Rare event probability estimation for groundwater inverse problems with a two-stage Sequential Monte Carlo approach

Lea Friedli[1,2] and Niklas Linde[1]

[1] Institute of Earth Sciences, University of Lausanne, Switzerland
[2] Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland

Abstract

Bayesian inversions followed by estimations of rare event probabilities are often needed to analyse groundwater hazards. Instead of focusing on the posterior distribution of model parameters, the main interest lies then in the distribution of a specific quantity of interest contingent upon these parameters. To address the associated methodological challenges, we introduce a two-stage Sequential Monte Carlo approach. In the first stage, it generates particles that approximate the posterior distribution; in the second stage, it employs subset sampling techniques to assess the probability of the rare event of interest. By considering two hydrogeological problems of increasing complexity, we showcase the efficiency and accuracy of the resulting PostRisk-SMC method for rare event probability estimation related to groundwater hazards. We compare the performance of the PostRisk-SMC method with a traditional Monte Carlo approach that relies on Markov chain Monte Carlo samples. We showcase that our estimates align with those of the traditional method, but the coefficients of variation are notably lower for the same computational budget when targeting more rare events. Furthermore, we highlight that the PostRisk-SMC method allows estimating rare event probabilities approaching one in a billion using less than one hundred thousand forward simulations. Even if the presented examples are related to groundwater hazards, the methodology is well-suited for addressing a wide range of topics in the geosciences and beyond.

1 Introduction

Decision-making processes concerning groundwater and other environmental systems are subject to uncertainty. Consequently, decision-making often involves the identification and avoidance of hazards while assessing associated risks. While a hazard represents a dangerous phenomenon itself, risk considers the resulting potential of harm for human individuals or economic assets (Ward et al., 2020). Risk assessment plays a crucial role in the context of groundwater management, as fresh and uncontaminated groundwater is a prerequisite for global water security (Famiglietti, 2014) and as remediation of contaminated aquifers is extremely costly and time-consuming. Groundwater contamination and over-exploitation have not only direct adverse consequences for humans, but also for ecosystems and ecosystem services.

Our focus is on a particular aspect of risk assessment, namely, the estimation of the probability of a hazard occurring. This hazard is defined by a quantity of interest that takes on critical values. For a precise analysis of hazard occurrence, it is essential to consider the uncertainty associated with the parameters of a conceptual model. Hence, in the field of hydrogeology, Monte Carlo approaches for sampling uncertain hydrological model parameters have been widely employed (e.g., Lahkim and Garcia 1999, Khadam and Kaluarachchi 2003, Benekos et al. 2007, Siirila et al. 2012, Enzenhoefer et al. 2012). Such approaches can be challenging to apply in practice since hazards often fall under the category of rare events, requiring specialized modeling techniques to accurately represent the tail behavior of the quantity of interest. In this context, classical Monte Carlo estimation is impractical as it requires an excessively large sample (Cérou et al. 2012). One approach to mitigate the computational burden is to combine

Monte Carlo methods with surrogate modeling (e.g., Li and Xiu 2010), thereby speeding up the computation time of forward evaluations. Another option is to employ importance sampling in order to focus the sampling on critical regions of the quantity of interest. However, selecting a well-working importance density for high-dimensional problems is often difficult (Au and Beck, 2003a). Extreme value theory (e.g., Brodin and Klüppelberg 2008), relying on fitting an extreme value distribution to represent the distribution of the quantity of interest, offers yet another alternative and is widely used to predict probabilities of environmental hazards such as extreme floods (Morrison and Smith, 2002). Extreme value theory necessitates sizable sample sizes for distribution fitting, is contingent on the chosen distribution's shape, and does not offer simulations of the rare event (e.g., Diebold et al. 2000). An alternative data-intensive method for estimating extremes is based on the 'Peaks over Threshold' technique (POT; Leadbetter 1991). In this approach, extreme events are analysed by focusing on values that exceed a certain threshold.

We perform rare event probability estimation for the case when indirect site-specific data y are available (e.g., from tracer or pumping tests). We employ a Bayesian framework in which the hydrogeological parameters θ are characterized by a posterior probability density function (PDF) $p(\theta|y)$, given by the distribution of θ conditioned on measurements y. Compared to a standard Bayesian inversion problem in which the end-product is an approximation of the posterior PDF, we interrogate the distribution of a quantity of interest depending on the parameters through a non-linear relationship $\theta \mapsto \mathscr{R}(\theta)$, for instance, in the probability of this quantity exceeding a critical threshold (considered as the hazard). In practical scenarios, the presence of non-linearity frequently precludes the availability of an analytical formula for the distribution of the quantity of interest when conditioned on the data y.

In structural engineering, similar problems have been addressed by performing probabilistic updating of system parameters using dynamic data and subsequently updating the estimation of the system's reliability (e.g., Papadimitriou et al., 2001). In this context, Straub (2011) introduced the so-called Bayesian Updating with Structural reliability methods (BUS; e.g. Straub and Papaioannou 2015). For the Bayesian analysis, BUS can be interpreted as an extension of rejection sampling (Ripley 2009). To extend BUS for posterior rare event probability estimation, Straub et al. (2016) present an approach targeting both the posterior and the rare event by using reliability methods. A challenge of this method is the selection of the constant employed in the extended rejection sampling, as its choice can impact overall performance. In a similar approach targeting 'updated robust reliability measures', Jensen et al. (2013) rely on transitional MCMC (Ching and Chen, 2007) to derive a set of posterior samples followed by subset sampling for the reliability analysis. A very different approach enabling the combination of inference and rare event estimation that has been explored in the geosciences is Bayesian Evidential Learning (BEL; Hermans et al. 2016), which aims to learn a direct relationship between measurements and quantity of interest by sampling from the prior distribution (e.g., Thibaut et al. 2021). For higher-dimensional parameter spaces and non-linear relationships, it can be difficult for BEL to capture the full joint distribution with a reasonable number of samples.

We propose a two-stage application of Sequential Monte Carlo (SMC; Doucet et al. 2001), which we refer to as the Posterior Risk Sequential Monte Carlo (PostRisk-SMC) method. Bayesian inversion in hydrogeology and other environmental fields is often addressed using Markov chain Monte Carlo (MCMC) methods. For high-dimensional problems with non-linear forward solvers, standard MCMC methods often have difficulties in approximating the posterior PDF within realistic computational constraints. This happens as the Markov chains may be trapped in local minima for long times or have insignificant probabilities of switching between posterior modes (e.g., Neal 2001, Amaya et al. 2022). To overcome these challenges, methods

based on so-called power posteriors have been introduced. In a power posterior, the likelihood function is downweighted by exponentiating it with the inverse of a temperature greater than one, a process known as tempering. This facilitates more straightforward exploration at higher temperatures. Parallel tempering (Earl and Deem, 2005) is an MCMC approach where interacting chains target different power posteriors, allowing states sampled at higher temperatures to propose states in chains targeting the posterior distribution. In geophysical inversion, Sambridge (2014) illustrated that parallel tempering significantly enhances sampling efficiency, enabling a more extensive exploration of the parameter space in comparison to conventional MCMC methods. Similar to traditional MCMC techniques, parallel tempering approximates the posterior using states sampled post burn-in. In contrast, annealed importance sampling (Neal, 2001) relies on sequential importance sampling. While in MCMC, the accuracy of posterior estimates relies on the precise identification of the burn-in period for the chains, annealed importance sampling ensures asymptotic correctness through importance steps, even when errors occur in approximating intermediate distributions (Neal, 2001). The SMC method (Doucet et al., 2001) is based on annealed importance sampling. As a particle method, it provides a weighted sample of particles for posterior approximation by simulating a sequence of power posteriors transferring the prior PDF to the posterior PDF by successively increasing the weight of the likelihood (Del Moral et al. 2007). While the SMC method is extensively used in science and engineering, it has only seen limited use in the geosciences (i.e., Vrugt et al. 2013, Linde et al. 2017). We build our PostRisk-SMC method on an adaptive version of the SMC method by Zhou et al. (2016), which automatically tunes the cooling sequence between power posteriors. Recently, adaptive SMC methods have been employed successfully for geophysical (Amaya et al., 2021; Davies et al., 2023) and hydrogeological (Amaya et al., 2022) inversion problems, demonstrating superior performance compared with state-of-the-art MCMC methods.

Relying only on a particle approximation of the posterior PDF is insufficient when estimating rare event probabilities. As a relatively small number (tens or hundreds, sometimes thousands) of particles is used in practice, this means that no particle is likely to be associated with the rare event that might, for instance, have a probability of one in a million. To address this, a new SMC formulation has emerged that specifically targets rare events by employing a sequence of nested sets pertaining to the hazard scenario. This sequence refers to a hierarchical structure of sets with each set being a subset of the set above it. In a scenario targeting the probability of the quantity of interest exceeding a critical threshold, the nested sets are related to intervals $[T_k, \infty)$ with thresholds T_k increasing from minus infinity to the threshold of interest. This approach relies on the fact that the small probability of the rare event can be expressed as a product of larger conditional probabilities involving the intermediate nested sets. Such a splitting technique was first introduced as 'subset sampling' by Au and Beck (2001) in the context of reliability analysis and has been applied, for instance, in the context of radioactive waste management (e.g., Cadini et al. 2012) and earthquake engineering (e.g., Au and Beck 2003b). In the SMC literature, subset sampling is presented by Del Moral et al. (2006) and Johansen et al. (2006). Cérou et al. (2012) and Botev and Kroese (2008) extended the existing methods by using an adaptive method that optimally selects the subsets on the fly. Subset sampling has been further leveraged by employing surrogates (Bourinet et al., 2011) or by employing a multilevel approach (Ullmann and Papaioannou, 2015). While all of these applications rely on uncertain parameters θ following a 'prior' PDF, we here adapt this approach to rare event estimation with respect to a posterior PDF that is first approximated by adaptive SMC. The resulting PostRisk-SMC method relies on the same principles as the approach of Jensen et al. (2013) but within the theoretical formulation of particle methods and SMC. Further, Jensen et al. (2013) consider engineering applications and dynamic data, while we introduce the PostRisk-SMC in the context of hydrogeological rare event probability estimation. In addition, we perform resampling of the particles

only occasionally (during the posterior phase), while the transitional MCMC approach applied by Jensen et al. (2013) does so in every iteration. Since resampling impacts the variance of estimates (Douc and Cappé, 2005), it is usually beneficial to resample only when the variation in the particle weights becomes too high.

For comparison purposes, we consider a conventional Monte Carlo approach for the rare event probability estimation, as applied for instance by Dall'Alba et al. (2023) for risk assessment of groundwater inflow in the context of tunnel construction. In our inversion setting, we rely on MCMC samples approximating the posterior PDF for the Monte Carlo estimation. Our first example consists of a simplified one-dimensional flow scenario where we utilize pumping tests to estimate the probability of high flow rates. Subsequently, we consider a more realistic two-dimensional flow and transport problem, focusing on assessing the probability of contamination breakthrough. The remainder of the manuscript is organized as follows: Section 2 gives a methodological overview of the considered setting and introduces the PostRisk-SMC method; Section 3 presents the one-dimensional flow example and Section 4 the two-dimensional transport example; finally, the study ends with a discussion and conclusions in Sections 5 and 6, respectively.

2 Methodology

2.1 Notation

We target an unknown property vector $\theta \in \mathbb{R}^P$ representing a model domain from which we obtain measurements $y \in \mathbb{R}^M$. We consider a setting where measurements are realizations of the random variable $Y = \mathscr{G}(\theta) + \varepsilon_{\mathscr{O}}$, with $\mathscr{G} : \mathbb{R}^P \to \mathbb{R}^M$ referring to the forward operator and $\varepsilon_{\mathscr{O}}$ to the observational noise. Assuming independent Gaussian observational errors, we express the likelihood as $p(y|\theta) = \varphi_M(y; \mathscr{G}(\theta), \Sigma_Y)$, with $\varphi_M(\cdot; \mathscr{G}(\theta), \Sigma_Y)$ denoting the PDF of a M-variate normal distribution with the mean $\mathscr{G}(\theta)$ and the diagonal covariance matrix Σ_Y of the observational errors. While we have opted for the simplicity of assuming independent Gaussian observational errors, the methodology remains applicable in a broader context, accommodating alternative error assumptions.

We consider a quantity of interest $R = \mathscr{R}(\theta)$ derived from θ via some function $\mathscr{R} : \mathbb{R}^P \to \mathbb{R}$. More specifically, we target a rare set $A = \{\theta \in \mathbb{R}^P : \mathscr{R}(\theta) \in \mathscr{T}\}$ for some interval $\mathscr{T} \subseteq \mathbb{R} \cup \{\infty, -\infty\}$. If we target the exceedance probability $\mathbb{P}(\mathscr{R}(\theta) \geq T)$ for some real number T, we assign $\mathscr{T} = [T, \infty)$. We are interested in $\mathbb{P}(\theta \in A|y)$ for θ distributed according to the posterior PDF $p(\theta|y)$ and write,

$$\mathbb{P}(\theta \in A|y) = \int_A p(\theta|y)d\theta. \tag{1}$$

2.2 Bayesian inversion and Metropolis–Hastings

In Bayes' theorem, the posterior PDF is given by,

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{2}$$

with the prior PDF $p(\theta)$ of the model parameters, the likelihood function $p(y|\theta)$ and the evidence $p(y)$. As in practice, it is often not possible to sample directly from the posterior when

the forward solver $\theta \mapsto \mathscr{G}(\theta)$ is non-linear, sampling methods such as MCMC and SMC can be applied.

