Stochastic Environmental Research and Risk Assessment (2020) 34:1591–1606
https://doi.org/10.1007/s00477-020-01865-2

ORIGINAL PAPER

Impact of geostatistical reconstruction approaches on model calibration for flow in highly heterogeneous aquifers

Martina Siena[1] · Monica Riva[1]

Published online: 2 September 2020

Abstract

Our study is aimed at assessing the extent at which relying on differing geostatistical approaches may affect characterization of the connectivity of geomaterials (or *facies*) and, in turn, model calibration outputs in highly heterogeneous aquifers. We set our study within a probabilistic framework, by relying on a numerical Monte Carlo (MC) approach. The reconstruction of the spatial distribution of geomaterials and flow simulations are patterned after a field scenario corresponding to the aquifer system serving the city of Bologna (Northern Italy). Two collections of MC realizations of facies distributions, conditional on available lithological data, are generated through two alternative geostatistically-based techniques, i.e., Sequential Indicator and Transition-Probability simulation. Hydraulic conductivity values of the least- and most-conductive facies are estimated within each MC simulation in the context of a Maximum Likelihood (ML) approach by considering available piezometric data. We provide evidence that the choice of the facies reconstruction technique (1) impacts the degree of connectivity of facies whose proportions are close to the percolation threshold while (2) is not sensibly affecting the connectivity associated with facies whose proportions are much larger than the percolation threshold. By relying on the unique (lithological and hydrological) data-set at our disposal, we also explore the performance of ML-based model identification criteria to (1) discriminate amongst competitive facies reconstruction geostatistical models and (2) quantify the (posterior probabilistic) weight associated with each model. We then show that ML-based model averaging provides estimates of hydraulic heads which are slightly more in agreement with available data when compared to the best-performing realization in the T-PROGS set than considering its counterpart associated with the SISIM-based collection.

Keywords Geostatistical reconstruction · Connectivity metrics · Groundwater flow model calibration · Bayesian model averaging · Transition probability

1 Introduction

Probabilistic reconstruction of complex aquifer systems is nowadays considered a common practice to account for uncertainty resulting from our lack of knowledge of highly-heterogeneous subsurface structures and spatial distribution of attributes of aquifer systems. In a stochastic framework, aquifer heterogeneity can be conceptualized by considering system attributes, such as hydraulic conductivity, as random functions. Widely employed geostatistical methods (e.g., Sequential Gaussian Simulation; Deutsch and Journel 1998) describe hydraulic properties as multivariate Gaussian random fields and quantify the degree of aquifer heterogeneity upon relying on a covariance (or variogram) structure that can be inferred from available data. In the presence of sharp interfaces between high and low-conductivity regions, a composite media approach, considering the system as composed by diverse lithological units (or facies), is typically adopted (Guadagnini et al. 2003; Winter et al. 2006). Categorical rather than continuous random fields can be considered in order to represent the distributions of facies. For this purpose, Sequential Indicator Simulation (SISIM; Goovaerts 1997; Deutsch and Journel 1998), which relies on indicator variograms (inferred, e.g., from borehole lithological data), is one of the most extensively applied methods (Deutsch 2006; Felletti et al. 2006; Marini et al. 2019). While being characterized

✉ Martina Siena
martina.siena@polimi.it

[1] Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Piazza L. Da Vinci 32, 20133 Milan, Italy

by ease of implementation, this method may not honor volumetric proportions of facies, as inferred from available data, because it tends to underestimate less-prevalent facies (Emery 2004; He et al. 2009), and/or may not fully capture patterns of facies connectivity (Gomez-Hernandez and Wen 1998; Kerrou et al. 2008), an issue which is particularly noticeable in three-dimensional systems (Dell'Arciprete et al. 2012, 2014). An alternative approach has been embedded in the Transition Probability Geostatistical Simulation approach, T-PROGS (Carle and Fogg 1996, 1997). The latter makes use of available lithological/sedimentological data to evaluate transition probabilities between facies, which are then interpreted through Markov-chain modeling techniques. This approach, as well as other Markov-chain based methods (e.g., Elfeki and Dekking 2001; Li 2007; Langousis et al. 2018), also (1) enables one to include information on spatial juxtapositional tendencies and soft conditioning data and (2) is deemed as more accurate than its covariance-based counterparts to represent volumetric proportions, mean lengths, and connectivity patterns driven by facies distributions (Weissmann et al. 1999; He et al. 2014, 2015; Koch et al. 2014).

Characterization of connectivity is key to quantify the degree of flow heterogeneity and the ensuing early (or late) breakthrough of dissolved chemicals at critical targets (Cvetkovic et al. 2014; Henri et al. 2015; Zinn and Harvey 2003), with direct implications in several contexts, including, e.g., remediation of contaminated sites and environmental risk assessment. Increasing efforts have been devoted to investigate the effects of various geostatistical approaches on connectivity metrics (Bakshevskaia and Pozdniakov 2016; Dell'Arciprete et al. 2012, 2014; Lee et al. 2007; Mohammadi et al. 2020; Sharifzadehlari et al. 2018; Vassena et al. 2010). Other studies addressed the impact different geostatistical methods on the accuracy of estimated facies distribution (He et al. 2009, Kessler et al. 2013; Marini et al. 2019; Park et al. 2007), hydraulic head and flux fields (Lee et al. 2007; Bianchi et al. 2015), or spreading of dissolved chemicals migrating in aquifer systems (Maghrebi et al. 2015; Siirila-Woodburn and Maxwell 2015).

Here, we investigate the extent at which the degree of connectivity resulting from a selected reconstruction method may alter the outcomes of a model calibration process. In this context, we rely on suites of numerical simulations of groundwater flow in the aquifer system serving the city Bologna (Northern Italy), whose internal architecture (in terms of spatial arrangement of geomaterials, or facies) is modeled within a probabilistic Monte Carlo (MC) framework. In this context, we rely on generating collections of equally-probable realizations of spatial distributions of facies conditional on a unique and extensive dataset of lithological information. We rest on the approaches associated with SISIM and T-PROGS and generate two collections (or ensembles) of MC realizations. We then evaluate (1) the variability of connectivity within each of the two sets of facies distributions (hereafter termed *intra-ensemble variability*), and (2) the extent at which the generation method affects facies connectivity (i.e., the *inter-ensemble variability*). A groundwater flow model is then calibrated for each MC realization by making use of available piezometric data. Ensuing estimates of hydraulic conductivity, K, are obtained through a Maximum Likelihood (ML) inverse modeling approach (Carrera and Neuman 1986). Our primary goal is to assess whether intra- or inter-ensemble variability of connectivity may have an effect on estimated values of K and their uncertainty. We rely on the ML framework to evaluate the model identification criterion proposed by Kashyap (1982) to (1) rank realizations within each ensemble according to their relative skill to interpret available data and (2) compute the likelihood associated with each realization which is then used for the implementation of a multi-model approach. The latter scheme relies on viewing each of the alternative MC realizations as a candidate hydrogeological model of the investigated site.

An additional goal of our study is the identification of effects that the adopted geostatistical reconstruction method might have on (1) the uncertainty associated with a desired modeling goal and (2) on the predictive performance of the outcomes of a Maximum Likelihood Bayesian Model Averaging (MLBMA; Ye et al. 2004; Neuman 2003), a framework of analysis that has been increasingly adopted to cope with the high degree of uncertainty related to field-scale flow and transport models (see, e.g., Lu et al. 2015; Samani et al. 2018). In essence, MLBMA enables the joint use of multiple candidate models by considering a weighted average of their predictions of the target modeling goal and making use of the ML-based posterior probability of each model. While such an approach might improve our predictive capability, with a more robust assessment of prediction uncertainty with respect to the output of a single model (Ye et al. 2004), it has also been observed (Winter and Nychka 2010) that the average of an ensemble of models can actually provide predictions of higher quality than those associated with best performing single model only if the individual models in the ensemble give rise to a wide range of outcomes.

2 Materials and methods

2.1 Site description

The study area lies between the Apennines margin and the alluvial plain surrounding the city of Bologna, in the

Emilia Romagna Region, Northern Italy (Fig. 1a). The area is part of the Po River Basin, which is a syntectonic sedimentary wedge forming the infill of the Pliocene–Pleistocene foredeep (Ricci Lucchi 1984).

As shown schematically in Fig. 1b, the basin is structured into three main groundwater bodies: Group A (with elevations ranging between 10 m a.s.l. and 150–200 m a.s.l.; which is, in turn, composed by four sub-units, namely A1, A2, A3 and A4); Group B (from 150–200 m to 300–450 m a.s.l.); and Group C (below 450 m a.s.l.). A recent classification (Regione Emilia-Romagna 2010), in compliance with Directive 2000/60/CE, considers the exposure to anthropogenic impacts and the paleo-geographical signatures of the diverse groups and identifies (1) an upper confined aquifer, formed by A1 and A2; and (2) a lower confined unit, which includes A3, A4, B and C. These aquifers are separated by discontinuous aquitards of variable thickness and composed by diverse lithological units (see also Guadagnini et al. 2004; Short et al. 2010; Molinari et al. 2012).

The domain investigated here extends across a surface of 20 × 23 km^2 in the horizontal plane and from − 450 m to 100 m a.s.l. along the vertical direction. Lithostratigraphic information are available at about 1300 boreholes, whose planar location is displayed in Fig. 2a. Borehole depths range between 10 and 600 m (see Fig. 2b). These data allowed identifying four main lithofacies within the area. These correspond to clay, gravel, silt, and sand, whose volumetric proportion (p_k, with $k = c$, g, si, sa, for clay, gravel, silt and sand, respectively) are $p_c = 52.3\%$, $p_g = 28.1\%$, $p_{si} = 13.3\%$, and $p_{sa} = 6.3\%$. The least- (clay) and the most- (gravel) permeable facies are associated with the two largest p_k values, encompassing (in total) about 80% of the aquifer. The area is characterized by an intense anthropic activity, water supplies being drawn mainly from the underlying aquifer. The intense exploitation of groundwater resources has led to the formation of a massive cone of depression superimposed to the mean natural subsurface flow, which is aligned (on average) along a South-West (i.e. from the Apennines) - North-East (i.e., towards the alluvial plain of the Po River) direction.

