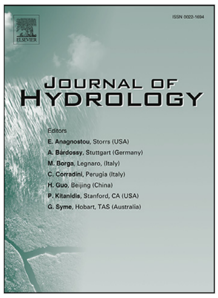
Journal Pre-proof



A Gaussian-mixture based stochastic framework for the interpretation of spatial heterogeneity in multimodal fields

Martina Siena, Chiara Recalcati, Alberto Guadagnini, Monica Riva

PII: S0022-1694(22)01419-6
DOI: https://doi.org/10.1016/j.jhydrol.2022.128849
Reference: HYDROL 128849

To appear in: *Journal of Hydrology*

A Gaussian-Mixture based stochastic framework for the interpretation of spatial heterogeneity in multimodal fields

Martina Siena[1], Chiara Recalcati[1], Alberto Guadagnini[1], Monica Riva[1*]

*corresponding author: monica.riva@polimi.it

Abstract

We provide theoretical formulations enabling characterization of spatial distributions of variables (such as, e.g., conductivity/permeability, porosity, vadose zone hydraulic parameters, and reaction rates) that are typical of hydrogeological and/or geochemical scenarios associated with randomly heterogeneous geomaterials and are organized on various scales of heterogeneity. Our approach and ensuing formulations embed the joint assessment of the probability distribution of a target variable, Y, and its associated spatial increments, ΔY, taken between locations separated by any given distance (or lag). The spatial distribution of Y is interpreted through a bimodal Gaussian mixture model. The modes of the latter correspond to an indicator random field which is in turn related to the occurrence of different processes and/or geomaterials within the domain of observation. The distribution of each component of the mixture is governed by a given length scale driving the strength of its spatial correlation. Our model embeds within a unique theoretical framework the main traits arising in a stochastic analysis of these systems. These include (i) a slight to moderate asymmetry in the distribution of Y and (ii) the occurrence of a dominant peak and secondary peaks

in the distribution of ΔY whose importance changes with lag together with the moments of the distribution. This causes the probability distribution of increments to scale with lag in way that is consistent with observed experimental patterns. We analyze the main features of the modeling and parameter estimation framework through a set of synthetic scenarios. We then consider two experimental datasets associated with different processes and observation scales. We start with an original dataset comprising microscale reaction rate maps taken at various observation times. These are evaluated from Atomic Force Microscopy (AFM) imaging of the surface of a calcite crystal in contact with a fluid and subject to dissolution. Such recent high resolution imaging techniques are key to enhance our knowledge of the processes driving the reaction. The second dataset is a well established collection of Darcy-scale air-permeability data acquired by Tidwell and Wilson (1999)[Water Resour Res, 35, 3375–3387] on a block of volcanic tuff through minipermeameters associated with various measurement scales.

1 Introduction

An assumption that often underlies stochastic analyses of hydrogeological, geochemical, or other Earth system quantities of interest is that these can be depicted as Gaussian random fields. Otherwise, sets of observations of a broad range of variables are characterized by sample probability distributions associated with distinctive traits that are not compatible with those typical of Gaussian fields. Of particular interest to our study are the following documented patterns: (i) the multimodal behavior of probability density function (PDF) of a given quantity, Y, emerging at certain scales of inspection, and (ii) the observation that the shape of the PDF of (spatial) increments of Y, $\Delta Y(\mathbf{s}) = Y(\mathbf{x}+\mathbf{s}) - Y(\mathbf{x})$ evaluated over the separation distance (or lag) $\mathbf{s}$, tends to change with lag.

In this context, a scaling behavior of the sample distribution of increments has been observed for several variables. For example, permeability (Painter, 1996; Riva et al., 2013), porosity (Painter, 1996; Guadagnini et al., 2014), hydraulic conductivity (Liu and Molz, 1997; Guadagnini et al., 2013; Meerschaert et al., 2004), and mineral dissolution rates observed at the microscale (Siena et al., 2021) are documented to be characterized by distributions of incremental values (ΔY) whose moments and main traits vary with $\mathbf{s}$ in a way that is not consistent with the assumption that Y be modeled as a Gaussian field. With specific reference to hydrogeological settings, when considering a porous medium whose internal architecture comprises different zones (or regions), each associated with a given geomaterial, attributes such as conductivity/permeability within each region can be viewed as characterized by a unimodal distribution (see, e.g., Winter et al., 2003 and references therein). In this context, Riva et al. (2015) and Guadagnini et al. (2018) suggested that the way the PDF of spatial increments of porosity and permeability scale with lag can be captured through a Generalized Sub-Gaussian (GSG) model. The latter embeds the Gaussian model as a special case. A similar approach has been adopted by Siena et al. (2020, 2021) in the context of their statistical analyses of mineral dissolution rates observed at the micro-scale through the use of modern Vertical Scanning Interferometry (VSI) and Atomic Force Microscopy (AFM).

Otherwise, conductivity fields in geologic media have sometimes been modeled upon relying on a multimodal distribution (see, e.g., Winter et al., 2003 and references therein). Since the early work of Journel (1983), Desbarats (1987), Rubin and Journel (1991), or Rubin (1995), this behavior has been recognized to arise from a homogenization within a unique population of conductivity values that are otherwise linked to regions characterized by differing geological attributes. In this framework, a stochastic approach based on a model encompassing a unique scale of heterogeneity might not be adequate to represent such composite media within which various processes and/or geomaterials coexist across a given spatial window of observation.

A description of a spatial random field as a statistically stationary system characterized through a multimodal model entails considering (i) the random geometry of the various regions (or clusters) identified across the system and (ii) the spatial distribution of the quantity of interest within each of these regions (Winter et al., 2003). In this setting, Rubin and Journel (1991) view the random function of interest, $Y(\mathbf{x})$, as a sum of $m = 1, ..., M$ Gaussian components, $Y_m(\mathbf{x})$, each weighted by a (statistically homogeneous random) indicator function. These authors associate the latter with the spatial distribution of the M zones/clusters across the domain. Each random field $Y_m(\mathbf{x})$ is described through its spatial structure and is typically assumed to be independent from the others and from the indicator function. Rubin (1995) considers a porous system composed by $M = 2$ distinct geomaterials and provides analytical formulations for the first two statistical moments (i.e., mean and variance) and for the covariance of $Y(\mathbf{x})$. Lu and Zhang (2002) further extend the above mentioned studies to include in the theoretical formulation a relationship between the covariance structure of the indicator and a characteristic length describing the spatial arrangement of the zones associated with the various geomaterials. Such a length scale is characterized using a Markov chain model, as expressed by Carle and Fogg (1997). Recent applications of these concepts are illustrated by, e.g., Dai et al. (2020), to assess the impact of the internal architecture of a sedimentary porous medium on solute plume dispersion; Gournelos et al. (2020), for the interpretation of the statistical behavior of monthly water discharge and suspended sediment load; and Jia et al. (2022) for the simulation of synthetic long-term time series of streamflow data.

Our study aims at extending the theoretical framework underpinning a stochastic description of a composite random field through a stationary multimodal distribution. In this broad framework, we derive rigorous formulations associated with the PDF of spatial increments of a given quantity of interest to embed the observed scaling tendencies of such distributions within a unique analytical modeling approach. In this sense, joint analysis of the PDF of data and their increments within a unique theoretical framework that ensures consistency between these two types of information yields improved characterization of the quantity under investigation. For the purpose of our study, we limit the theoretical formulations to a bimodal Gaussian mixture (GMIX). We then provide a general procedure for the estimation of all parameters embedded in the GMIX model. We explore the benefit of our approach upon applying the ensuing theoretical analysis to interpret two experimental datasets. These represent different processes and scales of observation and are characterized by stark bimodal traits.

The first experimental dataset considered comprises a collection of reaction rate maps that we obtain from direct observation of (microscale) surface topography of a calcite crystal subject to dissolution. Detailed fundamental knowledge about these types of dissolution/precipitation processes is usually demanded in the context of modeling of hydrogeochemical system dynamics. High resolution imaging techniques such as AFM or VSI enable direct observation of the processes taking place at the solid-fluid interface of a mineral. These experimental techniques have contributed to markedly enhance our understanding about reaction kinetics (see, e.g., Lüttge et al., 2019 and references therein). Experimental observations reveal that dissolution/precipitation processes at the microscale are affected by a wide variety of factors. These include, e.g., defects in the crystal lattice or inclusions (Fischer et al., 2014), which result in a remarkable heterogeneity in the reaction kinetics. Several authors (Lüttge et al., 2013, 2019; Fischer et al., 2012, 2014) suggest to rely on a stochastic approach and treat reaction rates as random fields. Some preliminary studies on stochastic characterizations are available. These are based, e.g., on a Generalized Extreme Value (Brand et al., 2017) or a GSG (Siena et al., 2021) model. Otherwise, further work is needed to comprehensively address the complex system behavior documented at such scales. In this setting, observed bimodal (or multimodal) traits of the PDF of reaction rates are linked to the occurrence of diverse mechanisms of dissolution of the solid surface in contact with the reacting fluid. As we show in Section 4.1, these mechanisms give rise to distinct regions across the observation domain.

The relative proportion of these zones evolves in time and imprints the parameters characterizing the ensuing bimodal random field of reaction rates.

We then consider a dataset representative of a Darcy-scale collection of air-permeability data (Tidwell and Wilson, 1999, 2002). These are acquired on the faces of a block of volcanic tuff through four minipermeameters having different inner radius. As such, each dataset is related to a given measurement/observation scale. Tidwell and Wilson (1999) show that log-permeability values sampled across the whole domain of investigation are characterized by a bimodal frequency distribution. The latter becomes more and more manifest as the scale of observation increases.

Our study is structured as follows. Section 2 illustrates the theoretical formulation of the GMIX model and our original developments associated with the probability distributions of incremental values of a bimodal Gaussian random field. In Section 3 we propose a parameter estimation procedure and test it on a collection of synthetically generated GMIX fields. Section 4 illustrates the analysis and interpretation of the two experimental datasets described above through the GMIX model. Our conclusions are presented in Section 5.

2 Stochastic model framework

2.1 Multimodal Gaussian Mixture

We consider $Y(\mathbf{x})$ to be a spatial random field exhibiting multimodal behavior across a given domain of interest, described as (e.g., Rubin, 1995; Lu and Zhang, 2002; Dai et al., 2020)

$$Y(\mathbf{x}) = \sum_{m=1}^{M} I_m(\mathbf{x}) Y_m(\mathbf{x}), \tag{1}$$

where M is the number of independent and mutually-exclusive modes (or components) of $Y(\mathbf{x})$, $Y_m(\mathbf{x})$ is the $m-th$ component evaluated at (vector) location $\mathbf{x}$, and I_m is an indicator random field independent of Y_m and defined as

$$I_m(\mathbf{x}) = \begin{cases} 1 & \text{if component } m \text{ occurs at } \mathbf{x} \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Note that I_m follows a Bernoulli distribution with mean $p_m(\mathbf{x}) = E\{I_m(\mathbf{x})\}$ (which corresponds to the relative proportion of I_m across the domain, under ergodic conditions, $E\{\cdot\}$ denoting ensemble expectation), and variance $\text{Var}\{I_m(\mathbf{x})\} = p_m(\mathbf{x})(1-p_m(\mathbf{x}))$. Note also that the following constraint is satisfied

$$\sum_{m=1}^{M} I_m(\mathbf{x}) = 1, \tag{3}$$

at any location $\mathbf{x}$ in the system.

Focusing on a bimodal field (i.e., $M = 2$), Eq. (1) reduces to

$$Y(\mathbf{x}) = I(\mathbf{x})Y_A(\mathbf{x}) + (1 - I(\mathbf{x}))Y_B(\mathbf{x}), \tag{4}$$

where subscripts A and B denote the two modes associated with the random field $Y(\mathbf{x})$. Setting $E\{I(\mathbf{x})\} = p$, the cumulative distribution function (CDF) and the probability density function (PDF) of $Y(\mathbf{x})$ are respectively defined as

$$F_Y(y) = \Pr\{Y \leq y\} = pF_{Y_A}(y) + (1-p)F_{Y_B}(y), \tag{5}$$

and

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y} = pf_{Y_A}(y) + (1-p)f_{Y_B}(y). \tag{6}$$

Here, $F_{Y_m}(y)$ and $f_{Y_m}(y)$ (with $m = A, B$) are the CDF and PDF of component m of the mixture, respectively.

