Advances in Water Resources 52 (2013) 275–291

Contents lists available at SciVerse ScienceDirect

Advances in Water Resources

journal homepage: www.elsevier.com/locate/advwatres

Distance-based kriging relying on proxy simulations for inverse conditioning

David Ginsbourger [a,*], Bastien Rosspopoff [b], Guillaume Pirot [c], Nicolas Durrande [d], Philippe Renard [c]

[a] Department of Mathematics and Statistics, University of Bern, Alpeneggstrasse 22, CH-3012 Bern, Switzerland
[b] Ecole Nationale Supérieure des Mines, FAYOL-EMSE, LSTI, F-42023 Saint-Etienne, France
[c] Centre of Hydrogeology and Geothermics, University of Neuchâtel, 11 Rue Emile Argand, CH-2000 Neuchâtel, Switzerland
[d] Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello, Sheffield S1 4DP, UK

ARTICLE INFO

Article history:
Received 16 May 2012
Received in revised form 26 November 2012
Accepted 29 November 2012
Available online 8 December 2012

Keywords:
Design of numerical experiments
Inverse problem
Proxy-based distances
Kriging
Metamodels
Optimization

ABSTRACT

Let us consider a large set of candidate parameter fields, such as hydraulic conductivity maps, on which we can run an accurate forward flow and transport simulation. We address the issue of rapidly identifying a subset of candidates whose response best match a reference response curve. In order to keep the number of calls to the accurate flow simulator computationally tractable, a recent distance-based approach relying on fast proxy simulations is revisited, and turned into a non-stationary kriging method where the covariance kernel is obtained by combining a classical kernel with the proxy. Once the accurate simulator has been run for an initial subset of parameter fields and a kriging metamodel has been inferred, the predictive distributions of misfits for the remaining parameter fields can be used as a guide to select candidate parameter fields in a sequential way. The proposed algorithm, *Proxy-based Kriging for Sequential Inversion* (ProKSI), relies on a variant of the *Expected Improvement*, a popular criterion for kriging-based global optimization. A statistical benchmark of ProKSI's performances illustrates the efficiency and the robustness of the approach when using different kinds of proxies.

1. Introduction

Inverse techniques are one of the corner stones of groundwater modeling. Their aim is to identify model structure and model parameter values from observed state variables. In practice, a wide range of approaches exist and have been compared extensively [1–6]. Often, the inverse problem is formulated in a least-square or maximum likelihood manner. A data misfit quantifies the difference between measured and calculated state variables. The aim is then to find a parameter field minimizing the misfit.

Less frequently in practice, the problem is solved in the Bayesian framework with the aim to recover an ensemble of representative samples (parameter fields) from the posterior probability distribution. This is particularly important when prior geological knowledge is available and can be expressed using geological models describing the parameter fields. Techniques such as multiple point statistics, object- or process-based geological simulations [7] are often used to express this prior knowledge, but then solving the inverse problem becomes very challenging because it is usually not possible to provide an explicit analytical expression of the posterior distribution. In such situations, one needs to rely on computational resources and statistical sampling techniques [8,4,9,10] such as Markov Chain Monte Carlo (MCMC) [11–17]. A practical difficulty in that approach is that evaluating the likelihood function, involving in itself a calculation of the misfit, is often computationally very demanding. This inhibits the user to let an MCMC procedure run for a sufficiently large number of iterations to enable convergence [18,15]. Similar computational issues arise in optimization problems related to groundwater management: if each evaluation of the objective function that has to be minimized requires a significant amount of computational resources, it becomes infeasible to reach the optimum in reasonable time.

To reduce the computational demand, one can use the concept of *metamodel* (or *response surface*). The response (e.g. the misfit) of the flow simulator is computed for a small set of inputs and can then be predicted by the metamodel for any other input. Various interpolation techniques can be employed such as radial basis functions, splines, or kriging [19–24]. An advantage of kriging is its ability to provide both a prediction of the possible response (kriging mean m) and a corresponding prediction uncertainty (kriging variance s^2). The prediction uncertainty drops to zero where the response has actually been computed with the simulator and increases when moving away from those input points. In the global optimization problem consisting in finding inputs minimizing the objective function, one can use m and s^2 to express a trade-off between the exploitation of the mean prediction (finding locations

* Corresponding author.
E-mail addresses: david.ginsbourger@stat.unibe.ch (D. Ginsbourger), bastien.rosspopoff@mines-saint-etienne.org (B. Rosspopoff), guillaume.pirot@unine.ch (G. Pirot), n.durrande@sheffield.ac.uk (N. Durrande), philippe.renard@unine.ch (P. Renard).

0309-1708/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.advwatres.2012.11.019

where m is low) and exploration of the design space (finding locations where the prediction is the most uncertain). This idea gave birth to the *Expected Improvement* (*EI*) criterion [25]: for every location within the input space, the kriging metamodel is used to derive a predictive distribution for the improvement that might be obtained by evaluating the objective function at that location. Here the term improvement refers to the difference between the best (i.e. the lowest) response observed so far and the response at the new location if this difference is positive, and 0 otherwise.

The input point with the highest *EI* is then chosen to run the numerical model again and update the metamodel. Such approaches based on kriging metamodels have been very successfully used for sequential design of computer experiments since the development of the *Efficient Global Optimization* algorithm [26] in the late 1990's. Several other criteria were later proposed for neighboring problems (see, e.g., [27]).

Another approach to reduce the computational demand is to use a concept of distance between parameter fields [28,29,9]. Several types of distances can be defined, but the important point is that the distance should be chosen such that it can be computed rapidly and help predicting if two parameter fields will lead to similar or different responses. For example, Suzuki et al. [30] used the Hausdorff distance to quantify the differences in the geometry of complex 3D models (having different fault systems, horizon geometries, etc.), coupled with the neighborhood algorithm [31] to search efficiently, within the prior ensemble, the models that match field observations of oil production. Scheidt and Caers [29] propose a general framework based on the concept of distance to quantify uncertainty. In their example, the problem consists in estimating oil recovery in a production well. The models all have the same geometry, but very different parameter fields (obtained using multiple-point statistics with different training images). The prior ensemble is large and the aim is to rapidly obtain a good estimation of the uncertainty on the forecast. For that purpose, Scheidt and Caers [29] define the square distance between two parameter fields as the integrated square difference between the responses computed for the two parameter fields with a fast streamline solver. The distances between every pair of parameter fields is computed and used as a base for mapping all the parameter fields in an abstract metric space in which it is possible to select a small number of them covering comprehensively the variability of the complete ensemble. Running the forward two-phase flow numerical simulator only on these selected geological models allows a fast and rather accurate uncertainty assessment. Going a step further, Caers et al. [32] use the same framework to formulate the inverse problem.

A promising direction for reducing the computational demand is the joint use of a pair of complex and simple models [33–37]. The distinction between the complex and simple models is not straightforward, but to remain general we can state that the complex model tends to account for all important and relevant physical processes as well as all the necessary geometrical complexity of the reservoir. On the opposite, the simple model neglects some aspects of this complexity with the aim of being much more computationally efficient. The simplification may be based on neglecting some physical processes, on reducing the problem dimension (2D instead of 3D), or on a coarse spatial or temporal resolution. In the remaining of this paper, we will use the terminology *accurate model* for the complex one, and *proxy* for the simple one. To use a combination of accurate and proxy models in practice, one needs to establish a link between the two. Several approaches can be devised. For example, Doherty and Christensen [36] identify some parameters of the proxy model by solving an inverse problem where the results of the accurate model have to be reproduced.

In this paper, we propose to link an accurate and a proxy model using a distance-based kriging metamodel. It allows to forecast outputs of the accurate model as it is done with traditional kriging metamodels. However, those methods are usually limited to parameter spaces of small dimensions. This makes their application for the identification of complete parameter fields impossible. The novelty of the proposed approach lies therefore in the way we define the covariance kernel at the core of the kriging metamodel. The concept is simple, we assume that the same parameter fields can be used as input data for the proxy and the accurate model. As suggested by Caers and his collaborators [28,29,32,9] we use a distance based on proxy responses, but we include that distance into the covariance kernel of the kriging equations. The consequence is a drastic reduction of the problem dimension, allowing to infer covariance parameters. Once the statistical relation between the proxy and the accurate model is established, it can be used to predict the misfit for any parameter field whose proxy response is known. It can also be updated when new runs of the accurate model become available. This general idea can be applied to a very wide range of problems.

One of the main aims of this paper is therefore to describe the concept of the distance-based kriging technique. We also illustrate how this technique can be used in a sequential algorithm aiming at quickly identifying a set of parameter fields whose responses computed with an accurate model match some reference data. In an inverse problem, the purpose is often not only be to find the global minimizer(s) but more to sample from a posterior distribution, and so we propose a variant of the *EI* criterion meant to spend more time exploring the possible various minima of the misfit function than *EI*. For illustration purpose, we consider a simple flow and solute transport problem. The geological heterogeneity is modeled using a multiple-point statistics technique [38] allowing to account for prior geological knowledge typical for a fluvio glacial environment. Numerous experiments with a randomization procedure are conducted to test the robustness of the method.

The paper is organized as follows. In Section 2 we give an overview of the sequential algorithm used to solve the inverse problem. In Section 3, we describe in detail the proposed kriging metamodel. The equations of ordinary kriging are recalled, with a focus on the role of the covariance kernel. The original kernel underlying our work is introduced, followed by a discussion on its interpretation and mathematical foundations. Some practical details follow on the estimation of covariance parameters. We end the presentation of the method in Section 4 by describing how the sequential search is driven. Sections 5 and 6 are dedicated to results and discussion. We first introduce a case study to illustrate the methodology. Then we present the results of a randomized experiment and statistically assess the method's performances based on a benchmark of 100 reference curves. We conclude and propose a few theoretical and practical perspectives in Section 7.

2. Overview of the sequential algorithm

The proposed sequential algorithm is named *Proxy-based kriging for Sequential Inversion* (ProKSI). Its aim is to identify rapidly, within a large ensemble of parameter fields, the ones whose responses computed with the accurate model fit a given reference curve. In practice, the algorithm consists in sequentially selecting among all the available parameter fields which one will be used as input for the accurate numerical model at the next iteration (Figs. 1 and 2). Before sketching the key phases of the algorithm, let us set a few notations.

Each candidate parameter field is denoted $\mathbf{x}_i \in E$ $(1 \leqslant i \leqslant N)$, where E is typically a space of dimension 10^4 to 10^6 when representing a discretization of the subsurface. In the following examples, $\mathbf{x}_i$ represents a categorical field obtained from multiple-point statistics simulation. But the proposed methodology is more general and can be applied without much modifications to models having various geometries or even based on different conceptual assumptions. The only requirement is that it is possible

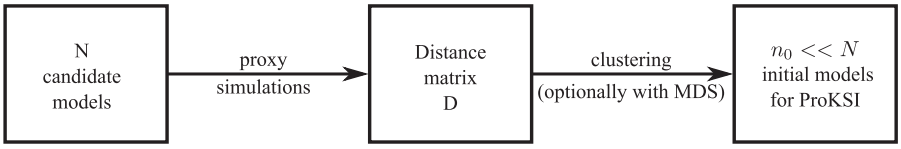

Fig. 1. Initialization steps of the ProKSI algorithm.

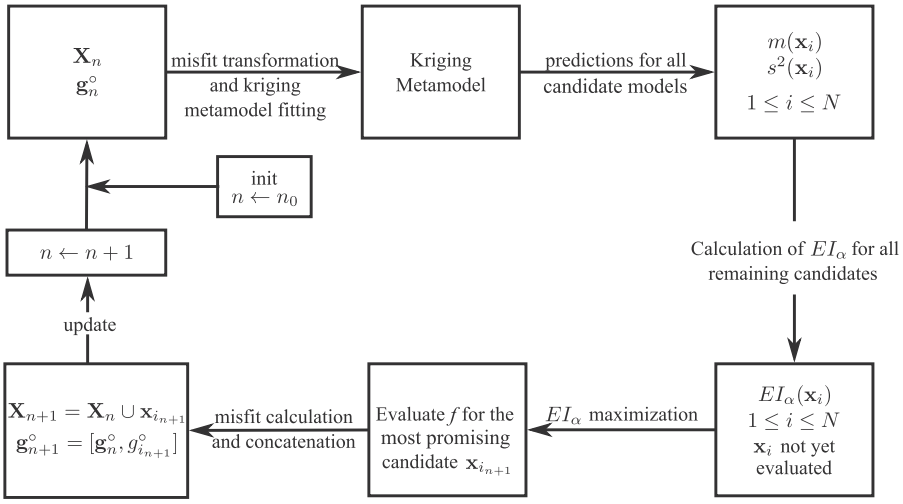

Fig. 2. Sequential loop of the ProKSI algorithm.

to compute the accurate and proxy responses for any of those input parameter fields.

The accurate numerical simulator is considered as a function f returning a vector of values. In the example, we assume more specifically that the simulator f returns for any parameter field $\mathbf{x} \in E$ a breakthrough curve $f_\mathbf{x}$ (concentration versus time):

$$f_\mathbf{x} : t \in [0,T] \rightarrow f_\mathbf{x}(t) \in [0,+\infty) \tag{1}$$

where t represents the time. The space of such curves is denoted by F.