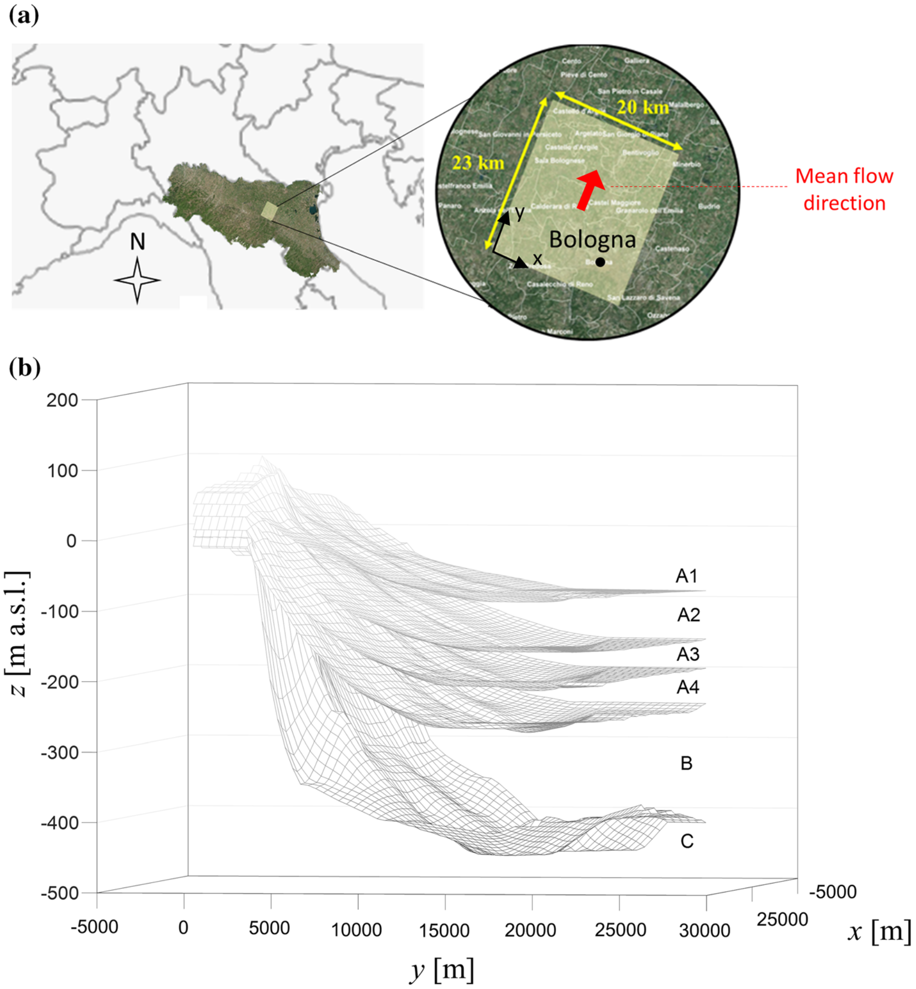


Fig. 1 **a** Location of the investigated domain (Emilia Romagna, Northern Italy); **b** Three-dimensional representation of the aquifer-group bases within the investigated domain

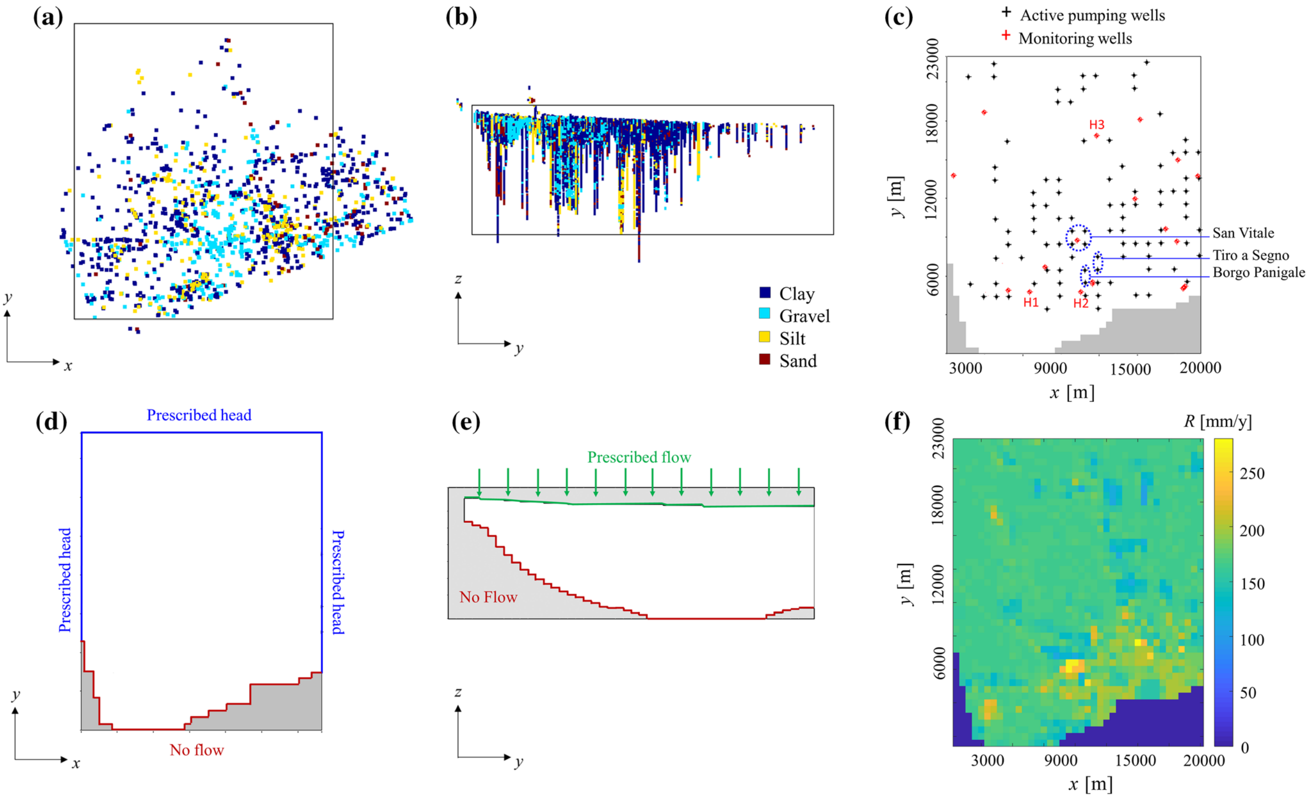


Fig. 2 Horizontal (**a**) and vertical (**b**) projections of the lithostratigrafic data available along the boreholes; **c** Location of pumping (black symbols) and monitoring (red symbols) wells within the domain of simulation (H1, H2 and H3 being those selected for validation in Sect. 3); **d–e** Boundary conditions of the groundwater flow model (grey areas indicate model inactive cells); **f** Contour map of surface recharge, *R*, computed according to Eq. (2)

The location of the three major well fields assisting urban water supply (i.e., Borgo Panigale, San Vitale and Tiro a Segno) is illustrated in Fig. 2c together with all pumping wells exploited for industrial/agricultural needs. The total volume of water withdrawn is estimated as 4.3×10^7 m^3/year and encompasses civil (78%), industrial (18%) and agricultural (4%) uses. Figure 2c also includes 20 monitoring wells, at which hydraulic head measurements are available and are employed for groundwater flow inverse modeling (see Sect. 2.4).

2.2 Geostatistical reconstruction

The geological structure of the aquifer has been reconstructed by relying on lithostratigraphic data available at 1300 boreholes (see Fig. 2a) and by making use of two geostatistically-based methods, corresponding to (1) a Sequential Indicator simulation (SISIM; e.g., Goovaerts 1997; Remy et al. 2009) and (2) a Transition Probability simulation (T-PROGS; e.g., Carle and Fogg 1996, 1997) approach. These techniques have been extensively used to reconstruct spatial distributions of a correlated categorical variable, $Z = \{1, \ldots, n_{facies}\}$ (n_{facies} being the number of facies/categories in the system). Here we summarize for each method the key steps of the generation procedure.

2.2.1 SISIM

1. We define $Z(\mathbf{x}) = \{1, 2, 3, 4\}$ as the discrete variable indicating the facies that can be found at location $\mathbf{x}$ and the indicator function, $I_k(\mathbf{x})$ ($k = 1, \ldots, 4$), such that $I_k(\mathbf{x}) = 1$ if $Z(\mathbf{x}) = k$, and $I_k(\mathbf{x}) = 0$ otherwise.
2. For each facies, we compute the sample directional variogram, $\gamma_{k,\alpha}(s_\alpha)$, where s_α is the separation distance along direction α (with $\alpha = x, y, z$).
3. The variogram $\gamma_{k,\alpha}(s_\alpha)$ is interpreted via a Maximum Likelihood (ML) approach with an exponential model (see Eq. 6).
4. Three-dimensional conditional Monte Carlo realizations of the spatial distribution of the four identified facies are obtained by making use of (1) the conditioning data available at 1300 boreholes included in the investigated domain (see Fig. 2a) and (2) the directional variogram model selected at step (3) via a sequential simulation approach (see, e.g., Remy et al. 2009).

2.2.2 T-PROGS

1. We define the transition probability matrix, $\mathbf{T}(\mathbf{s})$, with entries $t_{jk}(\mathbf{s}) = \Pr\{Z(\mathbf{x}+\mathbf{s}) = k | Z(\mathbf{x}) = j\}$, representing the probability for facies k to be found at $(\mathbf{x} + \mathbf{s})$ conditioned by the presence of facies j at $\mathbf{x}$.
2. Sample transiograms, representing the variation of $t_{jk}(\mathbf{s})$ with $\mathbf{s}$, are evaluated for all possible pairs of categories (j, k) on the basis of all available lithological data.
3. Sample transiograms are interpreted by a Markov-chain model as $\mathbf{T}(\mathbf{s}) = \exp(\mathbf{Rs})$, $\mathbf{R}$ being the transition rate matrix. Elements r_{jk} of $\mathbf{R}$ are inferred from sample transiograms as $r_{jk} = \partial t_{jk} / \partial s|_{s \to 0}$.
4. A sequential procedure is applied to assign a category in each unsampled point (see, e.g., Weissmann et al. 1999).
5. Facies distribution resulting after the sequential procedure is adjusted by a simulated quenching algorithm (Carle 1997) to minimize the discrepancy between the resulting experimental transiograms and the theoretical Markov-chain model inferred from the data.

For each geostatistical reconstruction technique, we generate a collection of n = 100 Monte Carlo (MC) realizations of facies distributions. We then investigate the impact of the type of geostatistical simulation approach on the degree of facies connectivity through the set of metrics defined in the following section.

2.3 Connectivity metrics

Let Ω be a three-dimensional domain which is discretized by n_{tot} (grid) cells and Ω_k be the subset of n_k cells associated with the k-th facies. Two cells of Ω_k, identified by the coordinates of their centroid (here denoted as $\mathbf{x}_A$ and $\mathbf{x}_B$, respectively) are defined as *connected* if there is a sequence of neighboring cells (from $\mathbf{x}_A$ to $\mathbf{x}_B$) completely included in Ω_k. A group of connected cells within Ω_k is denoted as a *cluster*. Connectivity indicators that are commonly used (e.g., Lee et al. 2007; Hovadik and Larue 2007; Renard and Allard 2013) for the characterization of three-dimensional subsurface systems include: (1) the total number of clusters composing Ω_k, here denoted as $N_{C,k}$; (2) $\Theta_k = C_{\max,k} / n_{tot}$, where $C_{\max,k}$ is the number of cells in the largest cluster of facies k; and (3) $N_{I,k}$, corresponding to the number of isolated cells of facies k. Clearly, the connectivity of facies k increases as Θ_k increases and as $N_{C,k}$ and $N_{I,k}$ decrease.