If each component m of Y is characterized by a Gaussian distribution with mean μ_m and variance σ_m^2, i.e., $Y_m \sim N(\mu_m, \sigma_m^2)$, the field $Y(\mathbf{x})$ is a bimodal GMIX, and Eq. (6) reads

$$f_Y(y) = \frac{p}{\sqrt{2\pi}\sigma_A} e^{-\frac{(y-\mu_A)^2}{2\sigma_A^2}} + \frac{(1-p)}{\sqrt{2\pi}\sigma_B} e^{-\frac{(y-\mu_B)^2}{2\sigma_B^2}}. \tag{7}$$

Making use of Eq. (6) the raw moment of Y of order q, $\langle Y^q \rangle$, can be computed as

$$\langle\, Y^q \,\rangle = p\langle\, Y_A^q \,\rangle + (1-p)\langle\, Y_B^q \,\rangle. \tag{8}$$

Therefore the mean of Y can be derived by setting $q = 1$ in Eq. (8), as

$$\mu_Y = p\mu_A + (1-p)\mu_B, \tag{9}$$

and central moments of order q of Y can be evaluated as

$$\langle\, Y'^q \,\rangle = \langle\, (Y-\mu_Y)^q \,\rangle = \sum_{j=0}^{q} \binom{q}{j} (-1)^j \mu_Y^j \langle\, Y^{q-j} \rangle\ . \tag{10}$$

In particular, variance, σ_Y^2, skewness, Sk_Y, and kurtosis, κ_Y, associated with a bimodal GMIX field are evaluated by setting in Eq. (10) $q = 2, 3, 4$, respectively, as

$$\sigma_Y^2 = \langle\, Y'^2 \,\rangle = p\sigma_A^2 + (1-p)\sigma_B^2 + p(1-p)(\mu_A - \mu_B)^2, \tag{11}$$

$$Sk_Y = \frac{\langle Y'^3 \rangle}{\sigma_Y^3} = \frac{p}{\sigma_Y^3}(1-p)(\mu_A - \mu_B)\Big[(1-2p)(\mu_A-\mu_B)^2 + 3(\sigma_A^2 - \sigma_B^2)\Big], \tag{12}$$

$$\begin{aligned}\kappa_Y =& \frac{\langle Y'^4 \rangle}{\sigma_Y^4} = \frac{1}{\sigma_Y^4}\Big\{3p(\sigma_A^4 - \sigma_B^4) + 3\sigma_B^4 \\ &+ p(1-p)(\mu_A-\mu_B)^2\Big[\big(1-3p(1-p)\big)(\mu_A-\mu_B)^2 + 6\big(\sigma_A^2 - p(\sigma_A^2 - \sigma_B^2)\big)\Big]\Big\}.\end{aligned} \tag{13}$$

Eqs. (12) and (13) clearly show that the PDF of a GMIX field can be (i) non-symmetric (i.e., $Sk_Y \neq 0$) even though each component Y_m of Y is symmetric and/or (ii) leptokurtic ($\kappa_Y > 3$, corresponding to a heavy tailed distribution) or platikurtic ($\kappa_Y < 3$), even through components Y_m are mesokurtic (i.e., characterized by $\kappa_Y = 3$). In the hydrogeological context, examples of bimodal features documented for quantities of interest observed across heterogeneous systems include porosity, conductivity, permeability, vadose zone hydraulic properties, and electrical resistivity (e.g., Zhang et al. (2005); Zhang (2009); Riva et al. (2013); Guadagnini et al. (2013, 2015); Russo (2002); Li et al. (2022)).

The extent of these deviations from a Gaussian behavior is controlled by the difference between the component means (i.e., $\mu_A - \mu_B$), the component variances (i.e., σ_A^2 and σ_B^2), and p.

In order to illustrate the main traits of the field considered, Fig. 1 shows the impact of p on the PDF (and related statistical moments) of a GMIX field characterized by $\mu_A - \mu_B = 2$, $\sigma_A^2 = 0.15$, and $\sigma_B^2 = 0.05$. The PDF of Y (see Fig. 1.a) exhibits two peaks and a local minimum located within the interval $y \in [\mu_B, \mu_A]$. As dictated by Eq. (9), the mean of Y varies linearly with p (see Fig. 1.b, note that μ_Y increases with p in our example, since $\mu_A > \mu_B$). The variance of Y exhibits a parabolic behavior with p (Fig. 1.c), as prescribed by Eq. (11). It attains a maximum value at $p = p_{max} = (1+\alpha)/2$, where $\alpha = (\sigma_A^2 - \sigma_B^2)/(\mu_A - \mu_B)^2$ (with $\alpha = 0.05$ in our example). This also implies that σ_Y^2 monotonically increases with p only when $\alpha > 1$. The skewness of Y (see Eq. (12) and Fig. 1.b) vanishes for the two trivial cases $p = 0$ (where $Y = Y_B$) and $p = 1$ (where $Y = Y_A$) and when $p = p_3^{Sk=0} = (1+3\alpha)/2$. Note that the PDF of Y is right-skewed ($Sk_Y > 0$) for $p \in (0, p_3^{Sk=0})$ and left-skewed ($Sk_Y < 0$) for $p \in (p_3^{Sk=0}, 1)$. When $|\alpha| \geq 1/3$, then $p_3^{Sk=0} \notin (0,1)$ and the PDF is right- (for $\alpha > 1/3$) or left- (for $\alpha < 1/3$) skewed regardless of the component proportions. Fig. 1.c also depicts the trend of the excess kurtosis ($E\kappa_Y = \kappa_Y - 3$) versus p. It can be shown from Eq. (13) that, in addition to the two trivial cases $p = 0, 1$, $E\kappa_Y$ vanishes for $p = p_{3,4}^{Ek=0} = 1/2 + \alpha \pm \sqrt{(1+6\alpha^2)/12}$. Hence, the PDF is platikurtic for $p \in (p_3^{Ek=0}, p_4^{Ek=0})$ and leptokurtic outside this range. If $|\alpha| \geq (3+\sqrt{6})/3$, then $p_{3,4}^{E\kappa=0} \notin (0,1)$ and the PDF is leptokurtic regardless of the component proportions.

2.2 Spatial increments of a Bimodal Gaussian Mixture

Let Y_1 and Y_2 denote the bimodal GMIX, $Y(\mathbf{x})$, at two (spatial) locations, $\mathbf{x}_1$ and $\mathbf{x}_2$, i.e.,

$$Y_1 = Y(\mathbf{x}_1) = I(\mathbf{x}_1)Y_A(\mathbf{x}_1) + (1 - I(\mathbf{x}_1))Y_B(\mathbf{x}_1), \tag{14a}$$

$$Y_2 = Y(\mathbf{x}_2) = I(\mathbf{x}_2)Y_A(\mathbf{x}_2) + (1 - I(\mathbf{x}_2))Y_B(\mathbf{x}_2). \tag{14b}$$

We extend the approach illustrated by Rubin (1995) and obtain the joint PDF of Y_1 and Y_2 as

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) =& \Pr\{I(\mathbf{x}_1) = 1, I(\mathbf{x}_2) = 1\} f_{Y_{A,1},Y_{A,2}}(y_1,y_2) \\ &+ \Pr\{I(\mathbf{x}_1) = 0, I(\mathbf{x}_2) = 0\} f_{Y_{B,1},Y_{B,2}}(y_1,y_2) \\ &+ \Pr\{I(\mathbf{x}_1) = 1, I(\mathbf{x}_2) = 0\} f_{Y_{A,1},Y_{B,2}}(y_1,y_2) \\ &+ \Pr\{I(\mathbf{x}_1) = 0, I(\mathbf{x}_2) = 1\} f_{Y_{B,1},Y_{A,2}}(y_1,y_2), \end{aligned} \tag{15}$$

where

$$f_{Y_{m,1},Y_{m,2}}(y_1,y_2) = \frac{e^{-r}}{2\pi\sigma_m^2\sqrt{1-\rho_m^2}}, \tag{16}$$

with

$$r = \frac{(y_1-\mu_m)^2 + (y_2-\mu_m)^2 - 2\rho_m(y_1-\mu_m)(y_2-\mu_m)}{2\sigma_m^2(1-\rho_m^2)} \text{ and } m = (A, B),$$

is the bivariate PDF of the Gaussian components of the mixture at the two locations. The joint PDF introduced in Eq. (16) is seen to depend on the spatial correlation $\rho_m = \rho_m(\mathbf{x}_1, \mathbf{x}_2)$ of each mode. We recall that, as mentioned above, the two components Y_A and Y_B are assumed to be uncorrelated. Hence, $f_{Y_{A,1},Y_{B,2}} = f_{Y_A}(y_1)f_{Y_B}(y_2)$ and $f_{Y_{B,1},Y_{A,2}} = f_{Y_B}(y_1)f_{Y_A}(y_2)$.

Therefore, Eq. (15) leads to

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) =& E\{I(\mathbf{x}_1)I(\mathbf{x}_2)\} f_{Y_{A,1},Y_{A,2}}(y_1,y_2) \\ &+ E\{[1-I(\mathbf{x}_1)][1-I(\mathbf{x}_2)]\} f_{Y_{B,1},Y_{B,2}}(y_1,y_2) \\ &+ E\{I(\mathbf{x}_1)[1-I(\mathbf{x}_2)]\} f_{Y_A}(y_1)f_{Y_B}(y_2) \\ &+ E\{[1-I(\mathbf{x}_1)]I(\mathbf{x}_2)\} f_{Y_B}(y_1)f_{Y_A}(y_2). \end{aligned} \tag{17}$$

Considering that

$$E\{I(\mathbf{x}_1)I(\mathbf{x}_2)\} = [E\{I(\mathbf{x})\}]^2 + C_I(\mathbf{x}_1,\mathbf{x}_2) = p^2 + C_I(\mathbf{x}_1,\mathbf{x}_2), \tag{18}$$

where $C_I(\mathbf{x}_1,\mathbf{x}_2)$ is the covariance of the indicator field $I(\mathbf{x})$, Eq. (17) can be rewritten as

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) = {} & [p^2 + C_I(\mathbf{x}_1,\mathbf{x}_2)] f_{Y_{A,1},Y_{A,2}}(y_1,y_2) \\ & + [(1-p)^2 + C_I(\mathbf{x}_1,\mathbf{x}_2)] f_{Y_{B,1},Y_{B,2}}(y_1,y_2) \\ & + [p(1-p) - C_I(\mathbf{x}_1,\mathbf{x}_2)]\{f_{Y_A}(y_1)f_{Y_B}(y_2) + f_{Y_B}(y_1)f_{Y_A}(y_2)\}. \end{aligned} \tag{19}$$

In the following we derive the analytical formulation for the PDF of the omnidirectional spatial increments $\Delta Y(s) = Y_1 - Y_2$ ($s = \mid \mathbf{x}_1 - \mathbf{x}_2 \mid$). Second-order stationarity is assumed for all random fields, i.e., $C_I(\mathbf{x}_1,\mathbf{x}_2) = C_I(s)$ and $\rho_m(\mathbf{x}_1,\mathbf{x}_2) = \rho_m(s)$. The probability distribution of $\Delta Y(s)$ can be obtained from the joint PDF of Y as

$$F_{\Delta Y}(\Delta y) = \Pr\{\Delta Y \le \Delta y\} = \int_{y_2=-\infty}^{+\infty}\int_{y_1=-\infty}^{\Delta y + y_2} f_{Y_1,Y_2}(y_1,y_2)\,dy_1 dy_2. \tag{20}$$

Making use of Eq. (19) and recalling that $f_{\Delta Y}(\Delta y) = \frac{dF_{\Delta Y}(\Delta y)}{d(\Delta y)}$ leads to

$$\begin{aligned} f_{\Delta Y}(\Delta y) = {} & \frac{p^2 + C_I(s)}{\sqrt{4\pi\sigma_A^2(1-\rho_A)}} e^{-\frac{\Delta y^2}{4\sigma_A^2(1-\rho_A)}} + \frac{(1-p)^2 + C_I(s)}{\sqrt{4\pi\sigma_B^2(1-\rho_B)}} e^{-\frac{\Delta y^2}{4\sigma_B^2(1-\rho_B)}} \\ & + \frac{p(1-p) - C_I(s)}{\sqrt{2\pi(\sigma_A^2+\sigma_B^2)}}\left(e^{-\frac{(\Delta y - \mu_A + \mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}} + e^{-\frac{(\Delta y + \mu_A - \mu_B)^2}{2(\sigma_A^2+\sigma_B^2)}} \right). \end{aligned} \tag{21}$$

The analytical expression of $f_{\Delta Y}(\Delta y)$ depends on (i) variances (σ_A^2 and σ_B^2) and correlation functions (ρ_A and ρ_B) associated with each of the mixture components, (ii) the difference between the component means ($\mu_A - \mu_B$) and (iii) mean (p) and covariance (C_I) of the indicator field. Fig. 2.a shows a graphical depiction of Eq. (21) for various lags, obtained upon relying on the exemplary set of parameters used for Fig. 1 and considering $p = 0.2$. For illustration purposes, we consider an isotropic exponential model to describe the above mentioned indicator covariance function, i.e., $C_I(s) = \sigma_I^2\rho_I(s) = \sigma_I^2 e^{-s/\lambda_I}$, ($\lambda_I$ and $\sigma_I^2 = p(1-p)$ being the correlation length and variance of I, respectively) and for ρ_m, i.e., $\rho_m = e^{-s/\lambda_m}(m = A, B)$, λ_m being the correlation length of component Y_m. Here, for illustration purposes, we set $\lambda_A = \lambda_B = 6$ and $\lambda_I = 6.4$. It can be noted that the PDF of ΔY is always (i) symmetrical with respect to zero; and (ii) characterized by a dominant central peak (located at $\Delta y = 0$ and controlled by the first two terms in Eq. (21)) and two lateral peaks (controlled by the last term in Eq. (21) and located at $\Delta y \approx \pm(\mu_A - \mu_B)$). Fig. 2.a also reveals that the relative importance of the lateral peaks increases (at the expense of the central peak) as lag increases. This behavior is driven by $C_I(s)$ and $-C_I(s)$ that are seen to multiply terms related to the central and lateral peaks in Eq. (21), respectively. As lag increases, $C_I(s)$ decreases and the difference between the height of the central and lateral peaks tends to be reduced. One can also see that the correlation functions ρ_A and ρ_B appear only within the first 2 terms of Eq. (21). Thus, their dependence on lag can only affect the central peak of the PDF of ΔY.

