Contaminant source localization via Bayesian global optimization

Guillaume Pirot[1], Tipaluck Krityakierne[2,3,4], David Ginsbourger[4,5,6], and Philippe Renard[7]

[1]Institute of Earth Sciences, University of Lausanne, Switzerland
[2]Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand
[3]Centre of Excellence in Mathematics, CHE, Bangkok, Thailand
[4]Oeschger Center for Climate Change Research, University of Bern, Switzerland
[5]Uncertainty Quantification and Optimal Design group, Idiap Research Institute, Martigny, Switzerland
[6]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland
[7]Centre for Hydrogeology and Geothermics, University of Neuchâtel, Switzerland

Correspondence to: G. Pirot (guillaume.pirot@unil.ch)

Abstract. A Bayesian optimization approach to localize a contaminant source is proposed. The localization problem is illustrated with two 2D synthetic cases that display sharp transmissivity contrasts and specific connectivity patterns. These cases generate highly non-linear objective functions that present multiple local minima. A derivative-free global optimization algorithm relying on a Gaussian Process model and on the Expected Improvement criterion is used to efficiently localize the minimum of the objective function, which corresponds to the contaminant source location. The generated objective functions are shared as a benchmark. This contribution is important because the functions present multiple local minima and are inspired from a practical field application. In addition, the computing time required to map the functions was on the order of two weeks. Sharing these complex objective functions provides a benchmark for global optimization techniques and should help designing new and efficient methods to solve this type of problems.

1 Introduction

The concept of polluter pays is not new (OECD, 1972) and holds for groundwater protection laws in many countries (USA, 1972; Swiss Confederation, 1983; European Union, 2000). A polluter can sometimes be identified by a specific chemical signature (Mansuy et al., 1997; Rachdawong and Christensen, 1997; Venkatramanan et al., 2016). However, when the signature is not unique, the ability to localize the contaminant source(s) can make defining responsibilities or reducing decontamination costs easier. The topic is not recent and several approaches have been developed and proposed in the last three decades to identify contaminant source characteristics such as source location or history release. In this paper, we show the ability of a Bayesian optimization approach to localize the source of a contaminant in a synthetic medium characterized by realistic property features.

In their review of mathematical methods for groundwater pollution source identification, Atmadja and Bagtzoglou (2001a) classify existing approaches into four categories: *1) Optimization approaches*, in which forward simulations are run successively and the simulated concentrations are compared to measured concentrations (e.g., Gorelick et al., 1983; Wagner, 1992; Datta et al., 2011); *2) Probabilistic and geostatistical approaches*, in which the Advection Dispersion Equation (ADE) are

solved backward in time based on the random walk particle methods (Bagtzoglou et al., 1992) or on stochastic differential equations (Wilson and Liu, 1994); *3) Analytical solution and regression approaches*, in which a set of equations can be solved analytically or whose parameters can be estimated by least-square regression (e.g., Ala and Domenico, 1992; Alapati and Kabala, 2000); and *4) Direct approaches*, in which the ADE are solved backward in time based on deterministic direct approaches such as Tikhonov regularisation (Skaggs and Kabala, 1994; Atmadja and Bagtzoglou, 2001b), quasi-reversibility (Skaggs and Kabala, 1995), minimum relative entropy (Woodbury and Ulrych, 1996) or the backward beam equation (Atmadja and Bagtzoglou, 2001b).

A complementary classification is proposed by Amirabdollahian and Datta (2013) in their overview on contaminant source characteristics identification. Their classification is rather based on computational complexity and refines the *Optimization approaches* class mentioned above into three sub-classes: *1) Response Matrix*, in which unit responses are assembled linearly (e.g., Gorelick et al., 1983); *2) Embedded Optimization*, in which the objective function embeds directly mathematical equations of flow and transport (e.g., Mahar and Datta, 2001) and *3) Linked Simulation-Optimization*, in which the optimization procedure calls numerical flow and transport simulators (e.g., Ayvaz, 2016).

The approach that we consider here is an example of the *Linked Simulation-Optimization* class as defined above, where the procedure driving successive simulator evaluations relies on Bayesian optimization principles. While Bayesian methods have been massively used throughout groundwater sciences and notably for contaminant source localization, let us emphasize that the term 'Bayesian optimization' does not refer to any arbitrary method that combines 'optimization' and 'Bayesian statistics'. Instead, the term refers to a specific family of optimization algorithms where a prior distribution is put on the objective function (See e.g. Shahriari et al., 2016, and references therein for an overview).

In practice, geological media are heterogeneous and analytical solutions are limited to homogeneous geological media to identify contaminant source characteristics. To simplify the classification proposed in the two reviews described above, we gather the different classes in two groups: backward or forward solver based approaches. Methods based on backward solvers consist of reversing the flow problem (Skaggs and Kabala, 1995; Milnes and Perrochet, 2007; Ababou et al., 2010). The ADE are solved backward in time. The transport physical processes are simulated 'backward' to localize the source and identify the release history. This classification regroups classes 2 and 4 as defined by Atmadja and Bagtzoglou (2001a). In this classification, both flow-field and contaminated plume are assumed perfectly known. Methods using forward solvers are based on an inverse problem formulation (Aral et al., 2001; Yeh et al., 2007; Mirghani et al., 2012). The source location and release history are inferred from concentration samples. Parameter models are proposed and used as input in a forward solver to simulate concentration breakthrough curves at the sample locations; when the mismatch between the simulated concentrations and the observed ones is within an acceptable level of error, the proposed model is accepted as a solution. This class of methods contains optimization methods as described by Atmadja and Bagtzoglou (2001a); Amirabdollahian and Datta (2013), but additionally contain posterior sampling methods which provide posterior probabilities of the solutions. In this class, less information about the contaminant plume is required and the method can be adapted to uncertain geology (Zhang et al., 2016).

Previous studies performed a characterization of contaminant sources in 1D (Woodbury and Ulrych, 1996), 2D (Singh and Datta, 2007) or 3D (Michalak and Kitanidis, 2004) modeling grids. In these examples, the source is often identified along

with other characteristics such as the release history(Aral and Guan, 1998), or the source geometry (Ayvaz, 2016). To the best of our knowledge, most existing studies consider the hydrogeological property field as homogeneous or multi-Gaussian like heterogeneous random field, which might not be the best representation of subsurface heterogeneity in flow and transport applications (Gómez-Hernández and Wen, 1998; Zinn and Harvey, 2003). One exception lies in the study conducted by Milnes and Perrochet (2007), reversing the flow, where the 2D synthetic aquifer is represented by channels and islands with a strong transmissivity contrast. So far, to the best of our knowledge, no geological medium featuring realistic property contrasts and connected features has been used in an inverse problem formulation of contaminant source characteristics identification.

Optimization approaches to contaminant source characterization usually consist of minimizing an objective function that relies on a misfit between simulated measurements and reference observations. The use of least square regression combined with linear programming (Gorelick et al., 1983) assumes a linear system, which is not adapted for the contaminant source localization problem. Classical non-linear optimization techniques following a gradient based approach (Mahar and Datta, 2000; Datta et al., 2011) present the risk of being stuck in local minima. Employing a tabu search algorithm (Yeh et al., 2007) presents the same inconvenience as it explores neighbor solutions. Combining a gradient descent algorithm with a genetic algorithm (Aral et al., 2001; Ayvaz, 2016) decreases the risk of becoming stuck in local minima, but the genetic algorithm may require longer parameter exploration if the mutations are not guided by a smart rule. A Levenberg-Marquardt iterative algorithm, that interpolates between the second order Gauss-Newton algorithm and the first order of a steepest descent algorithm (Hansen and Vesselinov, 2016), might offer strategies to prevent being trapped in a local minimum. Simulated annealing (Amirabdollahian and Datta, 2014) allows for a broader exploration but at a very high computational cost. Bayesian optimization is a powerful approach that limits the risk of being trapped in local minima and intelligently explores the parameter space by looking at figures of merit trading off exploitation of available results and space exploration such as the Expected Improvement (EI) criterion (Mockus, 1989; Jones et al., 1998a; Vazquez and Bect, 2010). To the best of our knowledge, the latter method has not yet been tried on contaminant source characterization problems.

