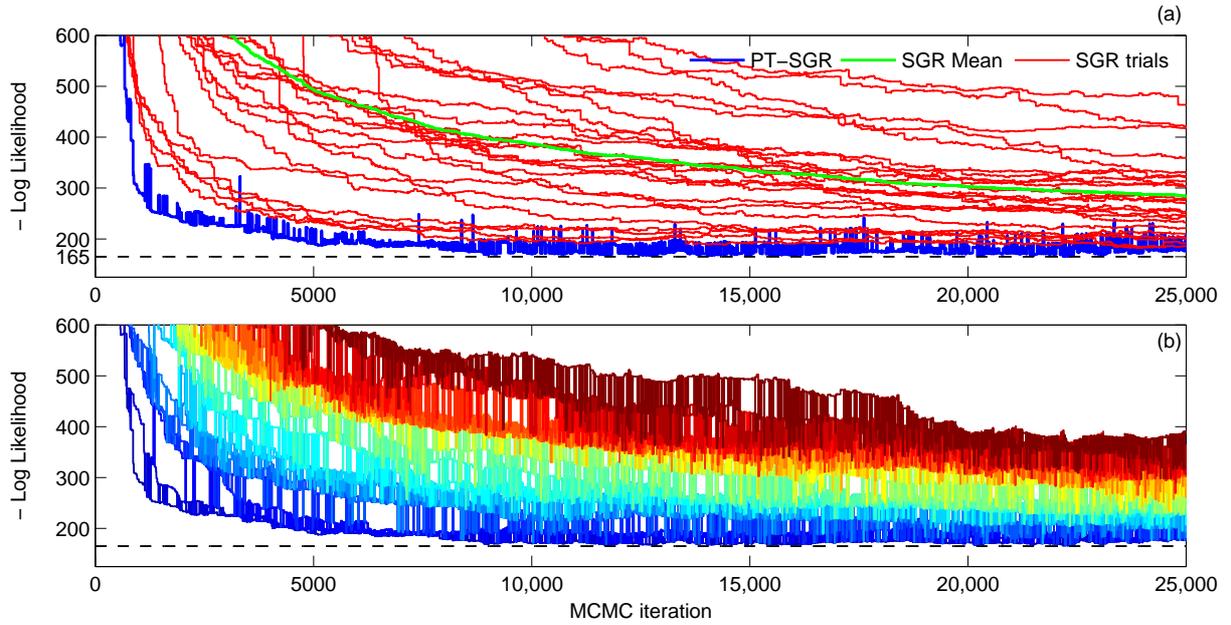


Figure 8: (a) Trace plot of the sampled negative log-likelihood values by the unit temperature chain evolved by PT-SGR (blue line) and the 24 independent SGR trials (red lines) for case study 2. The green line denotes the mean trajectory of the 24 SGR trials. (b) Trace plot of the sampled negative log-likelihood 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  $0.003 \text{ kg/m}^{-3}$ .