Now, given a reference curve $f_{\text{ref}} \in F$, the goal is to recover in a limited time which $\mathbf{x}_i$'s $(1 \leqslant i \leqslant N)$ minimize the misfit $g^\circ(\mathbf{x}) := d(f_{\text{ref}}, f_\mathbf{x})$, where d is some metric on F. For example, if we use the L^2 norm, the misfit will be expressed as:

$$g^\circ(\mathbf{x}) = \int_0^T (f_{\text{ref}}(t) - f_\mathbf{x}(t))^2 dt \tag{2}$$

Ideally, one wishes to describe the subset of input fields leading to a good fit, relying on a fixed number of evaluations $k < N$ dictated by computation time constraints. In addition to f, we assume that a "*proxy*" $p : E \rightarrow F$ is available, providing an approximate solution to the flow and transport equations significantly faster than f. p may stem for instance from an auxiliary simulator solving similar equations with simplified physics, or from degrading the accurate simulator f by reducing the time or spatial resolution.

The ProKSI algorithm starts with a series of initialization steps (Fig. 1):

1. A sample of parameter fields $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is drawn from a chosen prior distribution (e.g., by multiple-points statistics simulation).
2. The proxy responses $p(\mathbf{x}_i, t)$ are computed for all $\mathbf{x}_i$'s $(1 \leqslant i \leqslant N)$. The distances $d_{i,j}$ between the proxy responses of any pair of parameter fields are then computed:

$$d_{i,j}^2 = \int_0^T \left(p(\mathbf{x}_i, t) - p(\mathbf{x}_j, t)\right)^2 dt \tag{3}$$

This allows assembling a distance matrix D between all proxy responses.

3. A clustering technique (k-means) is used to group the parameter fields in n_0 classes. For each class, the parameter fields that are the closest to the centroïd are selected to get a subset $\mathbf{X}_{n_0} = \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_{n_0}}\}$ of n_0 initial models (See Fig. 6). Multidimensional scaling (MDS) is optionally used to map all the input parameter fields in a small-dimensional Euclidean space (Fig. 6).

For each of those n_0 selected parameter fields, the accurate response $f_{\mathbf{x}_{i_j}}$ is computed with the accurate numerical solver. We obtain a vector $\mathbf{g}^\circ = \{g_{i_1}^\circ, \ldots, g_{i_{n_0}}^\circ\}$ (where $g_{i_j}^\circ := g^\circ(\mathbf{x}_{i_j}),\ 1 \leqslant j \leqslant n_0$) containing the misfits for the n_0 parameter fields.

The values of $\mathbf{g}^\circ$ are transformed to obtain a sample $\mathbf{g}$ with a close-to-Gaussian distribution. Such a Gaussian transformation is important later when the kriging variance is used to represent the prediction uncertainty for the misfit. Different techniques can be used for the transformation such as normal score transform, or Gaussian anamorphosis. Here, a power-law transform $g_{i_j} = \left[g_{i_j}^\circ\right]^\gamma$ (Box–Cox type) is used as it is simple to implement and robust even when a small number of samples is available, and the value of γ is obtained by minimizing the skewness of the sample of transformed values $\{g_{i_j},\ 1 \leqslant j \leqslant n_0\}$.

A sequential loop (Fig. 2) then allows selecting a new parameter field at each iteration on which to run the accurate solver. This enables building progressively a set of parameter fields with low misfit values. The steps in that loop are the following (n is first set to n_0):

1. If not already done, apply a normalizing transform to the sample of misfits (See detail above). Estimate the covariance parameters τ, θ, and σ^2 as described in Section 3. Compute the kriging mean $m(\mathbf{x}_i)$ and the variance $s^2(\mathbf{x}_i)$ for all inputs $\mathbf{x}_i \notin \mathbf{X}_n$.
2. After having computed the value of the modified expected improvement criterion $EI_\alpha(\mathbf{x}_i)$ (see Section 4 for its definition) for all the remaining candidate models, Select a model with maximal EI_α value as next candidate, called $\mathbf{x}_{i_{n+1}}$.

3. Set $\mathbf{X}_{n+1} = \mathbf{X}_n \cup \{\mathbf{x}_{i_{n+1}}\}$. Compute $f_{\mathbf{x}_{i_{n+1}}}$ with the accurate numerical solver. Calculate the new corresponding misfit and append it to the vector of misfits: $\mathbf{g}^{\circ}_{n+1} = \{\mathbf{g}^{\circ}_n, g^{\circ}_{i_{n+1}}\}$. Go to step 1 and resume the search until a convergence criterion is met.

The algorithm stops when the EI_α reaches a prescribed lower threshold, or a desired number of evaluations has been done, for instance because the allocated search time is elapsed.

3. High-dimensional kriging with a proxy-based kernel

The most important difference between the existing methods and what we propose here is the distance-based kriging approach. It lies at the heart of sequential algorithm described earlier in Fig. 2. In this section, we will describe in detail how this step is performed. The main idea is to integrate the distance between proxy responses within the covariance kernel of the kriging metamodel (Fig. 3).

3.1. kriging for Computer Experiments

We adopt the framework of Gaussian processes [23] to model the transformed misfit between f_{ref} and the response of the accurate numerical simulator. The transformed misfit g is assumed to be one realization of a Gaussian process with high-dimensional index space $(G_{\mathbf{x}})_{\mathbf{x}\in E}$, with mean function μ and non-stationary covariance kernel k. We assume that μ is an unknown constant, as in the case of ordinary kriging. We denote by $\mathbf{g}$ the vector of known misfit values at the current design of experiments $\mathbf{X}_n := \{\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}\}$ $(n \geqslant n_0)$. The kriging mean $m(\mathbf{x}) = \mathbb{E}\left[G_{\mathbf{x}} | G_{\mathbf{x}_{i_1}} = g(\mathbf{x}_{i_1}), \ldots, G_{\mathbf{x}_{i_n}} = g(\mathbf{x}_{i_n})\right]$ and kriging variance s^2 at any arbitrary point $\mathbf{x} \in E$ are written:

$$m(\mathbf{x}) = \widehat{\mu} + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbf{g} - \widehat{\mu}\mathbf{1}) \tag{4a}$$

$$s^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}) + \frac{(1 - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{1})^2}{\mathbf{1}^T K^{-1}\mathbf{1}} \tag{4b}$$

where K is a $n \times n$ matrix with entries $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$, referred to as the *covariance matrix of observations*, $\mathbf{k}(\mathbf{x}) := (k(\mathbf{x},\mathbf{x}_1), \ldots, k(\mathbf{x},\mathbf{x}_n))'$ is a $n \times 1$ *covariance vector*, and $\widehat{\mu} = \frac{\mathbf{1}^T K^{-1}\mathbf{g}}{\mathbf{1}^T K^{-1}\mathbf{1}}$ is the best linear unbiased estimate of μ.

One of the attracting features of kriging is that m interpolates the observations (i.e. $\forall j \in \{1,\ldots,n\},\ m(\mathbf{x}_{i_j}) = g(\mathbf{x}_{i_j})$). Furthermore, s^2 vanishes at the design points ($s^2(\mathbf{x}_{i_j}) = 0$), and gives a quantification of the prediction uncertainty at unobserved points. A very important feature is that both properties remain valid whatever the chosen covariance kernel k. Hence, Eqs. (4a) and (4b) give a potentially infinite set of interpolating metamodels, and selecting k appropriately for the studied phenomenon appears to be a crucial issue in practice.

3.2. A new kernel for high-dimensional kriging based on fast proxies

Designing a suitable covariance kernel over $E \times E$ is very challenging because E is a space of parameter fields of typical dimensions ranging between 10^4 and 10^6. Hence, taking kernels usually employed in d-dimensional ($d \approx 10$) cases, e.g., an anisotropic power exponential kernel, will a priori not make sense in the present framework. Alternatively, uncovering features of the parameter fields $\mathbf{x} \in E$ leading to similar response curves would be ideal.

Here, we take advantage of the proxy responses in order to define a relevant measure of similarity. More precisely, we propose to use a covariance kernel of the following form:

$$k(\mathbf{x},\mathbf{y}) := \sigma^2 \exp\left(-\frac{1}{\theta^2}\int_0^T (p(\mathbf{x},t) - p(\mathbf{y},t))^2 dt\right) + \tau^2 \mathbf{1}_{\mathbf{x}=\mathbf{y}} \tag{5}$$

In words, the closer two proxy curves associated with two parameter fields $\mathbf{x},\mathbf{y}$ are, the closer the fits to the reference are expected to be when running the accurate simulator with those inputs. In addition to this transformed Gaussian kernel, the term $\tau^2\mathbf{1}_{\mathbf{x}=\mathbf{y}}$ stands for the nugget effect, and allows to model a possible dissimilarity between the accurate responses of the inputs $\mathbf{x},\mathbf{y}$, even if their associated proxy responses are close or even identical.

In fact, the proposed covariance kernel k can be seen as a standard stationary Gaussian kernel over $F \times F$, chained with the "proxy operator", that is with the function p:

$$k(\mathbf{x},\mathbf{y}) := \sigma^2 \exp\left(-\frac{1}{\theta^2}||p(\mathbf{x}) - p(\mathbf{y})||_F^2\right) + \tau^2 \mathbf{1}_{\mathbf{x}=\mathbf{y}} \tag{6}$$

where $||f||_F := \sqrt{\int_0^T f(t)^2 dt}$ $(f \in F)$ stands for the L^2 norm over F (the functions of F being further assumed continuous). This basic fact ensures that the proposed kernel is an admissible covariance. k is indeed positive definite over $E \times E$ (but not necessarily strictly) in virtue of the following property, for which a proof is proposed in appendix:

Property. *Let E and F be two arbitrary spaces. Given a positive definite kernel k_F over $F \times F$, the kernel k_E defined by*

$$k_E(\mathbf{x},\mathbf{y}) := k_F(p(\mathbf{x}), p(\mathbf{y})) \tag{7}$$

is positive definite over $E \times E$ for any function $p : E \to F$.

Note that in different contexts, similar methods relying on a change of variables within a positive definite kernel were already proposed, for example in [39] and subsequent works. Coming back to Eq. (5), the basis kernel k_F corresponding to Prop. 7 is none other

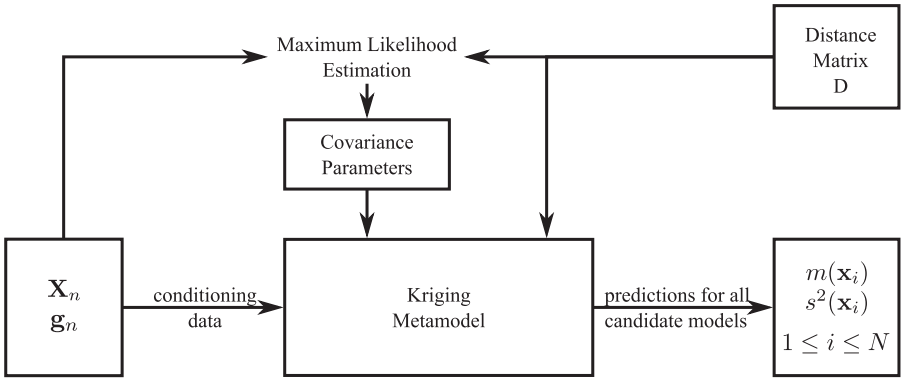

Fig. 3. Overview of the proxy-based kriging prediction workflow (after misfit transformation).

than an isotropic Gaussian kernel $k_F(\mathbf{u},\mathbf{v}) = \sigma^2 \exp\left(-\frac{1}{\theta^2}||\mathbf{u}-\mathbf{v}||_F^2\right)$, parametrized by a sill σ^2 and a range parameter $\theta > 0$. Note also that chaining the proxy operator with other kernels being positive definite in any dimension (e.g. kernels of the Matérn class) would be admissible too; a Matérn kernel may be preferred to a Gaussian one depending on the settings.

The next subsection focuses in detail on the chosen methodology for estimating the three parameters σ^2, θ, τ^2 from available data.

3.3. Parameter fitting for the proposed kriging model

Several methods can be considered for estimating σ^2, θ, and τ^2 based on available data. In particular, following the logic of Eq. (6), the problem boils down to a low-dimensional one thanks to the isotropy assumption relatively to the space F, and so usual variographic tools may well be applied in theory. However, most implemented variographic methods take low-dimensional vectors as inputs, while only the norm of the increments is available here. Alternative automatic estimation methods include cross-validation error minimization, and Maximum Likelihood Estimation (MLE).