The objective of this paper is threefold. First, to assess the performance of an inverse problem formulation in order to identify contaminant source characteristics on a synthetic case based on realistic hydrogeological property contrasts and connected structures. This is important because in spite of its advantages, inverse problem formulation to identify contaminant source characteristics has been employed only on multi-Gaussian type heterogeneities and the type of heterogeneities strongly influences mass transport. Second, to verify the efficiency of a Bayesian optimization algorithm which relies on expected improvement criteria in the formulated contaminant source identification problem. While Bayesian optimization has been applied to a variety of optimization problems, we believe that this is the first time the algorithm has been applied to the contaminant source identification problem. Third, to provide an open source black-box optimization benchmark that allows one to compare different optimization strategies on application driven objective functions, which are not currently available in the optimization community.

With these objectives, we propose an original application of an EI algorithm to infer, in a deterministic inverse problem formulation, the contaminant source location in a 2D heterogeneous aquifer that presents realistic property contrasts and connectivity structures. To allow for a comparison between the optimizer exploration and an exhaustive search of the discrete

parameter space, the model grid is limited to 2D to keep computational cost reasonable for flow and transport simulations. The 2D synthetic model is generated with a multiple-point statistics (Guardiano and Srivastava, 1993) algorithm called *DeeSse* (Mariethoz et al., 2010), from a training image representing the heterogeneous hydrogeological properties of a braided-river aquifer, which was generated by a pseudo-genetic algorithm (see Appendix A; Pirot et al., 2015). The hydrogeological properties and flow boundary conditions are assumed to be perfectly known. The flow and transport equations are solved numerically using the *Groundwater* software (Cornaton, 2007). The optimization is performed using the DiceKriging and DiceOptim R packages (Roustant et al., 2012). In addition, we provide a benchmark for optimization algorithms, which relies on an objective function generator that can be customized by choosing between 2 geological scenarios, 2 possible locations for the contaminant sources and by the selection of observations among 25 wells. The performance of the EI algorithm is assessed by 100 replications from different initial designs.

The paper is organized as follows: Section 2 describes the synthetic test case and the experimental setup. Section 3 explains the objective function generator. Section 4 details the steps of the EI algorithm. The results are presented in Section 5 and are discussed in Section 6. Conclusions are summed up in Section 7. The supplementary material provided online is listed in Appendix B.

2 Synthetic test cases

As different geological settings can lead to very different objective functions, and in order to test the robustness of the optimization method, we consider two synthetic cases corresponding to 5 m thick $\times$ 600 m long $\times$ 300 m wide braided river aquifers. Each aquifer is represented by a unique, supposedly known, 2D facies model (Figure 1) of 1 m by 1 m resolution to simplify the problem and to decrease the computing costs related to transport simulations. These 2D facies models (Figure 1), which present strong contrasts and realistic spatial structures, are obtained by MPS simulation, using the training image described in appendix (Figure A1). The hydrogeological properties associated to the facies are given in Table 1 and are inspired from analogs described in the literature (Jussel et al., 1994; Bayer et al., 2011). Note that the contaminant spreading is

facies	hydraulic conductivity $K(m/s)$	porosity	storage coefficient $S_s(m^{-1})$	molecular diffusion $D_m(m^2/s)$	longitudinal dispersivity $\alpha_L(m)$	transversal dispersivity $\alpha_{Th}(m)$
coarse sediments	10^{-1}	0.2	10^{-5}	10^{-9}	1	0.1
mixed sediments	10^{-3}	0.2	10^{-5}	10^{-9}	1	0.1
fine sediments	10^{-5}	0.2	10^{-5}	10^{-9}	1	0.1

Table 1. Hydrogeological parameters

mainly modeled by the explicit description of geological heterogeneity and therefore, the longitudinal dispersivity is taken as the smallest possible value with our mesh size. Another method to obtain 2D horizontal models of braided river aquifers from 3D models would have been to integrate vertically the hydraulic conductivity field, but since this smoothes out the hydraulic conductivity, the resulting 2D models present less contrasts and less realistic connected structures.

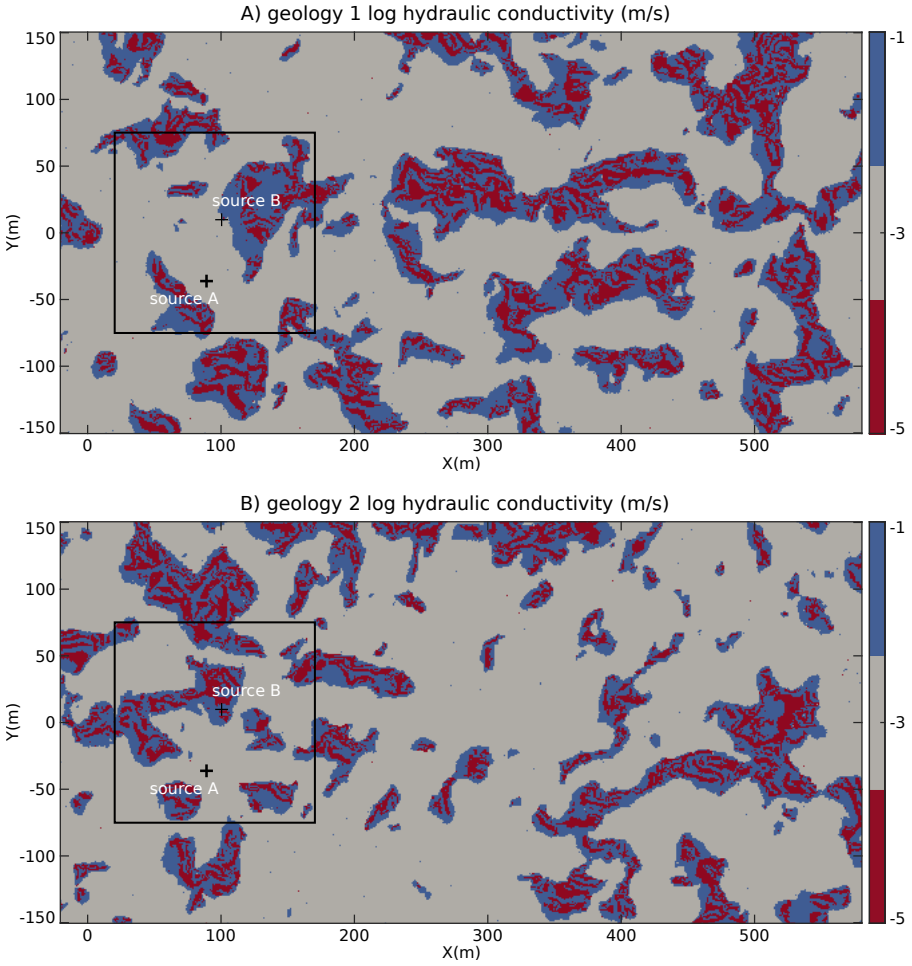

Figure 1. Experimental setup: 600m×300m 2D facies model of the aquifer; A) geology 1 and B) geology 2. The black square delimits the possible locations for the search of the contaminant source. The two reference source locations are identified by black crosses.

As boundary conditions for the flow and transport model, we impose a differential head of 2 m on the length of the model and no flow on the sides (Figure 2). We assume steady-state flow conditions to run transport simulations by solving the ADE with the finite elements code Groundwater (Cornaton, 2007).

