Advances in Water Resources 138 (2020) 103556

Contents lists available at ScienceDirect

Advances in Water Resources

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

The Traveling Pilot Point method. A novel approach to parameterize the inverse problem for categorical fields

Prashanth Khambhammettu [a,b,*], Philippe Renard [b], John Doherty [c]

[a] Arcadis, United States of America
[b] Centre for Hydrogeology and Geothermics, University of Neuchâtel, Switzerland
[c] Watermark Numerical Computing, Inc., Australia

ARTICLE INFO

Keywords:
Inverse problem
Categorical inversion
Traveling pilot points
Multiple-point statistics
Null Space Monte Carlo
Linear sub-space methods

ABSTRACT

Categorical parameter distributions are common-place in hydrogeological systems consisting of geologic facies/categories with distinct properties, e.g., high-permeability channels embedded in a low-permeability matrix. Parameter estimation is difficult in such systems because the discontinuities in the parameter space hinder the inverse problem. Previous research in this area has been focused on the use of stochastic methods. In this paper, we present a novel approach based on Traveling Pilot points (TRIPS) combined with subspace parameter estimation methods to generate realistic categorical parameter distributions that honor calibration constraints (e.g., - measured water levels). In traditional implementations, aquifer properties (e.g., hydraulic conductivity) are estimated at fixed pilot point locations. In the TRIPS implementation, both the properties associated with the pilot points and their locations are estimated. Tikhonov regularization constraints are incorporated in the parameter estimation process to produce realistic parameter depictions. For a synthetic aquifer system, we solved the categorical inverse problem by combining the TRIPS methodology with two subspace methods: Null Space Monte Carlo (NSMC) and Posterior Covariance (PC). A posterior ensemble developed with the rejection sampling (RS) method is compared against the TRIPS ensembles. The comparisons indicated similarities between the various ensembles and to the reference parameter distribution. Between the two subspace methods, the NSMC method produced an ensemble with more variability than the PC method. These preliminary results suggest that the TRIPS methodology has promise and could be tested on more complicated problems.

1. Introduction

Groundwater flow and contaminant transport models are commonly used to answer questions pertaining for example, to groundwater management and contaminant migration. These models solve the forward problem to answer the question under investigation. The forward problem involves model parameterization followed by solving a partial differential equation to obtain a state vector $\boldsymbol{d}$ (representing, for example, the groundwater head or contaminant concentration) in response to specified boundary conditions. The inverse problem, on the other hand, involves identifying the model parameter vector $\boldsymbol{m}$ from the state vector $\boldsymbol{d}$. Inverse problems in the groundwater modeling context have been studied extensively. Zhou et al. (Zhou et al., 2014) present a recent detailed discussion of the groundwater inversion problem and a review of historical and modern methods. The probabilistic formulation of the inverse problem (see, for example, Aster et al. (Aster et al., 2013)) can be expressed in terms of conditional probabilities as shown in Eq. (1).

$$q(\boldsymbol{m}|\boldsymbol{d}) = \frac{f(\boldsymbol{d}|\boldsymbol{m})p(\boldsymbol{m})}{c} \tag{1}$$

The term $q(\boldsymbol{m}|\boldsymbol{d})$ is the posterior probability density function and represents the probability of occurrence of a parameter vector conditioned by the observed measured dataset. The term $p(\boldsymbol{m})$ is known as the *prior* and represents the probability of occurrence of any model based only on the initial information such as geological knowledge without considering the measurements of the state variables $\boldsymbol{d}$. The term $f(\boldsymbol{d}|\boldsymbol{m})$ known as the likelihood represents the probability of simulating the measured data $\boldsymbol{d}$ given a model vector $\boldsymbol{m}$.

The categorical inverse problem is a special case of the groundwater inverse problem, pertaining to aquifers that consist of discrete geological facies/categories. For example, consider a two-categories aquifer with fluvial high-permeability channels incised in a low-permeability matrix.

* Corresponding author at :, 7550 Teague Road, Hanover, MD, 21076, United States of America
E-mail address: prashanth.khambhammettu@arcadis-us.com (P. Khambhammettu).

https://doi.org/10.1016/j.advwatres.2020.103556
Received 15 September 2019; Received in revised form 31 January 2020; Accepted 28 February 2020
Available online 29 February 2020

At any location in this aquifer, we would find only one of the two facies - channel or matrix. The inverse problem, in this case, requires us to generate categorical aquifer distributions when presented with prior geologic information about borehole logs (static data) and measurements of aquifer state (e.g., groundwater heads). Categorical problems are often more challenging to solve than their continuous counterparts because the parameter space is discontinuous. Linde et al. (Linde et al., 2015) presented an extensive review of existing methods for this class of problems. We summarize a few of them here.

The gradual deformation method (GDM) formulated by Hu et al. (Hu et al., 2001), (Hu, 2000) generates a sequence of model realizations that converge to matching the measured data. The key underlying concept in GDM is that the linear combinations of multiGaussian fields are also multiGaussian fields with similar statistics. It is, therefore, possible to explore a part of the model space by adjusting only one single parameter: a weight allowing to move between two pre-computed simulations. If the categorical field is obtained by truncation of one or several multi-Gaussian realizations, this process is straightforward and obtaining a discrete model that matches the measured observations can be treated as a usual continuous optimization problem. Caers and Hoffman (Caers and Hoffman, 2006) proposed the Probability Perturbation Method (PPM) when dealing with non-Gaussian priors and non-linear forward model responses. Rather than computing the posterior from the prior and likelihood, they instead decompose the posterior into a set of pre-posterior distributions containing facies and measurement data respectively. These pre-posterior probability distributions are perturbed until newer model realizations in the sequence increasingly converge to matching measurements. Ronayne et al. (Ronayne et al., 2008) applied the PPM to a transient aquifer test model and generated a distribution of permeable discrete channels embedded within less permeable deposits.

Alcolea and Renard (Alcolea and Renard, 2010) and Hansen et al. (Hansen et al., 2012) used an iterative Blocking Moving Window algorithm in conjunction with simulated annealing or a Markov chain based method to guide a multiple-point statistics (MPS) model in reproducing state variables and honor facies data known from prior knowledge. Mariethoz et al. (Mariethoz et al., 2010) proposed the Iterative Spatial Resampling (ISR) technique, a Markov chain-based method to sample from the posterior distribution. The transition from one element to the next in the Markov chain is based on sampling the values of the previous field at a set of random locations and using these points as conditioning data for the next iteration. While this procedure is straightforward and samples the posterior space in an unbiased manner, it is rather time-consuming. Jäggli et al. (Jaggli et al., 2017) proposed a faster approach named posterior population expansion (POPEX) expanding an initial ensemble of parameter models using MPS and local conditioning in such a manner that the new models are likely to belong to the posterior population. The POPEX approach was subsequently modified (Jäggli et al., 2018) to overcome predictive biases by combining machine learning techniques with an adaptive importance sampling strategy.

Several approaches to solve the categorical inverse problem based on pilot points have also been presented. Pilot points have been used to estimate heterogeneous hydrogeological parameter distributions for several decades ((Certes and de Marsily, 1991), (LaVenue and de Marsily, 2001), (Doherty, 2003)). Doherty et al. (Doherty et al., 2010) define pilot points as surrogate parameters in the inverse modeling process for representing heterogeneity in a lower-dimensional space. In these applications, a location-specific hydrogeological attribute (e.g., porosity, hydraulic conductivity) is associated with the pilot point. A pre-determined number of pilot points are placed at strategic locations along the model domain to capture the heterogeneity in the system. An iteration of the forward problem involves estimation of properties associated with each pilot point followed by spatial interpolation to create a spatially continuous parameter distribution from the discrete pilot point locations. The inverse problem involves the estimation of parameter values at the pilot point locations that honor the calibration constraints. Over the course of the parameter estimation, the locations of the pilot points remain static, but the parameters associated with the pilot points change. In the context of categorical fields, Li et al. (Li et al., 2003) used pilot points to guide an ensemble Kalman Filter approach to match dynamic (head) and geologic data simultaneously.

In this paper, we develop and test a new approach where pilot points are used in conjunction with linear subspace methods (Tonkin and Doherty, 2008) to solve the categorical inverse problem. The primary motivation here is that linear subspace methods are computationally inexpensive, and their application in the estimation of continuous real-world parameter fields has been well documented ((Keating et al., 2010), (Herckenrath et al., 2011)). We explore if these same approaches could be used to estimate discrete/categorical parameter fields. In our approach, we use pilot points in a non-traditional manner, that we refer to as the "Traveling Pilot Points (TRIPS)" approach. Rather than using pilot points for spatial interpolation, we iteratively adapt their positions to define the geometries of the discrete categories. In our opinion, there are two advantages to this approach. First, by using the positions of the pilot points, the categorical problem has been restated as a problem with continuous parameters which is easier to solve. Second, this approach allows us to infer the category geometries rather than to estimate them from spatial interpolation operations such as kriging indirectly.

The methodology described in this paper has only been tested so far on a synthetic problem with two categories and saturated two-dimensional groundwater flow. The technique might require additional modifications for more complex problems with multiple facies.

The subsequent sections of this paper are organized as follows. In Section 2, we present an overview of the TRIPS method and its applicability in the context of generating categorical parameter distributions. In Section 3, we present an overview of the various sampling methods used in Section 5. In Section 4, we present a synthetic groundwater problem with a categorical parameter distribution. In Section 5, we use the TRIPS approach to develop multiple likely parameter realizations for the synthetic problem. Finally, we present a summary of our findings in Section 6.

2. TRIPS Methodology

In Section 2.1, we introduce the principle and present the details of the implementation of *Traveling Pilot Points* (TRIPS) to solve the categorical inverse problem. In contrast to traditional pilot points, TRIPS are not fixed in location but instead can travel to locations of interest in the model domain.

