Solute transport in bounded porous media characterized by Generalized Sub-Gaussian log-conductivity distributions

Guillem Sole-Mari[1,2,3], Monica Riva[4], Daniel Fernàndez-Garcia[2,3], Xavier Sanchez-Vila[2,3], Alberto Guadagnini[4]

[1]Earth and Environmental Sciences Area, Geochemistry Department, Lawrence Berkeley National Laboratory

[2]Department of Civil and Environmental Engineering (DECA), Universitat Politècnica de Catalunya

[3]Associated Unit: Hydrogeology Group (UPC-CSIC)

[4] Department of Civil and Environmental Engineering (DICA), Politecnico di Milano

Corresponding author: Guillem Sole-Mari (gsolemari@lbl.gov)

Highlights:

- Log-conductivity (Y) is modeled as a Generalized Sub-Gaussian (GSG) field
- Average longitudinal spreading is smaller for GSG than for Gaussian Y fields
- Plume stretching in GSG fields is enhanced with respect to Gaussian fields
- Local dispersion masks identification of impacts of non-Gaussianity on transport
- The GSG nature of Y can be difficult to identify relying only on transport metrics

Abstract

There are increasing evidences that probability distributions and associated statistical moments of a variety of hydrogeological and soil science variables and their spatial increments display distinctive scale-dependent features that are not captured by a typical Gaussian model. A Generalized Sub-Gaussian (GSG) model is able to capture key aspects of this pattern. We present the results of a suite of computational analyses set in a Monte Carlo framework and aimed at assessing the impact of a GSG structure of log hydraulic conductivity (Y) on transport of a conservative solute through a three-dimensional bounded porous medium under steady-state saturated Darcy flow. Our results indicate that the longitudinal spreading of a plume is on average significantly smaller for Sub-Gaussian than for Gaussian Y fields. Otherwise, the velocity field arising from a Sub-Gaussian Y field induces enhanced plume stretching with respect to what can be observed in a Gaussian Y setting, this aspect potentially influencing the strength of solute mixing within these two types of conductivity domains. We also find that, in some cases, it may be difficult to identify the nature of the underlying conductivity field relying solely on observations of solute concentrations migrating within the system. In this regard, we show that the action of local dispersion tends to mask the influence of Sub-Gaussianity on major transport metrics.

Keywords: Porous media, flow and transport, heterogeneity, stochastic modeling, Sub-Gaussian models

1. Introduction

Flow and transport in porous and fractured media are strongly affected by spatial variability of hydraulic properties of the system. While it is virtually impossible to characterize all details of such heterogeneity, a large body of literature has framed the assessment of flow and transport processes taking place in heterogeneous porous media on a (geo)statistical description of continuum-scale system attributes, major emphasis being given on hydraulic conductivity (e.g., Dagan, 1984; Gelhar, 1993; Kitanidis, 1997; Rubin, 2003). Stochastic analyses of the effects of heterogeneity on flow and transport in porous media have been the subject of several investigations for over four decades. Among the variety of approaches developed in this context, studies embedded in a numerical Monte Carlo (MC) framework have enabled to explore subtle relationships between the heterogeneous features of the subsurface and salient elements characterizing flow and transport (e.g., Bellin et al., 1992; Naff et al., 1998a,b; Riva et al., 2008; de Dreuzy et al., 2012; Pedretti et al., 2017; Siena and Riva, 2018 and references therein). A typical Monte Carlo study comprises the analysis of a collection of outputs of interest that are generated from one or multiple models, given uncertainty in their inputs. In the context of flow and transport in heterogeneous porous media, this often involves the generation of a collection of realizations of random hydraulic conductivity fields.

Several studies consider the logarithm of hydraulic conductivity ($Y(\mathbf{x}) = \ln K(\mathbf{x})$, vector $\mathbf{x}$ denoting space location) to be described through a Gaussian distribution, characterized by a given variogram/covariance function describing the degree of spatial dependence based only upon the distance between points; i.e., assuming second-order stationarity. Rather, there is increasing evidence documenting the occurrence of distinct non-Gaussian features characterizing the distributions of a variety of hydrogeological and soil science variables. Key documented manifestations of such a behavior include the

observation that the distribution of spatial increments of such variables taken between two points separated by a given spatial distance (lag) tend to be symmetric and to develop heavier tails and sharper peaks as lag decreases. Such a behavior has been displayed (among others) by log-hydraulic conductivity and permeability (Painter, 1996, 2001; Liu and Moltz, 1997; Meerschaert et al., 2004; Siena et al., 2012, 2019; Riva et al., 2013a, 2013b; Guadagnini et al., 2018), electrical resistivity (Painter, 2001), vadose zone hydraulic properties (Guadagnini et al., 2012, 2013, 2014), neutron porosity (Riva et al., 2015a), sediment transport (e.g., Ganti et al., 2009), fully developed turbulence (Boffetta et al., 2008), and micro-scale geochemical data (Siena et al., 2020).

The Generalized Sub-Gaussian (GSG) model introduced in the literature in the last few years (Riva et al., 2015a, 2015b; Panzeri et al., 2016) has been shown to be capable of capturing all of these aspects. According to the GSG model, the departure of the distribution of a variable and its two-point increments from the Gaussian one is given by the action of a (spatially uncorrelated) subordinator on an otherwise spatially correlated Gaussian random field. This modeling strategy allows representing jointly within a unique framework the above-documented behavior (as described by probability distributions and/or moments) of a quantity and its incremental values and has been successfully applied to the interpretation of main features displayed by various subsurface attributes (Riva et al 2015a; Guadagnini et al., 2018; Siena et al., 2020, and references therein).

These concepts have already been employed in preliminary analytical and numerical studies of flow and transport in porous media whose log-conductivity is characterized through a GSG model. Riva et al. (2017) present lead-order analytical flow and transport solutions in unbounded GSG log-conductivity fields under mean-uniform flow. Libera et al. (2017) rely on a numerical Monte Carlo framework to analyze the joint

95 effects of a GSG heterogeneous log-conductivity field and a temporally variable pumping
96 rate on solute breakthrough curves (BTCs) detected at a pumping well operating in a two-
97 dimensional domain. Besides spatial dimensionality, three major limitations can be
98 identified for the analytical study by Riva et al. (2017): (i) macrodispersion is the only
99 transport metric analyzed, (ii) the joint impact of local dispersivity combined with the
100 heterogeneous advection driven by Y is not evaluated, and (iii) output uncertainty due to
101 finite size of the medium is not considered.

102 Starting from the above studies, here we focus on the influence of randomly
103 heterogeneous GSG log-conductivity fields on transport processes taking place therein
104 and address the following two questions: (1) At what extent does Sub-Gaussianity (i.e.,
105 the degree of departure from a Gaussian behavior) of Y impact main features and
106 observables of solute transport driven by advection and local dispersion in three-
107 dimensional bounded heterogeneous porous media?; and (2) Would the Sub-Gaussian
108 nature of Y be clearly identifiable upon relying solely on the analysis of solute
109 concentration fields in these settings or could it be disguised into estimates of apparent
110 structural parameters of a Gaussian model? To address these issues, here we rest on a
111 suite of computational analyses framed in a Monte Carlo context and explore the effects
112 of a Sub-Gaussian distribution of Y on observable transport metrics associated with a
113 solute migrating within a bounded three-dimensional porous medium under the action of
114 advective and dispersive mechanisms and subject to a uniform (in the mean) flow.

115 The work is structured as follows. Section 2 is devoted to a synthesis of the
116 theoretical framework underlying the Sub-Gaussian model and of the computational
117 analyses performed. Presentation and discussion of the results are included in Section 3.
118 Finally, in Section 4 we enumerate our main conclusions.

119

2. Theoretical framework, methodology and set-up of scenarios

2.1 Generalized Sub-Gaussian model and stochastic generation of three-dimensional log hydraulic conductivity fields

We consider a three-dimensional porous medium where log-conductivity, $Y(\mathbf{x}) = \ln K(\mathbf{x})$, is described as a stationary random function of space with constant ensemble mean, $\langle Y \rangle$, and zero-mean random fluctuation, $Y'(\mathbf{x}) = Y(\mathbf{x}) - \langle Y \rangle$. The latter forms a Generalized Sub-Gaussian process, GSG, defined as (Riva et al. 2015a, b)

$$Y'(\mathbf{x}) = U(\mathbf{x})G(\mathbf{x}). \quad (1)$$

Here, $\mathbf{x} = [x, y, z]^T$, $G(\mathbf{x})$ is a zero-mean Gaussian spatially correlated second-order stationary field, and $U(\mathbf{x})$ is a random non-negative spatially uncorrelated stationary field that is independent of $G(\mathbf{x})$. The statistical properties of $Y'(\mathbf{x})$ vary with the subordinator $U(\mathbf{x})$ and with the Gaussian field $G(\mathbf{x})$. Details about the mathematical formulation and theoretical framework of the GSG model can be found in Riva et al. (2015a) and Siena et al. (2020), who present analytical expressions for probability density functions (*pdf*) and statistical moments of Y' and of the corresponding increments (ΔY) evaluated at various separation lags, including formulations for covariance and variogram functions as well as integral scale of Y'. It is worth noting that, regardless the distributional form of U, the variogram of Y' is characterized by an integral scale which is always smaller than the one associated with the Gaussian field G and a nugget effect, rendered by the product of the variance of G and the variance of U.

Here, we follow Riva et al. (2015a, 2017) and consider Monte Carlo realizations of $Y'(\mathbf{x})$ associated with a log-normal subordinator, i.e., $\ln(U) \sim \mathcal{N}[0, (2-\alpha)^2]$, $\alpha < 2$. As such, $Y'(\mathbf{x})$ is characterized by two constant parameters (α, σ_G^2) and a correlation function ρ_G; σ_G^2 and ρ_G being variance and correlation function of $G(\mathbf{x})$, respectively. Note that

144 $Y'(\mathbf{x})$ becomes Gaussian and coincides with $G(\mathbf{x})$ when $\alpha \to 2$. For completeness,
145 Appendix A includes a brief summary of the key analytical formulations and properties
146 of the GSG model we consider in this study.

147 We generate three-dimensional realizations of $Y(\mathbf{x})$ of size $L_x \times L_y \times L_z$ by setting
148 $\langle Y \rangle = 0$, $\sigma_Y^2 = 1$ (where $\sigma_Y^2 = e^{2(2-\alpha)^2}\sigma_G^2$) and adopting an isotropic exponential
149 correlation function for $G(\mathbf{x})$, i.e. $\rho_G = e^{-s/I_G}$; s and I_G being the separation distance
150 (lag) between two locations and the integral scale of $G(\mathbf{x})$, respectively. We set $L_x =$
151 $L_y = 5I_Y$ and $L_z = 40I_Y$ in all realizations (z being the dominant direction of flow, see
152 Section 2.2), and $I_Y = e^{-(2-\alpha)^2}I_G$ denoting the integral scale of Y (see also Appendix A).
153 The spatial grid employed for the generation involves 125,000 cubic blocks of size $I_Y/5$.
154 Figure 1 depicts examples of $Y(\mathbf{x})$ realizations obtained for differing values of α. To
155 appreciate the nature of the random log-conductivity fields considered, we illustrate in
156 Appendix A the main features of the probability density functions associated with Y' and
157 its spatial increments ΔY evaluated at various separation lags for the fields depicted in
158 Figure 1. Figure A1 depicts the probability density function of Y', $f_{Y'}$, for the three values
159 of α considered in Figure 1 together with the Gaussian ($\alpha \to 2$) distribution.
160 Corresponding *pdfs* of spatial increments, $f_{\Delta Y}$, evaluated at short, intermediate and large
161 (with respect to I_Y) separation lags are depicted in Figure A2. The pattern associated with
162 the behavior of peaks and tails of these *pdfs* can be described quantitatively by analyzing
163 deviations from Gaussianity, clearly revealed by the excess standardized kurtosis,
164 reported in Figure A3. It is clear that the peak of the *pdf* of Y' or ΔY grows sharper and
165 the associated tails become heavier as α decreases, thus evidencing the deviations of the
166 fields we consider from a Gaussian behavior.

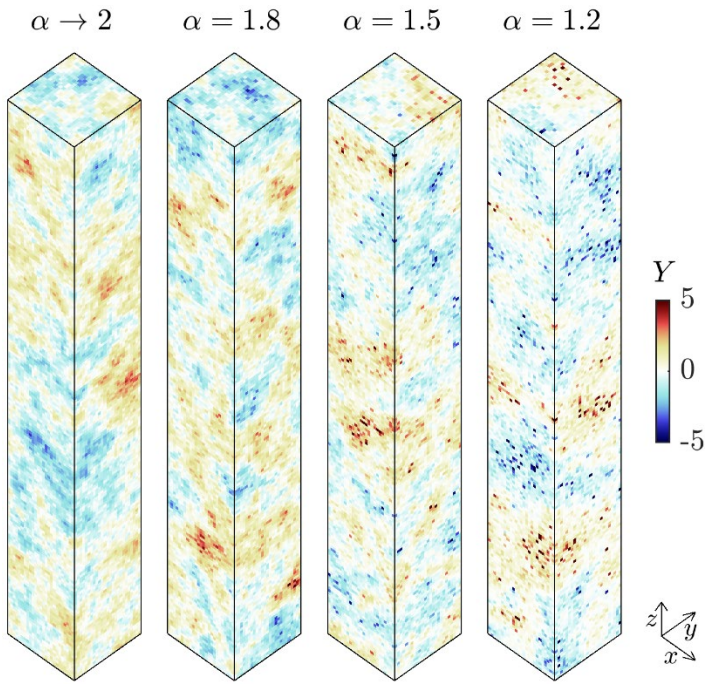


Figure 1: Example of $Y(\boldsymbol{x})$ individual realizations for various degrees of Sub-Gaussianity ($\alpha \to 2$ being equivalent to a Gaussian distribution).

As a remark, we note that the variance of Y in geologic porous media has sometimes been reported to attain markedly high values, *e.g.*, $\sigma_Y^2 \geq 4$ (Fogg & Zhang, 2016). It is well recognized that this stems from a homogenization of conductivity values within a unique population, while they actually pertain to diverse geological facies. One may alternatively represent the system as a composite medium, whose internal architecture is characterized by the presence of distinct facies, within which hydraulic conductivity can be spatially heterogeneous. As such, conductivity variance within a given geomaterial can be mild (Winter & Tartakovsky, 2000, 2002). Hence, the selected mild value of the log-conductivity variance in this work ($\sigma_Y^2 = 1$) is deemed as representative of the natural variability contained within a geological unit, which can potentially be depicted through statistically stationary heterogeneity models (Winter et al., 2003) like the one we consider. The relative size of the cross-section (i.e., $L_x = L_y = 5I_Y$) is set to be large enough to capture the main features of the log-conductivity distributions, while being small enough

184 to generate variability in the results between realizations with identical configuration,
185 similar to what one would expect from a collection of finite-sized samples belonging to
186 the same geological material or unit.

187 We further remark that here we focus on the analysis of the joint effects on transport
188 of the strength of (*i*) the departure of Y from a Gaussian model (as rendered by the
189 parameter α related to the subordinator U), and (*ii*) the relative importance of advective
190 and dispersive processes in Sub-Gaussian Y fields. Future studies can tackle also the
191 potential influence of the degree of heterogeneity, as expressed in terms of σ_Y^2, or consider
192 representing a log-conductivity field involving multiple subdomains, with diverse
193 geomaterials through a multimodal distribution, such as a mixture of Sub-Gaussian *pdfs*.

194 **2.2 Flow and transport model**

195 Steady-state single phase flow is described through mass balance and Darcy's
196 equations within the heterogeneous $K(\mathbf{x})$ fields generated as described in Section 2.1.

197 $$\nabla \cdot \mathbf{q}(\mathbf{x}) = 0, \qquad \mathbf{q}(\mathbf{x}) = -K(\mathbf{x})\nabla H(\mathbf{x}). \tag{2}$$

198 Here, $H(\mathbf{x})$ is hydraulic head and $\mathbf{q}(\mathbf{x})$ is Darcy flux. We set impermeable boundary
199 conditions on the lateral sides of the domain (x = 0, L_x; y = 0, L_y). Flow is forced along
200 direction z by setting a constant mean flow, $\bar{q}$, at the bottom of the sample, and by setting
201 a prescribed head at the top boundary. A homogeneous buffer zone of length $2I_Y$ and with
202 $Y = \langle Y \rangle$ is placed at both ends of the system (Figure 2) to stabilize the inflow condition.
203 These boundary conditions correspond to what one would typically impose in an
204 experimental set-up, head at the inlet being practically uniform in the transverse direction
205 due to the presence of the homogeneous buffer region. Imposing a fixed mean
206 longitudinal flow $\bar{q}$ for all realizations enables us to study transport phenomena which are
207 driven by relative spatial fluctuations of $q(\mathbf{x})$ around the same mean value.