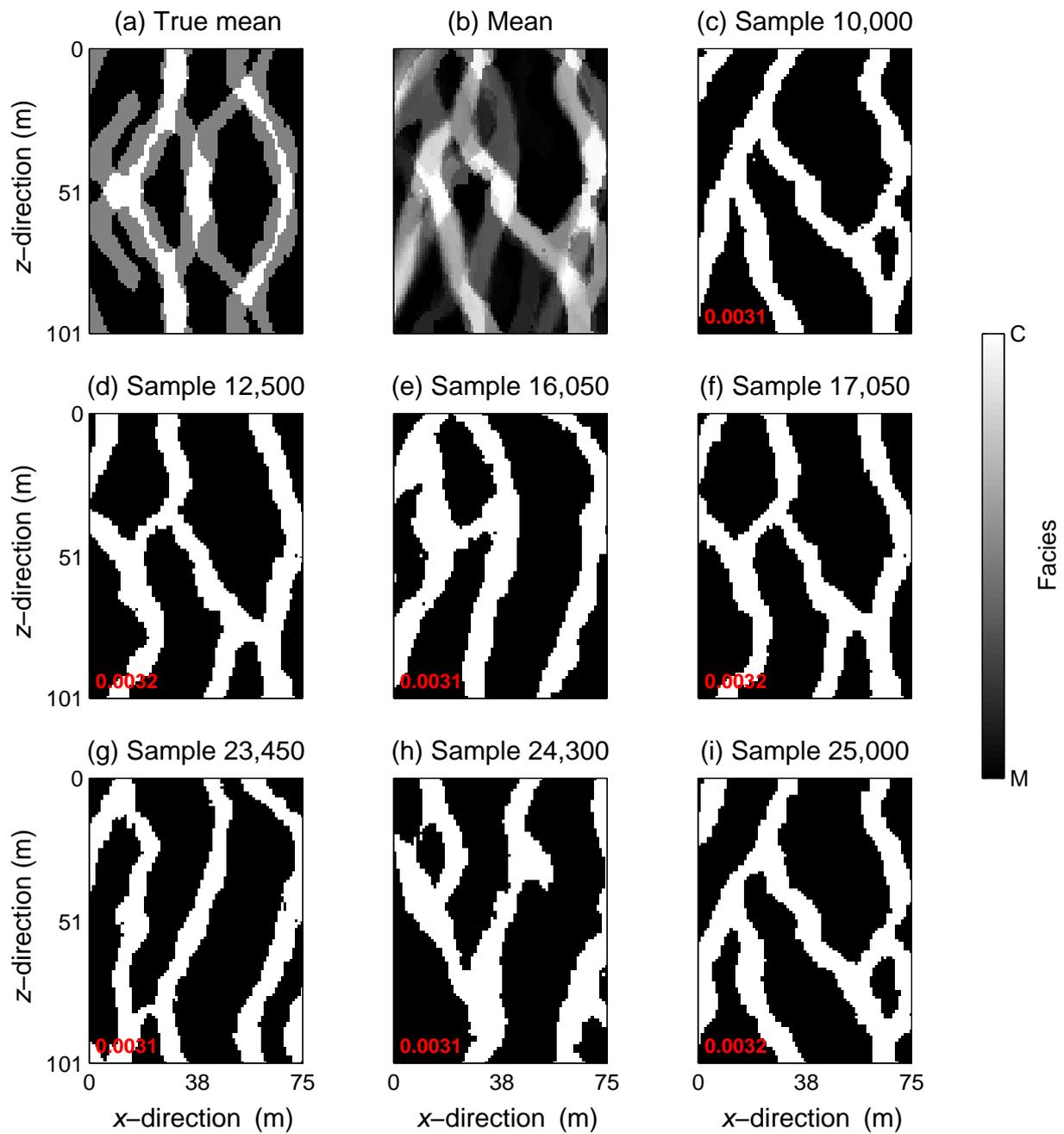


Figure 9: Mean and 8 successive model realizations taken at regular intervals throughout sampling for the PT-SGR trial of case study 2. The sample mean is computed on the basis of the samples produced by the unit temperature chains over iterations 10,000-25,000 and using a thinning factor of 50, thus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