The source of the contaminant is supposed to be unique, parameterized by the coordinates of its initial center of mass, and located within a search zone delimited by a 150 m × 150 m square-domain whose coordinates belong to $[20, 170] \times [-75, 75]$. To test the influence of the source location versus the geology, first on the misfit objective function and second on the ability of the proposed approach to deal with more or less complex objective functions, two reference locations (A and B) were chosen. Source A is located at ($x_s^A = 89, y_s^A = -36$). Source B is located at ($x_s^B = 100, y_s^B = 10$). To mimic way surface spills usually present some diffusion characteristics in their shape and can cover different geological features, the initial contaminant mass

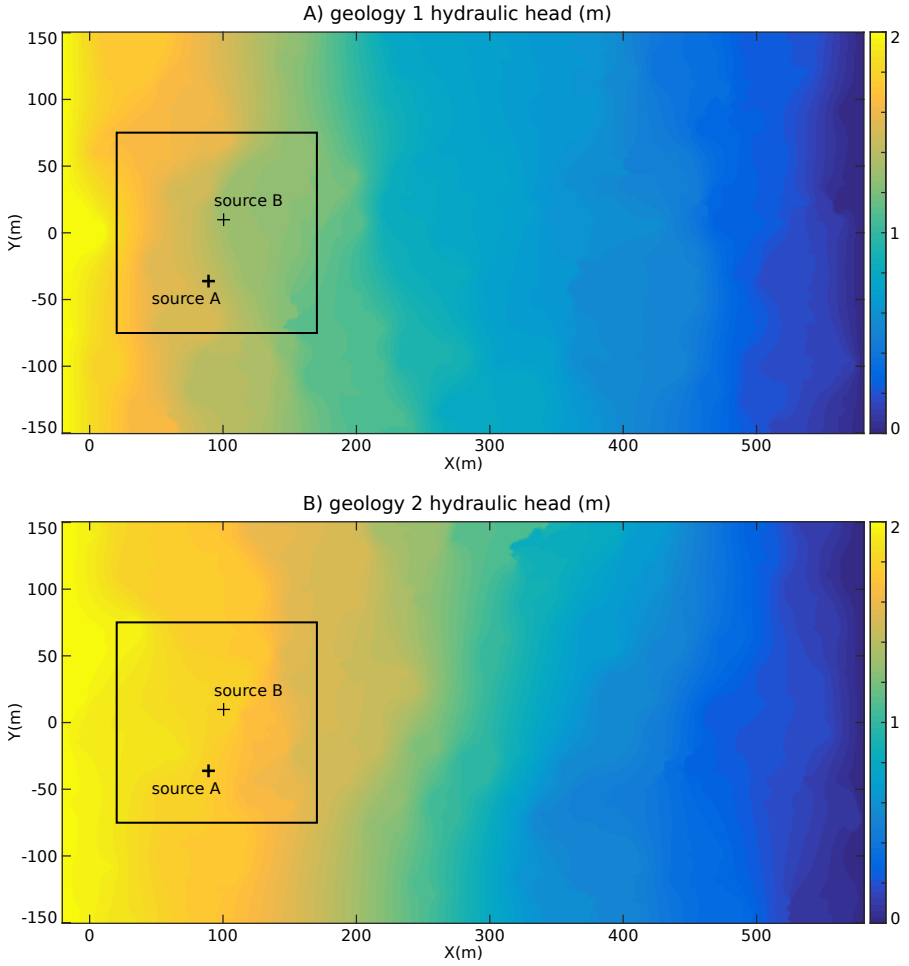

Figure 2. Steady state flow and boundary conditions for A) geology 1 and B) geology 2. The black square delimits the possible locations for the search of the contaminant source. The two reference source locations are identified by black crosses.

distribution at time 0 is chosen as a multi-Gaussian distribution centered on the source location and of standard deviation ($\sigma_x = 2.5$ m,$\sigma_y = 1.0$ m) for a total mass $m = 100$ kg. The reference concentration curves $c_{obs}(i,t)$ are obtained for $i = 1, \cdots, 25$ groundwater monitoring wells in Figure 3 and for times $t = 1, \cdots, T$ days. Three concentration breakthrough curves recorded at the well number 2, 16, and 22 are given as examples at the bottom of the figure.

The unknown location of the contaminant source is denoted $\mathbf{x} = (x_s, y_s)$. We define $c_{sim}(\mathbf{x}, i, t)$ as the simulated concentration level obtained at (i,t) when the contaminant source is located at $\mathbf{x}$. The aim is to find $\mathbf{x}$ that minimizes the following misfit objective function:

$$f(\mathbf{x}) = \left(\sum_{i=1}^{25} \sum_{t=1}^{T} |c_{obs}(i,t) - c_{sim}(\mathbf{x}, i, t)|^p \right)^{\frac{1}{p}}, \quad (1)$$

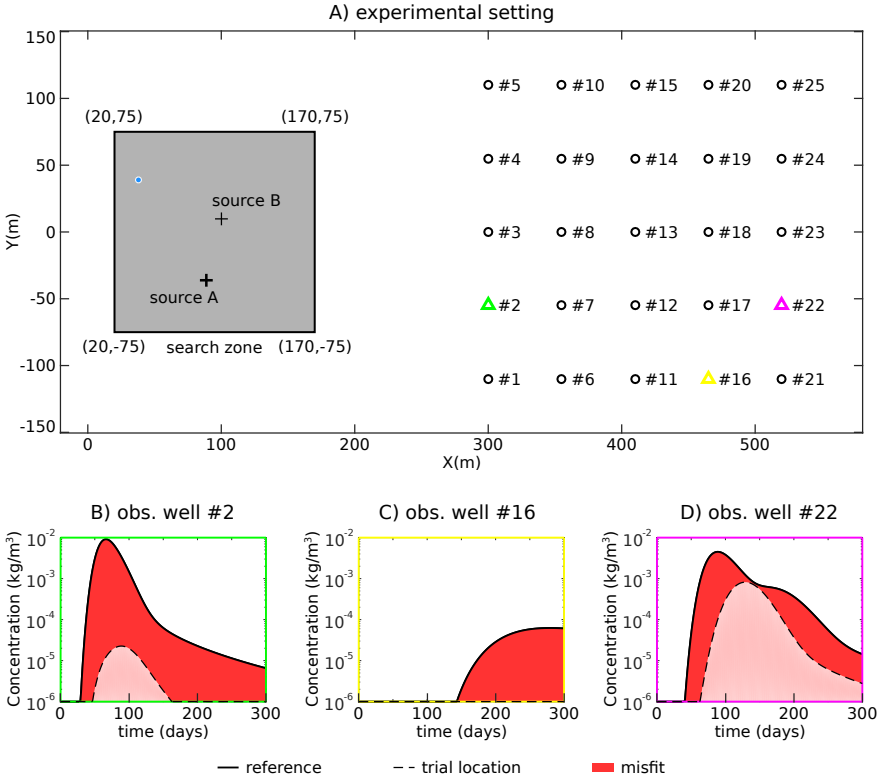

Figure 3. Misfit objective function settings; A) Location of the search zone (grey area), of the two reference contaminant sources and of the 25 groundwater monitoring wells (denoted by a circle or a triangle) within the hydrogeological model boundaries; the blue dot denotes the trial location of the contaminant; B), C) and D) misfit components at wells 2, 16, and 22 respectively, resulting from the comparison of the concentration breakthrough curves simulated at the trial location with the recorded ones for reference source A.

which corresponds to an ℓ^p norm. At the location of the reference source, the function reaches its minimum: 0. In this synthetic study, we neglect measurement errors on c_{obs} or conceptual or numerical errors on c_{sim} that may result from an incomplete knowledge of the transmissivity field or boundary conditions, which would be important to consider in a real field application.

The search zone is restricted to a discrete domain Z, using a regular grid of 3m resolution for three reasons. First, in practical applications, the location of the source is often restricted to an area thanks to historical information about industrial activities or accidents. Here, we apply the same principle but assuming a simple geometry. Second, this procedure and geometry allows us to provide an exhaustive computation of the objective function for the research community. Third, it is an interesting problem because most available optimization programs work either on continuous domains or are dedicated to specific classes of optimization problems (Integer programming, mixed linear integer programming), and few seem to be available for non-linear optimization over finite sets beyond metaheuristics used in combinatorial optimization (Rios and Sahinidis, 2013). In the case of our contaminant localization problem, by the nature of the problem, we do have a continuous structure (objective function)

where the domain is restricted to grid points. As an exhaustive evaluation of the objective function over Z is computationally expensive (depending on the mesh resolution), the aim of the optimization is to minimize the objective function f in the search zone within a limited number of iterations and for that purpose, we propose using an EI algorithm.

