Research

Author for correspondence:
M. Mousavi Nezhad
e-mail: m.mousavi-nezhad@warwick.ac.uk

Estimation of auto-covariance of log hydraulic conductivity from Generalized Sub-Gaussian porosity and particle size random fields

M. Harrison[1], M. Riva[3], M. Mousavi Nezhad[2,4] and A. Guadagnini[3]

[1]HetSys EPSRC Centre for Doctoral Training (CDT), and [2]School of Engineering, University of Warwick, Coventry, UK
[3]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy
[4]Department of Civil and Environmental Engineering, School of Engineering, University of Liverpool, Liverpool, UK

(iD) MMN, 0000-0002-0625-439X

We derive analytical formulations relating the spatial covariance (C_Y) of (log-transformed) hydraulic conductivities to auto- and cross-covariances of porosity (ϕ) and representative soil particle sizes within the framework of the classical Terzaghi model. The latter provides an empirical relationship which is widely used to obtain conductivity estimates. We frame the study within recent stochastic approaches and conceptualize appropriate transformations of ϕ and representative soil particle size as Generalized Sub-Gaussian (GSG) spatially cross-correlated random processes. Consistency of the theoretical framework against sample distributions of ϕ and particle size is assessed through the analysis of field data. A perturbation-based approach yields workable expressions of C_Y upon truncating the otherwise exact analytical solution at given orders of approximations. Our analytical (truncated) log-conductivity covariance is in agreement with its Monte Carlo-based counterpart. A Global Sensitivity Analysis relying on classical Sobol indices quantifies the relative importance of all parameters embedded in the formulation of C_Y. We show that parameters driving the GSG nature of the distribution of

THE ROYAL SOCIETY
PUBLISHING

(transformed) porosity are key to the main features of C_Y. We also document the relevance of properly capturing emergences of possible cross-correlations between ϕ and representative particle size to reconstruct conductivity fields.

1. Introduction

The assessment of key hydrogeological features associated with groundwater flow hinges on a reliable characterization of the spatial distribution of attributes of subsurface geomaterials/ lithotypes. These include, most notably, hydraulic conductivity (K). The latter can be considered as a parameter of major importance in a variety of branches of engineering and earth science. It is recognized to display a heterogeneous spatial distribution across the subsurface. Uncertainties associated with the quantification of the spatial variability of hydraulic conductivity inevitably propagate to model outputs. In this context, stochastic approaches are typically employed to assess subsurface system behaviour under uncertainty. Obtaining heterogeneous spatial distributions of K is challenging as it is hampered by the typically limited availability of direct observations. The latter are often obtained through interpretation of hydraulic tests. These might have an additional difficulty of being related to ill-defined observation scales. The associated data are usually complemented by other qualitative or quantitative information. These include, e.g. sedimentological information stemming from descriptive analyses of observations taken along exploration boreholes and/or particle size distributions. Depending on the type of available data, various stochastic approaches can be applied to estimate heterogeneous K fields across the domain of interest. These include the use of, e.g. (i) block classification of geomaterials through the use of facies delineation techniques, from which K values can then be assigned to each subdomain identified within the system [e.g. [1–3]]; (ii) (semi)empirical relationships relating geophysical data to K [e.g. [4,5]]; (iii) numerical generation (e.g. in a Monte Carlo context) of collections of (spatially random) K fields that are coherent with structural geologic patterns related to underlying sedimentological processes [e.g. [6–8]] or (iv) (two- or multi-point) geostatistics including data of various natures [e.g. [9,10]].

In this broad context, we note that hydrogeological quantities such as conductivity and/or porosity have been traditionally modelled as Gaussian random fields. Otherwise, there is strong evidence (e.g. [11] and references therein) that the probability density function (PDF) associated with spatial increments of several hydrogeological quantities (including, e.g. log-conductivity, porosity or electrical resistivity) change with the separation distance (or lag) at which increments are evaluated. Density functions of increments are documented to be characterized by distinct features that result in sharp peaks and heavy tails at short lags. These then tend to change (i.e. peaks decrease and tails become thinner) with increasing separation distances between locations at which increments are taken. Such a scaling feature results in increment density functions evolving towards Gaussian behaviour at large lags. Such an observation is also reflected by the behaviour of the kurtosis of increment PDFs, which tends to a value typical of a Gaussian distribution as lag increases. Note that the documented scaling of moments of increment PDFs is also consistent, for example, with the well-known observation that the variance of the population of increments, which is directly related to spatial correlation through the concept of variogram, evolves with lag.

It is then clear that the assumption of Gaussianity for a given variable of interest is not (in general) compatible with the simultaneous occurrence of the aforementioned characteristics of statistical scaling displayed by the sample probability distribution (and main statistical moments) of incremental values. The Generalized Sub-Gaussian (GSG) model is introduced by Riva *et al.* [12] to seamlessly capture all of these traits in a unique theoretical framework. The traditional approach that relies on the use of a Gaussian representation of the variable of interest is included in such a framework as a particular case. The GSG approach has been documented to provide a robust and flexible framework within which the spatial variability of a variety

of hydrogeological, geological, geophysical and Earth science quantities can be described (e.g. [11,13–15], and references therein).

Hydraulic conductivity estimates are routinely provided through empirical formulations. These typically include information related to Particle Size Curves (PSCs) and/or porosity. A variety of methods can be employed to evaluate PSCs from soil samples (e.g. laser diffraction; dry and wet sieving). We recall that soil particle size analyses are simple and enable one to characterize soil make-ups in terms of a number of effective grain diameters. These types of models were first introduced by Seelheim [16] and Hazen [17], who proposed a correlation between K and representative grain sizes. The reader is referred to [18] for a comprehensive review of existing empirical models relating K to information content embedded in PSCs. In addition, Rosas et $al.$ [18] assess the accuracy of these empirical models by analysing a dataset comprising more than 400 sediment samples. Relying on these formulations enables one to map spatial random fields of PSC representative grain sizes and/or porosity onto ensuing K fields. Uncertainty associated with key features of a PSC can then be propagated onto corresponding K values. In this context, Riva et $al.$ [19] introduce analytical formulations relating the first- and second-order (statistical) moments (i.e. mean and spatial covariance) of hydraulic conductivity to those describing representative particle diameters related to PSCs upon resting on the Beyer [20] formulation. They also include the influence of system porosity on these (statistical) moments, as described through the Kozeny-Carman [21] relationship. These authors then exemplify their results through a field application relying mainly on the Beyer formulation, due to the lack of direct porosity data at the investigated site. While noting that comprehensive datasets including porosity and particle size curves collected at the same location are not typically available in the literature, we also recall that a key advantage of available empirical formulations is that they are conducive to estimates of K at a moderate experimental cost and with a very limited computational effort.

Here, we extend the focus of Riva et $al.$ [19]. Key novelty elements of this work include the development of a set of original analytical formulations quantifying the relationship between the covariance of (natural) log-conductivity, representative soil particle sizes, and porosity. Our solutions are specifically designed to address common situations when the distribution of these variables does not follow a Gaussian pattern. The latter is included in our solutions as a particular case. We do so upon considering the Terzaghi empirical formulation [22]. The latter is widely used to infer estimates of hydraulic conductivity in aquifer systems whose main make-up comprises large grain sands [18]. Similar formulations available in the literature, relating hydraulic conductivity to d_{10} (i.e. the representative particle size diameter expressed in terms of percent in mass corresponding to the 10th percentile of a measured PSC) and porosity of a given soil sample, can be readily employed upon following the procedure we outline.

Consistent with the works of [11,12], we model appropriate functional transformations of d_{10} and porosity as GSG random quantities. Doing so enables us to imprint the traits typical of a GSG setting onto the main statistics driving the (heterogeneous) spatial distribution of conductivity. These descriptors can then be used in a typical estimation and/or (stochastic) simulation framework. After deriving analytical formulations that establish relationships between the characteristics of transformed porosity and d_{10} (GSG) fields with the log-conductivity covariance, a rigorous global sensitivity analysis (GSA) is performed. In line with the observation that (non-Gaussian) GSG traits of conductivities have a marked importance in subsurface flow and transport (e.g. [23,24]), GSA results enable one to ascertain the relative importance of the model parameters in shaping the spatial distribution of the log-conductivity field.

The study is structured as follows. In §2, we illustrate the theoretical basis underpinning the evaluation of the mean and covariance of log-conductivity on the basis of the aforementioned Terzaghi formulation. We present exact formulations as well as workable approximations upon relying on classical perturbation methods. Section 3 summarizes the analytical formulations for a bivariate GSG distribution, including the cross-covariance between GSG random quantities.

Table 1. List of main symbols.

symbol	refers to
a	dimensionless coefficient (see (5.1)); its value depends on the cross-correlation between B and C
α_ξ	shape parameter of the GSG model (with $\xi = L, B, C$)
B	function of porosity (see (2.3))
C	natural logarithm of d_{10}
C_ξ	covariance of ξ (with $\xi = Y, L, C$)
C_{BC}	cross-covariance between B and C
$C_Y^{[i]}$	analytical covariance of Y evaluated up to order i (see (2.12))
$C_Y^{(i)}$	contribution of order i to the analytical covariance of Y, (see (2.13)–(2.14))
d_{10}	characteristic particle size diameter
$f_{U_\xi}(u_\xi)$	PDF of U_ξ (with $\xi = L, B, C$)
$f_{G_B G_C}$	bivariate Gaussian PDF of (G_B, G_C)
$f_{B'C'}$	bivariate non-Gaussian PDF of (B', C')
G_ξ	Gaussian field underlying the GSG model (see (3.1), with $\xi = L, B, C$)
GSG	Generalized Sub-Gaussian model
h	separation distance or lag (i.e. $|\boldsymbol{x} - \boldsymbol{y}|$)
I_ξ	integral scale of ξ (with $\xi = Y, G_B, G_C, G_{BC}$)
K	hydraulic conductivity
l_ξ	correlation scale of ξ (with $\xi = G_L, G_B, G_C, G_{BC}$)
μ_B	mean of B
Y	natural logarithm of K
$\langle \xi \rangle$	mean of ξ (with $\xi = Y, L, B, C$)
ξ'	perturbation of ξ (with $\xi = Y, L, B, C$)
ϕ	porosity
PDF	probability density function
PSCs	particle size curves
σ_ξ	standard deviation of ξ (with $\xi = Y, L, B, C, G_B, G_C, G_L$)
ρ_ξ	correlation coefficient of G_ξ (with $\xi = Y, G_B, G_C, G_L, G_{BC}$)
U_ξ	subordinator of the GSG model (see (3.1) with $\xi = L, B, C$)
$\boldsymbol{x}$	space coordinates
$\langle \rangle$	ensemble mean operator

Section 4 introduces the extensive database we consider and the analyses we perform to assess the suitability of a GSG theoretical framework in interpreting the main traits emerging from this. Section 5 is devoted to the assessment of the reliability of our (perturbation-based) analytical formulations for various orders of approximations. To ensure transparent comparisons, we do so upon considering a suite of numerical Monte Carlo simulations which we take as reference. Finally, §6 provides a rigorous GSA aimed at assessing the relative importance of the key parameters embedded in the GSG modelling framework describing the fields of characteristic grain size of the particle size curve and porosity on the ensuing log-conductivity covariance. The key parameters and acronyms used in this study are listed in table 1.