Statistical moments of the incremental variables can then be readily evaluated from Eq. (21). Mean and all odd-order moments of ΔY are identically zero. Second moment of ΔY reads

$$\begin{aligned} \langle \Delta Y^2 \rangle = 2\big\{ & p^2\sigma_A^2(1-\rho_A) + (1-p)^2\sigma_B^2(1-\rho_B) \\ & + p(1-p)\left[(1-\rho_I)(\mu_A-\mu_B)^2 + \sigma_A^2(1-\rho_A\rho_I) + \sigma_B^2(1-\rho_B\rho_I)\right]\big\}. \end{aligned} \tag{22}$$

Since $C_Y = \sigma_Y^2 - \gamma_Y$, where $\gamma_Y = \langle \Delta Y^2 \rangle/2$ is the variogram of Y, Eqs. (11) and (22) allow evaluating the covariance of Y as

$$C_Y = p^2\sigma_A^2\rho_A + (1-p)^2\sigma_B^2\rho_B + p(1-p)\rho_I\left[(\mu_A-\mu_B)^2 + \sigma_A^2\rho_A + \sigma_B^2\rho_B\right] \tag{23}$$

The integral scale of Y, λ_Y, can be computed by integrating Eq. (23). As also discussed by Lu and Zhang (2002), λ_Y can be larger or smaller than the integral scale of the two modes and of the indicator, depending on the value of p, σ_A^2, σ_B^2 as well as on the correlation of the indicator field. Fig. 2.b depicts the variogram as a function of lag for various values of p. Each curve attains the corresponding sill, σ_Y^2 (dashed horizontal lines), for large lags. Consistent with Fig. 1.b, $\sigma_Y^2(p=0.6) > \sigma_Y^2(p=0.4) > \sigma_Y^2(p=0.8) > \sigma_Y^2(p=0.2)$. It can be noted that the same ordering holds also for the variogram values at any given lag.

The fourth-order moment of ΔY is

$$\begin{aligned}\langle \Delta Y^4 \rangle = 2\big\{&6p^2\sigma_A^4(1-\rho_A)^2 + 6(1-p)^2\sigma_B^4(1-\rho_B)^2 \\ &+ 6p(1-p)\rho_I\left[\sigma_A^4(1-\rho_A)^2 + \sigma_B^4(1-\rho_B)^2\right] \\ &+ p(1-p)(1-\rho_I)\left[(\mu_A-\mu_B)^4 + 3(\sigma_A^2+\sigma_B^2)(2(\mu_A-\mu_B)^2 + (\sigma_A^2+\sigma_B^2))\right]\big\}.\end{aligned} \tag{24}$$

The analytical expression for the kurtosis, $\kappa_{\Delta Y} = \langle \Delta Y^4\rangle/\langle \Delta Y^2\rangle^2$, associated with the increments of a Gaussian mixture can then be derived from Eqs. (22) and (24). We recall that this statistical moment quantifies the tailedness of the distribution and its dependence on lag is a distinctive element of the scaling behavior exhibited by the PDFs of increments of a GMIX field. Fig. 2.c depicts excess kurtosis, $E\kappa_{\Delta Y} = \kappa_{\Delta Y} - 3$, as a function of lag for various values of p. All curves exhibit a monotonic trend. The value of $E\kappa_{\Delta Y}$ is seen to increase (indicating tails that become heavier) as s decreases. This pattern is starkly consistent with the behavior observed for several Earth and environmental variables (Riva et al., 2015 and references therein). These results clarify that the increments of a GMIX field exhibit clear non-Gaussian traits, despite each component of the mixture being Gaussian. As noted above, they also show that the PDFs of increments tend to change with lag due the action of the degree of spatial correlation of the two Gaussian components of the mixture and of the indicator field. Values of $E\kappa_Y$ are also depicted in Fig. 2.c (dashed horizontal lines) and are such that $E\kappa_Y(p=0.2) > E\kappa_Y(p=0.8) > E\kappa_Y(p=0.4) > E\kappa_Y(p=0.6)$. The same relative order is maintained also by the values of $E\kappa_{\Delta Y}$. Note that values of $E\kappa_{\Delta Y}$ are negative (i.e., indicating platikurtic distributions of increments) at large values of s for $p > 0.2$.

2.3 Parameter estimation

In our analyses of synthetic and experimental datasets (Section 3 and 4), we assume a given set of N observations to be sampled from a GMIX field, Y. We infer the 5 parameters of Y (i.e., μ_A, μ_B, σ_A^2, σ_B^2, and p in Eq. (7)) by relying on a well-established Maximum Likelihood (ML) approach, implemented through an iterative Expectation-Maximization (E-M) procedure (McLachlan and Krishnan, 2008; Gournelos et al., 2020). According to the latter, each iteration consists of (i) the Expectation step (E-step), aimed at evaluating the (posterior) probability that each observation belongs to the mixture components, on the basis of an initial GMIX parameter set (or the parameter set obtained at the previous iteration); and (ii) the Maximization step (M-step), which uses the information from the E-step to estimate the GMIX parameter set maximizing the likelihood function. The algorithm stops when the increase of the likelihood function between two subsequent iterations is smaller than a prescribed threshold. Note that E-M suffers from the typical issues associated with ML approaches, i.e., uniqueness, identifiability, and stability (Carrera and Neuman, 1986). To address the issue of the sensitivity of results to parameter initialization, application of the E-M algorithm is repeated n times, each with a new set of initial parameters. Estimates of model parameters are then considered to correspond to the parameter set providing the highest likelihood among the n runs (McLachlan and Krishnan, 2008). The number of runs, n, is case specific and must be set through a stability analysis of the algorithm output. For the scenario here considered, the selection of $n = 40$ allows evaluating stable estimates of μ_m, σ_m^2, and p. Additional details are provided in the Supplementary material. Note that this procedure can be employed to obtain a fuzzy (or soft) clustering of the data: each observation is assigned to each mixture component with a given probability rather than to a unique component, as it would result from a hard-clustering approach. It is otherwise noted that the spatial arrangement of the categories must be known to estimate the parameters (λ_I, λ_A, and λ_B) of the correlation functions ρ_I, ρ_A, and ρ_B that drive the statistical behavior of the increments ΔY (see Eq. (21)). This can be achieved by (i) assigning each observation to the mixture component with which the largest posterior PDF is associated, (ii) compute the spatial increments $\Delta I(s)$ of the underlying indicator field and the corresponding sample correlation function $\tilde{\rho}_I(s)$, (iii) compute the spatial increments ΔY_m ($m = A, B$) within the regions associated with each Gaussian component and the corresponding sample correlation function $\tilde{\rho}_m(s)$.

An estimate of $\tilde{\lambda}_I$ can be obtained according to the following two approaches:

- method 1 - fit $\tilde{\rho}_I(s)$ with a suitable theoretical model (e.g., an exponential model or other);
- method 2 - evaluate the mean length of the indicator field (l_A, see Section 3) and estimate $\tilde{\lambda}_I$ as $\tilde{\lambda}_I = (1-p)l_A$ (Lu and Zhang, 2002).

An estimate of $\tilde{\lambda}_m$ can obtained by fitting $\tilde{\rho}_m(s)$ with a suitable theoretical model.

3 Synthetic case study

Multiple realizations of synthetic GMIX fields are generated to provide a transparent assessment of the reliability of the parameter estimation strategy described in Section 2.3 to be applied for the interpretation of the main traits displayed by the key statistics and empirical densities associated with experimental datasets (see Section 4). The generation procedure relies upon: (*i*) a Transition Probability simulation approach (which takes advantage of the widely tested code T-PROGS; e.g., Carle and Fogg, 1996, 1997) for the indicator field, $I(\mathbf{x})$; and (*ii*) a sequential Gaussian simulation framework (based on the broadly used and tested code SGSIM; e.g., Deutsch and Journel, 1998) for the Gaussian fields $Y_A(\mathbf{x})$ and $Y_B(\mathbf{x})$.

A set of $N = 100$ unconditional realizations of $I(\mathbf{x})$ are generated on a two-dimensional regular grid comprising 100 × 100 nodes by setting $p = 0.2$ and $l_A = 8.0$ (i.e., $\lambda_I = (1-p)l_A = 6.4$). The value of Y in each node of the grid is computed via Eq. (4) where Y_A and Y_B are obtained by two sets of N unconditional realizations of Gaussian random fields characterized by and exponential covariance function with $\lambda_A = \lambda_B = 6$, $\mu_A = 2.5$, $\mu_B = 0.5$, $\sigma_A^2 = 0.15$ and $\sigma_B^2 = 0.05$. Additional details are provided in the Supplementary material.

Fig. 3.a depicts an exemplary realization of the GMIX field obtained. Each realization is treated as a dataset to which the parameter estimation procedures detailed in Section 2.3 can be applied. Comparison between estimated and input model parameters enables us to assess the reliability of the proposed inference methodology.

Fig. 3.b depicts the binary categorical (i.e., indicator) field that is inferred from the E-M and clustering procedure applied to the synthetic dataset in Fig. 3.a. Given the indicator field and the ensuing sample correlation function $\tilde{\rho}_I(s)$, an estimate of $\tilde{\lambda}_I$ can then be obtained according to method 1 and/or 2 introduced in Section 2.3. Note that one of the categories needs to be characterized in terms of its mean length, l_m, when considering method 2. The white portion of the domain (category B) in Fig. 3.b corresponds to the so-called *background category*. In a categorical random field, this term is commonly adopted to identify the category that fills in the space within which other categories are distributed. As an example, in geostatistical applications associated with hydrogeological scenarios, categories are represented by the various lithofacies of a depositional environment and the background geomaterial is typically associated with the category characterized by the lowest deposition energy (Carle and Fogg, 1997). In our datasets, we evaluate the mean length l_A of the category that is not in the background (black regions in Fig. 3.b, associated with category A) by averaging over all of the connected sets of A (i) the length of the sides of the bounding box (see green lines in Fig. 3.b) and (ii) the diameter of the inscribed maximal balls (red circles in Fig. 3.b).

Fig. 4 collects values of GMIX parameters estimated for all the $N = 100$ synthetic realizations. The average of the estimates is always satisfactorily close to the corresponding input value. The mean squared relative deviation (MSRD) between input and estimated parameter values is evaluated over the whole collection of realizations. On the basis of this metric one can note that estimates are overall more accurate for (i) the parameters of the Gaussian distribution associated with category B

(except for the mean value), that occupies a larger portion of the domain as compared to category A and (ii) the correlation scale $\tilde{\lambda}_I$ obtained via method 2 as compared against its counterpart based on method 1. In light of these results, method 2 is considered for the estimation of λ_I in the context of the experimental datasets analyzed in Section 4.

4 Application to experimental scenarios

4.1 Microscale geochemical dataset

The first dataset we consider is an original collection of experimental microscale dissolution rate maps. The dataset is obtained from *in-situ* and real-time high-resolution measurements of topography, $z(x, y, t)$, of the $\{104\}$ crystallographic surface of a calcite sample subject to dissolution in deionized water via AFM imaging.

