Noname manuscript No.
(will be inserted by the editor)

Characterization of Flow through random media via Karhunen-Loève expansion: an Information Theory perspective

Aronne Dell'Oca · Giovanni M. Porta

Received: date / Accepted: date

Abstract We leverage on Information Theory to assess the fidelity of approximated numerical stochastic groundwater flow simulations. We consider flow in saturated heterogeneous porous media, where the Karhunen-Loève (KL) expansion is used to express the hydraulic conductivity as a spatially correlated random field. We quantify the impact of the KL expansion truncation on the uncertainty associated with punctual values of hydraulic conductivity and flow velocity. In particular, we compare the statistical dependence between variables by considering (a) linear correlation metrics (Pearson coefficient of correlation) and (b) metrics capable of accounting for nonlinear dependence (coefficient of uncertainty based on mutual information). We test the selected metrics by analyzing the relationship between hydraulic conductivity fields generated via Monte Carlo sampling with different levels of truncation of the KL expansion and the corresponding fluid velocity fields, obtained through the numerical solution of Darcy's flow. Our analysis shows that employing linear correlation metrics leads to a general overestimation of the correlation level and IT-based indicators are valuable tools to assess the impact of the KL truncation on the output velocity values. We then analyze the impact of the number of retained modes on the spatial organization of the velocity field. Results indicates that (i) as the number of modes decrease the spatial correlations of the velocity field increases; (ii) linear indicators of spatial correlation are again larger than the nonlinear counterparts.

A. Dell'Oca
Dipartimento di Ingegneria Civile ed Ambientale, Piazza Leonardo da Vinci 32, IT-20133 Milano, Italy
E-mail: aronne.delloca@polimi.it

G.M. Porta
Dipartimento di Ingegneria Civile ed Ambientale, Piazza Leonardo da Vinci 32, IT-20133 Milano, Italy
E-mail: giovanni.porta@polimi.it

Keywords Fluid Flow · Heterogeneous Media · Monte Carlo Simulation · Information Theory

1 Introduction

In the context of subsurface geology it is well known that the hydraulic properties of natural aquifers exhibit heterogeneity over diverse spatial length scales [7, 25]. This ubiquitous heterogeneity and the typical poor level of characterization of subsurface environments lead to a lack of knowledge about the hydraulic properties of the hosting formations, which is a major source of uncertainty. This motivates the interest towards applications of uncertainty quantification and propagation through numerical modeling of subsurface flow and transport processes.

In this work we consider the uncertainty stemming from incomplete knowledge of the hydraulic conductivity of fully saturated heterogeneous porous formations on the resulting steady state flow field. To address its parameterization under uncertainty, permeability (or hydraulic conductivity) is often described by random fields, where the heterogeneity structure can be characterized with geostatistical methodologies [7]. These methods rely on an assumed statistical distribution and spatial covariance function for the parameter under investigation. A classical assumption is to assign a lognormal distribution to hydraulic conductivity, although different models have been proposed in the literature [25]. When the log-conductivity is expressed by a multi-Gaussian random field, the well-known Karhunen-Loève (KL) [21] expansion can be used to reduce the stochastic dimension of the problem. This allows reducing the computational costs for uncertainty propagation, which can be achieved also through the implementation of model reduction techniques, e.g. [22, 20, 19]. KL-based approximations rely on a truncation of the exact expansion, which, in practice, is equivalent to selecting a number of terms in the series, also called modes. The number of modes retained in the KL expansion is typically determined by ensuring that a given fraction of the variance (considered as a proxy of the spatial variability) of the underlying field is retained. However, this approach does not ensure that a satisfactory level of variability of the output is also retained (see e.g., [33]) particularly in the presence of nonlinear relations between inputs and quantities of interest (QoI). Quantitative indicators are therefore required to assess the foreseen effect of such approximations on QoI. In this work our target is to characterize the propagation of uncertainty to fluid velocity in fully saturated heterogeneous porous media. The simulation of fluid velocity is fundamental to numerical studies aimed at assessing the impact of heterogeneity in fluid flow and solute transport processes taking place in heterogeneous aquifers and reservoirs. For instance, the velocity field and its spatial structure are key information to simulate and predict transport and dilution of solute mass in the subsurface [13, 32, 8]. To address these problems, general indicators that can connect parameters describing the spatial hetero-

geneity of geomaterials properties to flow and transport emerging features are sought in recent literature [2].

Developed after the work of Shannon [30], Information theory (IT) provides a suite of indicators that can be used to assess the similarity in information content of two distributions. Arguably the most widespread quantitative indicator derived from IT is Shannon entropy. In addition, IT provides quantitative indicators that can assess the mutual information between variables. These indicators can also be used to describe correlation within a spatial and temporal fields. For example, when nonlinear systems are of concern, the IT-based mutual information can be used as an alternative to linear correlation, as shown, e.g., by [26] in the context of the characterization of subsurface mineral distributions. IT-based indicators have been previously employed within model-based assessment of groundwater flow and solute transport. For example several studies have relied on the concept of entropy as an indicator of uncertainty within risk assessment procedures [23, 1] or to set up optimal experimental design for model discrimination [17]. IT mutual information has also been shown to be an indicator of the degree of nonlinearity existing between output variables in flow and transport simulations, with a particular focus on their spatial correlation [4]. As an alternative approach, the concept of entrogram was introduced in [3]. This latter corresponds to the fraction of the entropy of a variable sampled within a given spatial window and that of its counterpart associated with the whole spatial domain, as a function of the ratio between the spatial window and the whole domain sizes. The results in [3] indicate that the entropic scale (i.e., a measure of the overall persistence of a pattern of association) can be related to widely employed solute transport descriptors.

In this work, we apply IT-based metrics within stochastic groundwater flow simulations. We consider stochastic simulations based on Monte Carlo sampling of the KL expansion of hydraulic conductivity. The selection of the number of modes in the KL-expansion is typically guided by selecting a proportion of the variance of the original hydraulic conductivity field to be retained. Such proportion can be analytically determined and increases with the number of retained modes. However, for values of the variance of log-conductivity above unity the degree of nonlinearity between input (hydraulic conductivity) and output (flow velocity) increases [4]. A critical issue in this context is to control and constrain a priori the output approximation accuracy. Our work explores the possibility to employ IT-based metrics in the context of a quantitative assessment of the approximation resulting from KL series truncation. While IT has been employed with a similar objective in other fields [16], to the best of our knowledge this idea is explored here for the first time in the context of stochastic groundwater flow simulation. In particular, for flow fields associated with an increasing number retained modes in the KL expansion we (i) compare the resulting entropy of the flow field, to investigate the level of variability retained in the predicted quantity of interest; (ii) evaluate the mutual information (MI) between flow fields obtained at different levels of res-

olution; (*iii*) compare the behavior of the spatial auto MI of the flow fields, to investigate how the spatial organization of the diverse flow fields behaves.

In the following, section 2 presents the considered problem setup, the IT indicators employed in the analysis and the considered quantities of interest. Section (3) present the results obtained by the application to groundwater flow with three different levels of heterogeneity, i.e. three selected variances of log-conductivity.