2. Theoretical developments

The empirical formulation proposed by Terzaghi [22] yields hydraulic conductivity, K [m s^{-1}], at a given location x in space according to

$$K(x) = \frac{\beta g}{v} \left(\frac{\phi(x) - 0.13}{\sqrt[3]{[1 - \phi(x)]}} \right)^2 d_{10}^2(x). \tag{2.1}$$

Here, β [–] is a dimensionless quantity, typically ranging between $[6.1 - 10.3] \times 10^{-3}$ (its actual value should be estimated through calibration against available site-specific K data); g [m s^{-2}] corresponds to gravity; v [m^2 s^{-1}] is kinematic viscosity; ϕ [–] and d_{10} [mm] are porosity and the characteristic particle size diameter, respectively. Values of d_{10} are commonly obtained, e.g. from grain sieve analysis according to standardized protocols. Making use of (2.1), one can evaluate the natural logarithm of K, $Y(x) = \ln K(x)$, as

$$Y(x) = \gamma + 2 \ln[B(x) + 1] + 2C(x), \tag{2.2}$$

where $\gamma = \ln(\beta g / v)$, $C(x) = \ln[d_{10}(x)]$ and

$$B(x) = -1 + \frac{\phi(x) - 0.13}{\sqrt[3]{[1 - \phi(x)]}}. \tag{2.3}$$

Expanding the second term on the right-hand side of (2.2) in a Taylor series, assuming $|B(x)| < 1$, yields

$$\ln[B(x) + 1] = -\sum_{k=1}^{\infty} \frac{(-1)^k}{k} B^k(x). \tag{2.4}$$

Note that $|B(x)| < 1$ implies $0.13 < \phi < 0.93$, thus encompassing a wide range of porosity values and variety of practical applications, including scenarios associated with the type of geomaterials for which the Terzaghi formulation is most appropriate [18]. Making use of (2.1)–(2.4), log-conductivity can be expressed as

$$Y(x) = \gamma - 2 \sum_{k=1}^{\infty} \frac{(-1)^k}{k} B^k(x) + 2C(x). \tag{2.5}$$

The (ensemble) mean of $Y(x)$, $\langle Y(x) \rangle$, can be evaluated from (2.5) as

$$\langle Y(x) \rangle = \gamma - 2 \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \langle B^k(x) \rangle + 2 \langle C(x) \rangle. \tag{2.6}$$

Here and in the following, $\langle \rangle$ denotes the ensemble mean operator. From (2.5)–(2.6), $Y(x)$ perturbations around the mean, $Y'(x) = Y(x) - \langle Y(x) \rangle$, are evaluated as

$$Y'(x) = 2C'(x) - 2 \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \left(B^k(x) - \langle B^k(x) \rangle \right), \tag{2.7}$$

where $C'(x) = C(x) - \langle C(x) \rangle$. Making use of the binomial identity, one can write

$$\begin{aligned} B^k(x) - \langle B^k(x) \rangle &= (\langle B(x) \rangle + B'(x))^k - \left\langle (\langle B(x) \rangle + B'(x))^k \right\rangle \\ &= \sum_{p=0}^{k} \binom{k}{p} \langle B(x) \rangle^{k-p} B'^p(x) - \sum_{p=0}^{k} \binom{k}{p} \langle B(x) \rangle^{k-p} \langle B'^p(x) \rangle \\ &= \sum_{p=0}^{k} \binom{k}{p} \langle B(x) \rangle^{k-p} \left(B'^p(x) - \langle B'^p(x) \rangle \right). \end{aligned} \tag{2.8}$$

Substituting (2.8) into (2.7) leads to

$$Y'(x) = 2C'(x) - 2\sum_{k=1}^{\infty} \frac{(-1)^k}{k} \sum_{p=0}^{k} \binom{k}{p} \langle B(x)\rangle^{k-p} \left(B'^p(x) - \langle B'^p(x)\rangle\right).$$ (2.9)

One can then note that perturbations (or fluctuations) of the log-conductivity field depend (i) linearly on the perturbation of the $C(x)$ field (that is linked to d_{10}) and (ii) supra-linearly (up to order k) on the perturbation terms related to the B field (that is linked to porosity). Determination of $Y'(x)$ in (2.9) enables us to evaluate the covariance of the log-conductivity field, $C_Y(x, y) = \langle Y'(x)Y'(y)\rangle$, as

$$\frac{C_Y(x, y)}{4} = C_C(x, y) - \sum_{k=1}^{\infty} \frac{(-1)^k}{k} \sum_{p=0}^{k} \binom{k}{p} \langle B(x)\rangle^{k-p} \langle B'^p(x)C'(y)\rangle$$

$$- \sum_{j=1}^{\infty} \frac{(-1)^j}{j} \sum_{m=0}^{j} \binom{j}{m} \langle B(y)\rangle^{j-m} \langle C'(x)B'^m(y)\rangle + \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{(-1)^k}{k}\frac{(-1)^j}{j}$$

$$\times \sum_{p=0}^{k} \binom{k}{p} \sum_{m=0}^{j} \binom{j}{m} \langle B(x)\rangle^{k-p} \langle B(y)\rangle^{j-m} \left[\langle B'^p(x)B'^m(y)\rangle - \langle B'^p(x)\rangle\langle B'^m(y)\rangle\right],$$ (2.10)

where $C_C(x, y) = \langle C'(x)C'(y)\rangle$ is the covariance of the C field. In the following, we focus on statistically stationary random fields whereby (2.10) simplifies as

$$\frac{C_Y(h)}{4} = C_C(h) - 2\sum_{k=1}^{\infty} \frac{(-1)^k}{k} \sum_{p=1}^{k} \binom{k}{p} \mu_B^{k-p} \langle B'^p(x)C'(y)\rangle$$

$$+ \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{(-1)^{k+j}}{kj} \sum_{p=1}^{k} \binom{k}{p} \sum_{m=1}^{j} \binom{j}{m} \mu_B^{k-p+j-m} \left[\langle B'^p(x)B'^m(y)\rangle - \langle B'^p(x)\rangle\langle B'^m(x)\rangle\right],$$ (2.11)

where $h = |x - y|$ and $\mu_B = \langle B\rangle$. As shown in (2.11), C_Y depends on (i) the covariance of the C field, (ii) the expected value of the product between B'^p and C' (involving a perturbation term of order $p + 1$) and (iii) the expected value of the product between B'^p and B'^m (involving a perturbation term of order $p + m$). By incorporating terms denoted by superscripts p and m into (2.11), it becomes apparent that the covariance of the (log-transformed) hydraulic conductivity field is influenced by higher-order statistical moments that involve grain size distribution and porosity. One can then evaluate (2.11) to an order i of approximation according to

$$C_Y(h) \approx C_Y^{[i]}(h) = \sum_{n=2}^{i} C_Y^{(n)}(h).$$ (2.12)

Here, the superscript $[i]$ indicates that terms up to order i are included in the approximation. The lowest order contribution to C_Y is the second-order term, $C_Y^{(2)}$. The latter is evaluated from (2.11) as

$$\frac{C_Y^{(2)}(h)}{4} = C_C(h) - 2\langle B'(x)C'(y)\rangle \sum_{k=1}^{\infty} (-1)^k \mu_B^{k-1}$$

$$+ \langle B'(x)B'(y)\rangle \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} (-1)^{k+j} \mu_B^{k+j-2}$$

$$- \langle B'^2(x)\rangle^2 \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} (-1)^{k+j} \frac{(k-1)}{2}\frac{(j-1)}{2} \mu_B^{k+j-4}.$$ (2.13)

Considering (2.11), higher-order contributions ($C_Y^{(n)}$, with $n > 2$) to C_Y are then given by

$$
\begin{aligned}
\frac{C_Y^{(n)}(h)}{4} = &-2\langle B'^{n-1}(x)C'(y)\rangle \sum_{k=1}^{\infty} \frac{(-1)^k}{k}\binom{k}{n-1}\mu_B^{k+1-n} \\
&+ \sum_{k=1}^{\infty}\sum_{j=1}^{\infty} \frac{(-1)^{k+j}}{kj}\mu_B^{k+j-n}\left\{\sum_{\omega=1}^{n-1}\binom{k}{\omega}\binom{j}{n-\omega}\langle B'^{\omega}(x)B'^{n-\omega}(y)\rangle - \langle B'^n(x)\rangle\right. \\
&\left.\times\left[\binom{j}{n}\sum_{\omega=2}^{n-1}\binom{k}{\omega}\mu_B^{-\omega}\langle B'^{\omega}(x)\rangle + \binom{k}{n}\sum_{\omega=2}^{n}\binom{j}{\omega}\mu_B^{-\omega}\langle B'^{\omega}(x)\rangle\right]\right\}.
\end{aligned}
\tag{2.14}
$$

In §5, we analyse the accuracy of using (2.12) to approximate (2.11) by progressively increasing the order i of approximation.