3 Benchmark generator

A set of pre-computed objective functions was generated by considering misfit functions parameterized by our different geologies, contaminant source locations or norms. The set of pre-computed objective functions is used in this paper for testing a global optimization technique (EI algorithm). More precisely, time-varying misfit values at each of the 25 wells were calculated for 2 geological geometries, 2 contaminant source locations (A and B) and 2 types of norms used in the misfit objective function ($p = 1$ and $p = 2$) at a full factorial design of candidate points in the search zone Z. Allowing any combination of observation wells among the 25 leads to $2^3 \times (2^{25} - 1)$ possible test functions (i.e. more than 268×10^6 test cases). As these functions are known through their respective 51^2 values at the discretized source space Z, they can be re-interpolated (e.g. using splines) for continuous optimization purposes. Here we instead consider the discrete problem of selecting the optimal location among 51^2 candidates and for that goal, we will apply a straightforward discretized version of an EI algorithm as presented in the next section. The data and some R functions to generate benchmarks for any input parameters are provided on GitHub at https://github.com/gpirot/BGICLP. A brief description of the repository is given in Appendix B of this paper.

4 Optimization methodology

The optimization algorithm used hereafter to minimize $f(\mathbf{x})$ over the domain belongs to a class of Bayesian optimization algorithms (Mockus, 1989; Shahriari et al., 2016). The Bayesian aspect refers to placing a random process prior Y on the unknown function f (possibly computationally expensive) and updating its probability distribution thanks to available evaluation results. The optimization part relies on using conditional distributions of Y to iteratively choose points with the identification of f's global optimum/optimizer(s) in view. The crux is to fit adequate probabilistic models and also to design adapted *acquisition functions* (a.k.a *infill sampling criteria* in surrogate-based optimization) in order to drive algorithms to an efficient optimization.

A very popular class of probabilistic models used in such context rely on Gaussian Processes (GP), that are fully specified by a mean function $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$. In this work, we use ordinary kriging with a Matérn ($\nu = 3/2$) covariance function (See Roustant et al. (2012) for details) and the kernel parameters are estimated by maximum likelihood using the DiceKriging R package. While it is also possible to use a transformation of the response in GP-based optimization (e.g. Jones et al., 1998a), on the considered data it did not lead to substantial differences in optimization performance despite the non-negativity of the misfit.

Denoting training inputs and outputs as $\mathbf{X}_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\mathbf{f}_n = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)\}$, assuming a GP prior with a constant unkown mean (endowed with an improper uniform prior) leads to a Gaussian conditional distribution with the

following marginal predictive mean and variance:

$$m(\mathbf{x}) = \hat{\mu} + \mathbf{k}(\mathbf{x})^T K^{-1}(\mathbf{f}_n - \hat{\mu}\mathbf{1}) \tag{2}$$

$$s^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{k}(\mathbf{x}) + \frac{(1 - \mathbf{k}(\mathbf{x})^T K^{-1}\mathbf{1})^2}{\mathbf{1}^T K^{-1}\mathbf{1}}, \tag{3}$$

where K is the $n \times n$ prior covariance matrix (assumed invertible here) of responses at training inputs, with $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x},\mathbf{x}_1), ..., k(\mathbf{x},\mathbf{x}_n))^T$ is an $n \times 1$ covariance vector and $\hat{\mu} = \frac{\mathbf{1}^T K^{-1}\mathbf{f}_n}{\mathbf{1}^T K^{-1}\mathbf{1}}$ is the best linear unbiased estimate of μ.

The optimization algorithm typically starts with constructing a space-filling design $\mathbf{X}_{n_0} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_0}\}$ (See, e.g., (Dupuy et al., 2015)) and evaluating $f(\mathbf{X}_{n_0})$ to initialize the knowledge of the algorithm (e.g., 9 blue dots in the left panel of Figure 4A). Here the initial $\mathbf{X}_{n_0}$ is generated based on latin hypercube sampling (McKay et al., 1979). Then, the algorithm begins its iterations. In each iteration, the ensemble of n available evaluations $\mathbf{f}_n = \{f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n)\}$ is used to train the GP model and make predictions at yet unexplored decision space locations. The predictive distributions is then used to compute the so called Expected Improvement criterion (Mockus, 1989), which indicates at every point in the decision space how much the objective function value may be decreased by relative to $f_{\min} = \min \mathbf{f}_n$, in expectation:

$$\mathrm{EI}_n(\mathbf{x}) = \mathbb{E}_n\left[\max(0, f_{\min} - Y(\mathbf{x}))\right]. \tag{4}$$

The EI criterion offers a good balance between exploitation of regions with low predictive mean values and exploration of regions with high predictive means, which provides an efficient optimization search scheme (e.g., red dot in the right panel of Figure 4A). It turns out that EI can be calculated analytically (Mockus, 1989; Jones et al., 1998b). In our discrete settings with moderate number of search points, the EI can be computed at all unevaluated locations of f (e.g. right panels of Figure 4). The decision space location with the largest EI value is considered as the next point $\mathbf{x}_{n+1}$ (e.g. red dot on right panels of Figure 4) to evaluate f. The optimization is run using the DiceKriging and DiceOptim R packages developed by Roustant et al. (2012). Here the number of iterations is fixed in advance, so that it stops when the maximum number of iterations allowed is reached. Covariance parameters are updated after each iteration by Maximum Likelihood Estimation.

5 Results

Using information from the 25 observation wells, the optimization algorithm is applied over 4 configurations that depend on the retained geology and on the contaminant source location as described in Table 2, with the ℓ^2 norm taken for the objective function $f(x)$. Starting from a specific initial design, the exploration of the objective function by the EI algorithm (aiming at the contaminant source localization), are displayed in Figure 5 for each scenario. These objective functions display multiple local minima, narrow valleys and sometimes very flat bottoms. These characteristics are making the search for the global minimum rather challenging especially for gradient based techniques. The locations explored by the EI algorithm are plotted over the 3 m $\times$ 3 m discretization of the objective function f. The white and blue dots represent respectively the initial and then explored