2 Methods

2.1 Problem setup

We consider steady flow in a two dimensional heterogeneous saturated porous medium, through Darcy's law and fluid mass conservation

$$\nabla \cdot \mathbf{u} = 0 \qquad \mathbf{u} = -K(\mathbf{x})\nabla h \qquad \mathbf{x} \in \Omega \tag{1}$$

where $\mathbf{u}[\mathrm{LT}^{-1}]$ is the Darcy velocity, $K[\mathrm{LT}^{-1}]$ is hydraulic conductivity, h is hydraulic head, Ω is a two-dimensional spatial domain. Note that, even though our results are limited to a two-dimensional set-up, they can be extended to three dimensional systems following the same line of reasoning. In the latter case, we expect a different quantitative behavior (the dimensionality of the problem is a crucial factor in the flow organization), but we expect a similar qualitative behavior. The Darcy velocity has two components, i.e., v and u which are transversal and longitudinal with respect to to the main flow direction, respectively. We consider in this study flow taking place in a squared domain of unit size. We impose impermeable boundaries along the right and left edges, while we set a uniform value of v and of h along the top and bottom boundary, respectively. Equation (1) is numerically solved upon employing a mixed two-field finite element approach (see e.g., [9]) implemented within the FreeFem++ environment [14]. We employ a structured triangular spatial discretization, considering 313600 triangles in order to ensure the accurate resolution of the hydraulic conductivity spatial distribution (i.e., one correlation length of the conductivity field, see below, has been discretized by 16 elements [29]) and, in turn, of the ensuing velocity fields.

We assume here the log-conductivity $Y = \ln(K)$ as a spatially correlated Gaussian random field. Under such assumption, a reference field $\bar{Y}$ can be defined through the Karhunen-Loève expansion [21]

$$\bar{Y} = \sum_{n=1}^{\infty} \xi_n \lambda_n f_n(\mathbf{x}) \tag{2}$$

where ξ_n are orthogonal Gaussian random variables with zero mean, λ_n and $f_n(\mathbf{x})$ are the eigenvalues and eigenfunctions that can be used to approximate the Y field with spatial covariance $C_Y(\mathbf{x}, \mathbf{y})$. In the following we assume an exponential separable covariance function $C_{Y^{ref}}(\mathbf{x}, \mathbf{y}) = \sigma^2_{Y^{ref}} \exp(-|\mathbf{x_1} -$

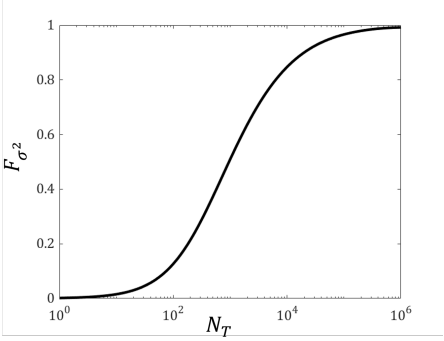

Fig. 1 Number of terms N_T to be included in (3) as a function of the target retained variance in the truncated field F_{σ^2}.

$\mathbf{y_1}|/\eta - |\mathbf{x_2} - \mathbf{y_2}|/\eta)$, being $\sigma^2_{Y^{ref}}$ the variance of the conductivity field, η [L] the correlation length and with subscript 1 and 2 indicating the longitudinal and transverse spatial directions, respectively. The correlation length η is chosen as 1/35 of the domain size. The analytical expressions provided in [34] are then used to compute the factors λ_n and the functions $f_n(\mathbf{x})$ in the expansion (2). The series (2) is employed in practice upon truncation to a finite number of terms, i.e.,

$$Y = \sum_{n=1}^{N_T} \xi_n \lambda_n f_n(\mathbf{x}) \approx \bar{Y} \tag{3}$$

The number of terms N_T included in the expansion can be selected upon requiring that the generated fields retain a given fraction F_{σ^2} of the reference field variance, i.e. $\sigma^2_{\bar{Y}} = F_{\sigma^2}\sigma^2_Y$, with $F_{\sigma^2} < 1$. In this context F_{σ^2} is often considered as an *a priori* indicator of the fidelity of the truncated expression in reproducing the original field. Figure 1 depicts F_{σ^2} versus N_T. Note that the value of F_{σ^2} increases monotonically with N_T, i.e. the number of terms increases with the required level of retained variance. Note that the Karhunen-Love expansion is valid for all variances of Y and can applied considering any dimensionality of the system.

In this work we analyze Monte Carlo samples of log-conductivity fields characterized by three distinct reference variances, i.e. $\sigma^2_{Y^{ref}} = 0.5, 1.5, 2.5$. For each of these three cases we consider a reference sample where the fields are generated upon truncating the expansion when the variance of the corresponding field attains a value of 0.97 of $\sigma^2_{Y^{ref}}$. Then, for each of the considered $\sigma^2_{Y^{ref}}$ we generate MC samples with F_{σ^2} comprised between 0.5 and 0.97 upon selecting a number of 1000 realizations per sample. In the following the results associated with $F_{\sigma^2} = 0.97$ are labeled as reference fields Y^{ref},u^{ref},v^{ref}. Variables obtained for $F_{\sigma^2} < 0.97$ are considered as truncated fields and indicated with Y^{tr},u^{tr},v^{tr}. Note that, even though low level of F_{σ^2} (e.g., smaller than 0.8) may be too restrictive in the context of practical applications, we here cover them in order to highlight the emergence of trends/patterns in the our results

2.2 Information Theory

Uncertainty in a discrete random variable, X, can be quantified via Shannon Entropy [30]

$$H(X) = \sum_{i=1}^{N_X} p_i \ln(p_i^{-1}) \tag{4}$$

where N_X is the number of bins used to discretize the sample probability distribution, and p_i is the probability associated with the i-th bin. Shannon Entropy can be interpreted as a measure of the uncertainty associated with X, i.e., $H(X)$ is maximum (assigned the binning) in case p_i is uniform across all the N_X bins while it equals zero when all the values of X fall into one single bin. Note that employing the natural logarithm leads to have *nats* as unit of measure for *H(X)*, other choices for the base of the logarithm are possible.

Statistical dependence between two random variables, i.e., X and Z, can be characterized by the reduction in the uncertainty that the knowledge of one variable entails for the other. This is formalized by the mutual information as

$$I(X;Z) = \sum_{i=1}^{N_X} \sum_{j=1}^{N_Z} p_{i,j} \ln(p_{i,j}/(p_i p_j)) \tag{5}$$

where N_Z is the number of bins associated with Z, p_j is the probability distribution for Z and $p_{i,j}$ is the joint probability distribution between X and Z. Mutual Information turns out to be null in case of independent random variables, while the equality $H(X) = H(Z) = I(X;Z)$ holds in case that the knowledge of one variable is sufficient to predicted the other one exactly. Mutual Information is again measured in *nats* as in (4). It is important to recall here that mutual information is a nonlinear dependence metric, i.e., it is capable of detecting dependence between random variables which are not induced by a linear relationship. On top of mutual information it is then convenient to define the dimensionless uncertainty coefficient (UC) [31]

$$UC(X;Z) = \frac{2I(X;Z)}{H(X) + H(Z)} \tag{6}$$

where UC is bounded between zero, for independent variables, and one, when there is an exact dependence between the two variables. In the following the UC will be compared against the well known linear correlation coefficient (or Pearson coefficient), i.e., $\rho(X;Z)$, which captures only the intensity of the linear dependence between two variables. Note that in principle $\rho(X;Z)$ can be both larger or smaller than $UC(X;Z)$, for a given pair of variables $(X;Z)$. The IT metrics employed in our work can be defined for continuous probability distributions where summations and probability mass functions are replaced by integrals and probability density functions, respectively. This approach would be characterized by a less intuitive and immediate interpretation, for example Entropy could be negative, infinite or impossible to evaluate, see, e.g.[6, 15]. Furthermore, the pdfs of the fluid velocity components are not associated with

a known analytical formulation but are here obtained as the result of numerical simuations. Employing a continuous approach would require subjecting these variables to quantization [6]. In general, the quality of IT metrics estimates increases (in a way which depends on the specific metric) with the level of quantization of the continuous variables [15]. These consideration lead us to treat the analyzed variables as discrete ones in line with several previous studies [28,12,24].