3. Generalized (cross-correlated) Sub-Gaussian fields

As shown in §2, the second-order contribution to the covariance of $Y(x)$, $C_Y^{(2)}(h)$, is fully determined by the mean and covariance of $B(x)$, the covariance of $C(x)$, and by the cross-covariance between $B(x)$ and $C(x)$ (see (2.13)). Following [12], here we assume that $B(x)$ and $C(x)$ can be interpreted through a GSG model. While we support this assumption on the basis of experimental data in §4, here we focus on (i) illustrating the key elements of the theoretical framework underpinning the GSG modelling approach and (ii) presenting an original expression enabling one to evaluate the cross-covariance between two GSG fields. The latter is then employed in (2.13) to assess the second-order contributions to the covariance (i.e. corresponding to $h > 0$) and variance ($h = 0$) of $Y(x)$.

In the context of a GSG modelling framework, perturbations of $B(x)$ and $C(x)$ around corresponding mean values can be written as

$$L'(x) = U_L(x)G_L(x), \tag{3.1}$$

where $U_L(x)$ (with $L = B, C$) represents statistically independent and identically distributed (iid) non-negative random quantities independent of $G_L(x)$. The latter corresponds to a zero-mean spatially correlated Gaussian field with standard deviation σ_{G_L} and auto-correlation ρ_{G_L}. Among the various possible models available to describe the probability density of U_L, f_{U_L} (see [25]), here, for illustration purposes and on the basis of experimental results (outlined in §4), we consider U_L to be lognormally distributed as

$$f_{U_L}(u_l) = \frac{e^{-\ln^2 u_l/(2(2-\alpha_L)^2)}}{\sqrt{2\pi}u_l(2-\alpha_L)} \quad \text{with } 0 < \alpha_L < 2, \tag{3.2}$$

where α_L is the shape parameter of the distribution. Here and in the following we use the classical notation of probability theory, in which random variables are denoted through uppercase letters and particular realizations of random variables are written in the corresponding lowercase letters. The subordinator parameter, α_L, is key in determining the shape of the GSG model. When $\alpha_L < 2$, f_{U_L} exhibits sharp peaks and heavy tails. As a special case of the GSG model, it can also be shown that the distribution of $L'(x)$ tends to Gaussian as $\alpha_L \to 2$ [12].

The study presented in [12] shows that the covariance $C_L(h)$ and variance σ_L^2 of $L'(x)$ are given by

$$C_L(h > 0) = e^{(2-\alpha_L)^2}\sigma_{G_L}^2\rho_{G_L}; \quad C_L(h = 0) = \sigma_L^2 = e^{2(2-\alpha_L)^2}\sigma_{G_L}^2. \tag{3.3}$$

Within (3.3), the influence of the subordinator, $U_L(x)$, in the determination of $C_L(h)$ is clearly evident through the inclusion of parameter α_L. Note that, according to (3.3), $C_L(h)$ exhibits a discontinuity at zero lag. The latter is commonly denoted as the nugget effect.

In the following, we derive original formulations to describe the cross-correlation between B and C. Since U_B and U_C are independent of each other and of G_L, the joint cumulative distribution

function (CDF) of B' and C' is

$$F_{B',C'}(b',c') = \iint_{D_b} \iint_{D_c} f_{U_B}(u_b)f_{U_C}(u_c)f_{G_B,G_C}(g_b,g_c)\,du_b du_c dg_b dg_c, \tag{3.4}$$

where $D_l = \{(u_l,g_l) \in \mathbb{R}^2 : u_l g_l < l\}$ (for $l = b,c$, $\mathbb{R}$ denotes real numbers), and f_{G_B,G_C} is the bivariate Gaussian PDF of (G_B, G_C), i.e.

$$f_{G_B,G_C}(g_b,g_c) = \frac{e^{-q(g_b,g_c)}}{2\pi \sigma_{G_B}\sigma_{G_C}\sqrt{1-\rho_{G_{BC}}^2}}. \tag{3.5}$$

Here, $\rho_{G_{BC}}$ is the coefficient of correlation between G_B and G_C and the exponent $q(g_b,g_c)$ in (3.5) is defined as

$$q(g_b,g_c) = \frac{g_b^2/\sigma_{G_B}^2 - 2\rho_{G_{BC}}g_b g_c/(\sigma_{G_B}\sigma_{G_C}) + g_c^2/\sigma_{G_C}^2}{2(1-\rho_{G_{BC}}^2)}. \tag{3.6}$$

Since U_L is non-negative, (3.4) reduces to

$$F_{B'C'}(b',c') = \int_{u_b=0}^{\infty}\int_{u_c=0}^{\infty}\int_{g_b=-\infty}^{b'/u_b}\int_{g_c=-\infty}^{c'/u_c} f_{U_B}(u_b)f_{U_C}(u_c)f_{G_B G_C}(g_b,g_c)\,dg_c dg_b du_c du_b. \tag{3.7}$$

The bivariate PDF of B' and C' can then be evaluated as

$$f_{B'C'}(b',c') = \frac{\partial^2}{\partial b' \partial c'}F_{B',C'}(b',c') = \int_{u_b=0}^{\infty}\int_{u_c=0}^{\infty} f_{U_B}(u_b)f_{U_C}(u_c)f_{G_B G_C}\left(\frac{b'}{u_b},\frac{c'}{u_c}\right)\frac{du_c du_b}{u_b u_c}. \tag{3.8}$$

Making use of (3.2) and (3.5), (3.8) becomes

$$f_{B'C'}(b',c') = \frac{1}{4\pi^2(2-\alpha_C)(2-\alpha_B)\sigma_{G_B}\sigma_{G_C}\sqrt{1-\rho_{G_{BC}}^2}}\int_{u_b=0}^{\infty}\int_{u_c=0}^{\infty}\frac{e^{-r}}{u_b^2 u_c^2}du_b du_c, \tag{3.9}$$

where,

$$r = \frac{1}{2}\left(\frac{\ln^2 u_b}{(2-\alpha_B)^2} + \frac{\ln^2 u_c}{(2-\alpha_C)^2} + \frac{b'^2/(\sigma_{G_B}^2 u_b^2) - 2\rho_{G_{BC}}b'c'/(\sigma_{G_B}\sigma_{G_C}u_b u_c) + c'^2/(\sigma_{G_C}^2 u_c^2)}{(1-\rho_{G_{BC}}^2)}\right). \tag{3.10}$$

Finally, making use of (3.9) one can evaluate the cross-covariance between the GSG fields $B(x)$ and $C(x)$ as

$$C_{BC}(h) = \int_{c=-\infty}^{\infty}\int_{b=-\infty}^{\infty} b'c' f_{B'C'}(b',c')\,db'dc' = \sigma_{G_B}\sigma_{G_C}\rho_{G_{BC}}e^{[(\alpha_B-4)\alpha_B + (\alpha_C-4)\alpha_C + 8]/2}. \tag{3.11}$$

Note that $C_{BC}(h)$ does not exhibit a nugget effect. Thus, it is continuous at zero separation distance, as opposed to the auto-covariance of $B(x)$ or $C(x)$ (see (3.3)).

The GSG expressions of both the auto-covariance, (3.3), and cross-covariance, (3.11), of GSG fields can then be embedded in (2.13) to evaluate the second-order approximation of the log-conductivity covariance and variance as

$$\frac{C_Y^{[2]}}{4} = \sigma_{G_C}^2 e^{(2-\alpha_C)^2}\rho_{G_C} - 2\sigma_{G_B}\sigma_{G_C}\rho_{G_{BC}}e^{\frac{1}{2}[(\alpha_B-4)\alpha_B + (\alpha_C-4)\alpha_C + 8]}\sum_{k=1}^{\infty}(-1)^k \mu_B^{k-1}$$

$$+ \sigma_{G_B}^2 e^{(2-\alpha_B)^2}\rho_{G_B}\sum_{k=1}^{\infty}\sum_{j=1}^{\infty}(-1)^{k+j}\mu_B^{k+j-2}$$

$$- (\sigma_{G_B}^2 e^{2(2-\alpha_B)^2})^2 \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}(-1)^{k+j}\frac{(k-1)}{2}\frac{(j-1)}{2}\mu_B^{k+j-4}, \tag{3.12}$$

and

$$\frac{\sigma_Y^{2[2]}}{4} = \sigma_{G_C}^2 e^{2(2-\alpha_C)^2} - 2\sigma_{G_B}\sigma_{G_C}\rho_{G_{BC}}\, e^{[(\alpha_B-4)\alpha_B+(\alpha_C-4)\alpha_C+8]/2}\sum_{k=1}^{\infty}(-1)^k\mu_B^{k-1}$$

$$+ \sigma_{G_B}^2 e^{2(2-\alpha_B)^2}\sum_{k=1}^{\infty}\sum_{j=1}^{\infty}(-1)^{k+j}\mu_B^{k+j-2}$$

$$- (\sigma_{G_B}^2\, e^{2(2-\alpha_B)^2})^2 \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}(-1)^{k+j}\frac{(k-1)}{2}\frac{(j-1)}{2}\mu_B^{k+j-4}. \tag{3.13}$$

Note that $C_Y^{[2]}$ exhibits a nugget effect, i.e. the nugget effect embedded in the auto-covariance of $B(x)$ and $C(x)$ propagates onto the auto-covariance of $Y(x)$. Finally, a second-order estimate of the integral scale of Y, $I_Y^{[2]}$ (i.e. a measure of the correlation scale of Y), can be evaluated from (3.12) and (3.13) as

$$I_Y^{[2]}\sigma_Y^{2[2]} = \sigma_{G_C}^2\, e^{(2-\alpha_C)^2}I_{G_C} - 2\sigma_{G_B}\sigma_{G_C}I_{G_{BC}}\, e^{[(\alpha_B-4)\alpha_B+(\alpha_C-4)\alpha_C+8]/2}\sum_{k=1}^{\infty}(-1)^k\mu_B^{k-1}$$

$$+ \sigma_{G_B}^2\, e^{(2-\alpha_B)^2}I_{G_B}\sum_{k=1}^{\infty}\sum_{j=1}^{\infty}(-1)^{k+j}\mu_B^{k+j-2}, \tag{3.14}$$

where I_{G_C}, $I_{G_{BC}}$ and I_{G_B} are the integral scales associated with fields G_C, G_{BC} and G_B, respectively.

Higher-order (statistical) moments and cross-moments of GSG fields are provided in appendix A. These can then be readily embedded in (2.14) to evaluate higher order contributions to the covariance of Y.