2.1. The Traveling Pilot Points principle

Let us consider an aquifer containing permeable channels embedded within an impermeable matrix. It is possible to generate such discrete geological fields with different geostatistical techniques. One could use for example, transition probabilities (TProGS), plurigaussian simulations, object-based models, or multiple-point statistics to model these structures. For all these techniques, it is possible to set a fixed number of locations where the type of geology is known (for example presence of a channel), but the locations themselves are unknown. Providing these locations as conditioning data to the geological simulation algorithm allows to change the parameterization of the geological simulation and to solve the inverse problem in this manner means to search for the optimal locations of these traveling pilot points. This approach modifies a discrete inverse problem into a continuous one and should, therefore, facilitate its resolution. This idea is very general and can have many applications.

2.2. An example of geological model

To test this idea in a simple situation, we consider a binary case with channels as illustrated in Fig. 1. To constrain the geometry of the chan-

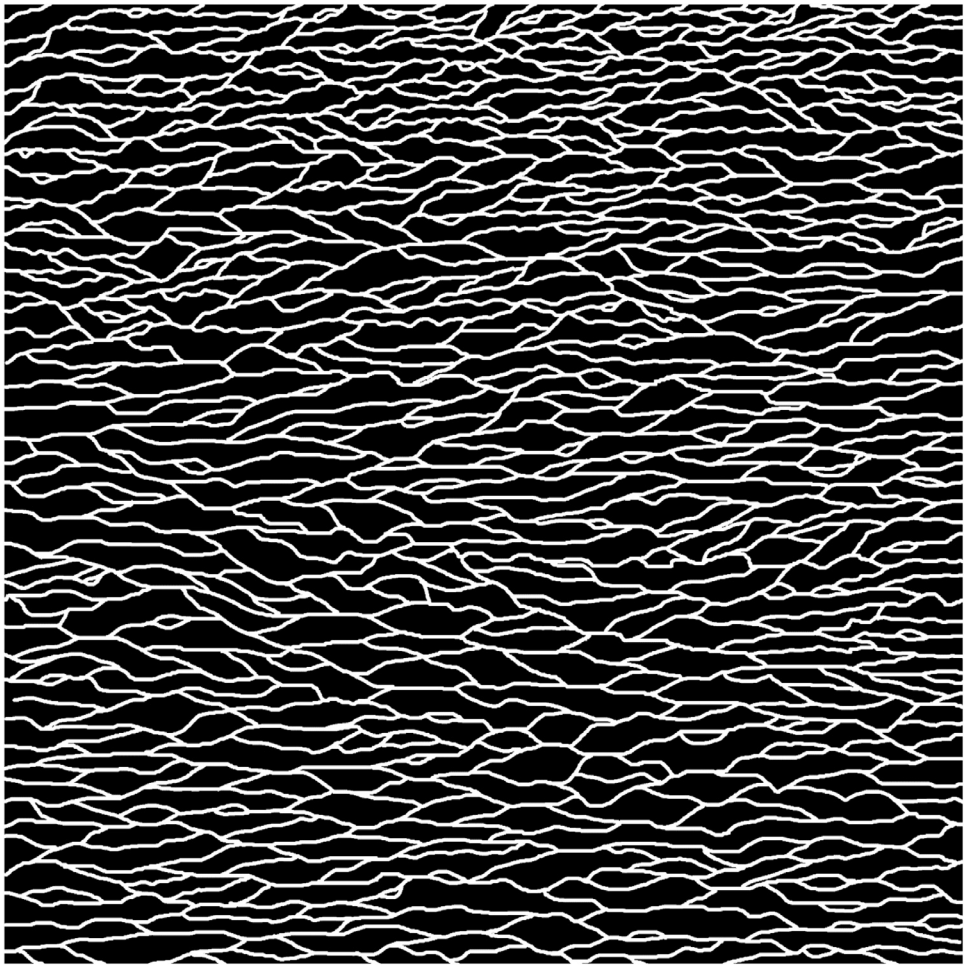

Fig. 1. Training Image representative of the geology in the synthetic aquifer. The image has a size of 2500 by 2500 pixels. The black pixels represent the matrix, while the white represent the channels. This image is borrowed from Laloy et al. 2018.

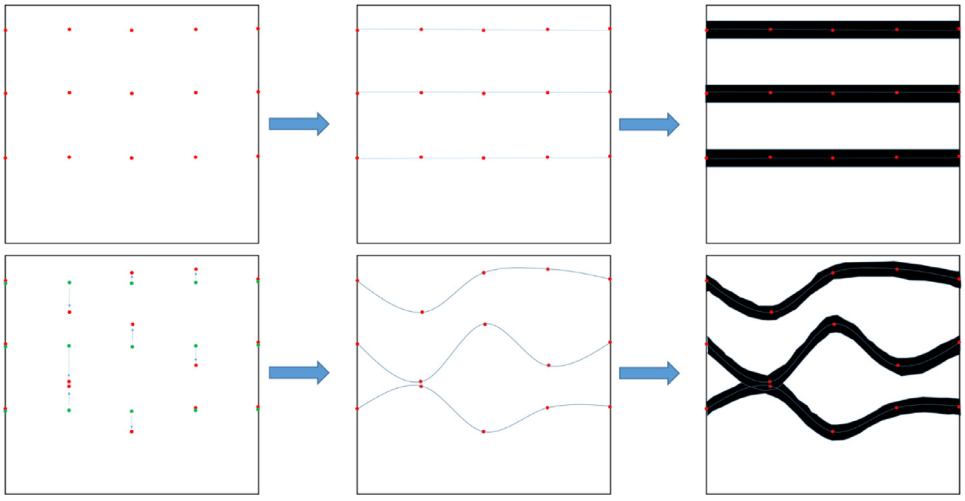

Fig. 2. Development of object-based model from pilot points with (top) initial position of TRIPS and (bottom) updated positions of TRIPS. Splines connecting related TRIPS are shown in the central panel and the channel objects created by buffering the splines are shown in the right panel.

nels in a simple manner, we use a two-step approach based on object-based simulations constrained by a training image.

On the one hand, the training image (Fig. 1) provides in a graphical manner the size of the channels, their sinuosity, their spacing, and so on. This image can be drawn by hand based on a geological concept. It offers flexibility and simplicity. On the other hand, the object-based model ensures that all the channels are continuous and that the geological models are generated very rapidly.

For the object-based model, we consider that there is a fixed number of channels crossing the area from left to right (Fig. 2). For each channel, we define a fixed number of traveling pilot points. For example, for a channel spanning an X distance of 100m, we can characterize it by 5 points spaced 20m apart. If the aquifer domain is 100 m × 100 m with a typical distribution of 3 channels, 15 points are used to track all the channels. To simulate the entire domain, the channel central lines are interpolated with a spline function using the position of the traveling pilot points as input. Then a constant thickness is applied along the central lines, and all pixels falling within this area are labeled as channel. We then have a simple function that relates the pilot point positions to channel geometry. To constrain the geometry of those channels in a

simple manner and make the link with the training image, we assumed that we could reproduce reasonably well the variability of the channels, their shapes and their relative positions using a multiGaussian distribution of the position of the traveling pilot points. In this manner, the prior geological model is fully determined by a set of mean values and a prior covariance matrix.

To estimate the prior covariance matrix, the training image has been cut into many sub-images having the same lateral extension as the simulation domain. The vertical extension was taken larger in order to account for channels that would enter the domain from the top or the bottom of the domain but not being entirely included in the simulation domain. For each sub-image, the positions of channels are tracked by recording the Y coordinate of the channel centerline at fixed intervals along the X axis. Then, the mean Y values $\bar{p}$ for every traveling pilot point and an empirical covariance matrix, $\boldsymbol{C_p}$, representing the covariance of their position along the Y axis is calculated from these recorded Y coordinates.

Once the covariance matrix is known, the generation of a geological model is obtained by first simulating a random vector $\boldsymbol{p}$ and then applying the procedure described above. The realizations of $\boldsymbol{p}$ are obtained using the discrete Karhunen-Loève expansion as shown for example by Sarma et al. (Sarma et al., 2008):

$$\boldsymbol{p} = \bar{p} + \boldsymbol{E}\boldsymbol{S}^{1/2}\boldsymbol{\rho} \tag{2}$$

In the above equation, **E** is the matrix of the eigenvectors of the covariance matrix $\boldsymbol{C_p}$, **S** is a diagonal matrix containing the eigenvalues of $\boldsymbol{C_p}$, and $\boldsymbol{\rho}$ is a vector of uncorrelated random normal variables (mean 0 and variance 1).

2.3. The Traveling Pilot Points approach

In the subsequent paragraphs, we present a more detailed description of the methodology. Let us consider a case where TRIPS are used to parameterize a property (e.g., hydraulic conductivity distribution) of a categorical aquifer containing f facies categories. Let n_i represent the number of TRIPS in facies i. The location of the j^{th} TRIP in the i^{th} facies in three-dimensional (3D) space is represented by (x_{ij},y_{ij},z_{ij}). The property value associated with the i^{th} facies category is represented by val_i. For example, if the x coordinates are known and the y and z coordinates are to be estimated, the vector $\boldsymbol{p}$, which contains all the unknowns (locations and category values) is represented by Eq. (3). This equation can be extended/modified for other problems with complex geometries.

$$\boldsymbol{p} = \left[y_{ij}, z_{ij}, val_i, \ldots\right] \; where \; i \in [1, f] \; and \; j \in \left[1, n_i\right] \tag{3}$$

The model parameter vector $\boldsymbol{m}$ is then determined by a spatial mapping/interpolation operation, as shown in Eq. (4). For example, $\boldsymbol{m}$ could represent the hydraulic conductivity field containing typically on the order of several tens of thousands of values which can be categorical while $\boldsymbol{p}$ contains only a few tens of continuous unknowns.

$$\boldsymbol{m} = \boldsymbol{Z}(\boldsymbol{p}) \tag{4}$$

In the above equation, the operator $\boldsymbol{Z}$ could represent a spatial interpolation method such as kriging or inverse distance weighted interpolation, for example. In this paper, this operator represents the mapping method illustrated in Fig. 2 and described in Section 2.2.