The most used MCMC method is the Metropolis–Hastings algorithm (MH algorithm; Metropolis et al. 1953; Hastings 1970). The MH algorithm is an iterative algorithm that, in each iteration, proposes a new set of model parameter values, which is then accepted or rejected based on the acceptance probability. The choice of the proposal density is crucial, as it has to balance the trade-off between exploration and exploitation. While standard Gaussian model proposals can be applied for a model space with reduced dimension, more high-dimensional parameter spaces present considerable challenges (e.g., Robert et al. 2018). To ensure robustness against different discretization choices and to maintain a reasonable step size while inferring thousands of unknowns, we rely on preconditioned Crank-Nicolson proposals that preserve the prior PDF (pCN; e.g. Cotter et al. 2013). The utilization of such prior-preserving proposals results in the acceptance probability being solely dependent on the likelihood values. In the field of geophysics, MCMC algorithms with model proposals that preserve the prior are known as extended Metropolis (Mosegaard and Tarantola, 1995). The pCN proposals have been utilized for instance in a parallel tempering approach by Xu et al. (2020).

2.3 From Sequential Monte Carlo to PostRisk-SMC

In this Section, we first introduce Sequential Monte Carlo for posterior inference (Section 2.3.1) and Sequential Monte Carlo for rare event estimation (Section 2.3.2). Subsequently, we introduce PostRisk-SMC , a novel sequential combination of both methods, designed to tackle the challenge of estimating rare event probabilities while accounting for posterior uncertainty (Section 2.3.3). For the methodology of the first stage (Section 2.3.1), we rely on the framework of Del Moral et al. (2007) and Zhou et al. (2016) and refer to their works for further details such as convergence behaviour. Likewise, for the second stage (Section 2.3.2), we follow the framework presented by Cérou et al. (2012) and suggest consulting their paper for additional information.

2.3.1 Sequential Monte Carlo for posterior inference

Posterior estimation with the SMC method is based on a particle approximation using N particles $\{\theta^{(1)}, \theta^{(2)}, ..., \theta^{(N)}\}$ with weights $\{W^{(1)}, W^{(2)}, ..., W^{(N)}\}$. If the particles are sampled according to the posterior, the weights are redundant and reduce to $1/N$. In practice, it is generally not possible to sample from the posterior and importance sampling using a density $\eta(\theta|y)$ is applied. Importance sampling generates samples from an importance distribution that assigns higher probabilities to regions where the target distribution is expected to have most of its mass, thereby reducing the variance of estimators (e.g. Owen and Zhou 2000). To achieve a well-working importance sampling approach for the posterior PDF, one should strive for a $\eta(\theta|y)$ as close as possible to $p(\theta|y)$. This can be achieved by building a sequence of K PDFs $\{p_0(\theta|y), p_1(\theta|y), ..., p_K(\theta|y)\}$ with $p_0(\theta|y) = p(\theta)$ and $p_K(\theta|y) = p(\theta|y)$, thus moving gradually from the prior PDF to the posterior PDF (Del Moral et al. 2007). The sequence is built on unnormalized power posteriors (Neal 2001),

$$p_k(\theta|y) = p(y|\theta)^{\alpha_k} p(\theta), \tag{3}$$

with $0 = \alpha_0 < \alpha_1 < ... < \alpha_K = 1$. With increasing exponent α_k, the relative influence of the likelihood on the power posterior grows. For a smaller exponent, the exponeniated term is
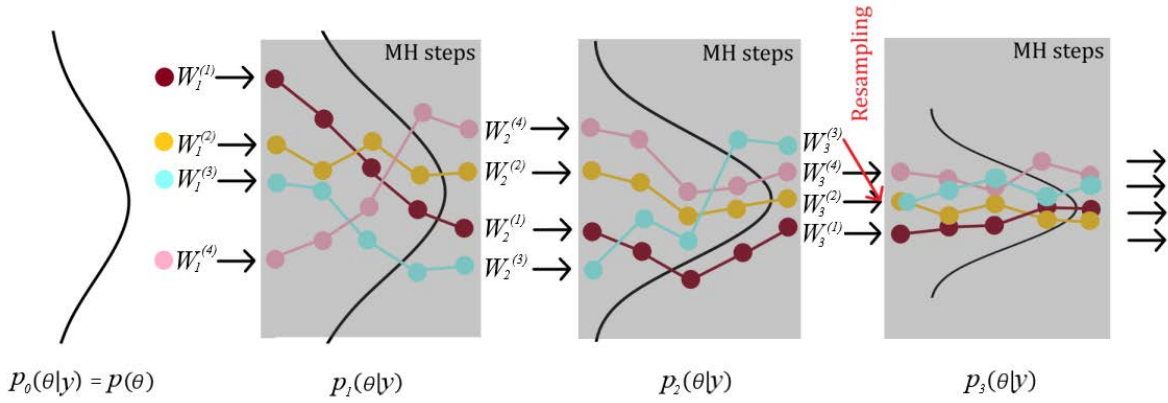
Figure 1: Illustration of the SMC method for posterior inference. We depict the first four power posteriors for an example with $N = 4$ particles and $s_P = 4$ MH steps.

'flatter' such that the power posterior is closer to the prior PDF. When using the importance density $\eta(\theta|y)$ to sample the particles $\theta^{(p)}$, the weights $W^{(p)}$ correspond to the normalized version of the importance weights $w^{(p)} = p(\theta^{(p)}|y)/\eta(\theta^{(p)}|y)$.

We start at iteration $k = 0$ with particles $\theta_0^{(p)}$ ($p = 1, 2, ..., N$) sampled from the prior PDF $p_0(\theta|y) = p(\theta)$ and initial weights $W_0^{(p)}$ being all equal to $1/N$. At iteration k of the SMC method, $p_k(\theta|y)$ is approximated by importance sampling based on the previously estimated power posterior $p_{k-1}(\theta|y)$. Therefore, the particles $\theta_{k-1}^{(p)}$ are assigned with incremental weights,

$$w_k^{(p)} = \frac{p_k\left(\theta_{k-1}^{(p)}|y\right)}{p_{k-1}\left(\theta_{k-1}^{(p)}|y\right)} = \frac{p\left(y|\theta_{k-1}^{(p)}\right)^{\alpha_k}}{p\left(y|\theta_{k-1}^{(p)}\right)^{\alpha_{k-1}}} = p\left(y|\theta_{k-1}^{(p)}\right)^{\alpha_k - \alpha_{k-1}}. \tag{4}$$

To account for the previous importance sampling steps, the cumulative normalized weights $W_k^{(p)}$ of the particles $\theta_{k-1}^{(p)}$ are defined as,

$$W_k^{(p)} = \frac{W_{k-1}^{(p)} w_k^{(p)}}{\sum_{j=1}^N W_{k-1}^{(j)} w_k^{(j)}}, \tag{5}$$

taking into account the history of weights and normalizing them to ensure their sum equals one. The particles $\theta_{k-1}^{(p)}$ approximating $p_{k-1}(\theta|y)$ are generated by propagating each particle $\theta_{k-2}^{(p)}$ according to a Markov kernel leaving $p_{k-1}(\theta|y)$ invariant (Neal 2001). This can be achieved by employing a finite number s_P of MH steps (Del Moral et al., 2007). In contrast to MCMC methods, the MH steps used within the SMC method do not have to converge as the importance sampling weights account for any possible sampling from the wrong distribution (Del Moral et al. 2007). The SMC procedure for posterior inference is illustrated in Figure 1.

When the (empirical) variance of the weights $W_k^{(p)}$ at iteration k becomes large, it is beneficial to resample the particles before propagation (Del Moral et al. 2007, Doucet and Johansen 2009). Resampling decreases the variance of the weights by discarding most particles with low weights and preferably reproducing those with high weights. Here, we use systematic resampling (Doucet and Johansen 2009). Subsequently, the weights $W_k^{(p)}$ are set to $1/N$, as the resampled particles are approximately distributed according to $p_k(\theta|y)$. Resampling increases the variance of the estimator, making it wasteful if the importance weights do not exhibit sig-

nificant variability (Del Moral et al., 2006). To decide when resampling is to be performed, the effective sample size (ESS; Kong et al. 1994),

$$ESS_k = \frac{\left(\sum_{p=1}^{N} W_{k-1}^{(p)} w_k^{(p)}\right)^2}{\sum_{o=1}^{N} \left(W_{k-1}^{(p)}\right)^2 \left(w_k^{(p)}\right)^2}, \tag{6}$$

is used. For instance, Del Moral et al. (2006) apply the decision rule of resampling if the ESS_k falls below 30 % of the number of particles N. To ensure that the final particles are a (unweighted) approximation of the posterior, we enforce a resampling step in the last iteration.

When defining the sequence of exponents α, one has to consider that too large differences between α_{k-1} and α_k lead to a large discrepancy between the power posteriors $p_{k-1}(\theta|y)$ and $p_k(\theta|y)$ and a subsequent high variance of the importance sampling estimator. However, if the difference is very small, an excessive number of steps are needed until $\alpha_k = 1$ is reached. It is natural to aim for a similar discrepancy between successive power posteriors (Zhou et al. 2016). To select the sequence of exponents α, we use the adaptive method of Zhou et al. (2016), based on the conditional effective sample size (CESS),

$$CESS_k = N \frac{\left(\sum_{p=1}^{N} W_{k-1}^{(p)} w_k^{(p)}\right)^2}{\sum_{p=1}^{N} W_{k-1}^{(p)} \left(w_k^{(p)}\right)^2}. \tag{7}$$

The $CESS_k$ quantifies the quality of $p_{k-1}(\theta|y)$ as an importance density to estimate expectations under $p_k(\theta|y)$ (Zhou et al. 2016). The *CESS* is equal to the *ESS* when resampling is conducted at each iteration. Zhou et al. (2016) show that using the *CESS* for the adaptive sequence leads to a reduction in estimator variance compared to an approach using the *ESS*. To define the next α_k, a binary search for the value for which the *CESS* is the closest to a pre-defined target value $CESS^*$ is performed (Zhou et al., 2016). A binary search operates by iteratively halving the interval containing the potential values, effectively reducing the search space with each step by comparing the target value to the middle element. If the target is less than the middle element, the search is restricted to the lower half of the interval; if it's greater, the search is limited to the upper half. The closer this target value $CESS^*$ is to N, the better the approximation, but the slower the algorithm becomes as the number of power posteriors grows. The SMC algorithm stops when α_k reaches one. Such an adaptive approach is expected to result in a more efficient algorithm compared to its non-adaptive counterpart. Importantly, it also leads to a more automated algorithm by minimizing the number of user-defined tuning parameters (Beskos et al., 2016). However, using an adaptive method for the selection of the exponents introduces a slight bias into the results. Beskos et al. (2016) explore the convergence behaviour for such adaptive approaches and establish that the output satisfies a weak law of large numbers and a central limit theorem. To indicate if we use an adaptive or fixed sequence of exponents, we specify the binary variable ADA_P as 1 for an adaptive and 0 for a predetermined selection. The full workflow of the SMC method for posterior inference is summarized in Figure 2.

