Stochastic Environmental Research and Risk Assessment (2021) 35:1039–1049
https://doi.org/10.1007/s00477-020-01943-5

ORIGINAL PAPER

Improving the computational efficiency of first arrival time uncertainty estimation using a connectivity-based ranking Monte Carlo method

Maria Morvillo[1] · Alessandra Bonazzi[1] · Calogero B. Rizzo[1] · Felipe P. J. de Barros[1]

Accepted: 19 November 2020 / Published online: 7 January 2021

Abstract
The first arrival time of a solute plume migrating from its source to an environmentally sensitive target is one of the key quantities of interest when assessing the risks of groundwater contamination. First arrival times are correlated with the hydraulic connectivity properties of spatially heterogeneous porous formations. Hydraulic connectivity leads to the presence of preferential flow paths which in turn control the transport dynamics of the leading edge of the solute plume and therefore first arrival times. In applications, these arrival times are subject to uncertainty given the lack of a detailed site characterization. The Monte Carlo method is commonly adopted to estimate the uncertainty of solute arrival times, however it leads to a computational burden. In this work, we build upon the existing knowledge regarding the correlation between connectivity and first arrival times to propose an innovative connectivity-based ranking Monte Carlo approach to quantify the uncertainty of first arrival times. The proposed method is tailored to predict first arrival times and allows to alleviate the computational costs when compared to the traditional Monte Carlo method. Our method consists of ranking the randomly generated spatially heterogeneous hydraulic conductivity fields according to their connectivity. The connectivity metric adopted is based on the concept of the minimum hydraulic resistance and can be obtained at a very low computational cost through the use of graph theory. We illustrate the methodology by analyzing the convergence rate of the first arrival time means and standard deviations. We compare the convergence rate of the first arrival times statistics obtained through the proposed methodology with those computed through the traditional Monte Carlo method. Overall, our results indicate the that the proposed methodology ensures a faster convergence of the considered quantities, thus reducing the time required for their estimation and the associated computational burden.

Keywords Stochastic hydrogeology · First arrival time · Uncertainty quantification · Connectivity · Contaminant transport · Graph theory

1 Introduction

Computing the uncertainty in first arrival times of a solute body at an environmentally sensitive location is a key component in subsurface risk analysis. The hydraulic properties characterizing the subsurface, such as the hydraulic conductivity K, are spatially heterogeneous over a multitude of length scales (Rubin 2003). The heterogeneous structure of the subsurface environment leads to the presence of preferential flow paths that can significantly impact the spatiotemporal dynamics of transport (Geng and Michael 2020). These preferential flow paths are an outcome of well-connected highly permeable porous materials which in turn control early solute breakthrough at a given location downstream from the source zone (Bianchi et al. 2011; Fuks et al. 2019). These first arrival times (or first passage times) are subject to significant amount of uncertainty given the limited resources for site characterization.

The impact of hydrogeological heterogeneity on solute arrival times has been topic of study in the past (Shapiro and Cvetkovic 1988; Rubin and Dagan 1992; Bellin et al. 1992; Riva et al. 2006; Gotovac et al. 2009; Fiori et al. 2011; Bianchi and Pedretti 2017). Highly connected channels in the spatially heterogeneous K field have been

✉ Felipe P. J. de Barros
fbarros@usc.edu

[1] Sonny Astani Department of Civil and Environmental Engineering, University of Southern California, 3620 S. Vermont Avenue, Los Angeles, CA 90089-2531, USA

shown as key factors in controlling contaminant early arrival times (Deutsch 1998; Trinchero et al. 2008; Tyukhova and Willmann 2016; Henri et al. 2015; Rizzo and de Barros 2017). Sahimi et al. (1983) showed how first arrival times in a disordered porous media in a two-phase flow were affected by the degree of saturation. Bianchi et al. (2011) investigated the geological connectivity at the Macrodispersion Experiment (MADE) site (Mississippi, USA) and how it impacted the asymmetry of the solute breakthrough curves. Harvey and Gorelick (1995) showed how the temporal moments of the concentration breakthrough curve were affected by heterogeneity. The importance of solute arrival times on the estimation of human health risk has also been reported in the literature (Andričević and Cvetković 1996; Maxwell and Kastenberg 1999; Henri et al. 2016; de Barros et al. 2016; Jabbari et al. 2017). For example, Henri et al. (2015, 2016) showed how preferential flow paths could be beneficial or detrimental to human health risks due to the presence of chlorinated solvents in groundwater. The impact of both permeability and porosity spatial heterogeneities on first and late arrival times was also topic of investigation (Libera et al. 2019). Within the context of groundwater management of non-point sources, it was shown that the homogenization of the hydraulic conductivity field heterogeneity affected the uncertainties on travel times (Henri and Harter 2019; Henri et al. 2020). The study of Andričević and Cvetković (1996) illustrates how travel times can be used as indicators to quantify how geologic heterogeneity influences the amount of released contaminant mass that crosses a control plane.

Several metrics have been employed to quantify connectivity in porous media (Sánchez-Vila et al. 1996; Knudby and Carrera 2005; Le Goc et al. 2010; Fiori and Jankovic 2012; Renard and Allard 2013; Gershenzon et al. 2015; Jimenez-Martinez and Negre 2017). Measures of connectivity fall within two categories: static and dynamic (Knudby and Carrera 2005). Static connectivity indicators are based on the K field whereas dynamic connectivity indicators use quantities based on flow and/or transport features. Geological objects have been defined by Deutsch (1998) in order to identify the preferential paths in a heterogeneous field. An alternative connectivity indicator was proposed by Trinchero et al. (2008) which is based on the hydraulic response time in a pumping test. Savoy et al. (2017) investigated the impact of different geological conceptualizations on both connectivity and early arrival times. Rizzo and de Barros (2017) used graph theory [i.e. using the Dijkstra's algorithm (Dijkstra et al. 1959)] to compute the least resistance path and used the minimum hydraulic resistance (Tyukhova et al. 2015; Tyukhova and Willmann 2016) as a connectivity metric. The approach proposed by Rizzo and de Barros (2017, 2019) is computationally efficient and was tested in both two- and three-dimensional heterogeneous porous media for generic source-to-receptor conditions (i.e. hydraulic connectivity between point source to control plane, point source to point receptor, etc). Knudby and Carrera (2006) also employed the Dijkstra's algorithm to study connectivity in a two-dimensional aquifer. The minimum hydraulic resistance has been shown to be an indicator of the first solute breakthrough times (Tyukhova and Willmann 2016; Tyukhova et al. 2015; Rizzo and de Barros 2017). For example, Rizzo and de Barros (2017) showed that the minimum hydraulic resistance is a lower bound estimation of the first arrival times. However, due to challenges in site characterization, the minimum hydraulic resistance (and therefore first arrival times) are subject to uncertainty. As highlighted in Chapter 9 of Rubin (2003), early arrival times are subject to the largest uncertainty (see also Zimmerman et al. 1998). To address this challenge, the graph theory-based minimum hydraulic resistance introduced in Rizzo and de Barros (2017) has been recently used to estimate the uncertainty of the minimum hydraulic resistance and improve data acquisition campaigns that aim to reduce the uncertainty in first arrival times (Rizzo and de Barros 2019).