A groundwater flow/transport model uses the model parameter field $\boldsymbol{m}$ in conjunction with site-specific initial and boundary conditions to produce an output vector **d** of simulated heads/velocities/concentrations as represented in Eq. (5). The operator $\boldsymbol{g}$ in Eq. (5), an abstraction for the groundwater model, acts upon the parameter vector **m** to produce the output vector **d** of simulated heads/concentrations.

$$\boldsymbol{d} = \boldsymbol{g}(\boldsymbol{m}) = \boldsymbol{g}[\; \boldsymbol{Z}(\boldsymbol{p})] \tag{5}$$

If the vector $\boldsymbol{d}_{obs}$ represents the measured counterparts to $\boldsymbol{d}$, the measurement objective function, $\emptyset_m$, which defines the misfit between the model and the measurements is calculated in Eq. (6) as

$$\emptyset_m = \left[\boldsymbol{d}_{obs} - \boldsymbol{g}(\boldsymbol{m})\right]^T \boldsymbol{C}_D^{-1} \left[\boldsymbol{d}_{obs} - \boldsymbol{g}(\boldsymbol{m})\right] \tag{6}$$

Where the ^{T} superscript represents the matrix transpose operation and $\boldsymbol{C}_D^{-1}$ is a diagonal matrix with element q_{ii} (element in the i$^{\text{th}}$ row and i$^{\text{th}}$ column) containing the weight associated with the (i$^{\text{th}}$) measurement and equal (in this paper) to the inverse of the measurement error variance (Doherty, 2010).

In Eq. (6), no consideration was given to the nature of the parameter vector. In cases where prior/preferred knowledge about the underlying parameter distribution exists, it is important to include that information to reduce the ill-posedness of the problem (Tonkin and Doherty, 2005). We incorporate a plausibility/regularization term $\emptyset_r$ in Eq. (7) to represent the deviation of the parameter set from the prior knowledge about their preferred values.

$$\emptyset_r = \left(\boldsymbol{p} - \boldsymbol{p}_i\right)^T \boldsymbol{C}_p^{-1} \left(\boldsymbol{p} - \boldsymbol{p}_i\right) \tag{7}$$

In the above equation, the vector $\boldsymbol{p}_i$ represents our knowledge about *preferred* conditions. Here, we take for $\boldsymbol{p}_i$ the vector containing the simulated initial differences of the Y coordinates of the traveling pilot points obtained from the procedure defined in Section 2.2. Eq. (7) ensures that the traveling pilot points can move around the initial position but in a manner that is compatible with the statistics derived from the analysis of the training image. Furthermore, to better constrain the relative positions of the traveling pilot points, we also considered the differences between the values in the regularization term. This is implemented by assuming that the differences between the updated and initial parameters should remain small. The covariance of the differences can be estimated from the covariance matrix of the parameter values as described in Appendix A.

Note that for the sake of keeping the above explanations as simple as possible, we did not describe how the covariances and mean parameter values were included in the parameter for the hydraulic conductivities. This is done in a straightforward manner by assuming that the parameter values were uncorrelated to the positions. The final covariance matrix contains in this case two independent blocks: one for the position, one for the parameter values.

Finally, the global objective function, $\emptyset_g$, includes both measurement and parameter misfit:

$$\emptyset_g = \emptyset_m + \mu^2 \emptyset_r \tag{8}$$

The above equations represent a technique of regularization that was implemented in the parameter-estimation software, PEST (Doherty, 2010). The factor μ^2 is a regularization weight multiplier, controlling the parameter misfit, and is explicitly estimated during the inversion process. The inverse problem in the current context is a constrained minimization problem where the global objective function $\emptyset_g$ is minimized while keeping the channel geometry compatible with our prior knowledge expressed through regularization.

In summary, the overall flowchart for the TRIPS algorithm is presented in Fig. 3. An initial vector $\boldsymbol{p}_i$ is generated using Eq. (2). This information is then transformed into a model parameter field (e.g., hydraulic conductivity) with the aid of the spatial interpolation operator. An initial forward model simulation is carried out. If the misfit is considered acceptable, the parameter estimation is stopped. Otherwise, an optimization method (in this paper, gradient optimization in the PEST software) is used to minimize the global objective function $\emptyset_g$ and obtain better values of the TRIPS. An updated field of model parameters is created, forward model simulation is carried out, and the objective function is re-evaluated. This process is repeated until the optimization objectives are met.

3. Generating ensembles of realizations

In this section, we describe three different approaches to generate ensembles of realizations. The first approach, *rejection sampling* (RS), is

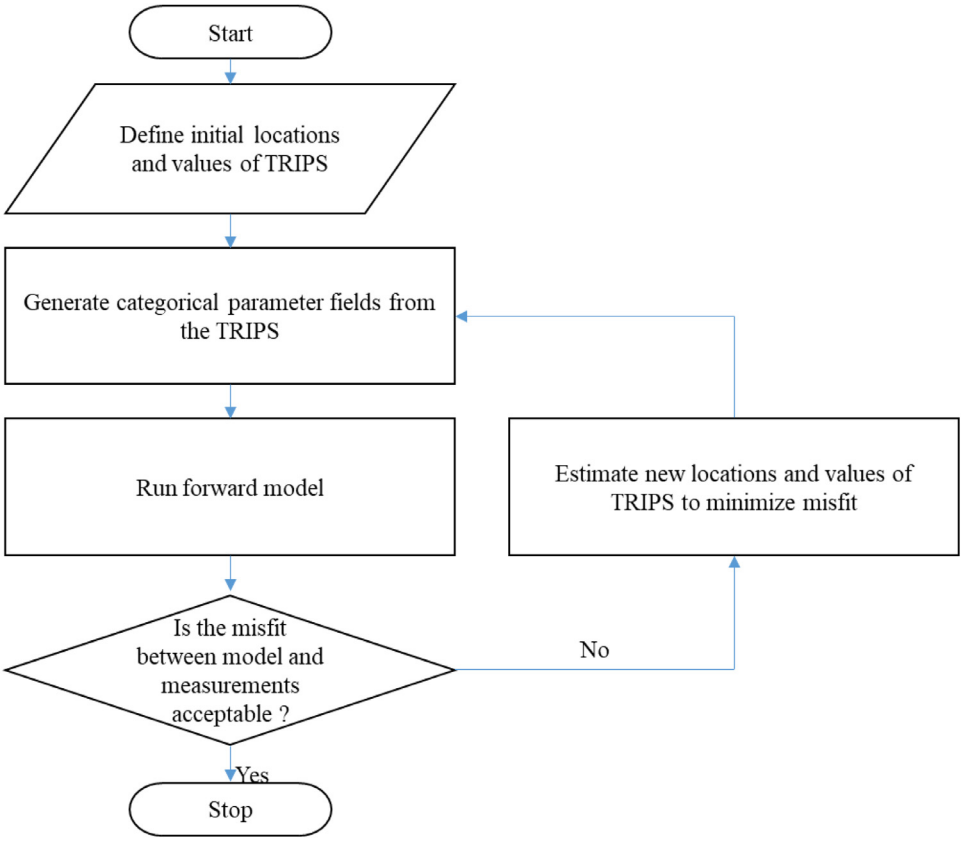

Fig. 3. Flowchart depicting the TRIPS algorithm.

a simple but computationally expensive approach to sample from the posterior distribution. Since this approach is capable of handling any kind of prior or posterior distributions, it serves as a benchmark for the other approaches which rely at least partly on a multiGaussian assumption. The second approach, *Null Space Monte Carlo* (NSMC), describes how the TRIPS methodology can be used in conjunction with subspace techniques which are computationally faster. The third approach, *Posterior Covariance* (PC), also a subspace technique, uses an alternate way to calculate the covariance matrix of the posterior and generates an ensemble rapidly. The second and third approaches are of interest in this paper as they cannot be applied to categorical inverse problems without using an indirect parameterization such as the one proposed here with TRIPS.

In summary, the TRIPS approach provides a framework for generating a channelized categorical aquifer field from pilot points spaced along the channel centerlines. The NSMC and PC methods use subspace techniques to sample the positions of these pilot points and the hydraulic conductivities of the aquifer categories.

These three approaches are applied and compared in Section 5 on a synthetic problem.

3.1. Rejection sampling

Rejection sampling (RS) described in (Mariethoz et al., 2010), (Tarantola, 2005), is a simple but computationally expensive way of sampling the posterior distribution. In this method, many candidate parameters, $\boldsymbol{p}$, are generated by sampling from the prior distribution, as described in Section 2.2. These parameters are converted into model parameters $\boldsymbol{m}$. Forward simulations are carried out, and misfit between the modeled and measured counterparts are tabulated. An acceptance probability, $P(\boldsymbol{m})$, defined in Eq. (9), is calculated for each candidate model based on the ratio of the likelihood function $L(\boldsymbol{m}) = f(\boldsymbol{d}|\boldsymbol{m})$ to the maximum possible value of the likelihood function $\boldsymbol{L_{max}}$.

$$P(\boldsymbol{m}) = \frac{\boldsymbol{L(m)}}{\boldsymbol{L_{max}}} \tag{9}$$

In this paper, $\boldsymbol{L_{max}}$ was determined as the maximum sampled value of the prior ensemble. The likelihood function is computed according to Eq. (10). It expresses the likelihood of a candidate model to reproduce the available data. It is inversely proportional to the measurement objective function and directly proportional to the standard deviation of the measurement error σ.

$$\boldsymbol{L(m)} \propto \exp\left[-\emptyset_m(\boldsymbol{m})\right] \tag{10}$$

For each candidate model, a random number from the Uniform distribution $U(0, 1)$ is concurrently generated along with the acceptance probability. If the acceptance probability is greater than this random number, the candidate model is accepted as a member of the posterior distribution. Otherwise, the candidate model is rejected. This method may reject many models, and therefore it is not computationally efficient, but the ensemble of accepted models represents the posterior distribution in an unbiased manner.

3.2. Null space Monte Carlo

The second method that we use in this paper is the Null Space Monte Carlo (NSMC) methodology described by Tonkin and Doherty (Tonkin and Doherty, 2008). It is a subspace-based pseudo-linear method capable of generating an ensemble of parameter realizations that have a reasonable fit with the data by construction. The NSMC method is described below.