Here we chose to base covariance parameter estimation on MLE, while keeping an eye on variographic tools for assisting MLE with initial values or bounds for the parameters. In ordinary kriging settings, MLE consists in maximizing the likelihood of σ^2, θ, τ^2 given $\mathbf{g}$ under the assumption that $G_{\mathbf{x_n}} \sim \mathcal{N}(\widehat{\mu}\mathbf{1}, K)$, or equivalently (See, e.g., [40]) in minimizing:

$$l(\sigma^2, \theta, \tau^2; \mathbf{g}) := \log(\det(K)) + (\mathbf{g} - \widehat{\mu}\mathbf{1})^T K^{-1} (\mathbf{g} - \widehat{\mu}\mathbf{1}) \tag{8}$$

where K and $\widehat{\mu}$ are functions of $(\sigma^2, \theta, \tau^2)$. Various global optimization procedures can be used for solving this non-convex optimization algorithm, ranging from Nelder and Mead's simplex to genetic algorithms using derivatives. Here we adopt a pragmatic one-at-a-time approach involving a concentration step on the variance parameter (described in appendix), for which convincing experimental results could be obtained as shown in Sections 5 and 6.

4. Sequential search driven by proxy-based kriging

For any candidate parameter field $\mathbf{x}_i$, the kriging metamodel of the previous section allows predicting the (transformed) misfit $g(\mathbf{x}_i)$ by $m(\mathbf{x}_i)$ with prediction variance $s^2(\mathbf{x}_i)$. Now, in a sequential procedure aiming at identifying the parameter fields with the lowest misfits such as considered here, m and s^2 can be used at any given iteration for selecting on which candidate parameter field to run the accurate numerical model next. For that purpose, we propose to use a variant of the *Expected Improvement* (*EI*) criterion, meant to spend more time exploring the basins of optima than the genuine *EI*.

By definition, *EI* is intended to point towards promising points, but also to foster space exploration. Hence, in *EI* algorithms like *EGO* [26], a typical behavior when evaluating the objective function at a good point (i.e. at a point becoming the current best) is to spend some additional iterations in its neighborhood, and then to get attracted by unexplored regions with higher kriging variances. This can be explained by coming back to *EI*'s formal definition. Let us denote by $g(\mathbf{X}_n)$ the vector of misfit values after n accurate evaluations of f, and by $\min(g(\mathbf{X}_n))$ the minimum misfit value found so far. The aim is now to find a parameter field $\mathbf{x}$ such that the magnitude of the improvement of $g(\mathbf{x})$ with respect to $\min(g(\mathbf{X}_n))$ be the highest in expectation. Let us remind the reader that the (transformed) misfit is modeled here as a Gaussian Process $(G_{\mathbf{x}})_{\mathbf{x}\in E}$. For any $\mathbf{x} \in E$, the difference between the current minimum and the unknown value of the misfit, $\min(G_{\mathbf{X}_n}) - G_{\mathbf{x}}$, is then a random variable. Only positive values are usually taken into account when one is not interested in regions with worse misfit, and the improvement is therefore defined as $(\min(G_{\mathbf{X}_n}) - G_{\mathbf{x}})^+ := \max(\min(G_{\mathbf{X}_n}) - G_{\mathbf{x}}, 0)$. The *EI* criterion for a candidate parameter field $\mathbf{x}$ is then defined as the expectation of this improvement conditional on $G_{\mathbf{X}_n} = g(\mathbf{X}_n)$:

$$EI(\mathbf{x}) := \mathbb{E}\left[(\min(G_{\mathbf{X}_n}) - G_{\mathbf{x}})^+ | G_{\mathbf{X}_n} = g(\mathbf{X}_n)\right] \tag{9}$$

where conditioning on the event $G_{\mathbf{X}_n} = g(\mathbf{X}_n)$ turns $\min(G_{\mathbf{X}_n})$ into $\min(g(\mathbf{X}_n))$, and leads to the well-known Gaussian conditional distribution for $G_{\mathbf{x}}$:

$$\mathcal{L}(G_{\mathbf{x}} | G_{\mathbf{X}_n} = g(\mathbf{X}_n)) = \mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x})) \tag{10}$$

Owing to this convenient property, the *EI* criterion offers the advantage of being analytically tractable (see [26]). Noting $T = \min(g(\mathbf{X}_n))$ and $f_{\mathcal{N}(m(\mathbf{x}),s^2(\mathbf{x}))}(\cdot)$ for the density of the $\mathcal{N}(m(\mathbf{x}), s^2(\mathbf{x}))$ distribution, we have:

$$\begin{aligned} EI(\mathbf{x}) &= \int_{-\infty}^{T} (T-u) f_{\mathcal{N}(m(\mathbf{x}),s^2(\mathbf{x}))}(u) du \\ &= (T - m(\mathbf{x}))\Phi\left(\frac{T - m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{T - m(\mathbf{x})}{s(\mathbf{x})}\right) \end{aligned} \tag{11}$$

where Φ and ϕ stand for the cumulative distribution function and the probability distribution function of the standard Gaussian distribution, respectively. Here we propose a variant of *EI* meant to put more emphasis on the exploration of basins of minimum while remaining tractable. Indeed, the aim in our motivating applications is not only to find the global minimizer(s) of g as quickly as possible, but also to find *a representative subset* of inputs leading to a response curve close to the reference, i.e. to a small misfit. The proposed trick to lower the repulsion effect of current best points is to replace $\min(g(\mathbf{X}_n))$ by a quantile of $g(\mathbf{X}_n)$ in the definition of *EI*. Calling α the level of this quantile, we define

$$EI_\alpha(\mathbf{x}) = (q_\alpha - m(\mathbf{x}))\Phi\left(\frac{q_\alpha - m(\mathbf{x})}{s(\mathbf{x})}\right) + s(\mathbf{x})\phi\left(\frac{q_\alpha - m(\mathbf{x})}{s(\mathbf{x})}\right) \tag{12}$$

where $q_\alpha = q_\alpha(\mathbf{X}_n)$ is the empirical α%-quantile of the sample of misfits $g(\mathbf{X}_n)$. Varying α allows tuning the criterion from normally explorative to very local. Indeed, when $\alpha = 0$, $q_{\alpha,n}$ coincides with the minimum of $g(\mathbf{X}_n)$, so that $EI_0 \equiv EI$. However, when tuning α to a strictly positive value (obviously smaller than 1), the tendency of *EI* to vanish near the observation points disappears. To prevent the algorithm from resampling at already explored points, we exclude them from the search. However, we are interested in points very close to the already explored points in terms of the proposed kernel, since they have similar proxy responses but may be very different in terms of parameter fields. Different values of α will be investigated in the application section, where the benefit of taking $\alpha > 0$ will be illustrated.

5. Illustration of the method through a case study

To illustrate the proposed approach, we consider a synthetic example. The input parameter fields $\mathbf{x}_i$ are categorical and describe a 2D vertical geological section (example Fig. 4(a)). The ensemble of fields $\mathbf{x}_i$ is generated using multiple-point statistics. The forward problem consists in computing a breakthrough curve $f_{\mathbf{x}} : t \in [0, T] \to f_{\mathbf{x}}(t) \in \mathbb{R}_+$ using an accurate numerical solver f. An example of a breakthrough curve is illustrated in Fig. 4(e).

The general aim, in this example is to test whether the proxy based kriging approach proposed in the previous section is able to help identifying efficiently a subset of fields $\mathbf{x}_j$ with relatively small misfit values $g^\circ(\mathbf{x}) := d(f_{\text{ref}}, f_{\mathbf{x}_j})$.

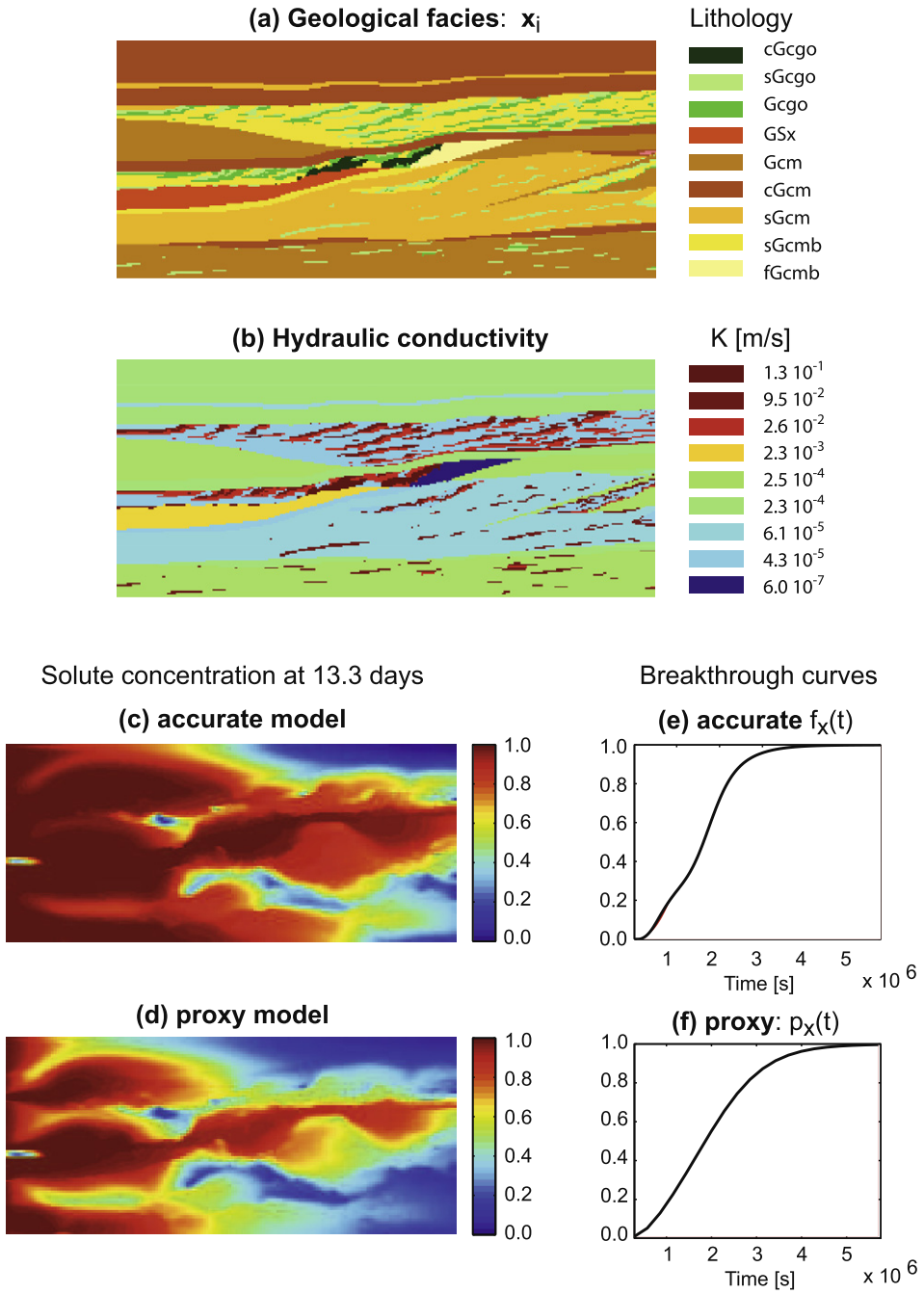

Fig. 4. Illustration of the main components of the illustrative example: (a) one simulation of a categorical field representing the geological facies. Each color corresponds to a type of sediment; (b) hydraulic conductivity field derived from the geological facies; (c,e) computation of the transient distribution of the concentration fields and breakthrough curve using an accurate numerical simulator; (d,f) approximation using a proxy simulator.

To test the reliability of the method, the procedure is repeated using different reference curves and configurations. The details of the generation of the geological models (i.e. the parameter fields), the forward numerical solver, and two proxy simulators are described in the following sections before presenting and discussing the results.

5.1. Geological facies simulations

The prior geological model for this case study is based on an aquifer analogue: the Herten site [41], typical of a glacio-fluvial environment, and on a multiple-point statistics geological model. Thousand realizations (the $\mathbf{x}_i$'s) of the geological medium are generated using the Direct Sampling (DS) multiple-point statistics method [38] using a geological section mapped from the Herten site as a training image. Fig. 4(a) shows one example. The grid has a size of 320 by 140 pixels covering an area of 16 m by 7 m. Each color corresponds to a geological facies. The parameters of the DS method are: a search neighborhood of 20 cells on each axis, a maximal number of neighboring nodes of 15, a distance threshold of 0.01, and a maximal scan fraction of 0.5. All the stochastic realizations are constrained by a secondary variable (describing the large scale sedimentary structures) in the training image and in the simulations, following the approach used by [42]. Fig. 5(a) displays 9 of those realizations. The variability between them is present only at small scales within the main sedimentary bodies.

