# A Pilot Point Guided Pattern Matching Approach to Integrate Dynamic Data into Geological 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

Methods based on multiple-point statistics (MPS) have been routinely used to characterize complex geological formations in the last decade. These methods use the available static data (for example, measured conductivities) for conditioning. Integrating dynamic data (for example, measured transient piezometric head data) into the same framework is challenging because of the complex non-linear relationship between the dynamic response and geology. The Ensemble PATtern (EnPAT) search method was recently developed as a promising technique to handle this problem. In this approach, a pattern is postulated to be composed of both parameter and state variables, and then, parameter values are sequentially (point-wise) simulated by directly sampling the matched pattern from an ensemble of training images of both geologic parameters and state variables. As a consequence, the updated ensemble of realizations of the geological parameters preserve curvilinear structures (i.e., non-multiGaussanity) as well as the complex relationship between static and dynamic data. Moreover, the uncertainty of flow and transport predictions can be assessed using the updated ensemble of geological models. In this work, we further modify the EnPAT method by introducing the pilot-point concept into the algorithm. More specifically, the parameter values at a set of randomly selected pilot point locations are simulated by the pattern searching procedure, and then a faster MPS method is used to complete the simulation by conditioning to the previously simulated pilot point values. This pilot point guided MPS implementation results in lower computational cost and more accurate inference of the parameter field. In addition, in some situations where there is sparsity of measured geologic static data, the EnPAT algorithm is extended to work only with the dynamic data. We employed a synthetic example to demonstrate the effectiveness of pilot points in the implementation of EnPAT, and also the capability of dynamic data to identify complex geologic structures when measured static data are not available.

*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

Mathematical modeling of subsurface flow and transport is essential for managing energy production and contaminant remediation. Aquifer parameters such as hydraulic conductivity or permeability, exhibit large spatial variation commonly over several orders of magnitude. Due to the scarcity of measurements, a geostatistical approach [e.g., 1, 2, 3, 4, 5, 6] is usually employed to represent the spatial heterogeneity of aquifer attributes. These geostatistical models are conditioned to the measured static data and yield multiple equiprobable realizations of the attributes. The uncertainty of model response is assessed subsequently by running a forward model on these multiple parameter realizations [7, 8, 9]. Besides static data (i.e., the hard data), dynamic data such as transient piezometric head and concentration can also be used to condition the models. The procedure of constructing aquifer models conditioned to dynamic data is termed inverse modeling where the objective is to identify the parameter values at unsampled locations by integrating those dynamic data into the model, and thus to improve the predictions of flow and transport in the future [e.g., 10, 11, 12, 13, 14, 15].

Inverse methods have been developed and used extensively to generate permeability or hydraulic conductivity models conditioned to dynamic data [e.g., 16, 17, 18, 19, 20, 21]. De Marsily et al. [16] developed the pilot-point method, in which the a conductivity map is determined by calibrating a few pilot-point locations followed by a kriging interpolation. Gómez-Hernández et al. [19] proposed the self-calibration stochastic inverse method, which is an extension of the pilot point method, aimed at the generation of multiple conductivity realizations, all matching the observed state data. The performance of the self-calibration method has been demonstrated for synthetic and real case studies [e.g., 22, 23, 24]. One key concern of this approach is how to determine the number of pilot points and their locations. Gómez-Hernández et al. [19] recommended two or three pilot points per correlation length. LaVenue and Pickens [25] placed the pilot points in the highest sensitivity regions. Wen et al. [26] proposed to randomly locate the pilot points such that the spacing between the pilot points is one correlation length. Wen et al. [27] coupled self-calibration with genetic algorithms to determine the optimal locations of pilot points. A code implementing the self-calibration model is available to the public [28].

Beside the pilot point-based inverse methods, Hu [20] proposed the gradual deformation method, in which a single deformation parameter controls the generation of conductivity fields such that the simulated state values match the observation data. Evensen [21] proposed the ensemble Kalman filter, a further extension of the extended Kalman filter, in which the covariance between the aquifer attribute at a location and the corresponding well response is calculated from an ensemble of realizations and is used subsequently to

3

update the ensemble so as to reflect the measured well response. Heidari et al. [29] proposed to update the conductivity fields at the pilot point locations using the EnKF, and then extrapolate the updated values to all locations in the aquifer by kriging.