The first step is to generate a single model with an acceptable level of misfit between the model and measurements. We do this using the gradient-based optimization method described in Section 2.3. The optimized parameter set from this model is denoted by the vector $\boldsymbol{p_c}$. The Jacobian matrix $\boldsymbol{X}$ is estimated. It contains the partial derivatives of the

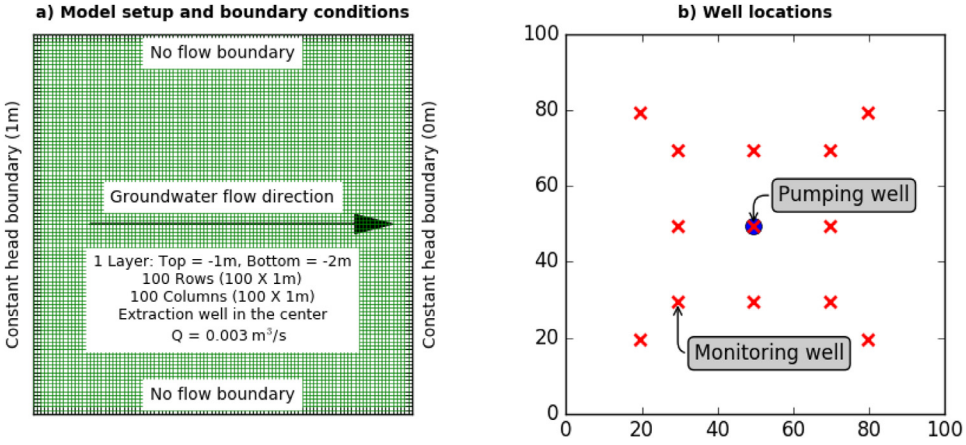

Fig. 4. Model setup, boundary conditions, and well locations for the synthetic problem. The locations of the pumping well and monitoring wells are shown in the plot to the right.

measured data with respect to the components of the vector $\boldsymbol{p_c}$. X_{ij} (representing the value in row i and column j) is calculated as the partial derivative of observation i with respect to parameter j. The weighted Jacobian matrix, $\boldsymbol{X^T C_D^{-1} X}$ is computed. The matrix $\boldsymbol{C_D^{-1}}$ contains observation weights as defined in Section 2.3. This weighted Jacobian matrix is decomposed using singular value decomposition (Tonkin and Doherty, 2008) as the product of three matrices in Eq. (11).

$$\boldsymbol{X^T C_D^{-1} X = U S V^T} \tag{11}$$

$\boldsymbol{U}$ is an orthonormal matrix containing the basis vectors for the range space of the weighted Jacobian; $\boldsymbol{S}$ is a rectangular diagonal matrix containing eigenvalues of the weighted Jacobian matrix; $\boldsymbol{V}$ is an orthonormal matrix containing the basis vectors for the parameter solution space and parameter null space. If there are n eigenvalues and the partition between the solution and null spaces is drawn after the first r eigenvalues, the matrix $\boldsymbol{V}$ from Eq. (11) can be thought of as $\boldsymbol{V} = [\ \boldsymbol{V_1}\quad \boldsymbol{V_2}\ ]$ where $\boldsymbol{V_2}$ has $(n - r)$ columns which form the basis vectors for the null space. Moore et al. (Moore and Doherty, 2005) present a discussion on the impact of this partition on predictive error variance.

Next, we generate multiple parameter vectors by sampling from the prior distribution following the methodology described in Section 2.2. These parameter vectors constitute the "uncalibrated parameters". The difference between each uncalibrated parameter set $\boldsymbol{p_u}$ and the calibrated parameter set is computed and projected into the parameter null space by multiplying with the null space projection matrix $\boldsymbol{V_2 V_2{}^T}$. This projected parameter set will lie in the parameter null space if the model were linear and if the null space was delineated accurately. The projected differences are added to the calibrated parameter set to create a new parameter set $\boldsymbol{p_{u\,-\,new}}$. This process is described by Eq. (12) where

$$\boldsymbol{p_{u-new} = p_c + V_2 V_2{}^T (p_u - p_c)} \tag{12}$$

Since the model is non-linear and there is uncertainty about the partition between the null and solution spaces, the parameter set from Eq. (12) does not often result in a calibrated model. Hence this parameter set is further updated using PEST (Doherty, 2010) until the measurement mismatch is acceptable.

3.3. Posterior Covariance Calculation

In this method, the posterior covariance matrix $\boldsymbol{C'}$ is estimated from the prior covariance matrix under the assumption of linearity (Tarantola, 2005).

$$\boldsymbol{C' = C_p - C_p X^T \left[X C_p X^T + C_D\right]^{-1} X C_p} \tag{13}$$

In Eq. (13), $\boldsymbol{C_p}$ is the prior covariance matrix. The second term on the right-hand side represents the impact of calibrating the model. The term $\boldsymbol{C_D}$ represents the covariance of the measurement errors. The matrix $\mathbf{X}$ represents the Jacobian matrix of the calibrated model. After calculating $\boldsymbol{C'}$, several parameter sets are randomly generated using a random parameter generator as described in Section 2.2. If the model were perfectly linear, each of these parameter sets would reproduce the observed data. An inspection of the likelihood functions revealed that it is not the case. Hence this parameter set is further updated using PEST (Doherty, 2010) until the measurement mismatch is acceptable.

4. Synthetic problem

A synthetic problem derived from Mariethoz et al. (Mariethoz et al., 2010) is analyzed in this paper. A constant discharge pump test is conducted in a square-shaped (100m × 100m) confined aquifer. The pumping well extracts 0.003 m^3/s from the center of the aquifer. The aquifer contains high-permeability fluvial channels embedded in a low-permeability matrix. Groundwater flow in the aquifer is two-dimensional flowing from left to right. A constant head boundary of 1m is located on the left edge, and a constant head boundary of 0m is located along the right edge. 12 monitoring wells are located around the pumping well. The aquifer schematic, boundary conditions, and well locations are shown in Fig. 4. Aquifer heads are recorded at the pumping and monitoring wells once the system reaches steady-state. The model representing this synthetic reality is referred to as the 'reference model'.

The facies distribution was developed following the approach described in Section 2.2. We used the training image (TI) introduced in Fig. 1. It represents channels and matrix in a 2500m × 2500m area. A large number (30,000) of sub-images were extracted from this TI. After visually inspecting a subset of these images, it was determined that there are typically three fluvial channels of width 13m in a 100m × 100m area. The covariance matrix $\boldsymbol{C_p}$ of the Y coordinates along the channel centerline was estimated according to the methodology described in Section 2.2. 15 points – 5 for each channel, were used to track the channels. The matrix scatter plot of the Y coordinates shown in Fig. 5 depicts the correlation between the various coordinates. In this plot, the variables y11 to y15 represent the coordinates of the top channel in a left to right direction. The variables y21 to y25 represent the coordinates of the middle channel in a left to right direction and the variables y31 to y35 represent the coordinates of the bottom channel in a left to right direction. The off-diagonal plots show the scatter between two coordinates and the diagonal plots show the histogram of a single coordinate. The plot shows that each point is strongly correlated with its neighbors along the same channel and weakly correlated with points in the other channels.

The discrete Karhunen-Loève expansion, expressed in Eq. 2, was then used to generate 100,000 parameter realizations based on $\boldsymbol{C_p}$. For each realization, three cubic B-splines (27) were used to connect the channel coordinates. A buffer of width 6.5m around each of the splines was created to represent a channel of 13m width. These channels were overlaid on the model grid and model cells fully covered by the channels were

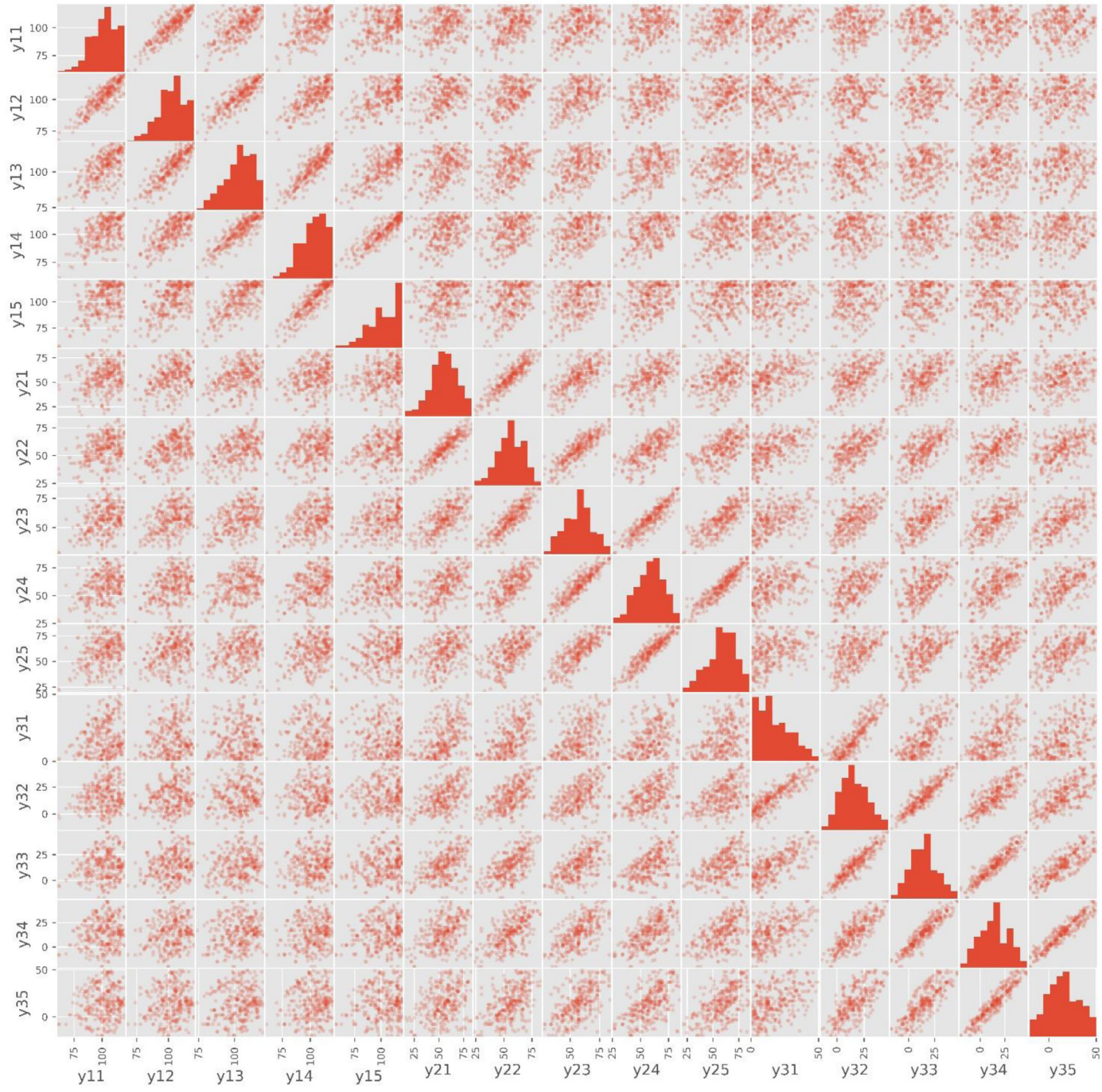

Fig. 5. Matrix Scatter Plot of the sampled Y Coordinates

assigned a hydraulic conductivity value randomly generated within a lognormal distribution with mean = -2 $\log_{10}$(m/s) and standard deviation = 0.1 $\log_{10}$(m/s). The remaining cells were assumed to be a part of the matrix and were assigned a hydraulic conductivity randomly generated based on a lognormal distribution with a mean = -4 $\log_{10}$(m/s), standard deviation = 0.1 $\log_{10}$(m/s).

A realization was randomly selected to represent the synthetic reality. For this selected realization, the hydraulic conductivity values of the channel and matrix were 8.7×10^{-3} m/s and 1.1×10^{-4} m/s respectively.

Reference head observations were obtained in the following manner. Steady-state groundwater flow was simulated for the aquifer described above using the USGS MODFLOW-NWT simulator (Niswonger et al., 2011). The facies distribution and the head distribution of the reference model are shown in Fig. 6. The calculated head distribution was sampled at the thirteen ((13)) observation wells. Normally distributed random noise (mean = 0 m, standard deviation = 0.05 m) was added to the sampled heads to simulate measurement error. These 13 adjusted heads constituted the reference head measurements.

5. Results

In this section, we generate an ensemble of conditional parameter realizations for the synthetic problem using the various methods described in Section 3 (RS, NSMC, and PC). For each ensemble, the cell-by-cell probability of finding a channel in the model domain, mean ensemble

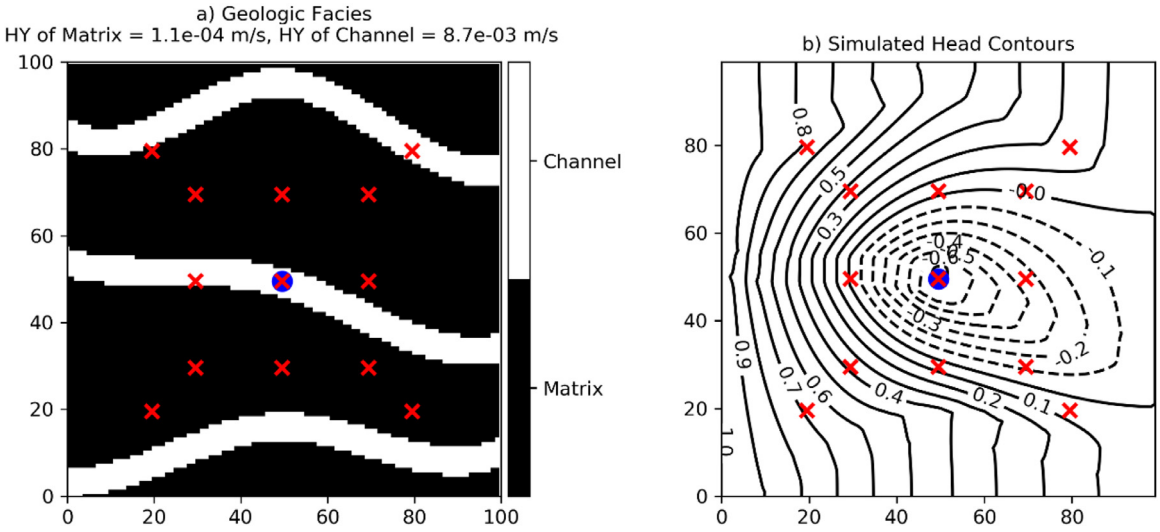

Fig. 6. Facies Distribution (a) and Head Distribution (b) of the Reference Model

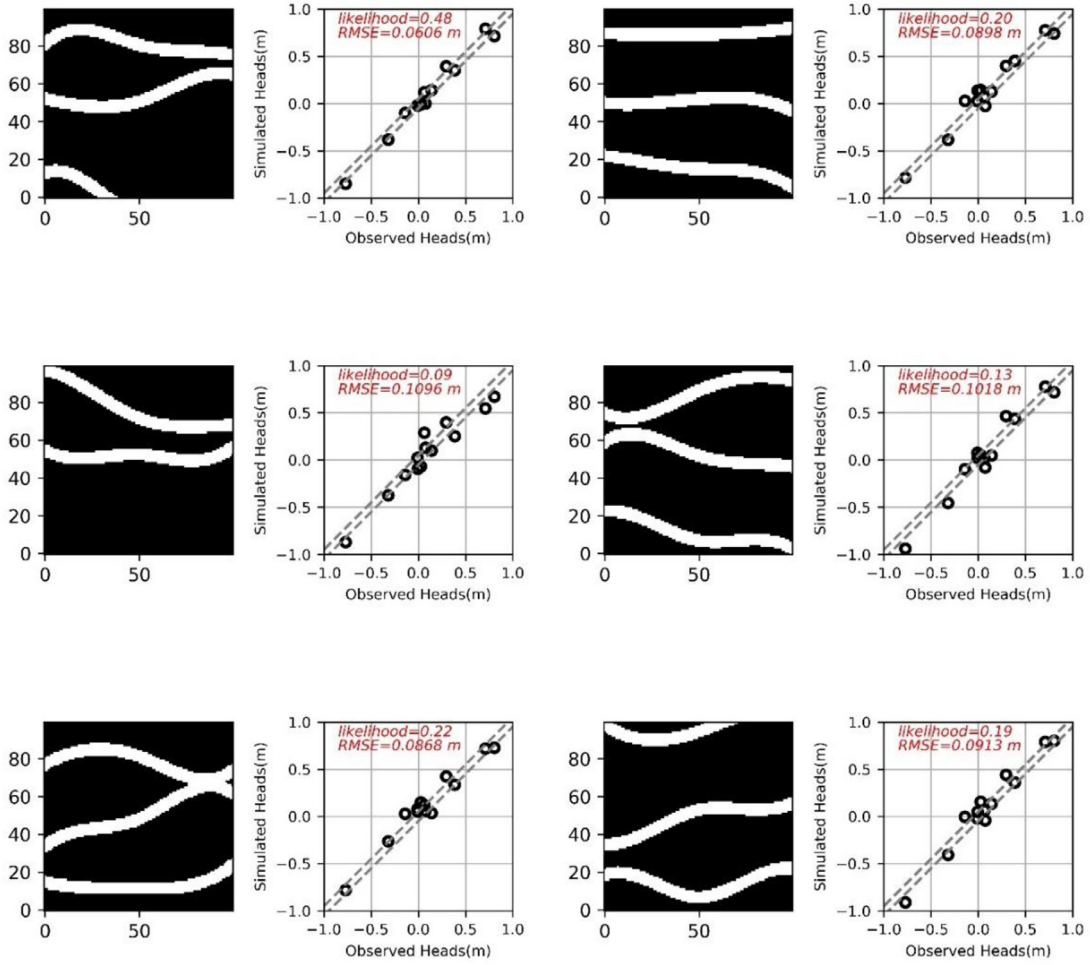

Fig. 7. Six randomly selected realizations from the 100 realizations chosen using Rejection Sampling. For each realization, the hydraulic conductivity distribution is shown on the left and the fit between the observed and simulated heads is shown on the right.

head distribution, and standard deviation of the simulated head distribution were calculated.

5.1. Rejection Sampling

A large set of parameter fields were generated based on the prior covariance matrix described in Section 4. Forward simulations were undertaken for each of these simulations, and the results evaluated under the rejection sampling methodology described in Section 3.1. We could obtain 100 models in the posterior distribution by evaluating 100,050 models. Six of these models were randomly selected and the hydraulic conductivity distributions and the corresponding fits between observed and simulated heads are presented in Fig. 7. These models demonstrate that several channel/hydraulic conductivity distributions can result in

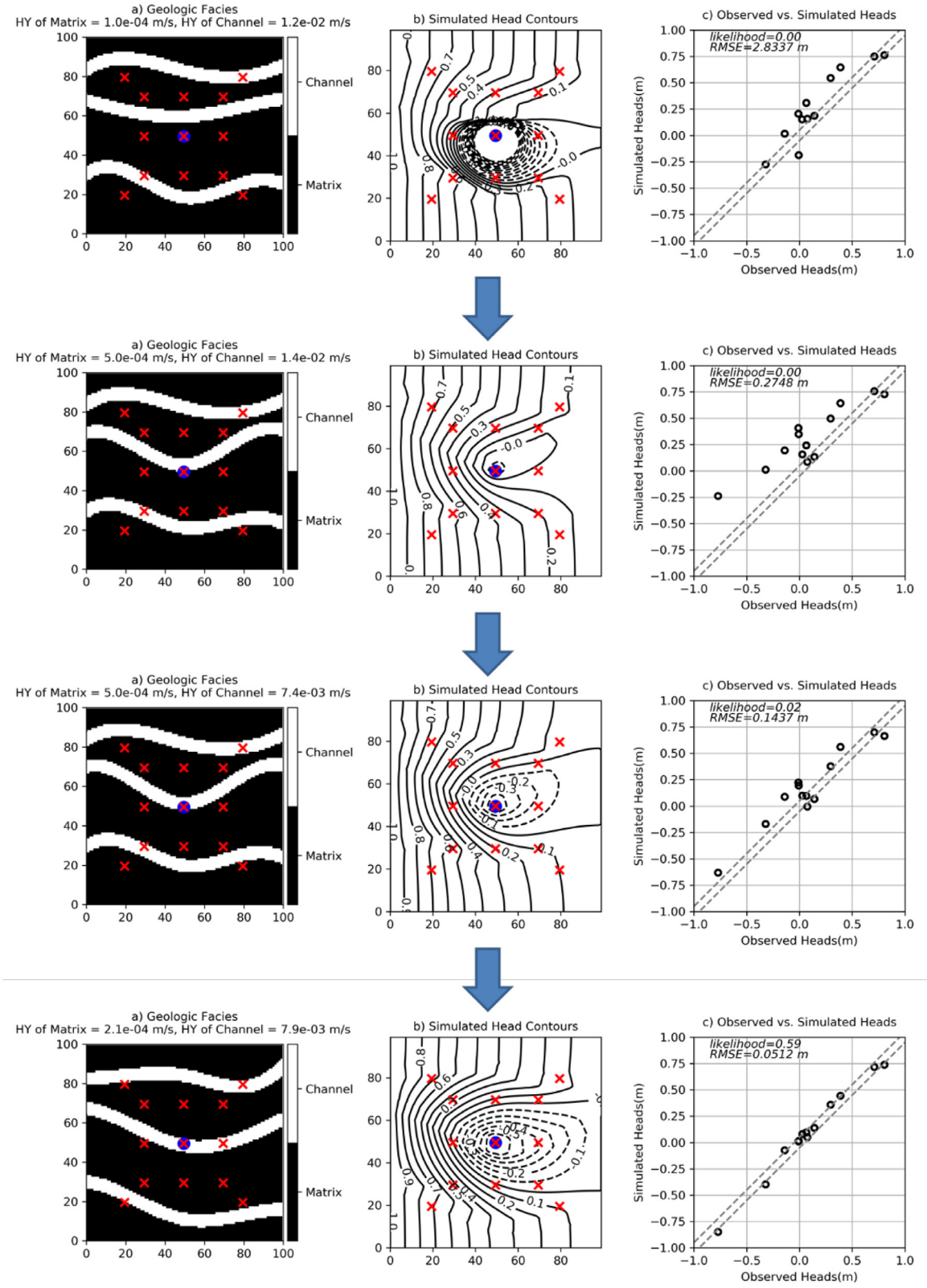

Fig. 8. Facies Distribution (a), Head Distribution (b), and Observed vs. simulated heads of the uncalibrated model (top panel), intermediate models (middle panels) and the calibrated model (bottom panel)

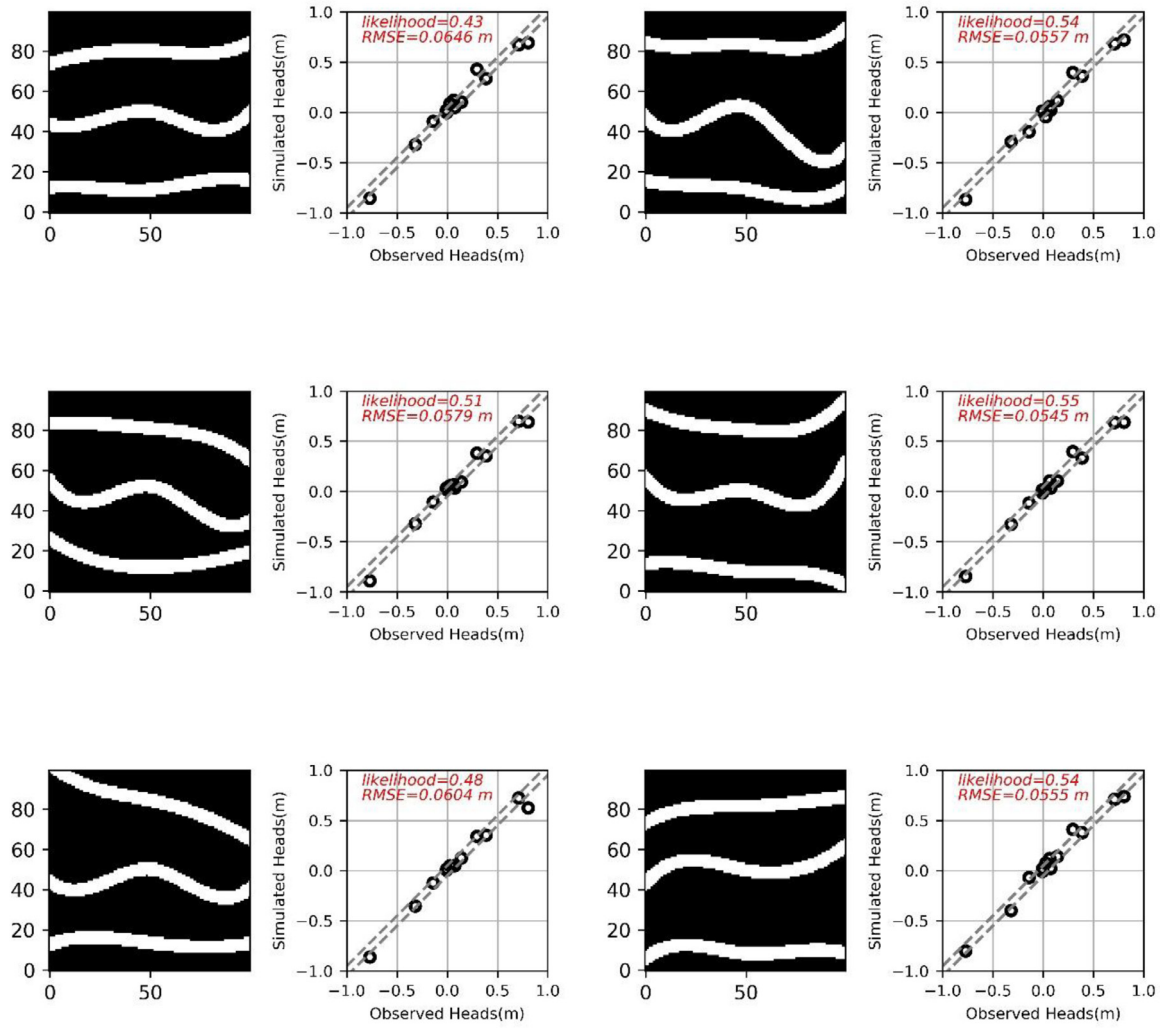

Fig. 9. Six randomly selected realizations from the 100 realizations generated using NSMC. For each realization, the hydraulic conductivity distribution is shown on the left and the fit between the observed and simulated heads is shown on the right.

reasonable matches to the measurements. Five of the six realizations have three channels whereas one realization has only two channels.

5.2. NSMC Method

The first step to sample a posterior distribution using the NSMC methods is to obtain a model corresponding to a maximum value of the likelihood (Section 3.2). Here, we describe how this model was obtained. A parameter set was first randomly generated from the prior covariance matrix $\boldsymbol{C_p}$. With the measurement and regularization constraints, parameter estimation was carried out by maximizing the likelihood function. The optimization was stopped when the modeled heads were considered acceptable. Facies distribution, head contours, measured vs. simulated heads for the initial, two intermediate models, and the calibrated model are presented in Fig. 8. This figure illustrates how the likelihood function increases as the central channel moves closer to the location of the pumping well (blue circle enclosing a red cross). A total of 484 groundwater flow model evaluations were required during this optimization to evaluate the misfit and the Jacobian. The resulting model is then used as the starting step for generating the ensemble.

The NSMC methodology described in Section 3.2 was used to generate 100 realizations. The computational cost for obtaining this ensemble was 12,754 forward model evaluations. Six of these models were randomly selected and the hydraulic conductivity distributions and the corresponding fits between observed and simulated heads are presented in Fig. 9. All the six realizations have three channels with the central channel exhibiting more curvature than the top/bottom channels.

5.3. PC Posterior Ensembles

The PC methodology described in Section 3.3 was used to generate another ensemble consisting of 100 realizations. This method is much more computationally efficient, since we generated 100 models with only 505 forward model evaluations. Six of these models were randomly selected and the hydraulic conductivity distributions and the corresponding fits between observed and simulated heads are presented in Fig. 10. All the six realizations have three channels.

5.4. Comparison of Posterior Ensembles

The characteristics of the parameter ensembles obtained using the various methods are presented in Fig. 11. Information is presented in a grid with four rows and four columns. Each row represents the characteristics of a parameter ensemble. In the first column of each row, cell-by-cell probabilities of finding a channel for that ensemble are depicted. Reddish colors imply a higher probability of finding a channel and bluish colors imply a lower probability. The probabilities for the prior distribution (first row) are low everywhere. When the head data

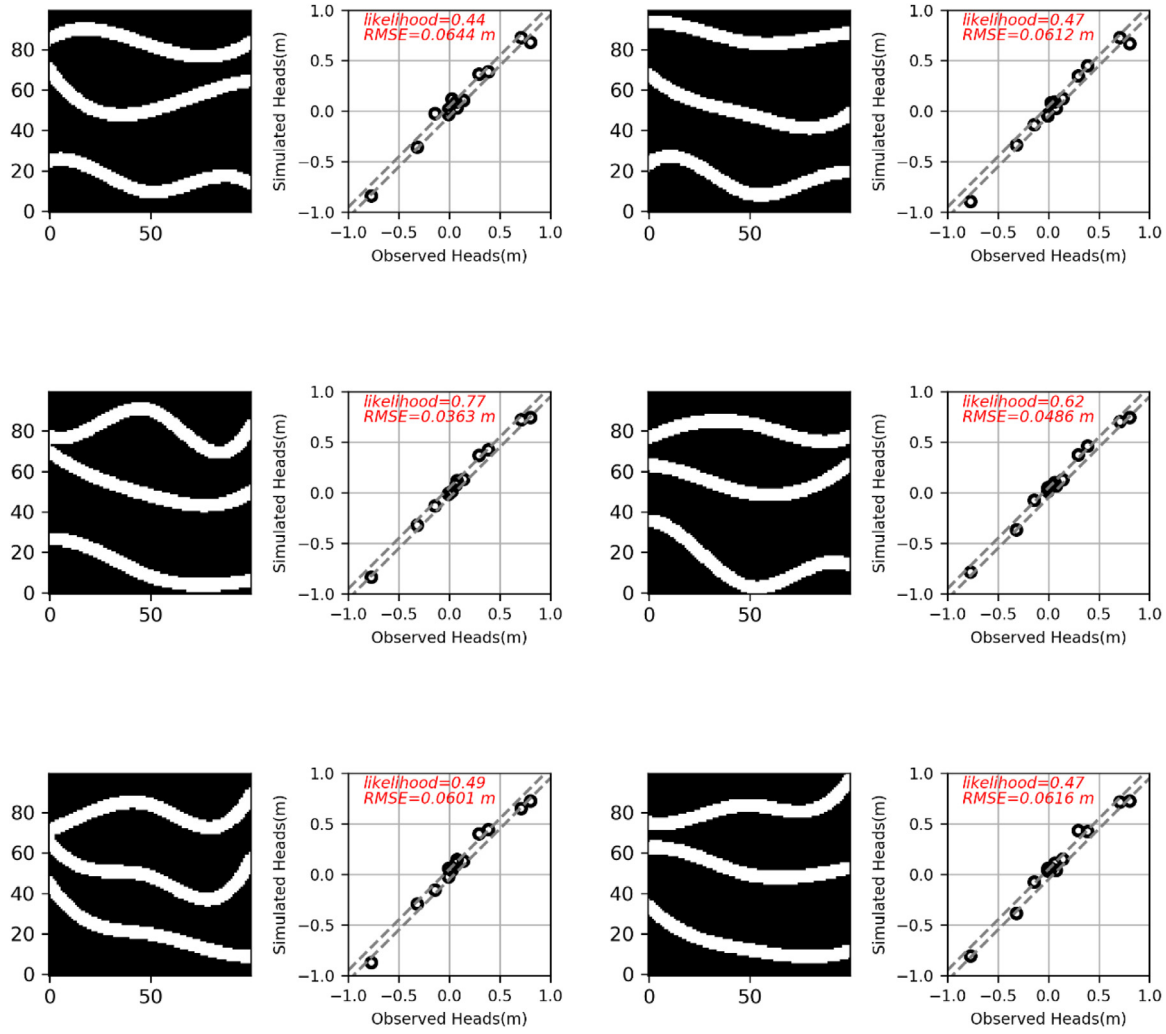

Fig. 10. Six randomly selected realizations from the 100 realizations chosen using the Posterior Covariance method. For each realization, the hydraulic conductivity distribution is shown on the left and the fit between the observed and simulated heads is shown on the right.

are not accounted for, the proposed geological model can place the channels anywhere in the domain in a uniform manner.

All the methods that are conditioned by the head data show that they can locate the presence of a channel at the location of the pumping well with a high probability. They also indicate a high probability to find matrix above and below the pumping well as well as 2 other channels in the top and bottom of the field. However, there are some differences in the values of the probabilities when moving away from the location of the pumping well.

The ensemble obtained with RS is unbiased and considered as the reference in this experiment. The RS and the NSMC ensembles exhibit more similarities than RS and PC. The PC method shows higher values for the probabilities than the RS and NSMC methods. It means that the subspace methods did not capture the complete variability of the posterior ensemble. They are much more efficiently numerically but this comes at the cost of an underestimation of the uncertainty. Some channel configurations that can reproduce the data and belong to the prior geological model are not identified in that case.

Histograms for the log-transformed hydraulic conductivities presented in the second column exhibit a bimodal distribution. The channel hydraulic conductivities are shown in red and the matrix hydraulic conductivities are in blue. The hydraulic conductivities of the reference model are shown as black dots on the histogram. The matrix hydraulic conductivities vary more than their channel counterparts. As with the ensemble probability, the PC method has narrower histograms implying a lack of variability. We also see on these graphs that all methods identify properly the hydraulic conductivity of the channels, while the matrix conductivity may be overestimated as compared to the reference by the NSMC and PC methods. A possible reason for this overestimation could be to compensate for deviations from the geometry in the reference field for maintaining the observed gradients.

Mean and standard deviation of the ensemble head distributions are in the third and fourth columns. The prior mean and variance are showing symmetry around the pumping well. The uncertainty on the head value is high as shown by the high values of the standard deviation. The effect of conditioning to the head values reduces significantly the uncertainty for the three ensembles. The ensemble mean head from the NSMC method resembles the ensemble mean head from the RS method and the head from the reference model (Fig. 6). The mean head estimated with the PC ensemble does not show as clearly as the two other methods the shift of the cone of depression toward the bottom of the image. Furthermore, in terms of standard deviation and uncertainty estimation the PC method has a very low ensemble standard deviation lending further credence to the lack of variability in realizations. As compared to the reference method (RS), the NSMC is closer to it but still underestimates the variability especially in the upper part of the domain.

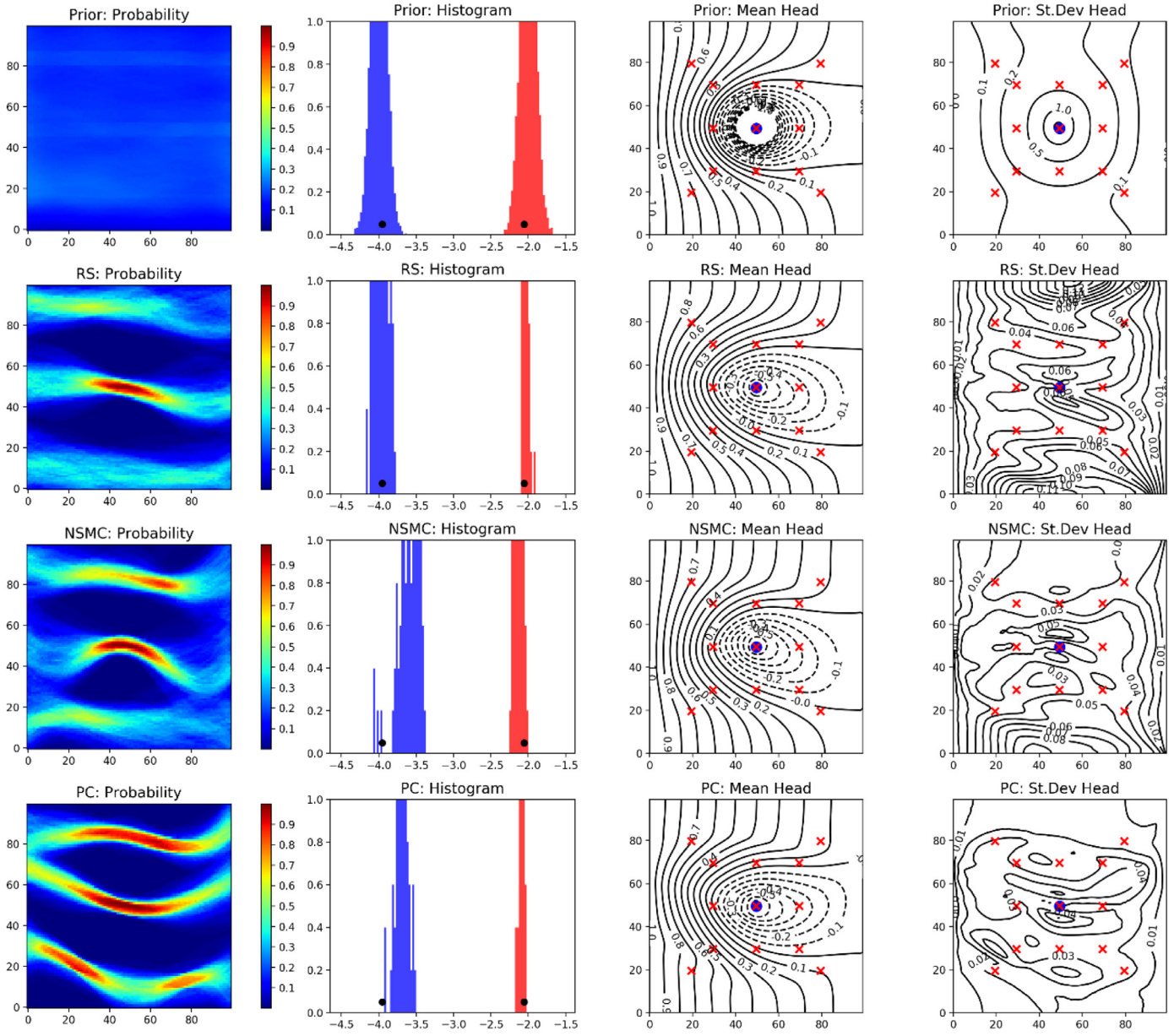

Fig. 11. Parameter ensemble characteristics from various methods, Prior (top row), Rejection sampling (second row), NSMC (third row) and PC (bottom row). In each row, ensemble parameter probability, histogram of ensemble conductivities, mean and standard deviation of simulated heads are shown.

6. Summary and Discussion

In this paper, we propose a new approach involving Traveling Pilot Points (TRIPS) and linear subspace methods to solve the categorical inverse problem in a probabilistic framework. We summarize some of the main findings below.

The first key proposition that we make in this paper is in letting pilot points travel and estimating both the locations of the channels and associated properties like hydraulic conductivity. We then propose to estimate the prior covariance matrix of the position of the pilot points from a training image. The advantage of that approach is in its simplicity. The user can provide an image of the type of channels that they want to model, and the covariance will be inferred directly. If the training image is too small, it is possible to use a multiple-point statistics simulation algorithm and generate an ensemble of simulations and derive the covariance matrix from the analysis of this simulation in the same manner as we analyze the sub-images in this work. In addition, we use first-order (difference) regularization constraints to preserve the curvature of the channels in the inversion process.

The proposed TRIPS parametrization was integrated in an optimization framework based on linear subspace methods allowing to obtain solutions of the categorical inverse problem for a synthetic aquifer. A posterior ensemble obtained with the rejection sampling method was considered to represent the reference solution and compared against the Null Space Monte Carlo (NSMC) and Posterior Covariance (PC) ensembles. The comparisons indicate that these parameter ensembles exhibit similarities with the reference distribution. The PC method was much more efficient in estimating members of the posterior ensemble. However, the variability was underestimated (Fig. 11, bottom row). The NSMC method was comparably slower because more model evaluations were required. However, the ensemble probability estimated by this method is closer to the ensemble from RS. The NSMC method provides a balance between computational efficiency and representation of the posterior ensemble. The number of model evaluations required by the NSMC method were comparable to stochastic approaches like ISR

(Mariethoz et al., 2010) and POPEX (Jaggli et al., 2017), while the PC method is much faster.

Overall, we believe the TRIPS methodology to be a promising entrant in the field of categorical inversion. While the example problem presented in this paper considers only two-dimensional channels traversing the domain, the methodology can be extended to real-world three-dimensional datasets with a larger number of facies and more complex geometries. For example, the TRIPS method could be used to estimate the complex channel framework at a real site such as the one discussed by Ronayne et al (Ronayne et al., 2008). For this case, the pilot points would represent the positions of channels in three dimensions and three-dimensional splines passing through the pilot points could be used to delineate the channels.

More generally, the extension of the proposed methodology is straightforward for all object-oriented geological modeling techniques (Pyrcz and Deutsch, 2014) since the positions of the objects are controlled in these models by seed points which can be considered as Traveling Pilot Points. The prior statistics on the number of objects and relative locations of these points can be derived from a set of initial simulations. The proposed algorithm described could then be used to update these locations and solve the inverse problem. The shape parameters concerning the three-dimensional size and orientation of the objects can be handled as well easily since these are continuous parameters that an inversion code like PEST can optimize. This step would be analogous to the identification of the hydraulic conductivity values within the channels as illustrated in the example treated in this paper. For objects having a flexible shape such as channels with varying width, traditional techniques such as the standard pilot points can be coupled with TRIPS: one can attach a width parameter to every traveling pilot point and interpolate the width along the channel length and update these parameters during the inversion.

The TRIPS method could also be used with pixel based geostatistical methods such as plurigaussian or MPS simulations (Pyrcz and Deutsch, 2014). Starting from one or a set of initial realizations, we could extract a set of conditioning locations and the corresponding categories from the realization. TRIPS would then proceed by moving the locations of these points, keeping the value of the categories and simulating again the complete field using these new conditioning data as input in the geostatistical algorithm. In this last case, developing the appropriate parameter covariance matrix remains a challenge.

Considering the subspace methods tested in this paper, one aspect which works to their favor is the large null space. As the size of the null space increases, TRIPS/NSMC methods could prove to be computationally parsimonious in comparison with other methods. We have shown in this paper, on a simple synthetic problem, that the gain in numerical efficiency comes at the cost of an underestimation of the overall uncertainty .

Declaration of Competing Interests

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

CRediT authorship contribution statement

Prashanth Khambhammettu: Conceptualization, Methodology, Software, Visualization, Validation, Writing - original draft, Writing - review & editing. **Philippe Renard:** Conceptualization, Methodology, Supervision, Validation, Writing - review & editing. **John Doherty:** Conceptualization, Methodology, Supervision, Validation, Writing - review & editing.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. The authors wish to thank the editor and the three anonymous reviewers whose comments contributed to an improved version of this paper. The authors are also thankful to Vivek Bedekar, Julien Straubhaar, and Christoph Jäggli for reviewing and commenting upon early drafts of this manuscript. The data and codes used to generate results for this paper can be obtained from the corresponding author (prashanth.khambhammettu@arcadis.com).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.advwatres.2020.103556.

Appendix A

A.1 Calculation of a covariance matrix for parameter differences

Given the covariance matrix $\boldsymbol{C}$ for a parameter vector $\boldsymbol{p}$, calculation of the covariance matrix $\boldsymbol{C}(\boldsymbol{p}-)$ for the parameter difference vector $\boldsymbol{p}-$ is described in this section with an example. If the parameter vector $\boldsymbol{p}$ had three parameters $y1$, $y2$, and $y3$, the parameter difference vector $\boldsymbol{p}-$ would have the differences $y1-y2$, $y1-y3$, *and* $y1-y2$. In matrix form, this relationship can be expressed by the equation

$$\begin{bmatrix} y1-y2 \\ y1-y3 \\ y2-y3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} \tag{1}$$

The above equation could be generalized for an arbitrary number of parameters/parameter differences by the equation

$$\boldsymbol{p}- = \boldsymbol{A}\boldsymbol{p} \tag{2}$$

If we have $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, Aster et al. (Aster et al., 2013) state that the covariance matrix of $\boldsymbol{y}$, $\boldsymbol{C}(\boldsymbol{y})$ can be calculated by the equation

$$\boldsymbol{C}(\boldsymbol{y}) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{x})\boldsymbol{A}^T \tag{3}$$

Combining Eqs. (2) and (3), the covariance matrix for the parameter difference vector can be calculated by the equation

$$\boldsymbol{C}(\boldsymbol{p}-) = \boldsymbol{A}\boldsymbol{C}(\boldsymbol{p})\boldsymbol{A}^T \tag{4}$$

References

Alcolea, A., Renard, P., 2010. Blocking Moving Window algorithm: Conditioning multiple-point simulations to hydrogeological data. Water Res. Res. 46 W08511.
Aster, R.C., Borchers, B., Thurber, C., 2013. Parameter Estimation and Inverse Problems, Second Edition Elsevier.
Caers, J., Hoffman, T., 2006. The Probability Perturbation Method: A New Look at Bayesian Inverse Modeling. Math. Geol. 38 (1), 81–100.
Certes, C., de Marsily, G., 1991. Application of the pilot-points method to the identification of aquifer transmissivities. Adv. Water Res. 14 (5), 284–300.
J. E. Doherty, M. N. Fienen and R. J. Hunt, "Approaches to Highly Parameterized Inversion: Pilot-Point Theory, Guidelines, and Research Directions, Scientific Investigations Report 2010-568," United States Geological Survey, 2010.
Doherty, J., 2003. Groundwater model calibration using pilot-points and regularization. Ground Water 41 (2), 170–177.
Doherty, J., 2010. PEST User's Manual, 5th Edition Watermark Numerical Computing, Brisbane, Australia.
Hansen, M.T., Cordua, K.S., Mosegaard, K., 2012. Inverse problems with non-trivial priors: efficient solution through sequential Gibbs sampling. Comput. Geosci. 16 (3), 593–611.
Herckenrath, D., Langevin, C.D., Doherty, J., 2011. Predictive uncertainty analysis of a saltwater intrusion model using null-space Monte Carlo. Water Res. Res. 47 (5).
Hu, L.Y., Blanc, G., Noetinger, B., 2001. Gradual deformation and iterative calibration of sequential simulations. Math. Geol. 33 (4), 475–489.
Hu, L., 2000. Gradual deformation and iterative calibration of Gaussian-related stochastic models. Math. Geol. 32 (1), 87–108.
Jäggli, C., Straubhaar, J., Renard, P., 2018. Parallelized Adaptive Importance Sampling for Solving Inverse Problems. Front. Earth Sci. 6, 15.
Jaggli, C., Straubhaar, J., Renard, P., 2017. Posterior population expansion for solving inverse problems. Water Res. Res. 53, 2902–2916.

Keating, E.H., Doherty, J., Vrugt, J.A., Kang, Q., 2010. Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality. Water Res. Res. 46 (10).

Laloy, E., Hérault, R., Jacques, D., Linde, N., 2018. Training-Image Based Geostatistical Inversion Using a Spatial Generative Adversarial Neural Network. Water Res. Res. 54 (1), 381–406.

LaVenue, A., de Marsily, G., 2001. Three-dimensional interference test interpretation in a fractured aquifer using the pilot-point inverse method. Water Res. Res. 37 (11), 2659–2675.

Li, L., Srinivasan, S., Zhou, H., Gomez-Hernandez, J., 2003. A pilot point guided pattern matching approach to integrate dynamic data into geological modeling. Adv. Water Res. 62, 125–138.

Linde, N., Renard, P., Mukherji, T., Caers, J., 2015. Geological realism in hydrogeological and geophysical inverse modeling. Adv. Water Res. 86–101.

Mariethoz, G., Renard, P., Caers, J., 2010. Bayesian inverse problem and optimization with iterative spatial resampling. Water Res. Res. 46 (11).

Moore, C., Doherty, J., 2005. Role of the calibration process in reducing model predictive error. Water Res. Res. 41 (5).

R. G. Niswonger, S. Panday and M. Ibaraki, "MODFLOW-NWT, A Newton formulation for MODFLOW-2005: U.S. Geological Survey Techniques and Methods 6–A37, 44 p.," U.S. Geological Survey, 2011.

Pyrcz, M.J., Deutsch, C.V., 2014. Geostatistical Reservoir Modeling. Oxford University Press.

Ronayne, M.J., Gorelick, S.M., Caers, J., 2008. Identifying discrete geologic structures that produce anomalous hydraulic response: An inverse modeling approach. Water Res. Res. 44 W08426.

Sarma, P., Durlofsky, L.J., Aziz, K., 2008. Kernel Principal Component Analysis for Efficient, Differentiable Parameterization of Multipoint Geostatistics. Math. Geosci. 40 (1), 3–32.

Scipy B-spline, [Online]. Available: https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.interpolate.splrep.html.

Tarantola, A., 2005. Inverse Problem Theory and Methods for Model Parameter Estimation. Paris.

Tonkin, M.J., Doherty, J., 2005. A hybrid regularized inversion methodology for highly parameterized environmental models. Water Res. Res. 41 (10), 16.

Tonkin, M., Doherty, J., 2008. Calibration-constrained Monte Carlo analysis of highly-parameterized models using subspace techniques. Water Res. Res. 45 W00B10.

Zhou, H., Gómez-Hernández, J.J., Li, L., 2014. Inverse methods in hydrogeology: Evolution and recent trends. Adv. Water Res. 63.