208 At the initial time ($t = 0$) a flux-averaged pulse injection of solute takes place
209 across the bottom of the heterogeneous region, i.e., along the horizontal plane $z = 0$.
210 Solute transport is described by the Advection Dispersion Equation (ADE)

211 $$\frac{\partial c}{\partial t}(\mathbf{x}, t) = \boldsymbol{\nabla} \cdot \big(\mathbf{D}(\mathbf{x})\boldsymbol{\nabla} c(\mathbf{x}, t)\big) - \boldsymbol{\nabla} \cdot \big(\mathbf{v}(\mathbf{x}) c(\mathbf{x}, t)\big), \quad (3)$$

212 where $c(\mathbf{x}, t)$ is resident solute concentration (i.e., solute per unit volume of the porous
213 medium), and $\mathbf{v}(\mathbf{x}) = \mathbf{q}(\mathbf{x})/\phi$, ϕ being porosity which we consider constant in this study.
214 In the context of modeling transport in porous media, local dispersion is an upscaled
215 apparent process that accounts for the combined effect of unresolved pore-scale velocity
216 variations and molecular diffusion (e.g., Bear, 1972). At high pore-scale Péclet numbers,
217 it is often considered to be proportional to the fluid velocity. To simplify the problem, we
218 assumed an isotropic local dispersion tensor $\mathbf{D}(\mathbf{x})$,

219 $$\mathbf{D}(\mathbf{x}) = D(\mathbf{x})\mathbf{1}_3, \qquad D(\mathbf{x}) = a|\mathbf{v}(\mathbf{x})|, \quad (4)$$

220 where $\mathbf{1}_3$ is the 3×3 identity matrix, a is a constant dispersivity, and $|\cdot|$ denotes the norm
221 of a vector. The advection and dispersion timescales for a characteristic distance I_Y can
222 be compared through the Péclet number, $\mathrm{Pe} = I_Y/a$, higher values of Pe corresponding
223 to a lower relative significance of the local dispersion process. In this study we explore
224 settings associated with values of Pe ranging from $\mathrm{Pe} = 10$ to $\mathrm{Pe} \to \infty$.

225 The steady-state flow problem (2) is solved by the widely tested finite difference
226 code MODFLOW 2005 (Harabaugh, 2005) using the Preconditioned Conjugate Gradient
227 method (PCG). The output cell-interface fluxes are then used to solve the ADE by means
228 of the Random Walk Particle Tracking method suggested by LaBolle et al. (2000). The
229 latter solves the so-called Generalized Stochastic Differential Equations (GSDE) and is
230 based on the following particle movement

$$\mathbf{X}_p(t+\mathrm{d}t) = \mathbf{X}_p(t) + \mathbf{v}\left(\mathbf{X}_p(t)\right)\mathrm{d}t + \sqrt{2D\left(\mathbf{X}_p(t)+\boldsymbol{\Lambda}_p(t)\right)\mathrm{d}t}\,\boldsymbol{\xi}_p(t), \tag{5}$$

$$\boldsymbol{\Lambda}_p(t) = \sqrt{2D\left(\mathbf{X}_p(t)\right)\mathrm{d}t}\,\boldsymbol{\xi}_p(t), \tag{6}$$

where $\mathbf{X}_p(t)$ is the position of particle p at time t, and $\boldsymbol{\xi}_p(t)$ is a vector of 3 (the number of spatial dimensions) independent uncorrelated standard normal random numbers. Velocities at any point inside a numerical cell Ω are evaluated by linear interpolation

$$\mathbf{v}(\mathbf{x}\in\Omega) = (1-\hat{\mathbf{x}})\odot\mathbf{v}_\Omega^- + \hat{\mathbf{x}}\odot\mathbf{v}_\Omega^+, \tag{7}$$

where $\hat{\mathbf{x}} = (\mathbf{x}-\mathbf{x}_\Omega^-)\oslash(\mathbf{x}_\Omega^+-\mathbf{x}_\Omega^-)$; operators "$\odot$" and "$\oslash$" correspond to the elementwise Hadamard product and division, respectively; $\mathbf{x}_\Omega^-$ and $\mathbf{x}_\Omega^+$ are vectors containing the lower and upper cell interface location (for each direction); and $\mathbf{v}_\Omega^-$ and $\mathbf{v}_\Omega^+$ are vectors whose entries are the corresponding cross-interface flow velocities.

The GSDE method (5) is particularly accurate and efficient when representing isotropic local dispersion characterized by strong local variations. Therefore, this method is suitable for reproducing transport features even within porous media characterized by local highly non-uniform flow conditions. We refer to Salamon et al. (2006) for a comprehensive review of random-walk methods for solving the advection-dispersion equation in spatially heterogeneous settings.

Solution of the flow problem (2) may require using a spatial grid finer than that introduced in Section 2.1 and according to which log-conductivity is generated. The grid block size, Δ, as well as the number of particles, n, and the time step, Δt, considered for the implementation of the GSDE (5) have been selected as detailed in the following. The spatial grid resolution (i.e., $1/\Delta$) and the number of particles n are increased until convergence is attained, i.e., until no significant changes are observed in the numerical solution (as quantified through the metrics/indicators detailed in Section 3). The time step

Δt is adapted dynamically, to fulfill the condition which is most restrictive between $\Delta t \leq \mathcal{C}\Delta/|\mathbf{v}|$ and $\Delta t \leq \mathcal{P}^2\Delta^2/2D$, $\mathcal{C}$ and $\mathcal{P}$ being numerical discretization parameters which are adjusted in the same fashion as Δ and n. The numerical model has been found to give an accurate solution of the flow and transport problem by setting $n = 2^{18}$ (= 262,144), $\mathcal{C} = \mathcal{P} = 0.125$, and subdividing each conductivity block into 64 elements, i.e., $\Delta = I_Y/20$, giving rise to a computational grid formed by 8 million cells.

All particles are assigned an identical mass M/n and are injected according to a flux-weighting scheme on the horizontal plane at $z = 0$. Flow and transport are simulated across $N = 5{,}000$ Monte Carlo realizations of $Y(\mathbf{x})$ (see Section 2.1) for α= 1.2, 1.5, 1.8, and $\alpha \rightarrow$ 2.0 (i.e., Gaussian field), and for Pe = 10, 100, 1000, and Pe$\rightarrow\infty$ (i.e., pure advection), for a total of 80,000 simulations. The latter have been performed using 180 cores from the Marconi A1 partition of the high-performance computer center CINECA (https://www.cineca.it/en). The total computational time required for our study is about 100,000 core hours.

In Section 3, we analyze the spatial and temporal moments, as well as additional key metrics (see also Appendix B for details), of

$$C_r(z,t) = \frac{\phi}{M}\int_0^{L_x}\int_0^{L_y} c_r(\mathbf{x},t)\,\mathrm{d}y\,\mathrm{d}x, \tag{8}$$

$$J_r(z,t) = \frac{\phi}{M}\int_0^{L_x}\int_0^{L_y} \mathbf{n}^{\mathrm{T}}\cdot[\mathbf{v}_r(\mathbf{x})c_r(\mathbf{x},t) - D_r(\mathbf{x})\nabla c_r(\mathbf{x})]\,\mathrm{d}y\,\mathrm{d}x, \tag{9}$$

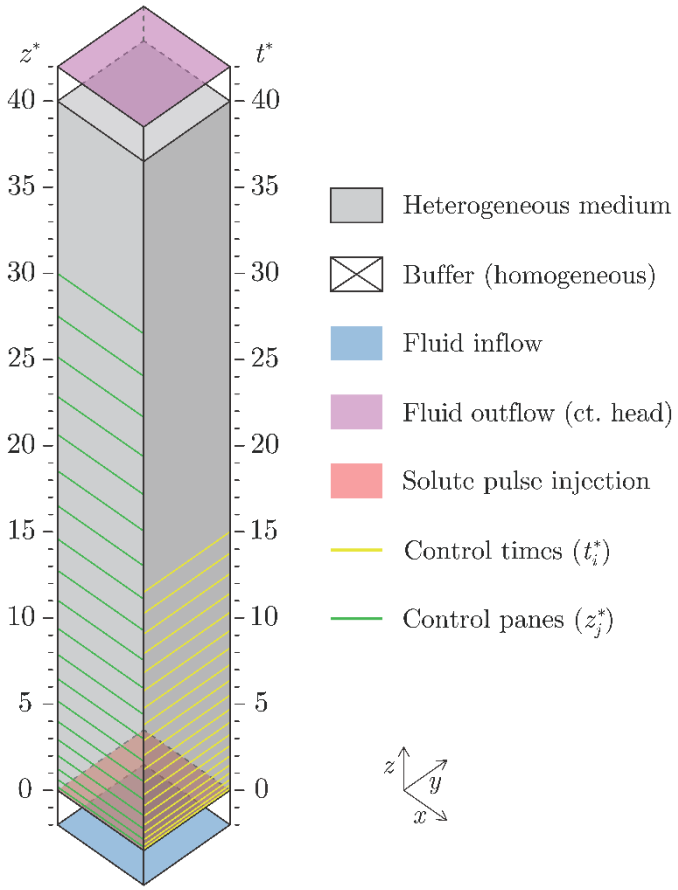
where $\mathbf{n}^{\mathrm{T}} = [0,0,1]$, and subscript r refers to the realization number, i.e., $r = 1, \ldots, N$. The quantity $C_r(z,t)$ is the longitudinal density of solute mass normalized by the total mass injected (hence it integrates to 1 along z). The quantity $J_r(z,t)$ is the net temporal flux of mass through a horizontal plane, also normalized by the total mass (hence it

integrates to 1 with t). Hereafter, we refer to $C_r(z,t)$ as the section-integral concentrations, and to $J_r(z,t)$ as the breakthrough curve (BTC).

Results shown in Section 3 are illustrated in terms of dimensionless times, t^*, and longitudinal positions, z^*, defined as

$$t^* = \frac{\bar{v}}{I_Y} t, \qquad z^* = \frac{z}{I_Y}, \tag{10}$$

where $\overline{v} = \overline{q}/\phi$. Note that t^* corresponds to the number of integral scales of Y traversed by the center of mass of the plume. The evolution of the solute plume is monitored by recording particle positions at 20 increasing dimensionless times t_i^*, $i \in \{1\,,2\,,\ldots,20\}$, with $t_1^* = 0.1$, $t_{20}^* = 15$, and $t_i^* = t_1^*\, i^{\log_{20}(t_{20}^*/t_1^*)}$. It has been verified that all particles are still located inside the heterogeneous domain in all MC realizations at t_{20}^*. We record particle passage times through selected planes perpendicular to the mean flow direction at distance z_j^* ($j \in \{1\,,2\,,\ldots,20\}$, with $z_1^* = 0.2$, $z_{20}^* = 30$, and $z_j^* = z_1^*\, j^{\log_{20}(z_{20}^*/z_1^*)}$) from the origin. The design of the setup is depicted in Figure 2.



289

290 *Figure 2: Design of the three-dimensional numerical simulations and scheme for*
291 *recording the results. The control times are shown graphically in terms of the*
292 *corresponding mean displacement of the fluid. Empty non-shaded regions at both ends*
293 *correspond to the homogeneous buffers.*

294

295 Our simulations yield a collection (obtained by sampling across MC realizations)
296 of (*i*) solute concentration profiles at a given observation time, $C_r(z, t_i)$, or (*ii*) solute
297 breakthrough curves (BTCs) at a given transverse section, $J_r(z_j, t)$. We follow Liu and
298 Müller (2004) and Lu and Stauffer (2012) and note that, in this context, a point-wise
299 average is not a proper estimate of the mean concentration profile (or curve). For instance,
300 such an average tends to overestimate the spread of the tails and may also underestimate
301 the peak. This issue is addressed by employing a percentile average over the cumulative
302 forms of all curves (see Lu and Stauffer (2012) for details). The shape of the ensuing
303 average curves is generally more similar than the point-wise average to the one of curves

corresponding to individual realizations. Here, we evaluate percentile averages $\bar{C}(z,t_i)$ and $\bar{J}(z_j,t)$ upon organizing simulation results (given in terms of particle positions and arrival times, respectively) in increasing order, and then averaging across realizations, to finally reconstruct the density of data by Kernel Density Estimation (KDE). Details on this procedure are offered in Appendix B.1 (see also Appendix D for details on KDE). For brevity, hereafter we refer to $\bar{C}(z,t_i)$ and $\bar{J}(z_j,t)$ as the average longitudinal concentration profile and the average BTC, respectively. Note that these percentile-averaged ensemble quantities are denoted with an overline, as opposed to angular brackets (i.e., $\langle\,\rangle$) indicating arithmetic ensemble average, which we employ for the scalar metrics analyzed (see eq. (11) in Section 3).

3. Results and discussion

We start by analyzing our results in terms of average spatial and temporal dynamics of the solute on the basis of the numerical Monte Carlo framework described in Section 2. We do so by considering the setting corresponding to pure advective mechanisms, i.e., $\mathrm{Pe} \to \infty$ (Section 3.1.1), to then include the assessment of the impact of considering a finite value for Pe (Section 3.1.2). Section 3.2 illustrates the analysis related to the complete set of statistical distributions of the modeling goals of interest.

3.1. Average space-time dynamics of the solute

Hereafter, we resort to the notation $\langle b \rangle$ to indicate the arithmetic average of a metric b of interest, as obtained across the collection of N Monte Carlo realizations, i.e.,

$$\langle b \rangle = \frac{1}{N}\sum_{r=1}^{N} b_r\,, \tag{11}$$

b_r being the value of metric b in the r^{th} realization.

3.1.1. Average spatial and temporal behavior under pure advection

We start by considering a fully advective transport scenario (Pe → ∞).

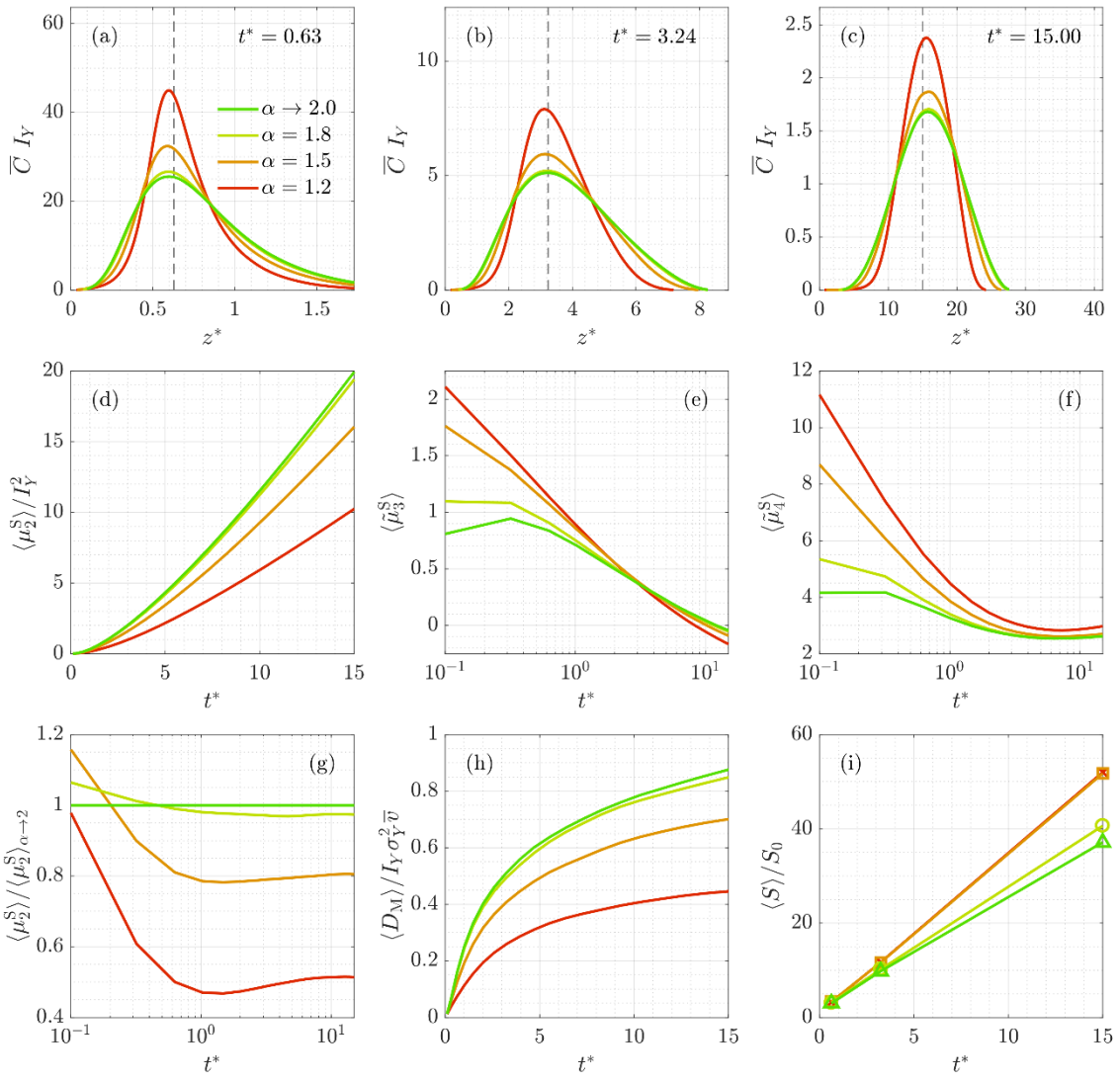


Figure 3: Average temporal evolution of the solute distribution in the longitudinal direction: (a-c) Longitudinal concentration profile at various times (dashed gray line corresponds to $z = \overline{v}t$), (d) spatial variance, (e) spatial skewness, (f) spatial kurtosis, (g) spatial variance normalized by the Gaussian case, (h) macrodispersion coefficient, and (i) advective stretching.