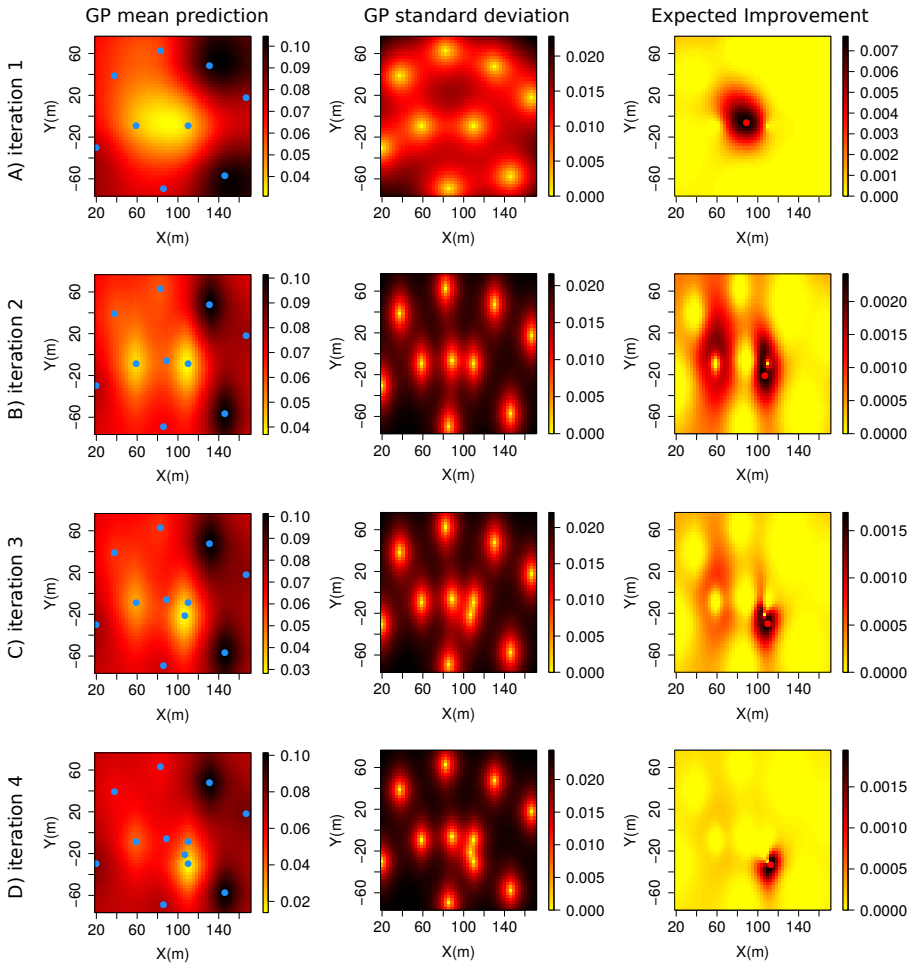

Figure 4. Illustration of the first four EI algorithm iterations for scenario 1; the sub-figures in the left column illustrate the prediction mean of f over the two-dimensional decision space at each iteration; the blue dots indicate the decision space locations where f was previously evaluated; the sub-figures in the center column illustrate the prediction variance of f over the two-dimensional decision space at each iteration; the sub-figures in the right column illustrate the expected improvement map over the two-dimensional decision space at each iteration; the red dot denotes the decision space location with the maximum EI value.

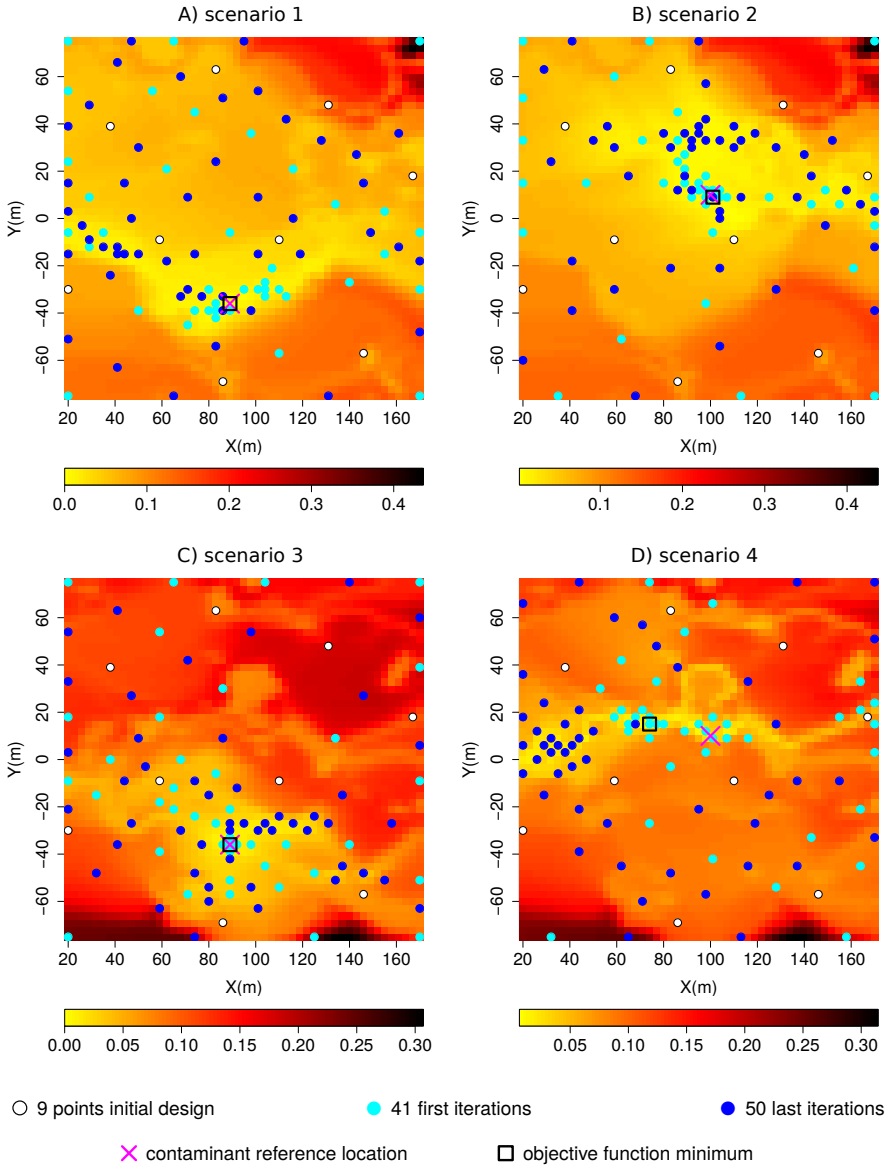

Figure 5. Solution exploration results for the 4 scenarios over the cost functions; A & B for geology 1; C & D for geology 2; A & C for initial contaminant location at $(89, -36)$; B & D for contaminant initial location at $(100, 10)$.

case	type of geology	source coordinate
1	geology 1	$(89, -36)$
2	geology 1	$(100, 10)$
3	geology 2	$(89, -36)$
4	geology 2	$(100, 10)$

Table 2. Description of the 4 configurations.

locations where the objective function was evaluated by the algorithm. In most cases, the minimum of the discretized objective function is reached in less than 50 evaluations. The geology seems to be the dominating factor for the global patterns of the objective function. Note that for scenarios 2 & 4, the contaminant source is located at $(100, 10)$, which is not on the discretized grid of the objective function; the closest point on the discretized grid is $(101, 9)$. For scenarios 4, the fact that the contaminant source is not located on a grid node implies that the contaminant reference source located at $(100, 10)$ and the minimum of the objective function located at $(80, 18)$ are 25 m apart.

The performance of the optimization algorithm is assessed on 100 replications. Each replication is characterized by a specific and uniformly drawn 9-point initial design. Each run is allowed a total budget of 100 evaluations of the objective function. The performance depends on the number of iterations required to locate the minimum of the objective function $\min_{\mathbf{x}} f(\mathbf{x})$. The performance can be assessed directly by looking at the optimality gap, i.e., the distance between the location of the best estimated minimum $f_{\min}$ of the objective function and the location of its true minimum $\min_{\mathbf{x}} f(\mathbf{x})$ as a function of the number of evaluations of f (Figure 6A to D). Another possibility is to look at the normalized best found minimum misfit between the true minimum $\min_{\mathbf{x}} f(\mathbf{x})$ and the best estimated minimum of the objective function $f_{\min}$ as a function of number of evaluations of f (Figure 6E to H). Both indicators behave similarly. Finally, the performance of the localization algorithms can be assessed by analyzing the distribution of the distance of the explored location that is closest to the true contaminant source over the 100 replications for a given number of iterations (Figure 7). Independently from the considered scenario, the bin counts for lowest values significantly increase when the number of iterations increase, and the bin counts for distances over 20 m rapidly come down to 0.

6 Discussion

Through successive kriging of the misfit between simulated and observed concentrations, guided by the expected improvement criterion, the proposed optimization algorithm localizes efficiently the source of a contaminant in a 2D geological environment representing realistic patterns and property contrasts. The algorithm requires only about 50 evaluations of the objective function instead of more than 2600 for an exhaustive evaluation on the discretized search zone ($\sim 1.9\%$).The total number of candidate points would increase exponentially in the number of dimensions of the parameter space, eliminating exhaustive search as an option from even moderate dimensions when assuming a high resolution.