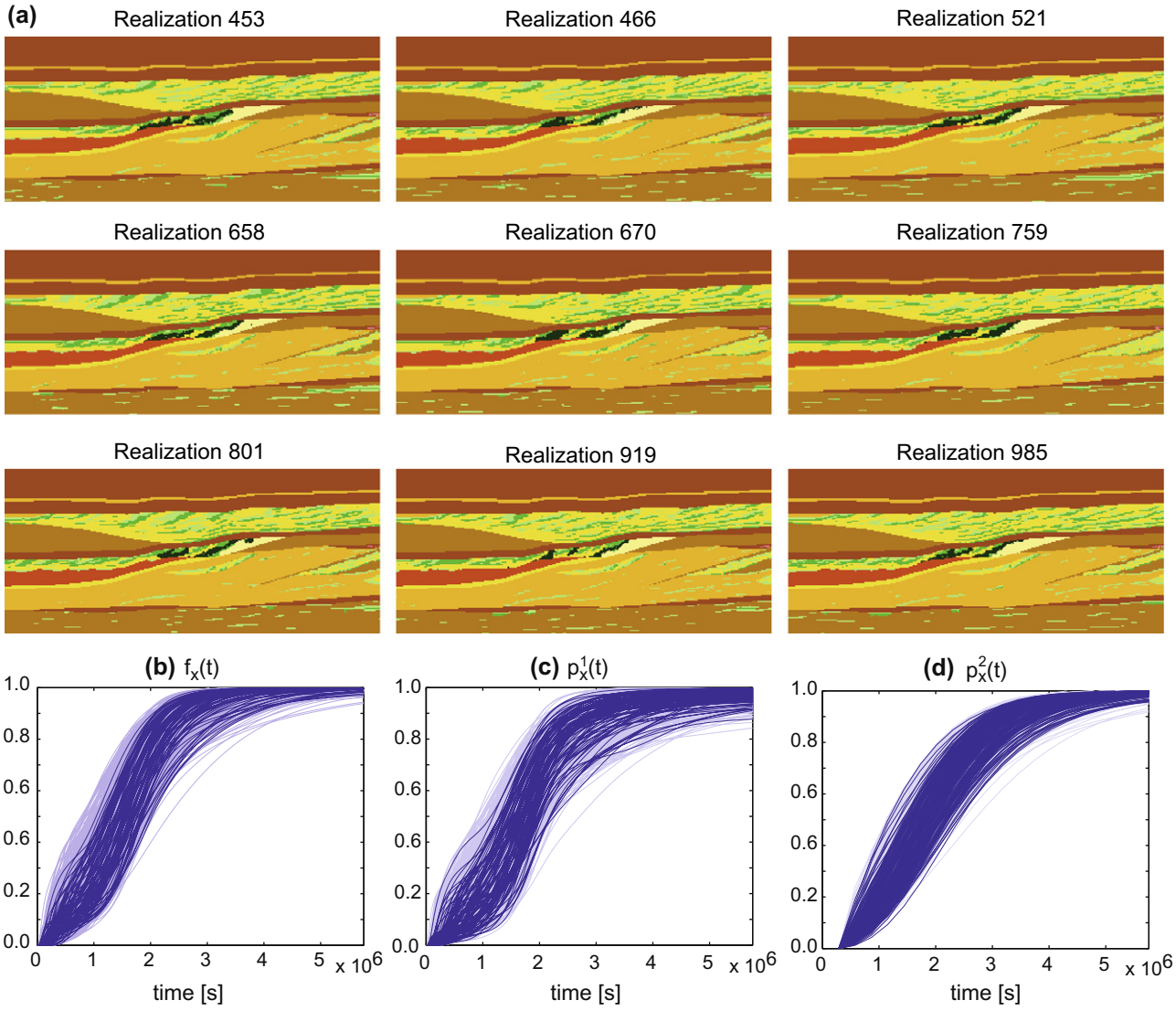

Fig. 5. (a) 9 realizations of the lithofacies. Because all the simulations are constrained by the large scale structure data, only the internal architecture within the main layers is displaying some variability between the simulations. (b) Ensemble of the breakthrough curves obtained with the accurate numerical model and the two proxies (c) and (d) for the 1000 parameter fields. To make the figure more readable, some breakthrough curves are represented in light gray color.

The large scale structures are identical in all realizations. The ensemble of those geological models constitutes a sample of our prior distribution on parameter fields.

5.2. Flow and transport simulations

The breakthrough curves are obtained by solving the classical advection–dispersion equation in transient state using a finite volume technique [43,44]. The spatial discretization is kept identical to the one used for the geological simulations. The boundary conditions and parameters are summarized in Table 1. A constant value of the hydraulic conductivity is assigned to each facies (Fig. 4(b)) according to the mean values obtained from laboratory experiments and described by Bayer et al. [41]. For the sake of simplicity, the porosity is considered homogeneous over all facies. A constant head is prescribed on the left (0.1 m) and right boundaries (0 m) and remain constant while the upper and lower boundaries are no flow boundaries. Those boundary conditions lead to a uniform steady-state flow from left to right.

The initial distribution of the solute concentration is set to zero everywhere in the domain. A fixed concentration of 1 is prescribed on the left boundary. The transport problem is solved in transient state. Fig. 4(c) shows the map of the solute concentration for the realization shown in Fig. 4(a) after 13.3 days of simulations. On the right boundary, the solute fluxes are integrated to compute the breakthrough curve $f_{\mathbf{x}}(t)$ representing the mean concentration at the outlet versus time (Fig. 4(e)).

Table 1
Parameter values for the solute transport model.

Parameter	Value
Porosity	0.35
Molecular diffusion	4.0×10^{-9} m/s
Longitudinal dispersivity (along x axis)	0.1 m
Transversely dispersivity (along z axis)	0.01 m
Total simulation time	1.44×10^7 s
Time steps length	1.44×10^4 s

Despite the apparent limited variability in the geological structure described above, a rather wide range of tracer breakthrough responses are obtained on the prior ensemble (Fig. 5(b)). This illustrates the importance of the internal heterogeneity of the high permeability features within the main sedimentary layers.

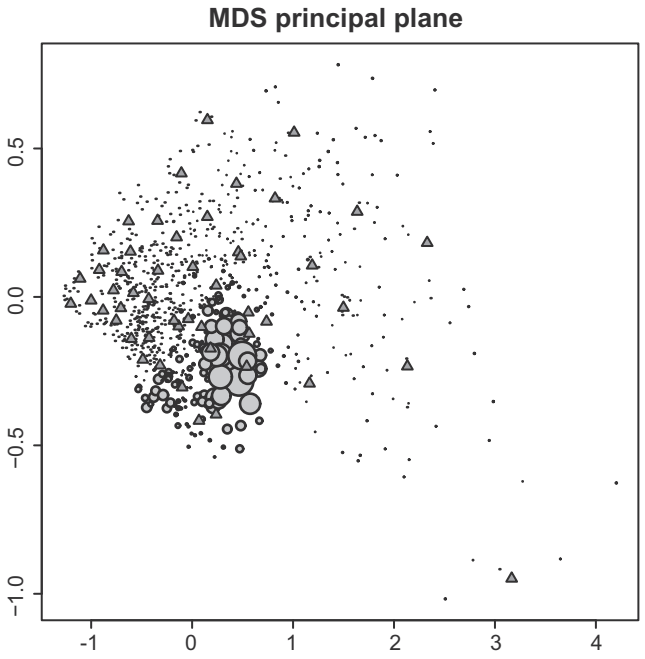

Fig. 6. Every point in the MDS space represents a parameter field. The triangles indicate the models that were selected by the *K*-means algorithm for the initial design of experiments, and the radius of the circles are proportional to the EI_z criterion.

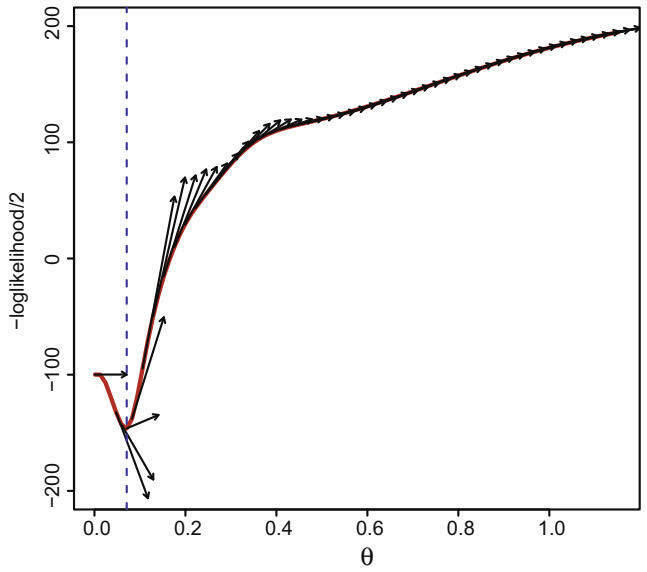

Fig. 8. Identification of θ by maximizing the concentrated log-likelihood function. The large curvature at the minimum indicates a well-identified parameter value. The arrows stand for the gradients of the log-likelihood function, which have been calculated analytically, and are used within the optimization procedure.

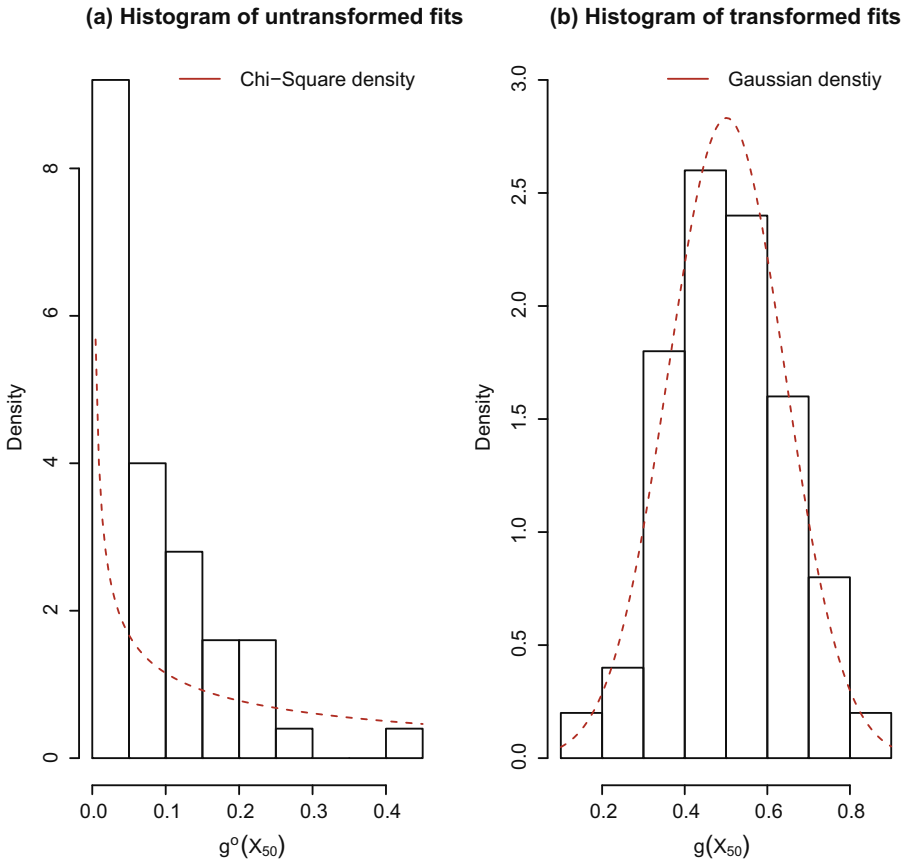

Fig. 7. Samples of untransformed (left) and transformed (right) misfit values obtained at a 50-point initial design of experiments in the case of a proxy with simplified physics. The histogram of the untransformed sample is closer to a chi-square distribution, whereas the one obtained by a power transformation, although remaining positive, is much more similar to the one of a Gaussian sample. The exponent used in the power transformation ($\gamma \approx 0.24$ here) is obtained by setting the skewness of the transformed sample to 0.

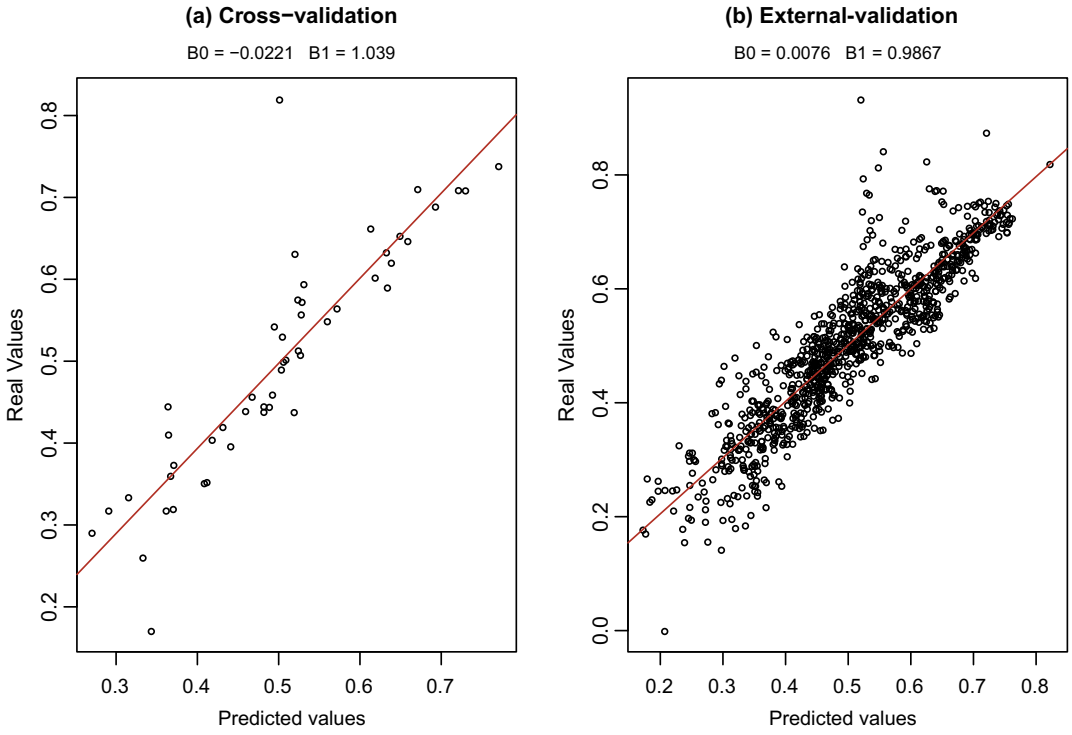

Fig. 9. (a) Cross validation and (b) External validation.