Figures 3(a-c) depict the average longitudinal concentration profile, $\bar{C}$, evaluated as detailed in Section 2.2 and Appendix B.1, for various values of α at increasing values of t^* (the vertical dashed gray lines correspond to the average fluid displacement, $z = \overline{v}t$). In the following, we discuss the effect of α on the temporal evolution of the plume by

analyzing these concentration profiles in combination with various metrics of interest depicted within Figures 3(d-i).

The longitudinal extent of the region where one can find (on average) significant concentration values tends to decrease with departure of Y from Gaussianity (i.e., decreasing α). A quantification of this qualitative behavior can be offered through the average centered second moment of the longitudinal profile of section-integral concentrations, $\langle \mu_2^{\mathrm{S}} \rangle$, which is an effective measure of longitudinal dispersion (the latter being different from the local dispersion, as quantified by **D** in (3)), and whose temporal evolution is depicted in Figure 3(d). Details on the definition and computation of spatial and temporal moments in each MC realization are provided in Appendix B.2. Stronger degrees of Sub-Gaussianity (i.e., lower values of α) are associated with decreased values of $\langle \mu_2^{\mathrm{S}} \rangle$. Figure 3(g) also reveals that the ratio between any of the $\langle \mu_2^{\mathrm{S}} \rangle$ curves and the one corresponding to the Gaussian case $\langle \mu_2^{\mathrm{S}} \rangle_{\alpha \to 2}$ tends to become constant after a few characteristic times (i.e., for $t^* > 1$). Quantities $1 - \langle \mu_2^{\mathrm{S}} \rangle / \langle \mu_2^{\mathrm{S}} \rangle_{\alpha \to 2}$ and $(2 - \alpha)^2$ display a nearly perfect linear correlation at late times (with a slope of ~ 0.78, details not shown). We note that dispersion in Gaussian Y fields is, under some assumptions, proportional to σ_Y^2 (Dagan, 1984). As such, our results imply that, at long times, relying solely on an analysis of the macrodispersive behavior of a solute transported by pure advection within a saturated porous formation may not allow to distinguish whether the underlying log hydraulic conductivity distribution is (*i*) Sub-Gaussian ($\alpha < 2$) with variance σ_Y^2, or (*ii*) Gaussian with some variance σ_Y^{*2} being lower than σ_Y^2. Yet, as shown in the following, these two models lead to quite different estimates in a number of observables.

Figure 3(h) depicts the evolution of the average longitudinal (macro)dispersion coefficient, $\langle D_{\mathrm{M}} \rangle$, computed for each realization as

$$D_{\mathrm{M},r}(t_i) = \frac{1}{2}\frac{\partial \mu^{\mathrm{S}}_{2,r}}{\partial t}\bigg|_{t_i} \cong \frac{\mu^{\mathrm{S}}_{2,r}(t_i) - \mu^{\mathrm{S}}_{2,r}(t_{i-1})}{2(t_i - t_{i-1})}. \quad (12)$$

This quantity is normalized by $\sigma_Y^2 I_Y \bar{v}$, which is the long-time asymptotic value of D_{M} in an infinite medium with Gaussian log-conductivities characterized by low spatial heterogeneity. Note that an asymptotic dispersion regime is not fully reached by $t^* = 15$. Riva et al. (2017) show that $D_{\mathrm{M}}/\sigma_Y^2 I_Y \bar{v} \to 1$ (at a rate which depends on α) in two-dimensional Sub-Gaussian Y fields and in the long-time limit, this result corresponding to analytical findings by Dagan (1982, 1984, 1989) and Rubin (1990a,b) for Gaussian Y fields under the same assumptions on flow. Assuming that a corresponding long-time limit holds also in three-dimensional unbounded systems, the results depicted in Figure 3(h) suggest that the impact of the presence of boundaries on the dynamics of D_{M} (i.e., on the departure of $D_{\mathrm{M}}/\sigma_Y^2 I_Y \bar{v}$ from 1) tends to increase with decreasing α.

Further inspection of Figures 3(a-c) reveals that low values of α yield a more pronounced peak for longitudinal concentration profiles. Additionally, the standardized distribution of vertical locations of the solute exhibits longer tails as α decreases (not shown). These features are markedly more evident at early times ($t^* = 0.63$). This result may be related to the observation that distributions of Y increments (which are related to spatial variations of velocities, that ultimately control longitudinal solute spreading) in Sub-Gaussian fields display heavy tails and sharp peaks at short distances (Riva et al., 2015a). Differences between curves associated with differing values of α become less apparent at late times ($t^* = 15$). One can also detect an increasing degree of asymmetry of the plume as α decreases, both at early and late times.

All of these qualitative observations are quantified by the analysis of the temporal evolutions of the normalized spatial moments of the plume, namely the spatial skewness $\tilde{\mu}^{\mathrm{S}}_3$ (Figure 3(e)) and kurtosis $\tilde{\mu}^{\mathrm{S}}_4$ (Figure 3(f)) (see Appendix B.2 for details about their

evaluation). We note that low α values tend to promote more skewed (i.e., asymmetric) plumes, being positively skewed at early times (a behavior mainly driven by high velocities) and (slightly) negatively skewed at late times (mainly driven by low velocities). Figure 3(f) suggests that lower values of α are associated (on average) with more kurtotic plumes. Differences between (average) spatial skewness and kurtosis of plumes evolving in Gaussian and Sub-Gaussian cases are substantial at early times and tend to become mild as time increases, consistent with the observation that the distribution of Y spatial increments in a Sub-Gaussian field tends to approach the Gaussian one as lag increases.

Figure 3(i) depicts the temporal evolution of the average advective stretching, i.e., $\langle S \rangle / S_0$. This metric quantifies hydrodynamic deformation through the evaluation of the aggregated temporal growth of a fluid surface, S, represented by a collection of particles displaced by advection and originally arranged on the horizontal plane $z = 0$ (i.e., corresponding to $S_0 = L_x L_y$). All details are provided in Appendix B.4. On average, the velocity field associated with a Sub-Gaussian *Y* field induces more stretched plumes than those evolving through a Gaussian setting. It is worth noting that this takes place despite the longitudinal spreading (represented by μ_2^S) being (sometimes significantly) smaller for Sub-Gaussian than for Gaussian fields. One may compare these results with those reported by Le Borgne et al. (2013, 2015) in two-dimensional Gaussian Y fields. Numerical simulations performed by these authors document that the mean elongation of a solute line displays a power-law temporal behavior with an exponent ranging between 0.65 and 1.25 for σ_Y^2 increasing from 0.25 to 4.0. Interpreting the temporal evolution of $\langle S \rangle / S_0$ through a power-law behavior (as in the above-mentioned Gaussian Y settings), its (best-fit) exponent increases for decreasing values of α (while $\sigma_Y^2 = 1$ is kept constant), from 0.85 in the Gaussian case to 0.95 for $\alpha \leq 1.5$. These results are probably

412 related to the observation that low values of α promote marked changes in longitudinal
413 velocity values at very short transverse distances (details not shown here for brevity), thus
414 inducing enhanced degrees of short-range fluid stretching. This effect seems to be capped
415 at strong degrees of Sub-Gaussianity (i.e., $\alpha = 1.2$), for which the average surface
416 increase is almost identical to the case of moderate-to-high Sub-Gaussian behavior (i.e.,
417 $\alpha = 1.5$).

418 The stark differences observed between the effects of a Sub-Gaussian Y field on the
419 longitudinal dispersion (Figure 3(d)) and on advective stretching (Figure 3(i)) could
420 indicate that the increased longitudinal dispersion associated with Gaussian Y fields with
421 respect to a Sub-Gaussian setting may not necessarily translate into enhanced mixing, the
422 latter being strongly linked to stretching (see, e.g., Dentz et al., 2011; Le Borgne et al.,
423 2013, 2015).

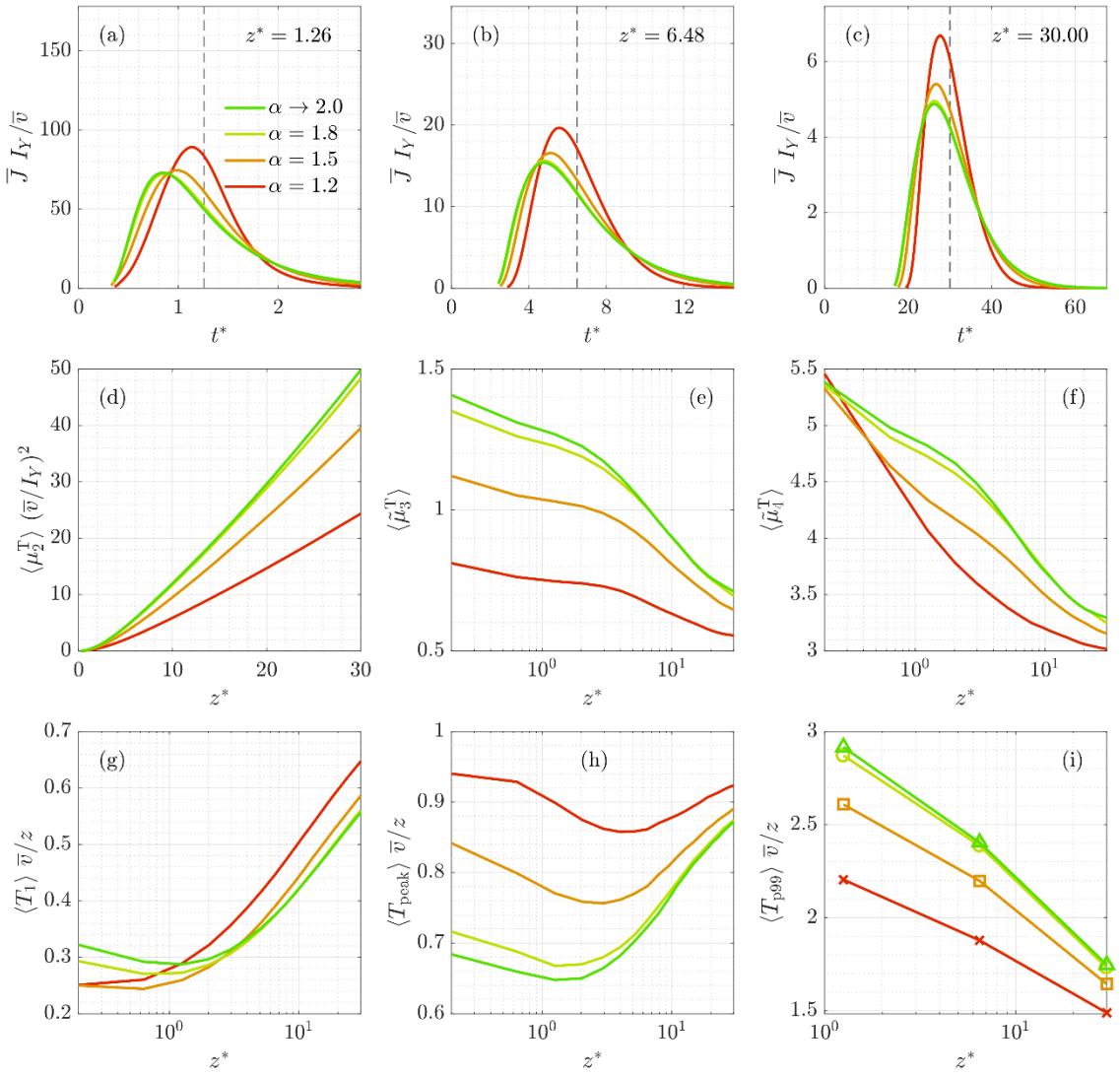


Figure 4: Average spatial evolution of the solute BTCs across planes orthogonal to the mean flow direction: (a-c) BTCs at various distances from the inlet (dashed gray line corresponds to $t = z/\overline{v}$), (d) temporal variance, (e) temporal skewness, (f) temporal kurtosis, (g) first arrival, (h) arrival of the peak, and (i) arrival of the 99% of the total mass.

Figures 4(a-c) depict the average temporal evolution of the solute mass flux, representing the solute breakthrough curve (BTC), for various values of α and at increasing distances from the inlet. The average BTCs are visibly narrower for lower values of α. This is related to the observation that the spatial evolution of the centered second temporal moment, depicted in Figure 4(d), has been found to be qualitatively very

similar to its temporal evolution counterpart (Figure 3(d)) in all cases analyzed, a feature which was also noted by Fernàndez-Garcia et al. (2005) for transport in three-dimensional, Gaussian Y fields. The average BTC is also visibly more symmetric for lower values of α, the arrival of the peak being less advanced in time with respect to the mean arrival time $z/\overline{v}$ (marked on Figures 4(a-c) by a gray vertical dashed line) than its counterpart based on Gaussian Y fields. This latter feature is particularly evident at short distances from the inlet ($z^* = 1.26$).

For all values of α, the BTCs become more symmetric as z^* increases. This behavior can be quantified through the (average) skewness of the BTCs, i.e., $\langle \hat{\mu}_3^{\mathrm{T}} \rangle$, whose spatial evolution along the column (Figure 4(e)) denotes a decreasing trend of the temporal skewness (*i*) with increasing distance from the inlet and (*ii*) with departure from Gaussian behavior of the underlying Y field. A similar feature is observed in Figure 4(f) for the average temporal kurtosis, $\langle \hat{\mu}_4^{\mathrm{T}} \rangle$.

Figures 4(g), 4(h) and 4(i) complement the description by depicting the average longitudinal distribution of key arrival times, corresponding to the time of arrival of the first particle, T_1, the BTC peak, T_{peak}, and the 99$^{\mathrm{th}}$ percentile, T_{p99}, respectively. These quantities are normalized by the mean arrival time (i.e., $z/\overline{v}$) for ease of interpretation. It is generally recognized that first arrival times are strongly linked to connectivity metrics, such as, e.g., the least resistance path, which is the path connecting two points or surfaces with the largest associated effective hydraulic conductivity (e.g., Gómez-Hernández and Wen, 1998; Rizzo and De Barros, 2017). In this context, Figure 4(g) suggests that Sub-Gaussian Y fields are more connected than their Gaussian counterparts at short distances ($z^* \leq 1$), the trend being reversed at longer distances ($z^* \gg 1$). This finding could be related to the aforementioned nugget effect displayed by the longitudinal velocity covariances associated with low α values, which would increase the likelihood of

velocities changing abruptly at very short distances, and possibly decrease the likelihood of occurrence of persistent high velocity paths.

As α decreases, the peak of the BTC (Figure 4(h)) is on average closer to the mean arrival time, a finding in agreement with the aforementioned reduced temporal skewness. This implies that in Sub-Gaussian fields the largest concentrations tend to be observed (on average) appreciably later than in Gaussian fields, especially for short distances (i.e., $z^* \approx 1$). The average longitudinal distribution of the 99$^{\text{th}}$ percentile of arrival times, T_{p99} (Figure 4(i)), shows that the BTC tends to be more compacted for lower values of α. All of these findings suggest that, as compared to a Gaussian distribution of log-conductivities, in the presence of a Sub-Gaussian Y field one can (on average) expect (*i*) a generally delayed arrival of the mass of a flux-injected solute to crossing planes located a few integral scales away from the injection plane (i.e., for $z^* > 5$), as expressed in terms of both the first arrival and the arrival of the peak; (*ii*) more pronounced peaks; and (*iii*) an earlier breakthrough of the majority (99%) of the solute mass.