Comparison of the different scenarios reveals that the geology controls the main features of the objective functions, which reinforce the importance of realistic geological structures in contaminant source localization problems. Of course, the shape

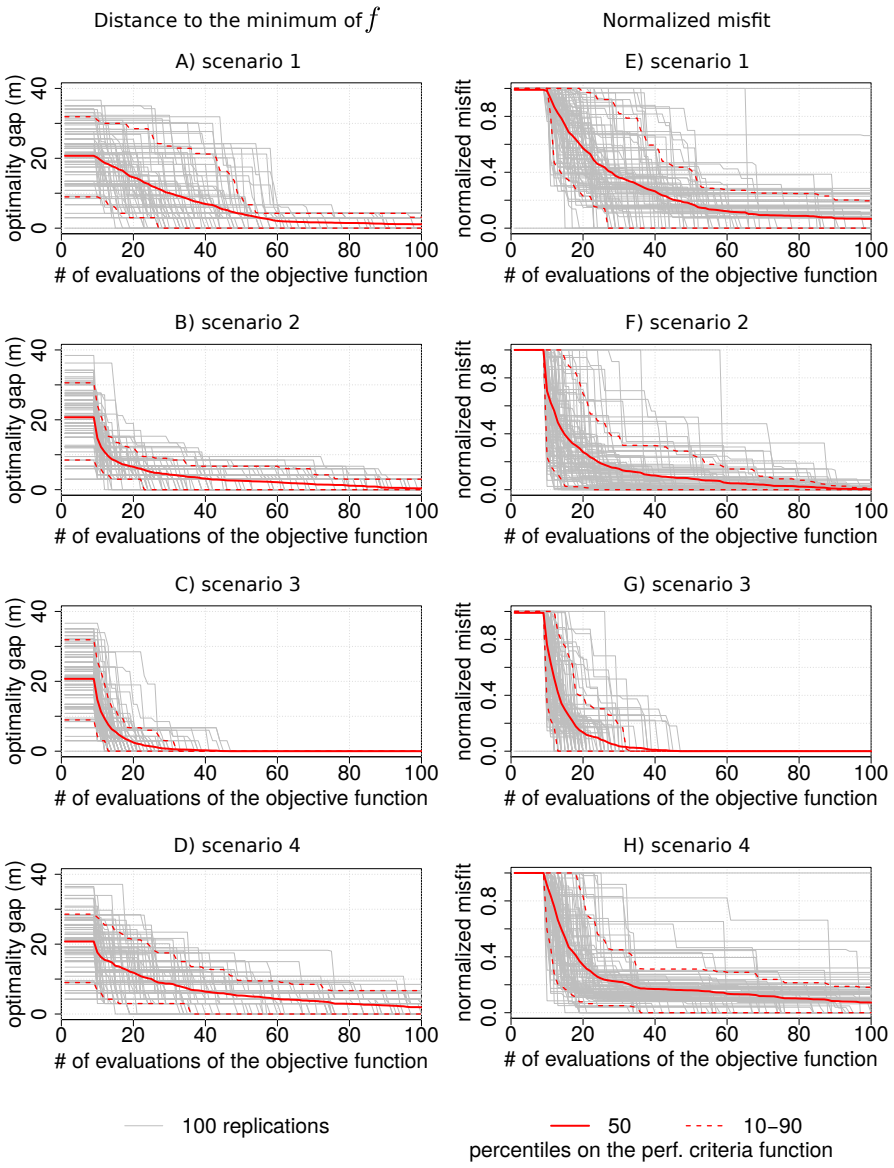

Figure 6. Performances of the EI optimization algorithm as a function of number of evaluations of the objective function for 100 different initial design; A), B), C) & D) distance of the best solution to the location of the objective function minimum; E), F), G) & H) normalized misfit; A) & E) scenario 1; B) & F) scenario 2 ; C) & G) scenario 3 ; D) & H) scenario 4.

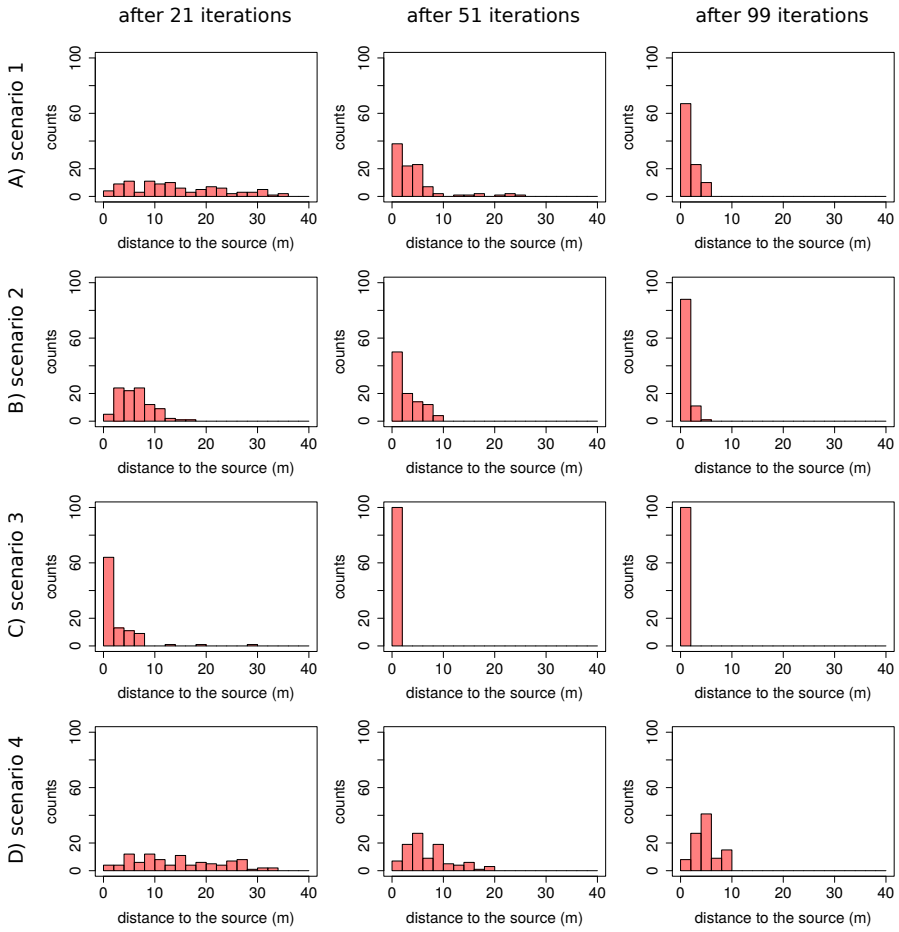

Figure 7. Distance to the contaminant source distribution for 100 replications for the best solution given by the EI algorithm ; row A) to D) for scenarios 1 to 4.

and location of lower values of the objective functions are controlled by the true location of the contaminant source. The results presented here are based on an objective function f computed with $p = 2$, which corresponds to an ℓ^2 norm. As the choice of the norm strongly influences the flat or deep aspect of valleys (low value zones) of the objective function, we additionally tested the EI algorithm on the 4 scenarios for ℓ^1 norm objective functions. We found that squared ℓ^2 norm lead to flatter wide valleys of low values for the objective functions, which might not favor the efficiency of the EI optimizer. However, the results and performances of the EI algorithm are very similar between the two norms tested.

By making source code of the function generator available for public use, we provide a benchmark of objective functions, driven by real hydrogeological applications, for testing and comparing optimization techniques. This benchmark will fill a gap for the community of applied mathematicians and statisticians who develop optimization algorithm and who want to test their tools on realistic objective functions. In addition, hydrogeologists will benefit from the code provided in the GitHub repository

so that they can implement the proposed optimization algorithm in their own applications. For the test case documented here and given the structure of the objective functions that are defined on a discrete domain, it does not seem relevant to apply off-the-shelf combinatorial algorithms. However it would be certainly of interest to compare to genetic/evolutionary algorithms compatible with such settings. Also, to enable comparisons with a broader class of derivative-free and also derivative-based algorithms, a pragmatic approach here would be to re-interpolate the data (with a careful inspection of the optima of the interpolator, i.e. a check that it is not perturbing the problem by too many potential artifacts) and conduct a benchmark involving Bayesian optimization (with EI and potentially also other infill sampling criteria) against a selection of state-of-the-art algorithms.