2.3.2 Sequential Monte Carlo for rare event estimation

The SMC method can be modified to enable simulation of rare events and estimation of their probabilities by using a sequence of not-so-rare nested events (Del Moral et al., 2006; Johansen et al., 2006; Cérou et al., 2012). It is assumed that θ is a random element on $\mathbb{R}^P$ with probability distribution $p(\theta)$ that can be sampled from. To estimate $\mathbb{P}(\theta \in A)$, the SMC method for rare
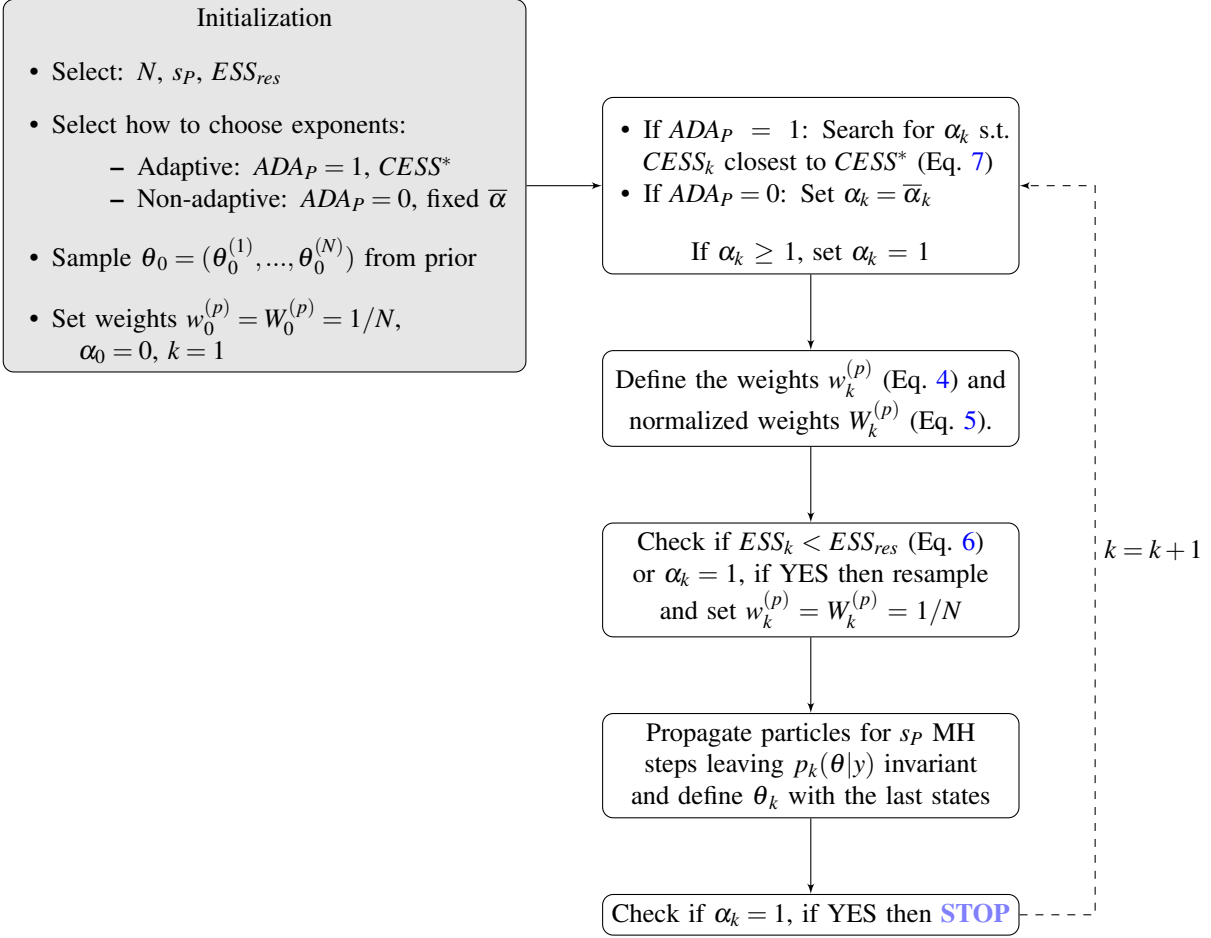
- If $ADA_P = 1$: Search for α_k s.t. $CESS_k$ closest to $CESS^*$ (Eq. 7)
- If $ADA_P = 0$: Set $\alpha_k = \overline{\alpha}_k$

  If $\alpha_k \geq 1$, set $\alpha_k = 1$

Define the weights $w_k^{(p)}$ (Eq. 4) and normalized weights $W_k^{(p)}$ (Eq. 5).

Check if $ESS_k < ESS_{res}$ (Eq. 6) or $\alpha_k = 1$, if YES then resample and set $w_k^{(p)} = W_k^{(p)} = 1/N$

Propagate particles for s_P MH steps leaving $p_k(\theta|y)$ invariant and define θ_k with the last states

Check if $\alpha_k = 1$, if YES then **STOP**

$k = k + 1$

Figure 2: Flow chart illustrating the SMC method for posterior inference.

event estimation employs a sequence of nested sets $A_k = \{\theta \in \mathbb{R}^P : \mathscr{R}(\theta) \in \mathscr{T}_k\}$, with $\mathbb{R}^P = A_0 \supset A_1 \supset ... \supset A_K = A$. It holds that,

$$\mathbb{P}(\theta \in A) = \prod_{k=1}^{K} \mathbb{P}(\theta \in A_k | \theta \in A_{k-1}). \tag{8}$$

If we are interested in $\mathbb{P}(\mathscr{R}(\theta) \geq T)$, the sequence of nested sets $A_k = \{\theta \in \mathbb{R}^P : \mathscr{R}(\theta) \in [T_k, \infty)\}$ corresponds to a sequence of increasing thresholds $\{T_0, ..., T_K\}$ with $T_0 = -\infty$ and $T_K = T$. For $\mathbb{P}(\mathscr{R}(\theta) \leq T)$, we employ $A_k = \{\theta \in \mathbb{R}^P : \mathscr{R}(\theta) \in (-\infty, T_k]\}$ using a sequence of decreasing thresholds with $T_0 = \infty$ and $T_K = T$.

The SMC method for rare event estimation starts by initializing N particles $\theta_0 = (\theta_0^{(1)}, ..., \theta_0^{(N)})$ sampled from $p(\theta)$. The first intermediate distribution $p_{A_0}(\theta) = p(\theta | \theta \in A_0)$ is equal to $p(\theta)$. To approximate the intermediate distribution $p_{A_k}(\theta) = p(\theta | \theta \in A_k)$ for $k \geq 1$, each particle $\theta_{k-1}^{(p)}$ is assigned a weight,

$$W_k^{(p)} = \begin{cases} 1/|I_k|, & \text{if } \theta_{k-1}^{(p)} \in A_k \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

with $I_k = \{p : \theta_{k-1}^{(p)} \in A_k\}$ and $|I_k|$ denoting its cardinality. Thereby, we are assuming that I_k is non-empty, otherwise the particle system dies. Subsequently, systematic resampling (Doucet and Johansen 2009) is employed such that particles which do not lie in A_k are replaced by
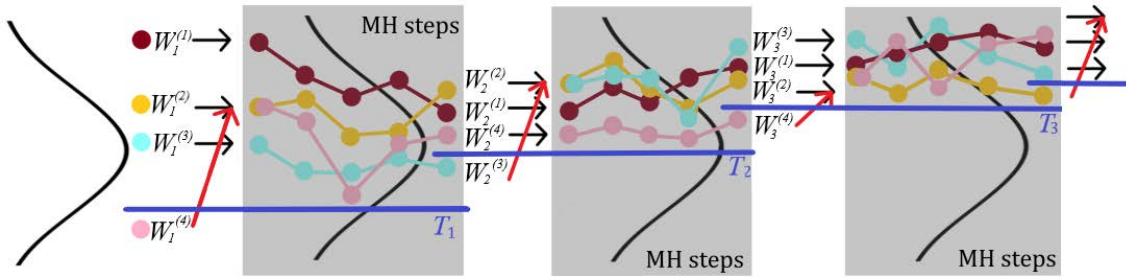
Figure 3: Illustration of the SMC method for rare events targeting $\mathbb{P}(\mathscr{R}(\theta) \geq T)$. We depict the first three thresholds for an example with $N = 4$ particles, $s_R = 4$ MH steps and a quantile of $\gamma = 0.25$.

particles that do. The resampled particles are propagated using a Markov kernel, leaving $p_{A_k}(\theta)$ invariant (Cérou et al. 2012). We are considering s_R steps with a MH algorithm whereby a transition is only accepted if θ stays in A_k. The procedure of SMC for rare event estimation targeting $\mathbb{P}(\mathscr{R}(\theta) \geq T)$ is illustrated in Figure 3.

We need to choose a sequence of nested sets such that $\mathbb{P}(\theta \in A_k | \theta \in A_{k-1})$ is reasonably high. Cérou et al. (2012) detail both a fixed and an adaptive algorithm. For $A_k = \{\theta \in \mathbb{R}^P : \mathscr{R}(\theta) \in [T_k, \infty)\}$, an adaptive method based on quantiles of $\mathscr{R}(\cdot)$ of the particles ensures that the asymptotic variance of the estimator is minimal (see Cérou et al. 2012). Utilizing the γ-quantile,

$$T_k = q_\gamma(\mathscr{R}(\theta_{k-1})),\tag{10}$$

guarantees that a ratio of $(1 - \gamma)$ of the particles survive. The adaptive algorithm's stopping criterion is met when the quantile surpasses the targeted threshold, at which point the last T_K is set equal to T. Then, $\mathbb{P}(\theta \in A)$ is estimated by multiplication of all $P_k = |I_k|/N$ for $k = 1, ..., K$. Due to the adaptiveness of the thresholds, the resulting estimator is biased given the finite number of particles N (Au and Beck, 2001). This bias is positive and becomes negligible compared to the variance of the estimator as the number of particles increases (Cérou et al., 2012). To circumvent this bias, one can either re-run the algorithm with the previously optimized sequence or use a predetermined fixed sequence of thresholds. With the binary variable ADA_R we indicate if we use fixed ($ADA_R = 0$) or adaptive ($ADA_R = 1$) sequences of thresholds. The work flow of the SMC method for rare event estimation is summarized in the flow chart in Figure 4.

2.3.3 Posterior Risk Sequential Monte Carlo method

To estimate $\mathbb{P}(\theta \in A|y)$, we introduce a sequential combination of the two SMC methods described in Sections 2.3.1 and 2.3.2 (PostRisk-SMC) . Let us write the k-th power posterior with respect to the subset A_k as,

$$p_k^A(\theta|y) = p(y|\theta)^{\alpha_k} p(\theta) \mathbb{1}\{\theta \in A_k\}.\tag{11}$$

While the first stage of the PostRisk-SMC algorithm generates particles distributed according to the posterior by increasing the exponent of the likelihood α_k with the subset A_k being held constant as $\mathbb{R}^p$, the second stage shrinks the subset while leaving the exponent of the power posterior at 1. For the rare event analysis, it is crucial that we start the second phase with a unweighted particle approximation of the posterior, ensured by the resampling step in the last step of the posterior inference stage. We denote as K_P the number of intermediate power posteriors, as K_R the number of thresholds and as $K = K_P + K_R$ their sum. Additionally, we define s_P as the number of MH steps employed between each importance sampling step in the
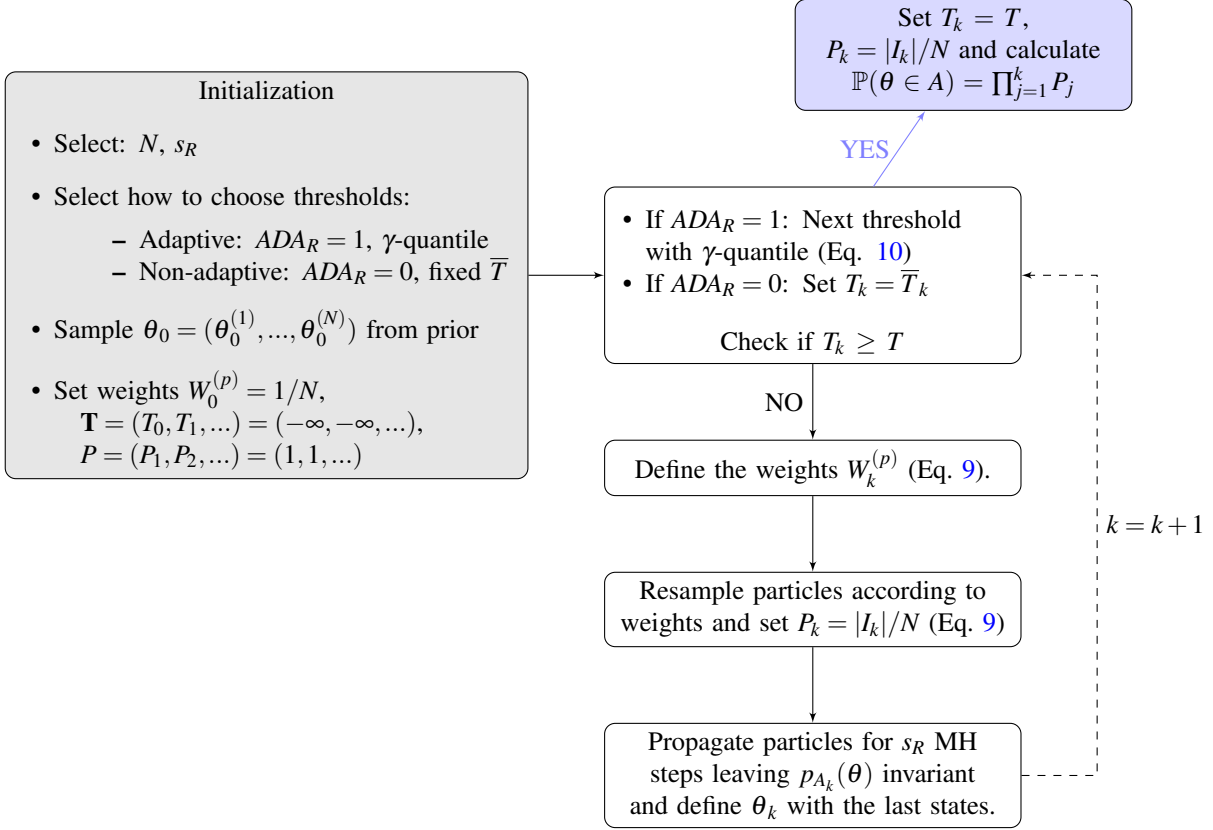
Figure 4: Flow chart illustrating the SMC method for rare event estimation of $\mathbb{P}(\mathscr{R}(\theta) \geq T)$.

posterior phase and s_R as the number between the subset sampling steps during the rare event phase. When the same number of steps is used for both, we denote it as $s = s_P = s_R$. The PostRisk-SMC method inherits the theoretical properties of the SMC methods utilized in the two stages, including any biases present in the estimators resulting from adaptive sequences of exponents and thresholds. The complete work flow of the PostRisk-SMC method is summarized in Figure 5.