Additional insights on facies connectivity can be offered by the analysis of the connectivity function (Renard and Allard 2013; Western et al. 2001), $\tau_{k,\alpha}(s_\alpha)$. The latter represents the probability for two cells of category/facies k separated by a distance s_α along direction α to be connected, i.e.,

$$\tau_{k,\alpha}(s_\alpha) = \frac{N(\mathbf{x}_A \leftrightarrow \mathbf{x}_B | \mathbf{x}_A \in \Omega_k, \mathbf{x}_B \in \Omega_k, \mathbf{x}_A - \mathbf{x}_B = s_\alpha \mathbf{e}_\alpha)}{N(\mathbf{x}_A \in \Omega_k, \mathbf{x}_B \in \Omega_k, \mathbf{x}_A - \mathbf{x}_B = s_\alpha \mathbf{e}_\alpha)} \quad with \;\; \alpha = \{x, y, z\} \tag{1}$$

where $\mathbf{e}_\alpha$ is the unit vector along direction α, $N(\mathbf{x}_A \in \Omega_k, \mathbf{x}_B \in \Omega_k, \mathbf{x}_A - \mathbf{x}_B = s_\alpha \mathbf{e}_\alpha)$ indicates the number of pairs of cells, $(\mathbf{x}_A, \mathbf{x}_B)$, belonging to category k that are separated by distance s_α along direction α and $N(\mathbf{x}_A \leftrightarrow \mathbf{x}_B | \mathbf{x}_A \in \Omega_k, \mathbf{x}_B \in \Omega_k, \mathbf{x}_A - \mathbf{x}_B = s_\alpha \mathbf{e}_\alpha)$ is the number of such pairs which also belong to the same cluster (this condition being expressed by $\mathbf{x}_A \leftrightarrow \mathbf{x}_B$).

The behavior of the connectivity function, as well as of the other indicators mentioned above, can be interpreted in the framework of percolation theory. The latter was originally formulated for uncorrelated Bernoulli random fields defined on infinite domains (e.g., Stauffer and Aharony 1992 and references therein) and considers a system composed by two phases (e.g., a solid and a void phase) in which, at any node of the grid, the field can be either 1 (void phase) or 0 (solid phase) with probability p and $(1 - p)$, respectively. A realization of such a system would hence be formed by Ω_1, i.e., the set of grid cells where the field is equal to 1, and its complementary set, Ω_0. The basic principle of percolation theory is that there is a critical value of p, termed *percolation threshold* (p_t), such that the probability P for Ω_1 to form a unique percolating cluster is equal to 1 if $p \geq p_t$ and P = 0 if $p < p_t$. As highlighted by Stauffer and Aharony (1992), the value of p_t decreases as (1) the grid dimension and (2) the number of neighbors of a grid cell increase. For a three-dimensional cubic grid with six neighbors, $p_t = 31\%$.

The application of percolation theory to geological settings is not straightforward, since facies distributions (1) are defined on a finite domain and (2) are spatially correlated. As a consequence, the probability of occurrence of a unique percolating cluster of facies k would not be a step function (i.e., P = 0 if $p_k < p_t$ and P = 1 if $p_k \geq p_t$), and would rather increase gradually from 0 to 1 over a range of p_k values. As illustrated by Harter (2005) and Hovadik and Larue (2007), this range widens as the analyzed set-up departs from the theoretical conditions, i.e., as the number of grid cells decreases and as the facies correlation length increases. Moreover, spatial correlation enhances connectivity, resulting in a decrease of the value of p_k at which percolation may occur. As highlighted by Renard and Allard (2013), percolation theory can also assist characterizing the behavior of the connectivity function on finite grids. For example one can note that (1) when $p_k < p_t$ then

$\tau_{k,j}$ will rapidly decrease to 0 as the separation distance s increases; and (2) when $p_k \geq p_t$, $\tau_{k,j}$ will tend asymptotically to $\tau_k^\infty = \left(C_{\max,k} \big/ n_k\right)^2$. All of these aspects are investigated in Sect. 3.

2.4 Groundwater flow model and model calibration

We evaluate three-dimensional steady state groundwater flow for each MC realization of facies distribution upon relying on the numerical code MODFLOW-2005 (Harbaugh 2005). On the basis of the available lithostratigraphic information, the system is discretized into $n_{tot} = 40 \times 46 \times 110$ cells of uniform size $\Delta x = \Delta y = 500$ m (along the horizontal) and $\Delta z = 5$ m (along the vertical). Figure 2d–e illustrate the simulation domain together with the adopted boundary conditions, where no-flow is imposed along the south-west boundary (i.e., corresponding to the Apennines margin) and along the bottom boundary, which coincides with the base of unit B (see Fig. 1b). Values of prescribed hydraulic head, h, are set along the remaining lateral boundaries (see Fig. 2d), as inferred from available piezometric data. Recharge, R, at the top of the domain is modeled as

$$R = P - Q - E + L \tag{2}$$

where P is rainfall, Q is surface runoff (land-use dependent, evaluated on the basis of the classical Curve Number method), E represents evapotranspiration (evaluated according to a modified Hargreaves method based on temperature data; Hargreaves and Allen 2003), and L quantifies water-pipe losses in the urban areas. The spatial distribution of R employed in the model is depicted in Fig. 2f.

A preliminary sensitivity analysis (details not shown) highlights that spatial distribution of hydraulic head is not significantly affected by conductivity values associated with silt, K_{si}, and sand, K_{sa} (the two categories with the smallest volumetric fraction, see Sect. 2.1). Therefore, we set $K_{si} = 10^{-6}$ m/s and $K_{sa} = 10^{-5}$ m/s (corresponding to intermediate values of conductivity for these geomaterials) and calibrate hydraulic conductivities of clay, K_c, and gravel, K_g, for each MC realization of geomaterial distribution on the basis of hydraulic head data, $\mathbf{h}^*$, available at $N_h = 20$ monitoring wells (see Sect. 2.1 and Fig. 2c).

Let $\mathbf{Y}_i^M = \left(Y_{c,i}^M, Y_{g,i}^M\right)$ be the vector collecting log-conductivities of clay, $Y_{c,i}^M = \ln K_{c,i}^M$, and gravel, $Y_{g,i}^M = \ln K_{g,i}^M$, of the ith realization (with $i = 1,\ldots n$) belonging to the ensemble $M = S, T$ (for the approach based on SISIM and T-PROGS, respectively). Maximum Likelihood (ML) estimates, $\hat{\mathbf{Y}}_i^M$, of $\mathbf{Y}_i^M$ are obtained by minimizing the Negative Log Likelihood criterion, NLL (Carrera and Neuman 1986). Minimization is performed through an iterative Levenberg–Marquardt algorithm embedded in PEST (Doherty 2002). In addition to yielding an estimate $\hat{\mathbf{Y}}_i^M$ of the (unknown) true vector $\mathbf{Y}_i^M$, the ML approach also provides (a Cramer-Rao lower bound approximation for) the covariance matrix, $\mathbf{Q}_i^M$, of the corresponding estimation errors.

We view the diverse MC realizations as a set of alternative/competing hydrogeological models of the investigated site. Model selection (or quality, information, discrimination) criteria have been proposed and compared to rank alternative models (e.g., Ye et al. 2008; Lu et al. 2011; Riva et al. 2011 and references therein). Among these, we focus on the KIC model identification criterion (Kashyap 1982). Upon dropping constant terms, the latter can be evaluated as

$$\mathrm{KIC}_i^M = N_h \ln \frac{J_i^M}{N_h} - N_P \ln(2\pi) - \ln\left|\mathbf{Q}_i^M\right| \tag{3}$$

where N_P is the number of model parameters (here, $N_P = 2$) and J_i^M is the sum of squared residuals for realization i of ensemble M, $J_i^M = \left(\mathbf{h}^* - \hat{\mathbf{h}}_i^M\right)^t \left(\mathbf{h}^* - \hat{\mathbf{h}}_i^M\right)$, where superscript t denotes transpose and $\hat{\mathbf{h}}_i^M$ is the vector of flow-model outputs (i.e., hydraulic head values evaluated at observation points) obtained using the ML-estimates $\hat{\mathbf{Y}}_i^M$. As compared against other model discrimination criteria, KIC has the unique advantage of enabling one to select a model by balancing (1) parsimony (via N_P), (2) model skill to reproduce observations (via J_i^M), and (3) (expected) information content per observation (via $\left|\mathbf{Q}_i^M\right|$) (Ye et al. 2008).

Making use of Eq. (3), one can (1) rank the alternative models and possibly select the one characterized by the minimum value of KIC_i^M as *best*, and/or (2) evaluate the posterior weight (or probability), $p(M_i|\mathbf{h}^*)$, of model i (within set M), as (see, e.g., Ye et al. 2004)

$$p(M_i|\mathbf{h}^*) = \frac{\exp\left(-\frac{1}{2}\left(\mathrm{KIC}_i^M - \mathrm{KIC}_{\min}^M\right)\right)p(M_i)}{\sum_{\ell=1}^{n}\left[\exp\left(-\frac{1}{2}\left(\mathrm{KIC}_\ell^M - \mathrm{KIC}_{\min}^M\right)\right)p(M_\ell)\right]} \tag{4}$$

where $\mathrm{KIC}_{\min}^M$ is the minimum value of KIC_i^M across all n models, and $p(M_i)$ is the prior probability of M_i. Here, we set $p(M_i) = 1/n$, as all realizations of set M are equally likely. Equation (4) allows evaluating the posterior estimate of hydraulic heads at observation wells as

$$\mathbf{h}_{\mathbf{POST}} = E\left[\hat{\mathbf{h}}^M\right] = \sum_{i=1}^{n} \hat{\mathbf{h}}_i^M p(M_i|\mathbf{h}^*). \tag{5}$$

In a similar way, it is also possible to evaluate the posterior probability $p(S_i, T_i|\mathbf{h}^*)$ of model i upon jointly considering both ensembles (i.e., T-PROGS and SISIM) while setting $\mathrm{KIC}^M_{\min}$ in Eq. (4) as the minimum between $\mathrm{KIC}^S_{\min}$ and $\mathrm{KIC}^T_{\min}$ and replacing n with $2n$.