Strong assumptions have been made to localize the contaminant source in the presented application. The hydrogeolological properties and the flow boundary conditions are assumed to be perfectly known and the hydrogeological model is spatially limited to two dimensions. Because of their expensive computing costs, three-dimensional applications will not allow for an exhaustive search of the solution; this is why they may require, in the near future, optimization algorithms such as the one proposed in this paper. Further research should also consider the uncertainty related to hydrogeological properties characterization and flow and transport boundary conditions. Some steps have already been made in that direction (Koch and Nowak, 2016), but were limited to multi-Gaussian conductivity fields. In addition, a regular grid discretization might compromise the ability to accurately locate the contaminant source in the presence of strong flow path. For example, in a real-world application, the contaminant source has a very low probability of being located on a grid node. This problem could be avoided by using adaptive meshing, which would require more computing resources.

7 Conclusions

The use of 2D hydraulic conductivity fields that present sharp contrasts and specific connectivity patterns produces complex objective functions with multiple local minima. The proposed benchmark tool produced from these complex functions offers challenging real-world test for developers of optimization algorithms. The EI algorithm used in this 2D study localized efficiently the contaminant source that is located on a grid node. More generally, the proposed algorithm is an interesting approach for combinatorial optimization algorithm. To improve the limitation imposed by a source centered on the nodes of a fixed mesh, which is independent of the optimization algorithm, future research could be conducted on optimization embedding adaptive meshing in flow and transport simulations; another possibility would be to relax the constraint on mass distribution of the initial plume as a way to deal with its related uncertainty. The good performance of the algorithm on this 2D case is encouraging to continue toward 3D applications and toward integration of geological uncertainty in contaminant source localization problems.

Code and data availability. The data and some R functions to generate benchmarks for any input parameters are provided on GitHub at https://github.com/gpirot/BGICLP. A brief description of the repository is given in the appendix of this paper.

Appendix A: Training Image

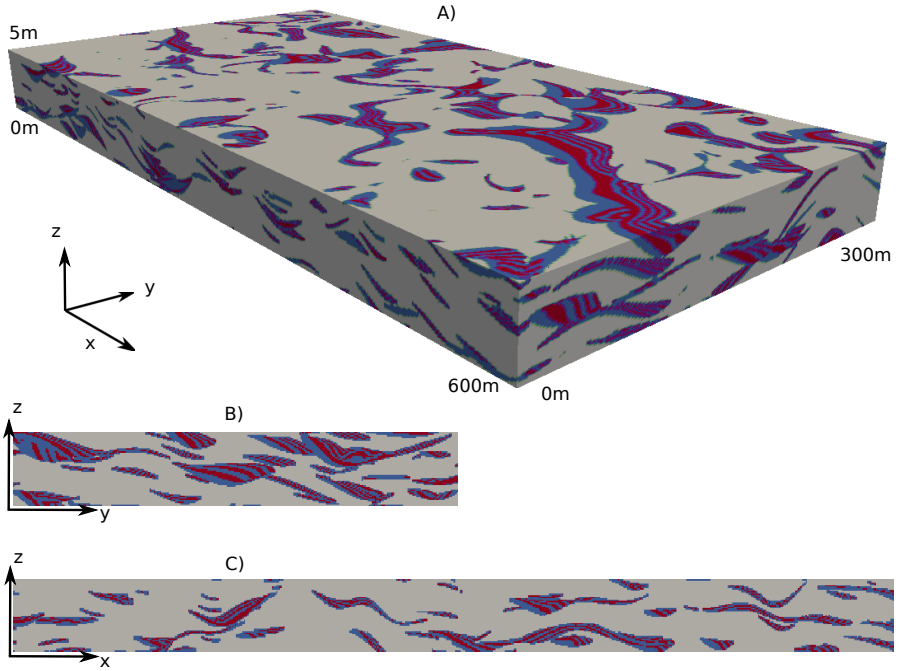

Figure A1. 600m $\times$300m $\times$5m training image with vertical scale exaggerated by 10; A) 3D representation; B) vertical section transversal to the main flow direction; C) vertical section longitudinal to the main flow direction. This three-dimensional model was generated by a pseudo-genetic algorithm proposed by Pirot et al. (2015). It is obtained by imitation of successive erosion and deposition events. Successive conditional simulations of topographies (Pirot et al., 2014) stacked together produce successive layers that are filled by heterogeneous geological facies according to a rule mimicking flow and sedimentation processes.

Appendix B: Supplementary material

The electronic supplementary material provided on the GitHub repository at https://github.com/gpirot/BGICLP with this paper contains 3 folders and 2 R-scripts.

The 'data' folder contains 1) the evaluation of sub-functions

$$f^p(\mathbf{x},i) = \sum_t |c_{obs}(i,t) - c_{sim}(\mathbf{x},i,t)|^p$$

over Z at $i = 1,\cdots,25$ observation wells for each of the 8 possible configurations (2 geologies, 2 sources, 2 norms) in the 'grid_25_wells_*.txt' files, 2) the $\mathbf{x}$ coordinates of the search zone Z and 3).

The 'figures' folder contains illustrations of $f(\mathbf{x})$ over Z for each of the 4 configurations when considering the 25 wells with the ℓ^2 norm.

The 'src' folder contains 3 R scripts. The 'image.scale.R' script, created by Pretty R at inside-R.org is used for graphic illustration purposes. The 'generate_lhs_on_grid.R' script allows generating initial point designs by latin hypercube sampling. The 'functionGenerator.R' script takes as arguments a selection of observation wells $\mathcal{W}$, a type of geology, the source coordinates and the type of norm used. It produces the evaluation of the function

$$f^p(\mathbf{x}, \mathcal{W}) = \sum_{i \in \mathcal{W}} f^p(\mathbf{x}, i)$$

over Z. When all wells are considered, $f^p(\mathbf{x}, \mathcal{W})$ resumes to $f(\mathbf{x})$ in Eq. 1.

The 'plotGeneratedFunction.R' script illustrates the use of the function generator and saves the plot in the 'figures' folder. The 'runEGO.R' script gives an example of how to use the proposed optimization algorithm.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to thank Fabien Cornaton for his support in the parameterization and use of Groundwater, Emily Voytek for her support in improving the reading of the manuscript, the anonymous reviewers and the editor Bill Hu for their comments and support. The second author would like to acknowledge support from the Oeschger Center for Climate Change Research (University of Bern), the Swiss Government Excellence Scholarship, as well as the Faculty of Science, Mahidol University.

References

Ababou, R., Bagtzoglou, A. C., and Mallet, A.: Anti-diffusion and source identification with the 'RAW' scheme: a particle-based censored random walk, Environmental Fluid Mechanics, 10, 41–76, 2010.

Ala, N. K. and Domenico, P. A.: Inverse analytical techniques applied to coincident contaminant distributions at Otis Air Force Base, Massachusetts, Groundwater, 30, 212–218, 1992.

Alapati, S. and Kabala, Z.: Recovering the release history of a groundwater contaminant using a non-linear least-squares method, Hydrological Processes, 14, 1003–1016, 2000.

Amirabdollahian, M. and Datta, B.: Identification of contaminant source characteristics and monitoring network design in groundwater aquifers: an overview, Journal of Environmental Protection, 4, 23–41, 2013.

Amirabdollahian, M. and Datta, B.: Identification of pollutant source characteristics under uncertainty in contaminated water resources systems using adaptive simulated anealing and fuzzy logic, International Journal of GEOMATE, 6, 757–763, 2014.

Aral, M. M. and Guan, J.: Identification of groundwater contaminant sources and release histories using genetic algorithms, vol. 1, Multimedia Environmental Simulations Laboratory, School of Civil and Environmental Engineering, Georgia Institute of Technology, 1998.

Aral, M. M., Guan, J., and Maslia, M. L.: Identification of contaminant source location and release history in aquifers, Journal of Hydrologic Engineering, 6, 225–234, 2001.