All the inverse methods mentioned above are optimal for multi-Gaussian geologic media. In other words, they perform well for conductivity fields following a multiGaussian random function such as those generated by two-point variogram-based geostatistical methods such as sequential Gaussian simulation [3]. However, traditional two-point covariance methods can not be used to describe fluvial depositions, which commonly display features of curvilinear geometry. The significance of a curvilinear feature on the flow and transport predictions has been discussed in the literature [30, 31, 32, 33, 34]. An alternative to two-point covariance methods is to use recently developed methods based on multiple-point statistics (MPS) to address this problem [4]. In this approach, instead of using the traditional variogram model, a training image that conceptually describes the salient geological features is used. A spatial template (i.e., a multiple-point configuration) is used to infer the experimental local conditional distributions [35]. A complete review of the training image based MPS method for aquifer modeling is presented in Hu and Chugunova [36]. Alternative approaches are available to generate non-multiGaussian field such as transition probabilities and copula methods [e.g., 37].

Inverse methods developed to work in conjunction with multiple-point-based simulation methods are relatively new in the literature [e.g, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. Caers [38] applied a probability perturbation method on the permeability fields generated from MPS. Alcolea and Renard [39] developed a block moving-window algorithm, an extension of the block Markov Chain Monte Carlo method by Fu and Gómez-Hernández [49] to condition MPS simulations to piezometric head data as well as connectivity data. Mariethoz et al. [40] proposed an iterative spatial resampling method working on the MPS simulation in the Bayesian framework. More recently, the ensemble Kalman filter (EnKF) is gaining popularity in petroleum engineering and hydrogeology because of its computational efficiency and real-time data assimilation features [e.g., 50, 51, 52, 53, 54, 55]. However, the classical EnKF, when implemented on models constrained to MPS, does not preserve the statistics or complexity of the models, because of the covariance-based updating used in the EnKF. The complex relationship between the spatial pattern of state variables and the flow response is approximated by a covariance, and the repeated updating results in a final ensemble that does not correctly exhibit the complex spatial connectivity of features such as channels, fractures, etc. Some variations of the EnKF have been proposed to overcome the limitation posed by the two-point covariance-based updating so that the connectivity of the permeability field is properly preserved. Sun et al. [41]

4

proposed to couple the EnKF and a mixture of Gaussian models as well as localization techniques in order to improve the performance for fluvial models. Sarma and Chen [42] developed a kernel EnKF to preserve the connectivity of the MPS-based permeability realizations. Zhou et al. [43, 56] proposed to first transform the parameter and state variables to marginal Gaussian distributions through the normal-score transformation before implementing the updating step in EnKF. Jafarpour and Khodabakhshi [44] suggested updating the ensemble mean of several MPS-based permeability realizations using EnKF and subsequently use the updated ensemble-mean values as soft data to regenerate updated models using the MPS approach. Hu et al. [45] proposed to update the uniform-score random numbers used to draw outcomes from the MPS conditional distributions, using the EnKF.

All the above-mentioned EnKF-based updating schemes may yield suboptimal representation of permeability variations in the aquifer because of the linearization of the transfer function model implied by the representation of the complex relationship between state and dynamic response variables using lower order moments (e.g., mean and covariance). Zhou et al. [47] proposed a pattern-search-based inverse method where the relationship between the conductivity field and the dynamic responses, in the form of corresponding patterns, is inferred from training images and used to simulate MPS-based conductivity fields. This process was subsequently extended in Li et al. [48] where an Ensemble PATten (EnPAT) algorithm was presented to integrate dynamic data within an ensemble-based multiple-point statistic framework. In this approach, model parameter and state values are simultaneously and sequentially estimated, which not only improves the characterization of the parameter field, but also makes it feasible to assimilate dynamic data in real-time, similar to other ensemble-based filtering approaches.