2.3 Quantities of interest

Our analysis considers three quantities of interest, i.e., (i) the log-conductivity Y and (ii) the two components of the velocity field v (longitudinal to the pressure gradient) and u (transverse). First, we focus on the analysis of punctual values of these variables, upon considering their distributions associated with diverse levels of truncation in the KL-expansion, as quantified by F_{σ^2}. This is obtained through two different types of analyses:

- we investigate the relationship between(i) the variance of the values of Y and that of v and u, as well as, between (ii) the Shannon Entropy, for the same variables, considering the diverse levels of F_{σ^2}.
- we analyse the relationship between the Linear Correlation Coefficient for (i) Y^{ref} and Y^{tr} and (ii) v^{ref}, u^{ref} and v^{tr}, u^{tr}, as a function of F_{σ^2}. We repeat the analysis by focusing on the corresponding uncertainty coefficients..

These statistics are evaluated at each location of the domain, we then exclude from the analysis locations which have a distance from the boundary smaller than eight correlation lengths, to avoid the influence of domain boundaries. Furthermore, in the following we report the spatial averages of the statistics of interest.

The Information Theory metrics detailed in Section 2.2 refer to discrete random variables. Hereafter, we treat Y and v,u as discrete quantities in (4)-(6), i.e., we consider their probability mass distributions resulting from empirical frequencies. For each of the three considered $\sigma^2_{Y^{ref}}$, we employed a regular binning, made of 15 bins, ranging from the minimum to maximum values of the targeted quantities Y^{ref}, v^{ref}, u^{ref}, obtained for the reference field (corresponding to $F_{\sigma^2} = 0.97$). The same bins are then employed to discretize the probability distributions associated with variables obtained for the same $\sigma^2_{Y^{ref}}$ but smaller F_{σ^2}. Note that the number of bins employed in the discretization of the inspected variables can be thought as the achievable characterization accuracy (e.g., as dictated by hypothetical instrumental apparatuses). For example, our choice of binning of the log-conductivity field could be thought as be equivalent to consider a system with 15 distinct hydrofacies, each identified by a bin.

3 Results

Our analysis is subdivided into two main parts. In the first part of our analysis, we aim at quantifying variations in the uncertainty in the punctual values as a function of the number of retained modes in the KL-expansion (2) and at assessing the degree of correlation and mutual correlation between corresponding values associated with diverse levels of truncation. Then, we assess the impact of the retained number of modes on the spatial organization of the two velocity components. To this end, we evaluate the uncertainty coefficient and the linear correlation coefficient between pairs of v and u separated by increasing spatial lags (along the mean flow direction), for a selected F_{σ^2}.

3.1 Analysis of punctual values of velocity and log-conductivity

We analyze here the relationship between the uncertainty of the input field, Y, and of the two velocity components, u and v. We firstly focus on the quantification of uncertainty as rendered by the variance (see e.g., [11]) and by Shannon entropy. Figure 2 displays σ^2_Y against σ^2_v and σ^2_u, for diverse values of F_{σ^2} (indicated by diverse markers) and variance of the reference field $\sigma^2_{Y^{ref}}$ (points indicated by diverse colors). As expected σ^2_v and σ^2_u are both directly proportional to σ^2_Y [11]. In particular, we find that the results associated with a single value of $\sigma^2_{Y^{ref}}$ follow a specific trend, where σ^2_v increases nonlinearly with σ^2_Y while σ^2_u display a trend close to linear (compare trends identified by symbols displayed in the same color in Figure 2a and b). This result is consistent with the idea that variations in the input log-conductivity are more deeply reflected by the longitudinal velocity rather than the transverse one. For both the velocity components the dependence of σ^2_v and σ^2_u on σ^2_Y is specific to the selected value of $\sigma^2_{Y^{ref}}$.

To demonstrate the possible limitations in using the variance to characterize heterogeneity of velocity distributions (see also [5,18]), Figure 3 reports the frequencies of the reference v for two values of $\sigma^2_{Y^{ref}}$ (note that values are normalized by the average longitudinal velocity). As the level of heterogeneity in Y increases the shape the pdf of velocity tends to display a increasing skewness towards high velocity values. For illustrative purposes, and to emphasize this transition, we compare the probability mass functions of v with that of Gaussian variables with equal mean and variance values. The change in the probability density shape is a result of the nonlinear relationship between log-conductivity and velocity, emerging for increasing $\sigma^2_{Y^{ref}}$, that has been reported in previous literature (e.g., [4]). The results shown in Figures 2-3 therefore suggest that the variance may not be an optimal indicator to characterize the velocity probability distributions for increasing heterogeneity.

For this reason, we resort to the Shannon entropy of v and u, which are shown as a function of σ^2_Y in Figure 4. This analysis reveals a lack of generality in the relationship between $H(v)$ and σ^2_Y (as well as, in those between $H(u)$ and σ^2_Y) which are strongly sensitive to the value of $\sigma^2_{Y^{ref}}$. In particular, the

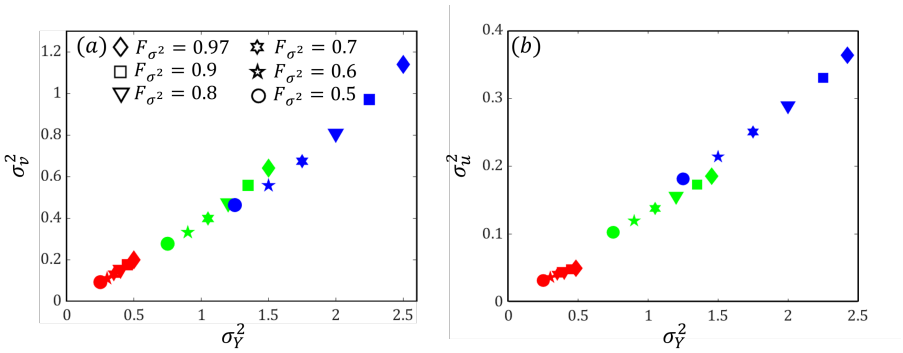

Fig. 2 Relationship between the variance of the log-conductivity, σ_Y^2, and of the longitudinal (a) σ_v^2 and (b) transverse velocity components σ_v^2, for $\sigma_{Yref}^2 = (0.5$ (red), 1.5 (green),2.5 (blue) and diverse values of F_{σ^2}, indicated by different markers.