3 Results and discussion

A preliminary analysis of the generated fields has been performed by evaluating the mean value, μ^M_k, and standard deviation, σ^M_k, (with $k = c, g, si, sa$) of facies volumetric proportions within each ensemble M (with $M = S, T$). The results of this analysis are listed in Table 1. Both ensembles render values of μ^M_k close to those inferred from the conditioning data, i.e., p_k. While T-PROGS realizations provide μ^T_k values that almost coincide with p_k, SISIM results slightly overestimate clay (percentage error $\sim$ 3%) and underestimate gravel (percentage error $\sim$ 5%) proportions. The SISIM ensemble is also characterized by a considerably larger variability (among the realizations, i.e., the *intra-ensemble variability*) of the facies volumetric proportions as compared to the T-PROGS counterpart, σ^T_k being two orders of magnitude smaller than σ^S_k.

Figure 3a, b depict facies distributions across a vertical cross-section in a sample realization generated with SISIM and T-PROGS, respectively. A qualitative comparison of these results reveals that the SISIM simulation is characterized by more fragmented lithological units than what can be observed within its T-PROGS counterpart. We assess quantitatively this aspect by computing the connectivity metrics introduced in Sect. 2.3 for both ensembles. We do so by focusing only on clay and gravel because (1) these two categories constitute about 80% of the aquifer system and (2) the connectivity of facies characterized by the largest/smallest conductivity values has been shown to play a major role in driving field-scale flow and transport processes (Wen and Gomez-Hernandez 1998 and references therein).

Figure 4a collects boxplots of $N^M_{C,k}$ values and highlights that both facies are considerably more fragmented (i.e., formed by a larger number of clusters) across the SISIM than the T-PROGS realizations. This feature is particularly evident for gravel, values of $N^S_{C,g}$ being almost one order of magnitude larger than those of $N^T_{C,g}$. Boxplots of Θ^M_k (rescaled by μ^M_k) are collected in Fig. 4b to provide a depiction of the relative size of the largest cluster with respect to Ω_k. Values of

Table 1 Facies proportions, p_k with $k = c, g, si, sa$, within the conditioning dataset. Mean, μ^M_k, and standard deviation, σ^M_k, of p_k evaluated over all MC realizations in the M ensemble ($M = S, T$) are also reported

Facies proportions	Clay p_c		Gravel p_g		Silt p_{si}		Sand p_{sa}	
Conditioning dataset	0.523		0.281		0.133		0.063	
Generated ensembles	μ^M_c	σ^M_c	μ^M_g	σ^M_g	μ^M_{si}	σ^M_{si}	μ^M_{sa}	σ^M_{sa}
SISIM, $M = S$	0.543	5×10^{-3}	0.267	6×10^{-3}	0.118	1×10^{-3}	0.071	5×10^{-3}
T-PROGS, $M = T$	0.523	2×10^{-5}	0.281	2×10^{-5}	0.133	1×10^{-5}	0.063	2×10^{-5}

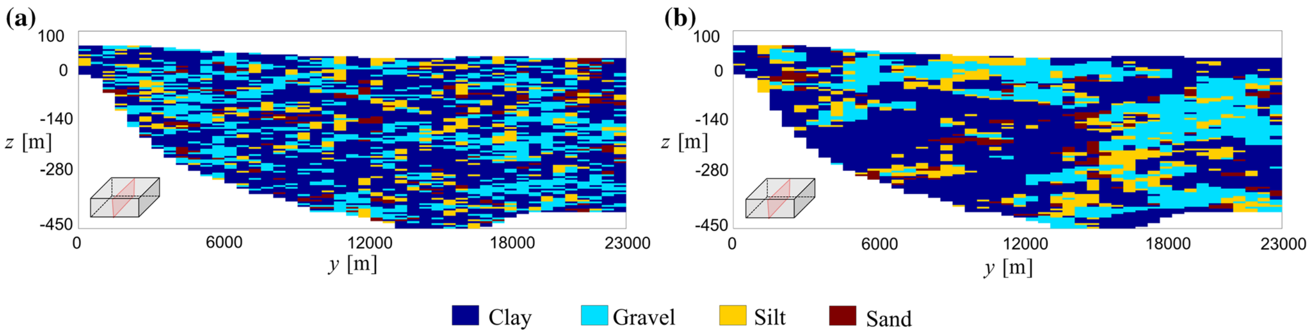
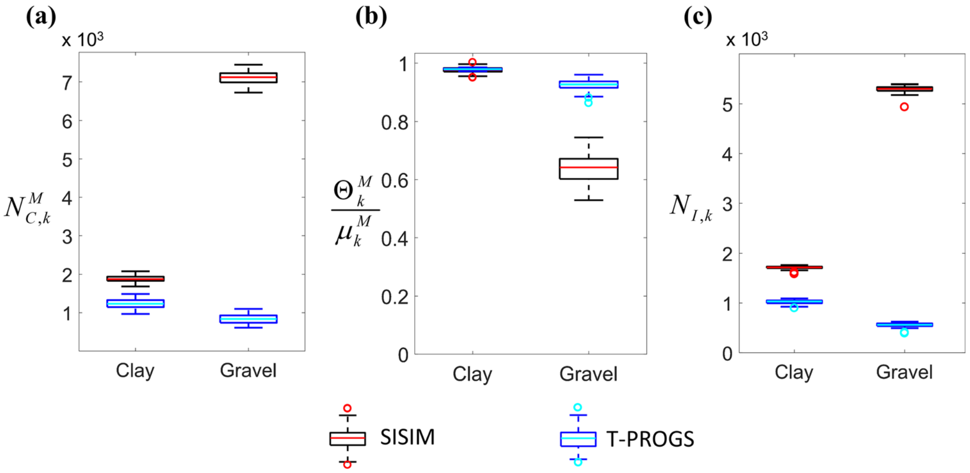


Fig. 3 Facies distribution along a vertical cross section in one realization generated with **a** SISIM and **b** T-PROGS

Fig. 4 Boxplots of the connectivity indices **a** $N^M_{C,k}$ **b** $\Theta^M_k \big/ \mu^M_k$ and **c** $N_{I,k}$ computed for clay and gravel units over all 100 realizations generated with SISIM (black and red symbols) and T-PROGS (blue and cyan symbols)



$$\Theta^M_c \big/ \mu^M_c = C^M_{max,c} \big/ \langle n^M_c \rangle$$

(brackets representing ensemble average) are close to one, suggesting that clay is essentially formed by a single cluster in all realizations for both generation methods. Otherwise, there is a non-negligible portion of cells associated with gravel that are not connected to the largest cluster (note that $\Theta^M_g \big/ \mu^M_g < 1$, in particular considering the SISIM set where $\Theta^S_g \big/ \mu^S_g$ ranges between 0.53 and 0.75). Moreover, Fig. 4c evidences that gravel can be found in a considerably larger number of isolated cells in the SISIM set, as compared to the T-PROGS ensemble. The SISIM set is characterized by a variability of all connectivity metrics that is slightly larger than that associated with the T-PROGS collection. This result is related to the different degree of variability of facies proportions across the realizations of the two ensembles, and resulting in $\sigma^S_k \gg \sigma^T_k$ (see Table 1).

In summary, all indicators inferred from the analysis of clusters document quantitatively that facies distributions generated upon relying on SISIM are characterized by a lower degree of connectivity than their T-PROGS counterparts. Values of $\Theta^M_k \big/ \mu^M_k$ for both ensembles indicate that clay is characterized by a higher degree of connectivity then gravel. These results are consistent with elements from percolation theory, as: (1) a standard variography analysis (see "Appendix") suggests that clay and gravel exhibit horizontal and vertical spatial correlation lengths which are larger for the T-PROGS than for the SISIM set; (2) clay has a high probability to form a unique percolating cluster, its volumetric fraction (p_c = 52.3%) being significantly larger than the (theoretical) percolation threshold (p_t = 31%, see Sect. 2.3). The spatial correlation of gravel contributes to yield a volumetric proportion (p_g = 28.1% $< p_t$) which is large enough for a percolating cluster to occur. However, consistent with the observation that p_g is close to (albeit smaller than) p_t, one can note a high variability of the connectivity of gravel from one realizations to the other (i.e., *intra-ensemble variability*). This issue has been further investigated by evaluating the connectivity function defined in Eq. (1). Figure 5 depicts $\tau^M_{k,\alpha}$ versus the (dimensionless) lag evaluated, within each realization (dotted curves), for clay and gravel along direction $\alpha = x,\ y,\ z$ and for both generated sets (i.e.,. $M = S,\ T$). Each plot in the figure also includes (1) the ensemble-averaged connectivity function (solid curves) and (2) the asymptotic sills (dashed horizontal lines) computed as $\overline{\tau^\infty_k} = \left\langle \left(C_{\max,k} \big/ n_k \right)^2 \right\rangle$, according to percolation theory. As expected, the connectivity function of clay is larger than the one of gravel regardless the generation approach. Considering clay, values attained by $\tau^S_{c,\alpha}$ and $\tau^T_{c,\alpha}$ along all directions are quite similar. The curves slightly deviate from the asymptotic value, $\overline{\tau^\infty_c}$, for large separation distances, an effect which is arguably related to the finite extent of the domain. On the other hand, values of $\tau^T_{g,\alpha}$ (Fig. 5b–d–f) for gravel are considerably larger than their $\tau^S_{g,\alpha}$ counterparts (Fig. 5a–c–e). This observation is consistent with the results depicted in Fig. 4 and provides additional evidence that gravel is less connected (more fragmented) in SISIM than in T-PROGS realizations. Figure 5 also suggests that gravel connectivity is characterized by a remarkably wider variability than what can be observed for clay for both generation methods, resulting also in a more pronounced deviation of the results from $\overline{\tau^\infty_g}$. We quantitatively assess this feature by evaluating the integral connectivity scale (Western et al. 2001), defined as $\Gamma_{\tau_{k,\alpha}} = \int_0^{N_\alpha \Delta\alpha} \tau_{k,\alpha}(s_\alpha) ds_\alpha$, for each realization of the two sets.