3.1.2. Analysis of the impact of local dispersion

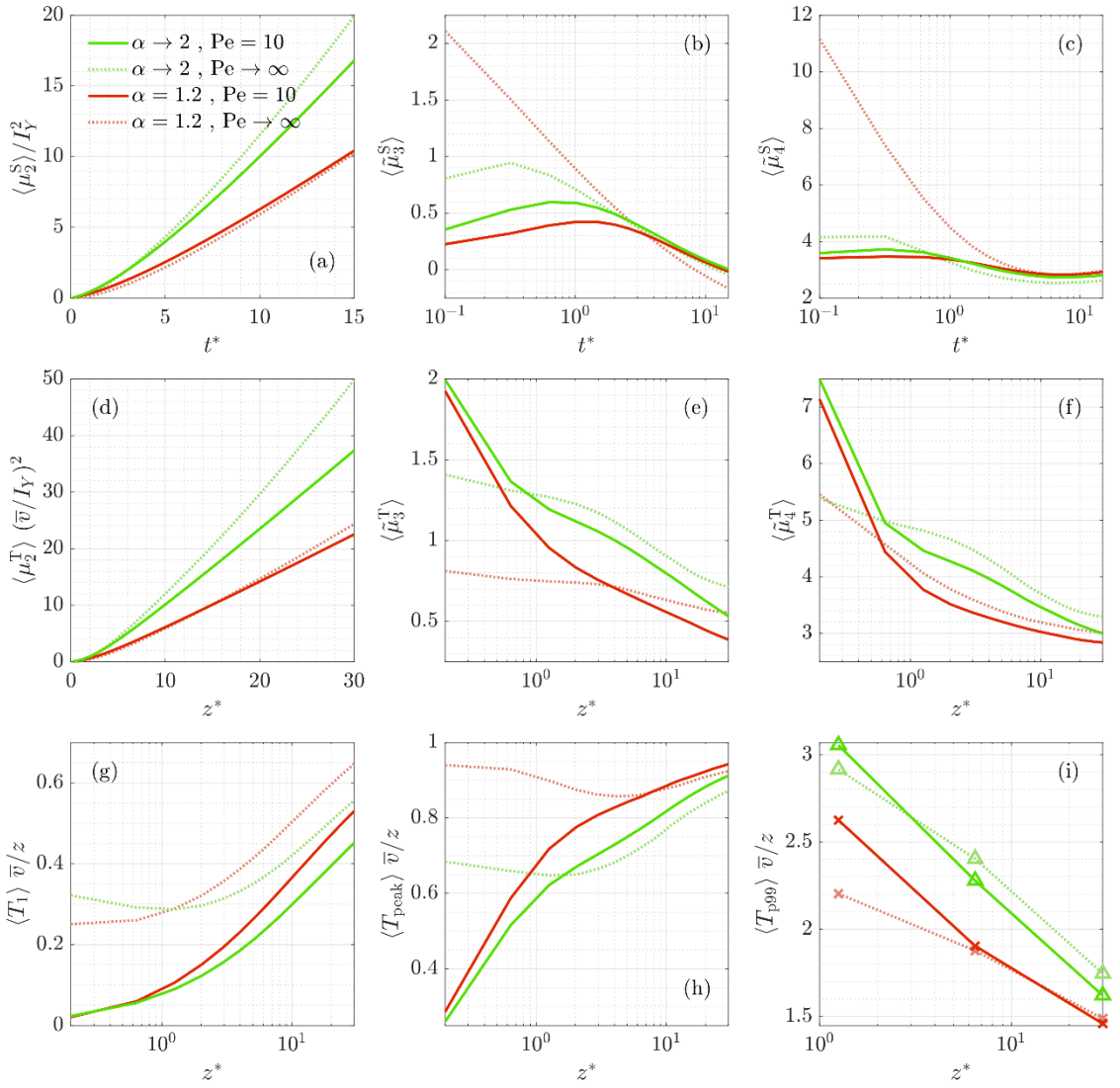


Figure 5: Effect of local dispersion ($Pe = 10$) compared to pure advection ($Pe \to \infty$) on the results with Gaussian and Sub-Gaussian ($\alpha = 1.2$) Y fields, for (a) spatial variance (macrodispersion), (b) spatial skewness, (c) spatial kurtosis, (d) temporal variance, (e) temporal skewness, (f) temporal kurtosis, (g) first arrival, (h) arrival of the peak, and (i) arrival of the 99% of the solute mass.

Here, we concentrate on the scenario corresponding to the lowest value of α considered (i.e., $\alpha = 1.2$; strongly non-Gaussian Y), compared to the Gaussian Y scenario (i.e., $\alpha \to 2$), to clearly illustrate the effects of local dispersion (i.e., finite values of Pe) on transport. In Figure 5 we compare the spatial and temporal variance, skewness and kurtosis as well as the longitudinal distribution of key arrival times (defined in Section

488 3.1.1) obtained for $\mathrm{Pe} \to \infty$ (dotted curves) with those corresponding to $\mathrm{Pe} = 10$ (solid
489 curves).

490 As one could expect, local dispersion typically reduces the difference between the
491 results for Gaussian and Sub-Gaussian settings. In the Gaussian case (green curves), one
492 can see that $\langle \mu_2^{\mathrm{S}} \rangle$ (Figure 5a) decreases due to the effect of local dispersion. This effect is
493 increasingly evident for decreasing values of Pe (not shown). Such a behavior has been
494 documented (analytically, numerically, and experimentally) in previous studies (e.g.,
495 Dentz et al., 2000, 2002, 2003; Dartois et al., 2018; Gist et al., 1990) and is related to the
496 observation that local dispersion favors particles to sample a wider range of velocities,
497 thus reducing the likelihood of sampling only extreme (low or high) velocity paths.
498 Conversely, this effect is not apparent in the strongly Sub-Gaussian fields we examine.
499 However, the aforementioned similarity between the temporal evolution of $\langle \mu_2^{\mathrm{S}} \rangle$ and the
500 spatial evolution of $\langle \mu_2^{\mathrm{T}} \rangle$ and the results of Figure 5(d) suggest that the reduction in
501 longitudinal dispersion observed for $\mathrm{Pe} = 10$ with respect to $\mathrm{Pe} \to \infty$ will eventually take
502 place in time also for $\alpha = 1.2$.

503 While it was noted that the average longitudinal skewness of the plume, $\langle \tilde{\mu}_3^{\mathrm{S}} \rangle$, tends
504 to deviate more strongly from zero both at early and late times as α decreases when $\mathrm{Pe} \to$
505 ∞, Figure 5(b) clearly documents that this trend is reversed at early times in the presence
506 of local dispersion. A similar behavior is documented in Figure 5(c) for the average spatial
507 kurtosis, $\langle \tilde{\mu}_4^{\mathrm{S}} \rangle$, the results corresponding to Gaussian and Sub-Gaussian settings tending
508 to become very similar under the action of local dispersion.

509 The average higher-order moments of the BTC, i.e., $\langle \tilde{\mu}_3^{\mathrm{T}} \rangle$ (Figure 5(e)) and $\langle \tilde{\mu}_4^{\mathrm{T}} \rangle$
510 (Figure 5(f)), are highly influenced by local dispersion at early times. At later times, local
511 dispersion appears to induce a (downward) shift in the results, with a pattern that is similar

for $\alpha \to 2$ and $\alpha = 1.2$. This feature is also observed for the first arrival time, $\langle T_1 \rangle$. Otherwise, the mass flux peak (Figure 5(h)) is delayed in time (on average) by the action of a finite Pe, which also reduces the difference between the cases corresponding to $\alpha \to 2$ and $\alpha = 1.2$. The effect of dispersion on the (average) 99$^{\text{th}}$ percentile arrival time, $\langle T_{\text{p99}} \rangle$, is analogous to that on $\langle \mu_2^{\text{T}} \rangle$. As such, one can see that these two metrics and $\langle \mu_2^{\text{S}} \rangle$ appear to contain very similar information about the effects of α and Pe on the processes analyzed.

3.2. Monte Carlo-based statistical distributions of metrics characterizing transport

Here, we discuss the results obtained in terms of the full distributions of quantities of interest across the collection of Monte Carlo realizations analyzed. We focus on the system behavior at early and late stages of the plume deformation and confine the discussion to the scenarios corresponding to $\text{Pe} \to \infty$ and $\text{Pe} = 10$. As a first result, we observe that the empirical MC-based distributions associated with most of the analyzed metrics are not well differentiated for the various values of α considered (i.e., there is some overlap between distributions). As such, we investigate the significance of relying on the metrics considered to discriminate the extent at which the log-conductivity field deviates from the Gaussian behavior (as reflected by the value of α) when only results from a single realization of Y are available. In other words, we analyze how significantly each metric of solute transport is impacted by the value of α characterizing a given Y realization. We do so upon relying on the Bayes classifier (e.g., James et al., 2013) combined with Kernel Density Estimation.

Following the developments detailed in Appendix C, we quantify the degree of overlap between sample distributions of a given metric by means of the index $P_{k\ell} \coloneqq P(\alpha = \alpha_k | \hat{\alpha} = \alpha_\ell)$, which represents the likelihood of the actual value of α being equal

to α_k, given that the classification method infers an estimate $\hat{\alpha} = \alpha_\ell$ (from comparing the result of a realization to the corresponding Monte Carlo collection of results). We consider the discrete set of α values corresponding to $\alpha_k, \alpha_\ell \in \{1.2\,, 1.5\,, 1.8\,, \to 2\}$. The values of index $P_{k\ell}$ are then seen as the entries of a (4×4) matrix, which we refer to as overlap matrix. Note that one can obtain an overlap matrix for each of the transport metrics analyzed. Values on the diagonal (i.e., $P_{k\ell}$ with $k = \ell$) of this matrix indicate the extent to which we can rely on the distribution of results for a given metric to identify the actual value of α associated with the underlying Y field. High values of the index concentrated on the diagonal of the overlap matrix indicate mild overlap (i.e., strong differentiation) of population densities, values evenly distributed across columns suggesting strong overlap (i.e., poor differentiation). The strongest overlaps are observed for high-order moments, whereas advective stretching is the best-differentiable metric. For most metrics, overlap is strengthened at later times and for lower values of Pe. A detailed analysis follows below.

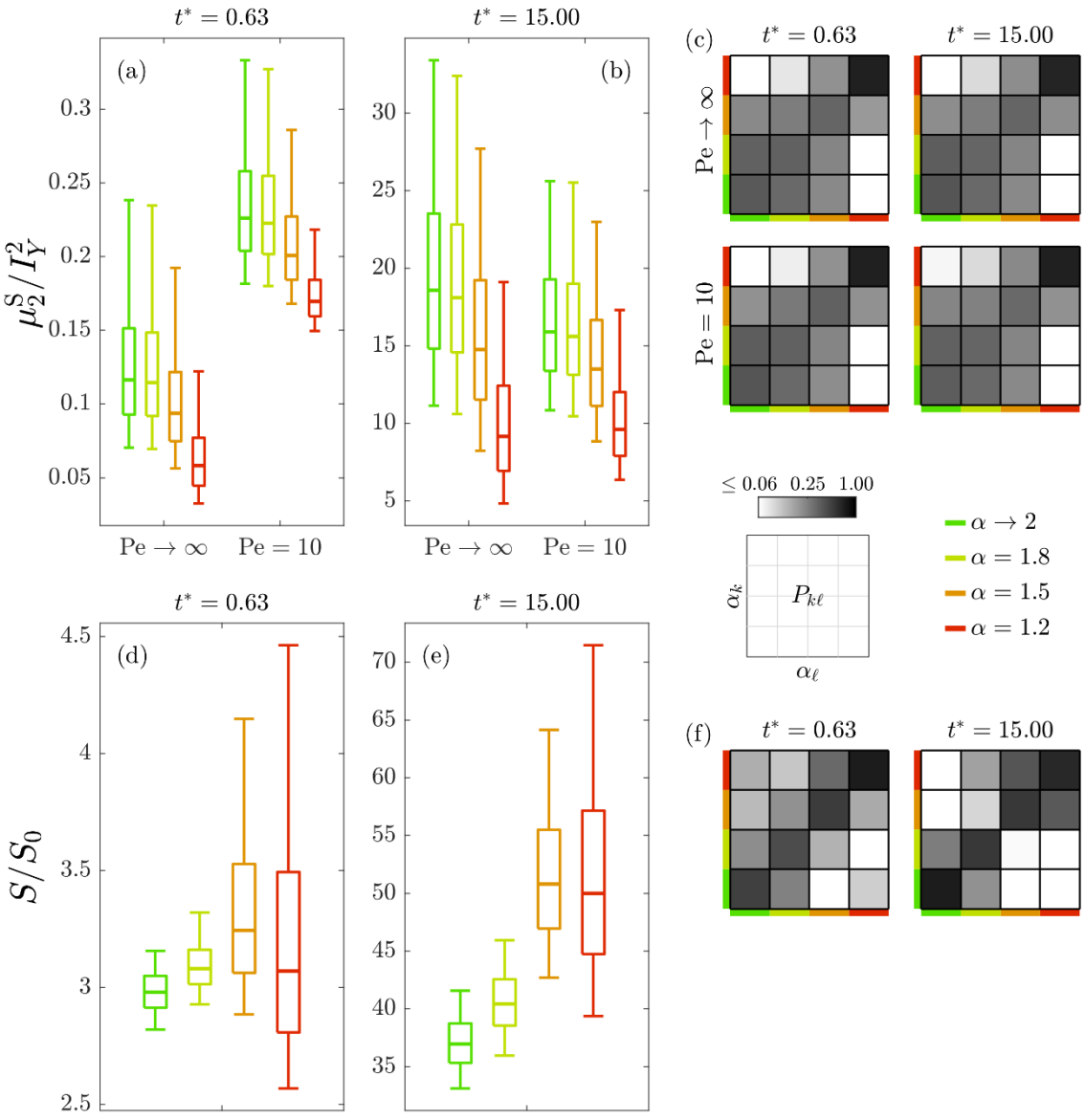


Figure 6: Box plot of (a)-(b) μ_2^S for $Pe \to \infty$ and Pe = 10; and (d)-(e) S/S_0 for $Pe \to \infty$ at an early and at a late time. The corresponding overlap quantification, $P_{k\ell}$, is also included (c, f).

Figure 6 depicts boxplots related to the populations of centered second moment of the section-integral concentrations μ_2^S (Figure 6(a-b)) and advective stretching S/S_0 (Figure 6(d-e)) for the various settings examined. These results are complemented by the graphical depiction of results stemming from the overlap analysis (see implementation details in Appendix C) for the two combinations of Pe (except for S/S_0, where Pe $\to \infty$

560 only) and various stages of plume development. Corresponding depictions related to other
561 selected transport metrics are presented in Figures 7 and 8.

562 Figure 6(a) suggests that the early-time ($t^* = 0.63$) distributions of values of μ_2^{S}
563 obtained for the analyzed values of α strongly differ for the two values of Pe considered.
564 This difference decreases as time increases. Otherwise, the degree of overlap between
565 distributions for diverse values of α, as quantified by the results in Figure 6(c), is not
566 significantly modified by local dispersion. The degree of overlap between the results for
567 Gaussian ($\alpha \to 2$) and slightly Sub-Gaussian ($\alpha = 1.8$) Y fields is markedly significant,
568 implying that the results of single realizations with $\alpha \to 2$ or $\alpha = 1.8$ are essentially
569 indistinguishable in terms of longitudinal dispersion. This tendency is documented for
570 most of the metrics analyzed. Results very similar to those corresponding to μ_2^{S} are
571 obtained in the analysis of the populations and overlaps of μ_2^{T} and T_{p99} (not shown here
572 for brevity). Figures 6(d) and 6(e) show that the range of values undertaken by the surface
573 growth due to hydrodynamic deformation, S/S_0, increases with the departure of Y from
574 Gaussianity, this metric being the most affected by slight differences in α. This is also
575 evidenced by the value of $P_{k\ell}$ in Figure 6(f). At late times ($t^* = 15$), there is only a slight
576 overlap between the moderately-to-strongly Sub-Gaussian cases ($\alpha \leq 1.5$) and the
577 Gaussian / slightly Sub-Gaussian cases ($\alpha \geq 1.8$).