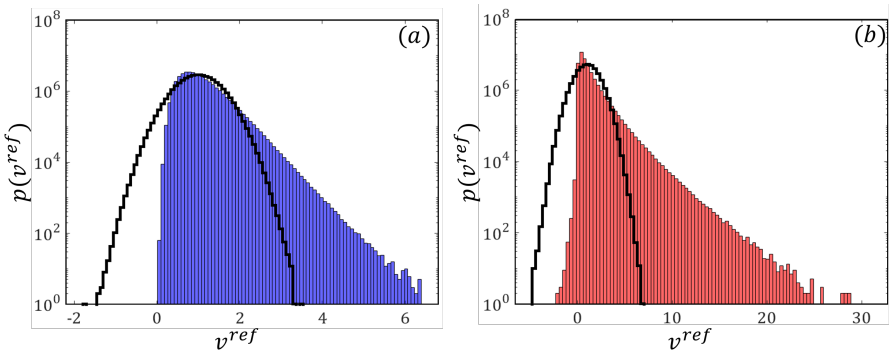

Fig. 3 Probability function for the longitudinal velocity component v^{ref}, i.e., v, for the reference field given (a) $\sigma_Y^2 = 0.5$ and (b) $\sigma_Y^2 = 2.5$. The probability function associated with a Gaussian variable with the same mean and variance is depicted (black curves) for each case.

relative change in $H(u)$ and $H(v)$ between the same levels of truncation F_{σ^2} largely differs as a function of σ_{Yref}^2. This result suggests that same relative change in the input variance is reflected in a different way on the velocity components pdfs, depending on the heterogeneity of the system.

To overcome this issue, we consider Shannon entropy H to describe the relation between Y and v, u. Figure 5a displays the relationship between the Shannon Entropy of the log-conductivity, $H(Y)$, and of the longitudinal velocity,$H(v)$, as a function of σ_{Yref}^2 and F_{σ^2}. Assigned a given value of σ_{Yref}^2, we observe that $H(v)$ increases linearly with $H(Y)$. This result shows that the uncertainty associated with v increases linearly with F_{σ^2}, or equivalently that $H(v)$ approaches its reference value as a linear function of $H(Y)$. However, Figure 5 also indicates that the slope of the $H(Y)$-$H(v)$ linear relationship is specific to the investigated σ_{Yref}^2 values. In particular, we observe that $H(v)$ is inversely proportional to σ_{Yref}^2 (assigned a value of F_{σ^2}). This result shows that the probability mass functions of v are more uniformly distributed across

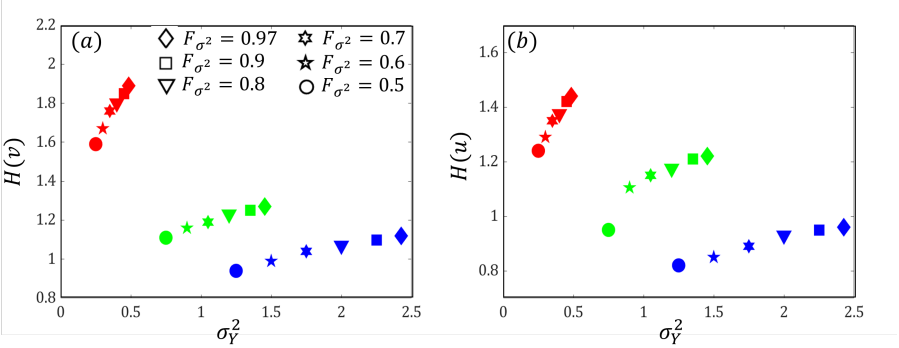

Fig. 4 Relationship between the variance of the log-conductivity, σ_Y^2, and the Shannon Entropy of (a) longitudinal $H(v)$ and (b) transverse $H(u)$ velocity components, for σ_{Yref}^2 = 0.5 (red), 1.5 (green),2.5 (blue) and diverse values of F_{σ^2}, indicated by different markers.

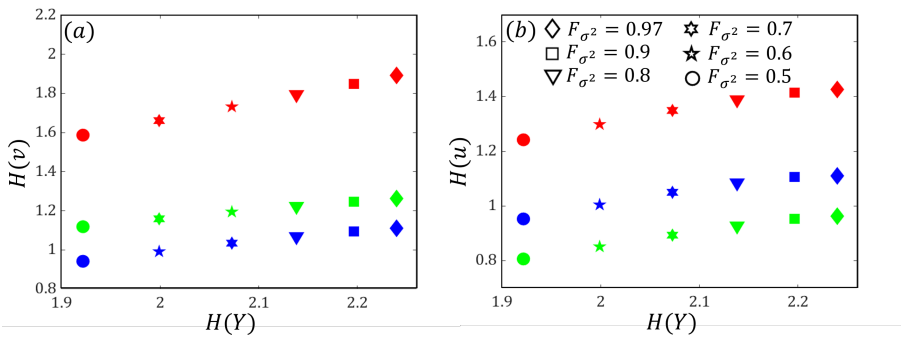

Fig. 5 Relationship between the Shannon Entropy of the log-conductivity, $H(Y)$, and of the (a) longitudinal $H(v)$ and (b) transverse $H(u)$ velocity components, for σ_{Yref}^2 = 0.5 (red), 1.5 (green),2.5 (blue) and diverse values of F_{σ^2}, indicated by different markers.

their supports as σ_{Yref}^2 decreases, thus rendering increasing $H(v)$ values. Note that this result is related to our choice of employing a constant number of bins for all σ_{Yref}^2, which implies that the bin size is highly variable across the different samples. In other words, this reflects the higher degree of spatial homogeneity of the values of v when the system heterogeneity is low, where the probability density displays relatively uniform values as a function of v. Qualitatively similar results are obtained for the transverse velocity, as shown in Figure 5b. Note, however, that $H(u)$ appears to be less sensitive than $H(v)$ to σ_{Yref}^2, i.e. markers identified with different colors are more close together in Figure 5b than in Figure 5a (note the different vertical axis scale). Moreover, the values of $H(u)$ tend to saturate to an asymptotic value for $F_{\sigma^2} > 0.9$. The analysis suggests that an increase in F_{σ^2} can be expected to have important effects on the resulting pdfs of v even for $F_{\sigma^2} > 0.9$, while the pdf of the transverse velocity u only displays a minor shape transition in the same range of values.

We analyze now the relationship between (i) the correlation between the truncated Y^{tr} and the reference Y^{ref} field and (ii) the correlation between the ensuing v^{tr}, u^{tr} (associated with the truncated Y) and the reference fields v^{ref}, u^{ref}. Our objective is to assess the significance of the linear correlation and of the uncertainty coefficients in quantifying the representativeness (or fidelity) of the velocity field obtained upon through a truncated field with respect to the reference one. Figure 6 shows the results obtained by associating each value of linear and nonlinear correlation coefficients computed for the log-conductivity with the those obtained for the corresponding velocity components. Figure 6a suggests that the relationship between $\rho(Y^{ref}; Y^{tr})$ and $\rho(v^{ref}; v^{tr})$ depends on $\sigma^2_{Y^{ref}}$, with increasing correlation detected for decreasing $\sigma^2_{Y^{ref}}$. On the other hand, Figure 6b shows that the relationship between $UC(Y^{ref}; Y^{tr})$ and $UC(v^{ref}; v^{tr})$ is insensitive to the degree of heterogeneity in the reference permeability field and that the two indicators practically coincide, i.e. $UC(Y^{ref}; Y^{tr}) \approx UC(v^{ref}; v^{tr})$. The linear correlation coefficients are considerably larger than the corresponding uncertainty coefficients, for any considered value of F_{σ^2} and $\sigma^2_{Y^{ref}}$. In particular for $F_{\sigma^2} = 0.9$ we obtain a $\rho(v^{ref}; v^{tr}) \approx 0.93$ and $UC(v^{ref}; v^{tr}) \approx 0.5$. A similar behavior is observed for the transverse velocity u, where the maximum $\rho(u^{ref}; u^{tr})$ approaches unity while the corresponding value of $UC(u^{ref}; u^{tr})$ is below 0.7. Such striking quantitative differences between the two indicators can be explained upon observing the bivariate distributions obtained for the pairs $(v^{ref}; v^{tr})$ and $(u^{ref}; u^{tr})$, reported in Figure 7. In particular, we observe that the largest velocity values detected in the reference velocity distributions are consistently not found within the truncated fields v^{tr}. This discrepancy becomes more evident for decreasing values of F_{σ^2} (see Figure 7b).This result is of potential relevance to a variety of applications in which the presence of large values of v (jointly with their spatial organization) plays a crucial role, e.g., early arrival times of dissolved chemicals at control planes (e.g., [10]), evaluation of the level of system connectivity (see e.g., [27]) and solute dispersion mechanisms (see e.g., [18]). Comparison of Figure 6c-d also shows that the truncated fields tend to underestimate the magnitude of the transverse velocity component. The implication of this result is that the tortuosity of the resulting streamlines would be likely underestimated for the truncated fields with respect to the reference ones.This behavior would propagate to solute transport dynamics (see e.g., [5]), possibly resulting in incorrect predictions of solute time arrivals at control planes and inaccurate quantification of uncertainty.