5.3. Two different proxies

A good proxy is faster than the accurate numerical model and allows to distinguish parameter fields that have similar or different responses in terms of tracer breakthrough. Such a proxy is generally not expected to provide an accurate simulation of the breakthrough or of solute concentration states. It should simply be a fast approximation allowing to discriminate parameter fields.

For this case study, we consider two different proxies and check their performances and reliability. The first one, $p_{\mathbf{x}}^1(t)$, is based on simplified physics. We use the same solver [43,44] and the same spatial and temporal resolution as for the accurate model based on the full physics, but we disregard diffusion and dispersion effects. The numerical simulation thereby only accounts for advection and numerical dispersion phenomena. The second proxy, $p_{\mathbf{x}}^2(t)$, is based on simply coarsening the time discretization of the accurate model. The number of time steps is reduced; their duration is increased to 2.88×10^5 s (i.e. a division by 20 of the number of time steps).

The breakthrough curves computed with the two proxies are displayed in Figs. 5(c) and (d). The first proxy gives breakthrough curves whose general shape resemble more the accurate model than the second proxy: some of the curves display a sigmoidal shape like the fine scale solution. The second proxy results in breakthrough curves that are more regular. For this proxy, the first arrivals of the tracer are almost identical for all geological models because of the coarse temporal resolution. The responses for $p_{\mathbf{x}}^2(t)$ present some variability, but less than $f_x(t)$ and the first proxy. For both proxies, the computational time is reduced by a factor of about 20. The accurate numerical solution takes about 7.5 min on a PC, while the two proxies run in about 20 s each.

5.4. Results

Let us now apply our kriging model to the problem of predicting the transformed misfit between the breakthrough curves of a given reference and the responses associated with the 1000 candidate geological media (i.e. parameter fields). The proxy used here is $p_x^1(t)$, the one with simplified physics. For now we arbitrarily choose one of the actual response curves (the realization with index 800) as a reference for illustration purposes. Note that more general results will be presented in Section 6, where statistics will be derived based on 100 randomly chosen reference curves.

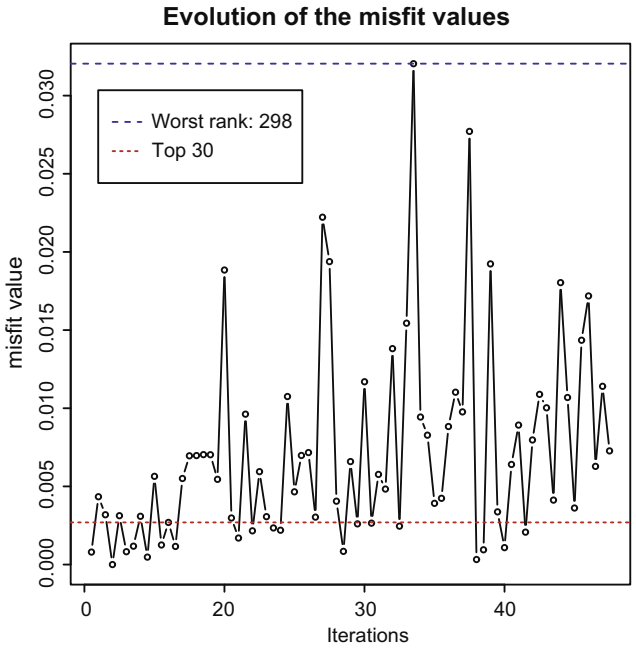

Fig. 10. Monitoring of the misfit values obtained for the parameter fields sequentially chosen by the ProKSI algorithm.

Among the 1000 considered inputs, 50 are chosen based on a clustering technique using proxy-induced distance (Fig. 6), in the flavor of Scheidt and Caers [29]'s approach. The actual response curves are calculated by using the accurate numerical model with the latter inputs, and the 50 corresponding values of misfit to the reference curve are calculated and stored in a vector, denoted by $g^{\circ}(X_{50})$ or $\mathbf{g}^{\circ}$, as in Section 3.

As shown on Fig. 7, a transformation is used to make the data misfits closer to Gaussian. For simplicity, we restrict the transformation to be a power transform, $g = (g^{\circ})^{\gamma}$. The ad hoc approach proposed here to determine the coefficient of this transform is to set the skewness of the transformed sample equal to zero. As will be presented in more detail in Section 6 (performance assessment), such transform significantly improves the predictivity of the kriging model, as well as the performances of the inversion algorithm proposed in the next section.

In a second step, we estimate the kernel parameters by maximum likelihood (MLE) based on the transformed sample of fits.

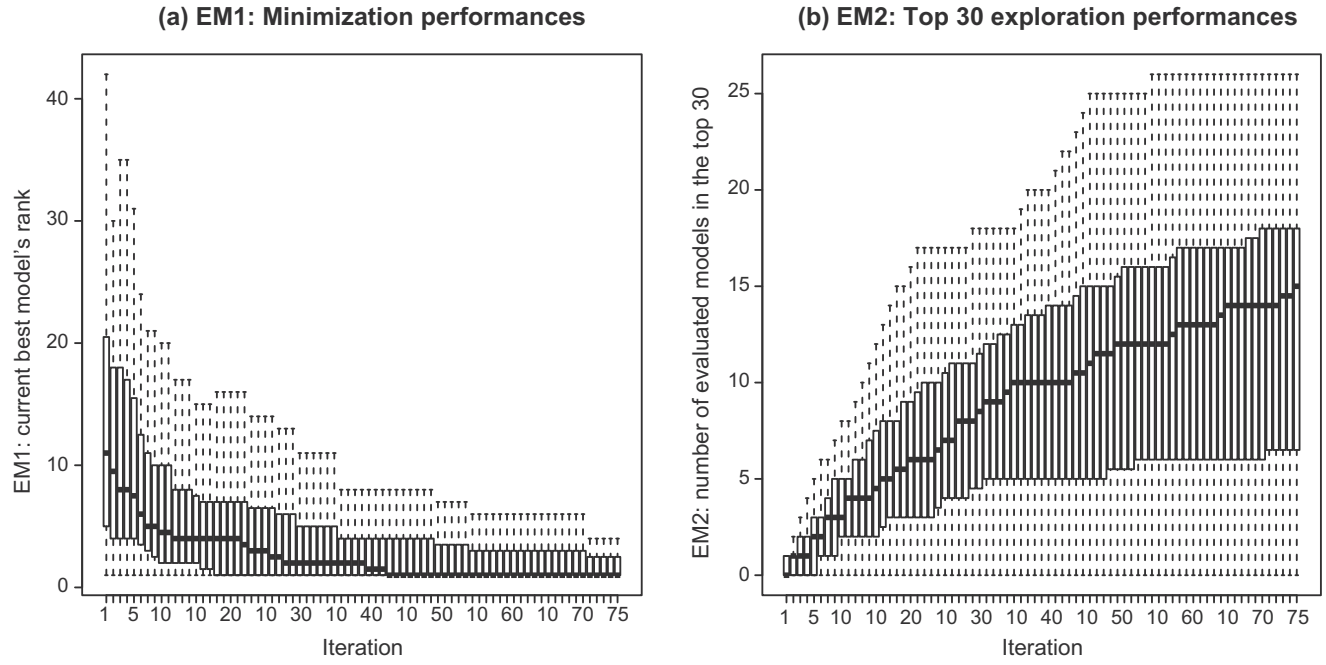

Fig. 11. Performances of the ProKSI algorithm (based on proxy 1) with a transformed misfit. (a) Box-plot of the EM1 criterion over the 100 numerical experiment replicates. (b) Box-plot of the EM2 criterion.

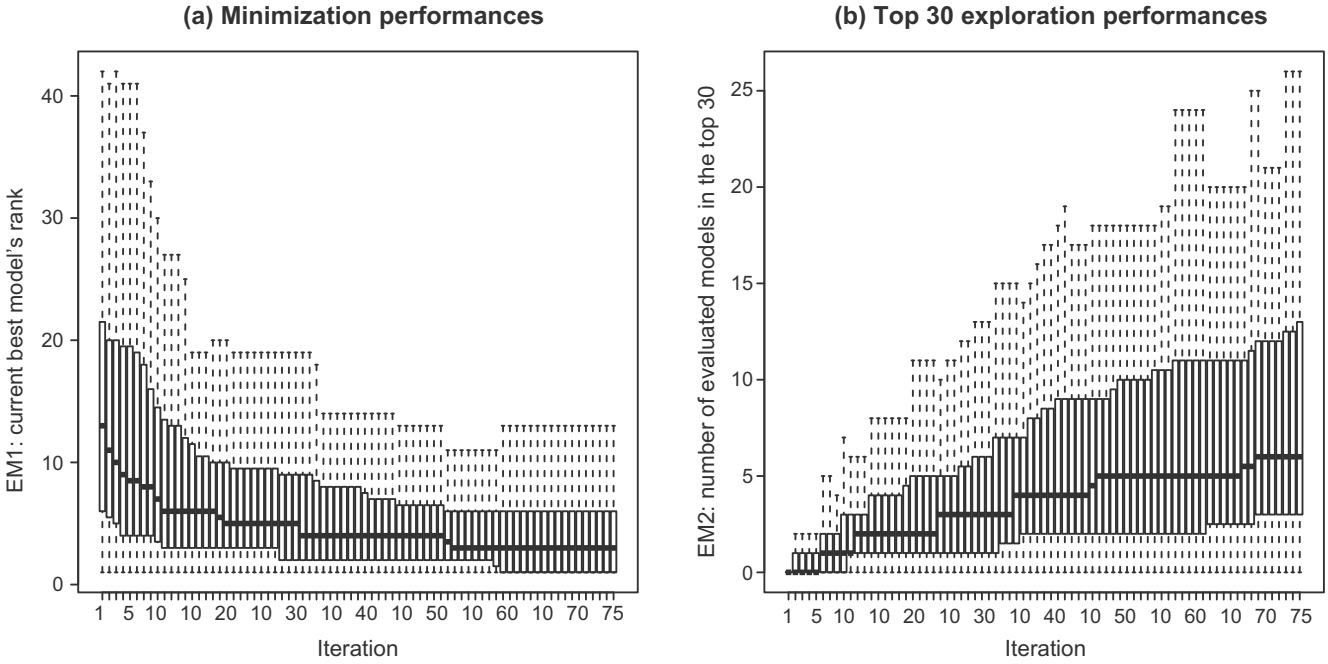

Fig. 12. Performances of the ProKSI algorithm (based on proxy 1) without power transform of the misfit. (a) Box-plot of EM1 over the 100 replicates of the numerical experiment. (b) Box-plot of EM2.

We can see in Fig. 8 that the optimal value of θ is very clearly defined since the log-likelihood curve has a large curvature at its minimum value.

The quality of the resulting kriging predictions is then evaluated: we first use a standard cross validation technique on the 50 samples used to build the kriging model (Fig. 9(a)) and then extend the comparison to an external validation on the complete ensemble of 1000 values (Fig. 9(b)). In both cases, the predicted values obtained by kriging are in good agreement with the true values; the regression line of predicted versus actual values has in intercept $B0$ close to zero and a slope $B1$ close to 1 (Fig. 9), indicating that the kriging predictions are not highly biased. Furthermore, one can see that the leave-on-out errors of (a) give a reasonable estimate of the prediction errors observed a posteriori on the exhaustive validation set.

The 50 iterative selections operated by the ProKSI algorithm provide candidate parameter fields that are assessed through the accurate numerical flow simulator. The misfit with the reference is plotted on Fig. 10.

6. Performance assessment

The good results obtained in the leading example (see Figs. 9 and 10) are of course conditioned by the chosen reference breakthrough curve f_{ref} (arbitrarily chosen as the one with index 800) and do not constitute a sufficient basis to appraise the ProKSI algorithm. Furthermore, the method is proxy-dependent, and it would make sense to test the sensitivity of the performances to both an improvement or a degradation in the proxy. In this section, we propose a more systematic benchmarking of the algorithm's performance by analyzing the results obtained with 100 different response curves, and for three different proxies, with a comparison to Monte Carlo random search in the case of the worse proxy. In that last situation, we will use a completely inadequate proxy model to test the robustness of the method. Furthermore, the effect of the power transform applied to the misfit function, as well as the effect of the replacement of the minimum by a quantile in the *EI* criterion are investigated. Before giving more details about the benchmark and the obtained results, let us first present the main performance evaluation metrics.