The spatial distribution of reaction rates, $R(x, y, t)$ [mmol cm^{-2} s^{-1}], can be obtained from the difference between two topography maps associated with two observation times separated by a temporal interval Δt as

$$R(x, y, t) = \frac{z(x, y, t) - z(x, y, t + \Delta t)}{V_m \Delta t}, \tag{25}$$

where $V_m = 36.9$ cm^{3}mol^{-1} is calcite molar volume.

Estimates of reaction rates via high resolution imaging techniques such as AFM or VSI provide remarkable insights on the complexity of mechanisms involved in these types of reactions. In this context, calcite $\{104\}$ is widely studied due to its abundance in natural environments and to its high reactivity (Heberling et al., 2010). When exposed to a solution, the crystal surface is affected by a variety of dissolution modes. This results in a marked spatial heterogeneity of the dissolution rate. Such variability can be aptly interpreted through a stochastic characterization (e.g., Fischer et al., 2012; Lüttge et al., 2019). The prevailing dissolution mechanism depends on the distance from chemical equilibrium. The latter is usually assessed in terms of the solution saturation state $\Omega = IAP/K_s$. Here, $IAP = a_{Ca^{2+}} \cdot a_{Co_3^{2-}}$ is the ion activity product, evaluated as the product of the activities of the species in the solution, and $K_s = a_{Ca^{2+},eq} \cdot a_{Co_3^{2-},eq}$ is the solubility product constant, given by the product of the species activities at equilibrium. It is noted that Ω approaches unity as the system tends to chemical equilibrium. We perform dissolution experiments in far-from-equilibrium conditions, i.e., $\Omega \in [0, 0.007]$ in our setting (Teng, 2004). Within this saturation range, calcite dissolves by nucleation of etch-pits, which may form randomly on the terraces of the crystal and/or in the presence of linear defects in the crystal lattice. For a detailed study of the dependence of the dissolution mode on Ω, we refer to Bouissonnié et al. (2018) and Teng (2004). Details on the experimental acquisition are provided in Appendix A. Prior to analysis, AFM data often require a signal processing phase, as reported by Marinello et al. (2010). Following Siena et al. (2021), we perform a preliminary detrend subtracting the best fitting second order polynomial function from the raw topography data. This enables us to remove the distortion induced by the AFM scanning and to obtain the fluctuation of calcite topography about its mean, i.e. $z'(x, y, t) = z(x, y, t) - \langle z(t) \rangle$. Fig. 5 collects maps of z' acquired during the dissolution experiment at constant time intervals $\Delta t = 13$ min. These provide a qualitative appraisal of the temporal evolution of the crystal surface during the reaction. We observe two main topography patterns. These are respectively related to: (i) the spreading of a multilayer (deep) etch-pit (MP) in the bottom left corner; and (ii) the nucleation, spreading and coalescence of several monolayer (shallow) etch-pits (mP) taking place on the terrace. As a consequence, topography maps can be subdivided into two regions, namely *Multilayer Region* and *Terrace Region*. These patterns are consistent with published literature analyses regarding dissolution in far-from-equilibrium conditions (e.g., Teng, 2004; Bouissonnié et al., 2018).

Similar to Siena et al. (2021), we view the dissolution rate as a random field. We express it as the sum of an average dissolution rate, $\langle R(t) \rangle = [\langle z(x, y, t) \rangle - \langle z(x, y, t + \Delta t) \rangle] / V_m \Delta t$, which is constant across the whole spatial domain of investigation, and a local fluctuation, $R'(x, y, t)$, which provides information about the spatial variability of the reaction rate $R(t)$, i.e., $R(x, y, t) = \langle R(t) \rangle + R'(x, y, t)$. Our statistical analysis is conducted on $R'(x, y, t)$, which is evaluated as

$$R'(x, y, t) = \frac{z'(x, y, t) - z'(x, y, t + \Delta t)}{V_m \Delta t}. \quad (26)$$

Figs. 6.A.a-d depict spatial distributions of $R'(x, y, t)$ resulting from the difference between z' maps separated by $\Delta t = 13\,\mathrm{min}$. The corresponding sample PDFs of R' are depicted in Figs. 6.B. All PDFs exhibit a prominent peak at $R' \approx 0$ and a secondary peak for $R' > 0$. From a qualitative standpoint, these results suggest that all points that belong to the same topography region (either Terrace or Multilayer) at times t_i and $t_i + \Delta t$ contribute to the highest peak. Otherwise, the secondary peak is driven by values of rate that are associated with locations that transition from one topography region to the other during the time interval Δt, following the spreading of the multilayer etch pit.

The observed bimodal trait of the sample PDFs of $R'(x, y, t)$ is consistent with an interpretation based on the GMIX stochastic framework introduced in Section 2. We denote hereafter as components A and B those associated with the peak at $R' > 0$ and at $R' \approx 0$, respectively.

We compute spatial increments of dissolution rate, ΔR, at various separation distances. Sample statistics are evaluated considering omnidirectional increments, with the only exception of the direction parallel to the AFM acquisition (denoted as x in Fig. 5.e), to avoid spurious correlation originated from measurement artifacts. This is consistent with the study of Marinello et al. (2010) who show that AFM measurements are often affected by *stripe noise*, i.e., a distortion of the signal occurring along the principal scanning direction. Figs. 6.C.a-d depict sample PDFs of increments ΔR for lags s = 16, 32, 64 dl (dl = 11.7 nm; see Appendix A), encompassing short and large distances relative to the size of the domain. All of these PDFs share some common features with their counterparts described in Section 2.2 in the context of the GMIX framework, i.e., they display (i) an overall symmetric behavior, (*ii*) the presence of a dominant peak coupled with lateral peaks, and (*iii*) a tendency to change their main traits with lag, denoting a scaling behavior.

The GMIX parameters as well as those associated with the distribution of $f_{\Delta R}$ are assessed according to the procedure detailed in Section 2.3, λ_I being estimated via method 2. The analytical formulations of $f_{R'}$ and $f_{\Delta R}$ (Eqs. (7) and (21)) obtained upon considering the GMIX parameters estimated at each time are juxtaposed to their sample counterparts in Figs. 6.Ba-d and 6.Ca-d, revealing a remarkably satisfactory agreement.

The analysis of the GMIX parameters at different times provides insights on the temporal evolution of the mechanisms driving the dissolution reaction. Temporal variations of parameter values are mainly linked to component A of the mixture. This is related to the observation that the area associated with category A is subject to higher relative variations than the corresponding one related to category B, with an average variation of $\sim 36\%$ and $\sim 4\%$, respectively, across the total temporal window analyzed. The overall temporal increase of p and λ_I (Fig. 7.a and Fig. 7.b) reflects a progressive growth of the area associated with category A. Such increasing trend is consistent with the (approximately) constant horizontal spreading rate, ν, of the MP (see Fig. 7.c) evaluated as (Ruiz-Agudo and Putnis, 2012)

$$\nu = \frac{1}{2}\left(\nu_{ac} + \nu_{ob}\right) = \frac{1}{2}\left(\frac{\Delta l_{ac}}{\Delta t} + \frac{\Delta l_{ob}}{\Delta t}\right), \tag{27}$$

where ν_{ac} and ν_{ob} are the horizontal spreading rate of acute and obtuse steps, respectively. These are estimated as the ratio of Δl_i [nm], i.e., the separation distance between etch pit edges at subsequent times (reported in Fig. 7.d), and the time step Δt [s]. In our case, we can only evaluate the spreading rate of MP acute steps ν_{ac} because no obtuse step fall inside the observation window. It can be noted that the order of magnitude of the results depicted in Fig. 7.c is consistent with existing results recently documented in the literature (e.g., Guren et al., 2020; Dong et al., 2020).

The mean of both components A and B remains almost constant with time (Fig. 7.e). Otherwise, a decreasing temporal trend is observed for the variance of component A, whereas σ_B^2 remains almost constant (Fig. 7.f). The documented pattern suggests that values of the second moment associated with component A progressively becomes more similar to that related to component B. This is also consistent with the observed temporal dampening of the multimodal behavior displayed by the PDF of R'. This trend is also revealed by an observed temporal decrease for Sk_R and $E\kappa_R$ (not shown).

The temporal evolution of the spatial correlation structure of each component of the mixture is inferred from the analysis of ρ_m ($m = A, B$). Fig. 8 depicts the sample spatial correlation associated with regions A and B. The following common traits can be noted at all times: (i) an oscillating behavior at large separation distances for ρ_A and (ii) the presence of a nugget effect for both ρ_A and ρ_B. We relate the oscillations in ρ_A to the small number of points separated by large lags for region A. Otherwise, the second trait could be attributed to the persisting stripe noise, which might especially influence short lags. We consider the exponential with nugget as interpretive model for ρ_m. Fig. 8 juxtaposes theoretical ρ_m values and experimentally-based counterparts. The analytical formulation enables one to grasp the main features associated with the experimental setting. Fig. 7.g depicts λ_A and λ_B versus time. An oscillatory behavior can be noticed, in particular for component B. We relate this trend to the dynamics of the monolayer etch-pits nucleating and spreading on the crystal terrace.

Fig. 9 juxtaposes the analytical curves associated with the GMIX correlation function (i.e., $\rho_R = 1 - (\langle \Delta R^2 \rangle / 2\sigma_R^2)$, $\langle \Delta R^2 \rangle$ being evaluated through Eq. (22)) and excess kurtosis, $E\kappa_{\Delta R}$ (evaluated making use of Eqs. (22) and (24)), of ΔR to their experimental sample counterparts. Theoretical spatial correlation structures shown in Figs. 9.A.a-d exhibit a satisfactory agreement with their sample counterparts, discrepancies being mainly visible at time $t_7 = 52$ min, for intermediate lags. Here, we notice that sample PDFs of ΔR evaluated for component A appear to deviate, albeit slightly, from a Gaussian behavior (not shown). Such a deviation from Gaussianity could also be at the basis of the imperfect agreement observed between sample and theoretical values of $E\kappa_{\Delta R}$. Given the purpose of the current study, we envision to further investigate these elements in future works. In this context, a candidate modeling approach could rely on considering mixtures of Generalized Sub-Gaussian processes. These have been analyzed by, e.g., Riva et al. (2015) and Siena et al. (2020), in the context of systems composed by a single region and include the Gaussian model as a particular case.

The results obtained here through the GMIX modeling framework are remarkably promising for the interpretation of high resolution geochemical data at the microscale. They show that model parameters are strictly linked to the temporal evolution of the surface features driving the dissolution reaction.

4.2 Permeability dataset

We consider a collection of air permeability data acquired by Tidwell and Wilson (1999). Data are sampled on the six faces of a $81 \times 74 \times 63$ cm^3 block of Topopah Spring Tuff. Each face extends across an area of 30×30 cm^2 and measurements are collected according to a uniform sampling grid of 36×36 points (horizontal resolution $\Delta = 0.85$ cm). Data collection relies on four air minipermeameters, each with a given tip-seal (inner minipermeameter radii, r, being $r_1 = 1.5$mm, $r_2 = 3.1$mm, $r_3 = 6.3$mm, $r_4 = 12.7$mm and outer radii being $2r_i$ $(i = 1, 2, 3, 4)$). We follow Tidwell and Wilson (1999) and Siena et al. (2012) and consider the inner radius to be representative of the measurement scale associated with the corresponding data. As observed by these authors, log-permeability data, $Y = \ln k$, display a bimodal character that becomes increasingly manifest as the tip-seal size increases. Tidwell and Wilson (1999) attribute this behavior to the nature of the tuff, where regions of high permeability (henceforth associated with component A of the mixture model we employ for data interpretation) are associated with pumice fragment, and areas of low permeability (hereafter associated with component B) are related to the background matrix.

Here, we focus on permeability data collected on face 1 of the block (see Tidwell and Wilson, 1999 and Figs. 10.A.a-d for a graphical depiction) and interpret $Y'(\mathbf{x}) = Y(\mathbf{x}) - \langle Y \rangle$ as a GMIX random field. Sample PDFs of Y' and ΔY are depicted in Figs. 10.B.a-d and Fig. 10.C.a-d, respectively, together with the GMIX-based solution. Note that there is a generally good agreement between the analytical model for both $f_{Y'}$ and $f_{\Delta Y}$ and their sample counterparts.