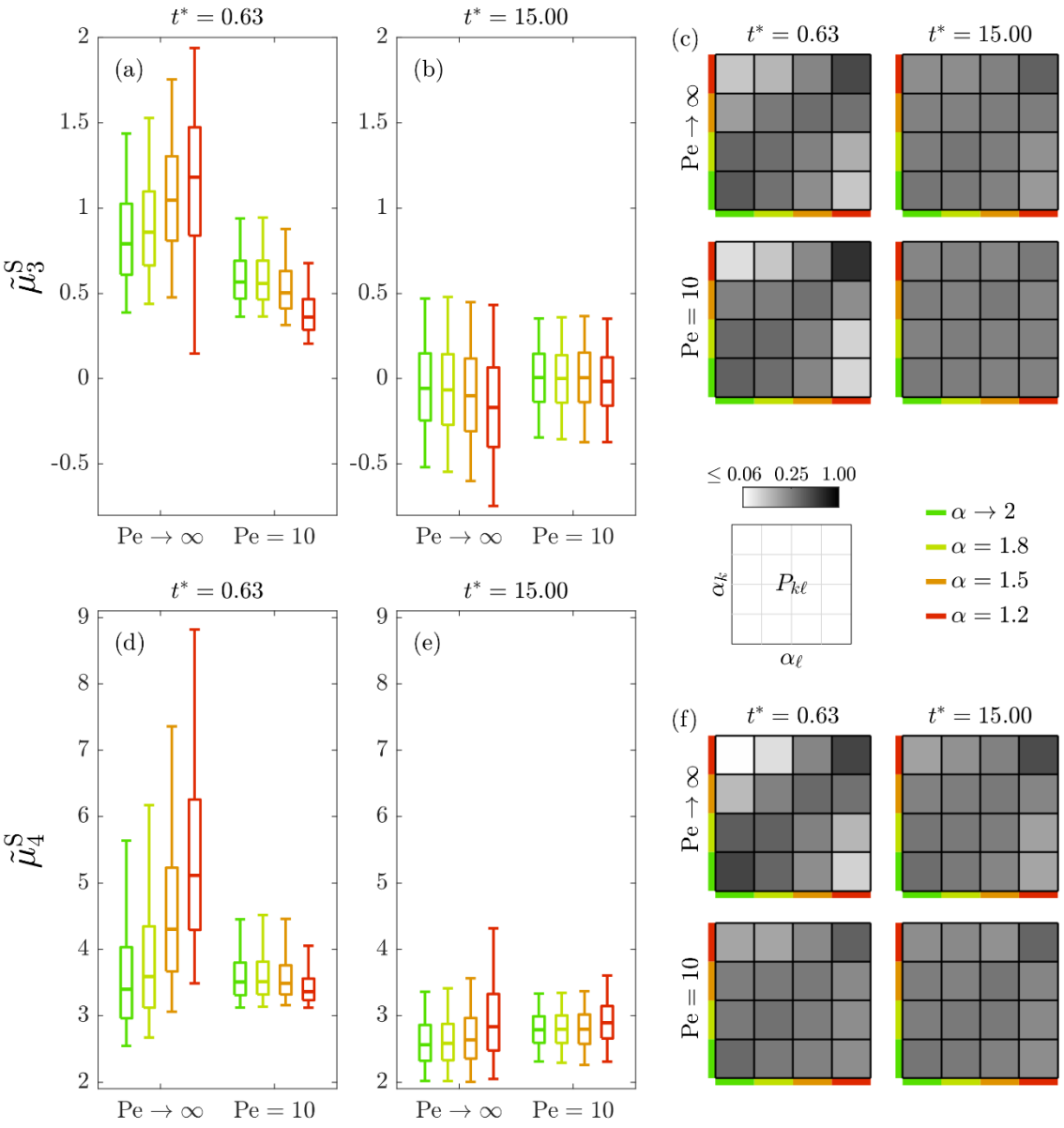


Figure 7: Box plot of (a)-(b) $\tilde{\mu}_3^S$ and (d)-(e) $\tilde{\mu}_4^S$ for $Pe \to \infty$ and Pe = 10 at an early and at a late time. The corresponding overlap quantification, $P_{k\ell}$, is also included (c, f).

Similar to what is observed for the centered second moment of the longitudinal concentration profile, a finite value of Pe narrows the range of values undertaken by the higher-order spatial moments across the collection of MC realizations (Figure 7). One can see that the distribution overlaps are significant (even for strongly Sub-Gaussian fields) for both skewness and kurtosis, especially at late times ($t^* = 15$) and for $\mathrm{Pe} = 10$. These results lead us to conclude that the average behavior of the higher-order moments

associated with diverse values of α which we document in Section 3.1.1 would most probably be undetectable from the analysis of the evolution of solute concentrations across individual realizations of the Y field. Similar conclusions are drawn from the analysis of the populations of higher-order temporal moments (not shown).

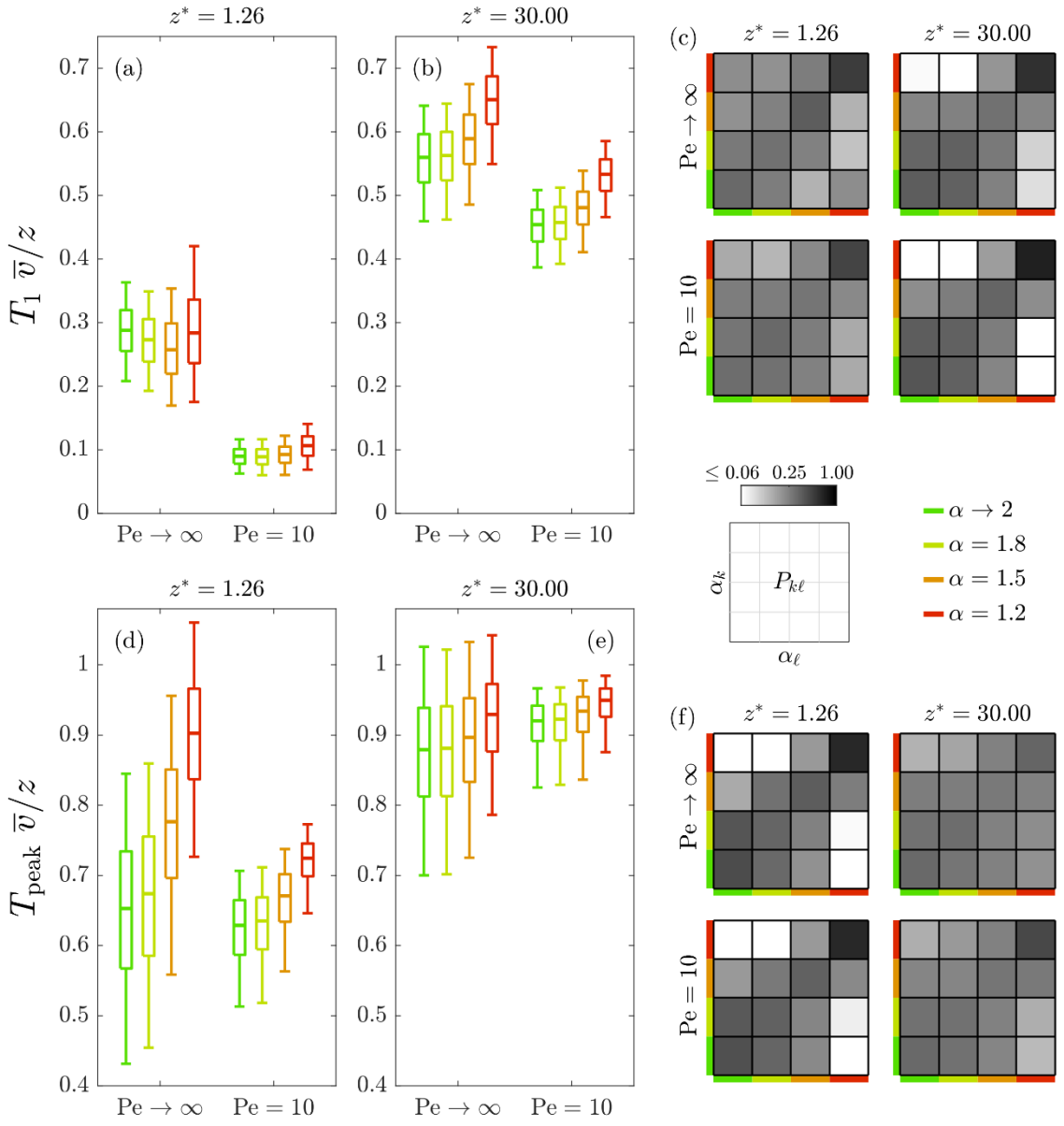
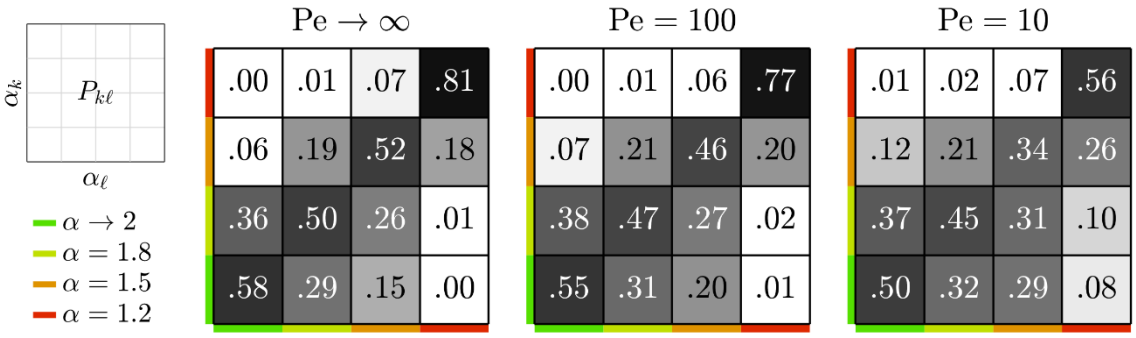


Figure 8: Box plot of (a)-(b) T_1 and (d)-(e) T_{peak} for $Pe \to \infty$ and Pe = 10 at a close and at a far distance from the inlet. The corresponding overlap quantification, $P_{k\ell}$, is also included (c, f).

Figure 8 depicts the boxplots and overlap quantification corresponding to the first arrival time (T_1) and the peak arrival time (T_{peak}). At close distance from the inlet ($z^* =$

598 1.26), the range of values of T_1 (i) narrows due to the effect of a finite Pe and (ii) is
599 virtually insensitive to the value of α. These results suggest that the behavior of T_1 is
600 chiefly dominated by local dispersion, this impact somewhat persisting for long travel
601 distances (i.e., $z^* = 30$). These results support the conjecture that, even as the overlap
602 with respect to the Gaussian case is not very high at long distances for moderate to strong
603 Sub-Gaussian Y fields, it would be hard to discriminate the effect of α from that of Pe if
604 both parameters are not known a priori. It is also worth noting that, while the time at
605 which the BTC peak occurs is (on average) closer to the mean travel time as α decreases
606 (Section 3.1), this tendency can hardly be detected at long distances from single
607 realizations. This is clearly seen in Figure 8(f), which shows that the overlap is nearly
608 total at $z^* = 30$.



609

610 *Figure 9: Overall overlap quantification, $P_{k\ell}$, between the distributions of the results*
611 *corresponding to various values of α and for three selected Péclet numbers. The*
612 *overlap is defined as $P_{k\ell} \coloneqq P(\alpha = \alpha_k | \hat{\alpha} = \alpha_\ell)$, where $\hat{\alpha}$ in this case is the result of a*
613 *Quadratic Discriminant Analysis (see Appendix C) combining the information embedded*
614 *in the evolution of all spatial and temporal moments as well as T_1, T_{peak}, and T_{p99}.*

615

616 Our analyses suggest that metrics that would typically be measurable in a real
617 setting, such as the (spatial or temporal) moments describing the evolution of the plume,
618 or the key arrival times, would not be able by themselves (i.e., without relying on
619 additional observations about the Y field) to reflect with certainty the Sub-Gaussian (or

620 Gaussian) nature of the heterogeneous porous medium across which solute migrates. This
621 is evidenced even assuming that σ_Y^2, I_Y, and Pe are known.

622 We further explore this aspect by analyzing the distributions of the results obtained
623 for all of these metrics jointly. We do so by relying on the classification technique known
624 as Quadratic Discriminant Analysis (QDA) to obtain an overall quantification of the
625 degree of differentiation of the sample distributions of the multiple metrics considered.
626 Details about the theoretical elements underpinning the approach are illustrated in
627 Appendix C. We ground QDA on the overall collections of results pertaining to spatial
628 and temporal moments (μ_2^{S}, $\tilde{\mu}_3^{\mathrm{S}}$, $\tilde{\mu}_4^{\mathrm{S}}$, μ_2^{T}, $\tilde{\mu}_3^{\mathrm{T}}$, $\tilde{\mu}_4^{\mathrm{T}}$), as well as the arrival times T_1, T_{peak} and
629 T_{p99}, and jointly considering results for $t^* = 0.63, 3.24, 15$ and $z^* = 1.26, 6.48, 30$, for
630 a total of 27 sample populations. We exclude S/S_0 from the analysis, as we focus on
631 solute transport metrics that could be detected directly in a real setting. This type of
632 analysis yields a global appraisal of our ability to discriminate results associated with a
633 unique realization through the assessment of an overlap matrix whose entries $P_{k\ell}$
634 represent the likelihood of the actual value of α being equal to α_k when the QDA
635 classification yields a value equal to α_ℓ. To broaden the range of scenarios, we perform
636 this analysis for $\mathrm{Pe} = 10$, 100, and $\mathrm{Pe} \to \infty$. The results are depicted in Figure 9. The
637 degree of differentiation between the overall behavior of individual realizations is visibly
638 higher than what one could infer by considering each of the metrics separately (Figures
639 6-8). Nevertheless, incorrect classifications are still abundant, especially for low values
640 of Pe. The results of our joint analysis of spatial moments, temporal moments and key
641 arrival times at multiple stages further support the observation that relying solely on
642 observations of concentrations and/or travel times for the considered sample size does not
643 allow to clearly and uniquely discriminate between values of α.

These results reinforce the conclusions drawn by Riva et al. (2017) and Libera et al. (2017), who found that manifestations of Y-field Sub-Gaussianity on moments/metrics associated with flow and transport tend to become virtually undetectable for $\alpha \geq 1.8$. This implies that (*i*) the Gaussian model may suffice to reproduce some observable transport features even when the nature of the underlying Y field is actually Sub-Gaussian, and (*ii*) it may be difficult in some cases to discern between non-Gaussian and Gaussian Y fields based solely on transport metrics. These conclusions are also in line with results obtained by Jankovic et al. (2017), who analyzed breakthrough curves resulting from transport of solute plumes through porous media with spatial distributions of log-conductivities relying on Gaussian and on various non-Gaussian statistical models. These authors concluded that the overall pattern of solute transport was relatively insensitive to the stochastic model employed, and that BTCs could be reasonably interpreted by relying solely on the knowledge of the mean, variance and integral scale of Y. As a consequence, we strengthen the concept that one is required to include a joint analysis of Y data and their increments (as suggested by Riva et al. (2015a) or Guadagnini et al. (2018)) in efforts aimed at heterogeneity characterization to detect possible signatures supporting an appropriate stochastic model representing the nature of the underlying Y field.

4. Conclusions

With reference to the assessment of the extent at which Sub-Gaussianity (i.e., the degree of departure from a Gaussian behavior) of Y can impact main features and observables of solute transport driven by advection and local dispersion in three-dimensional bounded heterogeneous porous media, we observe the following major elements.

- As compared to a Gaussian distribution of log-conductivities, a Generalized Sub-Gaussian (GSG) field of Y yields (on average) (*i*) a delayed first time of arrival of the solute mass to crossing planes located at a distance of a few integral scales from the solute injection plane; (*ii*) enhanced and delayed concentration peaks; and (*iii*) a reduced breakthrough time associated with the majority of the solute mass migrating in the system.
- On average, we document (*i*) an increasing degree of asymmetry and (*ii*) an enhanced kurtosis (resulting in heavier tails) of the plume with log-conductivity departing from a Gaussian behavior. Differences between the effects imprinted by Gaussian and strongly Sub-Gaussian ($\alpha \leq 1.5$) Y domains on spatial skewness and kurtosis of evolving plumes are marked at early times and tend to become mild as time increases.
- The velocity field associated with a GSG Y field promotes plume stretching while reducing longitudinal plume spreading compared to its Gaussian counterpart. Therefore, the increased longitudinal dispersion associated with Gaussian Y fields with respect to a GSG setting may not necessarily translate into enhanced solute mixing. This finding can have important implications on reactive transport; for example, in the context of mixing-limited reactive transport, it might imply that a higher mixing rate and reactivity can take place in Sub-Gaussian fields as compared to their Gaussian counterparts. Detailed analyses of this aspect will be subject of future research.

Concerning one's ability to clearly identify a possible underlying Sub-Gaussian nature of Y upon relying solely on the analysis of solute concentration fields in the settings analyzed, our study leads to the following major conclusions.

691 • Even as one has information about the values of variance and integral scale of Y,
692 the effects of a Sub-Gaussian nature of Y on hydrodynamic dispersion might
693 hardly be detectable in a (finite) heterogeneous setting, especially if characterized
694 by a low degree of Sub-Gaussianity (i.e., corresponding to $1.8 \leq \alpha < 2.0$ for the
695 type of subordinator considered). The action of local dispersion (i.e., finite values
696 of Pe) can then contribute to further mask the influence of Sub-Gaussianity on
697 major transport metrics, as also seen from a Quadratic Discriminant Analysis
698 performed on spatial and temporal moments and arrival times at various evolution
699 stages of the plume.
700 • While a GSG-based model can be consistent with scaling patterns of the
701 probability distribution of log-conductivity and its spatial increments at various
702 lags, relying solely on observations of solute concentrations migrating within the
703 system can in some cases hamper our ability to discern the nature of the
704 underlying conductivity field. To characterize the latter, one would need to
705 consider also a joint analysis of Y data and their increments.