4. Field evidence of GSG behaviour

The formulations illustrated in the previous sections consider perturbations of $B(x)$ and $C(x)$ as GSG fields (3.1) where $U_L(x)$ has a lognormal distribution. Here, we present experimental evidence to support this modelling concept.

We start by considering the Ramsey sandstone porosity dataset [26] consisting of 368 data obtained from cores taken from 12 vertical boreholes located in the East Ford Field, Texas, USA. These boreholes cover an area of about $4\,\text{km} \times 1\,\text{km}$. The sedimentary rock primarily consists of clay and silt with some sand. Porosity measurements are obtained through conventional core analysis on samples with length of about 30 cm, collected at depths ranging between 824 m and 855 m below ground level. Porosity values in the dataset range from 0.045 to 0.306. Thus, to ensure compliance with the assumptions of our theoretical framework (see §2), a few porosity values were excluded from the analysis. This leads to a final dataset size of 351. For a complete review of all available qualitative and quantitative hydrogeological data from the East Ford Field site, the reader is directed to [26]. We recall that the Terzaghi empirical model (2.1) is commonly applied to unconsolidated geomaterials. It is observed that the sandstone samples associated with the available dataset are characterized by significantly low volumetric content of calcite and chlorite cement. This element contributes to a weak lithification of the sandstones [27]. Furthermore, the sediment core description provided in [26] indicates that the sandstone is characterized as 'structureless or massive sandstones having few laminations'. This classification supports the potential applicability of the Terzaghi empirical model to interpret the data. Porosity measurements are characterized by a mean of B (μ_B) equal to -0.899 and exhibit a small standard deviation ($\sigma_B = 0.031$), thus yielding a coefficient of variation equal to 3.5%. The dataset exhibits a mild left tailing (skewness equal to -0.70), and a positive excess kurtosis (equal to 0.51). The corresponding porosity dataset (ϕ) has mean equal to 0.22, standard deviation equal to 0.028, and coefficient of variation equal to 12.6%. The porosity is characterized by a left-tailed distribution (skewness equal to -0.78) and excess kurtosis equal to 0.60. Figure 1a shows that the the overall
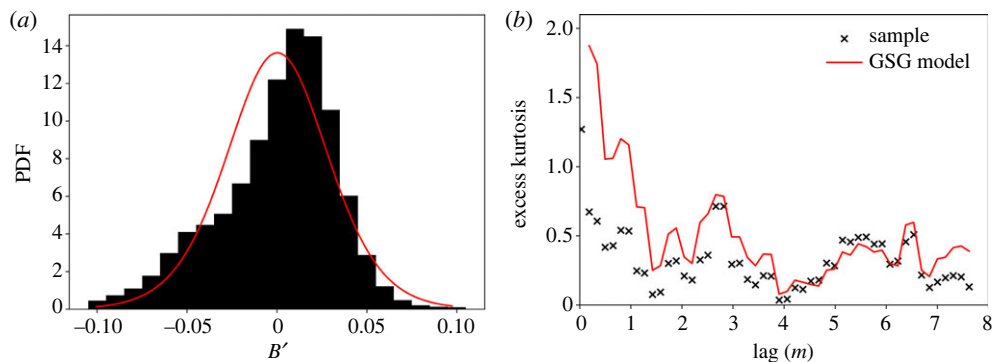
9

royalsocietypublishing.org/journal/rspa   Proc. R. Soc. A 480: 20230476

Figure 1. Sample and GSG-based (*a*) PDF (A 2) of B' and (*b*) excess kurtosis (A 8) of increments of B' associated with various separation distances (lags).

behaviour of the sample PDF of B' can be adequately captured by a GSG PDF with parameters $\alpha_B = 1.799$ and $\sigma_{G_B} = 0.0298$. These are estimated via a classical Maximum Likelihood (ML) approach. The resulting value for α_B corresponds to mild to moderate non-Gaussian behaviour [12]. Furthermore, figure 1*b* shows that the excess kurtosis of populations of increments of B' decreases with lag at which these are evaluated, i.e. the shape of the PDF of incremental values of B' varies from leptokurtic at low lags towards Gaussian at large lags. It is recalled that this finding is also consistent with the main traits documented for a variety of other environmental variables, as described in §1. These results suggest that these scaling features of B' increments can be consistently captured by the GSG model (3.1), as shown in figure 1*b*. Details about the ML approach employed to estimate the GSG model parameters are offered in [12].

We recall that extensive datasets associated with porosity and particle size curves collected at the same location along boreholes across a given site are not commonly available. Thus, we then consider PSC data associated with a dataset formed by 407 values of d_{10} evaluated via grain sieve analyses performed on drill core samples, which vary in length from 5 to 26.5 cm. The core samples were collected along 12 vertical boreholes at an experimental site near the city of Tuebingen, Germany [19]. The experimental site covers an area of about 400 m × 250 m, with a thickness of 8 m. The dataset primarily consists of alluvial deposits composed of gravel, sand and clay. This lithological composition renders it well-suited for application within the Terzaghi empirical framework. These data have also been previously employed to assess the potential of various geostatistically based aquifer characterization approaches (e.g. [19,28–31]).

Considering this information, we obtain $\mu_C = -0.477$ and $\sigma_C = 0.870$ (this corresponds to a large coefficient of variation, i.e. = 182.5%). It can be noted that C exhibits a very mild positive skewness (equal to 0.16) and negligible excess kurtosis (equal to -0.07). The corresponding d_{10} dataset exhibits a mean equal to 0.915 mm and a standard deviation equal to 0.917 mm, which results in a large coefficient of variation (equal to 100.2%). Furthermore, the d_{10} dataset is strongly non-Gaussian as evidenced by a right tail (skewness equal to 2.10) and high excess kurtosis (equal to 4.57). Figure 2*a* depicts the sample PDF of C' as well as the corresponding GSG-based counterpart. Estimates of the parameters of the latter (i.e. $\alpha_C = 1.996$ and $\sigma_{GC} = 0.87$) are obtained via the ML procedure described in [12]. Since $\alpha_C \approx 2$, the dataset exhibits negligible non-Gaussian traits. Nevertheless, the non-Gaussian features of C' are clearly evident in figure 2*b*. The graph reveals that the excess kurtosis of increments of C' is not constant across different lags (as it should be expected from a Gaussian PDF of C'), but rather decreases as lag increases. These distinctive features of C' can be effectively captured by our GSG model. These results, together with other available evidences– (see §1), contribute to strengthening our confidence in the ability of the GSG theoretical framework to describe the scaling behaviour of the statistics of the target random fields together with their increments. Note that the main purpose of the analysis of the datasets included here is related to support the applicability of a GSG modelling framework in capturing
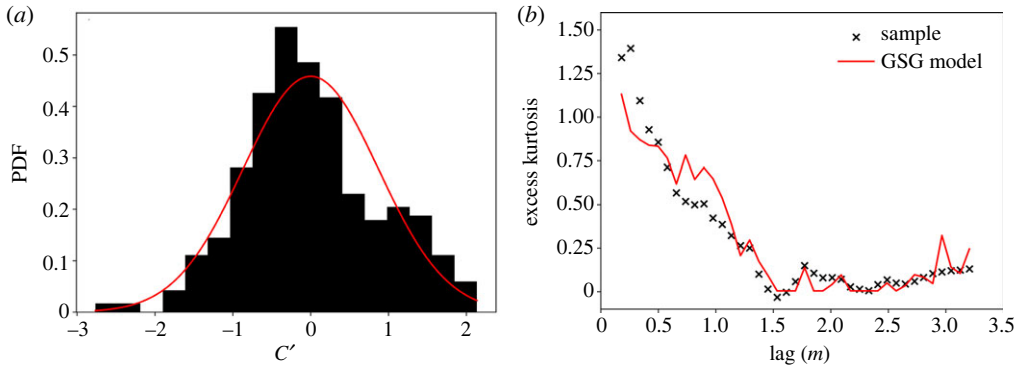
Figure 2. Sample and GSG-based (*a*) PDF (A 2) of C' and (*b*) excess kurtosis (A 8) of increments of C' evaluated at various separation distances (lags).

non-Gaussian traits exhibited by the B and C fields. The focus is not on characterizing a specific site. It is worth emphasizing that when examining a given dataset, it is important to jointly analyse the sample PDF of the data and their increments in order to detect non-Gaussian features.

5. Accuracy of the truncated analytical solution of C_Y

This section is devoted to the analysis of the accuracy of truncating the exact analytical expression (2.11) at given orders of approximation, as described through (2.12). We provide a transparent assessment of the accuracy of approximations of increasing order upon relying on synthetic numerical Monte Carlo (MC) simulations. The analysis is performed by evaluating $C_Y^{[i]}$ for increasing values of i (i.e. by increasing the order of approximation) for diverse levels of heterogeneity (as quantified by the variance) of the fields of B and C.

Synthetic realizations of $Y(x)$ are generated in a one-dimensional setting mimicking a scenario consistent with availability of borehole data. The simulation domain comprises 1000 elements, each of length equal to 0.01 (in arbitrary consistent units). We generate random fields of $Y(x)$ in three steps. First, Numpy's 'multivariate_normal' function [32] is used to generate Gaussian one-dimensional (auto- and cross-correlated) fields $G_L(x)$ (with $L = B, C$). Then a realization of $L'(x)$ is obtained upon making use of (3.1). Finally, the field $Y(x)$ is constructed upon resting on (2.2). The three-step procedure is repeated to yield a collection of $N_{MC} = 10\,000$ realizations of $Y(x)$. These enable us to obtain stable results for the quantities we analyse (not shown) and are then employed to compute the (ensemble) covariance, C_Y, and its corresponding correlation function. Reference values of the integral length scale of $Y(x)$, I_Y, are evaluated by numerically integrating the corresponding MC-based (ensemble) correlation function over the entire domain. These converged estimates are then used for meaningful comparisons with their corresponding analytical counterparts.