In high-dimensional scenarios characterized by complex posterior distributions, the process of particle propagation using a limited number of MH steps can become limiting. In such contexts, the frequency of particle resampling becomes important to monitor. In the rare event probability estimation phase, this aspect becomes even more critical as frequent resampling is unavoidable. This implies the need to ensure that a sufficient number of MH steps are used to prevent particle collapse following the resampling steps.

In groundwater settings where the rare event revolves around contamination hazards, the simulation of the quantity of interest often demands more computational resources than the forward model used to estimate the posterior PDF. To achieve computational speed-up under such situations (as exemplified in Section 4), we introduce a minor modification to the propagation step during the rare event phase of PostRisk-SMC . Instead of simulating both the forward response and quantity of interest in every step, we conduct first a series of ss_R posterior steps within each of the s_R steps. Subsequently, the last state is treated as a proposal from the posterior which is accepted or rejected based on whether it falls within the current subset.
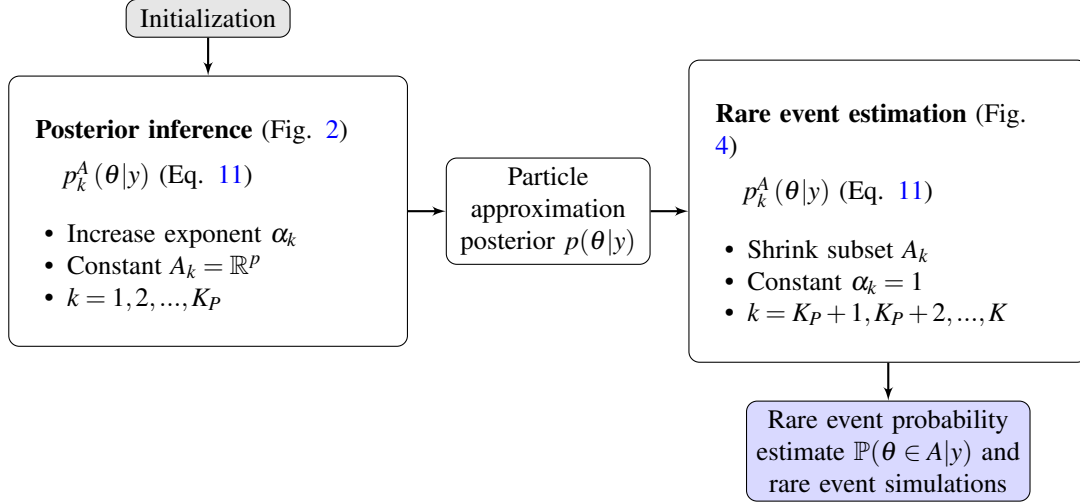
Figure 5: Work flow of the PostRisk-SMC method.

3 1D flow example

As a first example, we study a steady-state 1-D groundwater flow problem (diffusion equation). The chosen problem setting is inspired by a test case by Straub et al. (2016), which corresponds to the steady-state version of a test case introduced by Marzouk and Najm (2009). The fast run time of this simple toy example allows for a sensitivity analysis of the algorithmic parameters of the PostRisk-SMC method.

3.1 Synthetic setting

The model domain is the unit interval $D = [0,1]$ m and we consider the following steady-state equation,

$$\frac{d}{dx}\left(\theta(x)\frac{dh}{dx}\right) + b(x) = 0, \tag{12}$$

with hydraulic conductivity $\theta(x)$ [m/s], source $b(x)$ [1/s] and hydraulic head $h(x)$ [m].

The log-conductivity $\log\theta(x)$ is parameterized as a finite rank Gaussian random process expressed by,

$$\log\theta(x) = \mu_{\log\theta} + \sum_{i=1}^{n}\sqrt{w_i}v_i(x)Z_i, \tag{13}$$

with $\{w_i, v_i\}$ representing the first n eigenvalues and eigenfunctions from the Karhunen-Loève expansion (Loève, 1977) of a Gaussian process with mean $\mu_{\log\theta} = \log(10^{-5})$ and exponential covariance function $\kappa_{\log\theta}(\Delta x) = \sigma^2\exp(-\Delta x/l)$ with standard deviation $\sigma = 3$ and integral scale $l = 0.3$ m. Z_i denote independent standard normally-distributed variables. Following Straub et al. (2016), we employ a truncation after $n = 10$ terms. For the representation, we use a uniform grid with 40 intervals and under the assumption of the mean and covariance structure being known, we infer the ten first Z_i. The 'true' log-hydraulic conductivity values $\log\theta(x)$ are depicted in Figure 6a.

For the measurements, the source term $b(x)$ in Equation (12) is modelled using sources in the cells at 0.26, 0.51 and 0.76 m with identical strengths of 0.001 1/s. The measurements y are performed on the steady-state solution of $h(x)$ employing 7 sensors spaced uniformly on D
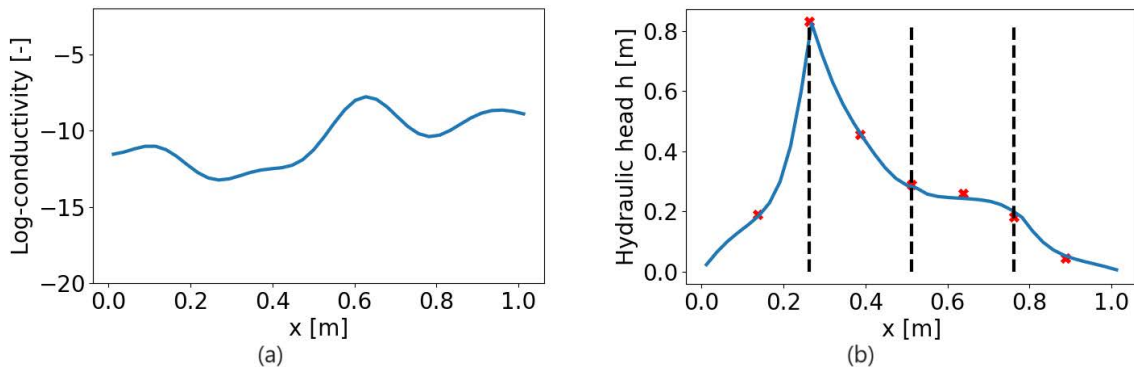
Figure 6: (a) 'True' log-hydraulic conductivity $\log \theta(x)$ on $D = [0,1]$ m and corresponding (b) steady-state solution $h(x)$ (solid line) for the diffusion equation including the pumping sources (source locations dashed) and the resulting noisy measurements y (crosses).

excluding the endpoints. To achieve this, Equation (12) is solved with linear finite differences on a uniform grid employing 40 cells and boundary conditions $h(0) = h(1) = 0$ m (Langtangen and Linge 2017). Finally, the synthetically-generated measurement values are contaminated with independent Gaussian errors having a standard deviation of 0.01 m (Fig. 6b).

For the rare event, we consider flow from the left to the right of the model domain and define the 'hazard' as the flow rate on the right boundary exceeding a critical value of T. To calculate the flow rate, we assume a hydraulic head difference of 1 m and take the harmonic mean of the conductivity values. To enable a comparison with MC estimation, we consider a first value of $T^* = 9 \times 10^{-6}$ m/s; the second value of $T^{**} = 9.5 \times 10^{-6}$ m/s is selected such that it targets a rare event with probability of one in a billion.

3.2 Results

We employ independent normal prior PDFs for the unknown Z_i of the KL-expansion representing the log-conductivity (Eq. 13). For the likelihood, we assume independent Gaussian measurement errors with the same standard deviation as used in the data generation process. We compare the results of the PostRisk-SMC method with those of a standard MH algorithm employing Gaussian proposals. To ensure an acceptance rate of approximately 30 %, the step width of the proposals is adjusted accordingly, taking into account the different scales of variation in the KL components (based on initial MH runs). The same configuration of the MH algorithm is used in the MH steps employed in each iteration of the PostRisk-SMC method.

For the PostRisk-SMC method, the following parameter choices have to be made: the number of particles N, the number of MH steps s in each iteration (here $s = s_P = s_R$), the selection of the exponents α_k (Eq. 7), the threshold ESS_{res} below which resampling is employed (Eq. 6) and the selection of the thresholds T_k (Eq. 10). Following Del Moral et al. (2006), we fix $ESS_{res} = 0.3 \times N$ for the resampling in the initial stage of posterior inference. We start by testing a configuration of PostRisk-SMC with $N = 40, CESS^* = 0.99 \times N$, $\gamma = 0.05$ and $s = 40$, employing adaptive schedules for the likelihood's exponents and the thresholds. Figure 7 depicts resulting particle approximations of the following distributions of the log-diffusivity profile: (a) prior $p_0^A(\theta|y) = p(\theta)$, (b) posterior $p_{K_P}^A(\theta|y) = p(y|\theta)p(\theta)$ and (c) posterior rare event $p_K^A(\theta|y) = p(y|\theta)p(\theta)\mathbb{1}\{\mathscr{R}(\theta) \geq T^*\}$. Considering our utilization of only 40 particles, it is not particularly problematic or unexpected that the true value may deviate outside the
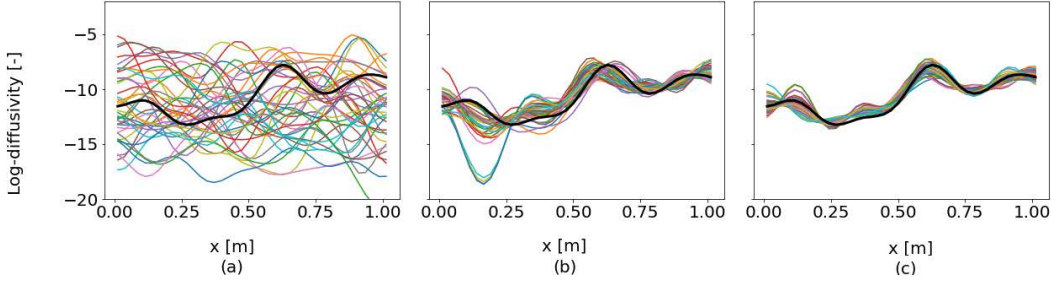
Figure 7: Results for the 1D flow example with the PostRisk-SMC method: Particle representation ($N = 40$) of the log-conductivity's (a) prior, (b) posterior and (c) posterior rare event (for T^*) distribution; the black lines depict the true profile and the coloured lines the particles.

expected range in specific areas of Figure 7c.

To explore the level of bias introduced by the adaptive schemes for our choice of $N = 40$ particles, we re-run the algorithm using the previously defined sequences as pre-defined values. The range of ten resulting estimates for T^{**} are depicted in Figure 8a (adaptive and re-run). The adaptive runs yield a mean estimate that is approximately 200 times greater than that of the re-runs. To circumvent bias while avoiding the computational burdens associated with increasing the number of particles or performing re-runs, we adopt in what follows a fixed sequence of thresholds for the rare event estimation part ($ADA_R = 0$ in Fig. 4). With K_P denoting the number of intermediate power posteriors and following the flow chart in Figure 5, the first threshold different from minus infinity is T_{K_P+1}. For the shape of the sequence, a suitable form can be determined, for example, by conducting an initial adaptive run (Fig. 8b). We use a logarithmic function,

$$f_T(k) = a \log(k) + T_{K_P+1}, \tag{14}$$

increasing from T_{K_P+1} to T^{**}. Therefore, we set the thresholds to $T_k = f_T(k - K_P)$ for $k = K_P + 1, ..., K$ and ensure that $T_K = f_T(K_R) = T^{**}$ by expressing $a = (T^{**} - T_{K_P+1})/\log(K_R)$. Finally, we change the closest value of T^* to this very value. For the first threshold, we test the choices of $T_{K_P+1} = 3, 5, 7 \times 10^{-6}$. The resulting threshold sequences are depicted in Figure 8b, together with the adaptive sequence utilizing $\gamma = 0.05$. The range and mean of ten estimates for T^{**} obtained with the different sequences are depicted in Figure 8a. We note that while the adaptive sequence leads to much higher values, the ones of the re-runs and the fixed sequences with the different T_{K_P+1} are comparable.

In our specific context, where the focus is on estimating the probability of rare events and the posterior of θ is rather smooth, the bias caused by the adaptive schedule in the first stage of posterior estimation is minimal. Tests (not shown) demonstrated that even when considering T^{**} and $N = 40$, the adaptive sequence for the posterior estimation resulted in an almost identical mean estimate compared to the re-runs (less than 0.02 % difference). As a result, we continue to use an adaptive sequence of exponents for the first stage of the algorithm ($ADA_P = 1$ in Fig. 2).