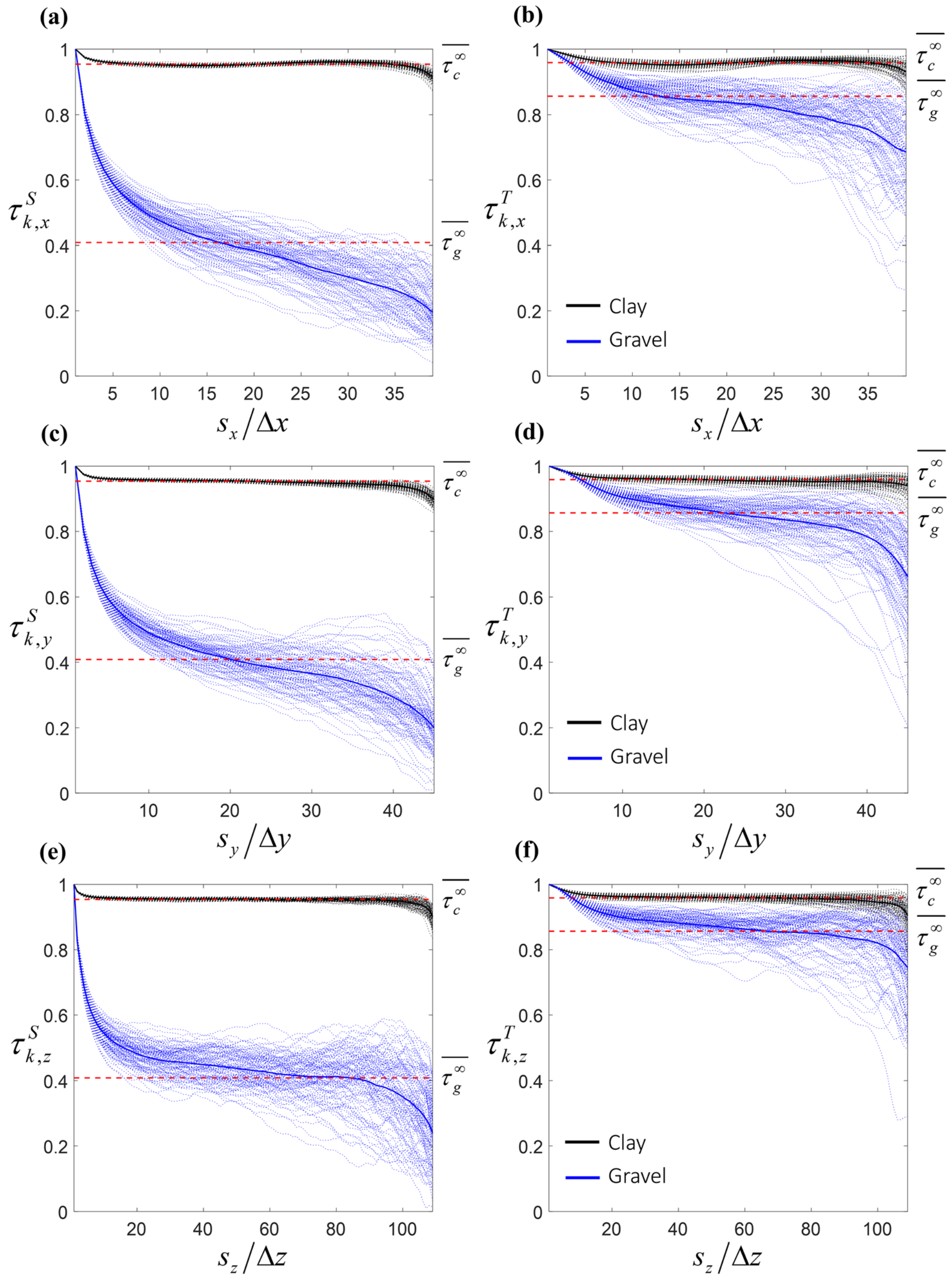


Fig. 5 Connectivity functions (dotted lines) computed in each single realization generated with SISIM (**a**–**c**–**e**) and T-PROGS (**b**–**d**–**f**) versus (dimensionless) separation distance along *x* (**a**–**b**), *y* (**c**–**d**) and *z* (**e**–**f**) axes, evaluated for clay and gravel. Ensemble-averaged connectivity functions (solid curves) and (mean) asymptotic values predicted by percolation theory (red dashed lines) are also reported

The variability of $\Gamma_{\tau_{k,\alpha}}$ within each ensemble is quantified by its coefficient of variation, $CV\left(\Gamma_{\tau_{k,\alpha}}\right)$. For both ensembles and both facies analyzed, we find that $CV\left(\Gamma_{\tau_{k,\alpha}}\right)$ is (almost) independent of direction (i.e., isotropic) and equal to 0.4% (SISIM) or 0.7% (T-PROGS) for clay and 10.5% (SISIM) or 4.5% (T-PROGS) for gravel.

Figure 6 collects estimates of clay, $\hat{K}^M_{c,i} = \exp\left(\hat{Y}^M_{c,i}\right)$, and gravel, $\hat{K}^M_{g,i} = \exp\left(\hat{Y}^M_{g,i}\right)$ conductivities obtained in each realization *i* of ensemble *M*, where $\hat{Y}_{c,i}$ and $\hat{Y}_{g,i}$ have been evaluated through the ML approach described in Sect. 2.4. The 95% confidence intervals (CIs) associated with each estimate (quantified by means of the diagonal terms of the covariance matrix **Q**) and the ensemble mean values, $\left\langle \hat{K}^M_c \right\rangle$ and $\left\langle \hat{K}^M_g \right\rangle$, are also depicted. Figure 6a suggests that the two ensembles are characterized by almost identical mean values of clay conductivity estimates ($\sim 4.6 \times 10^{-8}$ m/s) even as SISIM estimates are generally characterized by larger estimation errors (i.e., wider 95% CIs) than their T-PROGS counterparts. Considering gravel (Fig. 6b), $\hat{K}^S_{g,i}$ is (generally) larger than $\hat{K}^T_{g,i}$, mean values

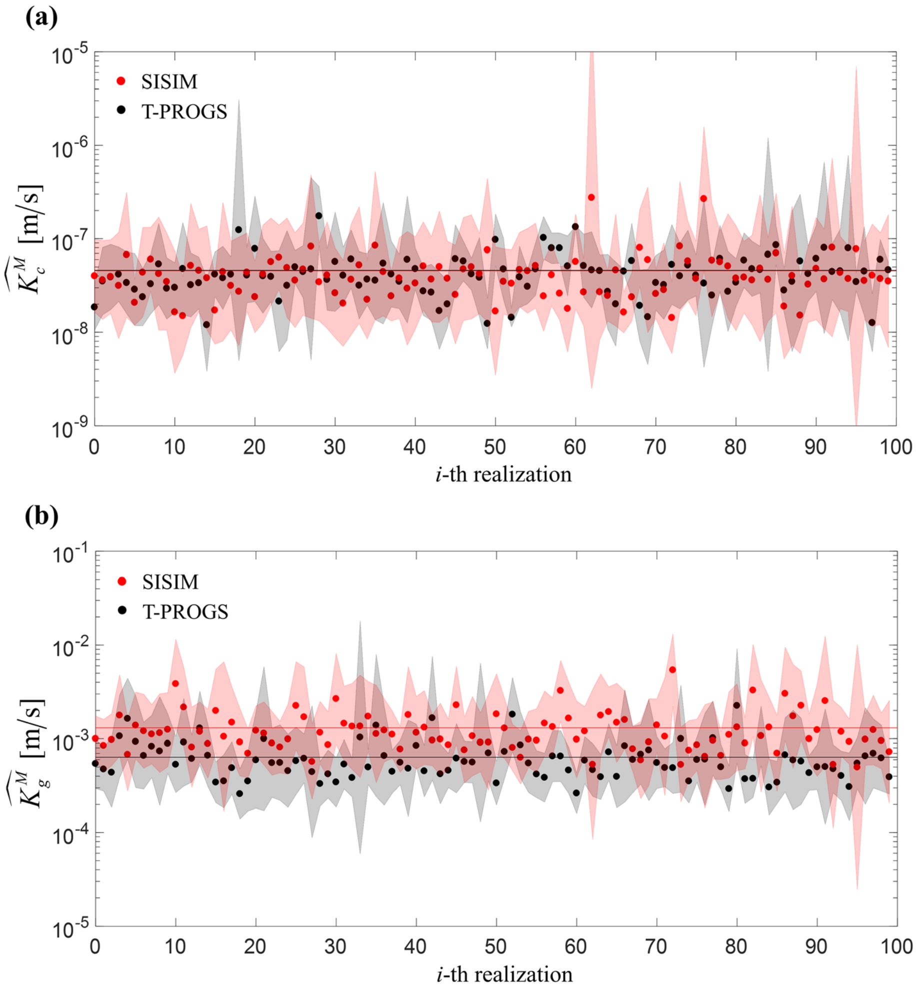


Fig. 6 Clay (**a**) and gravel (**b**) hydraulic conductivity estimates obtained for all realizations of SISIM and T-PROGS sets; 95% confidence intervals (shaded zones) and mean values (solid lines) are also reported

being equal to 1.3×10^{-3} m/s and 6.4×10^{-4} m/s, respectively. Results embedded in Fig. 6 are strictly related to the degree of facies connectivity observed for the two ensembles (as clarified by Figs. 4 and 5). Clay is characterized by similar connectivity as well as by similar values of hydraulic conductivity estimates in the two ensembles. Gravel exhibits a considerably smaller degree of connectivity in SISIM than in T-PROGS realizations. Consistently, conductivity estimates, $\hat{K}_{g,i}^{S}$, are significantly larger than $\hat{K}_{g,i}^{T}$. Conductivity estimates exhibit a larger intra-ensemble variability in the SISIM than they do in the T-PROGS set. This aspect is particularly evident for clay, the coefficients of variation $CV\left(\hat{K}_c^S\right)$ and $CV\left(\hat{K}_c^T\right)$ being equal to 80% and 54%, respectively. The intra-ensemble variability of gravel estimates is smaller, values of CV for both ensembles being about 50%. We observe that the intra-ensemble variability of conductivity is not related to the intra-ensemble variability of connectivity: for both sets, we obtain $CV\left(\hat{K}_c^M\right) > CV\left(\hat{K}_g^M\right)$ while $CV\left(\Gamma_{\tau_{c,\alpha}}\right) < CV\left(\Gamma_{\tau_{g,\alpha}}\right)$. This latter result is further supported by inspection of Fig. 7, where estimates $\hat{K}_{k,i}^M$ are plotted versus the integral connectivity scale evaluated along the mean flow direction, $\Gamma_{\tau_{k,y}}$. Results of similar quality (not shown) have been obtained by plotting $\hat{K}_{k,i}^M$ versus $\Gamma_{\tau_{k,x}}$ and $\Gamma_{\tau_{k,z}}$. The effect of variability of connectivity between the two ensembles (inter-ensemble variability) can be inferred by comparing the pattern of the points associated with the SISIM and T-PROGS realizations in Fig. 7: when considering clay (Fig. 7a), the results of the analysis yield two clouds of points which are practically overlapped, whereas the results for gravel (Fig. 7b) related to SISIM are characterized by values of $\Gamma_{\tau_{g,y}}$ and $\left\langle \hat{K}_g^M \right\rangle$ which are respectively lower and higher than the corresponding T-PROGS results. When analyzing intra-ensemble effects, one can see that the results of Fig. 7 do not provide a clear indication of correlation between connectivity and hydraulic conductivity within each ensemble. Table 2 summarizes the calibration results by listing the values of J, NLL and KIC obtained in the best realizations (i.e., with $\mathrm{KIC}_i^M = \mathrm{KIC}_{\min}^M$) within each of the two ensembles. All criteria appear to