As shown in (2.11), the auto-covariance of $Y(x)$ depends on the auto-correlation function of $B(x)$ and $C(x)$ as well as their cross-correlation. For illustration purposes, we adopt an exponential function to describe these auto- and cross-correlations, i.e.

$$\rho_{G_L}(h) = e^{-h/I_{G_L}} \quad \text{and} \quad \rho_{G_{BC}}(h) = a\,e^{-h/I_{G_{BC}}}. \tag{5.1}$$

Here, for illustration purposes, we set the characteristic length scales $I_{G_L} = I_{G_{BC}} = 1$. We then set the values of μ_B and α_L at the corresponding ML estimates obtained in §4. Note that, according to (2.11), the covariance of Y does not depend on the mean of C. To assess the accuracy of $C_Y^{[i]}$ to approximate C_Y, we consider diverse values of σ_{G_L} (and hence log-conductivity variance). We focus on four test cases, that cover a range of conductivity fields varying from mild to moderate heterogeneity, hereafter labelled as TC1–TC4. The GSG parameters used to model B and C fields for each test case are listed in table 2.
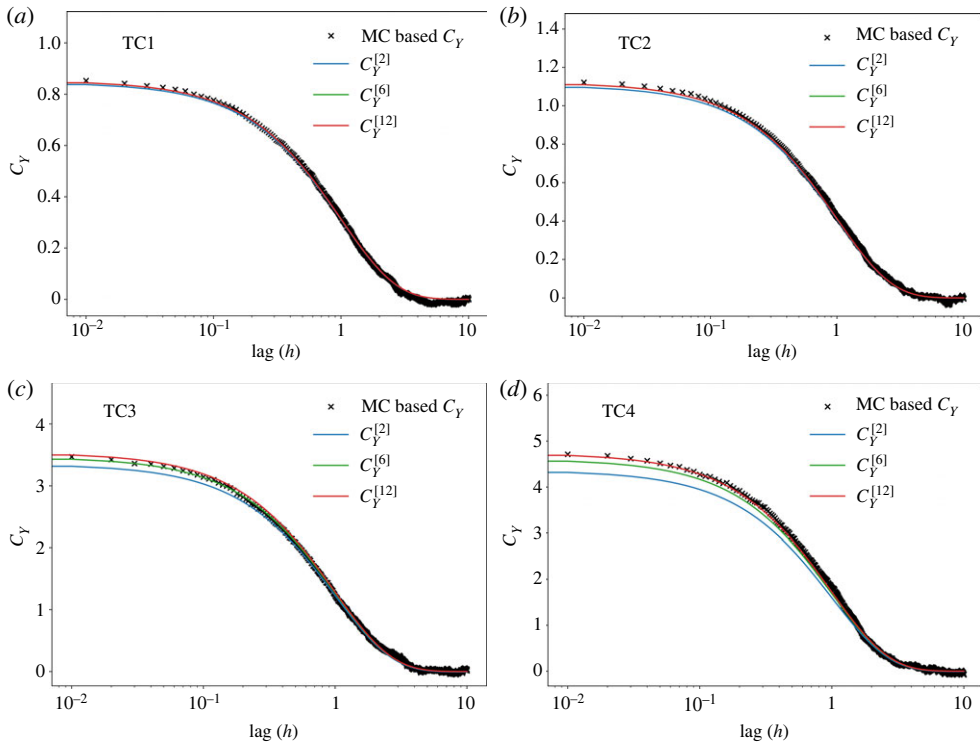
Figure 3. Monte Carlo-based C_Y and its counterpart evaluated through analytical approximations $C_Y^{[i]}$ ($i = 2, 6, 12$) (2.12) for TC1 (*a*), TC2 (*b*), TC3 (*c*) and TC4 (*d*).

Table 2. Values of the parameters employed in the Monte Carlo simulations scenarios. Resulting values of σ_Y are also included.

test case	α_B	σ_{G_B}	α_C	σ_{G_C}	a	σ_Y
TC1	1.799	0.015	1.996	0.435	0	0.926
TC2	1.799	0.015	1.996	0.435	0.5	1.061
TC3	1.799	0.0298	1.996	0.870	0	1.908
TC4	1.799	0.0298	1.996	0.870	0.5	2.193

We recall here that a geologic medium can be conceptualized as a system composed of disjoint regions of differing geomaterials, each with a given heterogeneous distribution of conductivity. In this context, the selected mild (TC1 and TC2) to moderate (TC3 and TC4) values of log-conductivity variance correspond to those that can be associated with the degree of heterogeneity characterizing a given geologic unit (e.g. [33,34]) and can be considered as suitable to appraise the accuracy of our analytical approximations. We further note that the $B(x)$ and $C(x)$ fields are uncorrelated in TC1 and TC3 and correlated in TC2 and TC4, to enable one to encompass a variety of scenarios.

Figure 3 juxtaposes MC-based values of C_Y (symbols) and $C_Y^{[i]}$ ($i = 2, 6, 12$) for all test cases analysed. As expected, the accuracy of the second-order analytical solution, $C_Y^{[2]}$, decreases as the level of heterogeneity of the log-conductivity fields increases. Differences between C_Y and $C_Y^{[2]}$ are virtually negligible for the smallest values of σ_Y analysed (i.e. TC1 and TC2). Higher orders of approximation are required for $C_Y^{[i]}$ for increasing values of σ_Y. This is particularly evident at small lag values. Furthermore, upon comparing figures 3*a*–*b* and figures 3*c*–*d* it can be

observed that the accuracy of $C_Y^{[i]}$ is unaffected by the cross-correlation between B and C. Tables 3 and 4 list, respectively, MC-based values of the variance and integral scales of Y against their analytical (truncated) approximations (for $i = 2, \ldots, 12$), as well as the percentage error associated with these. Percentage errors associated to TC1 and TC2 are (in general) similar. Also, the percentage error for TC3 and TC4 are characterized by similar trends. This quantitative analysis further strengthens the observation that the accuracy of the truncated solution is not significantly influenced by the cross-correlation between the B and C fields.

The results of this analysis enable us to conclude that, while the lowest order solution (i.e. $i = 2$) is sufficient to obtain accurate analytical results for mildly heterogeneous fields (TC1 and TC2), a higher order of approximation (up to $i = 12$) is required for increased heterogeneity of the system (TC3 and TC4). Despite the potential requirement for the use of a high order of approximation, the computational time required for executing these calculations remains negligible (less than 10 s when evaluating $C_Y^{[12]}$ on an Intel Core i9). Thus, there are no computational drawbacks associated with employing a higher order approximation. On this basis, we rely on $C_Y^{[12]}$ as a surrogate model for C_Y to perform the global sensitivity analysis illustrated in §6.

6. Global sensitivity analysis

It can be noted that the format of the analytical formulations derived do not enable a straightforward assessment of the relative effects of the parameters embedded therein. Being able to ascertain these elements is critical to enhancing our understanding of model functioning. In addition to this, we recall that our knowledge of model parameters is always hampered by uncertainty and a proper quantification of relationships between model inputs and outputs can provide valuable insights into the way such uncertainty propagates onto model outputs as well as guide future research efforts.

Thus, here, we rely on a global sensitivity analysis to assess the relative importance of the key parameters (i.e. μ_B, α_B, σ_{G_B}, I_{G_B}, α_C, σ_{G_C}, I_{G_C}, $I_{G_{BC}}$ and a) embedded in the GSG modelling framework describing the fields of characteristic grain size of the particle size curve and porosity on the ensuing log-conductivity covariance. We recall that α_L (with $L = B, C$) is a measure of the strength of departure of the GSG field from a Gaussian behaviour (see §3). The role of α_L on the covariance and excess kurtosis of a GSG random field can be seen in (3.3) and (A 8), respectively. Global sensitivity analysis is here performed upon relying on the well-known Sobol indices [35]. These global sensitivity metrics are widely employed to quantify the relative importance of uncertain model parameters on a model output of interest. Their formulation is based on the concept of variance apportioning, i.e. the highest sensitivity index represents the strongest contribution to the uncertainty (as quantified through the variance) of a target model output. Owing to their simplicity and intuitive nature, they have been considered in a variety of studies associated with field- and laboratory-scale hydrogeological scenarios (e.g. [36–42] and references therein).

For the purpose of our analysis, each uncertain model parameter is considered as a statistically independent Gaussian random variable with mean corresponding to the value used in §4 for TC4. Relying on these estimates is motivated by their consistency with the degrees of heterogeneity associated with our considered datasets and by the observation that TC4 incorporates cross-correlation between the B and C fields, enabling the assessment of the influence arising from these effects. A coefficient of variation equal to 1% is considered so that the possibility that some realizations violate the assumption underlying the derivation of (2.4) (i.e. $|B(\boldsymbol{x})| < 1$) is mitigated. The parameter space is sampled through a Monte Carlo approach. Here, we rely on 2000 random realizations for each parameter of the B and C GSG fields to ensure stable results (not shown).

Figure 4 depicts the sample PDF of $\sigma_Y^{2[12]}$ as well as the values of the total Sobol indices associated with each of the uncertain model parameters included in the formulation of $\sigma_Y^{2[12]}$. We note that μ_B is ranked as the model parameter that is most influential on the variability of $\sigma_Y^{2[12]}$, as reflected by the associated variance-based total Sobol index. It is then followed by σ_{G_C}

14

royalsocietypublishing.org/journal/rspa   *Proc. R. Soc. A* **480**: 20230476

Table 3. Monte Carlo-based values of σ_Y^2 and their counterparts corresponding to the analytical solution associated with various orders of truncation. Percentage errors are also included.

test case	MC σ_Y^2	$\sigma_Y^{2[2]}$	$\sigma_Y^{2[4]}$	$\sigma_Y^{2[6]}$	$\sigma_Y^{2[8]}$	$\sigma_Y^{2[10]}$	$\sigma_Y^{2[12]}$
TC1	0.860	0.850 (1.194%)	0.857 (0.379%)	0.858 (0.271%)	0.858 (0.248%)	0.858 (0.241%)	0.858 (0.238%)
TC2	1.125	1.111 (0.215%)	1.124 (0.030%)	1.126 (0.105%)	1.126 (0.131%)	1.126 (0.138%)	1.126 (0.140%)
TC3	3.656	3.363 (8.007%)	3.455 (5.480%)	3.501 (4.244%)	3.537 (3.258%)	3.578 (2.127%)	3.640 (0.425%)
TC4	4.781	4.361 (8.787%)	4.546 (4.918%)	4.621 (3.348%)	4.675 (2.221%)	4.732 (1.035%)	4.810 (0.608%)

Table 4. Monte Carlo-based values of I_Y and their counterparts corresponding to the analytical solution associated with various orders of truncation. Percentage errors are also included.

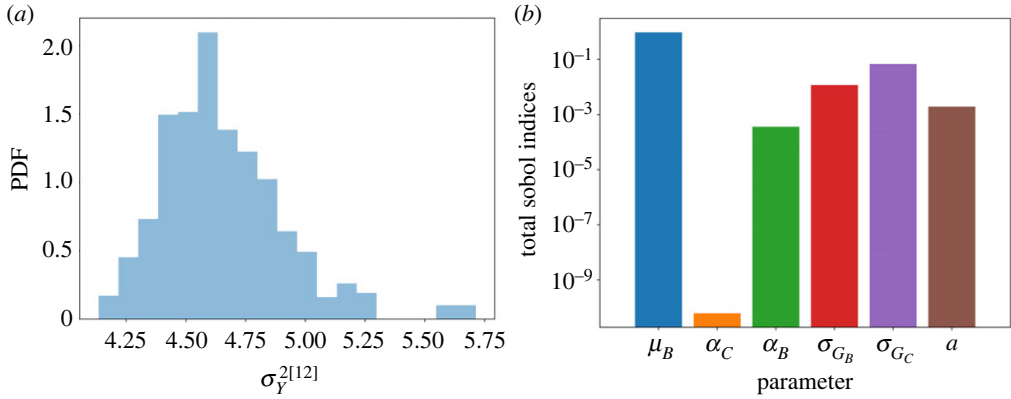| test case	MC I_Y	$I_Y^{[2]}$	$I_Y^{[4]}$	$I_Y^{[6]}$	$I_Y^{[8]}$	$I_Y^{[10]}$	$I_Y^{[12]}$
TC1	1.008	0.996 (1.206%)	0.995 (1.304%)	0.994 (1.328%)	0.994 (1.336%)	0.994 (1.339%)	0.994 (1.341%)
TC2	1.007	0.997 (0.997%)	0.996 (1.071%)	0.996 (1.089%)	0.996 (1.095%)	0.996 (1.098%)	0.996 (1.099%)
TC3	0.952	0.996 (4.650%)	0.993 (4.304%)	0.990 (3.992%)	0.986 (3.621%)	0.981 (3.056%)	0.971 (2.012%)
TC4	0.993	0.997 (0.350%)	0.994 (0.102%)	0.992 (0.122%)	0.990 (0.390%)	0.985 (0.799%)	0.978 (1.557%)

Figure 4. Sample PDF of $\sigma_Y^{2[12]}$ (a), and total Sobol indices associated with all uncertain parameters included in the formulation of $\sigma_Y^{2[12]}$ (b).
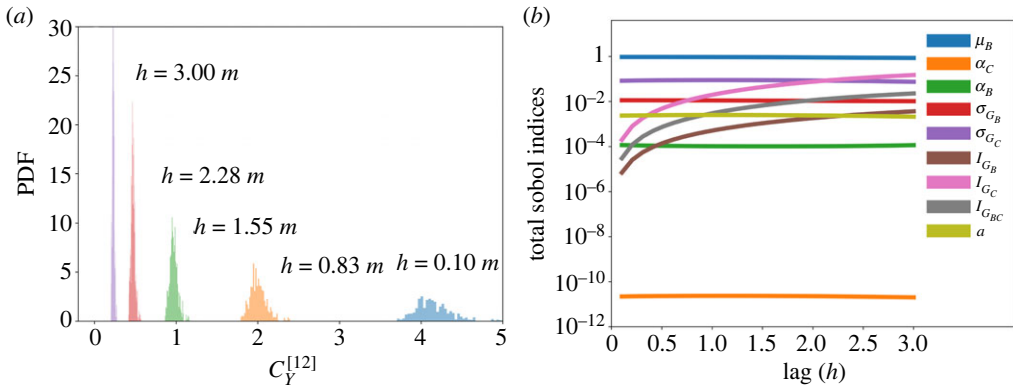

Figure 5. Sample PDF of $C_Y^{[12]}$ for representative lags, h (a); total Sobol indices associated with all uncertain parameters embedded in $C_Y^{[12]}$ versus lag (b).

and σ_{G_B}. This reveals that the uncertainty associated with the variance of the Gaussian fields underlying the GSG processes has a higher importance on σ_Y than the corresponding variance of the subordinator, U_L ($L = B, C$). Further to this, we note that the total Sobol indices related to σ_{G_C} and σ_{G_B} are very similar. These results imply that σ_Y is (virtually) equally influenced by the Gaussian component associated with porosity and d_{10} fields in the setting analysed. Notably, the strength of the non-Gaussian behaviour of the transformed porosity field, as quantified by α_B, plays a non-negligible role in driving the variance of Y. On the contrary, the impact of the departure from Gaussian of transformed d_{10} is less important.

Figure 5 depicts the sample PDF of $C_Y^{[12]}$ evaluated at five representative lags and the space evolution of total Sobol indices related to all parameters embedded in $C_Y^{[12]}$. These results provide further evidence that μ_B is the most influential parameter in our model, its impact being almost constant across the range of separation distances considered. They also show α_B to have a near constant (and non-negligible) effect on C_Y. The total Sobol index associated with α_C is very small and remains virtually constant with lag. These results suggest that accurately capturing heavy tails in the distribution of (transformed) porosity is critical to properly assess the main features of C_Y. Conversely, a precise reproduction of non-Gaussian features (as expressed through the range
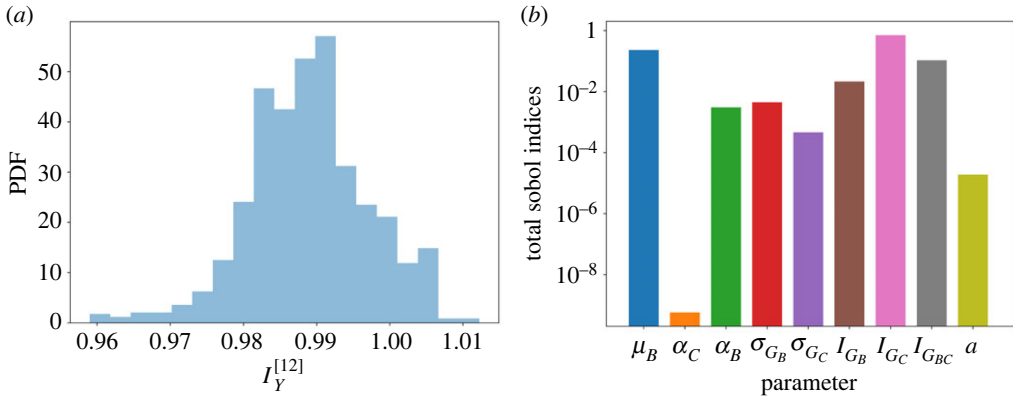
Figure 6. Sample PDF of $I_Y^{[12]}$ (*a*); total Sobol indices associated with all uncertain parameters contributing to $I_Y^{[12]}$ (*b*).

of values considered for α_C) associated with the (transform) d_{10} has a comparatively lesser impact on C_Y.

With reference to the integral scales I_{G_B}, I_{G_C} and $I_{G_{BC}}$ (associated with auto-correlation of B and C as well as to their cross-correlation, respectively), it is noted that their impact vanishes at zero lag, consistent with the observation that C_Y reduces to the variance of Y therein. Their importance for the uncertainty of C_Y tends to (monotonically) increase with lag. The latter result is in line with the conceptual meaning of these parameters; they are informative of the spatial extent of the correlation between two separate locations across the random field. Our results show that l_{G_C} has the strongest impact on C_Y. It is then followed by $I_{G_{BC}}$ and I_{G_B}.

Finally, figure 6 depicts the sample PDF of the integral scale of Y, $I_Y^{[12]}$, as well as the values of the total Sobol indices associated with each of the uncertain model parameters included in its formulation. Similar to what has been observed for $C_Y^{[12]}$ and $\sigma_Y^{2[12]}$, all parameters have a significant impact on $I_Y^{[12]}$, the only exception being α_C. We note that the parameter controlling the spatial extent of cross-correlation between B and C, $I_{G_{BC}}$, is associated with an important effect also on $I_Y^{[12]}$. This observation, together with the documented non-negligible importance of a, provides further evidence of the relevance of properly assessing possible cross-correlations between B and C in modelling I_Y. In other words, the spatial extent of cross-correlation between B and C has an important effect on the spatial extent of correlation observed in the hydraulic conductivity field.

It is remarked that $\alpha_C \approx 2$, as a result of the maximum-likelihood parameter estimation procedure. Thus, its effect on the model outcomes is limited in our analyses. However, the effect of α_B, whose values are distinctive of a strongly non-Gaussian random field (i.e. $\alpha_B \ll 2$, as shown in table 2), is relevant. In particular, the importance of such a parameter on $I_Y^{[12]}$ is similar to that of σ_{G_B} and I_{G_B}. This result supports the relevance of properly identifying possible deviations from Gaussian of the PDF model employed to describe a random quantity of interest.

7. Conclusion

The work presented in this study leads to the following main conclusions.

1. We provide a set of analytical formulations enabling one to relate the spatial covariance of randomly distributed log-conductivities to those of porosity and characteristic diameters related to particle size curves of soil samples. We rely on the classical Terzaghi formulation [22] that is commonly used to empirically infer conductivity of large-grain sands. A perturbation-based approach is then employed to obtain workable approximations of the otherwise exact formulations derived.

2. Our formulations are fully compatible with recent findings [11] documenting the suitability of a Generalized Sub-Gaussian (GSG) model to jointly represent the probability distribution of hydrogeological (and other environmental) quantities and their spatial increments associated with various separation distances (or lags). In this context, we present original formulations for cross-moments of bivariate GSG random fields (appendix A). These cross-moments are expressed in terms of the cross-covariance between the two GSG fields. We verify the appropriateness of treating porosity and representative PSC diameters as GSG random fields upon considering two established extensive datasets.

3. The effect of retaining only a finite number of terms in the perturbation-based expansion of the exact expression of log-conductivity covariance (C_Y) is tested through comparison against a collection of synthetic scenarios in a Monte Carlo framework. We find that relying on an order of approximation $i = 12$ for C_Y enables one to capture its Monte Carlo-based counterpart for all test cases analysed.

4. A Global Sensitivity Analysis grounded on the classical Sobol indices is performed to assess the relative importance of the parameters embedded in the formulation of C_Y. These model parameters correspond to those governing the GSG distribution of porosity and representative PSC diameter, including the length scales corresponding to the strength of the spatial auto- and cross-correlation. A key result of this study is that properly capturing the GSG nature of the distribution of (transformed) porosity is critical to effectively assess the main features of C_Y. Additionally, our findings underline the importance of properly assessing possible cross-correlations between porosity and representative soil particle sizes.

5. Our analytical approach can be readily transferred to derive corresponding formulations relying on other (semi)empirical formulations relating conductivity with porosity and PSC representative diameters. We recall that obtaining porosity and PSC information along boreholes in typical field settings is usually inexpensive and straightforward. Otherwise, extensive datasets comprising data of these quantities jointly collected at the same locations are not common. Our results and analyses support the benefit of having at our disposal these types of information to infer estimates of log-conductivity covariance. The latter can then be readily employed in the context of typical geostatistical frameworks for estimation or (stochastic) simulation of random conductivity fields.

Appendix A. Statistical moments of GSG fields

Here, we derive analytical expressions of high-order ($n > 2$) statistical moments of GSG fields (i.e. B and C in our setting) as well as their cross-moments. These can then be readily embedded in (2.14) for the evaluation of higher-order approximations of the covariance of Y.

(a) Statistical moments and cross-moments of GSG perturbations

The detailed derivation of the PDF of the GSG field L', defined in (3.1), is illustrated in [12]. When U_L follows a lognormal distribution, i.e.

$$f_{U_L}(u_L) = \frac{1}{\sqrt{2\pi}\, u_L (2 - \alpha_L)} \, e^{-\ln^2 u_L / (2(2-\alpha_L)^2)}, \tag{A 1}$$

one obtains

$$f_{L'}(l') = \frac{1}{2\pi(2 - \alpha_L)} \int_0^\infty \frac{1}{u^2} \, e^{-\left(\frac{1}{(2-\alpha_L)^2}\, \ln^2 \frac{u}{\sigma_{G_L}} + \frac{l^2}{u^2}\right)/2}. \tag{A 2}$$

Making use of (A 2), Riva *et al.* [12] shows that L' moments of odd orders vanish and those of even orders are given by

$$\langle L'^n \rangle = \frac{\sigma_{G_L}^n}{\sqrt{\pi}} 2^{n/2} \Gamma\left(\frac{n+1}{2}\right) e^{(2-\alpha_L)^2 n^2 / 2}. \tag{A 3}$$

The PDF of incremental values of L ($\Delta L = L(x) - L(y)$) evaluated between two locations x and y is given by Riva *et al.* [12]

$$f_{\Delta L}(\Delta l) = \frac{1}{2\pi^2 (2 - \alpha_L)^2} \sqrt{\frac{\pi}{2}}$$
$$\int_0^\infty \int_0^\infty \frac{e^{-[1/(2-\alpha_L^2)(\ln^2(u_1/\sigma_{G_L}) + \ln^2(u_2/\sigma_{G_L})) + \Delta l^2/(u_1^2 + u_2^2 - 2u_1 u_2 \rho_{G_L})]/2}}{u_2 u_1 \sqrt{u_1^2 + u_2^2 - 2u_1 u_2 \rho_{G_L}}} \, du_2 du_1. \tag{A 4}$$

Making use of (A 4), the cross-moment of order $(n + m)$ between $L'^n(x)$ and $L'^m(y)$ can be evaluated as,

$$\langle L'^n(x) L'^m(y) \rangle = \frac{1}{4\pi^2 (2-\alpha_L)^2 \sigma_{G_L}^2 \sqrt{1 - \rho_{G_L}^2}} \int_{u_1=0}^\infty \int_{u_2=0}^\infty \frac{e^{-(\ln^2 u_1 + \ln^2 u_2)/(2(2-\alpha_L)^2)}}{u_1^2 u_2^2}$$
$$\int_{-\infty}^\infty \int_{-\infty}^\infty l_1'^n l_2'^m e^{-[1/((1-\rho_{G_L}^2)\sigma_{G_L}^2)(l_1'^2/u_1^2 + l_2'^2/u_2^2 - 2\rho_{G_L} l_1' l_2'/(u_1 u_2))]/2} \, du_1 du_2 dl_1' dl_2'. \tag{A 5}$$

Integration of (A 5) leads to

$$\langle L'^n(x) L'^m(y) \rangle = I_1 \sigma_{G_L}^{m+n} e^{(2-\alpha_L)^2 (n^2 + m^2)/2}, \tag{A 6}$$

with

$$I_1 = \frac{1}{\sigma_{G_L}^{m+n+2} \sqrt{1 - \rho_{G_L}^2}} \int_{-\infty}^\infty \int_{-\infty}^\infty x^n y^m \, e^{-[(1/((1-\rho_{G_L}^2)\sigma_{G_L}^2)(x^2 + y^2 - 2\rho_{G_L} xy)]/2} \, dx \, dy. \tag{A 7}$$

According to (A 6)–(A 7), it is noted that the cross-moment vanishes if n is odd and m is even (and vice-versa). A closed-form solution of (A 7) can be derived for integer (odd or even) values of n and m. Table 5 lists the solutions of (A 7) used in our study to evaluate C_Y up to order 12 of approximation.

Table 5. Closed-form solutions of the integral (A 7) up to order 12 of approximation.

$n+m$	n	m	I_1
2	1	1	ρ_{G_L}
4	2	2	$1 + 2\rho_{G_L}^2$
	3	1	$3\rho_{G_L}$
6	3	3	$3\rho_{G_L}[3 + 2\rho_{G_L}^2]$
	4	2	$3[1 + 4\rho_{G_L}^2]$
	5	1	$15\rho_{G_L}$
8	4	4	$3[3 + 8\rho_{G_L}^2(3 + \rho_{G_L}^2)]$
	5	3	$15\rho_{G_L}[3 + 4\rho_{G_L}^2]$
	6	2	$15[1 + 6\rho_{G_L}^2]$
	7	1	$105\rho_{G_L}$
10	5	5	$15\rho_{G_L}[15 + 8\rho_{G_L}^2(5 + \rho_{G_L}^2)]$
	6	4	$45\rho_{G_L}[1 + 4\rho_{G_L}^2(3 + 2\rho_{G_L}^2)]$
	7	3	$315\rho_{G_L}[1 + 2\rho_{G_L}^2]$
	8	2	$105[1 + 8\rho_{G_L}^2]$
	9	1	$945\rho_{G_L}$
12	6	6	$45[5 + 2\rho_{G_L}^2(45 + 4\rho_{G_L}^2(2\rho_{G_L}^2 + 15))]$
	7	5	$315\rho_{G_L}[5 + 4\rho_{G_L}^2(5 + 2\rho_{G_L}^2)]$
	8	4	$315[1 + 16\rho_{G_L}^2(1 + \rho_{G_L}^2)]$
	9	3	$945\rho_{G_L}[3 + 8\rho_{G_L}^2]$
	10	2	$945[1 + 10\rho_{G_L}^2]$
	11	1	$10\,395\rho_{G_L}$

Considering (A 4) and sample-based incremental values of B and C at various lags yields (ML-based) estimates of α_L, σ_{G_L} and ρ_{G_L} are obtained as illustrated by Riva *et al.* [12]. The excess kurtosis of the increment distributions, $\kappa_{\Delta L} - 3$, is then evaluated from these estimates as [12]

$$\kappa_{\Delta L} - 3 = 3\,e^{2(2-\alpha_L)^2}\left[1 + \frac{1}{2}\left(\frac{e^{2(2-\alpha_L)^2} - 1}{e^{(2-\alpha_L)^2} - \rho_{G_L}}\right)\right] - 3. \tag{A 8}$$

ML-based values of excess kurtosis are then juxtaposed to their sample-based counterparts (figures 1 and 2).

(b) Cross-moments between B and C

The cross-moment of order $(n + 1)$ between $B'^m(x)$ and $C'(y)$ can be evaluated upon making use of (3.9) as

$$\langle B'^m(x)C'(y)\rangle = \frac{1}{4\pi^2(2-\alpha_B)(2-\alpha_C)\sigma_{G_B}\sigma_{G_C}\sqrt{1-\rho_{G_{BC}}^2}}\int_{u_b=0}^{\infty}\int_{u_c=0}^{\infty}\frac{1}{u_b^2 u_c^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}$$

$$b'^m c'\,e^{-[\ln^2 u_b/(2-\alpha_B)^2+\ln^2 u_c/(2-\alpha_C)^2+1/(1-\rho_{G_{BC}}^2)(b'^2/(\sigma_{G_B}^2 u_b^2)+c'^2/(\sigma_{G_C}^2 u_c^2)-2\rho_{G_{BC}}b'c'/(\sigma_{G_B}\sigma_{G_C}u_b u_c))]/2}\,db'\,dc'\,du_c\,du_b. \tag{A 9}$$

Integration of (A 9) leads to

$$\langle B''(\boldsymbol{x})C'(\boldsymbol{y})\rangle = 2^{\frac{n}{2}} n \frac{\sigma_{G_B}^n \sigma_{G_C}}{2\sqrt{2\pi}} [1 - (-1)^n] \rho_{G_{BC}} \Gamma\left(\frac{n}{2}\right) e^{\frac{1}{2}((2-\alpha_B)^2 + (2-\alpha_C)^2 n^2)}. \tag{A 10}$$

Note that (A 10) vanishes when n is even.