Overall results in Figure 6 suggest that a linear correlation metric overestimates the representativeness of the truncated fields with respect to the reference fields. Therefore, the use of linear metrics would be misleading if employed to assess the convergence of truncated fields to a higher fidelity approximations. The low values of $UC(v^{ref}; v^{tr})$ suggests that surrogate models based on a KL-based expansion of log-conductivity may be practically unable to faithfully reproduce the full velocity pdfs, particularly as for what concern the large velocity tail of the distribution.

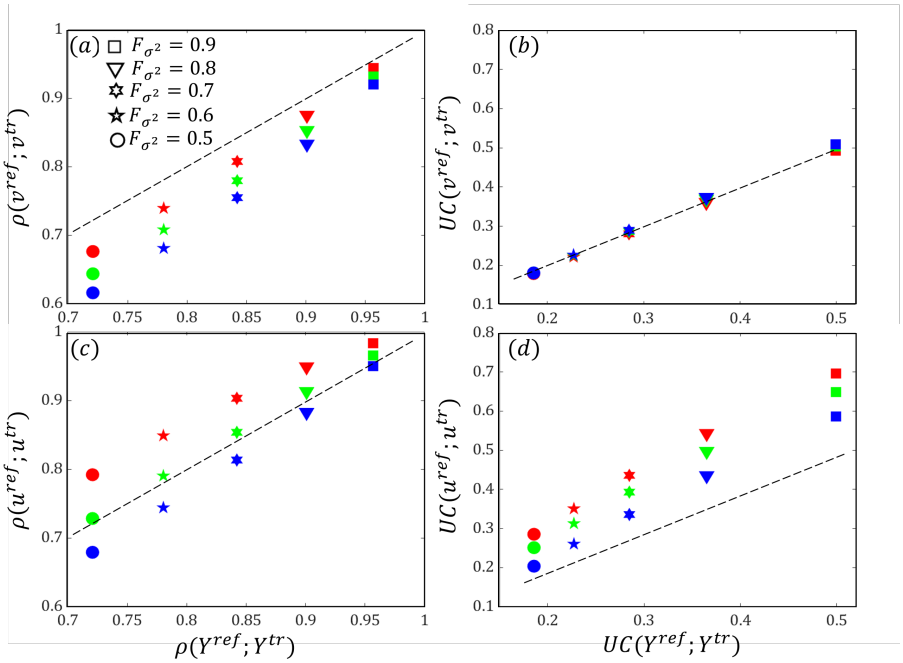

Fig. 6 Correlation analysis between truncated and references fields quantified by (a) linear correlation coefficients $\rho(Y^{ref};Y^{tr})$, $\rho(v^{ref};v^{tr})$, (b) uncertainty coefficients $UC(Y^{ref};Y^{tr})$, $UC(v^{ref};v^{tr})$, (c) linear correlation coefficients $\rho(Y^{ref};Y^{tr})$, $\rho(u^{ref};u^{tr})$, (b) uncertainty coefficients $UC(Y^{ref};Y^{tr})$, $UC(u^{ref};u^{tr})$. Results are for $\sigma^2_{Y^{ref}} = 0.5$ (red), 1.5 (green),2.5 (blue) and diverse values of F_{σ^2}, indicated by different markers. Dashed black lines indicate the axes bisectors.

3.2 Analysis of spatial correlation

We extend here the analysis to the characterization of the spatial auto-correlation of the velocity components v and u, as rendered by different values of F_{σ^2}, i.e. with increasing number of modes retained in the KL-expansion. To this end we evaluate distinct two-points correlation metrics, i.e., (i) the linear correlation coefficient $\rho(v;v(s))$, $\rho(u;u(t))$ and (ii) the uncertainty coefficient $UC(v;v(s))$, $UC(u;u(t))$ (see Section 2.2) where s and t are spatial lags in the longitudinal and transverse directions, respectively. These lags are here normalized by η, the correlation length assumed for Y^{ref}.

Figure 8 shows the trends of $\rho(v;v(s))$ and $UC(v;v(s))$ against s, for diverse values of F_{σ^2}, assigned $\sigma^2_{Y^{ref}} = 0.5$ and $\sigma^2_{Y^{ref}} = 2.5$. For both indicators we observe that spatial correlation tends to decrease with increasing F_{σ^2}. This result suggests that the use of truncated fields enforce larger spatial correlation than the reference one. We also observe that the linear correlation observed in the velocity components displays a slight inverse proportionality to $\sigma^2_{Y^{ref}}$, i.e. increasing heterogeneity reduces the linear correlation of the longitudinal velocity component, for a fixed lag. On the other hand, $UC(v;v(s))$ appears