We now keep $T_{K_P+1} = 5 \times 10^{-6}$ and explore the influence of the remaining parameters on the rare event estimation. As a baseline configuration, we use $N = 20, CESS^* = 0.9 \times N$ (resulting in $K_P = 40$), $K_R = 100$ and $s = 20$, requiring 55,000 forward simulations for T^{**}. Next, we multiply the computational budget by a factor of ten, allocating these additional computational resources successively to each of the parameters. This results in $N = 200, CESS^* = 0.9999 \times N$ (such that $K_P = 1250$), $K_R = 1330$ and $s = 200$. The resulting ranges of the rare event probability estimates for T^{**} using ten runs are depicted in Figure 9a and the means and coefficients of variation (COV; ratio of standard deviation to the mean) for both thresholds are summarized in
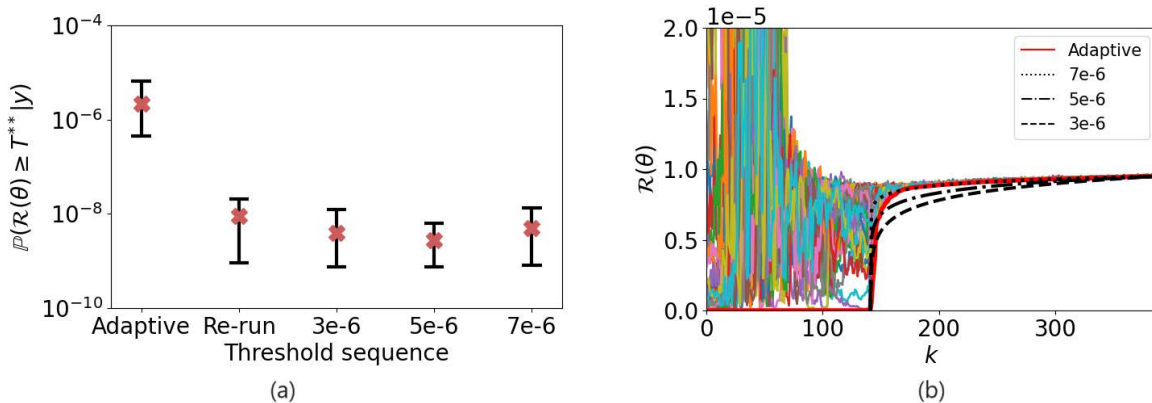
Figure 8: Illustration of the bias resulting from the adaptively determined threshold sequence within the PostRisk-SMC method for the 1D flow example: (a) Range of estimates $\mathbb{P}(\mathscr{R}(\theta) \geq T^{**}|y)$ using the different threshold sequences (ten runs each); the red crosses indicate the mean of the values and (b) evolving particle estimation of $\mathscr{R}(\theta)$ with the adaptive T_k-sequence (red) and the different fixed logarithmic sequences (black, Eq. 14 with different T_{K_P+1}).

Table 1: Table summarizing the different trials of the PostRisk-SMC and MH method applied to the 1D flow test case. The second column indicates the computational budgets used for the thresholds (in terms of the total number of forward and quantity of interest simulations); the mean and COV (coefficient of variation) are calculated based on 10 estimates of $\mathbb{P}(\mathscr{R}(\theta) \geq T|y)$ for T^* and T^{**}.

	T^*/T^{**} $[\times 10^3]$	N	$\frac{CESS^*}{N}$	s	K_R	Mean T^* $[\times 10^{-3}]$	COV T^*	Mean T^{**} $[\times 10^{-9}]$	COV T^{**}
PostRisk-SMC	40/55	20	0.9	20	100	2.28	0.71	3.88	1.72
PostRisk-SMC	400/550	200	0.9	20	100	2.45	0.27	4.81	0.35
PostRisk-SMC	510/550	20	0.9999	20	100	2.60	0.49	6.91	1.66
PostRisk-SMC	400/550	20	0.9	200	100	2.91	0.65	4.77	1.08
PostRisk-SMC	255/550	20	0.9	20	1330	2.72	0.44	4.31	0.79
MH	400/550	-	-	-	-	2.45	0.25	0	-

Table 1. While the means are comparable for all configurations, it is seen that the parameter with the most impact in reducing the COV for both thresholds is the number of particles N. In this test example, the optimal $CESS^*$ only has limited influence on the variance of the rare event estimate. Still, a high-quality representation of the posterior from the first stage leads to a smaller variance of the rare event estimate. Concerning the number of MH steps, we perform additional tests with values $s = 5, 10, 20, 200, 500$ (Fig. 9b for T^{**}). While there is high variance in the estimates for $s = 5$, the variance seems to stabilize from a value of $s = 20$ steps. Further increasing s to 200 or 500 necessitates a considerable number of additional forward operations, but leads to a much smaller improvement in the accuracy of the rare event estimate compared to increasing the number of particles. Furthermore, in the context of parallel computation, increasing the number of particles is more efficient compared to increasing s. Finally, when testing a value of K_R smaller than 100, we observed frequent failures due to the particle system dying. On the other hand, increasing the value to $K_R = 1330$ resulted in a decrease in the COV for both thresholds. Although this decrease was more significant than the effect of increasing the number of MH steps s, it still did not match the substantial improvement achieved by increasing the number of particles.

To enable comparison with a basic MH algorithm, we run 10 chains in parallel with one million iterations each. The resulting posterior median and 95% credible interval of the estimated log-diffusivity are shown in Figure 10a and the resulting samples of $\mathscr{R}(\theta)|y$ in Figure 10b. To visually compare this results with the SMC method, the credible interval in Figure 10a and the
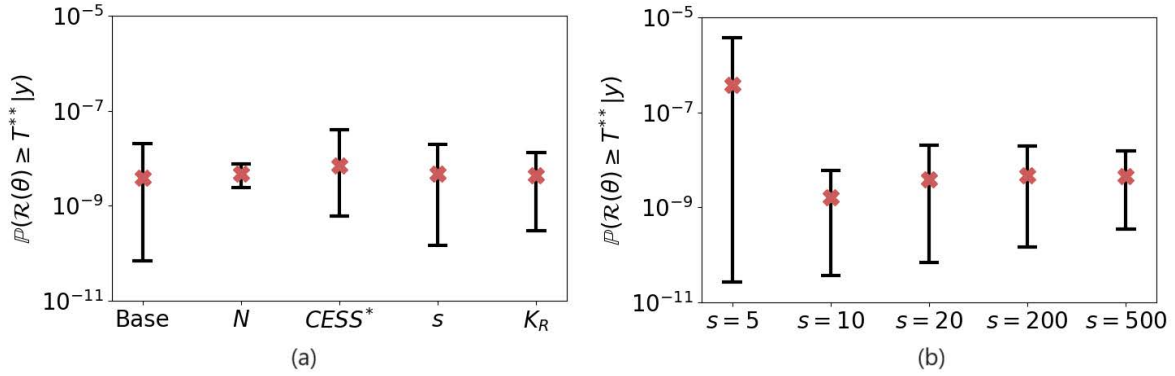
Figure 9: Impact of the configuration choices within the PostRisk-SMC method for the 1D flow example. (a) Range of the rare event probability estimates for T^{**} with the first bar corresponding to the base configuration and the following ones referring to the successive allocation of ten times more computational resources for either of the parameters with $N = 200, CESS^* = 0.9999 \times N$, $K_R = 1330$ and $s = 200$. (b) Range of the rare event probability estimates for T^{**} using different numbers of MH steps s. The red crosses in both plots indicate the mean values of the ten runs.
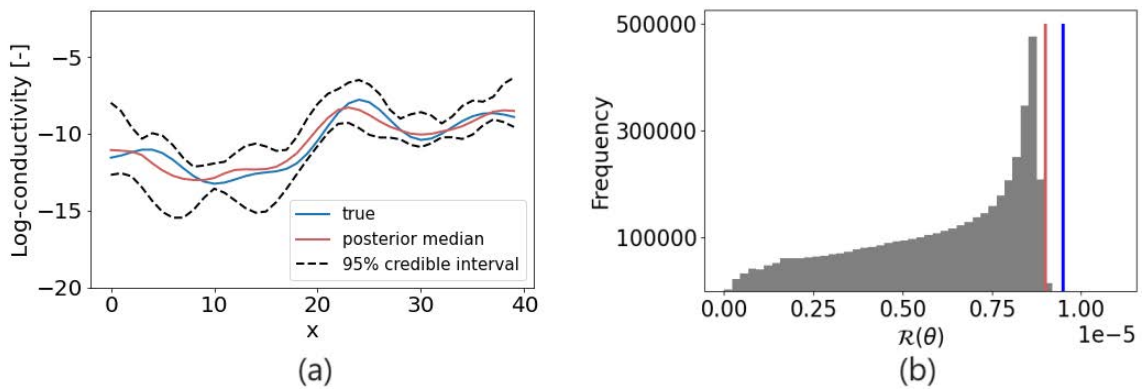

Figure 10: Results for the 1D flow example with the MH method: (a) Estimated posterior median (red) and credible interval (dashed) of the log-conductivity profile, together with the true profile (blue) and (b) transformed MH samples using $\theta \mapsto \mathscr{R}(\theta)$ with the thresholds of interest indicated (T^* in red and T^{**} in blue).

15

particle representation in Figure 7b can be considered. If we would perform MH running three chains in parallel, convergence according to the potential-scale reduction factor ($\hat{R}$-statistics using a target value of 1.2 for all parameters and the second half of the chains; Gelman and Rubin 1992) would be declared after 140'000 iterations and the resulting estimate would be 6.44×10^{-3} for T^* and zero for T^{**}. This indicates that with the computational budget of the basic version of PostRisk-SMC as shown in Table 1, we are unable to obtain any reliable estimates with MCMC using MH. With a higher budget of 400,000 for T^*, the mean of the ten estimates is 2.45×10^{-3}, and the COV is 0.25. The mean value matches the ones obtained with the PostRisk-SMC method. The comparable COV for the same computational budget of PostRisk-SMC ($N = 200$) is not surprising since the target probability enables enough samples in the MH chains. However, for T^{**}, all estimates obtained with MH are zero, even when using the full one million samples per chain.

Finally, we would like to highlight the power of including measurement data into this rare event estimation problem. As indicated in Figure 8b, for the prior distribution of the log-conductivity field ($k = 0$), $\mathscr{R}(\theta) \geq T$ is not a rare event for the considered thresholds. Therefore, we can easily estimate $\mathbb{P}(\mathscr{R}(\theta) \geq T)$ under the prior using a limited number of Monte Carlo samples, which gives us 0.23 for T^* and 0.22 for T^{**} (here employing 10,000 samples). We conclude that, compared to this previous prior probability of about one quarter, the pumping test measurements lead us to the assessment that the hazard occurrence can be specified as highly unlikely, especially for T^{**}.

4   2D flow and transport example

In the second test case, we infer a hydraulic transmissivity field θ using steady-state pressure data y from pumping tests. For the quantity of interest $\mathscr{R}(\theta)$, we consider the release of a contaminant on the left side of the model domain and observe the breakthrough of the concentration at a location on the right side of the domain. We are examining a hypothetical scenario where the contamination is expected to no longer pose a risk beyond a pre-defined time frame. That is, the hazard materializes if we observe a breakthrough at the considered location before this time has elapsed.

4.1   Problem setting

The aquifer under consideration has a size of $250 \times 250 \times 5$ m and we use a discretization on a grid with $51 \times 51 \times 1$ cells. We assume the properties to be uniform in the vertical direction, thereby simplifying the problem to two spatial dimensions. For the purpose of simulating both the data and the quantity of interest, we utilize the MODFLOW package implemented in Python, specifically the FloPy library (Bakker et al., 2016) and 'MT3D-USGS' (Bedekar et al., 2016) for the transport simulations.

We make the assumption that the system under investigation is confined. The unknown log-transmissivity field θ is assumed to be a Gaussian Random field (Chiles and Delfiner 2012). We assume a constant mean $\mu_{\log \theta} = \log(5 \times 10^{-5})$ with the transmissivity having units of m^2/s. For the isotropic covariance function, we employ an isotropic exponential covariance function in $\mathbb{R}^2$ with standard deviation $\sigma = 3$ and integral scale $l = 25$ m. In order to generate a realization of

the (51×51)-dimensional Gaussian random field, we utilize a pixel-based parameterization,

$$X = \mu_\theta + \Sigma_\theta^{1/2} Z, \tag{15}$$

where Σ_θ denotes the exponential covariance matrix and Z represents a (51×51)-dimensional random vector composed of independent and identically distributed (i.i.d.) standard normal variables. The 'true' log-transmissivity field is depicted in Figure 11a.

For the data y, we are considering a five-spot pumping test using a pumping well located in the middle of the model domain and local measurements of the log-transmissivity field at the well locations (Fig. 11b). For the pumping test, we consider a fixed hydraulic head at the left (2.5 m) and right (0 m) sides of the domain, no-flow boundaries on the other boundaries and pump with a rate of 5×10^{-4} m^3/s. For the data collection, we consider the steady-state of the system and measure the hydraulic head in four wells centered in the middle of the four quadrants of the domain. For the generation of the synthetic data, we add independent Gaussian observational errors with a standard deviation of 0.02 m. For the local measurements in the five wells, we assume a Gaussian measurement error with a standard deviation of 0.1 (log-scale). Then, we employ standard results for conditional Gaussian random fields, resulting in a mean and covariance matrix in Equation (15), which are conditioned on the local measurements and include their error.