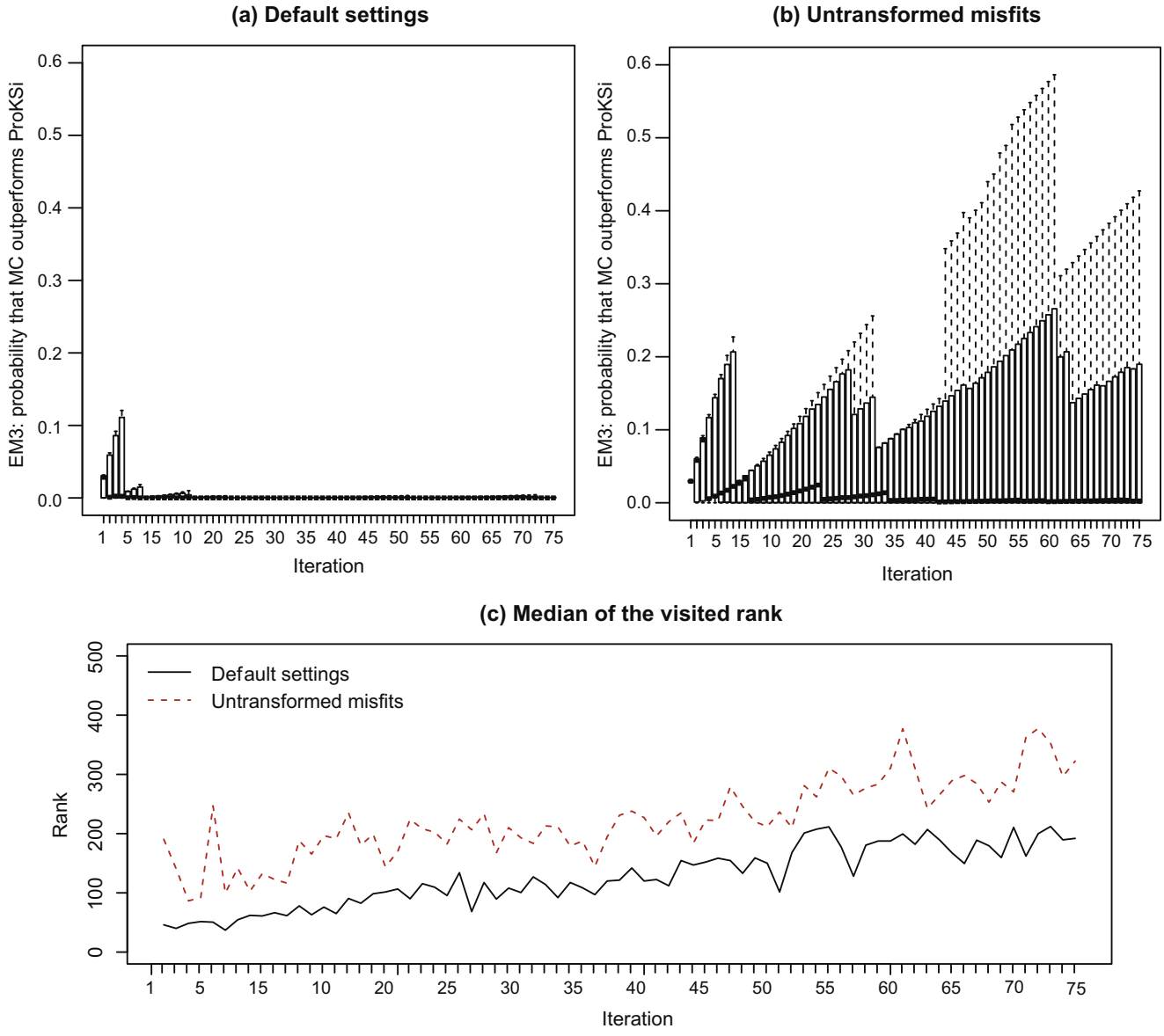

Fig. 13. Effect of the misfit transformation on the performances of the ProKSI algorithm in terms of its superiority with respect to a Monte-Carlo search and median rank of the evaluated models.

6.1. Performance evaluation metrics

EM1: current best model's rank.

One of the most natural way of evaluating an optimization method consists in monitoring the evolution of the misfit as a function of the number of iterations (Fig. 10). One can also plot the smallest misfit value achieved so far as function of the number of iterations. However, the curve obtained for such a metric would have a scale (on the y-axis) depending on the considered f_{ref}, which would prevent us from making comparisons between different tests. As a consequence, we choose to focus on the evolution of the rank of the current best model among the 1000 candidates. This rank would normally be unknown but here we can compute it because we evaluate the true misfit for all the candidate models (even those which are not selected by the ProKSI algorithm) in order to be able to test the efficiency of the method. Repeating the numerical experiment with multiple references, we can then plot some statistics of the rank as a function of the number of iterations (Fig. 11(a)).

EM2: number of evaluated models from the top 30.

The first metric (EM1) focuses on the capacity of the method to find at least one parameter field with a low misfit value, but not on its ability to explore the set of parameter fields with low misfit values. EM2 is meant to be a complement to EM1, by measuring the number of models of the top 30 (i.e. the 3% best models in terms of misfit value) evaluated along the algorithm. Though this proportion might seem rather arbitrary, EM2 gives a good picture of the algorithm's tendency to explore the possible multiple peaks of the posterior distribution of models. Again, the statistics of EM2 are plotted as a function of the number of iterations (Fig. 11(b)). Note that in the best configuration tested here (proxy 2 with $\alpha = 0.6$), 26 parameter fields among the top 30 were found in median after 75 iterations, so exploring all fields of top 30 out of 75 iterations appears as a kind of reachable best case (contrarily to exploring 75 among 75).

EM3: probability that random search outperforms the proposed algorithm.

It is expected that an elaborated algorithm like ProKSI (relying on a metamodel) performs better than random search, and at least not much worse in cases where the proxy is misspecified. The metric EM2 is well-adapted to base a comparison of ProKSI to a naive Monte Carlo (MC) algorithm, since the probability distribution of the number of points visited in the top 30 can be analytically derived for the case of a random search (this number then follows a hyper-geometric distribution). EM3 gives at each iteration of ProKSI the probability that MC sampling finds more points in the top 30.

6.2. Benchmark: design and implementation

6.2.1. Design of the benchmark

The aim of the benchmark was to assess the global performances of the ProKSI algorithm on the considered case study with the following specific questions in mind. How sensitive are the performances to: (Q1) the chosen proxy, (Q2) the value of the quantile α, and (Q3) the normalizing transform of the misfit values?

Consequently, we ran replications of the algorithm (by varying the reference curve) with different proxies, with or without power transform of the misfit function, and with different values of α. In order to obtain results based on solid statistical analysis, rather than on an arbitrary set of examples with a potentially low generalizability, we ran the ProKSI algorithm 100 times for each configuration (i.e. for each considered (proxy, transform, α) combination). For each considered proxy (p^1, p^2, and a third mismatched one described below), 50 parameter fields were chosen by Scheidt and Caers clustering technique, and 100 f_{ref} were randomly chosen among the 950 remaining candidates. Then, for any given configuration (in terms of transform and/or α value), 75 iterations of the ProKSI algorithm were run for the 100 chosen f_{ref}. The results are visualized in terms of box-plot sequences representing the statistical distributions of 100 values for the considered evaluation metric, evolving over the 75 iterations. Finally, for EM3, one sequence of 75 probabilities that a Monte Carlo algorithm would lead to more points in the top 30 than the proposed approach (one probability per iteration) can be produced for each replicate. We chose to summarize these

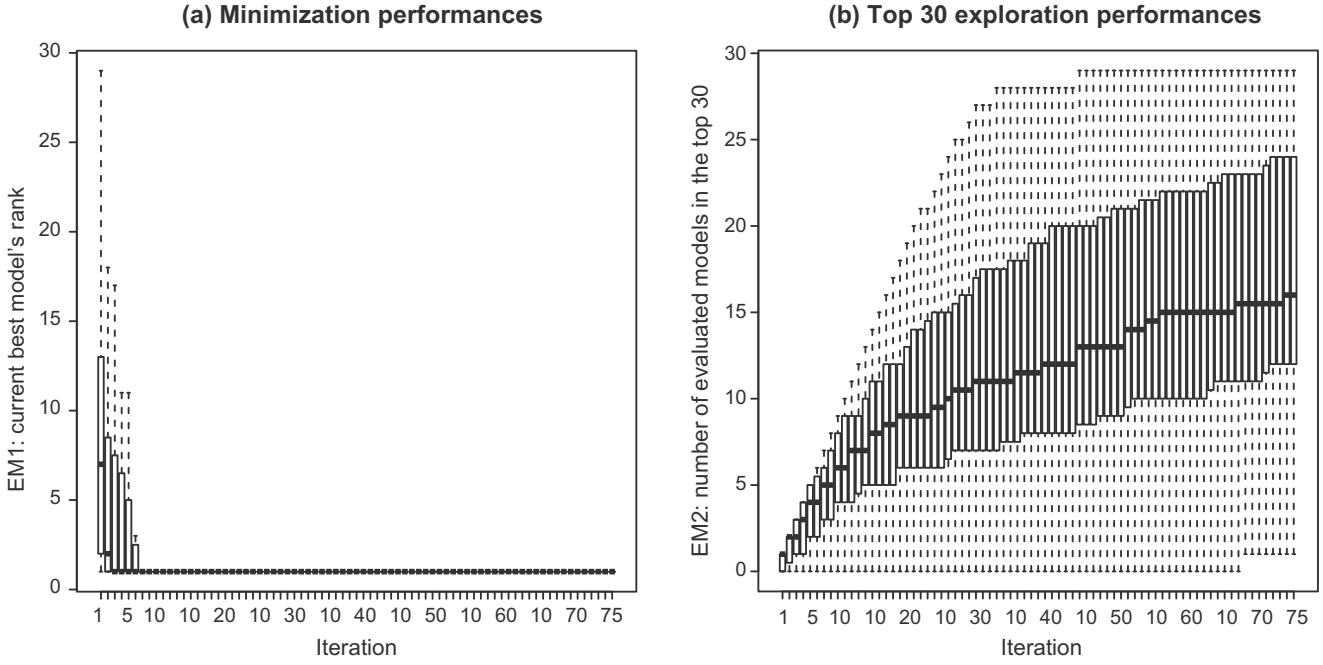

Fig. 14. Performances of the ProKSI algorithm (based on proxy 2) with default settings.

results by representing sequences of box-plots for the selected configurations.

6.2.2. Implementation of the benchmark

All the benchmark algorithm runs and the performance evaluation calculations were done using the open source statistical software R, based on the numerical simulation results obtained for the 1000 multiple-statistics simulations (see implementation details in Section 5). The R code, gathered in form of a package (*Pro-KSI*, forthcoming on the *Comprehensive R Archive Network*), was called for each task of the following loop, forming the basic brick of the benchmark for any fixed configuration:

Algorithm 1. Testing procedure for a proxy with a given algorithm configuration

1: **Choose** the initial design of experiment (**50** points using Scheidt et Caers approach)
2: **Choose 100** different simulations among the 950 remaining points
3: **for** $i = 1$ to $i = 100$ **do**
4: **Run 75** iterations of the algorithm on the *ith* reference
5: **Evaluate** the 3 EM's for each iteration of the *ith* run
6: **end for**