706 We note that the present study focuses on mildly heterogeneous bounded domains
707 (representing a spatially heterogeneous single geological unit) and is mostly based on
708 spatially or temporally aggregated transport metrics. Additional studies are required to
709 fully assess the extent to which the emergence of strictly local features associated with
710 transport behavior (such as those related to strong channeling effects within highly
711 heterogeneous, virtually unbounded media) might contribute to distinguish GSG log-
712 conductivity domains from their Gaussian counterparts.

713 **Acknowledgements**

714 The authors would like to thank the EU, MIUR, and MINECO for funding, in the frame
715 of the collaborative international Consortium (WE-NEED) financed under the ERA-NET

716 WaterWorks2014 Cofunded Call. This ERA-NET is an integral part of the 2015 Joint
717 Activities developed by the Water Challenges for a Changing World Joint Programme
718 Initiative (Water JPI). Part of the work was developed while Prof. A. Guadagnini was at
719 the University of Strasbourg with funding from Région Grand-Est and Strasbourg-
720 Eurométropole through the *Chaire Gutenberg*. The UPC authors acknowledge funding
721 from AGAUR Research Groups, 2017 SGR 1485.

722 Data available on Zenodo (https://doi.org/10.5281/zenodo.3356708).

723 References

724 Bear, J. (1972). *Dynamics of fluids in porous media*. Elsevier, New York.

725 Bellin, A., Salandin, P., & Rinaldo, A. (1992). Simulation of dispersion in
726 heterogeneous porous formations: Statistics, first-order theories, convergence of
727 computations. *Water Resources Research*, *28*(9), 2211–2227.
728 https://doi.org/10.1029/92WR00578

729 Boffetta, G., Mazzino,A., & Vulpiani, A. (2008). Twenty-five years of multifractals in
730 fully developed turbulence: A tribute to Giovanni Paladin, *J. Phys. A Math. Theor.*,
731 41, 363001

732 Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via
733 diffusion. *Ann. Statist.*, *38*(5), 2916–2957. https://doi.org/10.1214/10-AOS799

734 Dagan, G. (1982). Stochastic modeling of groundwater flow by unconditional and
735 conditional probabilities: 2. The solute transport. *Water Resources Research*,
736 *18*(4), 835–848. https://doi.org/10.1029/WR018i004p00835

737 Dagan, G. (1984). Solute transport in heterogeneous porous formations. *Journal of*
738 *Fluid Mechanics*, *145*, 151–177. https://doi.org/10.1017/S0022112084002858

739 Dagan, G. (1989). *Solute transport in porous formations*. Springer, Berlin.

740 Dartois, A., Beaudoin, A., & Huberson, S. (2018). Impact of local diffusion on
741 macroscopic dispersion in three-dimensional porous media. *Comptes Rendus*
742 *Mécanique*, *346*(2), 89–97. https://doi.org/10.1016/j.crme.2017.12.012

743 de Dreuzy, J.-R., Méheust, Y., & Pichot, G. (2012). Influence of fracture scale
744 heterogeneity on the flow properties of three-dimensional discrete fracture
745 networks (DFN). *Journal of Geophysical Research: Solid Earth*, *117*(B11).
746 https://doi.org/10.1029/2012JB009461

747 Dentz, M., Kinzelbach, H., Attinger, S., & Kinzelbach, W. (2003). Numerical studies of
748 the transport behavior of a passive solute in a two-dimensional incompressible
749 random flow field. *Physical Review E - Statistical Physics, Plasmas, Fluids, and*
750 *Related Interdisciplinary Topics*. https://doi.org/10.1103/PhysRevE.67.046306

751 Dentz, M., Kinzelbach, H., Attinger, S., & Kinzelbach, W. (2000). Temporal behavior
752 of a solute cloud in a heterogeneous porous medium: 1. Point-like injection. *Water*
753 *Resources Research*, *36*(12), 3591–3604. https://doi.org/10.1029/2000WR900162

754 Dentz, M., Kinzelbach, H., Attinger, S., & Kinzelbach, W. (2002). Temporal behavior
755 of a solute cloud in a heterogeneous porous medium 3. Numerical simulations.
756 *Water Resources Research*, *38*(7), 13–23. https://doi.org/10.1029/2001WR000436

757 Dentz, M., Le Borgne, T., Englert, A., & Bijeljic, B. (2011). Mixing, spreading and
758 reaction in heterogeneous media: A brief review. *Journal of Contaminant*
759 *Hydrology*. https://doi.org/10.1016/j.jconhyd.2010.05.002

760 Fernàndez-Garcia, D., & Sanchez-Vila, X. (2011). Optimal reconstruction of
761 concentrations, gradients and reaction rates from particle distributions. *Journal of*
762 *Contaminant Hydrology*, *120–121*, 99–114.
763 https://doi.org/10.1016/j.jconhyd.2010.05.001

764 Fernàndez-Garcia, D., Illangasekare, T. H., & Rajaram, H. (2005). Differences in the
765 scale-dependence of dispersivity estimated from temporal and spatial moments in
766 chemically and physically heterogeneous porous media. *Advances in Water*
767 *Resources*, *28*(7), 745–759. https://doi.org/10.1016/j.advwatres.2004.12.011

768 Fogg, G. E., & Zhang, Y. (2016). Debates—Stochastic subsurface hydrology from
769 theory to practice: A geologic perspective. *Water Resources Research*, *52*(12),
770 9235–9245. https://doi.org/10.1002/2016WR019699

771 Ganti, V., Singh, A.., Passalacqua, P., & Foufoula-Georgiu, E. (2009). Subordinated
772 Brownian motion model for sediment transport, *Phys. Rev*. E, 80, 011111,
773 doi:1539-5663755/2009/80(1)/011111(9)

774 Gelhar, L. W. (1993). *Stochastic Subsurface Hydrology*. Prentice-Hall.

775 Gist, G. A., Thompson, A. H., Katz, A. J., & Higgins, R. L. (1990). Hydrodynamic
776 dispersion and pore geometry in consolidated rock. *Physics of Fluids A: Fluid*
777 *Dynamics*, *2*(9), 1533–1544. https://doi.org/10.1063/1.857602

778 Gómez-Hernández, J. J., & Wen, X.-H. (1998). To be or not to be multi-Gaussian? A
779 reflection on stochastic hydrogeology. *Advances in Water Resources*, *21*(1), 47–61.
780 https://doi.org/10.1016/S0309-1708(96)00031-0

781 Guadagnini, A., Riva, M., & Neuman, S.P. (2012). Extended power-law scaling of
782 heavy-tailed random air-permeability fields in fractured and sedimentary rocks,
783 *Hydrol. Earth Syst. Sci*., 16, 3249-3260, doi:10.5194/hess-16-3249-2012

784 Guadagnini, A., Riva, M., & Neuman, S. P. (2018). Recent advances in scalable non-
785 Gaussian geostatistics: The generalized sub-Gaussian model. *Journal of*
786 *Hydrology*, *562*, 685–691. https://doi.org/10.1016/j.jhydrol.2018.05.001

787 Guadagnini, A., Neuman, S.P., Schaap, M.G., & Riva, M. (2013). Anisotropic statistical
788 scaling of vadose zone hydraulic property estimates near Maricopa, Arizona, *Water*
789 *Resour. Res*., 49, doi:10.1002/2013WR014286

790 Guadagnini, A., S.P. Neuman, M.G. Schaap, and M. Riva (2014), Anisotropic
791 Statistical Scaling of Soil and Sediment Texture in a Stratified Deep Vadose Zone
792 near Maricopa, Arizona, Geoderma, 214-215, 217-227,
793 doi:10.1016/j.geoderma.2013.09.008.

794 Harbaugh, A.W., 2005. MODFLOW-2005, The U.S. Geological Survey Modular
795 Ground-Water Model - the Ground-Water Flow Process. U.S. Geological Survey
796 Techniques and Methods 6-A16

797 James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical*
798 *Learning: With Applications in R*. Springer Publishing Company, Incorporated.

799 Jankovic, I., Maghrebi, M., Fiori, A., & Dagan, G. (2017). When good statistical models
800 of aquifer heterogeneity go right: The impact of aquifer permeability structures on
801 3D flow and transport. *Advances in Water Resources*, *100*, 199–211.
802 https://doi.org/10.1016/J.ADVWATRES.2016.10.024

803 Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications in Hydrogeology*.
804 Cambridge University Press. https://doi.org/10.1017/CBO9780511626166

805 LaBolle, E.M., Quastel, J.,Fogg, G.E., & Gravner, J. (2000). Diffusion processes in
806 composite porous media and their numerical integration by random walks:
807 Generalized stochastic differential equations with discontinuous coefficients,
808 *Water Resour. Res*., 36, 3, doi: 10.1029/1999WR900224

809 Le Borgne, T., Dentz, M., & E. Villemaux (2013). Stretching, coalescence, and mixing
810 in porous media. *Physical Review Letters*, 110, 20,
811 doi:10.1103/PhysRevLett.110.204501

812 Le Borgne, T., Dentz, M., & E. Villemaux (2015). The lamellar description of mixing in
813 porous media. *Journal of Fluid Mechanics*, 770, 458-498,
814 doi:10.1017/jfm.2015.117

815 Libera, A., de Barros, F. P. J., Riva, M., & Guadagnini, A. (2017). Solute concentration
816 at a well in non-Gaussian aquifers under constant and time-varying pumping
817 schedule. *Journal of Contaminant Hydrology*, *205*, 37–46.
818 https://doi.org/https://doi.org/10.1016/j.jconhyd.2017.08.006

819 Liu, H. H., & Molz, F. J. (1997). Comment on "Evidence for non-Gaussian scaling
820 behavior in heterogeneous sedimentary formations" by Scott Painter. *Water*
821 *Resources Research*, *33*(4), 907–908. https://doi.org/10.1029/96WR03788

822 Liu, X., & Müller, H.-G. (2004). Functional Convex Averaging and Synchronization for
823 Time-Warped Random Curves. *Journal of the American Statistical Association*,
824 *99*(467), 687–699. https://doi.org/10.1198/016214504000000999

825 Lu, Z., & Stauffer, P. H. (2012). On estimating functional average breakthrough curve
826 using time-warping technique and perturbation approach. *Water Resources*
827 *Research*, *48*(5). https://doi.org/10.1029/2011WR011506

828 Meerschaert, M. M., T. J. Kozubowski, F. J. Molz, and S. Lu (2004), Fractional Laplace
829 model for hydraulic conductivity, Geophys. Res. Lett., 31, L08501,
830 doi:10.1029/2003GL019320

831 Naff, R. L., Haley, D. F., & Sudicky, E. A. (1998a). High-resolution Monte Carlo
832 simulation of flow and conservative transport in heterogeneous porous media: 1.
833 Methodology and flow results. *Water Resources Research*, *34*(4), 663–677.
834 https://doi.org/10.1029/97WR02712

835 Naff, R. L., Haley, D. F., & Sudicky, E. A. (1998b). High-resolution Monte Carlo
836 simulation of flow and conservative transport in heterogeneous porous media: 2.
837 Transport results. *Water Resources Research*, *34*(4), 679–697.
838 https://doi.org/10.1029/97WR02711

839 Painter, S. (1996). Evidence for Non-Gaussian Scaling Behavior in Heterogeneous
840 Sedimentary Formations. *Water Resources Research*, *32*(5), 1183–1195.
841 https://doi.org/10.1029/96WR00286

842 Painter, S. (2001), Flexible scaling model for use in random field simulation of
843 hydraulic conductivity, *Water Resources Research*, *37*, 1155–1163.
844 https://doi.org/10.1029/2000WR900394

845 Panzeri, M., Riva, M., Guadagnini, A., & Neuman, S. P. (2016). Theory and generation
846 of conditional, scalable sub-Gaussian random fields. *Water Resources Research*,
847 *52*(3), 1746–1761. https://doi.org/10.1002/2015WR018348

848 Pedretti, D., Fernàndez-Garcia, D., Bolster, D., & Sanchez-Vila, X. (2013). On the
849 formation of breakthrough curves tailing during convergent flow tracer tests in
850 three-dimensional heterogeneous aquifers. *Water Resources Research*, *49*(7),
851 4157–4173. https://doi.org/10.1002/wrcr.20330

852 Pedretti, D., & Fernàndez-Garcia, D. (2013). An automatic locally-adaptive method to
853 estimate heavily-tailed breakthrough curves from particle distributions. *Advances*
854 *in Water Resources*, *59*, 52–65. https://doi.org/10.1016/j.advwatres.2013.05.006

855 Pedretti, D., Masetti, M., & Beretta, G. Pietro. (2017). Stochastic analysis of the
856 efficiency of coupled hydraulic-physical barriers to contain solute plumes in highly
857 heterogeneous aquifers. *Journal of Hydrology*, *553*, 805–815.
858 https://doi.org/10.1016/j.jhydrol.2017.08.051

859 Riva, M., Guadagnini, A., Fernàndez-Garcia, D., Sanchez-Vila, X., & Ptak, T. (2008).
860 Relative importance of geostatistical and transport models in describing heavily

tailed breakthrough curves at the Lauswiesen site. *Journal of Contaminant Hydrology*, *101*(1–4), 1–13. https://doi.org/10.1016/j.jconhyd.2008.07.004

Riva, M., Guadagnini, A., & Neuman, S. P. (2017). Theoretical analysis of non-Gaussian heterogeneity effects on subsurface flow and transport. *Water Resources Research*, *53*(4), 2998–3012. https://doi.org/10.1002/2016WR019353

Riva, M., Neuman, S. P., & Guadagnini, A. (2015a). New scaling model for variables and increments with heavy-tailed distributions. *Water Resources Research*, *51*(6), 4623–4634. https://doi.org/10.1002/2015WR016998

Riva, M., Panzeri, M., Guadagnini, A., & Neuman, S. P. (2015b). Simulation and analysis of scalable non-Gaussian statistically anisotropic random functions. *Journal of Hydrology*, *531*, 88–95. https://doi.org/10.1016/J.JHYDROL.2015.06.066

Riva, M., Neuman, S. P., & Guadagnini, A. (2013a). Sub-Gaussian model of processes with heavy-tailed distributions applied to air permeabilities of fractured tuff. *Stochastic Environmental Research and Risk Assessment*, *27*(1), 195–207. https://doi.org/10.1007/s00477-012-0576-y

Riva, M., Neuman, S. P., Guadagnini, A., & Siena, M. (2013b). Anisotropic Scaling of Berea Sandstone Log Air Permeability Statistics. *Vadose Zone Journal*, *12*. https://doi.org/10.2136/vzj2012.0153

Rizzo, C. B., & de Barros, F. P. J. (2017). Minimum Hydraulic Resistance and Least Resistance Path in Heterogeneous Porous Media. *Water Resources Research*, *53*(10), 8596–8613. https://doi.org/10.1002/2017WR020418

Rubin, Y. (1990a). Stochastic modeling of macrodispersion in heterogeneous porous media. *Water Resources Research*, *26*(1), 133–141. https://doi.org/10.1029/WR026i001p00133

Rubin, Y. (1990b). Correction to "Stochastic modeling of macrodispersion in heterogeneous porous media" by Yoram Rubin. *Water Resources Research*, *26*(10), 2631. https://doi.org/10.1029/WR026i010p02631

Rubin, Y. (2003). *Applied stochastic hydrogeology*. New York : Oxford University Press.

Salamon, P., Fernàndez-Garcia, D., & Gómez-Hernández, J. J. (2006). A review and numerical assessment of the random walk particle tracking method. *Journal of Contaminant Hydrology*, *87*(3–4), 277–305. https://doi.org/10.1016/j.jconhyd.2006.05.005

Siena, M., Guadagnini, A., Riva, M., & Neuman, S. P. (2012). Extended power-law scaling of air permeabilities measured on a block of tuff. *Hydrology and Earth System Sciences*, *16*(1), 29–42. https://doi.org/10.5194/hess-16-29-2012

Siena, M., & Riva, M. (2018). Groundwater withdrawal in randomly heterogeneous coastal aquifers. *Hydrology and Earth System Sciences*, *22*(5), 2971–2985. https://doi.org/10.5194/hess-22-2971-2018

Siena, M., Riva, M., Giamberini, M., Gouze, P., & Guadagnini, A. (2019). Statistical modeling of gas-permeability spatial variability along a limestone core*, Spatial Statistics*, *39*, 100249, doi: 10.1016/j.spasta.2017.07.007.