In this work, we further improve the performance of the EnPAT algorithm [48] by implementing the pilot-point concept as done in the sequential self-calibration method [16, 19]. More specifically, the conductivities at pilot point locations are generated through the EnPAT scheme, and then a fast MPS method is used to generate updates of the initial ensemble conditioned on the pilot-point parameter values. We demonstrate this algorithm on a synthetic data set. Moreover, in some cases, hard data (i.e., conductivity values used to condition the conductivity realizations) may be unavailable and only well responses may exist to generate the ensemble of aquifer models. We extend the EnPAT algorithm to condition only on the dynamic data in order to recognize curvilinear geologic structures. Lastly, we demonstrate the algorithm for conditioning to fully transient flow response information. In these demonstrations, we track the evolution of models as dynamic data is integrated sequentially in time.

The rest of the paper is organized as follows. Section 2 outlines the improved EnPAT methodology

and details the significance of the use of pilot points in this new algorithm. Next, in Section 3, numerical experiments are employed to demonstrate the proposed method. Then in Section 4, the influence of the number of pilot points on the results is analyzed. The paper ends with summary and conclusions.

## 2. Methodology

### 2.1. Algorithm

The ensemble pattern-search algorithm, an extension of the direct sampling (DS) method (a training image based MPS method) [5], was first proposed by Zhou et al. [47]. The algorithm extended the direct sampling method in two specific aspects: (1) the pattern is composed of not only the geologic parameters (e.g, conductivity) but also includes state variables (e.g., piezometric head); (2) an ensemble of realizations both for the geologic parameters as well as for state variables are used as multiple training images. Li et al. [48] further improved the algorithm by simultaneously estimating both geologic and state variables, resulting in a better characterization of permeability/conductivity at both large and small scales.

In this work, the ensemble pattern-search algorithm is coupled with the pilot-point concept commonly used in traditional inverse methods such as sequential self-calibration in order to improve computational efficiency. Figure 1 shows a flowchart of the EnPAT algorithm. As we can see, if all the gridblocks are pilot points, the algorithm is completely equivalent to the original implementation proposed by Zhou et al. [47]; on the contrary, if there are no pilot points, the algorithm would become the ensemble direct sampling MPS method as shown in the paper by Li et al. [48]. In this paper, we employ a finite set of pilot points in order to render the process of simultaneously searching for both the conducutivity and head pattern computationally fast. Thus, the basic idea is to use the original pattern search method to simulate the conductivity at the pilot points conditioned to the pattern of both conductivity and head data in the neighborhood of the pilot point. Subsequently, the ensemble direct sampling MPS (Zhou et al. [47]) is used to complete the subsurface models conditioned on the previously simulated pilot point values. To achieve this, once the number and location of pilot points are defined, the corresponding constraints (both k and h) on the nodes are enforced to simulate the conductivities. When the conductivities at all the pilot points are estimated, the simulation continues to simulate conductivity values at all remaining nodes conditioned to the values at the pilot point locations. The pilot point locations are sampled at random from the entire simulation domain.

The specific implementation of the EnPAT algorithm can be summarized as follows:

1. Initialization step. Generate an ensemble of $N_r$ prior geologic models conditioned to measured conductivity data using traditional MPS methods such as SNESIM [4]. It is assumed that the model is

composed of $N_n$ grid blocks. Here, the initial ensemble can be generated either using a single training image (a conceptual model of the aquifer [35]), or using several training images representing the prior uncertainty in the geological model.

2. Forecast step. For each conductivity model (realization), the transient groundwater flow model is solved from time zero to $t$ using a standard numerical simulator, i.e.,

$$\mathbf{Y}_k = f(\mathbf{X}_{k-1}, \mathbf{Y}_0) \tag{1}$$

$\mathbf{Y}_k$ and $\mathbf{Y}_0$ denote the simulated piezometric head at time $t = k$ and the initial head at time $t = 0$, respectively. $\mathbf{X}_{k-1}$ corresponds to the updated conductivity at time $t = k - 1$. The groundwater flow model, boundary conditions, sources and other unknown parameters are represented by the transfer function $f$.