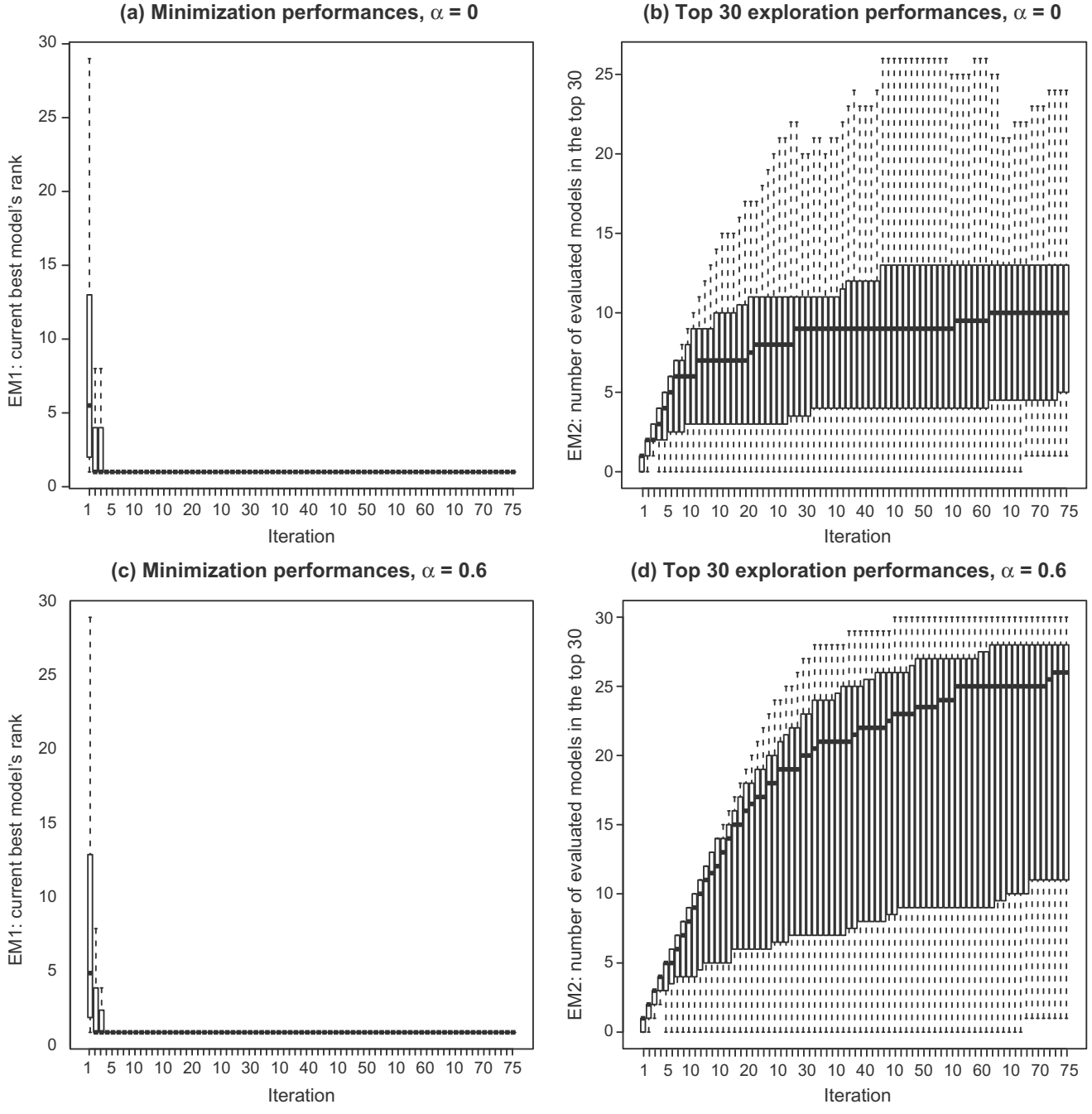

Fig. 15. Effect of the α parameter on the performances when using the second proxy.

6.3. Results

The first benchmark results, displayed on Fig. 11, deal with the performances on the ProKSI algorithm when applied to our test-case with proxy 1, and default settings concerning the normalizing transform and the *EI* variant (power transformation done, and $\alpha = 0.15$).

Fig. 11(a) represents the evolution of the statistics (box-plot) of EM1 over the 100 replicates, along the 75 iterations of the algorithm. We can see here that in 42 iterations, the actual best parameter field (out of 1000) has been found for more than 50% of the replications. In Fig. 11(b), the exploration performances are investigated in terms of EM2; it is found here that 15 parameter fields among the 30 best ones have been evaluated in median after 75 iterations of ProKSI. In total, these results show both how the proposed kriging metamodel helps reaching a fast convergence, and that ProKSI achieves a rather satisfying exploration of the set of best parameter fields in a limited number of iterations.

6.3.1. Effect of the misfit transformation on the algorithm performances

Fig. 12 represents the performances (in terms of EM1 and EM2) obtained by applying the ProKSI algorithm to our case study with default settings concerning the *EI* criterion ($\alpha = 0.15$) but without normalizing power transform for the misfit function.

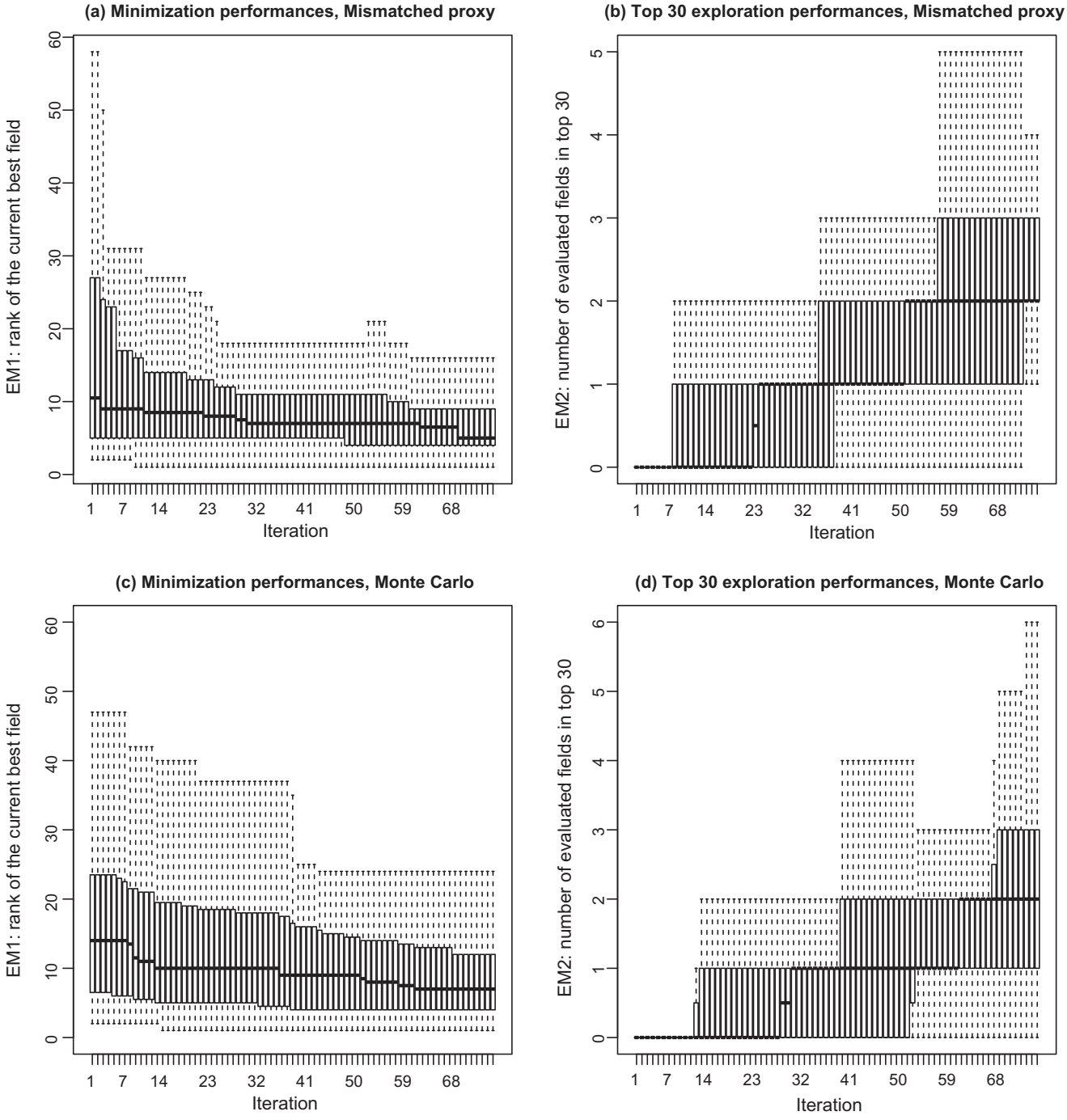

Fig. 16. Effect of a non-informative proxy on the performances.

The results appear to be clearly inferior to the ones obtained with the transformation: even after the 75 iterations, the median rank of the best evaluated parameter field is strictly above 1, which expresses a significantly slower convergence of ProKSI as with the transformed misfits. Similarly, the number of models forming the top 30 evaluated along the algorithm stagnates around 8 in median after the 75 iterations. The normalizing transform has thus clearly a positive effect on the efficiency of the algorithm, both in terms of fast convergence, and in terms of global exploration of the nearly optimal parameter fields.

However, as illustrated on Fig. 13, the results in terms of EM2 are good enough to outperform a pure random search (upper right graphic). On the lower graphic, the evolution of the median rank for the parameter fields evaluated by ProKSI with or without transform illustrate that the algorithm with transform spends more time in low misfit regions.

6.3.2. Effect of an improved proxy on the algorithm performances

Let us now present the results obtained when using the second proxy with default settings.

The most striking result when looking at Fig. 14 is the impressively fast convergence of the algorithm in terms of EM1 criterion. Indeed, in 7 iterations, the minimizer has been found in all considered cases (100 replicates). ProKSI successfully relies here on the information given by proxy 2 for uncovering the best point, only based on slightly more than the misfit values for the set of 50 initial models. What seems really outstanding in that case is that such a result is uniformly obtained for the 100 reference curves. To milden this success a bit, let us remark that the performances in terms of exploration are comparable to the first proxy, that is one half of the top 30 models were evaluated in median after termination.

6.3.3. Effect of the α parameter (from EI_α) on the algorithm performances

We investigate here the effect of the parameter α, tuning the quantile level in the proposed generalization of *EI*, on the performances of the algorithm. We obtained very different results for the two proxys. Indeed, the performances of ProKSI were not very sensitive to α when using the first proxy, so that we do not discuss this case here, and refer the interested reader to the appendix for more detail. However, α was found to be strongly influencing the algorithm's performances when using the second proxy, as illustrated on Fig. 15.

It is indeed observed on Fig. 15(a) and (b) that using ProKSI with the standard *EI* criterion ($\alpha = 0$) is less efficient compared to the considered default value $\alpha = 0.15$ (see Fig. 14): even though the algorithm convergence to the minimum is always comparably fast, the exploration performances are strongly affected by this change of criterion (median number of points in the top 30 after termination decreased from 15 to 10). On the other hand, increasing *alpha* to 0.6 was found to greatly improve the results in terms of exploration (again, without affecting the minimization performances, see Fig. 15(c)) since the median number of points in the top 30 jumped to 25, as can be seen on Fig. 15(d). In summary, introducing this parameter α was found beneficial for forcing the algorithm to spend more iterations (out of the 75 allocated ones) exploring the top candidates. Its optimal tuning is of course problem-dependent. The rather arbitrary default value $\alpha = 0.15$ chosen here gave improved results in both considered cases, even though better performances were reached by using a larger α value in the case of the second proxy.

6.3.4. Effect of a non-informative proxy on the algorithm performances

Finally, we propose to test the performances of ProKSI when using a completely inadequate proxy model. The idea is to see if the algorithm remains consistently applicable when the simplified model is poorly (or not at all) informative, and how using ProKSI in such degraded conditions would perform compared to a naive Monte Carlo search.

In order to emulate a non-informative proxy, we started from proxy 1, and randomly permuted the 1000 indices. We then ran the ProKSI algorithm with this "mismatched" proxy, and compared them to trajectories obtained by Monte Carlo (the whole replicated for the 100 reference curves).

As illustrated on Fig. 16, the performances of ProKSI with "mismatched" proxy are comparable to those of Monte Carlo. The algorithm hence appears reasonably robust to a proxy misspecification, while being potentially very efficient for well-chosen proxies.

7. Conclusion

Handling high resolution geological models in combination with complex physics solvers requiring heavy computational load to provide an accurate representation of a system while representing uncertainty is often mutually exclusive. Accurate complex models are often too computationally demanding to be used in the general framework of a Monte Carlo approach and analytical propagation of uncertainty is intractable. Resolving this issue is an important research topic both from a theoretical perspective and for a wide range of applications [45, e.g.], including hydrogeology.

In this paper, we propose a contribution which consists in coupling an accurate model, a simple model (the proxy), and a statistical metamodel. The statistical metamodel is used to link the results of the proxy with those of the accurate model. More precisely, this is achieved by developing a specific covariance kernel accounting for the difference in responses from the proxy models and allowing to predict by kriging the misfit between a given reference curve and the response of accurate model. One of the strengths of this idea is that the use of the distance between proxy responses permits to drastically reduce the dimension of the kriging problem and allows an efficient inference of the covariance parameters. The quality of the relation between the accurate and the proxy models is also directly taken into account via the covariance kernel. In addition, the chosen covariance kernel can be tailored to the practical problem that has to be solved (through the proxy, the kernel k_F, and more), which makes the approach quite flexible.

In the case study, we showed how such an approach can help in the case of an inverse conditioning problem where the prediction refers to the misfit between observations and the accurate model responses. As a first step, we propose here an iterative search algorithm. This example extends previous work done by Caers and colleagues [30,29,9] in which we add a step based on the kriging model described above to orient the search. We propose to guide the selection of a model during the search by defining a modified Expected Improvement criterion EI_α such that the algorithm will potentially explore multiple basins of minimum, if they exist.

The systematic analysis of the case study showed the following results.

- When the proxy is informative, the method is extremely efficient in finding the parameter fields that minimize the misfit.
- When the proxy is less informative, the method efficiency decreases remains competitive with respect to a random search.
- The proposed modified expected improvement criteria allows both identifying the global minimizer and exploring the existing basins of minimum.
- The method is more efficient if the misfit is properly transformed so as to get a close-to-Gaussian sample.