As shown by Tidwell and Wilson (1999), the increase in the tip-seal radius yields an overall spatial homogenization of the observations. This element is related to the smoothing effect of an increasing sampling scale on permeability. Boundaries of the pumice clusters become more evident as r increases. Such a trend is fully reflected by the sample PDFs and by the behavior of the GMIX parameters. We observe a decrease in the proportion parameter, p, with increasing r (Fig. 11.a). This is mainly related to the observation that increasing r essentially contributes to embed (i.e., homogenize) in region B isolated pixels attributed to component A, thus reducing the number of data assigned to category A. As a consequence, the length scale characterizing component A, i.e., l_A, increases with r. This leads to a corresponding increase of the correlation scale of the indicator variable, λ_I (Fig. 11.b). The progressive homogenization of the sample fields is also seen to be conducive to a decrease for both μ_A and σ_A^2, that tend to approach μ_B and σ_B^2, respectively (Fig. 11.c and Fig. 11.d).

We investigate the evolution of the spatial correlation of the log-permeability fields associated with each of the two identified clusters / components upon relying on a exponential model. Modeling results are depicted against sample data in Fig. 12. These results suggest that the modeling approach we propose yields an overall satisfactory agreement with available observations. We notice that sample data exhibit oscillations about zero at large lags. This behavior might be captured by more complex interpretive model, e.g., nested models entailing a hole effect contribution. While this would possibly lead to an increased number of model parameters, a rigorous analysis taking into account multiple interpretive models (in a multimodel framework) is beyond the scope of the current study and will be subject of future works. Fig. 11.e depicts the dependence of λ_m on r. As expected, we note an increase of the correlation length of both components, following the increase of measurement scale and ensuing spatial smoothing of the field.

Fig. 13.a depicts the sample correlation function, ρ_Y, inferred from the available data and its analytical counterpart based on the GMIX formulation. These results are complemented by Fig. 13.b which depicts the dependence of the experimentally- and modeling- based excess kurtosis of the increments with lag for the four available values of r. An overall good agreement between experimental and analytical results is observed. This strengthens our confidence about the ability of the model to capture the salient features of the system. One can clearly see that $E\kappa_{\Delta Y}$ increases as s decreases. At short separation distances, values of $E\kappa_{\Delta Y}$ associated with a given lag tend to increase with r. Thus, PDFs of increments at short lags are characterized by heavier tails (i.e., they strongly depart from a Gaussian behavior) as the observation scale r increases. This behavior is consistent with the enhanced importance of the secondary peaks displayed by the PDFs of the increments (see also Figs. 10.C) as well as by the observed increase of λ_I with r.

5 Conclusions

We focus on a stochastic characterization of spatial distributions of hydrogeological and hydrogeochemical quantities which views these as multimodal Gaussian random fields. Our work extends existing formulations (e.g., Lu and Zhang (2002) and references therein) to include in a unique theoretical framework the assessment of the probability distribution of a given quantity of interest and its spatial increments associated with various separation distances (or lags). This enables us to provide a robust characterization of key features of stochastic random fields which are organized on different scales of heterogeneity across the system. These include (i) a slight to moderate asymmetry in the distribution of the target quantity, Y, resulting from the presence of multiple peaks, and (ii) the occurrence of a dominant peak together with multiple secondary peaks in the distribution of increments ΔY. The relative importance of these peaks tends to vary with the lag at which increments of Y are taken across the system giving rise to observable scaling behaviors of the PDF of ΔY. We focus on the particular case of a bimodal Gaussian mixture, whose modes are identified through an indicator random field. The latter is related to the length scale governing the spatial arrangement of zones/regions within which the quantity of interest is randomly distributed. The presence of such regions can be linked to the occurrence of different processes and/or geomaterials within the domain of observation. Each component of the mixture is then characterized by a given length scale driving the spatial correlation of its values and spatial increments within each zone. In this sense, our theoretical framework enables one to infer distributions of quantities of interest through a joint analysis of data about values of the target quantity and its increments to ensures consistency between these two sets of observations. We propose a general procedure to estimate the model parameters, which includes partitioning the domain into the two components of the mixture, A and B. The robustness of the proposed methodology is assessed through extensive tests on a collection of synthetically generated random fields. The modeling framework is then applied to interpret two experimental datasets associated with different windows of observation. The first dataset is typical of geochemical applications and comprises an original collection of microscale reaction rate maps evaluated at various temporal instants from AFM topography measurements of the surface of a calcite crystal subject to dissolution. The second dataset is a collection of (Darcy-scale) air-permeability data acquired by Tidwell and Wilson (1999) on a block of volcanic tuff through minipermeameters associated with various measurement scales.

Our analysis yields the following key conclusions.

1. The analytical expression of the PDF of spatial increments, ΔY, of variables, Y, following a GMIX model displays a symmetric behavior, with a central dominant peak and two secondary peaks. The relative importance of these changes with the separation distance, hence reflecting a clear scaling behavior. This is also documented by the trend of the kurtosis of ΔY, $E\kappa_{\Delta Y}$, which increases as separation distance decreases.

2. The GMIX model captures the main traits exhibited by both experimental datasets considered. A satisfactory agreement is also observed between sample and analytical statistical moments associated with the increments (i.e., correlation function and excess kurtosis) at different lags. This is particularly evident for the permeability dataset. Otherwise, for the microscale geochemical dataset we observe some discrepancies between sample and modeled values for a specific range of separation distances. Here, we notice that sample PDFs of increments of reaction rates (ΔR) evaluated for component A appear to show slight deviations from a Gaussian behavior. These deviations could also be at the basis of the imperfect agreement observed between sample and theoretical values of $E\kappa_{\Delta R}$.

3. Analysis of the temporal behavior of the GMIX parameters provides quantitative insights on the experimental scenarios analyzed. In the case of the microscale geochemical dataset, the temporal behavior of the model parameters is seen to be closely related to the evolution of the observed dissolution patterns. Having at our disposal a tool capable of encapsulating the dynamics of the physical mechanisms taking place at the solid-liquid interface within a robust theoretical framework that can provide a joint accurate description of the statistics of the variable and its increments can be beneficial to transfer information to other spatial scales. Our analysis of Darcy scale permeabilities suggests that the trend of the GMIX parameters can also be related to the characteristic length scale associated with the observations. Here, the documented behavior of the model parameters reflects the main features associated with

the increase of a characteristic length of the measuring device that gives rise, in turn, to data smoothing and homogenization.

As detailed in Section 1, several key environmental variables driving contaminant fate and transport are documented to exhibit non-Gaussian features (even as monitored within the same geomaterial/cluster). While Gaussianity of the quantities associated with each cluster is an underlying hypothesis of the GMIX theoretical framework, we envision to extend our work to mixtures of Generalized sub-Gaussian fields. Such fields have been introduced by Riva et al. (2015) in the context of systems composed by a single cluster and include the Gaussian model as a particular case. Additional elements for future research include (a) the assessment of the uncertainty associated with measurement errors of AFM topography and the way these might affect dissolution rate estimates and/or (b) the exploration of candidate alternatives to be employed in the context of the clustering algorithm, which is a key aspect of the parameter inference framework. In this context, a candidate alternative methodology could rest on a Bayes classifier, which also provides an estimate of the level of uncertainty associated with the data partitioning.

References

Bouissonnié, A., Daval, D., Marinoni, M., and Ackerer, P. (2018). From mixed flow reactor to column experiments and modeling: Upscaling of calcite dissolution rate. Chemical Geology, 487:63–75.

Brand, A. S., Feng, P., and Bullard, J. W. (2017). Calcite dissolution rate spectra measured by in situ digital holographic microscopy. Geochimica et Cosmochimica Acta, 213:317–329.

Carle, S. and Fogg, G. (1996). Transition probability-based indicator geostatistics. Math Geol, 28(4):453–477.

Carle, S. and Fogg, G. (1997). Modelling spatial variability with one and multidimensional continuous-lag markov chains. Math Geol, 29(7):891–918.

Carrera, J. and Neuman, S. P. (1986). Estimation of aquifer parameters under transient and steady state conditions: 2. uniqueness, stability, and solution algorithms. Water Resources Research, 22(2):211–227.

Dai, Z., Zhan, C., Dong, S., Yin, S., Zhang, X., and Soltanian, M. R. (2020). How does resolution of sedimentary architecture data affect plume dispersion in multiscale and hierarchical systems? Journal of Hydrology, 582:124516.

Desbarats, A. J. (1987). Numerical estimation of effective permeability in sand- shale formations. Water Resources Research, 23:273–286.

Deutsch, C. V. and Journel, A. G. (1998). Gslib, geostatistical software library and user's guide. Oxford University Press, New York.

Dong, S., Berelson, W. M., Adkins, J. F., Rollins, N. E., Naviaux, J. D., Pirbadian, S., El-Naggar, M. Y., and Teng, H. H. (2020). An atomic force microscopy study of calcite dissolution in seawater. Geochimica et Cosmochimica Acta, 283:40–53.

Fischer, C., Arvidson, R. S., and Lüttge, A. (2012). How predictable are dissolution rates of crystalline material? Geochimica et Cosmochimica Acta, 98:177–185.

Fischer, C., Kurganskaya, I., Schäfer, T., and Lüttge, A. (2014). Variability of crystal surface reactivity: What do we know? Applied Geochemistry, 43:132–157.

Gournelos, T., Kotinas, V., and Poulos, S. (2020). Fitting a Gaussian mixture model to bivariate distributions of monthly river flows and suspended sediments. Journal of Hydrology, 590:125166.

Guadagnini, A., Neuman, S. P., Nan, T., Riva, M., and Winter, C. L. (2015). Scalable statistics of correlated random variables and extremes applied to deep borehole porosities. Hydrology and Earth System Sciences, 19(2):729–745.

Guadagnini, A., P.Neuman, S., M.G.Schaap, and Riva, M. (2013). Anisotropic statistical scaling of vadose zone hydraulic property estimates near maricopa, arizona. Water Resources Research, 49:8463–8479.

Guadagnini, A., P.Neuman, S., M.G.Schaap, and Riva, M. (2014). Anisotropic statistical scaling of soil and sediment texture in a stratified deep vadose zone near maricopa, arizona. Geoderma, 214-215:217–227.

Guadagnini, A., Riva, M., and Neuman, S. P. (2018). Recent advances in scalable non-Gaussian geostatistics: The generalized sub-Gaussian model. Journal of Hydrology, 562:685–691.

Guren, M. G., Putnis, C. V., Montes-Hernandez, G., King, H. E., and Renard, F. (2020). Direct imaging of coupled dissolution-precipitation and growth processes on calcite exposed to chromium-rich fluids. Chemical Geology, 552:119770.

Heberling, F., Trainor, T., Lützenkirchen, J., Eng, P., Denecke, M., and Bosbach, D. (2010). Structure and reactivity of the calcite(104) – water interface. Journal of Colloid and Interface Science, 354:848.

Jia, B., Zhou, J., Tang, Z., Xu, Z., Chen, X., and Fang, W. (2022). Effective stochastic streamflow simulation method based on Gaussian mixture model. Journal of Hydrology, 605:127366.

Journel, A. G. (1983). Nonparametric estimation of spatial distributions. Journal of the International Association for Mathematical Geology, 15:445–468.

Li, K., Wu, J., Nan, T., Zeng, X., Yin, L., and Zhang, J. (2022). Analysis of heterogeneity in a sedimentary aquifer using Generalized sub-Gaussian model based on logging resistivity. Stochastic Environmental Research and Risk assessment, 36:767–783.

Liu, H. H. and Molz, F. (1997). Comment on "evidence for non-Gaussian scaling behavior in heterogeneous sedimentary formations" by Scott Painter. Water Resources Research, 33:907–908.

Lu, Z. and Zhang, D. (2002). On stochastic modeling of flow in multimodal heterogeneous formations. Water Resources Research, 38(10):8–1–8–15.

Lüttge, A., Arvidson, R. S., and Fischer, C. (2013). A stochastic treatment of crystal dissolution kinetics. Elements, 9(3):183–188.

Lüttge, A., Arvidson, R. S., Fischer, C., and Kurganskaya, I. (2019). Kinetic concepts for quantitative prediction of fluid-solid interactions. Chemical Geology, 504:216–235.

Marinello, F., Carmignato, S., Voltan, A., Savio, E., and Chiffre, L. D. (2010). Error sources in atomic force microscopy for dimensional measurements: Taxonomy and modeling. Journal of Manufacturing Science and Engineering, 132.

McLachlan, G. J. and Krishnan, T. (2008). The EM algorithm and extensions, Wiley Series in Probability and Statistics.

Meerschaert, M. M., Kozubowski, T. J., and Fred J. Molz, S. L. (2004). Fractional laplace model for hydraulic conductivity. Geophysical Research Letters, 31.

Painter, S. (1996). Evidence for non-Gaussian scaling behavior in heterogeneous sedimentary formations. Water Resources Research, 32:1183–1195.