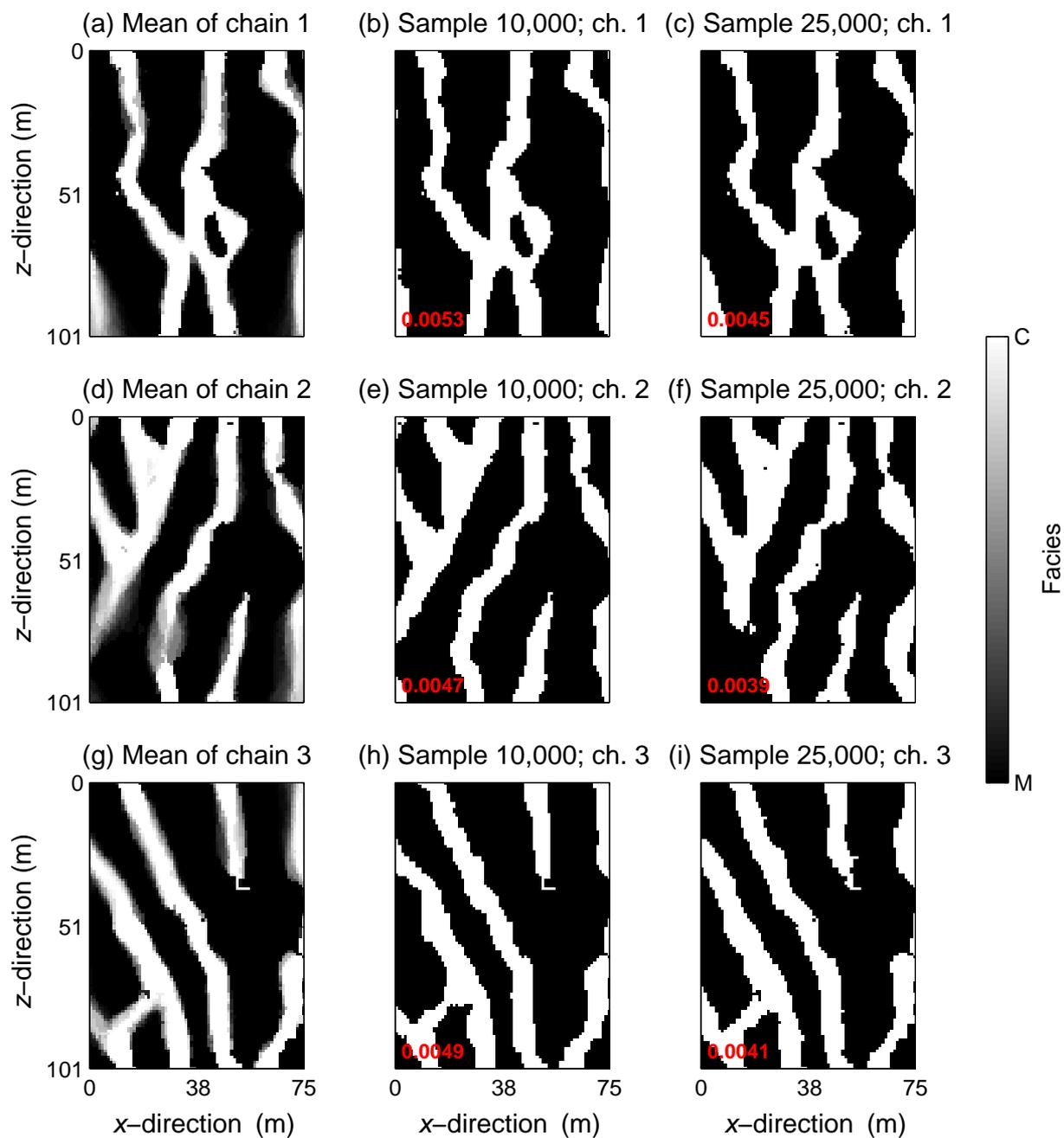


Figure 10: Mean sampled model over MCMC iterations 10,000-25,000 (using a thinning factor of 50), and sampled models after 10,000 and 25,000 MCMC iterations for 3 out of the 24 independent SGR chains and case study 2. The red number in the lower left corner of each plot is the corresponding RMSE (m)<sup>47</sup>.

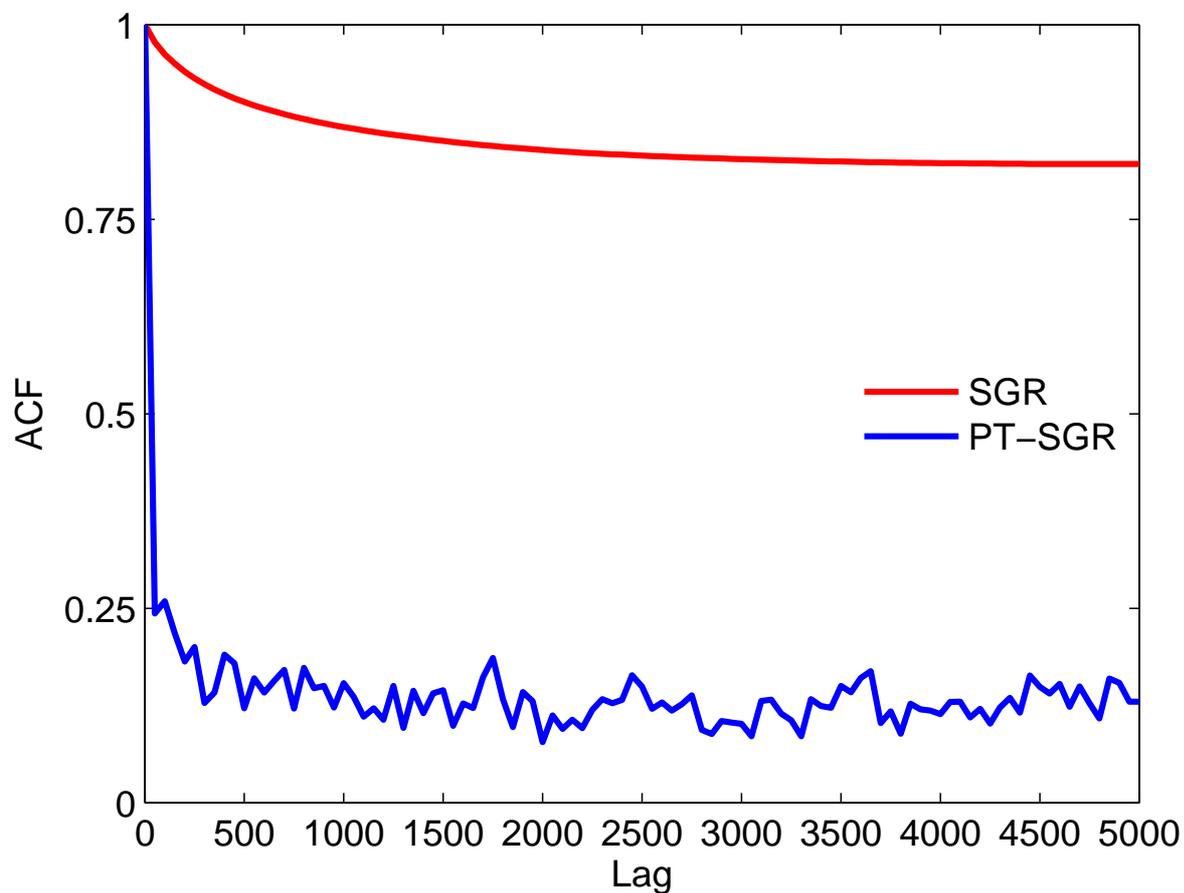


Figure 11: Mean autocorrelation function (ACF) of the 7575 conductivity grid values derived from PT-SGR (blue line) or SGR (red line) for lags 0-5000 and case study 1. The lag- $k$  autocorrelation is defined as the correlation between draws  $k$  lags apart. Listed statistics are computed for the last 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.