For the rare event, we examine a scenario where a contaminant is released on the left side of the model domain, while monitoring the concentration of the contaminant on the right side. Our primary focus lies in determining the time of breakthrough $\mathscr{R}(\theta)$ in a critical area in the middle of the right side of the model domain. The hazard is specified as a breakthrough before 60 days ($T = 60$ days), with the breakthrough being specified as the concentration being higher than or equal to 1 mg/l. To simulate this, we assume a constant concentration of 1 g/l on the left side, along with a fixed hydraulic head difference of 2.5 m between the left and right sides (as for the data collection). Additionally, we maintain a constant porosity of 0.3, an effective molecular diffusion coefficient of 10^{-9} m^2/s, a longitudinal dispersivity of 1 m, a ratio of the transverse to the longitudinal dispersivity of 0.1. Figure 11c illustrates the concentration distribution after 60 days from the start of the injection for the true field, and Figure 11d visualizes the corresponding contaminant front.

4.2 Results

We first investigate the occurrence of a contamination breakthrough without incorporating the data. Given the resource-intensive nature of the transport simulations, we adhere to a computational limit of approximately 15,000 evaluations of $\mathscr{R}(\cdot)$. When using the PostRisk-SMC method for this setting, we only employ the second phase and use $N = 40$ particles and $s_R = 10$ MH steps per subset (Fig. 4). Given that we have demonstrated significant bias when considering an adaptive sequence of thresholds in the one-dimensional flow example (Fig. 8), we choose to directly employ a fixed sequence in this test case. We employ a decreasing logarithmic sequence ranging from $T_1 = 3500$ days down to 100 days, utilizing 30 steps (according to Equation 14 with $K_P = 0$). As the conditional probability during the last steps becomes lower and the risk of the particle system dying is particularly high, we adapt the sequence to steps of five days from 100 days down to the 60 days of interest, leading to $K_R = 38$. For the propagation of the particles with MH, we use pCN proposals initialized with a $\rho = 1$ (independent proposals), which is then geometrically decreased by a factor of 0.9 in each subset. In this test case, we utilize the pCN proposals as we target a parameter space characterized by high dimensionality (51×51 variables). On the other hand, in the case of the one-dimensional flow
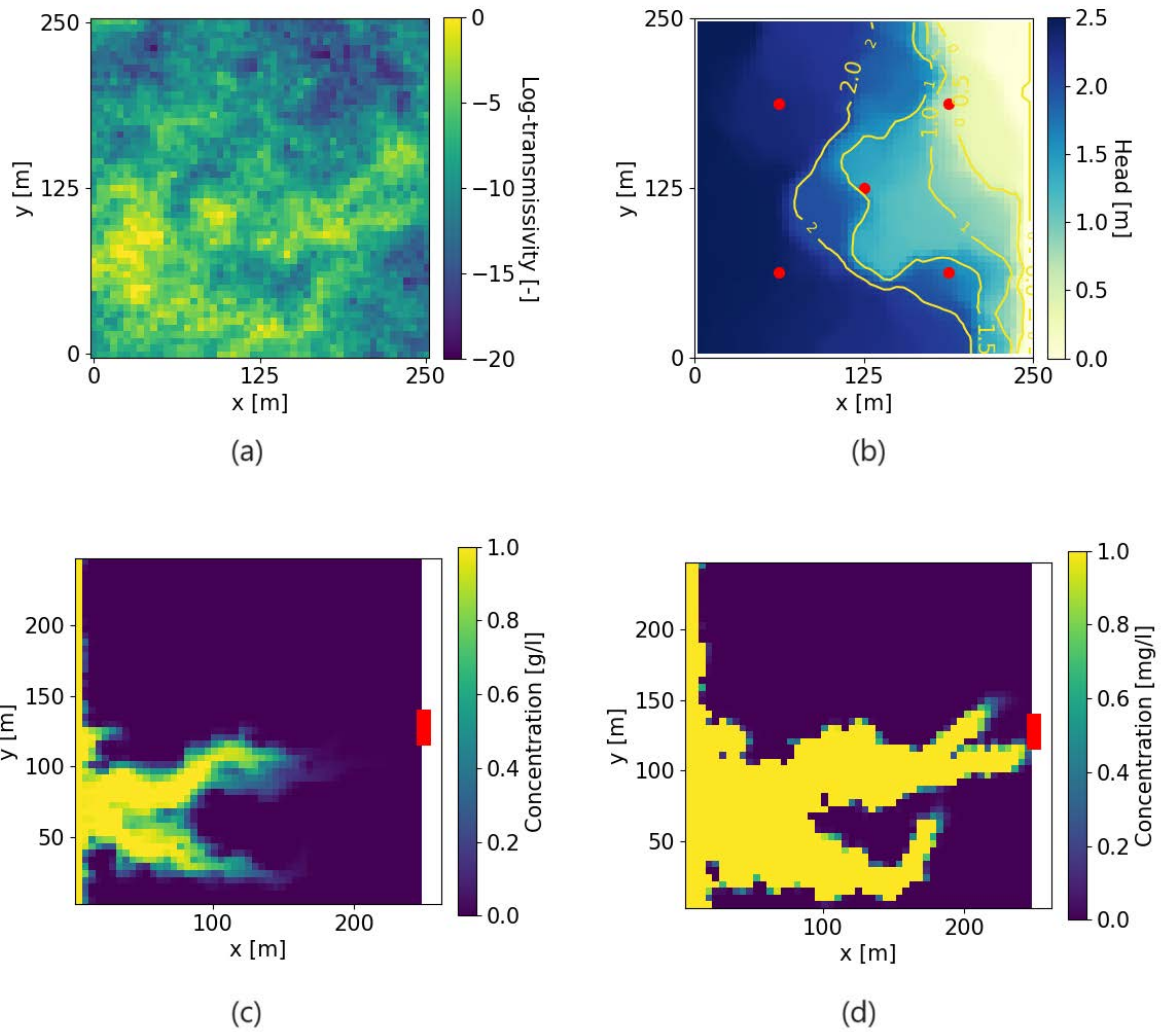
Figure 11: (a) 'True' log-hydraulic transmissivity field and corresponding (b) hydraulic heads resulting from the steady-state pumping test, the red dots indicate the well locations, (c) contamination field and (d) contaminant front after 60 days.
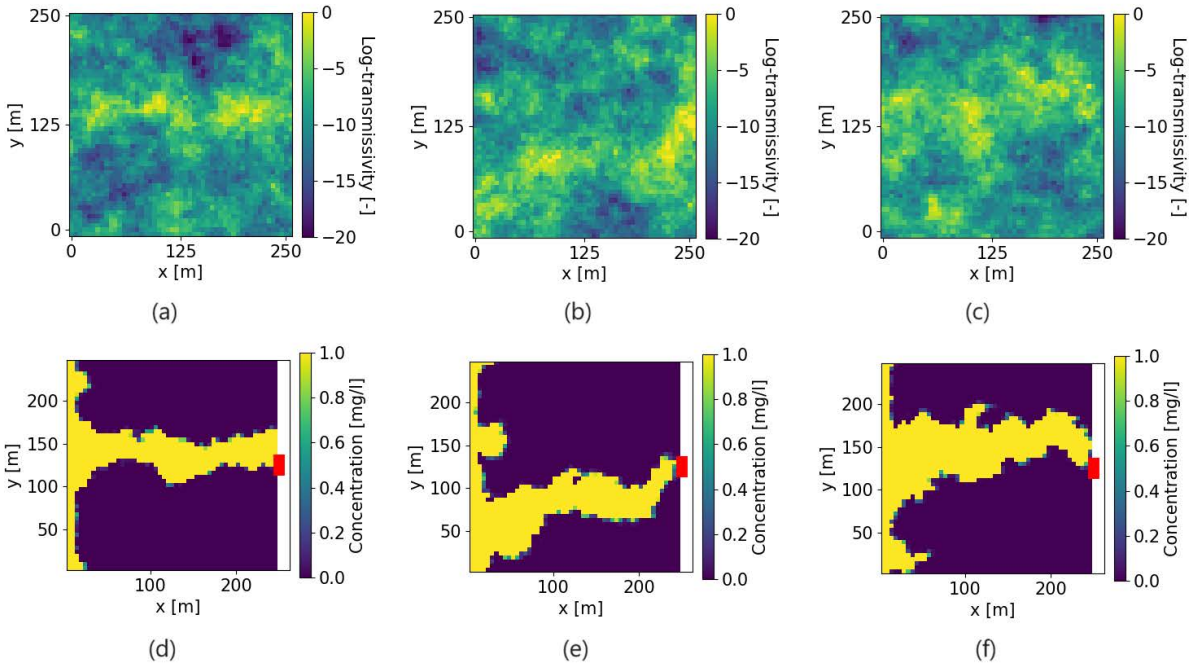
Figure 12: Rare event estimation for the 2D transport example with the PostRisk-SMC method (without inversion): (a-c) log-hydraulic transmissivity field examples from the final subset with $\mathscr{R}(\theta) \leq 60$ days and (d-f) their corresponding contaminant fronts.

example involving only 10 variables, standard Gaussian proposals proved to be effective. In Figure 12, we provide visual representations of three illustrative log-hydraulic transmissivity field realizations extracted from the final subset where $\mathscr{R}(\theta) \leq 60$ days. These examples are accompanied by their respective contamination fields. Figure 13a displays the mean transmissivity field of the particles. Running ten repetitions of the PostRisk-SMC method, we obtain a mean of 0.71×10^{-4} and a COV of 0.37 for $\mathbb{P}(\mathscr{R}(\theta) \leq 60$ d) (Table 2). With prior sampling and Monte Carlo estimation for the same computational budget, we obtain a mean of 0.87×10^{-4} and a COV of 0.60. While the Monte Carlo approach includes zero in the range of the ten probability estimates, the PostRisk-SMC method specifies the probability as being at least 0.24×10^{-4}.

We now consider the data. Figure 11d demonstrates that the hazard is occurring for the true log-hydraulic transmissivity field and we are interested to see if the integration of the local and pumping measurements helps to reflect this by increasing the rare event probability estimate. For the posterior inference part of PostRisk-SMC , we use a configuration with $N = 40$, $CESS^*/N = 0.99$ (leading to $K_P = 100$) and $s_P = 100$ MH steps per iteration (Fig. 2). A particle estimate of the posterior mean is depicted in Figure 13b. For the rare event phase of PostRisk-SMC , we implement the adaptation outlined in Section 2.3.3, wherein we conduct $ss_R = 100$ posterior steps within each of the $s_R = 10$ MH steps during the rare event phase of the algorithm. This implies that for every subset, we need to assess $\mathscr{R}(\cdot)$ ten times and $\mathscr{G}(\cdot)$ one thousand times. We use the same sequence of thresholds with $K_R = 38$ as described above. In total, this results in $N \times (K_P \times s_P + K_R \times s_R \times s_{RR}) = 1.92$ million evaluations of $\mathscr{G}(\cdot)$ and $N \times K_R \times s_R = 15,200$ evaluations of $\mathscr{R}(\cdot)$ (Table 2). For the propagation, the step size of the pCN proposals is adapted such that the 'posterior' steps have an acceptance rate of about 30 %. In Figure 14, we showcase three particles from the final posterior subset where $\mathscr{R}(\theta) \leq 60$ days, along with their corresponding contamination fields. Figure 13c shows the mean of the particles lying in the last posterior subset. Upon executing the PostRisk-SMC method ten times, we
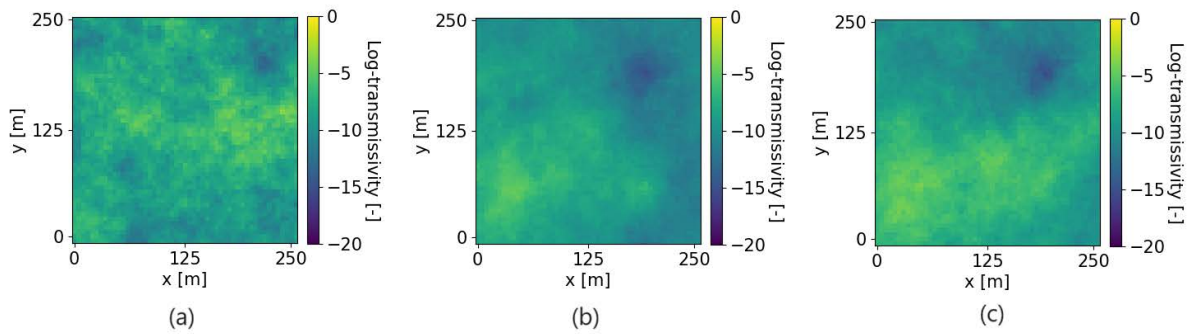
Figure 13: Results for the 2D transport example with the PostRisk-SMC method: Particle mean representing the log-hydraulic transmissivity field from the (a) prior subset where $\mathscr{R}(\boldsymbol{\theta}) \leq 60$ days, (b) posterior distribution and (c) posterior subset where $\mathscr{R}(\boldsymbol{\theta}) \leq 60$ days.

Table 2: Table summarizing the different trials of the PostRisk-SMC and MH method applied to the 2D transport test case under the prior and the posterior distribution. The second column shows the number of required simulations of the forward response $\mathscr{G}(\cdot)$ and quantity of interest $\mathscr{R}(\cdot)$ and mean, COV (coefficient of variation), min (minimum) and max (maximum) refer to the 10 estimates of the rare event probability.

	Method	$\mathscr{G}(\cdot)/\mathscr{R}(\cdot)$ $[\times 10^4]$	Mean $[\times 10^{-4}]$	COV	Min $[\times 10^{-4}]$	Max $[\times 10^{-4}]$	N	$\frac{CESS^*}{N}$	s_P	s_R	ss_R	K_R
Prior	PostRisk-SMC	- / 1.5	0.71	0.37	0.24	1.09	40	-	-	10	-	38
	Monte Carlo	- / 1.5	0.87	0.60	0	1.33	-	-	-	-	-	-
Posterior	PostRisk-SMC	192 / 1.5	4.56	0.21	3.55	6.64	40	0.99	100	10	100	38
	MH	192 / 1.5	5.64	0.49	2.01	12.75	-	-	-	-	-	-

compute an average of 4.56×10^{-4} and observe a COV of 0.21 for $\mathbb{P}(\mathscr{R}(\boldsymbol{\theta}) \leq 60$ days) (Table 2).

For a fair comparison with Monte Carlo estimation based on MCMC samples, we run ten chains with 1.92 Million steps and evaluate $\mathscr{R}(\cdot)$ for only 15,000 samples (per chain) that are obtained by thinning. We employ pCN proposals with an adjusted step size aiming for an acceptance rate of 30 %. We obtain a mean rare event probability estimate of 5.64×10^{-4} and a COV of 0.49 (Table 2). Using the first three chains, convergence with respect to the $\hat{R}$-statistics would be declared after 350,000 iterations. The corresponding merged 1,500 thinned samples per chain would specify the hazard occurrence probability as zero.

Similar to the one-dimensional flow example, we can observe that incorporating measurements leads to a shift in our estimation of the hazard occurrence probability. In the context of this two-dimensional transport example, the incorporation of local measurements and pumping data increases the estimated probability of hazard occurrence by a factor of about six compared with the estimate based on prior knowledge only. We observe that for the ten considered estimates, the range of the values for the prior and posterior can be clearly separated (for both PostRisk-SMC and Monte Carlo estimation).

5 Discussion

Sustainable groundwater management and assessment of associated hazards are pressing needs that are being accentuated under global change (e.g., Siebert et al. 2010, Famiglietti 2014, Gorelick and Zheng 2015). With the Posterior Risk Sequential Monte Carlo (PostRisk-SMC) method, we present an approach that combines Bayesian inversion and rare event probability
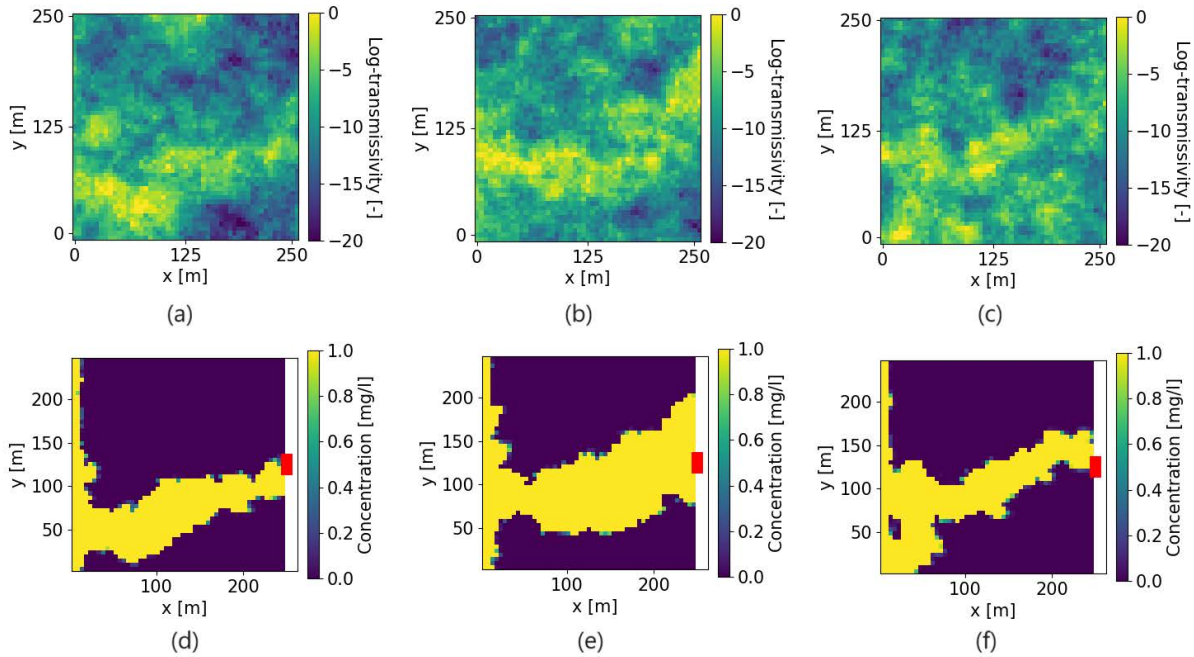
Figure 14: Rare event estimation for the 2D transport example with the PostRisk-SMC method (with inversion): (a-c) log-hydraulic transmissivity field examples from the final subset with $\mathscr{R}(\theta) \leq 60$ days and (d-f) their corresponding contaminant fronts.

estimation under uncertainty. It first generates a particle approximation of the posterior which is then propagated to provide an accurate estimation of the rare hazard probability. Thereby, the method relies on 'subset sampling' and aims to estimate a small probability as a product of larger conditional probabilities. In addition to probability estimation, the method also generates realizations of the rare event (as illustrated in Figs. 7 and 14), providing tangible representations of how the subsurface property field leading to the hazard could look like in practice. In this vein, the PostRisk-SMC approach aligns with the perspective of Ferré (2017), advocating for the communication of information to decision-makers regarding what is known, what is possible, and what remains unknown.

In the first phase of the PostRisk-SMC method, we employ adaptive SMC for Bayesian inference (Zhou et al., 2016), relying on power posteriors giving increasingly more weight to the likelihood. The adaptivity of the exponents introduces a slight bias in the results (Beskos et al., 2016), but its extent was found to be negligible in the considered test cases. This adaptive feature is attractive as it reduces the number of user-defined tuning parameters and contributes to a more efficient algorithm. The adaptively determined exponents rely on the choice of the target $CESS^*$ (Eq. 7). The closer this target value is to the number of particles N, the better the approximation, but the slower the algorithm becomes as the number of power posteriors grows. The optimal choice of the algorithmic variables $CESS^*$, s_P (number of MH steps) and N (number of particles) depends on the complexity of the posterior distribution, which is influenced by various factors such as the dimension of the parameter space and the underlying physics (Amaya et al., 2021). In their work, Amaya et al. (2021) suggest employing a combination of $CESS^*$ and s_P such that the weighted-mean likelihood of the particles is in agreement with the tempered likelihoods corresponding to the prescribed model throughout the entire run. Moreover, the authors suggest an algorithmic configuration that avoids too frequent resampling steps. To achieve this, they initially set a $CESS^*$ (for instance, $0.99N$), and subsequently fine-tune the number of MH steps (s_P) to ensure fitting the data and minimizing the need for resampling. This process can be

done employing a smaller number of particles N for preliminary runs, followed by employing a larger number of samples in the 'final' runs. While the accuracy of the approximation improves with an increase in the number of particles N, this enhancement comes at a computational cost. However, unlike many MCMC methods, the SMC method is particularly well-suited for parallel computation, as the particles can be distributed across multiple computing nodes.

In the second phase of the PostRisk-SMC method, we rely on subset sampling to estimate the rare event probabilities. The selection of intermediate thresholds involves a trade-off between the intermediate conditional probabilities and the number of particles (Au and Beck, 2001). If the threshold increases slowly, the conditional probabilities are large and a small number of particles is needed to ensure accurate estimation. On the other hand, more intermediate thresholds are needed until the target threshold is reached. If the thresholds increase faster, more particles are needed for an accurate estimation, which also increases the total number of simulations. Cérou et al. (2012) propose an adaptive sequence of thresholds based on quantiles to increase the efficiency of their algorithm. The negative aspect of introducing adaptive thresholds is a positive bias in the rare event probability estimate, which diminishes with an increasing number of particles (Cérou et al., 2012). Cérou et al. (2012) propose a correction factor for the bias, however, their analytical study assumes that the particles are independent, which is hard to guarantee in practice due to the resampling and the finite number of MH steps s_R.

In the one-dimensional flow example (Section 3), the bias resulting from the adaptive thresholds is far from negligible, especially when using a relatively small number of $N = 40$ particles and targeting a rare event with probability of one in a billion (Fig. 8). In light of this, we strongly caution against employing an adaptive scheme for the thresholds, particularly if not carefully assessing this bias by re-running with the previously defined sequence as pre-defined thresholds. To avoid bias, and the computational burden associated with re-running or increasing the number of particles, we employ a fixed sequence of thresholds (Eq. 14). When the choice of a suitable form for the fixed sequence is unclear, one option is to run an initial adaptive run that can provide valuable insights into appropriate functional forms of the sequence. Similarly, determining the number of subsets K_R can benefit from an initial adaptive run using a ratio of surviving particles guided by the literature (e.g., Cérou et al. 2012 recommend 75-80 %). A fixed sequence of thresholds leads to the possibility of the particle system "dying" during the rare event estimation process if no particles exceed the current threshold. We did not specifically consider this scenario, but one possible approach to address this issue is discussed by LeGland and Oudjane (2006). Their idea involves continuing to generate new particles until a specified count of particles has reached the given threshold.

In the context of the two-dimensional flow and transport example (Section 4), posterior exploration presents a challenge as strong non-uniqueness and underdetermination enable a wide range of solutions to accurately explain the observed data (Soueid Ahmed et al., 2014; Cotter et al., 2013). Hence, the number of resampling steps and the propagation through the MH steps play a crucial role in preventing particle collapse. This latter aspect gains even greater significance during the phase of rare event estimation, as resampling cannot be avoided. For this reason, we implement a slight adaptation of the PostRisk-SMC method outlined in Figure 5. Rather than simulating both the forward response and the quantity of interest at each iteration of the rare event phase, we first perform a sequence of ss_R posterior steps during each of the s_R MH steps. We then consider the last state as proposal from the posterior distribution and decide to accept or reject it depending on whether it lies within the current subset. In scenarios involving contamination simulations, where the computational cost of the contamination simulation typically surpasses that of the data simulation flow model, this strategy enhances particle propagation efficiency while simultaneously decreasing computational demands. We suggest

to verify that the chosen values for N, s_R and ss_R guarantee the generation of significantly new realizations through the MH steps, thereby preventing particle collapse. Visual inspection of particles at various stages of the algorithm can facilitate this assessment. Similar as for the posterior phase, elevating the number of particles N appears to be a suitable strategy for diminishing the variance of the rare event estimator. This proves advantageous, particularly considering the parallelizability of particles.

In both test examples, we investigate the significance of using the posterior instead of the prior PDF to determine the probability of hazard occurrence. In the context of the one-dimensional flow example, we showcase how the introduction of pumping test measurements in this scenario alters a rather likely event into a highly unlikely one. Indeed, the initial occurrence probability of roughly a quarter is after considering the data turned into a probability of one in a billion for T^{**}. In the case of the two-dimensional transport example, the situation is reversed: the inclusion of local measurements and pumping data helps in quantifying the probability of hazard occurrence as being six times higher than with prior knowledge alone. The integration of posterior inference serves as a clear demonstration of why it is crucial to design appropriate data acquisition strategies within the realm of risk assessment. Designing appropriate experimental designs for such tasks is a research area on its own, as exemplified by Li and Xiu (2010) and Nowak et al. (2012) for hydrological settings.

We compare the performance of the PostRisk-SMC method with a conventional Monte Carlo approach relying on prior or posterior samples obtained by the MH algorithm. In the one-dimensional flow example (Table 1), the estimates obtained with PostRisk-SMC align with those of the traditional method for the less rare event. For the more rare event with occurrence probability approaching one in a billion, the Monte Carlo approach fails in simulating the hazardous scenario. The PostRisk-SMC method, on the other hand, is able to specify the occurrence probability with a coefficient of variation of 0.35. In the two-dimensional transport example (Table 2), the PostRisk-SMC method successfully reduces the coefficient of variation by more than 50 % compared to Monte Carlo estimation based on MH samples (for the inversion setting). This comparison is established within a scenario where Monte Carlo estimation remains feasible. For rarer events, we anticipate complete failure of Monte Carlo estimation, as showcased by the one-dimensional flow example (Table 1).

It is worth noting that the two phases of the PostRisk-SMC method exhibit different dynamics. While in our 1D flow example, the adaptive procedure for the exponents defining the power posteriors leads to an exponential increase, the sequence of thresholds follows a logarithmic progression. In Section 4, we take an initial step in addressing this distinct difference in dynamics by using different numbers of MH steps for the two phases of the method. However, there is considerable potential for further exploration and refinement in this regard. So far, we only dealt with rare sets $A = \{\theta \in \mathbb{R}^P : \mathscr{R}(\theta) \in \mathscr{T}\}$ with $\mathscr{T} = [T, \infty)$ or $\mathscr{T} = (-\infty, T]$ for some real number T. If we would consider $\mathscr{T} = [T^*, T^{**}]$, one could gradually shrink the interval from both sides. Looking ahead, it could be interesting to incorporate surrogate modeling within the PostRisk-SMC method to tackle more complex and realistic problems. Surrogates (e.g. Razavi et al. 2012) in this context can serve as simplified models or approximations of the underlying system, allowing for faster evaluations and reducing the computational burden. Moreover, considering alternative approaches for the intermediate steps in both phases could be interesting, such as incorporating a method based on smoothed indicator functions and thermodynamic integration proposed by Xiao et al. (2019), in the second phase. Finally, exploring test cases that do not rely on Gaussian assumptions would be intriguing, as previously undertaken for the posterior part in Amaya et al. (2021) and Amaya et al. (2022).

6 Conclusions

The combination of Bayesian inversion and rare event estimation is very helpful for understanding groundwater hazards and their implications for humans and ecosystems. To overcome the challenges of rare event estimation in an inversion setting, we present a two-stage formulation of Sequential Monte Carlo, denoted as the PostRisk-SMC method. First, particles are generated to approximate the posterior distribution by adaptively increasing the exponent of the likelihood function. Second, subset sampling is employed to evaluate the probability of the rare event of interest. To showcase the efficacy and accuracy of the PostRisk-SMC method, we present a one-dimensional flow example and a two-dimensional flow- and transport example. The one-dimensional example demonstrates that the PostRisk-SMC method allows us to estimate rare event probabilities as low as one in a billion. In the two-dimensional example, we showcase the method's capability for rare event probability estimation in a more realistic and complex setting. In both examples, the PostRisk-SMC method successfully reduces the coefficient of variation of the rare event probability estimate compared to Monte Carlo estimation based on posterior samples. In both cases, the addition of the measurement data lead to a distinctly different assessment of the occurrence probability than relying on the prior only. Future work will also consider inclusion of surrogate modeling to speed up computations and applications to actual field settings.

Code availability

The code and test examples associated with this article are available in the following GitHub repository: https://github.com/LeaFrie/SMC_groundwater.

References

M Amaya, N Linde, and E Laloy. Adaptive sequential Monte Carlo for posterior inference and model selection among complex geological priors. *Geophysical Journal International*, 226 (2):1220–1238, 2021.

M Amaya, N Linde, and E Laloy. Hydrogeological multiple-point statistics inversion by adaptive sequential Monte Carlo. *Advances in Water Resources*, 166:104252, 2022.

Siu-Kui Au and James L Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001.

Siu-Kui Au and James L Beck. Important sampling in high dimensions. *Structural Safety*, 25 (2):139–163, 2003a.

Siu-Kui Au and James L Beck. Subset simulation and its application to seismic risk based on dynamic analysis. *Journal of Engineering Mechanics*, 129(8):901–917, 2003b.

Mark Bakker, Vincent Post, Christian D Langevin, Joseph D Hughes, Jeremy T White, JJ Starn, and Michael N Fienen. Scripting MODFLOW model development using Python and FloPy. *Groundwater*, 54(5):733–739, 2016.

Vivek Bedekar, Eric D Morway, Christian D Langevin, and Matthew J Tonkin. MT3D-USGS version 1: A US Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW. Technical report, US Geological Survey, 2016.

I D Benekos, C A Shoemaker, and J R Stedinger. Probabilistic risk and uncertainty analysis for bioremediation of four chlorinated ethenes in groundwater. *Stochastic Environmental Research and Risk Assessment*, 21(4):375–390, 2007.

A Beskos, A Jasra, N Kantas, and A Thiery. On the convergence of adaptive sequential Monte Carlo methods. *The Annals of Applied Probability*, 26(2):1111–46, 2016.

Z I Botev and D P Kroese. An efficient algorithm for rare-event probability estimation, combinatorial optimization, and counting. *Methodology and Computing in Applied Probability*, 10 (4):471–505, 2008.

J-M Bourinet, François Deheeger, and Maurice Lemaire. Assessing small failure probabilities by combined subset simulation and support vector machines. *Structural Safety*, 33(6):343–353, 2011.

E Brodin and C Klüppelberg. Extreme value theory in finance. In B. Everitt and E. Melnick, editors, *Encyclopedia of Quantitative Risk Analysis and Assessment*. Wiley, Chichester, 2008.

F Cadini, D Avram, N Pedroni, and E Zio. Subset simulation of a reliability model for radioactive waste repository performance assessment. *Reliability Engineering & System Safety*, 100: 75–83, 2012.

F Cérou, P Del Moral, T Furon, and A Guyader. Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, 22(3):795–808, 2012.

Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*, volume 713. John Wiley & Sons, 2012.

Jianye Ching and Yi-Chu Chen. Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. *Journal of Engineering Mechanics*, 133(7):816–832, 2007.

Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424 – 446, 2013.

Valentin Dall'Alba, Alexis Neven, Rob de Rooij, Marco Filipponi, and Philippe Renard. Probabilistic estimation of tunnel inflow from a karstic conduit network. *Engineering Geology*, 312:106950, 2023.

L Davies, AY Ley-Cooper, M Sutton, and C Drovandi. Bayesian detectability of induced polarisation in airborne electromagnetic data. *Geophysical Journal International*, 2023. ggad073.

P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

P Del Moral, A Doucet, and A Jasra. Sequential Monte Carlo for Bayesian computation. *Bayesian Statistics*, 8(1):34, 2007.

Francis X Diebold, Til Schuermann, and John D Stroughair. Pitfalls and opportunities in the use of extreme value theory in risk management. *The Journal of Risk Finance*, 1(2):30–35, 2000.

Randal Douc and Olivier Cappé. Comparison of resampling schemes for particle filtering. In *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, ISPA*, pages 64–69, 2005.

A Doucet and A M Johansen. A tutorial on particle filtering and smoothing: Fifteen years later. In Dan Crisan and Boris Rozovskii, editors, *The Oxford Handbook of Nonlinear Filtering*, volume 12, pages 656–704. Oxford University Press, New York, 2009.

A Doucet, N De Freitas, and N Gordon. An introduction to sequential Monte Carlo methods. In Arnaud Doucet, N. de Freitas, and N. Gordon, editors, *Sequential Monte Carlo methods in practice*, pages 3–14. Springer, New York, 2001.

D J Earl and M W Deem. Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 7(23):3910–3916, 2005.

R Enzenhoefer, W Nowak, and R Helmig. Probabilistic exposure risk assessment with advective–dispersive well vulnerability criteria. *Advances in Water Resources*, 36:121–132, 2012.

J S Famiglietti. The global groundwater crisis. *Nature Climate Change*, 4(11):945–948, 2014.

Ty PA Ferré. Revisiting the relationship between data, models, and decision-making. *Groundwater*, 55(5):604–614, 2017.

Andrew Gelman and Donald B Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

S M Gorelick and C Zheng. Global change and the groundwater management challenge. *Water Resources Research*, 51(5):3031–3051, 2015.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

Thomas Hermans, Erasmus Oware, and Jef Caers. Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data. *Water Resources Research*, 52(9):7262–7283, 2016.

HA Jensen, C Vergara, C Papadimitriou, and E Millas. The use of updated robust reliability measures in stochastic dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 267:293–317, 2013.

A M Johansen, P Del Moral, and A Doucet. Sequential Monte Carlo samplers for rare events. In *6th International Workshop on Rare Event Simulation*, pages 256–267, 2006.

I M Khadam and J J Kaluarachchi. Multi-criteria decision analysis with probabilistic risk assessment for the management of contaminated ground water. *Environmental Impact Assessment Review*, 23(6):683–721, 2003.

A Kong, J S Liu, and W H Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994.

M B Lahkim and L A Garcia. Stochastic modeling of exposure and risk in a contaminated heterogeneous aquifer. 1: Monte Carlo uncertainty analysis. *Environmental Engineering Science*, 16(5):315–328, 1999.

Hans Petter Langtangen and Svein Linge. Diffusion equations. In *Finite Difference Computing with PDEs: A Modern Software Approach*, pages 207–322. Springer, Cham, 2017.

Malcolm R Leadbetter. On a basis for 'peaks over threshold' modeling. *Statistics & Probability Letters*, 12(4):357–362, 1991.

François LeGland and Nadia Oudjane. A sequential particle algorithm that keeps the particle system alive. In *Stochastic Hybrid Systems: Theory and Safety Critical Applications*, pages 351–389. Springer, Berlin Heidelberg, 2006.

Jing Li and Dongbin Xiu. Evaluation of failure probability via surrogate models. *Journal of Computational Physics*, 229(23):8966–8980, 2010.

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

Michel Loève. *Elementary Probability Theory*. Springer, 1977.

Youssef M Marzouk and Habib N Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228 (6):1862–1902, 2009.

N Metropolis, A W Rosenbluth, M N Rosenbluth, A H Teller, and E Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

J E Morrison and J A Smith. Stochastic modeling of flood peaks using the generalized extreme value distribution. *Water Resources Research*, 38(12), 2002. WR000502.

Klaus Mosegaard and Albert Tarantola. Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research: Solid Earth*, 100(B7):12431–12447, 1995.

R M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Wolfgang Nowak, Yoram Rubin, and Felipe PJ de Barros. A hypothesis-driven approach to optimize field campaigns. *Water Resources Research*, 48(6), 2012.

A Owen and Y Zhou. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Costas Papadimitriou, James L Beck, and Lambros S Katafygiotis. Updating robust reliability using structural test data. *Probabilistic Engineering Mechanics*, 16(2):103–113, 2001.

Saman Razavi, Bryan A Tolson, and Donald H Burn. Review of surrogate modeling in water resources. *Water Resources Research*, 48(7), 2012. WR011527.

Brian D Ripley. *Stochastic Simulation*. John Wiley & Sons, 2009.

Christian P Robert, Víctor Elvira, Nick Tawn, and Changye Wu. Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(5):e1435, 2018.

Malcolm Sambridge. A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, 196(1):357–374, 2014.

S Siebert, J Burke, J Faures, K Frenken, J Hoogeveen, P Döll, and F T Portmann. Groundwater use for irrigation–a global inventory. *Hydrology and Earth System Sciences*, 14(10):1863–1880, 2010.

E R Siirila, A K Navarre-Sitchler, R M Maxwell, and J E McCray. A quantitative methodology to assess the risks to human health from CO2 leakage into groundwater. *Advances in Water Resources*, 36:146–164, 2012.

A Soueid Ahmed, Abderrahim Jardani, A Revil, and Jean-Paul Dupont. Hydraulic conductivity field characterization from the joint inversion of hydraulic heads and self-potential data. *Water Resources Research*, 50(4):3502–3522, 2014.

Daniel Straub. Reliability updating with equality information. *Probabilistic Engineering Mechanics*, 26(2):254–258, 2011.

Daniel Straub and Iason Papaioannou. Bayesian updating with structural reliability methods. *Journal of Engineering Mechanics*, 141(3):04014134, 2015.

Daniel Straub, Iason Papaioannou, and Wolfgang Betz. Bayesian analysis of rare events. *Journal of Computational Physics*, 314:538–556, 2016.

Robin Thibaut, Eric Laloy, and Thomas Hermans. A new framework for experimental design using bayesian evidential learning: The case of wellhead protection area. *Journal of Hydrology*, 603:126903, 2021.

Elisabeth Ullmann and Iason Papaioannou. Multilevel estimation of rare events. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):922–953, 2015.

Jasper A Vrugt, Cajo JF ter Braak, Cees GH Diks, and Gerrit Schoups. Hydrologic data assimilation using particle markov chain monte carlo simulation: Theory, concepts and applications. *Advances in Water Resources*, 51:457–478, 2013.

Philip J Ward, Veit Blauhut, Nadia Bloemendaal, James E Daniell, Marleen C de Ruiter, Melanie J Duncan, Robert Emberson, Susanna F Jenkins, Dalia Kirschbaum, Michael Kunz, et al. Natural hazard risk assessments at the global scale. *Natural Hazards and Earth System Sciences*, 20(4):1069–1096, 2020.

Sinan Xiao, Sebastian Reuschen, Gözde Köse, Sergey Oladyshkin, and Wolfgang Nowak. Estimation of small failure probabilities based on thermodynamic integration and parallel tempering. *Mechanical Systems and Signal Processing*, 133:106248, 2019.

Teng Xu, Sebastian Reuschen, Wolfgang Nowak, and Harrie-Jan Hendricks Franssen. Preconditioned Crank-Nicolson Markov chain Monte Carlo coupled with parallel tempering: An efficient method for Bayesian inversion of multi-Gaussian log-hydraulic conductivity fields. *Water Resources Research*, 56(8):e2020WR027110, 2020.

Y Zhou, A M Johansen, and J A D Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3): 701–726, 2016.