References

1. Ritzi Jr RW, Jayne DF, Zahradnik Jr AJ, Field AA, Fogg GE. 1994 Geostatistical modeling of heterogeneity in glaciofluvial, buried-valley aquifers. *Groundwater* **32**, 666–674. (doi:10.1111/j.1745-6584.1994.tb00903.x)
2. Guadagnini L, Guadagnini A, Tartakovsky DM. 2004 Probabilistic reconstruction of geologic facies. *J. Hydrol.* **294**, 57–67. (doi:10.1016/j.jhydrol.2004.02.007)
3. Anderson MP. 1989 Hydrogeologic facies models to delineate large-scale spatial trends in glacial and glaciofluvial sediments. *Geol. Soc. Am. Bull.* **101**, 501–511. (doi:10.1130/0016-7606(1989)101<0501:HFMTDL>2.3.CO;2)
4. McGrath R, Styles P, Thomas E, Neale S. 2002 Integrated high-resolution geophysical investigations as potential tools for water resource investigations in karst terrain. *Environ. Geol.* **42**, 552–557. (doi:10.1007/s00254-001-0519-2)
5. Allen-King RM, Halket RM, Gaylord DR, Robin MJ. 1998 Characterizing the heterogeneity and correlation of perchloroethene sorption and hydraulic conductivity using a facies-based approach. *Water Resour. Res.* **34**, 385–396. (doi:10.1029/97WR03496)
6. Koltermann CE, Gorelick SM. 1996 Heterogeneity in sedimentary deposits: a review of structure-imitating, process-imitating, and descriptive approaches. *Water Resour. Res.* **32**, 2617–2658. (doi:10.1029/96WR00025)
7. Partington D, Therrien R, Simmons CT, Brunner P. 2017 Blueprint for a coupled model of sedimentology, hydrology, and hydrogeology in streambeds. *Rev. Geophys.* **55**, 287–309. (doi:10.1002/2016RG000530)
8. Maurer T, Schneider A, Gerke HH. 2013 Scenario-based three-dimensional distributed sediment structures for a constructed hydrological catchment. *Vadose Zone J.* **12**, 4–18. (doi:10.2136/vzj2013.02.0047)
9. Troisi S, Fallico C, Straface S, Migliari E. 2000 Application of kriging with external drift to estimate hydraulic conductivity from electrical-resistivity data in unconsolidated deposits near Montalto Uffugo, Italy. *Hydrol. J.* **8**, 356–367. (doi:10.1007/s100400000083)
10. Patriarche D, Castro MC, Goovaerts P. 2005 Estimating regional hydraulic conductivity fields–a comparative study of geostatistical methods. *Math. Geol.* **37**, 587–613. (doi:10.1007/s11004-005-7308-5)
11. Guadagnini A, Riva M, Neuman SP. 2018 Recent advances in scalable non-Gaussian geostatistics: the generalized sub-Gaussian model. *J. Hydrol.* **562**, 685–691. (doi:10.1016/j.jhydrol.2018.05.001)
12. Riva M, Neuman SP, Guadagnini A. 2015 New scaling model for variables and increments with heavy-tailed distributions. *Water Resour. Res.* **51**, 4623–4634. (doi:10.1002/2015WR016998)
13. 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 Resour. Res.* **56**, e2020WR027436. (doi:10.1029/2020WR027436)
14. Li K, Wu J, Nan T, Zeng X, Yin L, Zhang J. 2022 Analysis of heterogeneity in a sedimentary aquifer using generalized sub-Gaussian model based on logging resistivity. *Stochast. Environ. Res. Risk Assess.* **36**, 1–17.
15. Siena M, Riva M, Giamberini M, Gouze P, Guadagnini A. 2019 Statistical modeling of gas-permeability spatial variability along a limestone core. *Spat. Stat.* **34**, 100249. (doi:10.1016/j.spasta.2017.07.007)
16. Seelheim F. 1880 Methoden zur bestimmung der durschlassigkeit des bodens [in German]. *Zeitschrift für Anal. Chem.* **19**, 387–418. (doi:10.1007/BF01341054)
17. Hazen A. 1983 Some physical properties of sand and gravel with special reference to their use in filtration. In *24th Annual Report, Massachusetts State Board of Health.* Boston, MA*:* Massachusetts State Board of Health.

18. Rosas J, Lopez O, Missimer TM, Coulibaly KM, Dehwah AH, Sesler K, Lujan LR, Mantilla D. 2014 Determination of hydraulic conductivity from grain-size distribution for different depositional environments. *Groundwater* **52**, 399–413. (doi:10.1111/gwat.12078)

19. Riva M, Sanchez-Vila X, Guadagnini A. 2014 Estimation of spatial covariance of log conductivity from particle size data. *Water Resour. Res.* **50**, 5298–5308. (doi:10.1002/2014WR015566)

20. Beyer W. 1964 Zur bestimmung der wasserdurchlässigkeit von kiesen und sanden aus der kornverteilungskurve. *WWT* **14**, 165–168.

21. Carman PC. 1937 Fluid flow through granular beds. *Trans. Inst. Chem. Eng.* **15**, 150–166.

22. Terzaghi K. 1925 Principles of soil mechanics. *Eng. News-Rec.* **95**, 19–32.

23. Riva M, Guadagnini A, Neuman SP. 2017 Theoretical analysis of non-Gaussian heterogeneity effects on subsurface flow and transport. *Water Resour. Res.* **53**, 2998–3012. (doi:10.1002/2016WR019353)

24. Ceresa L, Guadagnini A, Riva M, Porta GM. 2022 Macrodispersion in generalized sub-Gaussian randomly heterogeneous porous media. *Int. J. Heat Mass Transfer* **195**, 123117. (doi:10.1016/j.ijheatmasstransfer.2022.123117)

25. 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 Resour. Res.* **56**, 2020WR027436. (doi:10.1029/2020WR027436)

26. Dutton SP, Flanders WA, Barton MD. 2003 Reservoir characterization of a Permian deep-water sandstone, East Ford field, Delaware basin, Texas. *AAPG Bull.* **87**, 609–627. (doi:10.1306/10100201085)

27. Urban J. 2016 The geological constraints of the development of sandstone landforms in Central Europe, a case study of the Świętokrzyskie (Holy Cross) Mountains, Poland. *Geomorphology* **274**, 31–49. (doi:10.1016/j.geomorph.2016.09.014)

28. Riva M, Guadagnini L, Guadagnini A, Ptak T, Martac E. 2006 Probabilistic study of well capture zones distribution at the Lauswiesen field site. *J. Contam. Hydrol.* **88**, 92–118. (doi:10.1016/j.jconhyd.2006.06.005)

29. Riva M, Guadagnini A, Fernandez-Garcia D, Sanchez-Vila X, Ptak T. 2008 Relative importance of geostatistical and transport models in describing heavily tailed breakthrough curves at the Lauswiesen site. *J. Contam. Hydrol.* **101**, 1–13. (doi:10.1016/j.jconhyd.2008.07.004)

30. Barahona-Palomo M, Riva M, Sanchez-Vila X, Vazquez-Sune E, Guadagnini A. 2011 Quantitative comparison of impeller-flowmeter and particle-size-distribution techniques for the characterization of hydraulic conductivity variability. *Hydrol. J.* **19**, 603–612. (doi:10.1007/s10040-011-0706-5)

31. Menafoglio A, Guadagnini A, Secchi P. 2016 Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a Bayes space approach. *Water Resour. Res.* **52**, 5708–5726. (doi:10.1002/2015WR018369)

32. Harris CR *et al.* 2020 Array programming with NumPy. *Nature* **585**, 357–362. (doi:10.1038/s41586-020-2649-2)

33. Winter CL, Tartakovsky DM. 2002 Groundwater flow in heterogeneous composite aquifers. *Water Resour. Res.* **38**, 23–1—23–11. (doi:10.1029/2001WR000450)

34. Winter C, Tartakovsky D, Guadagnini A. 2003 Moment differential equations for flow in highly heterogeneous porous media. *Surv. Geophys.* **24**, 81–106. (doi:10.1023/A:1022277418570)

35. Sobol IM. 2001 Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **55**, 271–280. (doi:10.1016/S0378-4754(00)00270-6)

36. Zhang C, Chu J, Fu G. 2013 Sobol's sensitivity analysis for a distributed hydrological model of Yichun River Basin, China. *J. Hydrol.* **480**, 58–68. (doi:10.1016/j.jhydrol.2012.12.005)

37. Oladyshkin S, De Barros F, Nowak W. 2012 Global sensitivity analysis: a flexible and efficient framework with an example from stochastic hydrogeology. *Adv. Water Res.* **37**, 10–22. (doi:10.1016/j.advwatres.2011.11.001)

38. Ciriello V, Di Federico V, Riva M, Cadini F, De Sanctis J, Zio E, Guadagnini A. 2013 Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer. *Stochast. Environ. Res. Risk Assessm.* **27**, 945–954. (doi:10.1007/s00477-012-0616-7)

39. Gan Y, Duan Q, Gong W, Tong C, Sun Y, Chu W, Ye A, Miao C, Di Z. 2014 A comprehensive evaluation of various sensitivity analysis methods: a case study with a hydrological model. *Environ. Modell. Softw.* **51**, 269–285. (doi:10.1016/j.envsoft.2013.09.031)

40. Razavi S, Gupta HV. 2015 What do we mean by sensitivity analysis? The need for comprehensive characterization of 'global' sensitivity in Earth and Environmental systems models. *Water Resour. Res.* **51**, 3070–3092. (doi:10.1002/2014WR016527)
41. Maina FZ, Guadagnini A. 2018 Uncertainty quantification and global sensitivity analysis of subsurface flow parameters to gravimetric variations during pumping tests in unconfined aquifers. *Water Resour. Res.* **54**, 501–518. (doi:10.1002/2017WR021655)
42. Sahu Q, Fahs M, Hoteit H. 2022 Optimization and uncertainty quantification method for reservoir stimulation through carbonate acidizing. *ACS Omega* **8**, 539–554. (doi:10.1021/acsomega.2c05564)
43. Harrison M, Riva M, Mousavi Nezhad M, Guadagnini A. 2024 Estimation of auto-covariance of log hydraulic conductivity from Generalized Sub-Gaussian porosity and particle size random fields. USGS. (https://pubs.usgs.gov/of/2003/ofr-03-420/spreadsheets.htm)
44. Harrison M, Riva M, Mousavi Nezhad M, Guadagnini A. 2024 Estimation of auto-covariance of log hydraulic conductivity from Generalized Sub-Gaussian porosity and particle size random fields: Dataset. Zenodo. (https://zenodo.org/records/10150952)