The ensemble of conductivity realizations at $t = k - 1$ (i.e. $\mathbf{X}_{k-1}$) plus the corresponding ensemble of simulated piezometric head at $t = k$ (i.e. $\mathbf{Y}_k$) can be used as the ensemble of training images to simulate the ensemble conductivity at $t = k$ (i.e. $\mathbf{X}_k$) using the following pattern search method.

3. Ensemble pattern search step. Update the ensemble conductivity $\mathbf{X}_{k-1}$ based on the observed head data at time $t = k$.

   (a) Start the loop to estimate the conductivity realization $r = 1$ at $t = k$ (i.e., $\mathbf{X}_k$).

   (b) Generate a random path $P_i(x), i = 1, 2, \ldots, N_n$ visiting all the nodes of the realization, in which the first $N_p$ nodes are used as pilot points. These pilot points are located randomly within the domain.

   (c) Start the loop to estimate the conductivity $X_k^i$ and piezometric head $Y_k^i$.

      i. If the conductivity and head data both are known at node $i$ by measurement (e,g., hard data), go to step (d).

      ii. Build the conditioning pattern, which is composed of geologic parameter (e.g., conductivity) and state (e.g., head) variables (see Figure 2A). The conductivity and head data in the pattern can be either the measured data or the previous estimated data (as in traditional sequential simulation). The size of pattern is determined by the predefined maximum number of conductivity and head data and the search radius. When the conditioning data are sparse, the size of pattern is usually large, which is good for the estimation of conductivity at the large scale such as the channels (see Figure 2A); when the available conditioning data becomes dense (at later stages of the simulation), the pattern in a smaller area around the simulation

7

node will be searched, which is beneficial for the simulation of smaller scale features (Figure 2B). This scheme with flexible pattern size is thus equivalent to the multiple grid approach in the sequential Gaussian simulation algorithm proposed by Gómez-Hernández and Journel [3], in terms of reducing the artifacts in the simulation. Also, in order to reduce the computational cost, a spiral search strategy centered on the estimated location $i$ is considered [2].

Note that, if the simulation is at the early stages and no hard data are available, the conditioning pattern will be composed of only state variables (Figure 2E).

iii. Generate a random path $P_j, j = 1, 2, \ldots, N_r$ visiting all the realizations. Search for a match to the conditional pattern (see Figure 3). Start the loop with a randomly selected realization $j$.

A. Search the candidate pattern on the realization $j$. The search would be enforced at the node $i$ because the state (head) is affected by the presence of boundary conditions and source terms. However, this would require a large ensemble in order to find a matching pattern and is consequently computationally expensive. In order to alleviate the computational expense, the 8 nodes (26 nodes in three dimensions) surrounding the center grid block $i$ are also searched for a matching pattern. Despite using a small area surrounding the node $i$ to search for the matching pattern in the ensemble, the method has the same accuracy as the basic DS algorithm that uses a single training image (Figure 2C and D) to simulate the pattern of conductivity. At the pilot point location, conductivity is simulated constrained to both conductivity and piezometric head. At other locations, the simulation is conducted with the pattern of only conductivity; in other words, the ensemble DS is used to extrapolate the conductivity on the basis of the previously simulated pilot point.

B. Calculate the distances, $d_j^X$ for conductivity and $d_j^Y$ for head, between the conditional pattern and the candidate pattern found during the search. Specific distance functions will be described in section 2.2.