Riva, M., Neuman, S. P., and Guadagnini, A. (2015). New scaling model for variables and increments with heavy-tailed distributions. Water Resources Research, 51:4623–4634.

Riva, M., Neuman, S. P., Guadagnini, A., and Siena, M. (2013). Anisotropic scaling of berea sandstone log air permeability statistics. Vadose Zone Journal, 12.

Rubin, Y. (1995). Flow and transport in bimodal heterogeneous formations. Water Resources Research, 31(10):2461–2468.

Rubin, Y. and Journel, A. G. (1991). Simulation of non-Gaussian space random functions for modeling transport in groundwater. Water Resources Research, 27:1711–1721.

Ruiz-Agudo, E. and Putnis, C. V. (2012). Direct observations of mineral-fluid reactions using atomic force microscopy: the specific example of calcite. Mineralogical Magazine, 76(1):227–253.

Russo, D. (2002). Stochastic analysis of macrodispersion in gravity- dominated flow through bimodal heterogeneous unsaturated formations. Water Resour. Res., 38:1114.

Siena, M., Bussetti, G., Recalcati, C., Riva, M., Duò, L., and Guadagnini, A. (2021). Statistical characterization of heterogeneous dissolution rates of calcite from in situ and real-time afm imaging. Transport in Porous Media.

Siena, M., Guadagnini, A., Bouissonnié, A., Ackerer, P., Daval, D., and Riva, M. (2020). Generalized sub-Gaussian processes: Theory and application to hydrogeological and geochemical data. Water Resources Research, 56.

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

Teng, H. (2004). Controls by saturation state on etch pit formation during calcite dissolution. Geochimica et Cosmochimica Acta, 68:253–262.

Tidwell, V. C. and Wilson, J. L. (1999). Upscaling experiments conducted on a block of volcanic tuff: results for a bimodal permeability distribution. Water Resources Research, 35:3375–3387.

Tidwell, V. C. and Wilson, J. L. (2002). Visual attributes of a rock and their relationship to permeability: A comparison of digital image and minipermeameter data. Water Resources Research, 38:1261.

Winter, C., Tartakovsky, D., and Guadagnini, A. (2003). Moment differential equations for flow in highly heterogeneous porous media. Surveys in Geophysics, 24:81–106.

Zhang, Y. (2009). Hierarchical geostatistical analysis of an experimental stratigraphy. Math Geosci, 41:145–162.

Zhang, Y., Person, M., Paola, C., Gable, C. W., Wen, X.-H., and Davis, J. M. (2005). Geostatistical analysis of an experimental stratigraphy. Water Resour. Res., 41:W11416.

A Appendix

We employ an AFM (Keysight 5500 apparatus) in contact mode, equipped with silicon tips (Bruker, RESPA-40) and with an Al-coated cantilever (elastic constant $k = 5\,\mathrm{N/m}$) to sample the evolution in time of the topography of the crystal surface. The calcite sample (volume $V \sim 5 \times 3 \times 1$ mm^3) is cleaved from an Iceland spar (Mexico) along the $\{104\}$ plane using a razor blade. The calcite fragment is placed on a glass slide and subsequently fluxed with nitrogen to remove any remains from the cleavage process. The slide is then secured on the AFM support and a Viton O-ring is centered on the sample to seal the cell volume ($\sim 2\,\mathrm{mL}$). A portion of the crystal surface ($6 \times 6\ \mu\mathrm{m}^2$) is imaged with a constant scan frequency (1.41 Hz, scan time: 6 min) along a 512×512 uniform grid, with a horizontal resolution of 11.7 nm. We only consider scans in the top-down direction (Figure 5.e), stopping the acquisition between each imaging frame for $\sim 0.5\,\mathrm{min}$. Therefore, each topography image is obtained within a temporal window of width $\Delta t_a = 6.5\,\mathrm{min}$. Note that in Section 4.1 we evaluate dissolution rate maps across a temporal window $\Delta t = 2\Delta t_a = 13\,\mathrm{min}$. This enables us to detect a significant variation of the area associated with component A, yielding a robust assessment of incremental data and model parameters therein.

The cell is open to air on the top and is filled with the solution (deionized Milli-Q water $18.2\,\mathrm{M\Omega \cdot cm}$). A system of synchronized syringes is connected to the cell. This allows a complete substitution of the fluid in contact with the sample. A volume of 3 mL (~ 1.5 times the volume of the cell) of solution is replaced in the temporal interval between two consecutive frames. This yields (approximately) constant chemical conditions during the whole data acquisition, as documented by the constant spreading rate of the multilayer etch-pit (Fig. 7.d).

Acronyms

AFM Atomic Force Microscopy

PDF probability density function

CDF cumulative distribution function

GMIX Gaussian mixture

VSI Vertical Scanning Interferometry

ML Maximum Likelihood

GSG Generalized Sub-Gaussian

722 # Figures

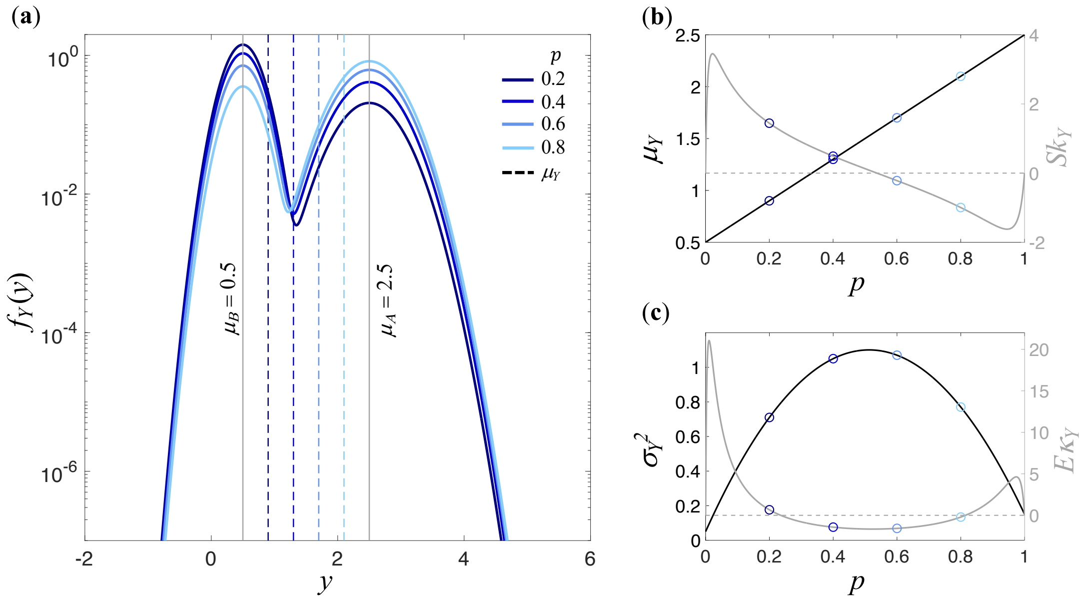


Figure 1: (a) Probability density functions (PDFs), $f_Y(y)$, of the GMIX model evaluated according to Eq. (7) for $\mu_A = 2.5$, $\mu_B = 0.5$, $\sigma_A^2 = 0.15$, $\sigma_B^2 = 0.05$, and four values of p. The associated (b) mean, μ_Y, and skewness, Sk_Y; (c) variance, σ_Y^2, and excess kurtosis, $E\kappa_Y$, are also depicted as a function of p. Empty circles in (b)-(c) correspond to statistical moments associated with the PDFs depicted in (a).

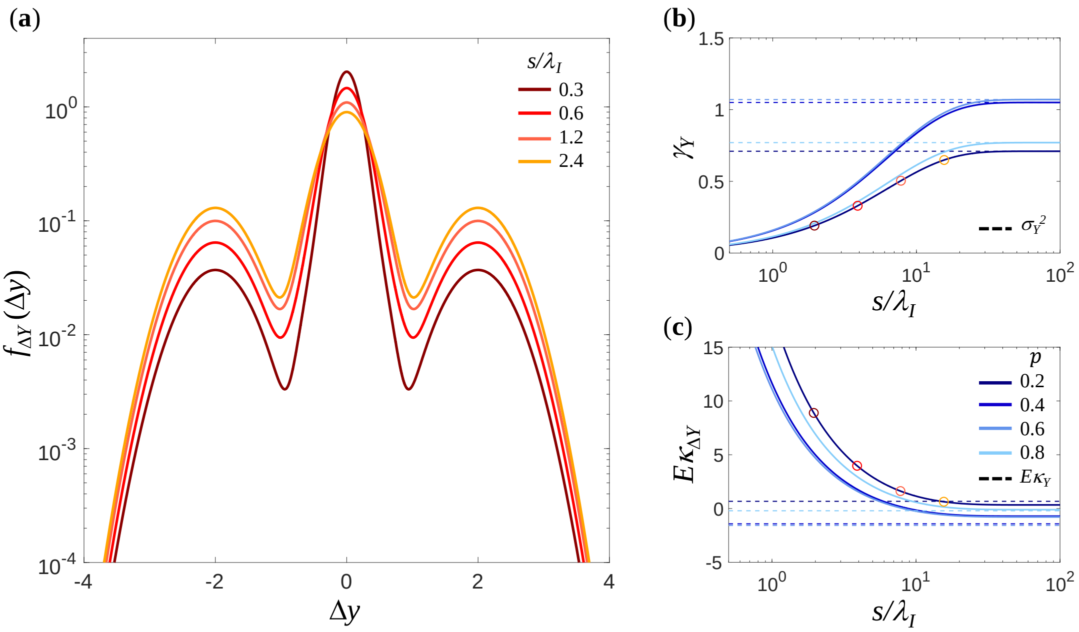


Figure 2: (a) Probability density functions (PDFs) of increments, $f_{\Delta Y}(\Delta y)$, evaluated according to the GMIX model Eq. (21) for $\mu_A = 2.5$, $\mu_B = 0.5$, $\sigma_A^2 = 0.15$, $\sigma_B^2 = 0.05$, $p = 0.2$, $\lambda_A = \lambda_B = 6$ and $\lambda_I = 6.4$, at four values of the dimensionless lags, s/λ_I. The (b) variogram, γ_Y and (c) excess kurtosis, $E\kappa_{\Delta Y}$, are also depicted versus s/λ_I for four values of p. Empty circles in (b)-(c) correspond to the statistical moments associated with the PDFs depicted in (a).

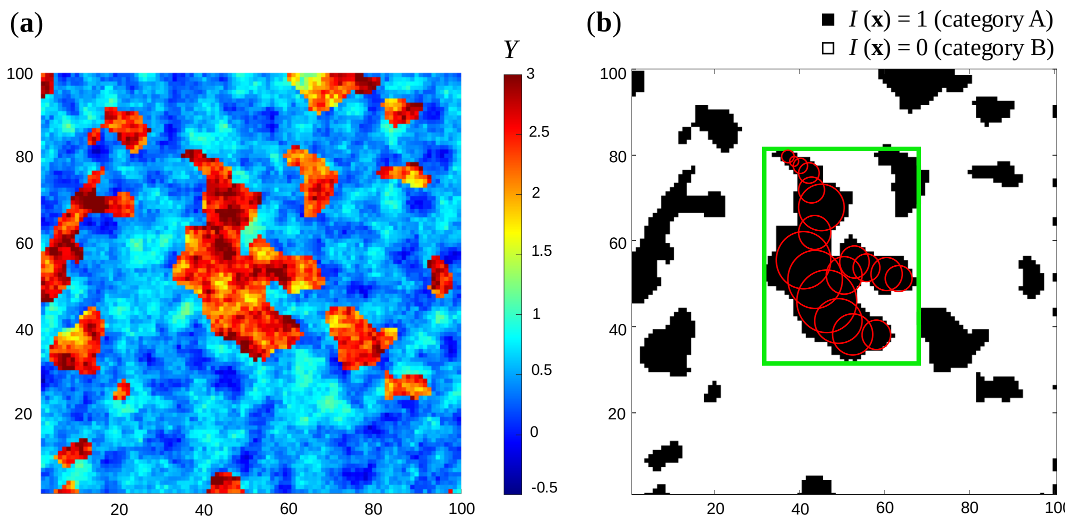


Figure 3: (a) Synthetic realization of a GMIX field with the set of parameters used in Fig. 2; (b) Indicator field associated with the GMIX realization depicted in (a). Green lines and red circles represent the sides of the bounding box and the inscribed maximal balls for a connected cluster, respectively.