Atmadja, J. and Bagtzoglou, A. C.: State of the art report on mathematical methods for groundwater pollution source identification, Environmental Forensics, 2, 205–214, 2001a.

Atmadja, J. and Bagtzoglou, A. C.: Pollution source identification in heterogeneous porous media, Water Resources Research, 37, 2113–2125, 2001b.

Ayvaz, M. T.: A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems, Journal of Hydrology, 538, 161–176, 2016.

Bagtzoglou, A. C., Dougherty, D. E., and Tompson, A. F.: Application of particle methods to reliable identification of groundwater pollution sources, Water Resources Management, 6, 15–23, 1992.

Bayer, P., Huggenberger, P., Renard, P., and Comunian, A.: Three-dimensional high resolution fluvio-glacial aquifer analog–Part 1: Field study, Journal of Hydrology, 405, 1–9, 2011.

Cornaton, F. J.: Ground water: a 3-D ground water and surface water flow, mass transport and heat transfer finite element simulator, reference manual, University of Neuchâtel, Neuchâtel, Switzerland, 2007.

Datta, B., Chakrabarty, D., and Dhar, A.: Identification of unknown groundwater pollution sources using classical optimization with linked simulation, Journal of Hydro-Environment Research, 5, 25–36, 2011.

Dupuy, D., Helbert, C., and Franco, J.: DiceDesign and DiceEval: Two R Packages for Design and Analysis of Computer Experiments, Journal of Statistical Software, 65(11), 2015.

European Union: Good-quality water in Europe (EU Water Directive), https://www.epa.gov/laws-regulations/summary-clean-water-act, 2000.

Gómez-Hernández, J. and Wen, X.: To be or not to be multi-Gaussian? A reflection on stochastic hydrogeology, Advances in Water Resources, 21, 47–61, 1998.

Gorelick, S. M., Evans, B., and Remson, I.: Identifying sources of groundwater pollution: an optimization approach, Water Resources Research, 19, 779–790, 1983.

Guardiano, F. and Srivastava, R.: Multivariate geostatistics: beyond bivariate moments, in: Geostatistics Troia 1992, edited by Soares, A., pp. 133–144, Springer Netherlands, Dordrecht, 1993.

Hansen, S. K. and Vesselinov, V. V.: Contaminant point source localization error estimates as functions of data quantity and model quality, Journal of Contaminant Hydrology, 193, 74–85, 2016.

Jones, D., Schonlau, M., and Welch, W.: Efficient Global Optimization of Expensive Black-Box Functions, Journal of Global Optimization, 13, 455–492, 1998a.

Jones, D. R., Schonlau, M., and Welch, W. J.: Efficient global optimization of expensive black-box functions, Journal of Global optimization, 13, 455–492, 1998b.

Jussel, P., Stauffer, F., and Dracos, T.: Transport modeling in heterogeneous aquifers: 1. Statistical description and numerical generation of gravel deposits, Water Resources Research, 30, 1803–1817, 1994.

Koch, J. and Nowak, W.: Identification of contaminant source architectures-A statistical inversion that emulates multiphase physics in a computationally practicable manner, Water Resources Research, 52, 1009–1025, 2016.

Mahar, P. S. and Datta, B.: Identification of pollution sources in transient groundwater systems, Water Resources Management, 14, 209–227, 2000.

Mahar, P. S. and Datta, B.: Optimal identification of ground-water pollution sources and parameter estimation, Journal of Water Resources Planning and Management, 127, 20–29, 2001.

Mansuy, L., Philp, R. P., and Allen, J.: Source identification of oil spills based on the isotopic composition of individual components in weathered oil samples, Environmental Science and Technology, 31, 3417–3425, 1997.

Mariethoz, G., Renard, P., and Straubhaar, J.: The Direct Sampling method to perform multiple-point geostatistical simulations, Water Resources Research, 46, W11536, 2010.

McKay, M. D., Beckman, R. J., and Conover, W. J.: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code, Technometrics, 21, 239–245, 1979.

Michalak, A. M. and Kitanidis, P. K.: Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling, Water Resources Research, 40, W08302, 2004.

Milnes, E. and Perrochet, P.: Simultaneous identification of a single pollution point-source location and contamination time under known flow field conditions, Advances in Water Resources, 30, 2439–2446, 2007.

Mirghani, B. Y., Zechman, E. M., Ranjithan, R. S., and Mahinthakumar, G.: Enhanced simulation-optimization approach using surrogate modeling for solving inverse problems, Environmental Forensics, 13, 348–363, 2012.

Mockus, J.: Bayesian Approach to Gobal Optimization, vol. 37, Kluwer Academic Pub, Springer, The Netherlands, 1989.

OECD: Guiding Principles Concerning International Economic Aspects of Environmental Policies, Recommendation, http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=4&InstrumentPID=255&Lang=en&Book=, c(72)128, reprinted in 11 I.L.M. 1172, 1972.

Pirot, G., Straubhaar, J., and Renard, P.: Simulation of braided river elevation model time series with multiple-point statistics, Geomorphology, 214, 148–156, 2014.

Pirot, G., Straubhaar, J., and Renard, P.: A pseudo genetic model of coarse braided-river deposits, Water Resources Research, 51, 9595–9611, 2015.

Rachdawong, P. and Christensen, E. R.: Determination of PCB sources by a principal component method with nonnegative constraints, Environmental Science and Technology, 31, 2686–2691, 1997.

Rios, L. M. and Sahinidis, N. V.: Derivative-free optimization: a review of algorithms and comparison of software implementations, Journal of Global Optimization, 56, 1247–1293, 2013.

Roustant, O., Ginsbourger, D., and Deville, Y.: Dicekriging, Diceoptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization, Journal of Statistical Software, 51, 54p, 2012.

Shahriari, B., Swersky, K., Wang, Z., Adams, R., and de Freitas, N.: Taking the human out of the loop: A review of bayesian optimization, Proceedings of the IEEE, 104(1), 148–175, 2016.

Singh, R. M. and Datta, B.: Artificial neural network modeling for identification of unknown pollution sources in groundwater with partially missing concentration observation data, Water Resources Management, 21, 557–572, 2007.

Skaggs, T. H. and Kabala, Z.: Recovering the history of a groundwater contaminant plume, Water Resources Research, 30, 71–79, 1994.

Skaggs, T. H. and Kabala, Z.: Recovering the history of a groundwater contaminant plume: Method of quasi-reversibility, Water Resources Research, 31, 2669–2673, 1995.

Swiss Confederation: Federal Act on the Protection of the Environment, https://www.admin.ch/opc/en/classified-compilation/19830267/index.html, 1983.

USA: Clean Water Act, https://www.epa.gov/laws-regulations/summary-clean-water-act, 1972.

Vazquez, E. and Bect, J.: Convergence properties of the expected improvement algorithm with fixed mean and covariance functions, Journal of Statistical Planning and Inference, 140:11, 3088–3095, 2010.

Venkatramanan, S., Chung, S. Y., Kim, T. H., Kim, B.-W., and Selvam, S.: Geostatistical techniques to evaluate groundwater contamination and its sources in Miryang City, Korea, Environmental Earth Sciences, 75, 1–14, 2016.

Wagner, B. J.: Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling, Journal of Hydrology, 135, 275–303, 1992.

Wilson, J. L. and Liu, J.: Backward tracking to find the source of pollution, Water Management Risk Remediation, 1, 181–199, 1994.

Woodbury, A. and Ulrych, T. J.: MINIMUM RELATIVE ENTROPY: theory and application to recovering the release history of a groundwater contamninant, Water Resources Research, 32, 2671–2681, 1996.

Yeh, H.-D., Chang, T.-H., and Lin, Y.-C.: Groundwater contaminant source identification by a hybrid heuristic approach, Water Resources Research, 43, w09420, 2007.

Zhang, J., Li, W., Zeng, L., and Wu, L.: An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems, Water Resources Research, 52, 5971–5984, 2016.

Zinn, B. and Harvey, C. F.: 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 Resources Research, 39, 1051, 2003.