C. Compare the calculated distance values with the predefined tolerance values. Specifically:

$$
\begin{cases}
\text{Case } 1 : X_k^i = X_{k-1}^j, Y_k^i = Y_k^j, & \text{if } d_j^X < \xi^X \text{ and } d_j^Y < \xi^Y. \\
\text{Case } 2 : X_k^i = X_{k-1}^j, & \text{if } d_j^X < \xi^X . \\
\text{Case } 3 : Y_k^i = Y_k^j, & \text{if } d_j^Y < \xi^Y .
\end{cases}
\tag{2}
$$

where $\xi^X$ and $\xi^Y$ are the predefined tolerance values for the conductivity and head, respectively. Case 1 denotes simulation at the pilot point locations (Figure 2A and B), in which case tolerances are applied to both the conductivity and piezometric head patterns in order to find the match. Case 2 corresponds to simulation of the non-pilot point locations (Figure 2C and D), in which case a tolerance is applied only to the conductivity pattern in order to find the closest match. Case 3 occurs when hard data are not available and only the head data are used to construct the pattern (Figure 2E and F). In that case a tolerance is only applied to the pattern of piezometric head in order to find the closest match.

If the computed distance is less than the specified tolerance, go to step (d).

D. Otherwise, set $j = j + 1$, and go to step A.

iv. If at the end of previous steps the tolerance criteria are not met, then the pattern with the smallest distances is retained and the corresponding grid block ($n$) conductivity value is assigned as the simulated one, i.e.,

$$
\begin{cases}
\text{Case 1}: X_k^i = X_{k-1}^n, Y_k^i = Y_k^n, & \text{if } d_{min}^X = d_n^X \text{ and } d_{min}^Y = d_n^Y . \\
\text{Case 2}: X_k^i = X_{k-1}^n, & \text{if } d_{min}^X = d_n^X . \\
\text{Case 3}: Y_k^i = Y_k^n, & \text{if } d_{min}^Y = d_n^Y .
\end{cases}
\tag{3}
$$

where $d_{min}^X$ and $d_{min}^Y$ are the smallest distances for the conductivity and piezometric head, respectively. The simulated values will be used as the conditioning data for the subsequent simulation.

(d) Set $i = i + 1$ and go to step (c) until all nodes are estimated.

(e) Set $r = r + 1$ and go to step (a) until all realizations are simulated.

4. Loop back. The forecast and pattern search loop is applied again on the updated conductivity field for the next time step of observations (i.e., set $k = k + 1$ and go to step 2 until all the time step observation data are conditioned).

Even though both conductivity and head are simultaneously simulated using the pattern-search approach in order to capture the variability of the aquifer attributes at multiple scales [48], only the simulated conductivity is kept for updating at the next time step. This is because the updated piezometric heads might not be consistent with those simulated by the full-physical simulator because of complexity of well/boundary

9

conditions, etc. The training images for piezometric head for pattern search at all subsequent time steps are obtained by re-running the forward simulator from time step zero. This does add to the computational cost, but it is similar to the implementation in EnKF-based methods such as static EnKF and re-start EnKF [e.g., 54, 44, 57, 45].

Unlike in previous implementations [47, 48] where the pattern search constrained to both conductivity and head is performed at all nodes in the domain, in this new algorithm, pattern search is performed only at predefined pilot points. At all other non-pilot point locations, the pattern is composed of only conductivity and the simulation is conducted by pattern search through the criteria presented in Case 2 above. The implementation of the full pattern search at pilot point locations not only accelerates the simulation, but also improves the quality of updated conductivity field in terms of preserving the large scale connectivity. The pilot points are randomly located within the field, and they are changed from one realization to the next as well as from one time step to the next.

### 2.2. Distance functions

In the algorithm, distance functions play a key role to quantify the candidate pattern. Follow the same approach as in Zhou et al. [47]. The Manhattan distance is computed for categorical variables (i.e., rock facies or homogeneous conductivities within facies) and the weighted Euclidian distance function is computed for continuous variables (for example, piezometric head or continuous conductivities).

- Manhattan distance for the categorical variables:

$$
d\{d(x_n), p(x_n)\} = \frac{1}{n} \sum_{i=1}^{n} a_i \quad d \in [0, 1]
$$

$$
a_i = \begin{cases} 0, & \text{if } d(x_i) = p(x_i) \\ 1, & \text{otherwise} \end{cases} \tag{4}
$$

where $n$ is the numbers of nodes in the pattern, $d(x_n)$ is the conditioning pattern, $p(x_n)$ is the matching pattern, $x_i$ represents the members of the pattern.