Siena, M., Guadagnini, A., Bouissonnié, A., Ackerer, P., Daval, D., & Riva, M. (2020). Generalized Sub-Gaussian processes: theory and application to hydrogeological and geochemical data, *Water Resources Research*, WRCR24714, doi:10.1029/2020WR027436

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. *ChapMan & HALL/CRC* (Vol. 37). https://doi.org/10.2307/2347507

Sole-Mari, G., & Fernàndez-Garcia, D. (2018). Lagrangian Modeling of Reactive Transport in Heterogeneous Porous Media With an Automatic Locally Adaptive Particle Support Volume. *Water Resources Research*, *54*(10). https://doi.org/10.1029/2018WR023033

Sole-Mari, G., Fernàndez-Garcia, D., Rodríguez-Escales, P., & Sanchez-Vila, X. (2017). A KDE-Based Random Walk Method for Modeling Reactive Transport With Complex Kinetics in Porous Media. *Water Resources Research*, *53*(11). https://doi.org/10.1002/2017WR021064

Sole-Mari, G., Bolster, D., Fernàndez-Garcia, D., & Sanchez-Vila, X. (2019). Particle Density Estimation with Grid-Projected and Boundary-Corrected Adaptive Kernels. *Advances in Water Resources*, 103382. https://doi.org/10.1016/j.advwatres.2019.103382

Winter, C. L., Tartakovsky, D. M., & Guadagnini, A. (2003). Moment Differential Equations for Flow in Highly Heterogeneous Porous Media. *Surveys in Geophysics*, *24*(1), 81–106. https://doi.org/10.1023/A:1022277418570

Winter, C. L., & Tartakovsky, D. M. (2000). Mean Flow in composite porous media. *Geophysical Research Letters*, *27*(12), 1759–1762. https://doi.org/10.1029/1999GL011030

Winter, C. L., & Tartakovsky, D. M. (2002). Groundwater flow in heterogeneous composite aquifers. *Water Resources Research*, *38*(8), 11–23. https://doi.org/10.1029/2001WR000450

Appendix A: Analytical formulation of the GSG model and key properties

Riva et al. (2015a,b) introduce the Generalized Sub-Gaussian (GSG) model according to which the random function $Y'(\mathbf{x})$ defined by (1) is described considering the

subordinator $U(\mathbf{x})$ to be lognormally distributed according to $\ln(U) \sim \mathcal{N}[0, (2-\alpha)^2]$, i.e.,

$$f_U(u) = \frac{e^{-\frac{(\ln u)^2}{2\,(2-\alpha)^2}}}{\sqrt{2\pi}u(2-\alpha)}, \tag{A1}$$

with $\alpha < 2$. Here, we include a brief summary of the key analytical formulations and scaling properties of the isotropic GSG model introduced by Riva et al (2015a). Corresponding formulations for anisotropic fields have been derived by Riva et al (2015b).

The marginal *pdf* of $Y'(\mathbf{x})$, $f_{Y'}(y')$, is fully defined by α and the standard deviation of $G(\mathbf{x})$, σ_G, as

$$f_{Y'}(y') = \frac{1}{2\pi\sigma_G(2-\alpha)} \int_0^\infty e^{-\frac{1}{2}\left[\left(\frac{ln\,u}{2-\alpha}\right)^2 + \left(\frac{y}{\sigma_G u}\right)^2\right]} \frac{du}{u^2}. \tag{A2}$$

The latter corresponds to a normal-lognormal (NLN) distribution whose variance, σ_Y^2, and standardized kurtosis, $\kappa_{Y'}$, of $Y'(\mathbf{x})$ are given by

$$\sigma_Y^2 = \sigma_G^2 e^{2(2-\alpha)^2}, \qquad \kappa_{Y'} = 3e^{4(2-\alpha)^2}. \tag{A3}$$

It is noted that the lognormal distribution (A1) tends to a delta function when $\alpha \to 2$ and (A2) reduces to the Gaussian distribution. Figure A1 depicts $f_{Y'}$ for the three values of α examined in our study and by setting $\sigma_Y^2 = 1$. The Gaussian *pdf* having the same mean and variance as $Y'(\mathbf{x})$ and corresponding to $\alpha \to 2$ is also depicted for completeness. While $f_{Y'}$ is close to the Gaussian *pdf* for $\alpha = 1.8$, one can observe that $f_{Y'}$ exhibits peaks and tails which become sharp and long, respectively, as α decreases. These features, which represent a clear deviation from a Gaussian distribution, is also quantified by the excess kurtosis, $EK_{Y'} = \kappa_{Y'} - 3$, whose value increases (deviating from zero, which corresponds to the scenario for which $\alpha \to 2$) as α decreases.

The *pdf* of incremental values of $Y'(\mathbf{x})$, $\Delta Y = Y'(\mathbf{x}) - Y'(\mathbf{x}+\mathbf{s})$, is given by

$$f_{\Delta Y}(\Delta y) = \frac{1}{2\pi\sigma_G(2-\alpha)^2\sqrt{2\pi}}\int_0^\infty\int_0^\infty e^{-\frac{1}{2}\left[\frac{ln^2 u_1 + ln^2 u_2}{(2-\alpha)^2}+\frac{\Delta y^2}{\sigma_G^2 r^2}\right]}\frac{du_2 du_1}{u_2 u_1 r}. \quad \text{(A4)}$$

Here, $r = \sqrt{u_1^2 + u_2^2 - 2\rho_G u_1 u_2}$ and $\rho_G(s)$ is the correlation function of $G(\mathbf{x})$ evaluated at separation distance (or lag) $s = |s|$. The standardized kurtosis of ΔY is rendered as

$$\kappa_{\Delta Y} = 3e^{2(2-\alpha)^2}\left[1 + \frac{1}{2}\left(\frac{e^{2(2-\alpha)^2} - 1}{e^{(2-\alpha)^2} - \rho_G}\right)^2\right]. \quad \text{(A5)}$$

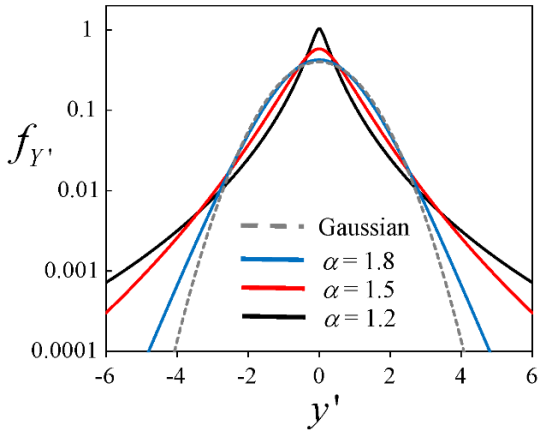
It is remarked that the dependence of $\rho_G(s)$ on lag induces a corresponding dependence of the shape of $f_{\Delta Y}$. For example, $\kappa_{\Delta Y} \to 3$ and the distribution of ΔY tends to the Gaussian one as $\alpha \to 2$. Otherwise, the shape of $f_{\Delta Y}$ scales with lag (or, equivalently, with ρ_G) with peak which tends to sharpen and tails to become heavier (i.e., corresponding to increased $\kappa_{\Delta Y}$) as a function of ρ_G. Figure A2 depicts $f_{\Delta Y}$ for three exemplary lags and α = 1.8 (Fig. A2a), 1.5 (Fig. A2b), and 1.2 (Fig. A2c). All GSG fields are characterized by the same variance ($\sigma_Y^2 = 1$) and integral scale ($I_Y = I_G e^{-(2-\alpha)^2}$, I_G being the integral scale of $G(\mathbf{x})$), and lags are normalized with respect to I_Y. Also shown for comparison are Gaussian distributions having the same variance as ΔY. In all cases, $f_{\Delta Y}$ exhibits sharp peaks and heavy tails at small lags. Increasing lag, $f_{\Delta Y}$ tends to become Gaussian when α is large (i.e., α = 1.8), otherwise remaining heavy tailed for smaller values of α. This feature is also illustrated in Figure A3, depicting excess kurtosis of Y' and ΔY versus lag for the three values of α considered in our study. The excess kurtosis of ΔY, $EK_{\Delta Y} = \kappa_{\Delta Y} - 3$, decreases as lag increases, rendering the peak of $f_{\Delta Y}$ less sharp and its tails lighter. When α = 1.8, $EK_{\Delta Y}$ is seen to tend to an asymptotic value which is << 1, so that $f_{\Delta Y}$ becomes virtually Gaussian. Included in Figure A3 are horizontal lines depicting the excess of kurtosis of Y', $EK_{Y'}$. In all investigated cases, $EK_{Y'} > EK_{Y'}$ at small lags, i.e. $f_{\Delta Y}$ has sharper peaks and heavier tails than does $f_{Y'}$, the opposite being true at large lags.

The variogram, γ_Y, of Y', is given by

$$\gamma_Y = \sigma_G^2 e^{(2-\alpha)^2}\left(e^{(2-\alpha)^2} - 1\right) + e^{(2-\alpha)^2}\gamma_G, \quad \text{(A6)}$$

983 γ_G being the variogram of $G(\mathbf{x})$. Note that according to (A6) γ_Y is discontinuous at the
984 origin, i.e., at $s = 0$, thus exhibiting a nugget effect. Figure A4 compares (on arithmetic
985 and logarithm scales) γ_Y computed with three values of α (= 1.2, 1.5, and 1.8) and setting
986 $\sigma_Y^2 = 1$ and $I_Y = 1$. Also shown for comparison is the variogram obtained within a
987 Gaussian field (i.e., corresponding to $\alpha \to 2$) characterized by the same variance and
988 integral scale as Y' and by an exponential correlation function. The variogram associated
989 with $\alpha = 1.8$ virtually coincides with its counterpart associated with a Gaussian field
990 everywhere, with the exception of small lags where the GSG variogram is characterized
991 by a nugget (clearly visible on logarithm scale).

992



993

994 *Figure A1: Probability density function of Y' for $\sigma_Y^2 = 1$ and three values of α*
995 *(solid curves). Also shown is the Gaussian pdf having the same mean and variance as Y'*
996 *(dashed curve).*

997

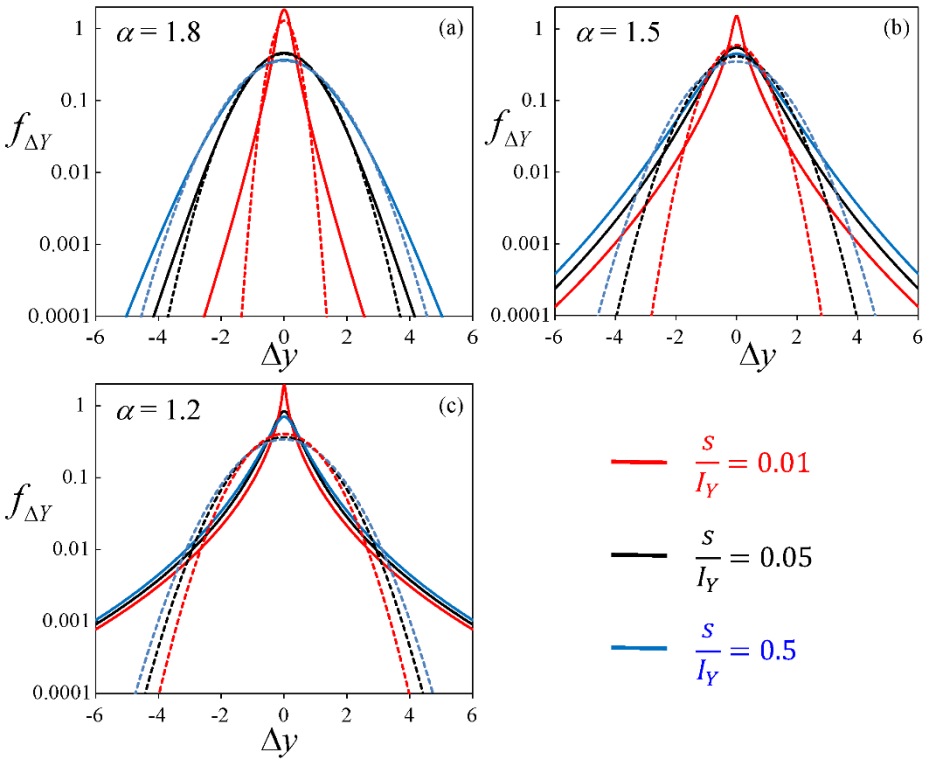


Figure A2: Probability density function of ΔY for $\sigma_Y^2 = 1$, three lags and (a) $\alpha = 1.8$, (b) 1.5, and (c) 1.2 (solid curves). Also shown are Gaussian pdfs having the same variance as ΔY (dashed curves). Lags are normalized with respect to I_Y.

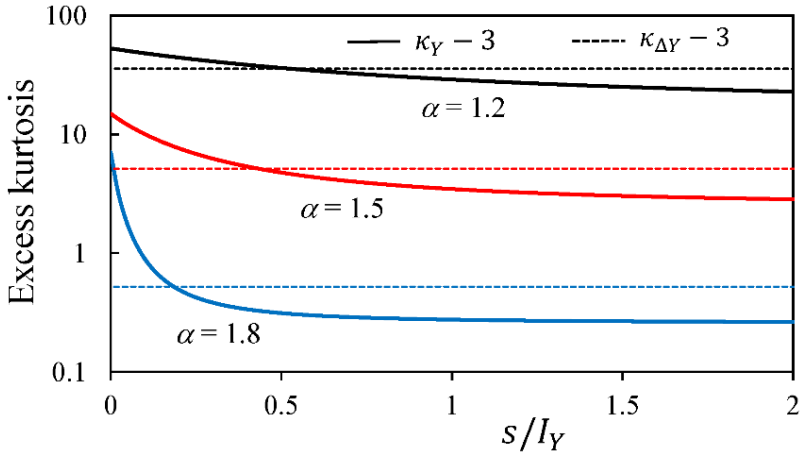


Figure A3: Excess kurtosis of Y′ (dashed horizontal lines) and ΔY (solid curves) versus normalized lag for $\alpha = 1.2$, 1.5, and 1.8.

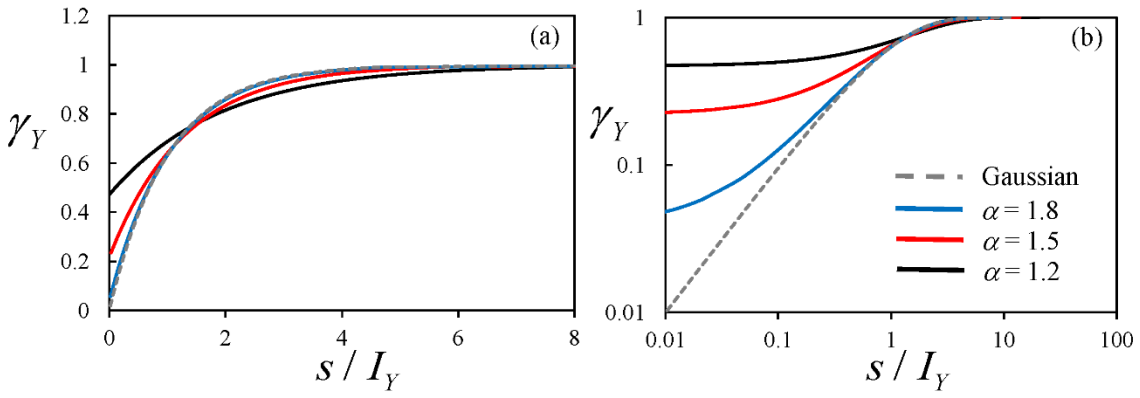


Figure A4: Variogram of Y' obtained for α = 1.2, 1.5, and 1.8 on (a) arithmetic and (b) logarithmic scales (solid curves). Also shown for comparison the variogram associated with a Gaussian field (and corresponding to $\alpha \to 2$) characterized by the same variance and integral scale as Y' (dashed curve).

Appendix B: Transport Metrics

B.1. Percentile average of concentration curves

We obtain percentile-averaged curves over all MC realizations of (*i*) section-integral longitudinal profiles of concentration at a given observation time and (*ii*) temporal distributions of mass flux through a given horizontal plane (or BTCs), here denoted as $\bar{C}$ and $\bar{J}$, respectively. Following the terminology used by Lu and Stauffer (2012) for BTCs, the percentile average results from averaging over the times of the cumulative distribution of $\bar{J}$ at fixed percentiles. By analogy, here we also extend this technique to longitudinal concentration profiles.