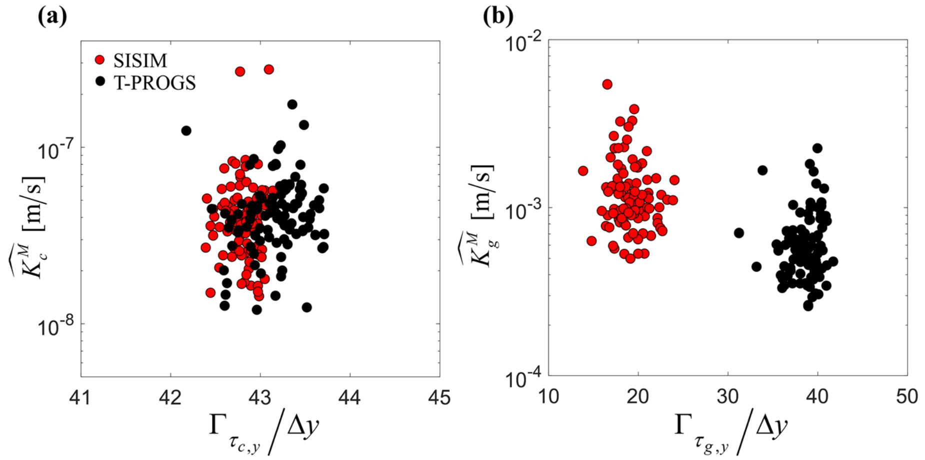


Fig. 7 Clay (**a**) and gravel (**b**) hydraulic conductivity estimates versus integral connectivity scales along the *y* axis obtained for SISIM and T-PROGS sets

Table 2 Values of *J*, NLL and KIC obtained in the best realization (i.e., with $\mathrm{KIC}_i^M = \mathrm{KIC}_{\min}^M$) in the two ensembles

Criterion	SISIM set	T-PROGS set
J (m^2)	346	307
NLL	122	118
KIC	125	123

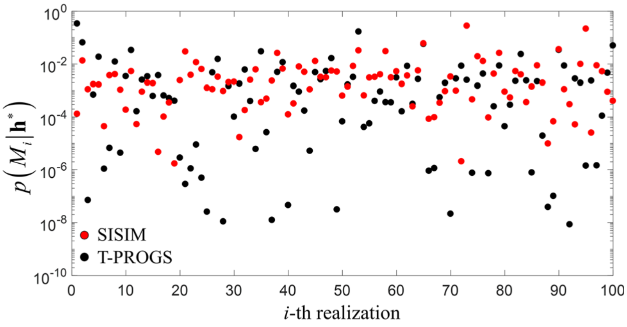


Fig. 8 Values of KIC-based posterior probabilities, $p(M_i|\mathbf{h}^*)$, evaluated within each ensemble of MC realizations

favor the best T-PROGS realization over its SISIM counterpart. Figure 8 illustrates the posterior probability, $p(M_i|\mathbf{h}^*)$, evaluated for both ensembles according to Eq. 4. It can be observed that the range of variability of $p(M_i|\mathbf{h}^*)$ is narrower for the SISIM than for the T-PROGS set, suggesting that the performance of diverse realizations associated with the former approach is more uniform than in the latter case. The best model in the T-PROGS set is associated with a larger posterior probability (34%) than its SISIM counterpart (28%). Figure 9a, b depict the hydraulic heads evaluated at all observation wells using the ML estimates of model parameters, $\hat{\mathbf{h}}_i^M$ (with $i = 1, \ldots n$), versus observed values, $\mathbf{h}^*$, for the SISIM and T-PROGS set, respectively. These figures also include (1) the posterior estimates, $\mathbf{h}_{\mathbf{POST}}$, evaluated according to Eq. (5), (2) the prior estimates, $\mathbf{h}_{\mathbf{PRIOR}}$, evaluated by replacing $p(M_i|\mathbf{h}^*)$ in Eq. (5) with the (uniform) prior probability $p(M_i)$, and (3) the sum of squared residuals computed with $\mathbf{h}_{\mathbf{POST}}$ and $\mathbf{h}_{\mathbf{PRIOR}}$ (here denoted as J_{POST}^M and J_{PRIOR}^M, respectively). Figure 9c provides a depiction of the results obtained by evaluating the posterior probability $p(S_i, T_i|\mathbf{h}^*)$ of model *i* upon jointly considering T-PROGS and SISIM collections of realizations (see Sect. 2.4). These analyses clearly document that (1) results obtained through ML-based posterior model probabilities are more accurate than those evaluated with their prior counterparts independent of the ensemble considered (note that $J_{POST}^M \ll J_{PRIOR}^M$); and (2) the T-PROGS set yields smaller values of *J* than SISIM. The model-averaged results (i.e., $\mathbf{h}_{\mathbf{POST}}$; see Eq. (5)) obtained by analyzing jointly the two ensembles (Fig. 9c) are very similar to those obtained upon relying solely on the T-PROGS set.

Figure 10 allows comparing (in terms of *J*) the predictive skill of $\mathbf{h}_{\mathbf{POST}}$ against results of single SISIM (Fig. 10a) and T-PROGS (Fig. 10b) realizations. Posterior model averaging provides slightly more/less accurate predictions than the best-performing single realization in the T-PROGS/SISIM set. This result is related to the observation that, as inferred from Fig. 8, the performance of diverse models in the SISIM ensemble is more uniform that what can be observed with reference to the T-PROGS set. As such, our results provide further documentation that when individual models provide results which are very similar, the performance of model-averaged results is worse than the one of the most skillful single model, whose effectiveness is essentially dampened in the averaging process (see also Winter and Nychka, 2010). The predictive capability in terms of flow-model outputs associated with the two aquifer reconstruction approaches is also tested through a validation procedure, as described in the following. Hydraulic head measurements at three

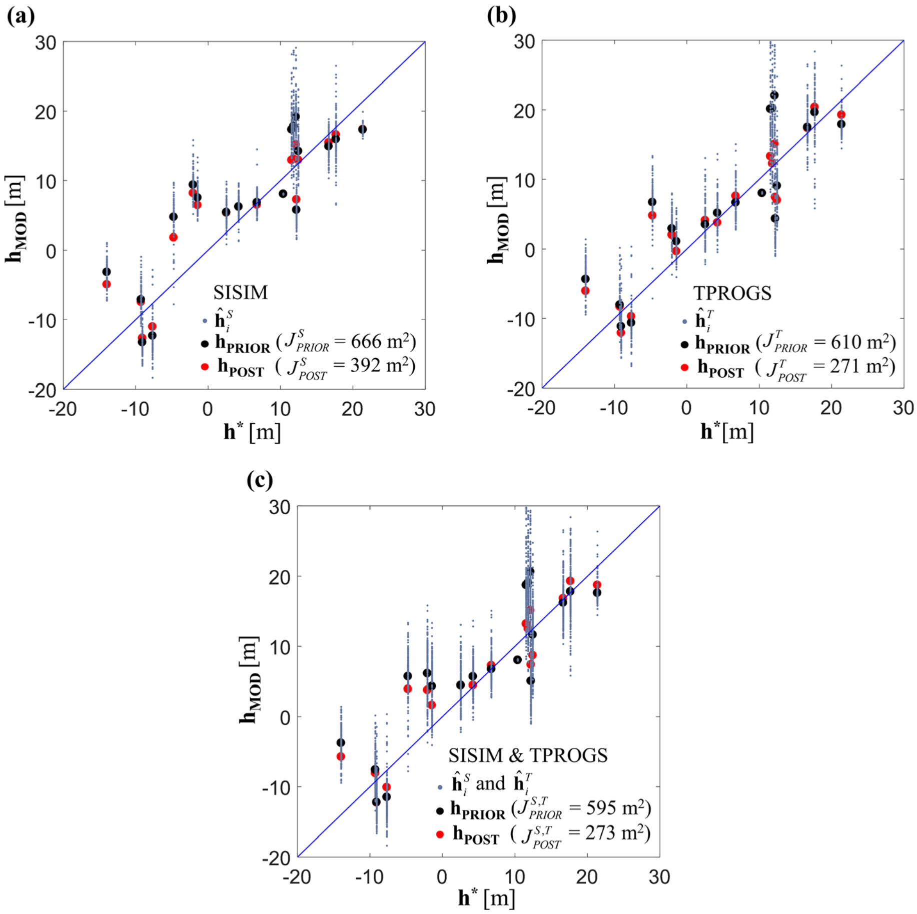


Fig. 9 Prior, $\mathbf{h_{PRIOR}}$, and posterior, $\mathbf{h_{POST}}$, hydraulic head estimates versus observed hydraulic heads. Results are obtained from **a** SISIM set, **b** T-PROGS set and **c** evaluating the posterior probability of model *i* upon jointly considering T-PROGS and SISIM (see Sect. 2.4). The estimates obtained in each realization are also reported

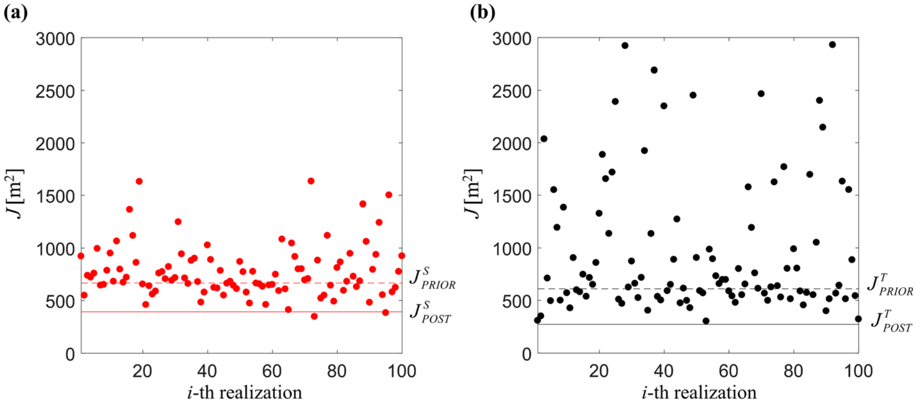


Fig. 10 Sum of squared residuals, *J*, obtained for single realizations of SISIM (**a**) and T-PROGS (**b**) ensembles. Values obtained for the averaged models with prior (dashed lines) and posterior (straight lines) weights are also reported

monitoring wells have been taken out (one at a time) from the calibration dataset to compare the local estimate, $\mathbf{h_{MOD}}$, resulting from the numerical model with the value ($\mathbf{h}^*$) observed at these locations. The selection of the monitoring wells to be used for validation is aimed at encompassing different regions of the aquifer. In this sense, we consider the data associated with locations which are close to the major well fields (see Fig. 2c) and corresponding to the shallowest (H1, with $z = -$ 21.1 m) and the deepest (H2, with $z = -$ 191.7 m) locations and one, farther from the well fields, at an intermediate depth (H3, with $z = -$ 104 m). We repeat the process of model calibration considering the resulting data-sets on a subset of 30 realizations from each ensemble. Figure 11 depicts, in the form of box plots, the results obtained at these three selected locations across all realizations of the selected SISIM and T-PROGS subsets. It can be observed that T-PROGS clearly outperforms SISIM for the prediction of head at the H1 monitoring well and slightly outperforms SISIM in the