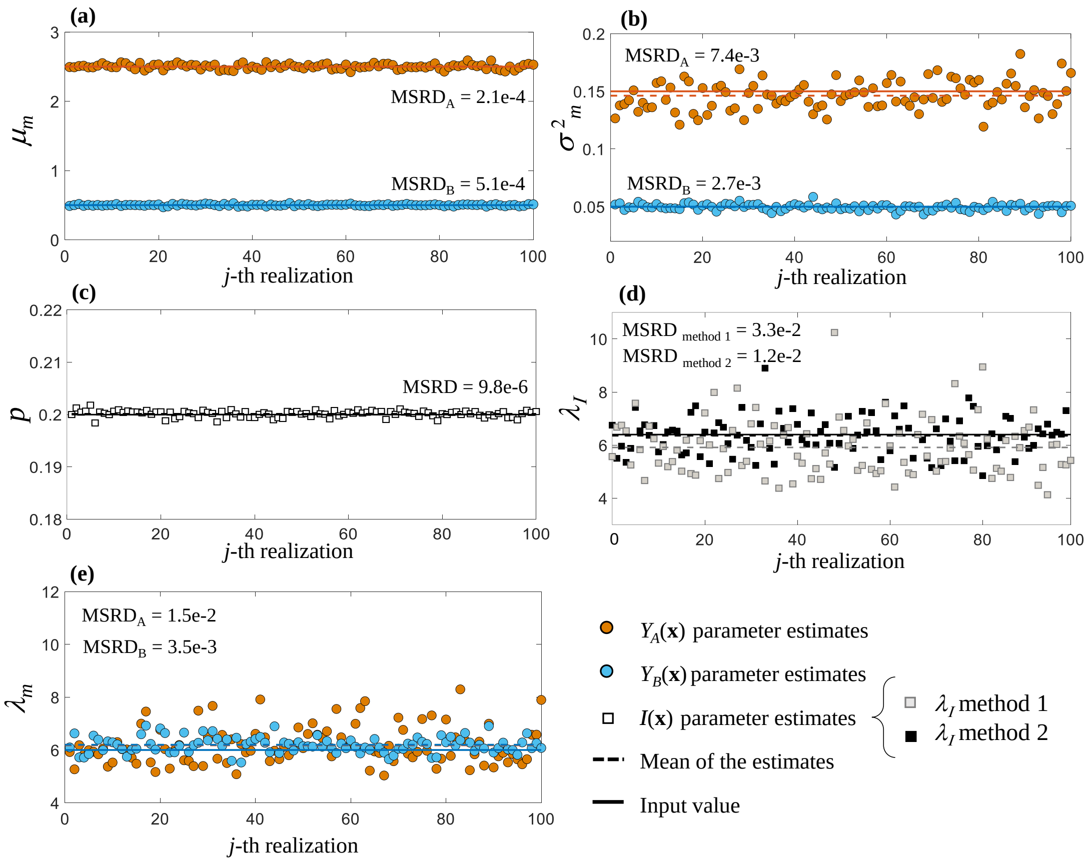


Figure 4: Results of the GMIX parameter estimation approach applied on the collection of synthetic datasets generated with the set of parameters used in Fig. 2. Mean of the estimates (dashed lines), input values used in the generation (solid lines) and mean squared relative deviation (MSRD) between input and estimated parameter values are also reported.

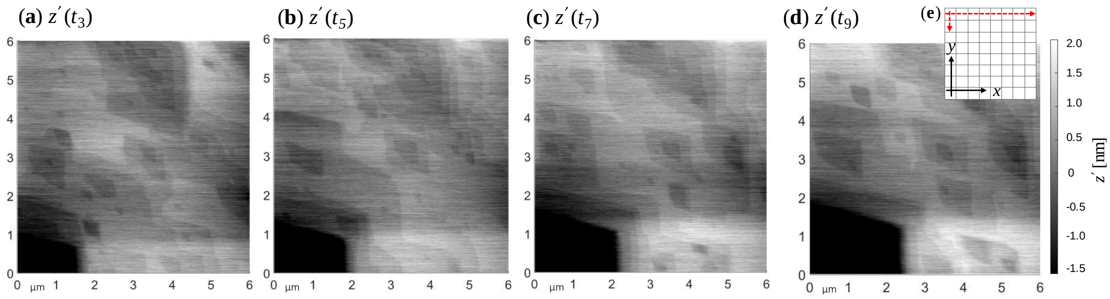


Figure 5: Images of fluctuation of calcite topography about its mean, $z'(t_i)$, acquired via AFM at (a) $t_3 = 26$ min, (b) $t_5 = 39$ min, (c) $t_7 = 52$ min and (d) $t_9 = 65$ min from the beginning of the experiment. Scanning directions are depicted in (e).

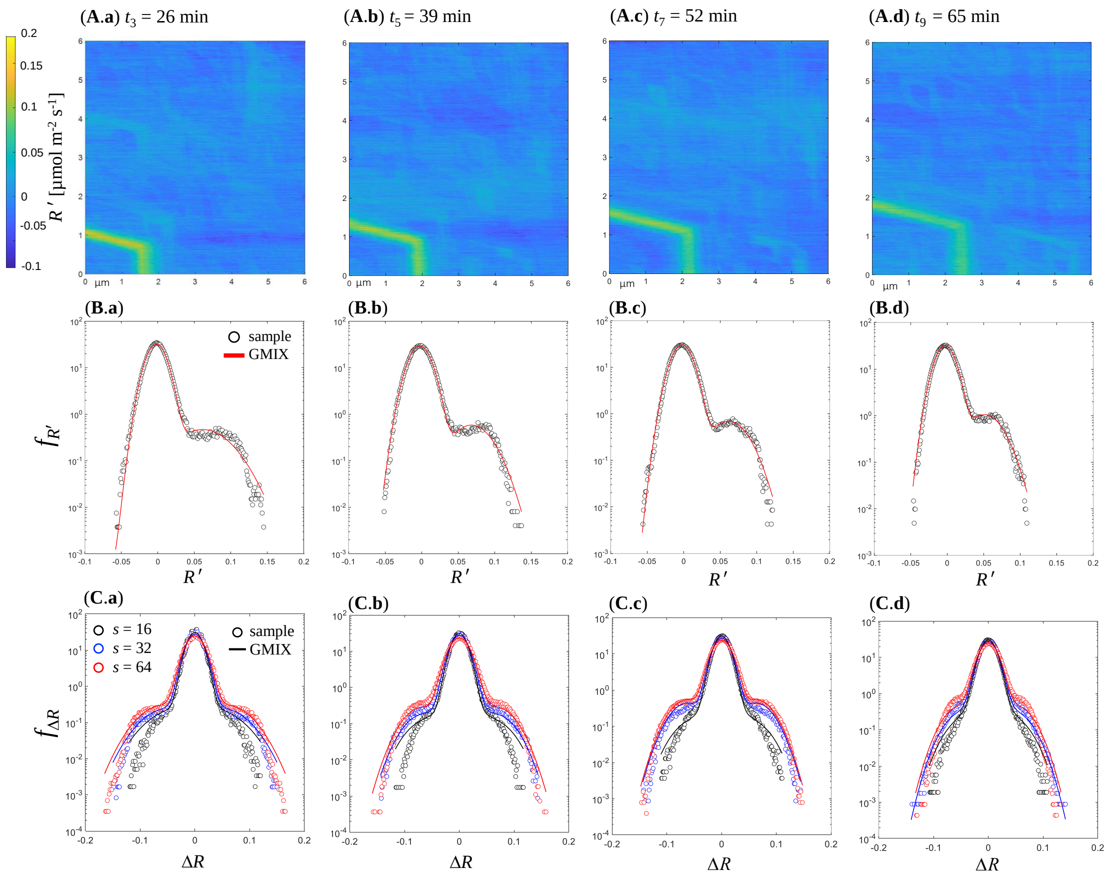


Figure 6: Dataset and statistical results associated with observed dssolution rate maps, $R'(t_i)$. Colormaps of $R'(t_i)$ (A) evaluated with Eq. (26) from AFM topography measurements depicted in Fig. 5 are shown for the considered times (a-d). Sample PDFs of $R'(t_i)$ (B) and $\Delta R(t_i)$ (C) are also depicted. Analytical results for the PDFs of reaction rate (Eq. (7)) and its spatial increments (Eq. (21)) are juxtaposed to the experimental data.

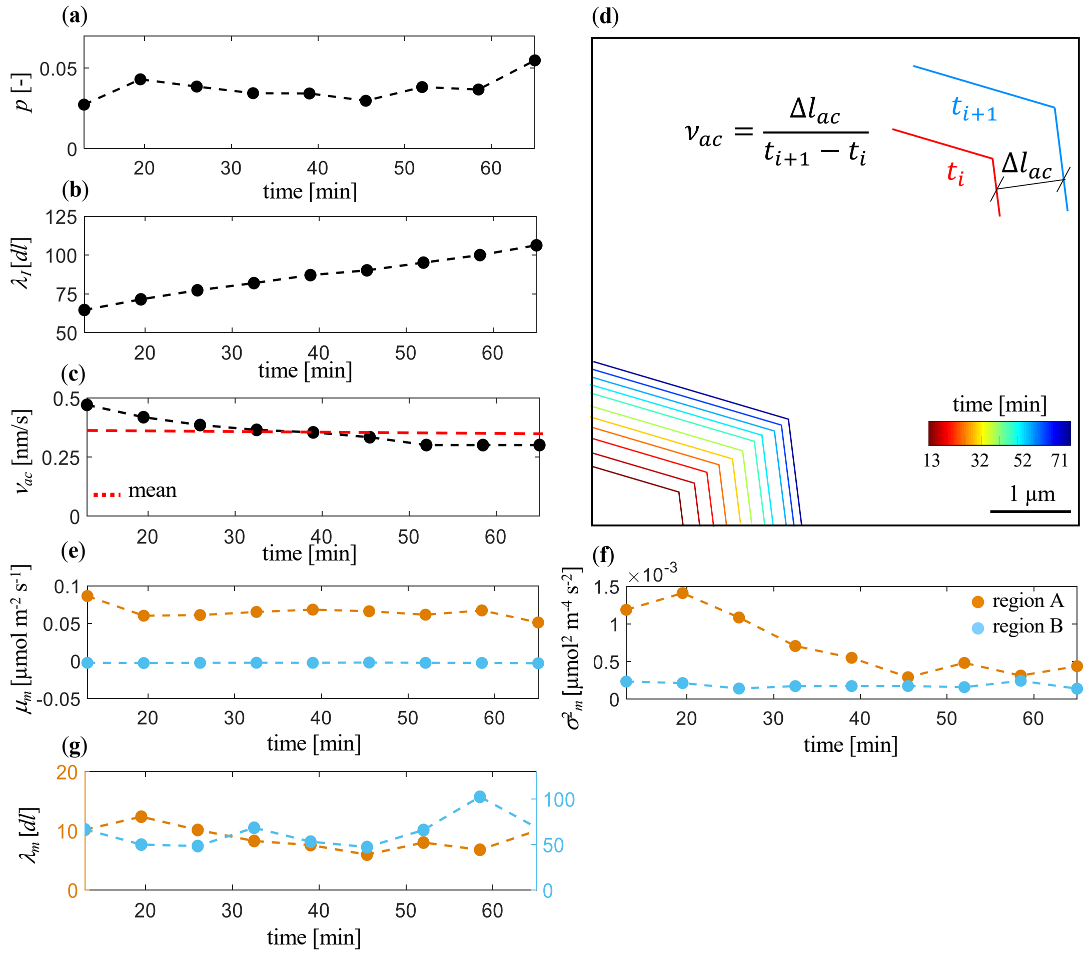


Figure 7: Temporal trend of parameters of the GMIX model. The evolution of the parameters associated with the indicator random field (p and λ_I) is depicted in (a) and (b), respectively. The behavior of mean, μ_m (e), variance, σ_m^2 (f), and correlation length, λ_m (g) for component $m = (A, B)$ is illustrated. Panel (c) illustrates the horizontal spreading rate of acute steps, ν_{ac}, associated with the etch-pit edges shown in (d) at various times.