The first step of the procedure is to rearrange in ascending order the collection of particle positions / arrival times corresponding to each realization, i.e.,

$$Z_{p,r}(t_i) < Z_{p+1,r}(t_i), \qquad \forall p \in \{1,2,\dots,(n-1)\}, \tag{B1}$$

$$T_{p,r}(z_j) < T_{p+1,r}(z_j), \qquad \forall p \in \{1,2,\dots,(n-1)\}, \tag{B2}$$

By computing the arithmetic average of these quantities across MC realizations as in (11), each particle index p is associated with a value of $\langle Z_p(t_i)\rangle$ or $\langle T_p(z_j)\rangle$. According to the above-mentioned definition of percentile average, these values correspond to regular increments in the cumulative (i.e., integral) forms of $\bar{C}(z,t_i)$ and $\bar{J}(z_j,t)$, i.e.,

$$\int_0^z \bar{C}(\zeta,t_i)\mathrm{d}\zeta = \frac{1}{n}\sum_{p=1}^{n}\mathcal{H}\left(\langle Z_p(t_i)\rangle - z\right), \tag{B3}$$

$$\int_0^t \bar{J}(z_j,\tau)\mathrm{d}\tau = \frac{1}{n}\sum_{p=1}^{n}\mathcal{H}\left(\langle T_p(z_j)\rangle - t\right), \tag{B4}$$

where $\mathcal{H}$ is the Heaviside function. Differentiation of (B3) and (B4) yields

$$\bar{C}(z,t_i) = \frac{1}{n}\sum_{p=1}^{n}\delta\left(\langle Z_p(t_i)\rangle - z\right), \tag{B5}$$

$$\bar{J}(z_j,t) = \frac{1}{n}\sum_{p=1}^{n}\delta\left(\langle T_p(z_j)\rangle - t\right), \tag{B6}$$

where δ is the Dirac delta. Since n is a finite number, these expressions need to be smoothed to produce adequate estimates. This is accomplished by replacing δ in (B5) - (B6) with a smoothing kernel W characterized by a smoothing bandwidth h. This constitutes the basis of Kernel Density Estimation (KDE), which is summarized in Appendix D.

B.2. Moments

We study the evolution of the spatial and temporal moments of the section-integral concentrations (C_r) and the breakthrough curves (J_r), respectively, in each realization r, as rendered through

$$\mu_{1,r}^{\mathrm{S}}(t_i) = \int_0^{L_z} z\, C_r(z,t_i)\mathrm{d}z, \qquad \mu_{\kappa,r}^{\mathrm{S}}(t) = \int_0^{L_z} \left(z - \mu_1^{\mathrm{S}}(t_i)\right)^{\kappa} C_r(z,t_i)\mathrm{d}z, \qquad \kappa \geq 2, \text{(B7)}$$

$$\mu_{1,r}^{\mathrm{T}}(z_j) = \int_0^{\infty} t\, J_r(z_j,t)\mathrm{d}t, \qquad \mu_{\kappa,r}^{\mathrm{T}}(z) = \int_0^{\infty} \left(t - \mu_1^{\mathrm{T}}(z_j)\right)^{\kappa} J_r(z_j,t)\mathrm{d}t, \qquad \kappa \geq 2, \text{(B8)}$$

where κ is the moment order, and superscripts S and T stand for *spatial* and *temporal*, respectively. In practice, integrals in (B7) and (B8) are estimated directly from the discrete particle positions and arrival times, respectively, as

$$\mu_{1,r}^{\mathrm{S}}(t_i) = \frac{1}{n}\sum_{p=1}^{n} Z_{p,r}(t_i), \qquad \mu_{\kappa,r}^{\mathrm{S}}(t_i) = \frac{1}{n}\sum_{p=1}^{n}\left[Z_{p,r}(t_i) - \mu_{1,r}^{\mathrm{S}}(t_i)\right]^{\kappa}, \qquad \kappa \geq 2, \quad \text{(B9)}$$

$$\mu_{1,r}^{\mathrm{T}}(z_j) = \frac{1}{n}\sum_{p=1}^{n} T_{p,r}(z_j), \qquad \mu_{\kappa,r}^{\mathrm{T}}(z_j) = \frac{1}{n}\sum_{p=1}^{n}\left[T_{p,r}(z_j) - \mu_{1,r}^{\mathrm{T}}(z_j)\right]^{\kappa}, \qquad \kappa \geq 2. \text{(B10)}$$

We also define the standardized moments (i.e., skewness for $\kappa = 3$, kurtosis for $\kappa = 4$)

$$\tilde{\mu}_{\kappa,r}^{\mathrm{S}}(t_i) = \frac{\mu_{\kappa,r}^{\mathrm{S}}(t_i)}{\left[\mu_{2,r}^{\mathrm{S}}(t_i)\right]^{\frac{\kappa}{2}}}, \qquad \tilde{\mu}_{k,r}^{\mathrm{T}}(z_j) = \frac{\mu_{\kappa,r}^{\mathrm{T}}(z_j)}{\left[\mu_{2,r}^{\mathrm{T}}(z_j)\right]^{\frac{\kappa}{2}}}, \qquad \kappa \geq 3. \qquad \text{(B11)}$$

We noted that temporal moments computed in our study from the full set of particle arrival times (see (B10)) were mainly controlled by the tails of the associated distributions. BTCs in heterogeneous porous media often display heavy tails that decay slowly with a power-law behavior (e.g., Pedretti et al., 2013). Even truncated power laws may have very high values of higher-order moments that would hardly be accurately inferred from simulations with a reasonable number of particles. To overcome this issue, temporal moments were computed by excluding the 1% highest arrival times.

All realizations were checked to fulfill the following expressions, corresponding to a flux-weighted injection, for the spatial and temporal first moments at all evaluation times t_i and all evaluation positions z_j, respectively:

$$\frac{\partial \mu_{1,r}^{\mathrm{S}}}{\partial t}(t_i) \geq \bar{v}, \qquad \mu_{1,r}^{\mathrm{T}}(z_j) = \frac{z_j}{\bar{v}}. \tag{B12}$$

B.3. Key arrival times

As a complement to temporal moments, we evaluate some additional BTC features that may provide further insight on transport behavior. Namely, we compute: the first arrival time ($T_{1,r}$), the arrival time corresponding to the peak of the corresponding concentration curve ($T_{\mathrm{peak},r}$) and the 99$^{\mathrm{th}}$ percentile arrival time ($T_{\mathrm{p99},r}$). The peak is identified in each realization r from the KDE reconstruction of the BTC (see Appendix D), whereas the two other measures are obtained directly from particle arrival times.

B.4. Advective stretching

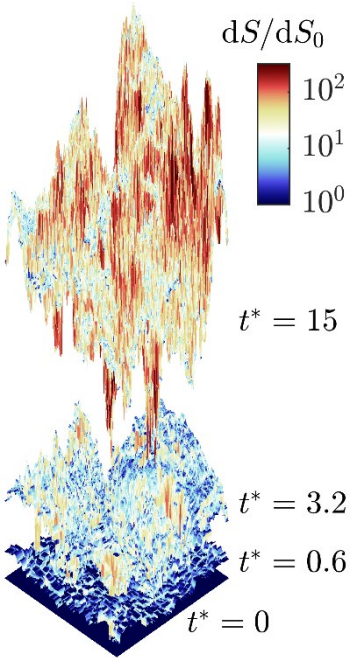


Figure B1. Graphical depiction of the evolving shape of a material plane under heterogeneous advection, and mapping of the local surface growth.

We quantify hydrodynamic deformation (advective stretching) by tracking a mesh of triangular surfaces defined by the initial relative position of particles in the scenario corresponding to $\mathrm{Pe} \to \infty$. Figure B1 illustrates the local growth of the individual triangular surfaces $\mathrm{d}S$ at three observation times in a selected realization of Y. In Section 3 we focus on the ratio S/S_0, where S is the sum of all triangular $\mathrm{d}S$, and $S_0 = L_x L_y$.

Appendix C: Quantification of Distribution Overlaps

Here, we provide a synthesis of the basic concepts of Bayesian classification, focusing on the methods we employ in our study.

C.1. The Bayes classifier

The Bayes classifier assigns observations of a (multi-dimensional random) vector $\mathbf{R}$, with values $\mathbf{r}$, to a class K (which has a finite number N_k of possible values k) according to data-driven estimates of the conditional probability density functions $f_k(\mathbf{r}) = p(\mathbf{R} = \mathbf{r}|K = k)$. Ideally, it is the classification method with the least associated error provided that the true *pdf*s, f_k, are known or reproduced correctly. Given an observation with a value $\mathbf{r}$, this method assigns it to the category k for which $g_k(\mathbf{r}) = P(K = k|\mathbf{R} = \mathbf{r})$ is highest among all possible k. According to Bayes theorem

$$g_k(\mathbf{r}) = \frac{p_k f_k(\mathbf{r})}{\sum_{\ell=1}^{N_\mathrm{k}} p_\ell f_\ell(\mathbf{r})}, \tag{C1}$$

where $p_k = P(K = k)$ is the prior probability of any observation to belong to class k. If no prior information is available, all $p_\ell = p_k$ ($\forall \ell \neq k$), and the classifier will assign the observation $\mathbf{r}$ to the class k for which f_k is highest. If we assume that all f_k are Gaussian, with mean $\widehat{\mathbf{m}}_k$ and covariance matrix $\widehat{\mathbf{\Sigma}}_k$ estimated from data, then the classification is equivalent to assigning observation $\mathbf{r}$ to the class k for which

$$\delta_k(\mathbf{r}) = -(\mathbf{r} - \widehat{\mathbf{m}}_k)^\mathrm{T} \widehat{\mathbf{\Sigma}}_k^{-1} (\mathbf{r} - \widehat{\mathbf{m}}_k) - \log\left|\widehat{\mathbf{\Sigma}}_k\right| + 2\log p_k \tag{C2}$$

is highest. This approach is known as Quadratic Discriminant Analysis (QDA).

For one-dimensional variables, R, it may in some cases be feasible to reconstruct $f_k(r)$ from data, instead of assuming that it is Gaussian. We do so in our study upon relying on Kernel Density Estimation (see Appendix D). Each observation r is then assigned to the class k with the highest $g_k(r)$ (see (C1)). We term this approach as *KDE discriminant*.

C.2. Confusion matrix and Overlap matrix

In discriminant analysis, the Confusion Matrix compares the obtained classification (rows) with the actual classes (columns) of a number of random observations, so that the entries on the diagonal of the matrix correspond to correct classifications. Normalization of the columns by the total number of observations of each class yields a normalized confusion matrix. The latter informs us on the likelihood of correct (diagonal) and incorrect (outside diagonal) classification. In other words, if we define variable $\widehat{K}$ as the estimated class (and K as the true class), the entries of the normalized confusion matrix are $P'_{\ell k} = P\left(\widehat{K} = \ell \middle| K = k\right)$.

We define the overlap of a *pdf,* f_k, (conditioned to $K = k$) with f_ℓ (conditioned to $K = \ell$) as the likelihood that, if the class given by the discriminant analysis, $\widehat{K}$, is ℓ, the actual class, K, is k; i.e., $P_{k\ell} = P\left(K = k \middle| \widehat{K} = \ell\right)$. The quantity $P_{k\ell}$ (and in particular, its values for $k = \ell$) indicates the level of reliability associated with relying on the *pdfs* to estimate the actual class of an individual realization. Mapping $P'_{\ell k}$ onto $P_{k\ell}$ is accomplished through Bayes theorem

$$P_{k\ell} = \frac{p_k P'_{\ell k}}{\sum_{m=1}^{N_k} p_m P'_{\ell m}} \tag{C3}$$

C.3. Implementation

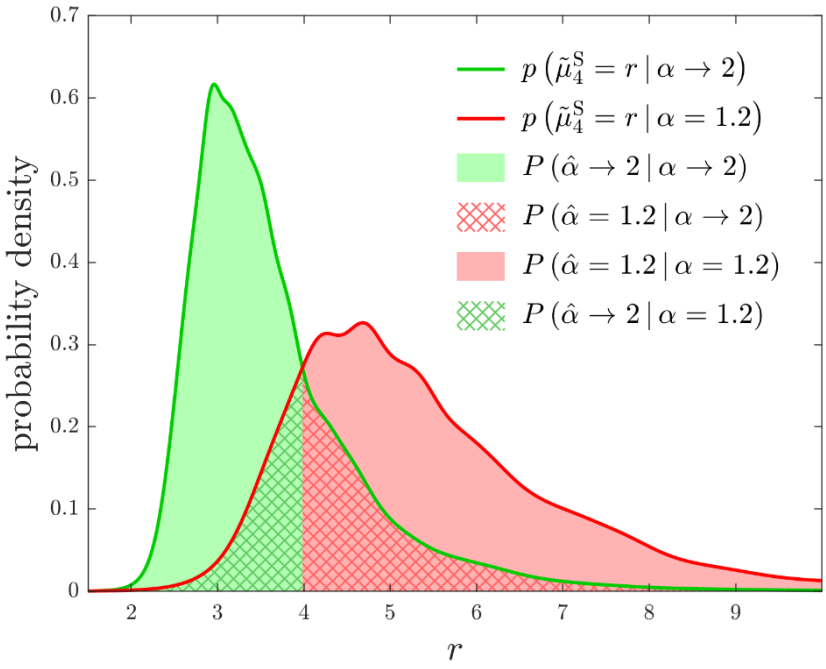


Figure C1: Graphical depiction exemplifying the integration procedure for the evaluation of the $P'_{\ell k}$ entries of the normalized confusion matrix using the KDE classifier for the early-time spatial kurtosis.

In our study, the classes are the four studied values of α, and the random observations are the selected metrics evaluated from the solute transport results. We use the KDE discriminant for the analysis of each of these metrics, whereas we apply QDA for the joint analysis of several metrics. In both cases, the final result of interest is the overlap matrix defined in Section C.2, obtained from the normalized confusion matrix by (C3). In the case of the KDE discriminant, the entries of the normalized confusion matrix can be obtained by numerical integration of the relevant parts of the *pdf*s (see Figure C1 as a graphical example), whereas the confusion matrix for the QDA is obtained by explicitly classifying all observations.

Note that the KDE approach and the QDA approach require a training dataset (for the KDE reconstruction of the *pdf* in the former and for the estimation of the mean and

covariance in the latter). Relying on the same dataset for training and for testing (i.e., for building the confusion matrix) could slightly bias the results towards correct classification (that is, towards a non-overlap result). However, we expect this effect to be negligible given the rather large size of the dataset (i.e., 5000 observations of each class).

Appendix D: Kernel Density Estimation

Here, we provide a brief introduction to univariate Kernel Density Estimation (KDE), which we use to reconstruct curves, $f(x)$, which are (or can be treated as) probability density functions, relying on a limited number of samples X_p, $p = 1, \dots, n$. The KDE is formulated as

$$f(x) \cong \hat{f}(x) = \frac{1}{n}\sum_{p=1}^{n} W\left(X_p - x; h\right), \tag{D1}$$

where W is a kernel or smoothing function, which we choose to be Gaussian, i.e.,

$$W(u; h) = \frac{1}{\sqrt{2\pi}h}\exp\left(-\frac{u^2}{2h^2}\right), \tag{D2}$$

h being the kernel bandwidth.

When X_p are particle longitudinal positions or arrival times, $f(x)$ is the longitudinal density of solute mass (8) or the breakthrough curve (9), respectively (Fernàndez-Garcia and Sànchez-Vila, 2011; Sole-Mari et al., 2017, 2019; Sole-Mari and Fernàndez-Garcia, 2018; Pedretti and Fernàndez-Garcia, 2013). Alternatively, the collection of samples X_p can be formed by evaluations of a scalar metric that varies across MC realizations and in this case $f(x)$ is the probability density function of the metric x (see Appendix C).

A key aspect of the accuracy of KDE is to select an optimal value for the bandwidth h. A widely popular technique employed for this aim relies on minimizing the Asymptotic

Mean Integrated Squared Error (AMISE). It can be shown that, for a Gaussian W, in the limit when $nh \to \infty$ and $h \to 0$ (e.g. Silverman, 1986)

$$\mathrm{AMISE} \approx \left(2\sqrt{\pi}nh\right)^{-1} + \frac{h^4}{4}\int_{\mathbb{R}} \left(\frac{\partial^2 f}{\partial x^2}\right)^2 dx\,. \tag{D3}$$

By setting $\frac{\partial \mathrm{AMISE}}{\partial h} = 0$, the absolute minimum AMISE can be shown to be obtained for the following choice of bandwidth

$$h = \left[2\sqrt{\pi}\, n \int_{\mathbb{R}} \left(\frac{\partial^2 f}{\partial x^2}\right)^2 \mathrm{d}x\right]^{-\frac{1}{5}}. \tag{D4}$$

Since $\int_{\mathbb{R}} \left(\frac{\partial^2 f}{\partial x^2}\right)^2 \mathrm{d}x$ is *a priori* unknown, several iterative methodologies have been developed to estimate it from the data and solve equation (D4). Here, we apply one such methodology, namely the Improved Sheater-Jones (ISJ) algorithm developed by Botev et al. (2010).