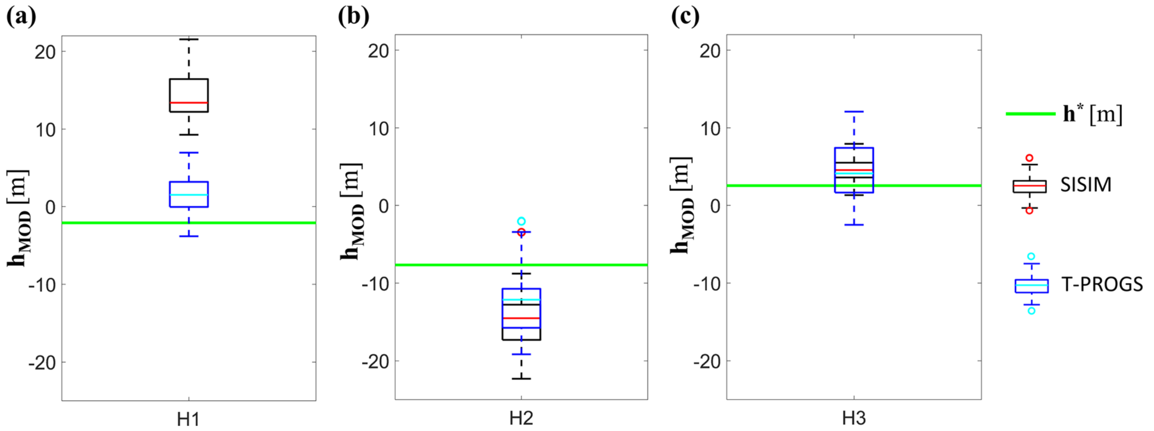


Fig. 11 Validation results: box plots of $\mathbf{h_{MOD}}$ obtained at **a** H1, **b** H2, **c** H3 (see Fig. 2c) for the SISIM (black) and T-PROGS (blue) subsets, the green line corresponding to the measured value, **h***

estimation of head at H2. The two methods provide results of similar quality in predicting heads at H3.

4 Conclusions

We explore the extent at which the geostatistical reconstruction method may affect the calibration of hydraulic parameters and the predictive capability of a groundwater flow model. For this purpose, we perform flow simulations on a domain patterned after the Bologna aquifer system by relying on two Monte Carlo ensembles of equally-likely realizations of facies distributions, respectively obtained on the basis of (1) indicator variograms (SISIM) and (2) transition probabilities (T-PROGS) inferred from lithological data. The latter reveal a high degree of heterogeneity in the aquifer, in which the least- (clay) and the most- (gravel) conductive facies represent almost 80% of the total aquifer. We take advantage of a Maximum Likelihood, ML, framework and of model discrimination criteria to (1) calibrate clay and gravel hydraulic conductivities in each realization of the two ensembles, on the basis of available piezometric data; (2) compute the posterior probability associated to each model and (3) evaluate the performance of ML-based model averaging approaches. Our results can be summarized as follows:

1. Clay, whose degree of connectivity is similar between the two ensembles, is also characterized by similar results in terms of calibrated conductivities.
2. Gravel exhibits an appreciably lower degree of connectivity in SISIM realizations compared to T-PROGS ones. As a consequence, hydraulic conductivity estimates of gravel obtained in the SISIM realizations are generally larger than their T-PROGS counterparts.
3. The degree of variability exhibited by connectivity indicators among realizations of the same ensemble seems not to be directly related with the variability of conductivity estimates.
4. The variability of the performance (in terms of sum of squared residuals) of alternative models in the SISIM set is smaller than in the T-PROGS set. This results in smaller variability of the posterior probability of models belonging to the SISIM ensemble respect to T-PROGS one.
5. The best individual model (as identified by model discrimination criteria) as well as the ML-based average model in the T-PROGS set are more skillful (in reproducing hydraulic head data) than their counterparts obtained for the SISIM set.

The findings of this study form the basis upon which one can obtained enhanced understanding of complex flow and transport dynamics. These elements are currently under study and will be the subject of future works.

5 Appendix

Correlation lengths of clay and gravel have been evaluated in SISIM and T-PROGS sets by means of a three-dimensional variographic analysis. Figure 12 collects ensemble directional indicator variograms, $\gamma_{k,\alpha}$ (with $\alpha = x, y, z$), of clay and gravel computed over all MC realizations in each set along the x ($\gamma_{k,x}$, Fig. 12a), y ($\gamma_{k,y}$, Fig. 12b), and z ($\gamma_{k,z}$, Fig. 12c) axes. All indicator variograms can be interpreted by exponential models, defined as

$$\gamma_{k,\alpha}(s_\alpha) = \sigma_k^2\left(1 - \exp\left(-\frac{3s_\alpha}{r_\alpha}\right)\right), \tag{6}$$

where σ_k^2 is the sill of facies k, which is very close to μ_k^M (1-μ_k^M), being μ_k^M the mean proportion of facies k in ensemble

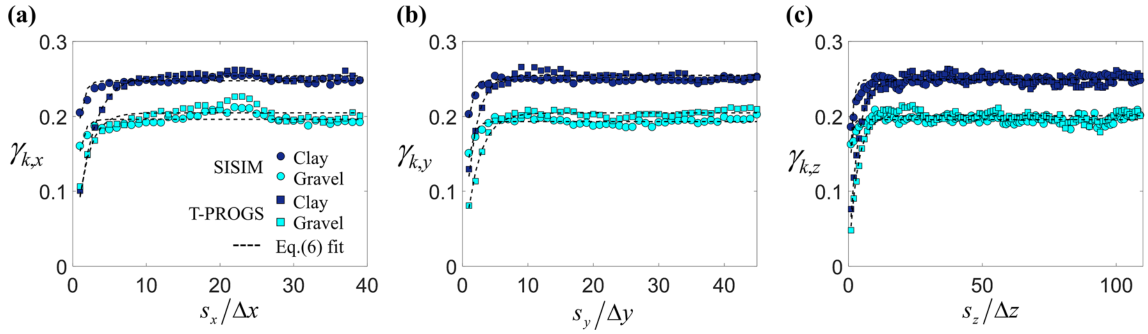


Fig. 12 Ensemble indicator variograms, $\gamma_{k,\alpha}$, of clay and gravel versus (dimensionless) separation distance along **a** x, **b** y, and **c** z axes computed over all MC realizations in each set. Interpreted models evaluated according to a ML-fit of Eq. 6 are also depicted (dashed lines)

Table 3 ML-estimates of horizontal and vertical ranges

Range estimates	Clay			Gravel		
	$r_x/\Delta x$	$r_y/\Delta y$	$r_z/\Delta z$	$r_x/\Delta x$	$r_y/\Delta y$	$r_z/\Delta z$
SISIM	1.9	1.9	2.9	2.0	2.2	2.0
T-PROGS	6.7	4.7	9.8	4.9	6.3	9.8

M (see Table 1); s_α and r_α respectively are lag and range along direction α. Results in Fig. 12 and in Table 3 highlight that clay and gravel exhibit horizontal and vertical correlation lengths which are larger for the T-PROGS than for the SISIM set.

Acknowledgements The authors would like to thank the EU and MIUR for funding, in the frame of the collaborative international consortium (WE-NEED) financed under the ERA-NET WaterWorks 2014 Cofunded Call. This ERA-NET is an integral part of the 2015 Joint Activities developed by the Water Challenges for a Changing World Joint Programme Initiative (Water JPI).

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

References

Bakshevskaia VA, Pozdniakov SP (2016) Simulation of hydraulic heterogeneity and upscaling permeability and dispersivity in sandy-clay formations. Math Geosci 48(1):45–64. https://doi.org/10.1007/s11004-015-9590-1

Bianchi M, Kearsey T, Kingdon A (2015) Integrating deterministic lithostratigraphic models in stochastic realizations of subsurface heterogeneity. Impact on predictions of lithology, hydraulic heads and groundwater fluxes. J Hydrol 531:557–573. https://doi.org/10.1016/j.jhydrol.2015.10.072

Carle SF (1997) Implementation schemes for avoiding artifact discontinuities in simulated annealing. Math Geol 29(2):231–244. https://doi.org/10.1007/BF02769630

Carle SF, Fogg GE (1996) Transition probability-based indicator geostatistics. Math Geol 28(4):453–477. https://doi.org/10.1007/BF02083656

Carle SF, Fogg GE (1997) Modelling spatial variability with one and multidimensional continuous-lag Markov chains. Math Geol 29(7):891–918. https://doi.org/10.1023/A:1022303706942

Carrera J, Neuman SP (1986) Estimation of aquifer parameters under transient and steady state conditions: maximum likelihood method incorporating prior information. Water Resour Res 22(2):199–210. https://doi.org/10.1029/WR022i002p00199

Cvetkovic V, Fiori A, Dagan G (2014) Solute transport in aquifers of arbitrary variability: a time-domain random walk formulation. Water Resour Res 50(7):5759–5773. https://doi.org/10.1002/2014WR015449

Dell'arciprete D, Bersezio R, Felletti F, Giudici M, Comunian A, Renard P (2012) Comparison of three geostatistical methods for hydrofacies simulation: a test on alluvial sediments. Hydrogeol J 20:299–311. https://doi.org/10.1007/s10040-011-0808-0

Della'rciprete D, Vassena C, Baratelli F, Giudici M, Bersezio R, Felletti F (2014) Connectivity and single/dual domain transport models: tests on a point-bar/channel aquifer analogue. Hydrogeol J 22(4):761–778. https://doi.org/10.1007/s10040-014-1105-5

Deutsch CV (2006) A sequential indicator simulation program for categorical variables with point and block data: BlockSIS. Comput Geosci 32(10):1669–1681. https://doi.org/10.1016/j.cageo.2006.03.005

Deutsch CV, Journel AG (1998) GSLIB, Geostatistical software library and user's guide. Oxford University Press, New York

Doherty J (2002) PEST: model independent parameter estimation, user manual, 4th edn. Watermark Numer. Computing, Corinda

Elfeki A, Dekking M (2001) A Markov chain model for subsurface characterization: theory and applications. Math Geol 33(5):569–589. https://doi.org/10.1023/A:1011044812133

Emery X (2004) Properties and limitations of sequential indicator simulation. Stoch Envir Res and Risk Ass 18:414–424. https://doi.org/10.1007/s00477-004-0213-5