- The parameter α –defining the quantile of the misfit distribution below which a parameter field is considered as an interesting candidate – allows to control the degree of exploration of the method. For low values of α, the algorithm tends to rapidly leave regions of parameter fields with proxy responses similar to those of already evaluated fields. For higher values of α, the algorithm spends more iterations in regions of fields with proxy responses close to the ones of evaluated fields with the lowest misfit values, hence performing a more extensive exploration of top candidates.

The results obtained so far are very encouraging and show that the use of a kriging technique to couple a complex and a simple model will open a broad range of new perspectives. Note that in the presented example, selecting the misfit minimizer among parameter fields generated by multiple-points simulation provides a useful parameter reconstruction without any regularization. If one wants to obtain not only the best solution but an ensemble of models, then the selection criterion and the iterative search procedure will have to be modified in order to ensure that the final ensemble will be a representative sample of the posterior distribution. The method can also be extended in a relatively straightforward manner to allow generating new candidate models by coupling it, for example, with the Iterative Spatial Resampling method [15]. It is also very clear that this type of approach can be parallelized to improve the numerical performances [46].

As a final note, the model selection approach illustrated in the case study is applicable when detailed geological knowledge is available (type of geological environment, sufficient data to build a good geological model, and known properties for the various facies) and when this information can be used to generate a prior ensemble of geological models. An obvious limitation of that approach is that the algorithm is only searching within a predefined ensemble of parameter fields. If all the proposed candidates are very unrealistic, it may happen that none of the candidates will lead to a reasonable fit to the observations. However, the approach is very flexible: users can apply it with many different geological priors, and explore their relevance in an efficient manner with the proposed method. There is no limit in the type of prior on the parameter field used to generate the initial ensemble (e.g. boolean model, truncated plurigaussian, pseudo-genetic model, etc.). Coping with such a variety of priors is usually not possible with more traditional inversion methods.

Acknowledgments

The work presented in this paper is part of the *Integrated methods for stochastic ensemble aquifer modeling (ENSEMBLE)* project supported by the Swiss National Science Foundation under the contract CRSI22_122249/1. The authors want to thank I. Lunati, who provided the finite volume matlab code to solve the flow and transport equations and helped to setup the flow problem and proxy simulations, as well as N. Linde, P. Brunner and five anonymous reviewers for their constructive comments.

Appendix A. Proof that a p.d. kernel chained with a proxy is p.d.

Property. *Let E and F be two arbitrary spaces. Given a positive-definite kernel k_F over $F \times F$, the kernel k_E defined by*

$$k_E(\mathbf{x},\mathbf{y}) := k_F(p(\mathbf{x}),p(\mathbf{y})) \tag{A.1}$$

is positive-definite over $E \times E$ for any function $p : E \to F$.

Proof. Let $n \in \mathbb{N}, \mathbf{x}_1,\ldots,\mathbf{x}_n \in E$, and $\alpha_1,\ldots,\alpha_n \in \mathbb{R}$. Then

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_E(\mathbf{x}_i,\mathbf{x}_j) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_F(p(\mathbf{x}_i),p(\mathbf{x}_j)) = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k_F(\mathbf{y}_i,\mathbf{y}_j) \geqslant 0$$

by using the definition of positive-definiteness applied to k_F with the points $\mathbf{y}_i := p(\mathbf{x}_i) \in F$ $(1 \leqslant i \leqslant n)$ and the coefficients $\alpha_1,\ldots,\alpha_n$ as above. □

Appendix B. On the approximate MLE used here for the covariance parameters

When $\tau^2 = 0$, it is known [40] that $\widehat{\mu} = \frac{\mathbf{1}^T R(\theta)^{-1}\mathbf{g}}{\mathbf{1}^T R(\theta)^{-1}\mathbf{1}}$, and the optimal value of σ^2 can be expressed as a function of θ only:

$$\sigma^{2*}(\theta) := \frac{1}{N}(\mathbf{g}-\widehat{\mu}\mathbf{1})^T R(\theta)^{-1}(\mathbf{g}-\widehat{\mu}\mathbf{1}), \tag{B.1}$$

where $R(\theta)$ is the correlation matrix of $G_{\mathbf{x}_n}$. Minimizing l is then equivalent to the one-dimensional minimization over θ of the so-called *concentrated*(or *profile*) log-likelihood:

$$l_c(\theta;\mathbf{g}) := l(\sigma^{2*}(\theta),\theta,0;\mathbf{g}). \tag{B.2}$$

When $\tau^2 > 0$, it is not possible to concentrate on σ^2 as in Eq. (B.1) because the correlation matrix itself then depends on σ^2 too.

Here we approach the problem sequentially, and preserve the concentration step at the price of a minor approximation. First, an estimate of τ^2 is derived based on variographic considerations (variance of differences corresponding to increments with smaller norm values). Then, a first guess of σ^2, say σ_0^2, is made using a similar approach. This guess may also stem from a previous iteration in the case of a sequential design of experiments. Based on τ^2 and σ_0^2, an approximate formula – analogue to Eq. (B.1) – is then proposed for the optimal variance as a function of the range:

$$\widetilde{\sigma^{2*}}(\theta) := \frac{1}{N}(\mathbf{g}-\widehat{\mu}(\theta)\mathbf{1})^T\left(R(\theta)+\frac{\tau^2}{\sigma_0^2}I\right)^{-1}(\mathbf{g}-\widehat{\mu}(\theta)\mathbf{1}), \tag{B.3}$$

θ is then tuned by optimizing the following approximate concentrated likelihood:

$$\widetilde{l}_c(\theta;\mathbf{g}) := l(\widetilde{\sigma^{2*}}(\theta),\theta,\tau^2;\mathbf{g}) \tag{B.4}$$

References

[1] de Marsily G, Delhomme JP, Delay F, Buoro A. 40 years of inverse problems in hydrogeology. CR Acad Sci Ser II A Earth Planet Sci 1999;329(2):73–87.
[2] Kaipio J, Somersalo E. Statistical and computational inverse problems. Applied mathematical sciences, vol. 160. Springer-Verlag; 2005.
[3] Carrera J, Alcolea A, Medina A, Hidalgo J, Slooten L. Inverse problem in hydrogeology. Hydrogeol J 2005;13(1):206–22.
[4] Tarantola A. Inverse problem theory and model parameter estimation. SIAM; 2005.
[5] Franssen HJ, Alcolea A, Riva M, Bakr M, van der Wiel N, Stauffer F, et al. A comparison of seven methods for the inverse modelling of groundwater flow. Application to the characterisation of well catchments. Adv Water Resour 2009;32(6):851–72.
[6] Oliver D, Chen Y. Recent progress on reservoir history matching: a review. Comput Geosci 2011;15:185–221.
[7] de Marsily G, Delay F, Goncalves J, Renard P, Teles V, Violette S. Dealing with spatial heterogeneity. Hydrogeol J 2005;13(1):161–83.
[8] Mosegard K, Tarantola A. Monte carlo sampling of solutions to inverse problems. Water Resour Res 1995;100(B7):12431–47.
[9] Caers J. Modeling uncertainty in the earth sciences. Wiley-Blackwell; 2011.
[10] Vrugt JA, Ter Braak CJF. Dream(d): an adaptive markov chain monte carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. Hydrol Earth Syst Sci 2011;15:3701–13.
[11] Oliver DS, Cunha LB, Reynolds AC. Markov chain monte carlo methods for conditioning a permeability field to pressure data. Math Geol 1997;29(1):61–91.
[12] Liu JS. Monte Carlo strategies in scientific computing. Springer-Verlag; 2008.

[13] Fu JL, Gomez-Hernandez JJ. A blocking markov chain monte carlo method for inverse stochastic hydrogeological modeling. Math Geosci 2009;41(2):105–28.
[14] Alcolea A, Renard P. Blocking moving window algorithm: conditioning multiple-point simulations to hydrogeological data. Water Resour Res 2010;46.
[15] Mariethoz G, Renard P, Caers J. Bayesian inverse problem and optimization with iterative spatial resampling. Water Resour Res 2010;46(W11 530).
[16] Mondal A, Efendiev Y, Mallick B, Datta-Gupta A. Bayesian uncertainty quantification for flows in heterogeneous porous media using reversible jump markov chain monte carlo methods. Adv Water Resour 2010;33(3):241–56.
[17] Cui T, Fox C, O'Sullivan MJ. Bayesian calibration of a large-scale geothermal reservoir model by a new adaptive delayed acceptance metropolis hastings algorithm. Water Resour Res 2011;47.
[18] Keating EH, Doherty J, Vrugt JA, Kang QJ. Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. Water Resour Res 2010;46(W10517).
[19] Bliznyuk N, Ruppert D, Shoemaker C, Regis R, Wild S, Mugunthan P. Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. J Comput Graph Stat 2008;17(2):270–94.
[20] Mugunthan P, Shoemaker CA. Assessing the impacts of parameter uncertainty for computationally expensive groundwater models. Water Resour Res 2006;42(10).
[21] Sacks J, Welch W, Mitchell T, Wynn H. Design and analysis of computer experiments. Stat Sci 1989;4(4):409–35.
[22] Matheron G. Principles of geostatistics. Econ Geol 1963;58:1246–66.
[23] Rasmussen C, Williams K. Gaussian processes for machine learning. M.I.T. Press; 2006.
[24] Paciorek C. Nonstationary gaussian processes for regression and spatial modelling. Ph.D thesis, Carnegie Mellon University, 2003.
[25] Mockus J. Bayesian approach to global optimization. Kluwer Academic Publishers; 1988.
[26] Jones D, Schonlau M, Welch W. Efficient global optimization of expensive black-box functions. J Global Optim 1998;13:455–92.
[27] Bect J, Ginsbourger D, Li L, Picheny V, Vazquez E. Sequential design of computer experiments for the estimation of a probability of failure. Stat Comput 2012;22(3):773–93.
[28] Suzuki S, Caers J. A distance-based prior model parameterization for constraining solutions of spatial inverse problems. Math Geosci 2008;40(4):445–69.
[29] Scheidt C, Caers J. Representing spatial uncertainty using distances and kernels. Math Geosci 2009;41(4):397–419.
[30] Suzuki S, Caumon G, Caers J. Dynamic data integration for structural modeling: model screening approach using a distance-based model parameterization. Comput Geosci 2008;12:105–19.
[31] Sambridge M. Geophysical inversion with a neighborhood algorithm-I: Searching a parameter space. Geophys J Int 1999;138(2):479–94.
[32] Caers J, Park K, Scheidt C. Handbook of geomathematics. Modeling uncertainty of complex earth systems in metric space, vol. Part 5. Springer; 2011. chap. p. 865–889.
[33] Kaipio J, Somersalo E. Statistical inverse problems: discretization, model reduction and inverse crimes. J Comput Appl Math 2007;198(2):493–504.
[34] Kennedy MC, O'Hagan A. Predicting the output from a complex computer code when fast approximations are available. Biometrika 2000;87(1):1–13.
[35] Lodoen OP, Tjelmeland H. Bayesian calibration of hydrocarbon reservoir models using an approximate reservoir simulator in the prior specification. Stat Model 2010;10(1):89–111.
[36] Doherty J, Christensen S. Use of paired simple and complex models to reduce predictive bias and quantify uncertainty. Water Resour Res 2011;47(W12534).
[37] Tossavainen OP, Percelay J, Stacey M, Kaipio JP, Bayen A. State estimation and modeling error approach for 2-d shallow water equations and Lagrangian measurements. Water Resour Res 2011;47.
[38] Mariethoz G, Renard P, Straubhaar J. The direct sampling method to perform multiple-point geostatistical simulations. Water Resour Res 2010;46.
[39] Guttorp P, Sampson P. Nonparametric estimation of nonstationary spatial covariance structure. J Am Stat Assoc 1992;87(417):108–19.
[40] Santner T, Williams B, Notz W. The design and analysis of computer experiments. Springer; 2003.
[41] Bayer P, Huggenberger P, Renard P, Comunian A. Three-dimensional high resolution fluvio-glacial aquifer analog – Part 1: Field study. J Hydrol 2011;405:1–9.
[42] Comunian A, Renard P, Straubhaar J, Bayer P. Three-dimensional high resolution fluvio-glacial aquifer analog – Part 2: Geostatistical modeling. J Hydrol 2011;405:10–23.
[43] Künze R, Lunati I. A matlab toolbox to simulate flow through porous media. Tech. Rep. University of Lausanne, Switzerland, 2011.
[44] Künze R, Lunati I. An adaptive multiscale method for density-driven instabilities. Journal of Computational Physics 2012;231(17):5557–70.
[45] Christie M, Cliffe A, Dawid P, Senn S. Simplicity, complexity, and modelling. Statistics in practice. Chichester: John Wiley & Sons, Ltd.; 2011.
[46] Ginsbourger D, Le Riche R, Carraro L. Computational intelligence in expensive optimization problems. Kriging is well-suited to parallelize optimization. Studies in evolutionary learning and optimization. Springer-Verlag; 2010. chap. p. 1867–4534.