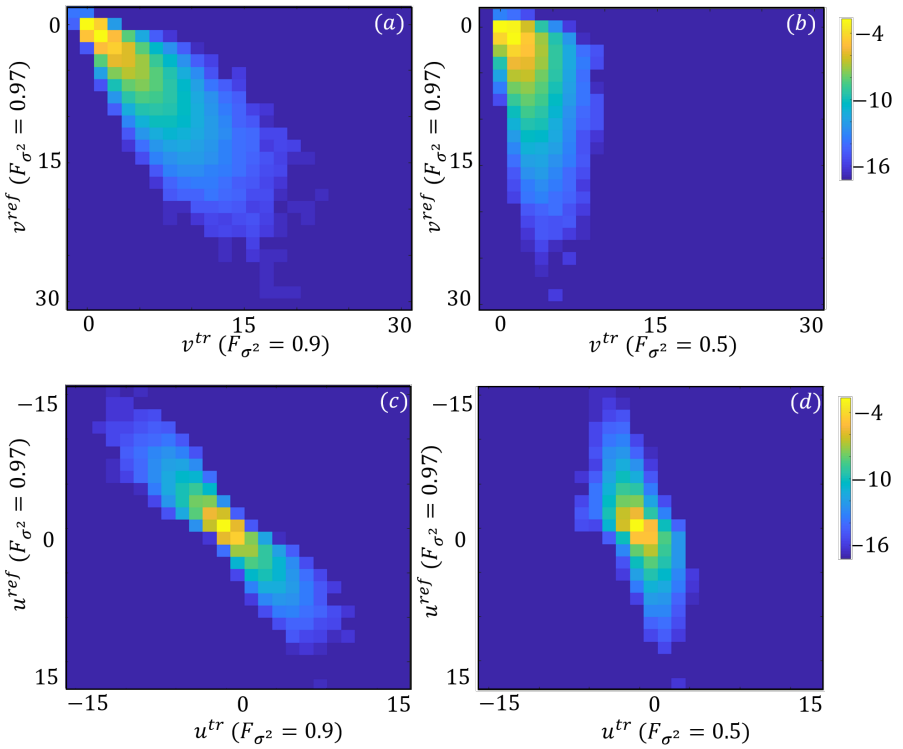

Fig. 7 Bivariate log-probability distributions between reference and truncated fields of the longitudinal v (a)-(b) and transverse u (c)-(d) velocity component. The truncated fields are obtained with (a) and (c) $F_{\sigma^2} = 0.9$, (b) and (d) for $F_{\sigma^2} = 0.5$, for $\sigma^2_{Y^{ref}} = 2.5$.

to be less sensitive than $\rho(v; v(s))$ to the value assumed by $\sigma^2_{Y^{ref}}$. Figure 8 also shows that $\rho(v; v(s))$ tends to be larger than $UC(v; v(s))$ assigned a given lag s. This result indicates that the linear correlation metric always scores higher than the nonlinear one, in line with results in Figure 6. To explain this result, Figure 9 displays the joint pdfs of $v; v(s)$ evaluated at spatial locations separated by lag of (a) one, (b) three and (c) six for $\sigma^2_{Y^{ref}} = 0.5$ and (d-f) for $\sigma^2_{Y^{ref}} = 2.5$. The joint probability distributions depicted in Figure 9 show two consistent features: (i) a high density in the upper left corner, aligned with the upper left - bottom right diagonal, and (ii) a fairly symmetrical dispersion of the probability density around this diagonal, whose intensity increases with spatial lag. The definition of $\rho(v; v(s))$ is such that feature (i) turns out to be the dominant one, i.e., $\rho(v; v(s))$ reflects the marked tendency of the joint probability density $p_{i,j}$ of displaying dominant terms when $i = j$. The linear correlation metric is instead poorly sensitive to the actual dispersion of the joint pdf. On the other hand, $UC(v; v(s))$ detects the dispersion of the joint pdf, which reflects non negligible probability of having uncorrelated pairs of v. This feature is emphasized as the spatial lag increases (e.g., compare panels a

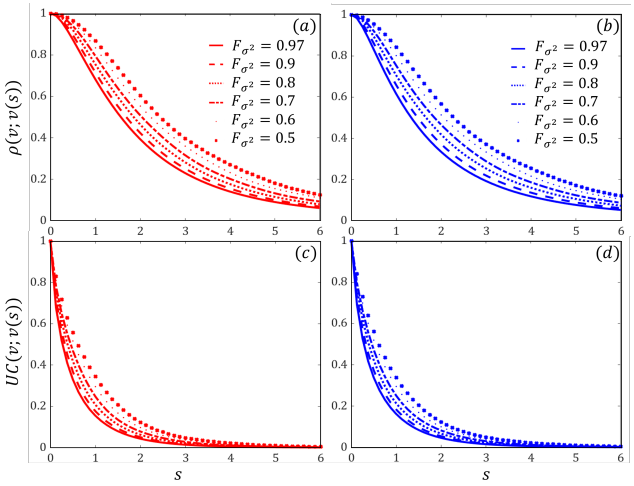

Fig. 8 (a)-(b) Linear Correlation Coefficients and (c-d) Uncertainty coefficients for pairwise values of the longitudinal component of velocity, i.e., $\rho(v; v(s))$, as a function of the spatial lag s, along the mean flow direction for (a,c) $\sigma^2(Y^{ref}) = 0.5$ and (b,d) $\sigma^2(Y^{ref}) = 2.5$. Lags are aligned with v and normalized by the Y^{ref} correlation length η. Diverse level of truncation (i.e., F_{σ^2}) of the KL-expansion are depicted, with different line types.

and c in Figure 9). These observations highlight the relevance of considering suitable summary metrics (such as the uncertainty coefficient) in order to characterize the whole behavior of the v and $v(s)$ joint pdf.

Figure 10 displays spatial correlation of the transverse flow velocity component u. As previously noted for v (see Figure 8), we observe that the spatial correlation of u measured by means of $\rho(u; u(t))$ or $UC(u; u(t))$ tends to decrease with F_{σ^2}, even though in a less marked fashion with respect to v. Figures 10a-b reveal that $\rho(u; u(t))$ assumes negative values, a feature related with the conservation of mass imposed by the flow equation (1). Note that, for $F_{\sigma^2} = 0.5, 0.6$ and t larger than two (approximately) the value of $\rho(u; u(t))$ displays some mild oscillations. This feature is associated with the fact that as fewer modes are retained in the KL-expansion the ensuing Y^{tr} fields exhibit a periodic behaviour in space determined by the shape of the retained modes. Inspection of Figure 10c-d suggests that $UC(u; u(t))$ is smaller than $\rho(u; u(t))$ for assigned $\sigma^2(Y^{ref})$ and F_{σ^2}, in line with the results obtained for v (see Figure 8). Furthermore, as fewer modes are retained in the KL-expansion (i.e, small F_{σ^2}) the values of $UC(u; u(t))$ increases, given t. Inspection of the inserts in 10c-d allows observing that also $UC(u; u(t))$ has a fluctuating behaviour as a function of t for $F_{\sigma^2} = 0.5, 0.6$ (see previous discussion of Figure 10 a-b). We note that the correlation detected by $UC(u; u(t))$ is very similar for the two considered $\sigma^2_{Y^{ref}}$. This result reflects the fact the joint log-probability of pairwise u values display a fairly similar shape in the two cases, as shown in Figure 11. A close comparison of the curves in Figure 10c-d reveals that $UC(u; u(t))$ tends to zero faster for $\sigma^2_{Y^{ref}} = 0.5$ than for $\sigma^2_{Y^{ref}} = 2.5$ (see results in log-

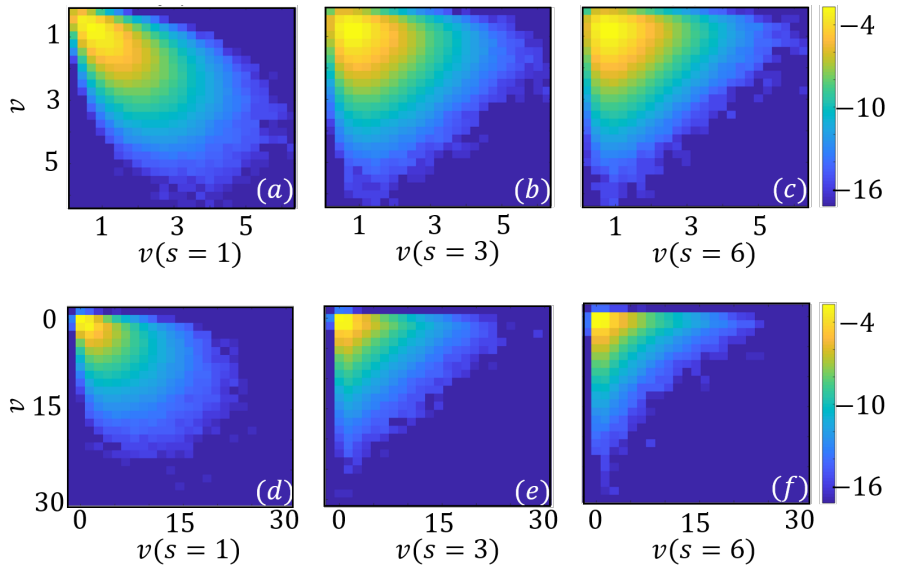

Fig. 9 Joint log-probability distribution of pairwise values of the longitudinal velocity, i.e., v, separated by different spatial lags s for $\sigma^2(Y^{ref}) = 0.5$ and (d-f) $\sigma^2(Y^{ref}) = 2.5$. Lags are aligned with v and normalized by the Y^{ref} correlation length η, results are for $F_{\sigma^2} = 0.97$.

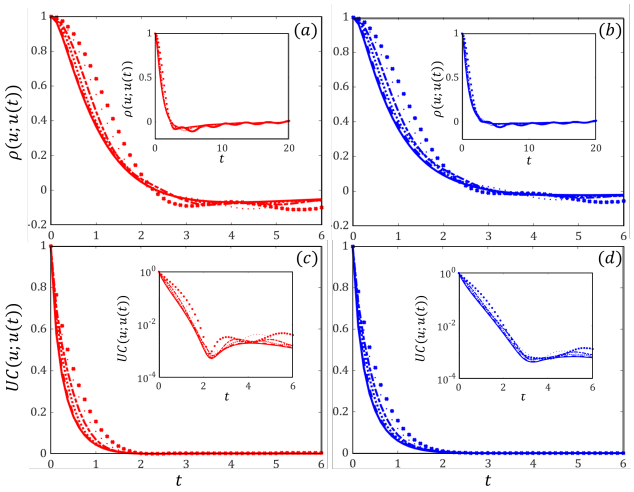

Fig. 10 (a)-(b) Linear Correlation Coefficients and (c-d) Uncertainty coefficients for pairwise values of the longitudinal component of velocity, i.e., $\rho(u; u(t))$, as a function of the spatial lag t, along the transverse direction for (a,c) $\sigma^2(Y^{ref}) = 0.5$ and (b,d) $\sigma^2(Y^{ref}) = 2.5$. Lags are aligned with u and normalized by the Y^{ref} correlation length η. Diverse level of truncation (i.e., F_{σ^2}) of the KL-expansion are depicted, with different line types (see legend in Figure 8). Inserts in (a)-(b) extend results to the interval $t \in [0, 20]$, in (c)-(d) display UC values with a log-axis scale.

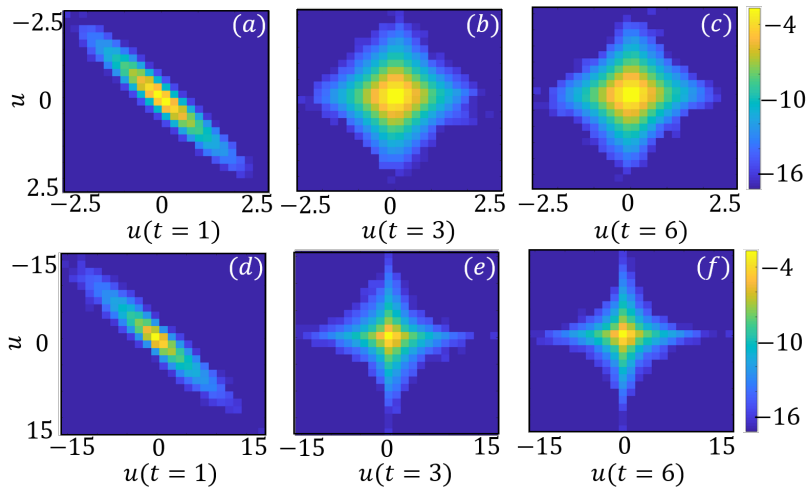

Fig. 11 Joint log-probability distribution of pairwise values of the transverse velocity, i.e., u, separated by different spatial lags t for $\sigma^2(Y^{ref}) = 0.5$ and (d-f) $\sigma^2(Y^{ref}) = 2.5$. Lags are aligned with u and normalized by the Y^{ref} correlation length η, results are for $F_{\sigma^2} = 0.97$.

scale reported in the inserts). This result is in line with the distribution $p_{i,j}$ being more evenly distributed across the bins in Figure 11b-c than in Figure 11d-e. These differences appear subtle but may play a role in determining the spatial correlation of transverse displacement in solute transport models, which play a relevant role in the upscaling of solute mixing, transport dynamics and associated risk assessment (e.g., [13, **?**]). The spatial organization of the transversal flow component plays a crucial role in the transport dynamics of dissolved chemical (either passive or reactive ones). As such, it appears crucial to rely on summary metrics that can quantify the characteristics of joint pdfs of interest.

4 Conclusions

In the context of fluid flow within two-dimensional heterogeneous porous media we assess the impact of diverse degree of truncation of the Karhunen-Loève (KL) expansion of the log-conductivity (Y) field in terms of the ensuing (i) variability of the velocity field components and (ii) of their correlation in space. Analysis of point wise velocity components indicates that the variance can be an incomplete descriptor of the uncertainty associated with the velocity field and suggests to focus on the Shannon entropy as a metric to quantify the degree of uncertainty associated with both velocity components. In this context, one of the objectives of our study is to quantify the similarity of truncated input and output fields to a reference probability distribution. To this end, we tested

linear and nonlinear correlation metrics. We find that the linear correlation coefficients, for both Y and the velocity components, are always larger than the corresponding Uncertainty coefficients. This result shows that the linear metric overestimates the degree of similarity with respect to the nonlinear one. A direct implication of this comparison is that linear correlation metrics may lead to misleading results when comparing velocity fields associated with different levels of truncation of the KL-expansion. Our results also suggest that truncated velocity and log-conductivity have the same uncertainty coefficients if compared against their reference fields, a result which can be valuable to assess the relation between input and output variables in the approximation of stochastic groundwater flow outputs.

In line with these results, our analysis also shows that the Linear Correlation Coefficient typically exceeds the Uncertainty Coefficient when evaluating the spatial correlation between pairwise velocity components. Furthermore, we found that as the number of modes in retained in the KL-expansions decreases both velocity components become more correlated in space (according to both the linear and nonlinear metrics, even though with different degrees). These results could be relevant in assessing the correlation structure in solute mass velocity distributions, which are employed in upscaled or effective transport models.

Overall our results suggest that it is beneficial to assess the impact of diverse degree of truncation of the input KL-expansion by means of Information Theory metrics. The latter can provide a more comprehensive assessment of the degree of uncertainty and of dependence (either among point-wise, across levels of truncation, and pairwise velocity components, across space lags) than standard metrics (e.g., variance and Linear Correlation coefficient).

IT metrics are capable of capturing salient features of point-wise velocity components and their spatial organization and therefore they may help in quantifying the impact of approximations (such as those entailed by the KL truncation) on analyses of practical relevance. These include, for example, uncertainty quantification and risk assessment for flow and solute transport scenarios. Moreover, our results indicate that IT metrics could be employed upscaling methodologies aimed at characterizing solute transport and mixing in heterogeneous media, which are routinely constrained to observed point-wise and two-points correlation indicators of the flow components.