Felletti F, Bersezio R, Giudici M (2006) Geostatistical simulation and numerical upscaling, to model ground-water flow in a sandygravel, braided river, aquifer analogue. J Sedim Res 76:1215–1229. https://doi.org/10.2110/jsr.2006.091

Gomez-Hernandez JJ, Wen XH (1998) To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology. Adv Water Res 21(1):47–61. https://doi.org/10.1016/S0309-1708(96)00031-0

Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, New York

Guadagnini A, Guadagnini L, Tartakovsky DM, Winter CL (2003) Random domain decomposition for flow in heterogeneous stratified aquifers. Stoch Environ Res Risk Assess 17:394–407. https://doi.org/10.1007/s00477-003-0157-1

Guadagnini L, Guadagnini A, Tartakovsky DM (2004) Probabilistic reconstruction of geologic facies. J Hydrol 294(1–3):57–67. https://doi.org/10.1016/j.jhydrol.2004.02.007

Harbaugh AW (2005) MODFLOW-2005, the US Geological Survey modular ground-water model: the ground-water flow process. US Department of the Interior, US Geological Survey, Reston, pp A6–A16

Hargreaves GH, Allen RG (2003) History and evaluation of Hargreaves evapotranspiration equation. J Irrig Drain Eng 129(1):53–63. https://doi.org/10.1061/(ASCE)0733-9437(2003)129:1(53)

Harter T (2005) Finite-size scaling analysis of percolation in three-dimensional correlated binary Markov chain random fields. Phys Rev E 72:026120. https://doi.org/10.1103/PhysRevE.72.026120

He Y, Hu K, Li B, Chen D, Suter HC, Huang Y (2009) Comparison of sequential indicator simulation and transition probability indicator simulation used to model clay content in microscale surface soil. Soil Sci 174(7):395–402. https://doi.org/10.1097/SS.0b013e3181aea77c

He X, Koch J, Sonnenborg TO, Jørgensen F, Schamper C, Refsgaard JC (2014) Transition probability-based stochastic geological modeling using airborne geophysical data and borehole data. Water Resour Res 50(4):3147–3169. https://doi.org/10.1002/2013WR014593

He X, Højberg AL, Jørgensen F, Refsgaard JC (2015) Assessing hydrological model predictive uncertainty using stochastically generated geological models. Hydrol Process 29(19):4293–4311. https://doi.org/10.1002/hyp.10488

Henri CV, Fernandez-Garcia D, de Barros FPJ (2015) Probabilistic human health risk assessment of degradation-related chemical mixtures in heterogeneous aquifers: risk statistics, hot spots, and preferential channels. Water Resour Res 51(6):4086–4108. https://doi.org/10.1002/2014WR016717

Hovadik JM, Larue DK (2007) Static characterizations of reservoirs: refining the concepts of connectivity and continuity. Petrol Geosci 13(3):195–211. https://doi.org/10.1144/1354-079305-697

Kashyap RL (1982) Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans Pattern Anal 4(2):99–104. https://doi.org/10.1109/tpami.1982.4767213

Kerrou J, Renard P, Franssen HJH, Lunati I (2008) Issues in characterizing heterogeneity and connectivity in non-multiGaussian media. Adv Water Resour 31(1):147–159. https://doi.org/10.1016/j.advwatres.2007.07.002

Kessler TC, Comunian A, Oriani F, Renard P, Nilsson B, Klint KE, Bjerg PL (2013) Modeling fine-scale geological heterogeneity—examples of sand lenses in tills. Groundwater 51(5):692–705. https://doi.org/10.1111/j.1745-6584.2012.01015.x

Koch J, He X, Jensen KH, Refsgaard JC (2014) Challenges in conditioning a stochastic geological model of a heterogeneous glacial aquifer to a comprehensive soft data set. Hydrol Earth Syst Sci 18(8):2907–2923. https://doi.org/10.5194/hess-18-2907-2014

Langousis A, Kaleris V, Kokosi A, Mamounakis G (2018) Markov based transition probability geostatistics in groundwater applications: assumptions and limitations. Stoch Environ Res Risk Assess 32:2129–2146. https://doi.org/10.1007/s00477-017-1504-y

Lee SY, Carle SF, Fogg GE (2007) Geological heterogeneity and comparison of two geostatistical models: sequential Gaussian and transition probability-based geostatistical simulation. Adv Water Resour 30(9):1914–1932. https://doi.org/10.1016/j.advwatres.2007.03.005

Li W (2007) A fixed-path Markov chain algorithm for 1 conditional simulation of discrete spatial variables. Math Geol 39(2):159–176. https://doi.org/10.1007/s11004-006-9071-7

Lu D, Ye M, Neuman SP (2011) Dependence of Bayesian model selection criteria and Fisher information matrix on sample size. Math Geol 43(8):971–993. https://doi.org/10.1007/s11004-011-9359-0

Lu D, Ye M, Curtis GP (2015) Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models. J Hydrol 529:1859–1873. https://doi.org/10.1016/j.jhydrol.2015.07.029

Maghrebi M, Jankoyic I, Weissmann GS, Matott LS, Allen-King RM, Rabideau AJ (2015) Contaminant tailing in highly heterogeneous porous formations: sensitivity on model selection and material properties. J Hydrol 531:149–160. https://doi.org/10.1016/j.jhydrol.2015.07.015

Marini M, Felletti F, Beretta GP, Terrenghi J (2019) Three Geostatistical methods for hydrofacies simulation ranked using a large borehole lithology dataset from the Venice Hinterland (NE Italy). Water 10(7):844. https://doi.org/10.3390/w10070844

Mohammadi HS, Mohammad JA, Faramarz DA (2020) CHDS: conflict handling in direct sampling for stochastic simulation of spatial variables. Stoch Environ Res Risk Assess 34(6):825–847. https://doi.org/10.1007/s00477-020-01801-4

Molinari A, Guadagnini L, Marcaccio M, Guadagnini A (2012) Natural back-ground levels and threshold values of chemical species in three large-scale groundwater bodies in Northern Italy. Sci Total Environ 425:9–19. https://doi.org/10.1016/j.scitotenv.2012.03.015

Neuman SP (2003) Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. Stoch Environ Res Risk Assess 17(5):291–305. https://doi.org/10.1007/s00477-003-0151-7

Park E, Elfeki AMM, Song Y, Kim K (2007) Generalized coupled Markov chain model for characterizing categorical variables in soil mapping. Soil Sci Soc Am J 71:909–917. https://doi.org/10.2136/sssaj2005.0386

Regione Emilia-Romagna (2010). Council Decree (Delibera di Giunta) n. 350 of 8/02/2010, Approval of the activities of the Emilia Romagna Region related to the implementation of Directive 2000/60/EC aiming at the design and adoption of the Management Plans of the hydrographic districs Padano, Appennino settentrionale and Appennino centrale http://ambiente.regione.emilia-romagna.it/acque/temi/piani%20di%20gestione Accessed 20 Mar 2019

Remy N, Boucher A, Wu J (2009) Applied geostatistics with SGeMS: a user's guide. Cambridge University Press, New York

Renard P, Allard D (2013) Connectivity metrics for subsurface flow and transport. Adv Water Resour 51:168–196. https://doi.org/10.1016/j.advwatres.2011.12.001

Ricci Lucchi F (1984) Flysh, molassa, clastic deposits: traditional and innovative approaches to the analysis of north Apennine basins (Flysch, molassa, cunei clastici: tradizione e nuovi approcci nell'analisi dei bacini orogenici dell'Appennino settentrionale). Cento Anni di Geologia Italiana, Volume Giubilare 1 centenario Soc. Geol. Ital 279–295

Riva M, Panzeri M, Guadagnini A, Neuman SP (2011) Role of model selection criteria in geostatistical inverse estimation of statistical data- and model-parameters. Water Resour Res 47:W07502. https://doi.org/10.1029/2011WR010480

Samani S, Moghaddam AA, Ye M (2018) Investigating the effect of complexity on groundwater flow modeling uncertainty. Stoch Environ Res Risk Assess 32(3):643–659. https://doi.org/10.1007/s00477-017-1436-6

Sharifzadehlari M, Fathianpour N, Renard P, Amirfattahi R (2018) Random partitioning and adaptive filters for multiple-point stochastic simulation. Stoch Environ Res Risk Assess 32(5):1375–1396. https://doi.org/10.1007/s00477-017-1453-5

Short M, Highdon D, Guadagnini L, Guadagnini A, Tartakovsky DM (2010) Predicting vertical connectivity within an aquifer system. Bayesian Anal 5(3):557–582. https://doi.org/10.1214/10-BA522

Siirila-Woodburn ER, Maxwell RM (2015) A heterogeneity model comparison of highly resolved statistically anisotropic aquifers. Adv Water Resour 75:53–66. https://doi.org/10.1016/j.advwatres.2014.10.011

Stauffer D, Aharony A (1992) Introduction to percolation theory, 2nd edn. Taylor & Francis, London

Vassena C, Cattaneo L, Giudici M (2010) Assessment of the role of facies heterogeneity at the fine scale by numerical transport experiments and connectivity indicators. Hydrogeol J 18(3):651–668. https://doi.org/10.1007/s10040-009-0523-2

Weissmann GS, Carle SF, Fogg GE (1999) Three-dimensional hydrofacies modeling based on soil surveys and transition probability geostatistics. Water Resour Res 35(6):1761–1770. https://doi.org/10.1029/1999WR900048

Wen XH, Gomez-Hernandez JJ (1998) Numerical modeling of macrodispersion in heterogeneous media: a comparison of multi-Gaussian and non-multi-Gaussian models. J Contam Hydrol 30(1–2):129–156. https://doi.org/10.1016/S0169-7722(97)00035-1

Western A, Bloschl G, Grayson RB (2001) Toward capturing hydrologically significant connectivity in spatial patterns. Water Resour Res 37:83–97. https://doi.org/10.1029/2000WR900241

Winter CL, Nychka D (2010) Forecasting skill of model averages. Stoch Environ Res Risk Assess 24:633–638. https://doi.org/10.1007/s00477-009-0350-y

Winter CL, Guadagnini A, Nychka D, Tartakovsky DM (2006) Multivariate sensitivity analysis of saturated flow through simulated highly heterogeneous groundwater aquifers. J Comput Phys 217:166–175. https://doi.org/10.1016/j.jcp.2006.01.047

Ye M, Neuman SP, Meyer PD (2004) Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour Res 40(5):W05113. https://doi.org/10.1029/2003WR002557

Ye M, Meyer PD, Neuman SP (2008) On model selection criteria in multimodel analysis. Water Resour Res 44:W03428. https://doi.org/10.1029/2008WR006803

Zinn B, Harvey CF (2003) When good statistical models of aquifer heterogeneity go bad: a comparison of flow, dispersion, and mass transfer in connected and multivariate Gaussian hydraulic conductivity fields. Water Resour Res 39(3):1051. https://doi.org/10.1029/2001WR001146

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.