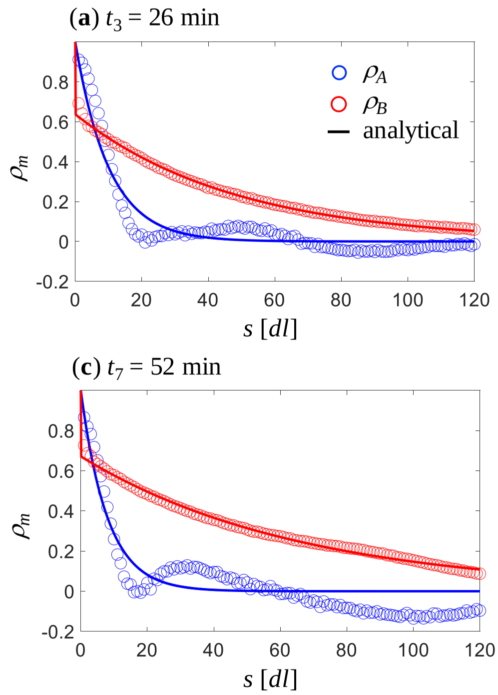


Figure 8: Spatial correlation of components A and B associated with the rate maps (a-d) depicted in Fig. 6.A as a function of separation distance, s. The analytical interpretive model, i.e., exponential model with nugget, is also depicted.

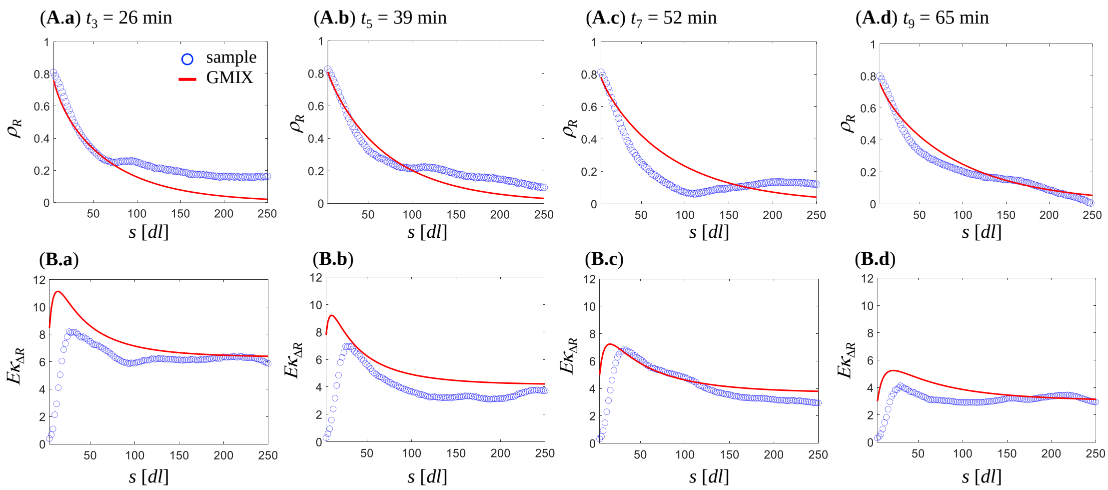


Figure 9: Statistical moments associated with the spatial increments of R'. Analytical expressions resulting from the GMIX formulation are juxtaposed to the sample correlation function, ρ_R, (A) and excess kurtosis, $E\kappa_{\Delta R}$ (B) associated with the rate maps (a-d) shown in Fig. 6.A.

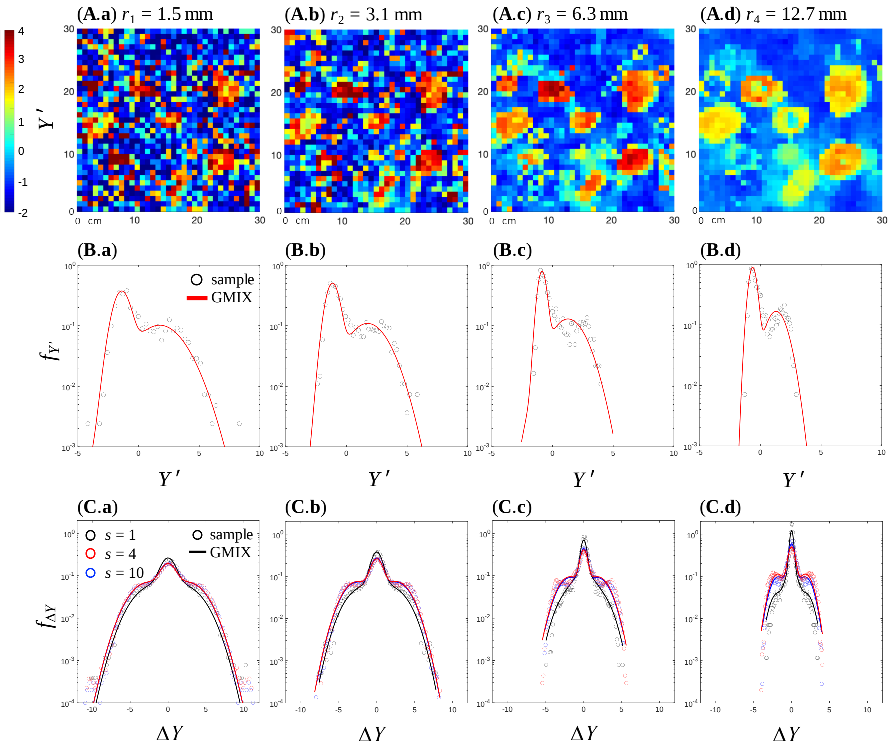


Figure 10: Data and results of statistical analyses of (log) air-permeability, $\ln k$, acquired by Tidwell and Wilson (1999) across Face 1 of a Topopah Spring Tuff block. Colormaps of $Y' = \ln k - \langle \ln k \rangle$ (A) for the different values of the inner tip-seal radius of the minipermeameters employed in the experiments, i.e., (a) $r_1 = 1.5$ mm, (b) $r_2 = 3.1$ mm, (c) $r_3 = 6.3$ mm, (d) $r_4 = 12.7$ mm. Results of the GMIX analytical model are juxtaposed to the sample statistics of Y' data (B) and of their increments, ΔY (C).

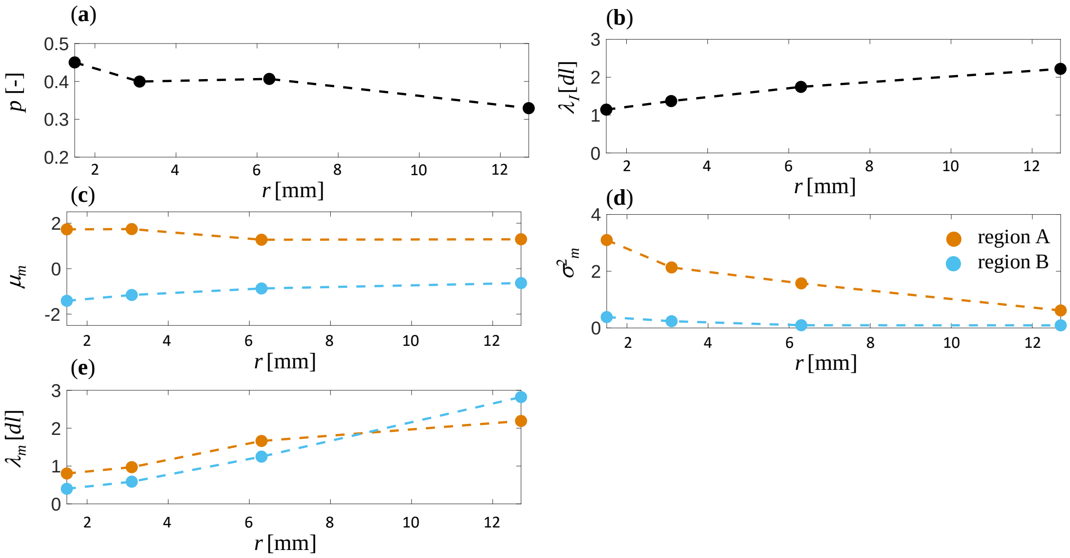


Figure 11: Parameters of the GMIX model versus tip-seal inner radius, r. Panels (a) and (b) depict parameters associated with the indicator random field, i.e., p and λ_I, while (c), (d) and (e) depict the trend of mean, μ_m, variance, σ_m^2, and correlation length, λ_m $(m = A, B)$, of each of the two components.

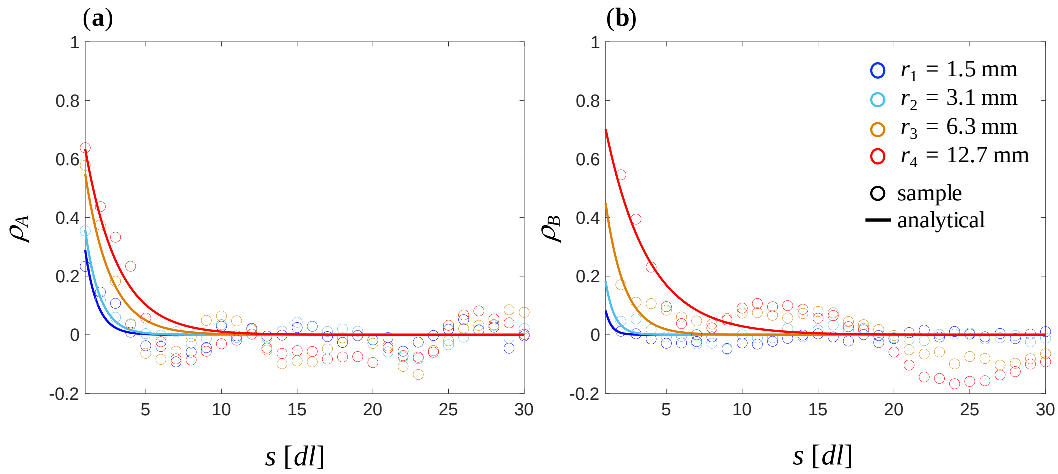


Figure 12: Spatial correlation of components A and B as a function of the separation distance, s, for the considered tip-seal inner radii. Results based on the interpretive model, i.e. single parameter exponential model, are also depicted.

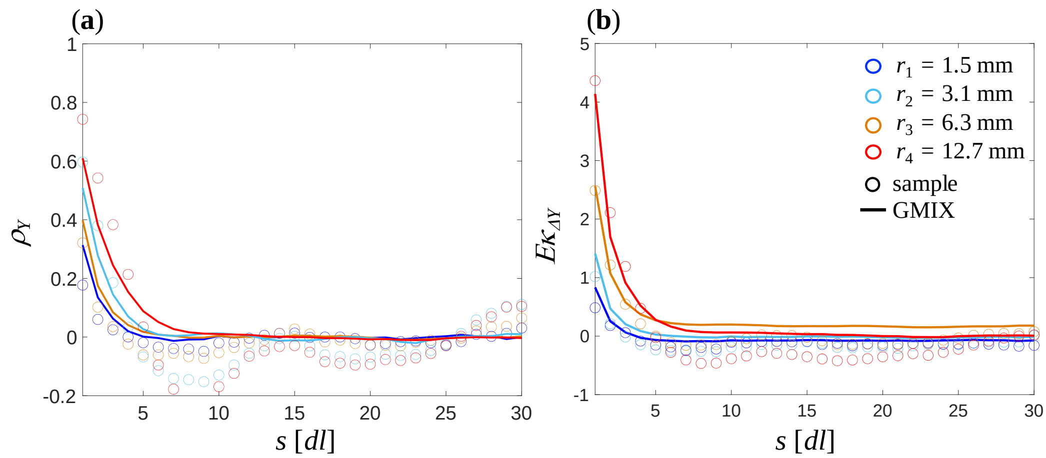


Figure 13: Statistical moments associated with the spatial increments of the air permeability data. Theoretical and sample correlation function, ρ_Y, (a) and excess kurtosis, $E\kappa_{\Delta Y}$, (b) are depicted for the considered tip-seal inner radii, r_i

- A unique framework for Gaussian Mixtures (GMIX) and their increments is derived
- Probability distributions of GMIX increments scale with separation distance
- A GMIX parameter estimation method is developed and tested
- The GMIX model captures the main traits exhibited by hydrogeochemical datasets
- Temporal trends of GMIX parameters reflect dynamics of mineral dissolution patterns

Credit Author Statement

The contribution, gender and position of each Author is detailed in the following.
Dr. Martina Siena (Female). Assistant Professor. Tutoring activity for the laboratory experiments. Contribution: research design (10%), data collection (45%), data analysis (30%), modeling (35%), writing (25%), interpretation (20%), editing (20%).
Chiara Recalcati (Female). PhD Student. data collection (55%), data analysis (30%), modeling (25%), writing (25%), interpretation (20%), and editing (20%)
Dr. Alberto Guadagnini (Male). Professor. Co-Supervisor of Chiara Recalcati: research design (45%), data analysis (20%), modeling (20%), writing (25%), interpretation (30%), editing (30%).
Dr. Monica Riva (Female). Professor. Supervisor of Chiara Recalcati. research design (45%), data analysis (20%), modeling (20%), writing (25%), interpretation (30%), editing (30%).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: