A Local-Global Pattern Matching Method for Subsurface Stochastic Inverse Modeling

Liangping Li[a,*], Sanjay Srinivasan[a], Haiyan Zhou[a], J. Jaime Gómez-Hernández[b]

[a] *Center for Petroleum and Geosystems Engineering Research, University of Texas at Austin, 78712, Austin, USA*
[b] *Research Institute of Water and Environmental Engineering, Universitat Politècnica de València, 46022, Valencia, Spain*

Abstract

Inverse modeling is an essential step for reliable modeling of subsurface flow and transport, which is important for groundwater resource management and aquifer remediation. Multiple-point statistics (MPS) based reservoir modeling algorithms, beyond traditional two-point statistics-based methods, offer an alternative to simulate complex geological features and patterns, conditioning to observed conductivity data. Parameter estimation, within the framework of MPS, for the characterization of conductivity fields using measured dynamic data such as piezometric head data, remains one of the most challenging tasks in geologic modeling. We propose a new local-global pattern matching method to integrate dynamic data into geological models. The local pattern is composed of conductivity and head values that are sampled from joint training images comprising of geological models and the corresponding simulated piezometric heads. Subsequently, a global constraint is enforced on the simulated geologic models in order to match the measured head data. The method is sequential in time, and as new piezometric head become available, the training images are updated for the purpose of reducing the computational cost of pattern matching. As a result, the final suite of models preserve the geologic features as well as match the dynamic data. This local-global pattern matching method is demonstrated for simulating a two-dimensional, bimodally-distributed heterogeneous conductivity field. The results indicate that the characterization of conductivity as well as flow and transport predictions are improved when the piezometric head data are integrated into the geological modeling.

Keywords: multiple-point geostatistics, conditional simulation, inverse modeling, global matching, uncertainty assessment

---

*Corresponding author
Email addresses: liangpingli@utexas.edu (Liangping Li), sanjay.srinivasan@engr.utexas.edu (Sanjay Srinivasan), haiyanzhou@utexas.edu (Haiyan Zhou), jaime@dihma.upv.es (J. Jaime Gómez-Hernández)

1. Introduction

Inverse modeling is a mathematical approach to identify parameters such as permeability or hydraulic conductivity at unsampled locations such that flow and transport modeling using the estimated parameters match observed state variables such as piezometric head or concentration data. Predictions for groundwater flow and solute transport made using the estimated parameters would then be more accurate. The fact that the number of observed state variables is much smaller than the number of unknown parameters implies that the solution of inverse problem will be non-unique (Carrera and Neuman, 1986) especially when heterogeneous subsurface systems are considered. In order to represent this non-uniqueness, stochastic inverse modeling seeks to generate multiple likely representations of parameter fields that are all conditioned to both direct measurements of the parameters at specific locations and dynamic data (Gómez-Hernández et al., 1997). The multiple calibrated models obtained by applying stochastic inversion methods could be used to assess the uncertainty in predictions based on the available data. Reliable models for uncertainty are required by decision-makers For a review of the evolution and recent trends of inverse methods in hydrogeology, the reader is referred to Zhou et al. (2014).

In cross-bedded aquifers or fluvial geologies, aquifer properties such as hydraulic conductivity exhibit connectivity along curvilinear paths. This complex connectivity significantly affects the flow and transport of fluids and chemical species (Gómez-Hernández and Wen, 1998; Renard and Allard, 2011). Reproduction of the curvilinear geometry can be achieved using Multiple-Point Statistics (MPS) based stochastic simulation methods (Strebelle, 2002). MPS simulation was developed to overcome the limitation of traditional two-point variogram-based methods, which cannot capture strong connectivities in the subsurface aquifer. The higher moments (i.e., multiple-point statistics) are introduced into the simulation by borrowing patterns from a training image (Guardiano and Srivastava, 1993). Although MPS provides an avenue to simulate complex formations, stochastic inverse modeling within the framework of MPS simulations is extremely challenging because of the difficulty in maintaining the complex curvilinear connectivity geological structures while simultaneously honoring dynamic data that are related to conductivity through a strongly non-linear transfer function.

In the literature, stochastic inverse methods can be classified into two groups. In the first group, an objective function is first constructed based on the discrepancy between observed data and simulated values. This objective function is subsequently minimized by iteratively perturbing the parameter values until a sufficiently close match is attained. Preservation of the prior geological structures is not explicitly considered during this process of optimization. Examples of this data-driven stochastic inverse method are sequential self-calibration (Gómez-Hernández et al., 1997; Hendricks Franssen et al., 2003), the pilot-point method (de Marsily, 1978) and the ensemble Kalman filter (EnKF) (Evensen, 2003). It has been proven that these methods yield optimal estimates for multiGaussian conductivity fields. Some variants were proposed to handle non-multiGaussian conductivity fields. For example, Capilla et al. (1999) proposed the application of self-calibration method to local conditional probabilities defining the uncertainty in conductivity, instead of calibrating the conductivities directly. Later, Capilla and Llopis-Albert (2009) coupled the gradual deformation method and the optimization of the probability fields in order to improve the efficiency of the previous proposal. In a similar way, Hu et al. (2013) proposed to consider the uniform random number used to draw the MPS realizations as part of the state variable set in EnKF. Sun et al. (2009) coupled Gaussian mixture models and EnKF to handle non-Gaussian conductivity fields. Jafarpour and Khodabakhshi (2011) proposed to first update the ensemble of MPS-generated conductivities to derive local probabilities, and then, to re-simulate the conductivities using the probability maps as soft data. Zhou et al. (2011) developed a normal-score EnKF to handle non-Gaussianity within the ensemble Kalman filtering framework.

In the second group of inverse modeling approaches, data integration is achieved using Bayes' theorem. The posterior models are sampled from the prior models by assessing first a likelihood function. A typical example of this model-driven stochastic inverse method is rejection sampling (Tarantola, 2005). The likelihood of a model sampled from a prior set is assessed, and the model is rejected depending on a likelihood threshold. The prior geological structures will be preserved in this process, because the posterior set of models is simply a subset of the prior set. However, like the particle filtering approach, this method is computationally expensive and is inapplicable in most practical cases because tens of thousands of models need to be evaluated. To improve the computational efficiency, Mariethoz et al. (2010a) proposed an iterative spatial resampling method in which the candidate models are generated by conditioning to data sampled from previous accepted models, thus resulting in less computational cost because of faster convergence to a posterior set that exhibits the desired dynamic characteristics. Another popular Bayesian approach to inverse modeling is the Markov chain Monte Carlo method (McMC) (Metropolis et al., 1953; Oliver et al., 1997) in which the parameter model is first locally perturbed for a gridblock or for a set of gridblocks (i.e., the transition kernel) and then the forecast model is run to judge whether the new candidate model will be accepted (e.g., the Metropolis-Hastings rule). The problems with these McMC methods are: (1) the acceptance rate of new models is dependent on the transition kernel used; (2) a long chain is usually required before the posterior distribution can be correctly sampled, and (3) a large number of perturbed models have to be generated and evaluated. An extensive description of the mathematical framework for the McMC method and recent advances can be found in the review paper by Liu et al. (2010).

The Ensemble PATtern matching (EnPAT) stochastic inverse method was first proposed by Zhou et al. (2012) with the aim to create multiple conductivity fields honoring both measured conductivity and piezometric head data as well as the prior geological structures. The EnPAT is inspired by the Direct Sampling (DS) MPS method developed by Mariethoz et al. (2010b). In DS, the conductivity patterns are directly sampled from a training image without storing the entire pattern database in memory. This results in fast simulation and the possibility to simulate continuous variables such as hydraulic conductivity. Zhou et al. (2012) borrows the concept of DS and expands the conductivity pattern to include the pattern of piezometric heads for the purpose of inverse modeling. Correspondingly, multiple MPS-simulated conductivity models and the corresponding head models obtained by running the forward simulator are jointly used as the training images for learning during the simulation. Conductivities are simulated by matching joint patterns from the training image sets. As a result, the simulated conductivity models are not only conditioned to the measured conductivity and piezometric data, but also preserve the prior geological structures. Li et al. (2013a) developed a hybrid of the EnPAT and the pilot point/self-calibration method (Gómez-Hernández et al., 1997) to reduce the computational cost and to improve the characterization of conductivity connectivity during the dynamic data assimilation process.

In this paper, we propose a local-global pattern matching method to integrate dynamic data into geologic models. In the previous implementation of the EnPAT, a local pattern is considered for ensemble matching, but that does not guarantee that the updated model matches the observed global dynamic data because of the non-linearity of the forecast function as well as the existence of complex boundary conditions. To address this issue, we implement an additional step in which we simulate the global response of the updated models and select those that best fit the observed data after the process of local pattern matching. As a consequence, updated models will preserve the geological structures and the dynamic data, although at a computational cost because of the additional forward simulations in the rejected models. In order to mitigate the computational demand and to accelerate the learning process, the training image sets are refined by progressively replacing the worst models in the prior training set with the newly accepted models. The method therefore borrows the concept of iterative resampling proposed by Mariethoz et al. (2010a). A ranking scheme is implemented to identify the poor initial models. The proposed methodology is demonstrated on a synthetic example for which predictions of flow and transport are considered.

The remainder of the paper will be organized as follows. In section 2, the implementation of the ensemble pattern matching method is described, with emphasis on the significance of global constraints on the predictions of flow and transport. In section 3, a synthetic example is used to demonstrate the effectiveness of the proposed method. Then, in section 4, we discussed the computational efficiency of the EnPAT by continuously refining the training images. In section 5, there is a general discussion. The paper ends with a summary and conclusions.

2. Methodology

In the EnPAT method two steps are performed at each time step: the forecast step (i.e., solving the flow equation based on the current hydraulic conductivities to derive the piezometric head) and updating step (i.e., updating both conductivity and head through a pattern matching approach).

During the updating step, patterns are constructed for the updating of each gridblock in each realization by searching within a predefined search neighborhood for static parameter values such as conductivities and dynamic variable values such as heads. Suppose that for the updating of the conductivity and the head at gridblock i and realization j for time t, we have found conductivities ($K = k_1, k_2, \cdots, k_n$) and heads ($H = h_1, h_2, \cdots, h_m$). Denote this as the conditioning pattern ($\mathbf{P}_{t,j,i}$) (see Fig 1),

$$\mathbf{P}_{t,j,i} = \begin{bmatrix} K \\ H \end{bmatrix}_{t,j,i} \tag{1}$$

within the context of sequential simulation (see, for instance, Gómez-Hernández and Journel (1993)) the conductivity and head components in the pattern can be the observed data and/or the previously estimated values. The number of conductivity (n) and head data (m) in the pattern must be less than a maximum conditioning data specified by the user and fall within a predefined maximum search radius around the simulation node. The conditioning pattern is dependent on the location of the gridlock, the particular stage within the simulation, and the time step. EnPAT extends traditional MPS method in two important ways, first, the patterns contain not only the parameters such as conductivities, but also state variables, and second, an ensemble of joint training images is used. When head data are included in the pattern, the MPS method becomes multi-variable co-simulation. In other words, the simulated conductivity is constrained by the surrounding conductivities and heads, and thus the multipoint cross-correlation between both variables can be preserved in the simulation.

Pattern matching is initialized by generating an ensemble of prior conductivity fields and the corresponding ensemble of simulated heads. In this paper, the initial ensemble of conductivity fields is generated using the direct sampling MPS method, using a common training image for a fluvial aquifer. In the forecast step, the flow simulator is run for each conductivity realization until the time step for which new measured head data (h^t_{obs}) are available. The ensemble of head realizations obtained by running the forward model, plus the ensemble of conductivities will be used as the joint training images to update both conductivity and head, given any observed head data. The pattern matching scheme has the following steps:

- Build the conditioning pattern $\mathbf{P}$ at the first node to be simulated, conditioned to the measured n conductivity data and m piezometric head data.

- Search for candidate patterns in the joint training images. Calculate the distance between the candidate pattern $\hat{\mathbf{P}}$ found in the joint training images and the conditioning pattern around the simulation node $\mathbf{P}$ (see Fig 1). In this research, distance is measured by computing a weighted Euclidean distance function:

$$d_{(\mathbf{P},\hat{\mathbf{P}})} = \left[\frac{1}{\sum_{i=1}^{p} h_i^{-1}} \sum_{i=1}^{p} h_i^{-1} \frac{(\mathbf{P} - \hat{\mathbf{P}})^2}{d_{max}^2} \right]^{1/2} \tag{2}$$

  where p is the number of data in the pattern; h is the Euclidean distance between the gridblock to be simulated and the conditioning data; and d_{max} is the maximum absolute difference of conductivities or heads observed in the pattern. The standardized distance between the candidate and conditional patterns lies within the range of 0 to 1. The searching process for candidate patterns is limited to a small area around the gridblock to be estimated because the calculated head depends on boundary conditions and sources. If the search radius is specified to be large then the influence of global boundary conditions becomes more pronounced.

- If the resulting distances, computed independently for conductivities and for heads are both smaller than predefined threshold values ($d^k_{(\mathbf{P},\hat{\mathbf{P}})} < \zeta_k$ and $d^h_{(\mathbf{P},\hat{\mathbf{P}})} < \zeta_h$), the conductivity value and the piezometric head value of the matching pattern at the location of the simulation node is retained; if no pattern is found meeting these criteria, the values from the closest pattern are retained (see Fig 1). The retained conductivity and head values become conditioning data for the simulation of the next gridblocks.

- Repeat the three previous steps until all nodes in the domain are simulated.

The simulated conductivity and head values represent a realization of the updated conductivity and head field, conditioned to the measured conductivity and piezometric head data. Multiple realizations of updated conductivity and head values can be obtained by visiting the unknown gridblocks along different random paths. Furthermore, as more observation data become available at the next time step, the simulated models at the previous time get updated.

The EnPAT procedure can be combined with the use of pilot points to reduce the computational cost. More specifically, the pattern search step is only applied to a set of predefined pilot point locations, and then traditional MPS is used to complete the non-simulated gridblocks. This variation of the EnPAT has been implemented in the study by Li et al. (2013a).

The EnPAT algorithm can use any flow simulator as a black box because only the outputs, such as piezometric heads, are required. Compared to most of the traditional gradient-based stochastic inverse methods that require access to quantities such as Jacobian of the flow model (i.e., the self calibration method (Gómez-Hernández et al., 1997)), the coding of the EnPAT is much simpler (Li et al., 2013b). Another advantage of the EnPAT is that, like the EnKF, the updated ensemble of conductivities can be used to assess the residual uncertainty associated with predictions. In addition, the updated conductivity fields preserve the curvilinear geometry exhibited by the training image. In fluvial deposits, for instance, it is very significant to preserve the connectivity of the geology in order to make accurate predictions of flow and transport (Gómez-Hernández and Wen, 1998).

However, there is a potential problem in the updating of conductivity and head values using the pattern matching approach. The updated conductivity might be inconsistent with the updated piezometric head because the transfer function relating the two is not explicitly accounted for. In other words, the simultaneously simulated conductivity and head might not honor Darcy's law (i.e., the mass balance equation). To handle this problem, in this paper, a global constraint is enforced on the updated conductivity. The updated conductivity is evaluated by running the flow model and only those models that match the observed global responses (such as piezometric head or concentration data) at the corresponding time step will be retained. By doing so, we ensure that the updated conductivities not only preserve complex connectivity but also honor the global dynamic data. The cost of the local-global pattern matching will be higher than the original implementation of the EnPAT, however to alleviate that cost, we will propose a learning scheme that is described next. This global match step makes the new algorithm similar to the iterative EnKF or confirmed EnKF used in petroleum engineering (Wen and Chen, 2006).

In order to mitigate the computational cost, a learning process is integrated into the pattern matching scheme after the global matching step. Specifically, the mismatch between the observed and simulated response data will be used to rank the accepted models. The worst models in the training images in terms of the mismatch between predicted and observed heads will be replaced with the new accepted models. The set of training images will thus be refined using the ranking scheme (Bayer et al., 2010), which will result in a faster matching during the next local pattern searching process.

Fig 2 is the flowchart of the improved EnPAT algorithm accounting for global constraints that incorporates the process of refining the training image sets. The ensemble of conductivity training images and the corresponding simulated heads as well as the observed head data are the inputs to the algorithm. The pattern matching is performed at a few randomly selected pilot points first, and then the results are extrapolated using a multiple point simulation technique such as the direct sampling technique. The simulation at the pilot point locations starts with a search of conditioning data in the vicinity of the simulation node. The conditioning pattern comprises both the pattern of conductivity data as well as the pattern of head data. The conditioning data pattern is determined by the search radius and the maximum number of conditioning nodes specified by the user. The training image ensemble is searched in order to find the matching pattern. The distance between the conditioning data pattern and the pattern in the training image is calculated, and, if the distance is lower than a tolerance value, the outcome at the node corresponding to the simulation node is retained as the simulated value. After all the pilot point locations are simulated, the remaining nodes are simulated using the MPS technique. Once the simulated realization is complete, flow simulation is performed and the global match to the observed head data is assessed. If the match is within a tolerance value, the updated realization is assimilated into the training image set. The training image with the worst match to the observed data is dropped from the training image set. It is evident that the computational cost is mainly dependent on the predefined tolerance value used to judge if the response of the updated conductivity model matches the history. The tolerance value can be linked to the likelihood function describing the head measurement error (Mariethoz et al., 2010a). In the subsequent example, the computational efficiency of the training image refining scheme will be evaluated.

The EnPAT algorithm is coupled with the groundwater flow modeling program MODFLOW (Harbaugh et al., 2000) and the direct sampling MPS method (Mariethoz et al., 2010b), and programmed in C++.

3. Synthetic Example

3.1. Example Setup

Given the training image shown in Fig 3A, the reference logconductivity field is generated using the direct sampling MPS method (Mariethoz et al., 2010b) (see Fig 3B). The model is discretized into $50 \times 50 \times 1$ gridblocks with cell size $1m \times 1m \times 1m$. The logconductivity values follow a bimodal histogram with mean and standard deviation of $0.12m/d$ and $2.51m/d$, respectively. The reference conductivity field exhibits curvilinear features with high conductivity sand channels and low conductivity mudstone zones. We assume that an injection well ($Q = 25m^3/d$) is located at the center of the aquifer and there are 8 observation wells (Fig 3B). This reference logconductivity model will be regarded as the true model, and the aim of the stochastic inverse simulation is to generate a suite of models that are as close to the true model as possible, conditioned to the observed piezometric head data.

The aquifer is assumed confined with constant head boundaries on the eastern and western sides and no flow boundaries at the remaining faces (Fig 3B). The specific storage is set constant and equal to 0.01. The total simulation time is 30 days discretized into 10 time steps that follow a geometric series with ratio of 1.2. The observed data of the first five time steps will be used as the conditioning data. We assume that there are 9 measured conductivity data at the locations shown in Fig 3B. Five hundred initial conductivity models are generated using the direct sampling MPS method with the same training image and simulation parameters (for example, the distance threshold and the number of conditioning data in the pattern) as the reference, conditioned to the measured conductivities. The initial head is assumed to be zero in the whole aquifer. We only consider the uncertainty in conductivity. The boundary conditions and other parameters are assumed known with certainty.

The parameters used in the EnPAT are listed as follows. The search radius is set as 25 m for both conductivity and head. The maximum number of elements in the pattern for conductivity and head are specified to be 10 for both. The weighted Euclidean distance is used to calculate the distance between the conditional and candidate patterns. Distance tolerances for the head and conductivity are both set to zero. The number of pilot points is specified to be 300. The threshold value used to judge the global convergence of updated models is set at 0.1. Apart from the threshold parameter, the sensitivity of results to other parameters are extensively investigated either in the context of the direct sampling MPS method (Meerschman et al., 2013) or in the pattern matching scheme (Li et al., 2013b) because both these methods have some parameters in common.

3.2. Results

In order to assess the uncertainty of the updated models before and after integrating the observed piezometric head data, the multidimensional scaling method is used to visualize the geological models in metric space, although a range of other approaches is proposed in the paper by Bennett et al. (2013). Multidimensional scaling (Borg and Groenen, 2005) is a data analysis method used to segment the model space on the basis of dissimilarity between models. For example, the difference in piezometric head between any pair of models can be used to construct the distance matrix of head data. Applying the multidimensional scaling method, the distance matrix is projected to an equivalent metric space. In this transformed space, closer points imply similar response behaviors, which might imply similar geological structures.

Figure 4 shows the geological models in metric space for the different cases. For the case of the prior models (i.e., before conditioning to piezometric head data) (see Fig 4A), the geological models exhibit large variations in terms of simulated head. Specifically, two selected models far from the true model in metric space have distinctively different spatial pattern of hydraulic conductivity, compared with the reference. When the piezometric head data are integrated into the geological models using EnPAT (i.e., the local pattern matching approach) (see Fig 4B), the geological models converge to the reference in the metric space. It is evident that the posterior uncertainty of the geological models is smaller. However, a few models still have a minor deviation from the reference model in terms of the simulated head. It can be interpreted that the local pattern matching does not guarantee a global history match. This might be because of the limited size of the training ensemble that is insufficient to estimate the multipoint correlations between parameter and state accurately. The arbitrarily specified distance tolerance values might also contribute to the incorrect global match. For the case of the updated models using EnPAT with the global constraint, all the geological models are close to the reference (see Fig 4C). If we look at the two individual models, they show very similar spatial geologic patterns to the reference.

Figure 5 displays the simulated head by rerunning the forward simulator from time zero for wells #3 and #6 for the different cases. When the observed head data are not integrated, the simulated head values for different models exhibit a large spread and the ensemble average of head values departs from the reference. When the head measurements are integrated into the geological models using EnPAT, the uncertainty (i.e., spread) of simulated heads is reduced and the ensemble average is also close to the reference. As is evident from the updated geological models shown in Fig 4, the simulated head values for some models still deviate significantly from the reference. In order to remove such models, the global constraint is enforced and the resulting set of models yield simulated head values in a tighter range.

3.3. Flow and transport predictions

In channelized aquifers, the connectivity of high conductivities impacts flow and transport of solutes. Here, we use the updated conductivity models obtained previously to predict the flow and transport behaviors by subjecting then to modified boundary conditions. Specifically, the western boundary condition is changed from a constant head $h = 0$ to $h = 5$ m and the injection well is removed. Flow in the aquifer is simulated at steady-state. The longitudinal and transverse dispersion coefficients are set as 0.5 m and 0.1 m, respectively. A conservative tracer is injected at the left side of the aquifer (see Fig 6), and a control plane located at $x = 45$ m is used to record the travel time of particles. The random walk particle tracking algorithm (Salamon et al., 2006) is utilized to solve the transport equation.

Figure 7 displays the cumulative breakthrough curves (BTCs) for different cases. In models that are not constrained to the observed piezometric head data (Fig 7A), the BTCs show a large spread and the reference curve is close to the 5^{th} percentile of ensemble BTCs. This indicates that the initial models do not exhibit adequate connectivity of high conductivity regions in the vicinity of the injection face resulting in later arrival times to the control plane. When the models are conditioned to head data (Fig 7B), the spread of BTCs is reduced and the ensemble average of BTCs is much closer to the reference. We also observe that the breakthrough profile of the reference is more than the 5^{th} percentile of BTCs, which implies that the ensemble connectivity of the updated geological models is no longer underestimated in the vicinity of the injection face. When a global constraint on the updated geological models is enforced (Fig 7C) the BTCs have a smaller spread and the ensemble average is close to the reference.

4. Computational Efficiency

One of improvements of the EnPAT algorithm presented in this work is the introduction of the learning process by refining the set of training images by replacing the worst models in the prior set with the newly accepted models. In this way, the training images will be close to the "true" model when more and more accepted models are reached. Additionally, the local pattern search process becomes faster for finding the matched candidate pattern because the uncertainty of training images is correspondingly reduced.

Fig 8 shows the evolution of maximum root mean square error (RMSE) of the training image models in terms of the mismatch between the simulated head and the observation data, at the first time step. As we see, after 500 models are generated, the maximum RMSE is close to the predefined tolerance value for the global constraint. In other words, the training image models will reflect the observed data at this stage.

Fig 9 displays the comparison of the computational efficiency of the original EnPAT with global constraint but without the training image replacement and the improved algorithm. It clearly shows that after the training image is refined using newly accepted models, the number of evaluations for each new generated model is reduced significantly, which results in lower computational cost.

5. Discussion

In this paper, we propose a local-global pattern matching method to characterize heterogeneous hydraulic conductivity field using observed piezometric head data. In previous studies (e.g., Zhou et al., 2012), only a local pattern matching approach is considered and the updated geological models might not be consistent with the observed piezometric head data because the relationship between the parameter and state variables may be inaccurately represented by the pattern searching procedure. This typically occurs when the ensemble size is not large enough to explore the non-linear relationship between the parameter and state variables. Another source of error might be the distance tolerance values that may be specified to be too large in order to find the matched pattern. To address this issue, a global constraint is carried out to select the updated models which fit the observed data by running the forward simulator. Specifically, the updated conductivity model will be accepted if the root mean square of absolute error between the simulated head and observed data is smaller than the predefined tolerance value. A key issue associated with this approach is the increased computational cost incurred to ensure that the global constraint is satisfied. It is evident that the computational expense is dependent on the magnitude of the defined tolerance value. In order to reduce the computational cost, a second level learning process is integrated into the EnPAT. Specifically, the training image models are refined by replacing the worst models in the training set with the newly accepted model. By doing so, the matching process is much faster and the number of evaluation of forward model is reduced.

There are some similarities between EnPAT with global constraint and rejection sampling (Tarantola, 2005). In rejection sampling, the likelihood of the data given a particular model is computed and the model is accepted based on that likelihood exceeding a threshold. In the current implementation of EnPAT with the global constraint, a particular updated model is accepted based on the mismatch between the observation and the simulation being below a threshold. If the threshold is large, more models are accepted. The correspondence between the posterior set of models obtained using the scheme outlined in this paper and a classical implementation of rejection sampling using the likelihood function will be assessed in a later publication. The new candidate model is generated by MPS method using the training set that is composed of both the conductivity models and the corresponding simulated piezometric heads. In this way, the acceptance ratio could be much higher and the sampling scheme could be more effective, which is similar with the iterative spatial resampling proposed by Mariethoz et al. (2010a) for which the new candidate model is regenerated by conditioning on a set of hard data sampled from the previous accepted model.

The local-global pattern matching approach could be extended to integrate other sources of data such as flow-rate and concentration data within the same framework. Including additional variables into the joint pattern could make the pattern matching process more challenging, because the patterns found from the training set may not represent the relationship between the parameter and state variables accurately. An alternative is that, the joint pattern composed of conductivity and head is locally matched through the pattern searching scheme, and then the flow-rate or concentration data could be matched in a global match step. This multi-level patten matching could be more effective than a one-step implementation and will be extensively investigated in the future.

In the current implementation, we assume that the training image is known and there is no uncertainty about the training image. In practice, the training image may have some degree of uncertainty. It is straightforward to integrate the uncertainty of training image into the EnPAT. Specifically, the uncertainty of training images could be handled by assembling multiple realizations generated by different training images in the ensemble used for the local pattern match.

The performance of the EnPAT is dependent on the information available. If the conditioning data could not reflect the potential geological structures, the EnPAT could not identify them, accordingly.

6. Conclusions

In complex geological systems such as fluvial aquifers, carbonate systems and naturally fractured aquifers, multiple-point statistics-based modeling methods are required to characterize complex and curvilinear features. Parameter identification with MPS requires an effective inverse method that yields models that not only honor the observed dynamic data, but also preserve curvilinear geological features that impact hydrocarbon recovery and aquifer remediation.

In this paper, a hybrid of EnPAT and global matching method is developed for developing models that honor multiple point statistics defining reservoir connectivity as well as the observed dynamic data. Specifically, the updated models through the local pattern matching approach are forward simulated to verify if they match the observed dynamic data. In other words, global pattern matching is conducted after the local pattern matching (i.e., the EnPAT) so that the resultant models will be conditioned to dynamic data and the curvilinear geometry will be preserved as well. In addition, to accelerate the local and global match, the training image models are refined by integrating the new matched models. We tested the local-global pattern matching approach to characterize a bimodally distributed heterogeneous conductivity field. The results indicate that the characterization of conductivity and flow and transport predictions are improved after the integration of the global constraint into the EnPAT algorithm. Also, the computational cost is reduced when a ranking scheme is introduced into the algorithm.

Acknowledgements. The authors gratefully acknowledge the financial support by DOE through projects

DE-FE0004962 and DE-SC0001114. The last author acknowledges the support of the Spanish Ministry of Economy and Competitiveness through project CGL2011-23295. We greatly thank the three anonymous reviewers for their comments, which substantially improved the manuscript.

References

Bayer, P., de Paly, M., Bürger, C. M., 2010. Optimization of high-reliability-based hydrological design problems by robust automatic sampling of critical model realizations. Water Resources Research 46 (5).

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., et al., 2013. Characterising performance of environmental models. Environmental Modelling & Software 40, 1–20.

Borg, I., Groenen, P. J., 2005. Modern multidimensional scaling: Theory and applications. Springer.

Capilla, J., Llopis-Albert, C., 2009. Gradual conditioning of non-Gaussian transmissivity fields to flow and mass transport data: 1. Theory. Journal of Hydrology 371 (1-4), 66–74.

Capilla, J. E., Rodrigo, J., Gómez-Hernández, J. J., 1999. Simulation of non-gaussian transmissivity fields honoring piezometric data and integrating soft and secondary information. Math. Geology 31 (7), 907–927.

Carrera, J., Neuman, S., 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. Water Resources Research 22 (2), 199–210.

de Marsily, G., 1978. De l'identification des systemes hydrogeologiques. Ph.D. thesis.

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics 53 (4), 343–367.

Gómez-Hernández, J., Wen, X., 1998. To be or not to be multi-Gaussian? a reflection on stochastic hydrogeology. Advances in Water Resources 21 (1), 47–61.

Gómez-Hernández, J. J., Journel, A. G., 1993. Joint simulation of multiGaussian random variables. In: Soares, A. (Ed.), Geostatistics Tróia '92, volume 1. Kluwer, pp. 85–94.

Gómez-Hernández, J. J., Sahuquillo, A., Capilla, J. E., 1997. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 1, Theory. Journal of Hydrology 203 (1–4), 162–174.

Guardiano, F., Srivastava, R., 1993. Multivariate geostatistics: beyond bivariate moments. In: Soares, A. (Ed.), Geostatistics-Troia. Kluwer Academic Publ, Dordrecht, pp. 133–144.

Harbaugh, A. W., Banta, E. R., Hill, M. C., McDonald, M. G., 2000. MODFLOW-2000, the U.S. Geological Survey modular ground-water model. U.S. Geological Survey, Branch of Information Services, Reston, VA, Denver, CO.

Hendricks Franssen, H., Gómez-Hernández, J., Sahuquillo, A., 2003. Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data. Journal of Hydrology 281 (4), 281–295.

Hu, L. Y., Zhao, Y., Liu, Y., Scheepens, C., Bouchard, A., 2013. Updating multipoint simulatings using the ensemble Kalman filter. Computers & Geosciences 51, 7–15.

Jafarpour, B., Khodabakhshi, M., 2011. A probability conditioning method (PCM) for nonlinear flow data integration into multipoint statistical facies simulation. Mathematical Geosciences 43 (2), 133–164.

Li, L., Srinivasan, S., Zhou, H., Gómez-Hernández, J. J., 2013a. A pilot point guided pattern matching approach to integrate dynamic data into geological modeling. Advances in Water Resources 62, 125–138.

Li, L., Srinivasan, S., Zhou, H., Gómez-Hernández, J. J., 2013b. Simultaneous estimation of both geologic and reservoir state variables within an ensemble-based multiple-point statistic framework. Mathematical Geosciences 46, 597–623.

Liu, X., Cardiff, M. A., Kitanidis, P. K., 2010. Parameter estimation in nonlinear environmental problems. Stochastic Environmental Research and Risk Assessment 24 (7), 1003–1022.

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

Mariethoz, G., Renard, P., Straubhaar, J., 2010b. The direct sampling method to perform multiple-point geostatistical simulaitons. Water Resources Research 46 (11).

Meerschman, E., Pirot, G., Mariethoz, G., Straubhaar, J., Meirvenne, M., Renard, P., 2013. A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. Computers & Geosciences 52, 307–324.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., Teller, E., 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21 (6), 1087–1092.

Oliver, D., Cunha, L., Reynolds, A., 1997. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Mathematical Geology 29 (1), 61–91.

Renard, P., Allard, D., 2011. Connectivity metrics for subsurface flow and transport. Advances in Water Resources 51, 168–196.

Salamon, P., Fernàndez-Garcia, D., Gómez-Hernández, J. J., 2006. A review and numerical assessment of the random walk particle tracking method. Journal of Contaminant Hydrology 87 (3-4), 277–305.

Strebelle, S., 2002. Conditional simulation of complex geological structures using multiple-point statistics. Mathematical Geology 34 (1), 1–21.

Sun, A. Y., Morris, A. P., Mohanty, S., 2009. Sequential updating of multimodal hydrogeologic parameter fields using localization and clustering techniques. Water Resources Research 45 (7).

Tarantola, A., 2005. Inverse problem theory and methods for model parameter estimation. siam.

Wen, X., Chen, W., 2006. Real-time reservoir model updating using ensemble kalman filter with confirming option. SPE Journal 11 (4), 431–442.

Zhou, H., Gómez-Hernández, J., Hendricks Franssen, H., Li, L., 2011. An approach to handling non-gaussianity of parameters and state variables in ensemble kalman filtering. Advances in Water Resources 34 (7), 844–864.

Zhou, H., Gómez-Hernández, J., Li, L., 2012. A pattern-search-based inverse method. Water Resources Research 48 (3).

Zhou, H., Gómez-Hernández, J., Li, L., 2014. Inverse methods in hydrogeology: evolution and recent trends. Advances in Water Resources 63, 22–37.

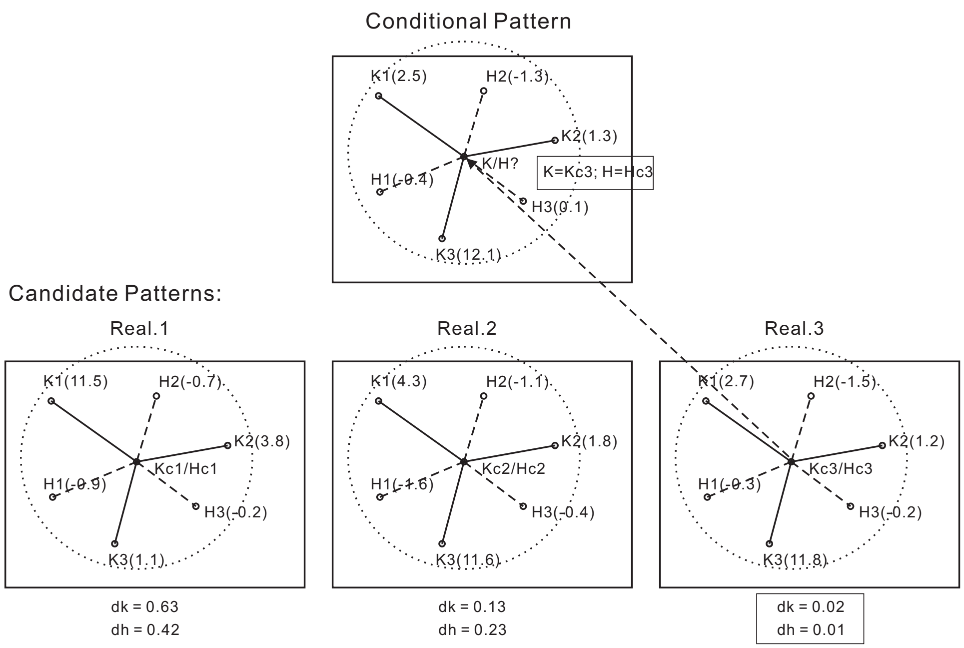

Figure 1: Scheme of pattern matching. The gridblock conductivity and head are sampled as the estimated values if its pattern has distance values smaller than thresholds or minimum distance values.

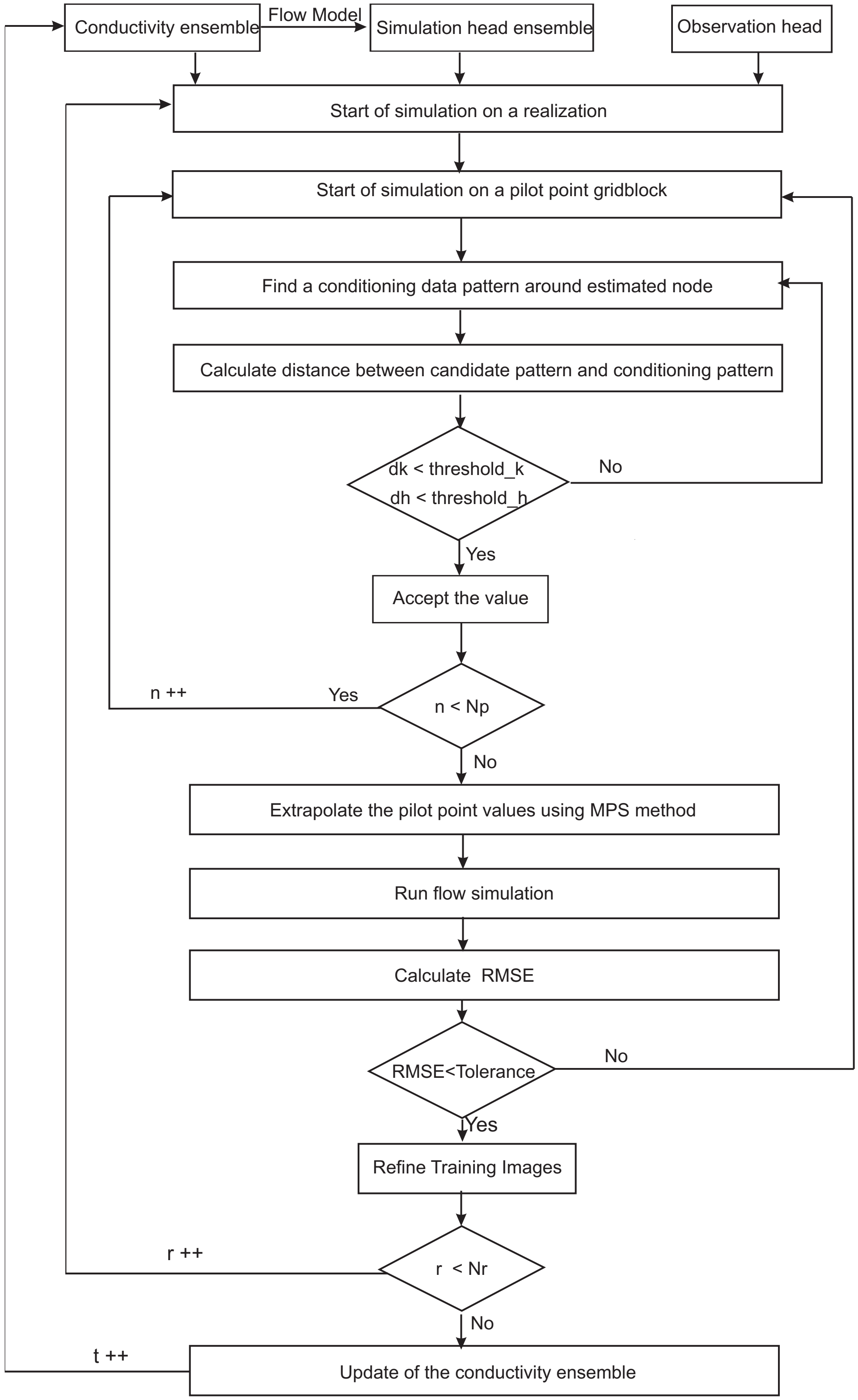

Figure 2: Flowchart of the EnPAT. dk and dh indicate the threshold values of distance for the conductivity and head, respectively; n means the number of gridlock simulated in a realization; Np is the number of pilot points; r denotes the number of realization in the ensemble; Nr is the total number of realizations; t is the number of time step for the simulation.

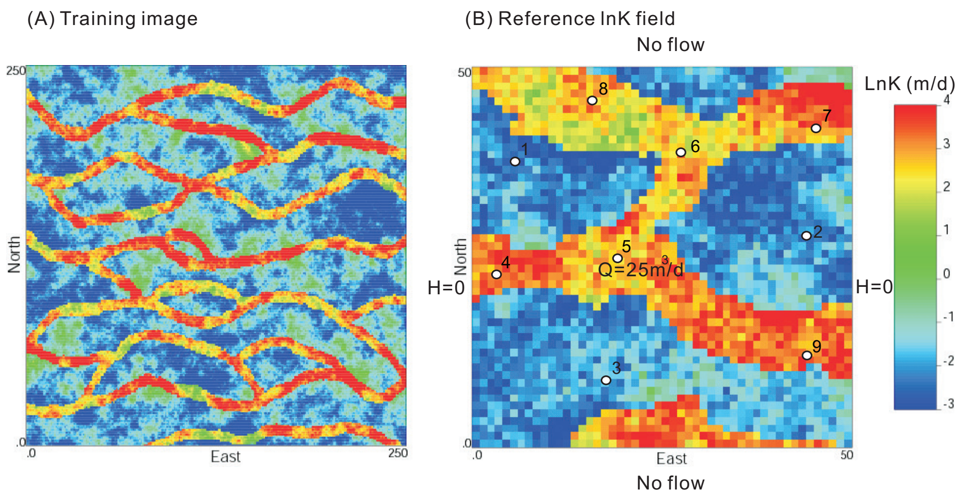

Figure 3: (A) Training image (B) Reference conductivity, boundary conditions of flow model and observation wells.

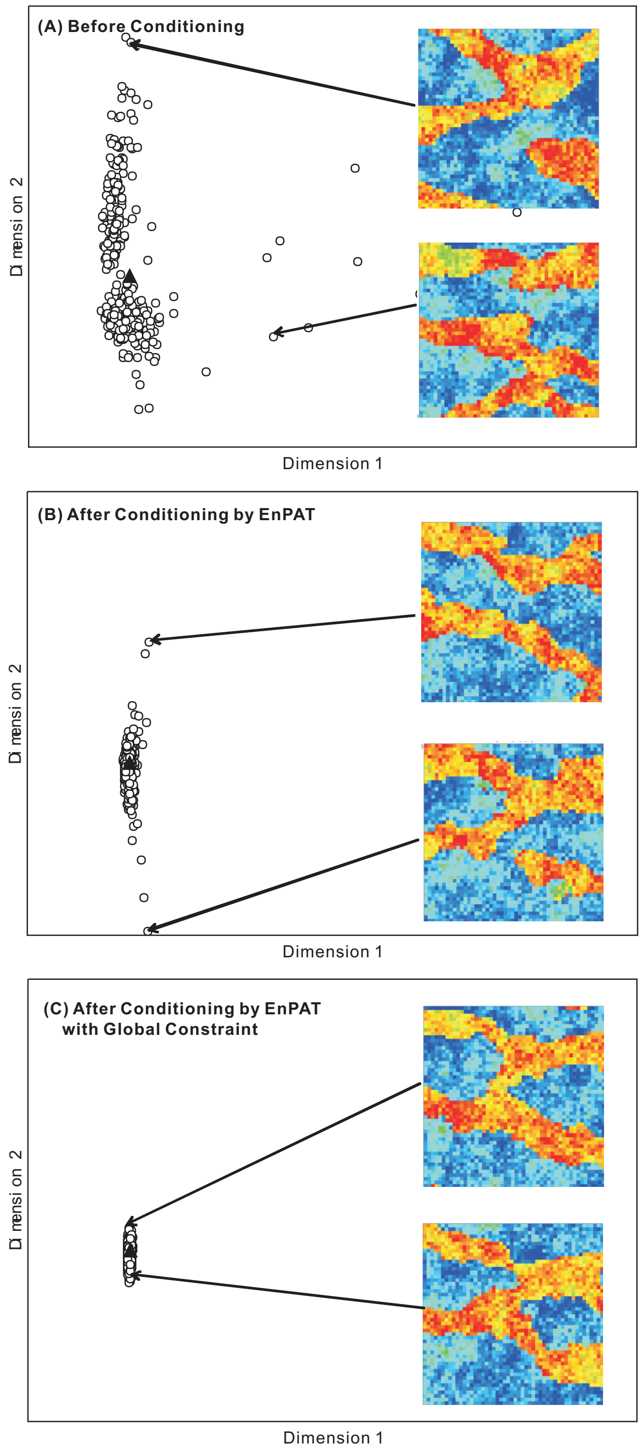

Figure 4: Visualization of geological models in terms of the simulated heads of the last time step at the well locations using the multidimensional scaling method. The open circle denotes geological model, and the triangle indicates the true model.

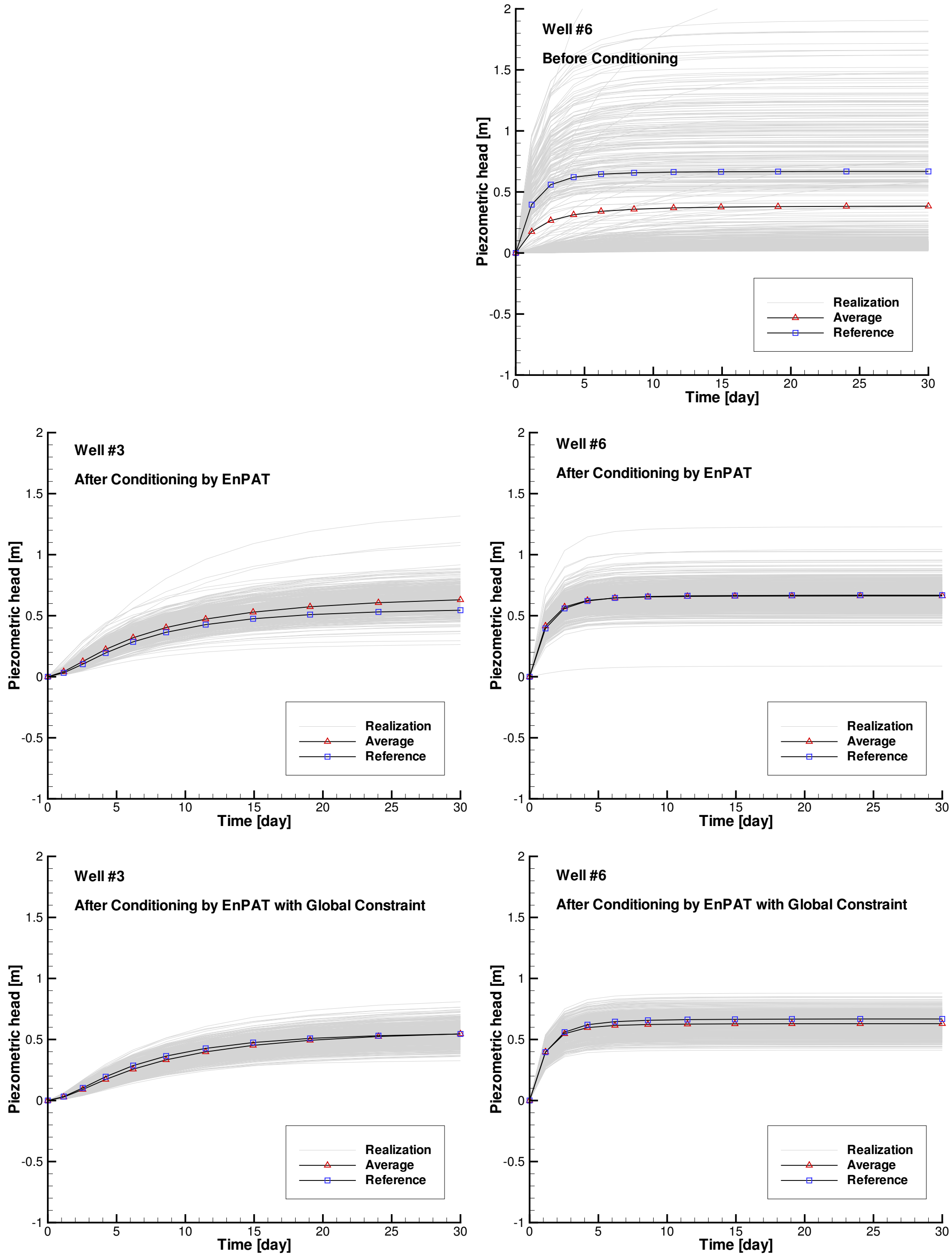

Figure 5: The simulated head at two wells before and after the data conditions using the EnPAT with and without global constraint.

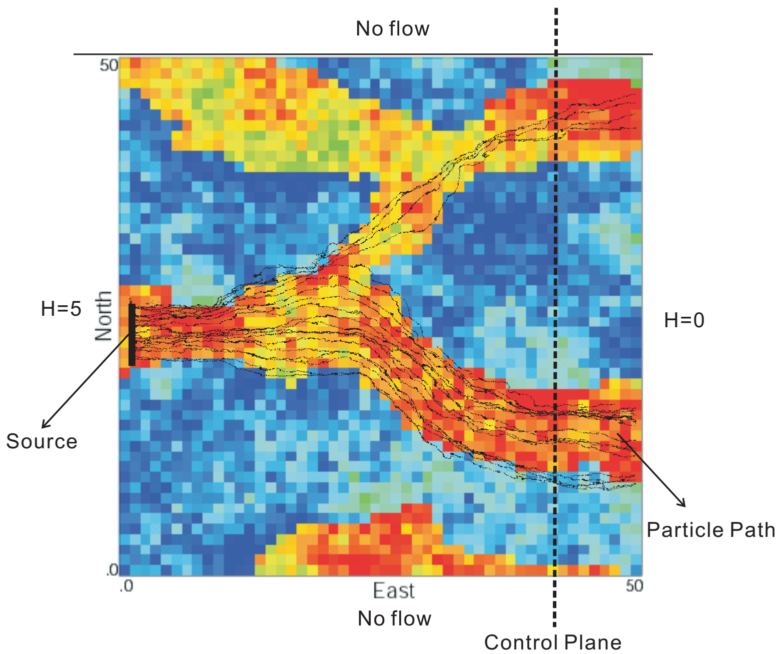

Figure 6: Transport configuration

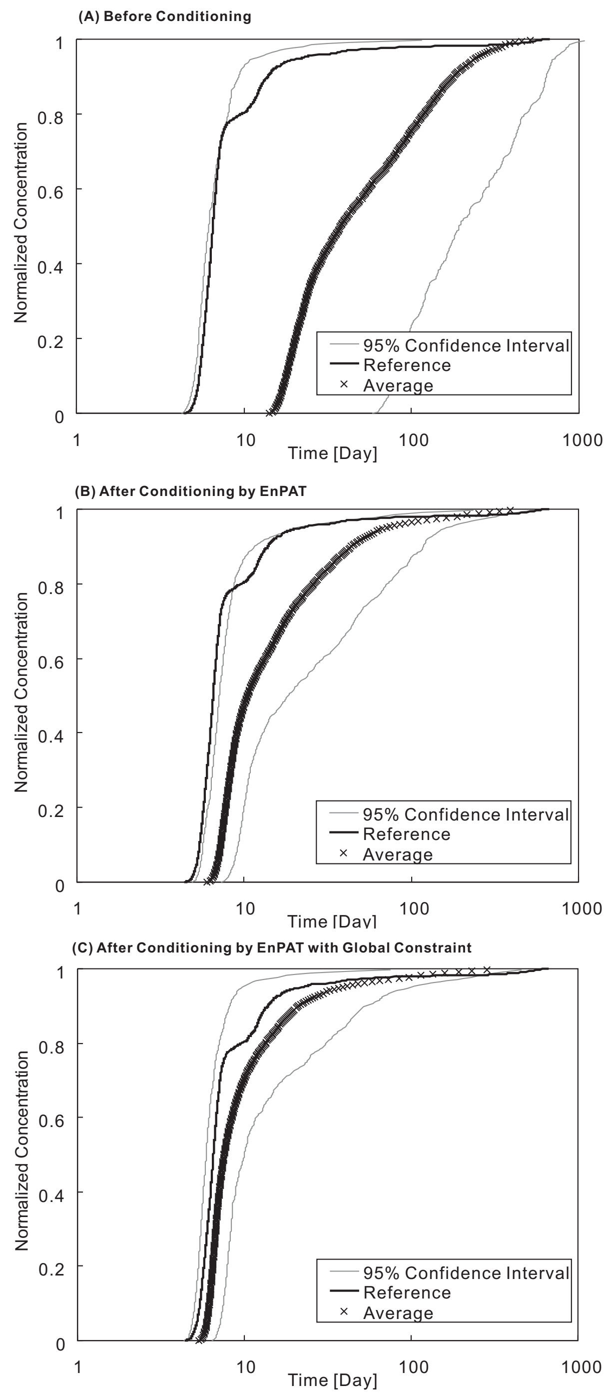

Figure 7: The simulated cumulative breakthrough curves before and after the data conditions using the EnPAT with and without global constraint

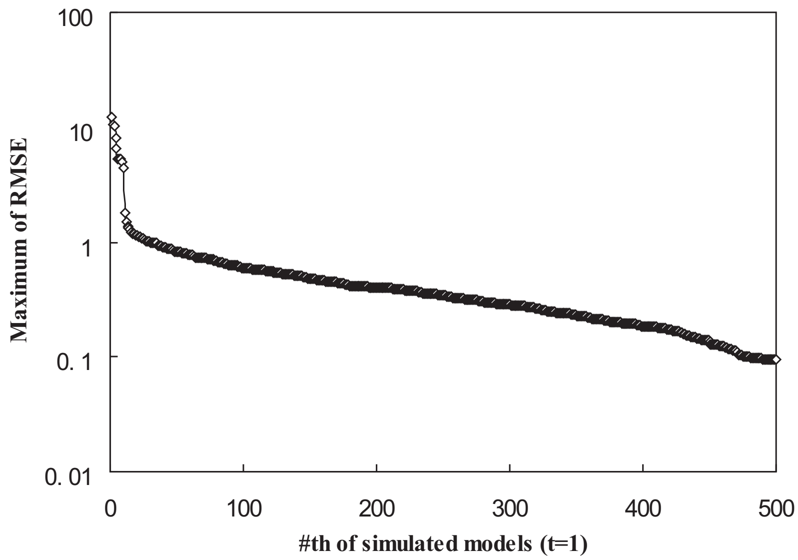

Figure 8: The maximum of RMSE for the training image models ($t = 1$)

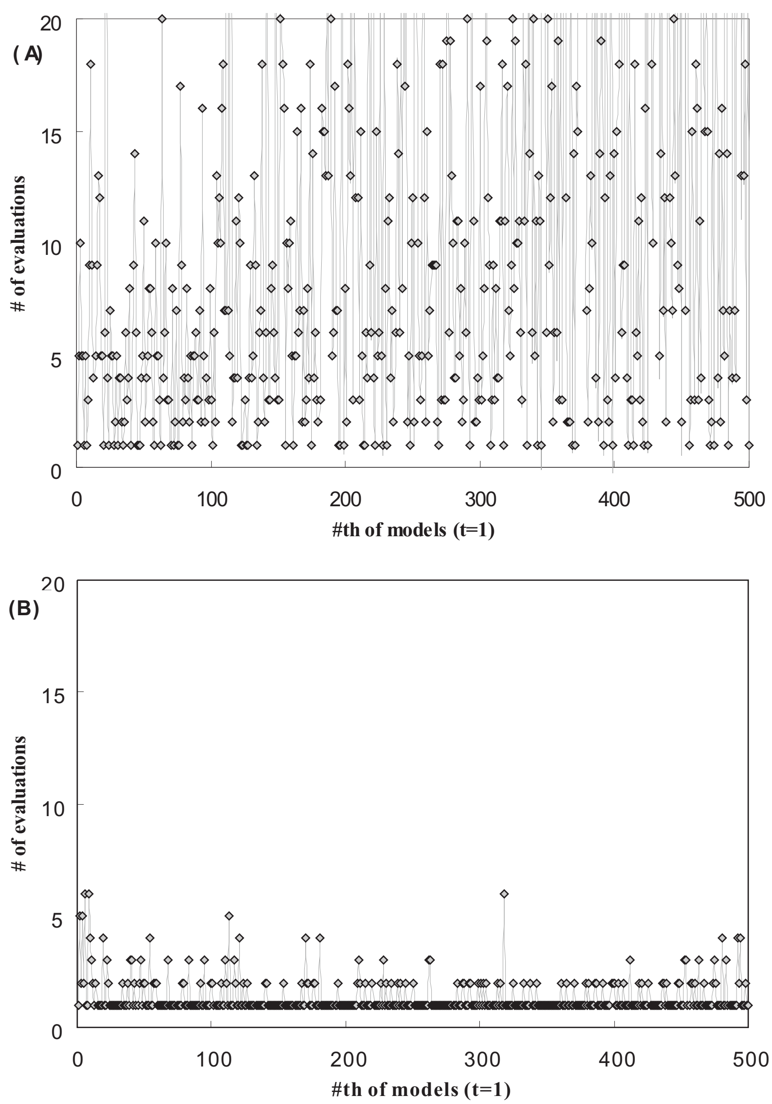

Figure 9: The number of evaluations for each simulated models using EnPAT (A) and improved EnPAT (B) ($t = 1$)