References

1. de Barros, F.P.J., Rubin, Y.: A risk-driven approach for subsurface site characterization. Water Resources Research **44**(1) (2008). DOI 10.1029/2007WR006081
2. Bianchi, M., Pedretti, D.: Geological entropy and solute transport in heterogeneous porous media. Water Resources Research **53**(6), 4691–4708 (2017). DOI 10.1002/2016WR020195
3. Bianchi, M., Pedretti, D.: An entrogram-based approach to describe spatial heterogeneity with applications to solute transport in porous media. Water Resources Research **54**(7), 4432–4448 (2018). DOI 10.1029/2018WR022827

4. Butera, I., Vallivero, L., Ridolfi, L.: Mutual information analysis to approach nonlinearity in groundwater stochastic fields. Stochastic Environmental Research and Risk Assessment **32**(10), 2933–2942 (2018). DOI 10.1007/s00477-018-1591-4
5. Comolli, A., Hakoun, V., Dentz, M.: Mechanisms, upscaling, and prediction of anomalous dispersion in heterogeneous porous media. Water Resources Research **55**(10), 8197–8222 (2019). DOI 10.1029/2019WR024919
6. Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, USA (2006)
7. Dagan, G.: Flow and Transport in Porous Formations (1989). OCLC: 1053810248
8. Dell'Oca, A., Porta, G., Guadagnini, A., Riva, M.: Space-time mesh adaptation for solute transport in randomly heterogeneous porous media. Journal of Contaminant Hydrology **212**, 28–40 (2018). DOI 10.1016/j.jconhyd.2017.07.001
9. Esfandiar, B., Porta, G., Perotto, S., Guadagnini, A.: Anisotropic mesh and time step adaptivity for solute transport modeling in porous media. In: P. S., F. L. (eds.) SEMA SIMAI Springer Series,New Challenges in Grid Generation and Adaptivity for Scientific Computing, pp. 1–25. Springer, Milan (2014)
10. Fiori, A., Jankovic, I.: On preferential flow, channeling and connectivity in heterogeneous porous formations. Mathematical Geosciences **44**(2), 133–145 (2012). DOI 10.1007/s11004-011-9365-2
11. Gelhar, L.: Stochastic Subsurface Hydrology. Prentice-Hall (1993)
12. Goodwell, A.E., Kumar, P.: Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. Water Resources Research **53**(7), 5920–5942 (2017). DOI 10.1002/2016WR020216
13. Hakoun, V., Comolli, A., Dentz, M.: Upscaling and Prediction of Lagrangian Velocity Dynamics in Heterogeneous Porous Media. Water Resources Research (2019). DOI 10.1029/2018WR023810
14. Hecht, F.: New development in freefem++. J. Numer. Math. **20**(3-4), 251–265 (2012)
15. Kaiser, A., Schreiber, T.: Information transfer in continuous processes. Physica D: Nonlinear Phenomena **166**(1), 43 – 62 (2002). DOI https://doi.org/10.1016/S0167-2789(02)00432-3
16. Katsoulakis, M.A., Plechaĉ, P.: Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. The Journal of Chemical Physics **139**(7), 074,115 (2013). DOI 10.1063/1.4818534
17. Kikuchi, C.P., Ferré, T.P.A., Vrugt, J.A.: On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models. Water Resources Research **51**(6), 4454–4481 (2015). DOI 10.1002/2014WR016795
18. Le Borgne, T., de Dreuzy, J.R., Davy, P., Bour, O.: Characterization of the velocity field organization in heterogeneous media by conditional correlation. Water Resources Research **43**(2) (2007). DOI 10.1029/2006WR004875
19. Lin, G., Tartakovsky, A.: An efficient, high-order probabilistic collocation method on sparse grids for three-dimensional flow and solute transport in randomly heterogeneous porous media. Advances in Water Resources **32**(5), 712–722 (2009). DOI 10.1016/j.advwatres.2008.09.003
20. Liu, Z., Liu, Z., Peng, Y.: Dimension reduction of Karhunen-Loeve expansion for simulation of stochastic processes. Journal of Sound and Vibration **408**, 168–189 (2017). DOI 10.1016/j.jsv.2017.07.016
21. Loève, M.: Probability theory. 2: ..., 4. ed edn. No. 46 in Graduate texts in mathematics. Springer, New York (1978). OCLC: 310756087
22. Marzouk, Y.M., Najm, H.N.: Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. Journal of Computational Physics **228**(6), 1862–1902 (2009). DOI 10.1016/j.jcp.2008.11.024
23. Moslehi, M., de Barros, F.: Uncertainty quantification of environmental performance metrics in heterogeneous aquifers with long-range correlations. Journal of Contaminant Hydrology **196**, 21–29 (2017). DOI 10.1016/j.jconhyd.2016.12.002
24. Nearing, G.S., Ruddell, B.L., Clark, M.P., Nijssen, B., Peters-Lidard, C.: Benchmarking and process diagnostics of land models. Journal of Hydrometeorology **19**(11), 1835–1852 (2018). DOI 10.1175/JHM-D-17-0209.1

25. Neuman, S.P., Guadagnini, A., Riva, M., Siena, M.: Recent Advances in Statistical and Scaling Analysis of Earth and Environmental Variables. In: P.K. Mishra, K.L. Kuhlman (eds.) Advances in Hydrogeology, pp. 1–25. Springer New York, New York, NY (2013). DOI 10.1007/978-1-4614-6479-21
26. Oberst, S., Niven, R.K., Lester, D.R., Ord, A., Hobbs, B., Hoffmann, N.: Detection of unstable periodic orbits in mineralising geological systems. Chaos: An Interdisciplinary Journal of Nonlinear Science **28**(8), 085,711 (2018). DOI 10.1063/1.5024134
27. Renard, P., Allard, D.: Connectivity metrics for subsurface flow and transport. Advances in Water Resources **51**, 168 – 196 (2013). DOI https://doi.org/10.1016/j.advwatres.2011.12.001
28. Ruddell, B.L., Kumar, P.: Ecohydrologic process networks: 1. identification. Water Resources Research **45**(3) (2009). DOI 10.1029/2008WR007279
29. Salandin, P., Fiorotto, V.: Solute particles in highly heterogeneous aquifers. Water Resources Research **34**(5), 949–961 (1998)
30. Shannon, C.: A mathematical theory of communication. Bell System Technical Journal **27**(3), 379–423 (1948). DOI 10.1002/j.1538-7305.1948.tb01338.x
31. Theil, H.: Statistical decomposition analysis. North-Holland Publishing Co (1972). DOI 10.1017/S0770451800006916
32. Wright, E., Sund, N., Richter, D., Porta, G., Bolster, D.: Upscaling Mixing in Highly Heterogeneous Porous Media via a Spatial Markov Model. Water **11**(1), 53 (2018). DOI 10.3390/w11010053
33. Zahm, O., Constantine, P., Prieur, C., Marzouk, Y.: Gradient-based dimension reduction of multivariate vector-valued functions. arXiv:1801.07922 [math] (2018). URL http://arxiv.org/abs/1801.07922. ArXiv: 1801.07922
34. Zhang, D., Lu, Z.: An efficient, high-order perturbation approach for flow in random porous media via KarhunenLove and polynomial expansions. Journal of Computational Physics **194**(2), 773–794 (2004). DOI 10.1016/j.jcp.2003.09.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0021999103005072