There are several approaches that aim to quantify uncertainty of model predictions in the subsurface environment. Existing approaches consist of perturbation methods, Monte Carlo simulations and polynomial chaos expansion [for a review of the use of these methods in subsurface hydrology, see Zhang et al. (2010)]. Although conceptually straightforward, the Monte Carlo approach is computationally demanding since the statistical accuracy of its predictions depends on the number of realizations used. Given its conceptually straightforward approach, the Monte Carlo method is considered the standard for estimating uncertainty. However it suffers from slow convergence and the statistical accuracy of a given model prediction depends on the number of K field realizations utilized (Loll and Moldrup 1998; Ballio and Guadagnini 2004) and the degree of heterogeneity of the subsurface environment (Leube et al. 2013; Moslehi et al. 2015). As a consequence, full-blown stochastic modeling based on exhaustive Monte Carlo leads to a significant computational burden. Works have reported methodologies to reduce the computational time associated with the uncertainty estimation of solute arrival times. Berrone et al. (2020) proposed to combine a multilevel Monte Carlo (MLMC) and a graph based primary subnetwork identification algorithm (see Hyman et al. 2017) to estimate the mean and variance of first passage times (i.e. first arrival times) in fractured media. Gotovac et al. (2020) employed a maximum entropy algorithm based on Fup basis functions within a Monte Carlo framework to characterize the uncertainty in solute travel times.

In this paper, we propose a hydraulic connectivity-based methodology to speed up Monte Carlo simulations that aim to predict the uncertainty of first arrival times of a solute plume in a spatially heterogeneous porous media. To achieve our goals, our methodology combines elements from the graph theory-based connectivity concept proposed in Rizzo and de Barros (2017) with the connectivity ranking concepts introduced in Deutsch (1998) in order to accelerate the Monte Carlo convergence of the statistical moments of the first arrival times.

2 Minimum hydraulic resistance

We are interested in computing the first arrival times of a solute released in a spatially heterogeneous flow through a porous medium. Heterogeneity stems from the spatially variable, locally isotropic, hydraulic conductivity field $K(\mathbf{x})$ where $\mathbf{x} = (x_1, \ldots, x_d)$ represents the Cartesian coordinate system and d is the dimensionality of the flow domain. The first arrival times are associated with solute particles that travel through well-connected zones of high K values. In order to identify these well connected zones, we will use the concept of the *Minimum Hydraulic Resistance* (MHR) (Tyukhova and Willmann 2016; Rizzo and de Barros 2017). The MHR is a static connectivity measure (Tyukhova and Willmann 2016) that relies solely on the K field and has units of time. The hydraulic resistance is an indicator of the resistance found by a solute particle traveling along a path denoted by Γ. If the path Γ follows a segment of well-connected high conductivity zones, it will likely be associated with a low hydraulic resistance. That implies that the time needed for a solute particle (following the path Γ) to reach a given receptor will be low.

Let us consider a solute source zone characterized by a volume $\mathcal{V}_S$ and a receptor (i.e. an environmentally sensitive target) of volume $\mathcal{V}_T$. The MHR between $\mathcal{V}_S$ and $\mathcal{V}_T$ is mathematically expressed as

$$\mathcal{R}_m = \min_{\Gamma \in \mathcal{P}_{\mathcal{V}_S}^{\mathcal{V}_T}} \int_\Gamma \frac{1}{K} \mathrm{d}\gamma, \tag{1}$$

where $\mathcal{P}_{\mathcal{V}_S}^{\mathcal{V}_T}$ denotes all the existing paths from every point within $\mathcal{V}_S$ to every point within $\mathcal{V}_T$. The path $\hat{\Gamma}$, through which the hydraulic resistance is minimized, is defined as the *Least Resistance Path* (LRP) and its hydraulic resistance value is equal to the MHR $\mathcal{R}_m$ [see Rizzo and de Barros (2017) for further details]. The MHR $\mathcal{R}_m$ in (1) is equivalent to the shortest path problem and can be computed through a graph theory framework and solved with a variation of the Dijkstra's algorithm (Rizzo and de Barros 2017). Details regarding the algorithm and the procedure to transform the K field into a graph can be found in Rizzo and de Barros (2017, 2019). The tool denoted as "Lazy Mole" was used to compute the $\mathcal{R}_m$ between two given locations in a heterogeneous K field and is shared openly on GitHub (https://github.com/GerryR/lazymole) (Rizzo and de Barros 2017).

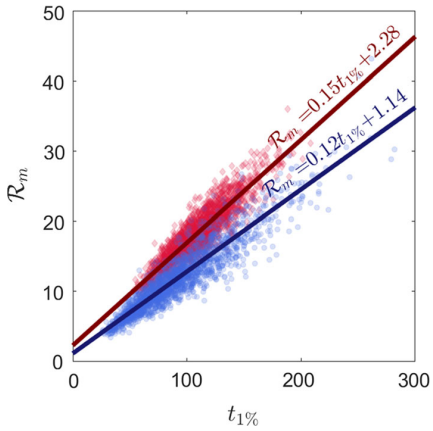

Fig. 1 Illustration of the correlation between first arrival times $t_{1\%}$ and the minimum hydraulic resistance $\mathcal{R}_m$. An ensemble of 2047 spatially random two-dimensional hydraulic conductivity fields were used for each level of heterogeneity. Heterogeneity is measured through the log-conductivity variance σ_Y^2. Results are obtained for $\sigma_Y^2 = 3$ (blue) and $\sigma_Y^2 = 1$ (red), using the parameters reported in Table 1

As initially showed in Tyukhova et al. (2015) and Tyukhova and Willmann (2016), there is a correlation between the $\mathcal{R}_m$ and the solute first arrival time. As depicted in Fig. 4 of Tyukhova and Willmann (2016) and Fig. 7 of Rizzo and de Barros (2017), larger values of $\mathcal{R}_m$ correspond to larger solute first arrival times. For the sake of completeness, this correlation is plotted in Fig. 1. Figure 1 plots $\mathcal{R}_m$ versus the first arrival times (i.e. the time in which 1% of the solute mass arrives at a control plane) for two levels of heterogeneity in the conductivity fields. For every K field, there is one value of $\mathcal{R}_m$ and a corresponding first arrival time computed through a flow and transport simulator.

3 Methodology

3.1 Generating a connectivity-based ranked hydraulic conductivity ensemble

As mentioned in Rizzo and de Barros (2019), one of the key advantages of using a static connectivity metric such as $\mathcal{R}_m$ is that important features of the hydrogeological system can be extracted without resorting to flow and transport simulations and therefore, without much computational costs. This is a relevant advantage when dealing with stochastic systems that rely on the use of a Monte Carlo

framework (Ballio and Guadagnini 2004). The computational costs associated with Monte Carlo simulations are particularly high when the quantity of interest is the first arrival time which is subject to largest uncertainty (Rubin 2003).

Given the positive correlation between $\mathcal{R}_m$ and first arrival times (see Fig. 1), in the present study we propose a methodology which consists of ranking the hydraulic conductivity fields based on their $\mathcal{R}_m$ values with the goal of speeding up the statistical convergence of the mean and variance of the first arrival times ($t_{1\%}$), defined as the time required to the first 1% of contaminant mass to reach a given location downstream from the source zone. Deutsch (1998) introduced the idea of combining the usage of static quantities (related to connectivity) and ranking process to estimate specific dynamic field characteristics. The idea consisted of the following steps: (1) defining an indicator variable that depends on several geological parameters such as lithofacies, porosity and permeability in order to define which geo-objects within the domain were well-connected; and (2) ranking the given geostatistical realizations according to their connectivity level.

In our work, we propose to employ Deutsh's ranking process (Deutsch 1998) in order to improve Monte Carlo convergence of the statistical moments of $t_{1\%}$. Our methodology is based on the following steps:

1. *Hydraulic conductivity field generation* We start by randomly generating a large ensemble of size N of log-conductivity fields, i.e. $Y = \log(K)$. These log-conductivity fields can be generated using any random space function model and has no restrictions regarding its statistical nature, i.e. multi-Gaussian or non multi-Gaussian.
2. $\mathcal{R}_m$ *computation* Once the solute source $\mathcal{V}_S$ and the target $\mathcal{V}_T$ zones are defined (which are identical for all the N realizations of the conductivity field), the "Lazy Mole" tool (Rizzo and de Barros 2017) is employed to evaluate the $\mathcal{R}_m$ value between those two locations for each generated conductivity field.
3. *Ranking* Based on the previously computed values of $\mathcal{R}_m$ (see item 2), the set of N conductivity fields is reordered according to the following procedure:
   (i) sort, in descending or ascending order, the N conductivity fields according to their minimum hydraulic resistance $\mathcal{R}_m$;
   (ii) rearrange the N conductivity fields into a *balanced binary search tree* (i.e., for each node of the binary search tree, the height of the left and right sub-trees differ by at most 1 for each node) (Booth and Colin 1960);
   (iii) reorder the fields by traversing the *balanced binary search tree* using a Breath-First Search (BFS).
4. *Subsets generation* This step consists of generating subsets from the ranked hydraulic conductivity field ensemble. From the ranked set of N hydraulic conductivity fields, α subsets are generated, where $\alpha = \log_2(N+1)$; each of the α subsets includes the first $2^n - 1$ conductivity fields, with $n = 1, \ldots, \alpha$.
5. *Flow and transport simulations* For each generated hydraulic conductivity field, groundwater flow and solute transport are simulated. For each K field, we compute the first arrival time $t_{1\%}$.
6. *Evaluation of the first arrival time statistics* We compute the mean and variance of the first arrival times for every given subset of the hydraulic conductivity field ensemble. The convergence is analyzed by utilizing the α subsets generated from the ranked set of N hydraulic conductivity fields.

To test the performance of the methodology, we will compare the convergence rate of the mean and variance of $t_{1\%}$ between the ranked approach described above and the traditional Monte Carlo method without ranking (i.e. running flow and transport simulations on the non-ranked K fields, see item 1 of the list above). Additional details regarding the ranking and the creation of the subsets are provided in Sect. 3.2.

3.2 Implementation

The goal of our ranking procedure is to obtain subsets of K fields that will lead to a quicker convergence of the statistical moments of $t_{1\%}$. In order to achieve this goal, a ranking procedure based on the $\mathcal{R}_m$ value (see Eq. 1) of each K field is applied such that the subsets are balanced. A balanced subset is defined as the subset where, to the number of K fields with an higher value of $\mathcal{R}_m$, corresponds an equal number of counterparts of K fields with lower $\mathcal{R}_m$ values.

To illustrate the implementation of the methodology, Fig. 2 shows the procedure to rank $N = 7$ hydraulic conductivity fields following item 3 listed in Sect. 3.1. As shown in Fig. 2a, initially the $\mathcal{R}_m$ values of the randomly generated K fields do not follow any specific order. The first step of our ranking procedure is represented in Fig. 2b, where the hydraulic conductivity fields are sorted in increasing order of $\mathcal{R}_m$ values; after that, a *balanced binary search tree* is constructed from the fields in Fig. 2b starting from the center and continuing with the center of the left and right subsets, creating the tree structure reported in Fig. 2c. Each subset is finally created by searching the tree with the BFS algorithm (red dotted line in Fig. 2c), as

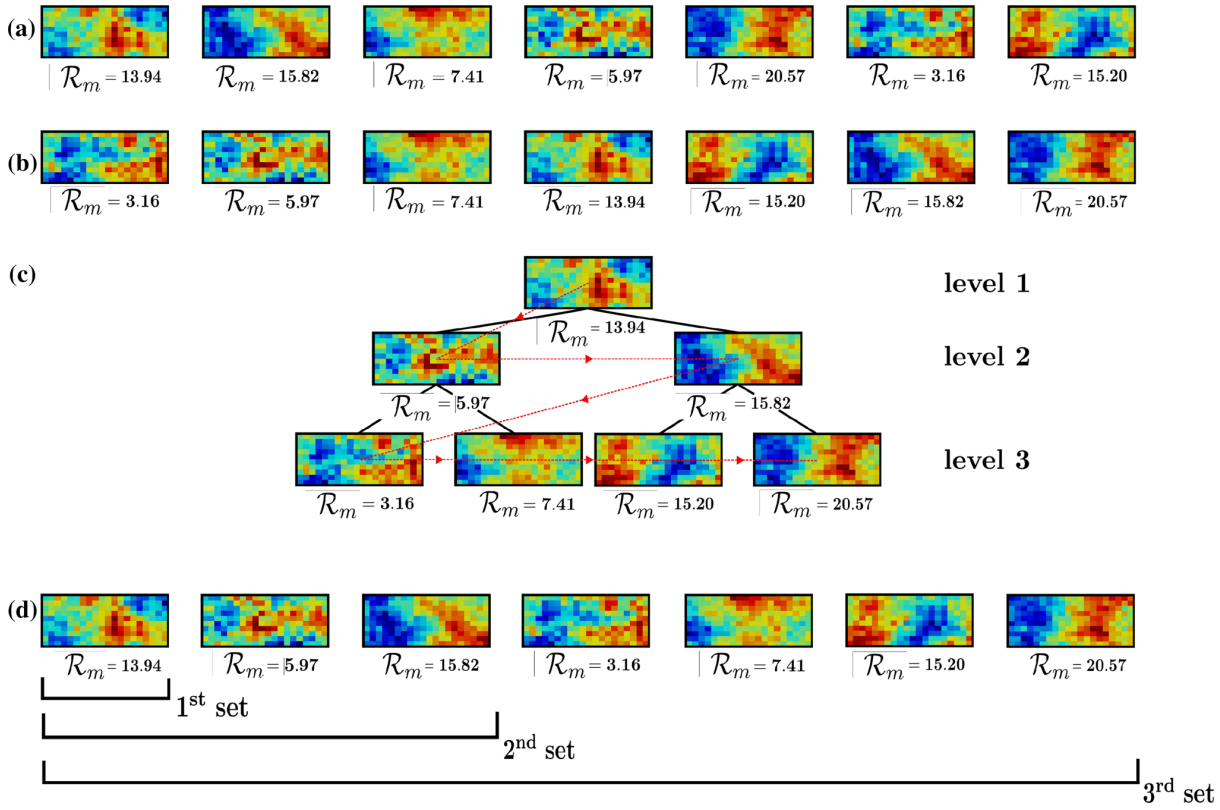

Fig. 2 Example of the connectivity-based ranked distribution methodology for $N = 7$. **a** The randomly generated K fields, each one of them characterized by a value of minimum hydraulic resistivity $\mathcal{R}_m$. **b** The K fields are reordered for increasing $\mathcal{R}_m$ values. **c** A *balanced binary search tree* is built. **d** The subsets are created searching the tree through BFS algorithm that screens the tree following the directions indicated by the red dotted line in **c**. The three generated subsets are highlighted by the horizontal square brackets in **d**

highlighted by the horizontal square brackets displayed in Fig. 2d. If in the *tree sort* algorithm (Knuth 1968) the elements of the generated binary search tree are sorted following the in-order traversal, then by adopting the BFS, the sorting process (1) starts from the root of the tree (see level 1 in Fig. 2), (2) then moves to the next tree level (see level 2 in Fig. 2) and (3) proceeds as indicated by the red dotted line in Fig. 2d.

To better grasp the rational behind the proposed connectivity-based ranking methodology, we can consider that the first subset contains one single conductivity field, the one whose value of $\mathcal{R}_m$ is the median of the whole distribution shown in Fig. 2b. Based on the relationship between $\mathcal{R}_m$ and $t_{1\%}$ (see Fig. 1), we can assume that to this field corresponds a value of $t_{1\%}$ close to the mean value of $t_{1\%}$ of the K ensemble of size N. To build the second subset, we consider the remaining fields as partitioned in two groups: those with $\mathcal{R}_m$ values lower than the median value (the one associated with the field that constitutes the first subset), and those with higher $\mathcal{R}_m$ values; in each of these two groups, the hydraulic conductivity field with the median value of $\mathcal{R}_m$ is added to the first subset to generate the second one. In general, we can assume that the β fields contained in a subset divide the complete set of fields (in increasing order of $\mathcal{R}_m$) in $\beta + 1$ groups. The field corresponding to the median value of $\mathcal{R}_m$ of each one of these groups is added to the existing subset to create the subsequent one.

4 Illustration

4.1 Physical set-up and input data

For the upcoming illustrations, we will test our methodology on two- and three-dimensional (2D and 3D) random K fields. The porous domain has dimensions ℓ_i along the ith direction where $i = 1, \ldots, d$. The log-conductivity field $Y = \log K$ is modelled as a multi-Gaussian process characterized by its mean value μ_Y, variance σ_Y^2, correlation length along the ith direction λ_i and an exponential spatial correlation function.

Table 1 Input parameters used in the simulations

Symbol	Value	Units
$Y = \log K$ random generator		
$\ell_1 \times \ell_2$ (2D)	205×100	[m]
$\ell_1 \times \ell_2 \times \ell_3$ (3D)	$181 \times 91 \times 31$	[m]
μ_Y	1.6	[m/day]
$K_G = \exp[\mu_Y]$	5	[m/day]
σ_Y^2 (2D)	1, 3, 4	[–]
σ_Y^2 (3D)	3	[–]
λ_1, λ_2 (2D)	8, 8	[m]
λ_1, λ_2, λ_3 (3D)	10, 10, 5	[m]
Flow simulations		
h_{in}, h_{out}	10, 0	[m]
$\Delta x_1 \times \Delta x_2$ (2D)	1×1	[m]
$\Delta x_1 \times \Delta x_2 \times \Delta x_3$ (3D)	$1 \times 1 \times 1$	[m]
Transport simulations		
α_1, α_2 (2D)	0.01, 0.001	[m]
α_1, α_2, α_3 (3D)	0.01, 0.001, 0.001	[m]
D_m	8.6×10^{-5}	[m^2/day]
s_1, s_2 (2D)	1, 100	[m]
s_1, s_2, s_3 (3D)	1, 91, 31	[m]
L_1 (2D)	200	[m]
L_1 (3D)	180	[m]
N_p (2D)	10^5	[–]
N_p (3D)	10^7	[–]

The K fields are randomly generated utilizing the SGeMS tool (Remy et al. 2009) which is based on a sequential Gaussian simulation model. All the generated Y fields are characterized by the same value of μ_Y, σ_Y^2 and λ_i. All values used to generate the conductivity fields are reported in Table 1.

In order to compute the first arrival times, we simulate flow and transport on the randomly generated K fields. The governing equations for flow and transport are provided in the "Appendix". Flow is considered to be at steady state and in the absence of sinks and sources with permeameter-like boundary conditions, i.e. prescribed hydraulic heads at the entrance and exit of the domain and no-flow conditions at the remaining boundaries. The hydraulic head at the entrance and exit of porous medium are denoted by h_{in} and h_{out} respectively. The flow field is simulated numerically using MODFLOW (Harbaugh 2005) together with the Python interface Flopy (Bakker et al. 2016).

An inert solute is instantaneously injected along a source zone of area $\mathcal{A}_o = s_1 \times s_2$ (for the 2D case) or volume $\mathcal{V}_o = s_1 \times s_2 \times s_3$ (for the 3D case). We consider both advective and local dispersive mechanisms for the transport problem. The solute plume is simulated through the used of a random walk particle tracking (RWPT) code denoted as PAR^2 (Rizzo et al. 2019). PAR^2 is an open source GPU-accelerated RWPT simulator and it can be downloaded following the instructions provided in Rizzo et al. (2019). For this work, we compute the mass cumulative breakthrough curve at a control plane located at a longitudinal distance L_1 from the solute source zone. From the mass breakthrough curve, we extract the first arrival time, denoted here as $t_{1\%}$.

Values for the parameters used in the flow and transport simulations are provided in Table 1. Table 1 also provides information on the grid resolution and the number of particles used in the numerical simulations.

4.2 Test cases

We demonstrate the capability of the proposed methodology to speed-up the convergence of the first two statistical moments of $t_{1\%}$. In order to evaluate the performance of the ranked-based methodology, we compare the results with the ones obtained through the classic Monte Carlo approach. We start our analysis by considering a 2D computational domain with characteristic dimensions listed in Table 1. For all upcoming results, we set $\sigma_Y^2 = 3$ unless stated otherwise.

Prior to illustrating the features of the proposed connectivity-based ranked distribution methodology, we need to ensure that the mean and standard deviation of $t_{1\%}$ are converged in the classic Monte Carlo approach. Figure 3 presents the convergence analysis of the $t_{1\%}$ cumulative distribution function (CDF). Convergence is assumed to occur when the fluctuations in the probability $\Pr[t_{1\%} < \tau]$ (with τ = 60, 90, 100, 120 and 200 days) become negligible. The results depicted in Fig. 3 also show that the

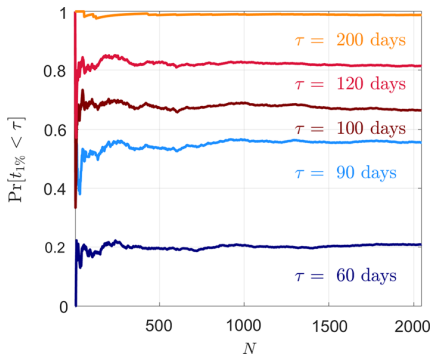

Fig. 3 Convergence analysis of the first arrival ($t_{1\%}$) CDF. The probability that $t_{1\%}$ will take a value less than a specified value τ as a function of the number of Monte Carlo realizations N (where $1 \leq N \leq 2047$) is shown for spatially heterogeneous hydraulic conductivity fields characterized by $\sigma_Y^2 = 3$. All curves reach their respective asymptotic value around $N = 1500$

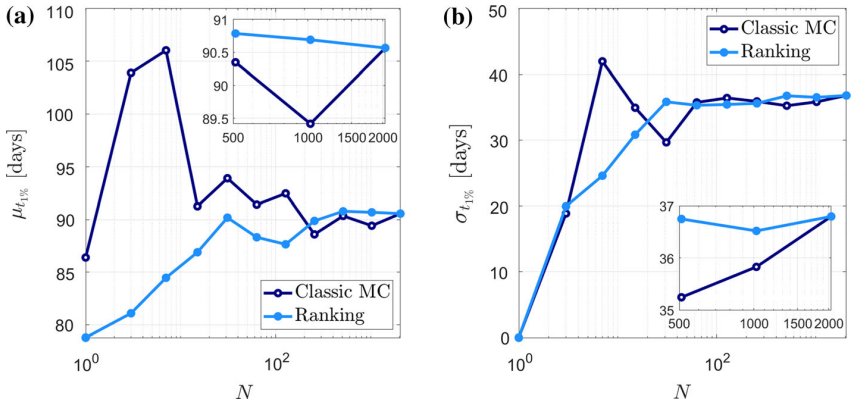

Fig. 4 Convergence analysis for the **a** mean $\mu_{t_{1\%}}$ and **b** standard deviation $\sigma_{t_{1\%}}$ of $t_{1\%}$ obtained from the connectivity-based ranked Monte Carlo (light blue) and for the classic Monte Carlo (dark blue). Results obtained for hydraulic conductivity fields characterized by $\sigma_Y^2 = 3$

probabilities evaluated for the selected values of τ do not change significantly for $N \gtrsim 1500$. With a conservative mindset, we consider that the first two statistical moments of $t_{1\%}$ are converged for $N \gtrsim 2000$. Moreover, the number of K field realizations has been chosen in order for the last balanced subset to include all the N fields according to the relationship for α provided in item 4 of Sect. 3.1. For such reasons, N has to be compliant with:

$$N > 2000 \tag{2}$$

$$N = 2^{\alpha} - 1, \tag{3}$$

thus we set $\alpha = 11$ and consequently $N = 2047$. Therefore, for all upcoming results, we set the mean and standard deviation of $t_{1\%}$ evaluated at $N = 2047$ to be the *reference* values in order to test the performance of the connectivity-based ranked distribution methodology presented in Sect. 3. From the aforementioned $N = 2047$ realizations, we can obtain the number of evaluated subsets, namely $\alpha = 11$ (see Sect. 3.2) which is needed to apply the proposed connectivity-based ranking Monte Carlo method. To keep the comparison with the classic (non-ranked) Monte Carlo unbiased, we only report values of the mean and standard deviation of $t_{1\%}$ obtained from the Monte Carlo samples of size equal to the number of realizations in the α subsets.

In Fig. 4 we show how the first two statistical moments of $t_{1\%}$ at the control plane (see Table 1) vary as a function of sample size N of the selected subsets of K fields. The results reported in Fig. 4 show that both the mean $\mu_{t_{1\%}}$ (Fig. 4a) and standard deviation $\sigma_{t_{1\%}}$ (Fig. 4b) reach convergence at $N \approx 500$ with the proposed connectivity-based ranked distribution methodology, while the results obtained from the classic Monte Carlo present a more unstable trend. The insets shows how the fluctuations of the classic Monte Carlo approach are more pronounced.

To analyze the error associated the Monte Carlo simulations, we define the error metric $\mathcal{E}_\Omega$ as:

$$\mathcal{E}_\Omega(N) = \left| \frac{\Omega|_{N=2^n-1} - \Omega|_{N=2047}}{\Omega|_{N=2047}} \right| \times 100, \tag{4}$$

with $\Omega = [\mu_{t_{1\%}}, \sigma_{t_{1\%}}]$ and $n = 1, \ldots, \alpha$. Note that Eq. (4) represents the relative error measure with respect to the reference values obtained for the statistical moments (i.e. using $N = 2047$, see Fig. 3).

The behaviour of the relative error given by Eq. (4) is illustrated in Fig. 5, where its value for the ranked

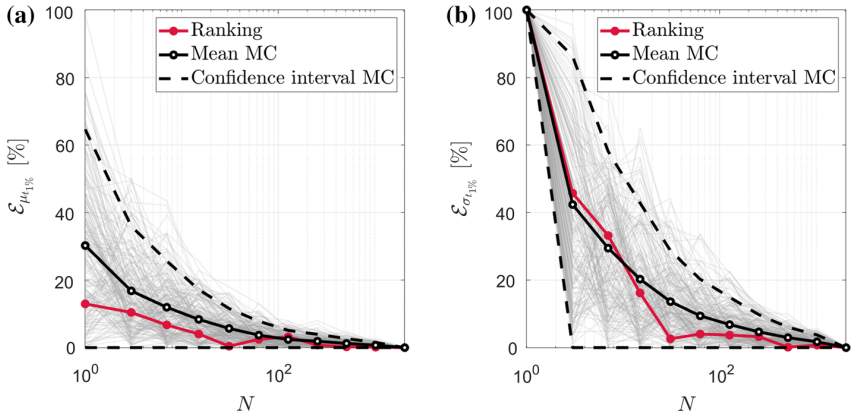

Fig. 5 Relative error $\mathcal{E}$ of **a** $\mu_{t_{1\%}}$ and **b** $\sigma_{t_{1\%}}$ as a function of the ensemble size N. Results for the connectivity-based ranked Monte Carlo (red solid line) and for the classic Monte Carlo ensemble randomly shuffled 200 times (gray lines). The black solid line corresponds to the mean of the relative error over all 200 shuffles, while the black dotted line represents the confidence interval

methodology is compared to the values of the classic Monte Carlo approach. Note that for this error analysis, the classic Monte Carlo ensemble was randomly shuffled 200 times, i.e. all $N = 2047$ K field realizations have been randomly reordered in 200 different ways. This is done in order to account for of all possible convergence behaviours since the randomness of the K field generation process can impact the convergence rate of a given quantity of interest. For example, each "gray" line depicted in Fig. 5 corresponds to a convergence rate of the first arrival time statistics for a specific random sampling of the K field. For completeness, we include the average value and confidence interval of the convergence rate obtained over all randomly reordered fields. Close inspection of Fig. 5 reveals that the proposed methodology leads to a faster error reduction for both the mean and standard deviation of the first arrival times. Furthermore, the connectivity-based ranked distribution methodology removes the randomness effect of the convergence rate of the output statistics (compare "red" solid curve with the "gray" curves in Fig. 5a, b).

The trend of the relative error (4) obtained for $\mu_{t_{1\%}}$ is reported in Fig. 5a. It can be seen that our methodology leads to a much faster convergence of $\mu_{t_{1\%}}$, having an error consistently lower than the average error obtained from the traditional Monte Carlo approach. The only exception is observed for one N value for which the values of the error for the proposed connectivity-based ranked scheme and the average error curve obtained from the classic Monte Carlo shuffled ensembles coincide (see overlapping red and black solid lines). Figure 5b shows that for $N = 31$ the error for $\sigma_{t_{1\%}}$ is kept under 5% for the ranked methodology (red solid line), while the mean of the traditional Monte Carlo shuffled ensembles (black solid line) needs more realizations ($N \approx 255$) to reach the above mentioned error value. Both Fig. 5a, b show that the proposed connectivity-based ranked scheme leads to a smaller error (in the average sense) when compared to the traditional Monte Carlo method.

Similar to Fig. 3, we now compute the probability $\Pr[t_{1\%} < \tau]$ for τ = 60, 90, 100, 120 and 200 days using the connectivity-based ranking method and the classic Monte Carlo approach (see Fig. 6). Figure 6 displays the convergence rate of the probabilities for both approaches. We remark that the connectivity-based ranking method leads to a faster convergence of the probabilities. With $N \approx 100$, the probabilities obtained by the proposed method are almost converged. As a consequence, we can assume that the shape of the CDF reaches a stable structure at reasonably small N values, meaning that not only the first two statistical moments ($\mu_{t_{1\%}}$ and $\sigma_{t_{1\%}}$), but also the higher order moments such as skewness and kurtosis can be evaluated

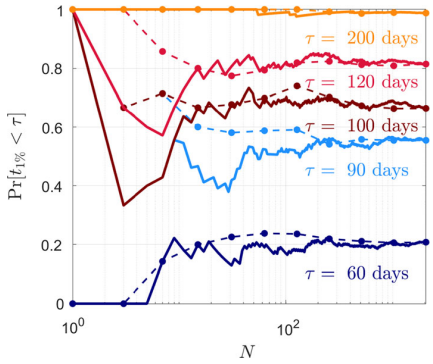

Fig. 6 Comparison of the convergence rate of the probabilities obtained from the traditional Monte Carlo method (solid line) and the proposed connectivity-based ranked Monte Carlo (dashed line) method

using a smaller number of realizations N if the proposed methodology is applied.

Next, we apply our methodology for porous formations displaying different levels of heterogeneity in the hydraulic conductivity field. Heterogeneity is epitomized by the log-conductivity variance σ_Y^2. Figure 7 reports the convergence analysis for $\sigma_Y^2 = 1$, 3 and 4. Results displayed in Fig. 7 reveal that the performance of proposed connectivity-based ranking method improves when σ_Y^2 increases. For example, Fig. 7a, b show that both $\mu_{t_{1\%}}$ and $\sigma_{t_{1\%}}$ computed for $\sigma_Y^2 = 4$ tend to their converged values at lower N (i.e. $N \approx 100$) (and with less oscillations) for the proposed connectivity-based ranking scheme when compared to the values obtained via the traditional Monte Carlo approach. The key reasons for this are as follows: when heterogeneity increases, (1) the likelihood of the occurrence of well-connected highly conductive channels increases and (2) the corresponding least resistance paths are better delineated given the higher contrasts in K values within the domain. Therefore, the performance of the graph theory methodology used to estimate $\mathcal{R}_m$ (Rizzo and de Barros 2017, 2019) improves, which is the foundation of the proposed connectivity-based ranked scheme, see Sect. 3. Additionally, there is more variability in the first arrival time predictions when considering larger values of σ_Y^2. On one hand, this leads to more oscillations on the convergence rate of the statistical output of the first arrival times when the traditional Monte Carlo approach is adopted. On the other hand, for higher heterogeneity, the ranking procedure (which makes use of balanced subsets) reduces this noisy effect leading to faster convergence of the first two statistical moments.

For completeness, we now briefly illustrate the performance of the methodology for a 3D case. Parameter values used in the 3D simulations are reported in Table 1. Similar

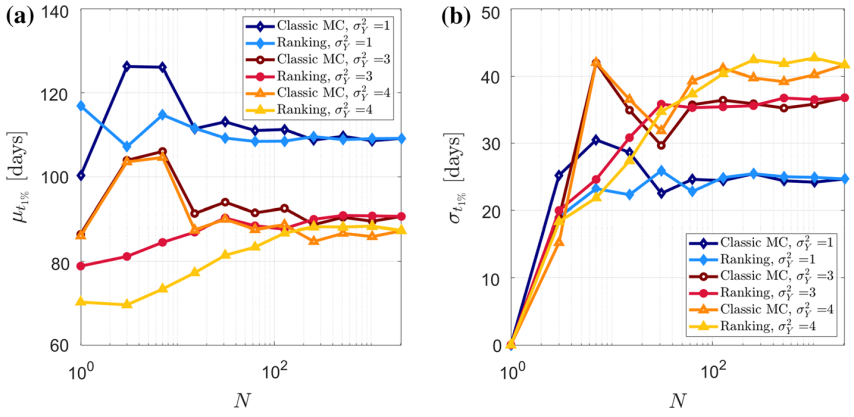

Fig. 7 Convergence rate for the **a** mean and **b** standard deviation of $t_{1\%}$ for $\sigma_Y^2 = 1, 3$ and 4. Comparison between the proposed connectivity-based Monte Carlo method (filled markers) with the traditional Monte Carlo method (void markers)

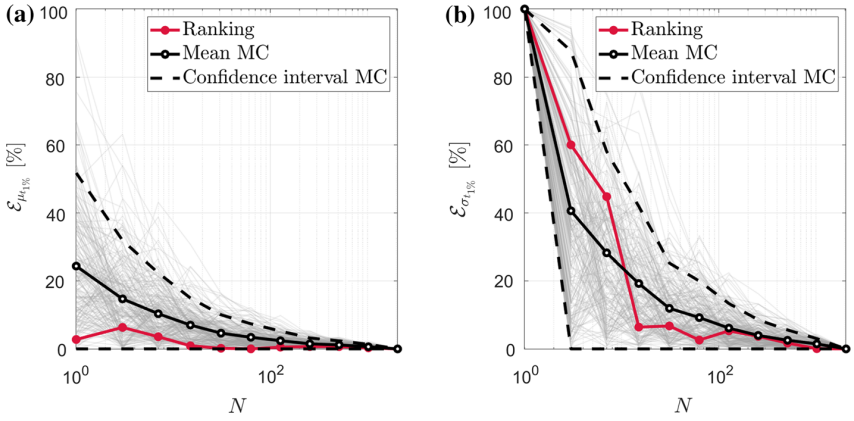

Fig. 8 Relative error $\mathcal{E}$ of **a** $\mu_{t_{1\%}}$ and **b** $\sigma_{t_{1\%}}$ as a function of the ensemble size N of 3D K fields. Results for the connective-based ranked Monte Carlo (red solid line) and for the classic Monte Carlo ensemble randomly shuffled 200 times (gray lines). The black solid line corresponds to the mean of the relative error over all 200 shuffles, while the black dotted line represents the confidence interval

to the 2D case, the 3D scenario analysis takes into account an ensemble of $N = 2047$ hydraulic conductivity fields with $\sigma_Y^2 = 3$. In Fig. 8 we present the same error analysis employed for the 2D scenario (Fig. 5). As shown in Fig. 8a, the adoption of our ranked methodology grants that the error is consistently lower than the mean value of the shuffled classic Monte Carlo ensembles for $\mu_{t_{1\%}}$. Figure 8b reveals that the error for $\sigma_{t_{1\%}}$ is lower than the average (over all shuffled classic Monte Carlo ensembles) error for $N > 15$. As previously mentioned, it is important to note that the ranking procedure in our approach removes the randomness effect of the sampling on the convergence of the model output statistics. The results depicted in Fig. 8 show the potential of the proposed methodology in reducing the computational burden associated with the estimation of first arrival times in 3D simulations.

5 Summary

This paper provides a new methodology to improve the computational efficiency of Monte Carlo simulations aimed at estimating the uncertainty of first arrival times of a solute plume in a spatially heterogeneous porous formation. As shown in Chapter 9 of Rubin (2003), first arrival times are subject to largest uncertainty.

The methodology proposed in this work is based on ranking the randomly generated hydraulic conductivity fields according to their connectivity and re-ordering the ranked ensemble through a balanced binary tree sampling procedure. We employ the minimum hydraulic resistance as a connectivity metric given (1) its strong correlation with first arrival times and (2) that it can be obtained at a very low computational cost through graph theory [see details in Rizzo and de Barros (2017)]. Due to the latter, the processing time needed to perform the ranking procedure is negligible.

We test the proposed methodology against the results obtained through the traditional Monte Carlo method (i.e. non ranked). Our results show that the mean and standard deviation of the first arrival times (obtained from the ranked and re-ordered ensemble) converge faster for both 2D and 3D domains and with less oscillations. Furthermore, we show how the level of heterogeneity of the porous formation impacted the performance of the connectivity-ranked Monte Carlo simulations. The results

reported in this work show that the performance of the proposed method increases with the increasing level of heterogeneity since the connectivity metric employed in our work becomes a better indicator of the first arrival time (Rizzo and de Barros 2017).

Future research directions consist of applying this methodology to real sites. In this work, the proposed methodology was illustrated in an ensemble of unconditional hydraulic conductivity fields. As shown in Rizzo and de Barros (2019), the graph theory-based connectivity metric is also a random variable and can be obtained in both unconditional and conditional random fields. Details regarding the statistical features of the minimum hydraulic resistance can be found in the literature (Rizzo and de Barros 2019). If the geostatistical model is uncertain, our approach can be expanded by adopting Bayesian averaging concepts to generate the conductivity fields (Neuman 2003) needed to estimate the connectivity metric ensemble.

Acknowledgements The authors acknowledge the financial support provided by the National Science Foundation (Grant Number 1654009).

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

Appendix: Flow and transport formulations

We consider a steady-state flow in a spatially heterogeneous aquifer. The flow field is governed by:

$$\nabla \cdot (K \nabla h) = 0, \tag{5}$$

with h denoting the hydraulic head and K the hydraulic conductivity.

For all our 2D and 3D simulations, we consider permeameter-like boundary conditions to ensure the flow is uniform-in-the-mean. That is achieved by setting Dirichlet boundary conditions on the inflow and outflow of the domain and no-flow Neumann conditions on the remaining boundaries.

An inert solute is instantaneously released along a line source perpendicular to the mean flow direction. Transport is assumed to be governed by the advection-dispersion equation:

$$\frac{\partial c}{\partial t} + \mathbf{u} \cdot \nabla c = \nabla \cdot (\mathbf{D} \nabla c). \tag{6}$$

where c is the resident concentration, $\mathbf{u}$ is the velocity field, $\mathbf{D}$ is the local-dispersion tensor assumed to be anisotropic and defined as:

$$\mathbf{D} = (\alpha_T |\mathbf{u}| + D_m)\mathbf{I} + \frac{\alpha_L - \alpha_T}{|\mathbf{u}|}\mathbf{u}\mathbf{u}^T \tag{7}$$

where D_m is the molecular diffusion, α_L is the longitudinal dispersivity and α_T is the transverse dispersivity, where in our study x_1 denotes the longitudinal dimension and x_2 and x_3 the transverse ones. Transport is solved through a random walk particle tracking (RWPT) code. The RWPT code used in our work is GPU-based and denoted as PAR^2 (Rizzo et al. 2019).

References

Andričević R, Cvetković V (1996) Evaluation of risk from contaminants migrating by groundwater. Water Resour Res 32(3):611

Bakker M, Post V, Langevin CD, Hughes JD, White J, Starn J, Fienen MN (2016) Scripting MODFLOW model development using Python and FloPY. Groundwater 54(5):733

Ballio F, Guadagnini A (2004) Convergence assessment of numerical monte carlo simulations in groundwater hydrology. Water Resour Res 40(4):e2003WR002876

Bellin A, Salandin P, Rinaldo A (1992) Simulation of dispersion in heterogeneous porous formations: statistics, first-order theories, convergence of computations. Water Resour Res 28(9):2211

Berrone S, Hyman J, Pieraccini S (2020) Multilevel Monte Carlo predictions of first passage times in three-dimensional discrete fracture networks: a graph-based approach. Water Resour Res 56(6):e2019WR026493

Bianchi M, Pedretti D (2017) Geological entropy and solute transport in heterogeneous porous media. Water Resour Res 53(6):4691

Bianchi M, Zheng C, Wilson C, Tick GR, Liu G, Gorelick SM (2011) Spatial connectivity in a highly heterogeneous aquifer: from cores to preferential flow paths. Water Resour Res 47(5):e2009WR008966

Booth AD, Colin AJ (1960) On the efficiency of a new method of dictionary construction. Inf Control 3(4):327

de Barros F, Bellin A, Cvetkovic V, Dagan G, Fiori A (2016) Aquifer heterogeneity controls on adverse human health effects and the concept of the hazard attenuation factor. Water Resour Res 52(8):5911

Deutsch CV (1998) Fortran programs for calculating connectivity of three-dimensional numerical models and for ranking multiple realizations. Comput Geosci 24(1):69

Dijkstra EW et al (1959) A note on two problems in connexion with graphs. Numer Math 1(1):269

Fiori A, Jankovic I (2012) On preferential flow, channeling and connectivity in heterogeneous porous formations. Math Geosci 44(2):133

Fiori A, Jankovic I, Dagan G (2011) The impact of local diffusion upon mass arrival of a passive solute in transport through three-dimensional highly heterogeneous aquifers. Adv Water Resour 34(12):1563

Fuks O, Ibrahima F, Tomin P, Tchelepi HA (2019) Analysis of travel time distributions for uncertainty propagation in channelized porous systems. Transp Porous Media 126(1):115

Geng X, Michael HA (2020) Preferential flow enhances pumping-induced saltwater intrusion in volcanic aquifers. Water Resour Res 56(5):e2019WR026390

Gershenzon NI, Soltanian MR, Ritzi RW Jr, Dominic DF, Keefer D, Shaffer E, Storsved B (2015) How does the connectivity of open-framework conglomerates within multi-scale hierarchical fluvial

architecture affect oil-sweep efficiency in waterflooding? Geosphere 11(6):2049
Gotovac H, Cvetkovic V, Andricevic R (2009) Flow and travel time statistics in highly heterogeneous porous media. Water Resour Res 45(7):e2008WR007168:
Gotovac H, Cvetkovic V, Andricevic R (2010) Significance of higher moments for complete characterization of the travel time probability density function in heterogeneous porous media using the maximum entropy principle. Water Resour Res 46(5):e2009WR008220
Harbaugh AW (2005) MODFLOW-2005, the US Geological Survey modular ground-water model: the ground-water flow process. US Department of the Interior, US Geological Survey, Reston
Harvey CF, Gorelick SM (1995) Temporal moment-generating equations: modeling transport and mass transfer in heterogeneous aquifers. Water Resour Res 31(8):1895
Henri C, Harter T (2019) Stochastic assessment of nonpoint source contamination: joint impact of aquifer heterogeneity and well characteristics on management metrics. Water Resour Res 55(8):6773
Henri C, Fernàndez-Garcia D, de Barros F (2015) Probabilistic human health risk assessment of degradation-related chemical mixtures in heterogeneous aquifers: risk statistics, hot spots, and preferential channels. Water Resour Res 51(6):4086
Henri C, Fernàndez-Garcia D, de Barros F (2016) Assessing the joint impact of DNAPL source-zone behavior and degradation products on the probabilistic characterization of human health risk. Adv Water Resour 88:124
Henri CV, Harter T, Diamantopoulos E (2020) On the conceptual complexity of non-point source management: impact of spatial variability. Hydrol Earth Syst Sci 24(3)
Hyman JD, Hagberg A, Srinivasan G, Mohd-Yusof J, Viswanathan H (2017) Predictions of first passage times in sparse discrete fracture networks using graph-based reductions. Phys Rev E 96(1):013304
Jabbari N, Aminzadeh F, de Barros FP (2017) Hydraulic fracturing and the environment: risk assessment for groundwater contamination from well casing failure. Stoch Environ Res Risk Assess 31(6):1527
Jimenez-Martinez J, Negre CF (2017) Eigenvector centrality for geometric and topological characterization of porous media. Phys Rev E 96(1):013310
Knudby C, Carrera J (2005) On the relationship between indicators of geostatistical, flow and transport connectivity. Adv Water Resour 28(4):405
Knudby C, Carrera J (2006) On the use of apparent hydraulic diffusivity as an indicator of connectivity. J Hydrol 329(3–4):377
Knuth DE (1968) The art of computer programming, volume 1: fundamental, algorithms. Addison-Wesley, Reading
Le Goc R, de Dreuzy JR, Davy P (2010) Statistical characteristics of flow as indicators of channeling in heterogeneous porous and fractured media. Adv Water Resour 33(3):257
Leube PC, de Barros FPJ, Nowak W, Rajagopal R (2013) Towards optimal allocation of computer resources: trade-offs between uncertainty quantification, discretization and model reduction. Environ Model Softw 50:97
Libera A, Henri C, de Barros F (2019) Hydraulic conductivity and porosity heterogeneity controls on environmental performance metrics: implications in probabilistic risk analysis. Adv Water Resour 127:1
Loll P, Moldrup P (1998) A new two-step stochastic modeling approach: application to water transport in a spatially variable unsaturated soil. Water Resour Res 34(8):1909
Maxwell R, Kastenberg W (1999) Stochastic environmental risk analysis: an integrated methodology for predicting cancer risk from contaminated groundwater. Stoch Environ Res Risk Assess 13(1–2):27
Moslehi M, Rajagopal R, de Barros FPJ (2015) Optimal allocation of computational resources in hydrogeological models under uncertainty. Adv Water Resour 83:299
Neuman SP (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch Environ Res Risk Assess 17(5):291
Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press, Cambridge
Renard P, Allard D (2013) Connectivity metrics for subsurface flow and transport. Adv Water Resour 51:168
Riva M, Sánchez-Vila X, Guadagnini A, De Simoni M, Willmann M (2006) Travel time and trajectory moments of conservative solutes in two-dimensional convergent flows. J Contam Hydrol 82(1–2):23
Rizzo CB, de Barros FP (2017) Minimum hydraulic resistance and least resistance path in heterogeneous porous media. Water Resour Res 53(10):8596
Rizzo C, de Barros F (2019) Minimum hydraulic resistance uncertainty and the development of a connectivity-based iterative sampling strategy. Water Resour Res 55(7):5593
Rizzo C, Nakano A, de Barros F (2019) Par2: parallel random walk particle tracking method for solute transport in porous media. Comput Phys Commun 239:265
Rubin Y (2003) Applied stochastic hydrogeology. Oxford University Press, Oxford
Rubin Y, Dagan G (1992) Conditional estimation of solute travel time in heterogeneous formations: impact of transmissivity measurements. Water Resour Res 28(4):1033
Sahimi M, Davis HT, Scriven L (1983) Dispersion in disordered porous media. Chem Eng Commun 23(4–6):329
Sánchez-Vila X, Carrera J, Girardi JP (1996) Scale effects in transmissivity. J Hydrol 183(1–2):1
Savoy H, Kalbacher T, Dietrich P, Rubin Y (2017) Geological heterogeneity: goal-oriented simplification of structure and characterization needs. Adv Water Resour 109:1
Shapiro AM, Cvetkovic VD (1988) Stochastic analysis of solute arrival time in heterogeneous porous media. Water Resour Res 24(10):1711
Trinchero P, Sánchez-Vila X, Fernàndez-Garcia D (2008) Point-to-point connectivity, an abstract concept or a key issue for risk assessment studies? Adv Water Resour 31(12):1742
Tyukhova AR, Willmann M (2016) Connectivity metrics based on the path of smallest resistance. Adv Water Resour 88:14
Tyukhova AR, Kinzelbach W, Willmann M (2015) Delineation of connectivity structures in 2-D heterogeneous hydraulic conductivity fields. Water Resour Res 51(7):5846
Zhang D, Shi L, Chang H, Yang J (2010) A comparative study of numerical approaches to risk assessment of contaminant transport. Stoch Environ Res Risk Assess 24(7):971
Zimmerman D, De Marsily G, Gotway CA, Marietta MG, Axness CL, Beauheim RL, Bras RL, Carrera J, Dagan G, Davies PB et al (1998) A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resour Res 34(6):1373