- Weighted Euclidean distance for the continuous variables

$$
d\{d(x_n), p(x_n)\} = \left( \frac{1}{\sum_{i=1}^{n} h_i^{-1}} \sum_{i=1}^{n} h_i^{-1} \frac{|d(x_i) - p(x_i)|^2}{d_{max}^2} \right)^{1/2} \quad d \in [0, 1] \tag{5}
$$

<sub>227</sub> where $h_i$ is the Euclidean distance between node $i$ and the node being simulated, therefore giving more

<sub>228</sub> weight to dissimilarities of the closest nodes, $d_{max}$ is the maximum absolute difference $|d(x_i) - p(x_i)|$.

## <sub>229</sub> 3. Synthetic example

### <sub>230</sub> 3.1. Reference field

<sub>231</sub> A single-phase transient groundwater flow problem is set up to demonstrate the effectiveness of the

<sub>232</sub> proposed method. The aquifer is discretized into $50 \times 50 \times 1$ grid blocks, with the block dimensions of 1

<sub>233</sub> $m \times 1 \ m \times 1 \ m$. The training image from Strebelle [4] is used as the conceptual geologic model with high

<sub>234</sub> conductivity sand and low conductivity shale (see Figure 4A). The SNESIM code developed by Strebelle [4]

<sub>235</sub> is employed to generate the reference facies map, in which the proportion of sand is set as 30%. Constant

<sub>236</sub> hydraulic conductivity values with 10 and $10^{-4}$ $m/d$ are assigned to sand and shale, respectively. The

<sub>237</sub> variance of log-conductivities is about 5.3. The reference model is shown in Figure 4B.

<sub>238</sub> The aquifer is assumed to be confined and modeled with constant head boundaries at $x = 0 \ m$ and $x = 50$

<sub>239</sub> $m$ (i.e., $h = 0 \ m$ at those boundaries) and with no flow boundaries at the remaining model faces (see Figure

<sub>240</sub> 4B). The initial head is assumed zero over the entire domain. Specific storage is set as constant and equal

<sub>241</sub> to 0.01 $m^{-1}$. A pumping well (well #2) located at the center of the domain extracts water at a constant

<sub>242</sub> flow rate of 25 $m^3/d$. There are three observation wells located within the domain. The spatial locations

<sub>243</sub> of the 4 wells are shown in Figure 4B. The simulation time is 30 $days$ and is discretized into 10 time steps

<sub>244</sub> following a geometric sequence of ratio 1.2.

<sub>245</sub> The finite-difference code MODFLOW [58] is used to solve the transient flow equation (1). The piezo-

<sub>246</sub> metric head data for the first five time steps (8.6 $days$), collected from the 4 wells are used as conditioning

<sub>247</sub> data to update the initial model parameters. No measurement error is considered as well as no measured

<sub>248</sub> conductivity data in this example. The measurement error could be implicitly accounted for by increasing

<sub>249</sub> the distance tolerance value for each variable.

<sub>250</sub> Five hundred unconditional facies realizations are generated using SNESIM and then populated with

<sub>251</sub> constant conductivity values. The parameters used in the EnPAT are listed in Table 1. Note that: (1) the

<sub>252</sub> tolerance values both for the conductivity and head are defined as zero so that the best simulated value

<sub>253</sub> (corresponding to the minimum distance) is achieved at the pilot point locations, even though this increases

<sub>254</sub> the computational cost; (2) The number of pilot point is set at 500. Sensitivity of the results to the number

<sub>255</sub> of pilot point locations will be studied in section 4. The sensitivity to other parameters such as tolerance

<sub>256</sub> and number of conditioning data in the pattern is referred to the work by Meerschman et al. [59].

*3.2. Evaluation criteria*

A set of metrics are computed in order to quantify the performance of proposed method.

1. Check the ensemble mean and variance of conductivity maps, that can be used to check the reproduction

of the main patterns observed in the reference and to display the spatial uncertainty depicted by the

updated ensemble of conductivity realizations.

2. Calculate the average absolute error ($AAE$) and average ensemble spread ($AES$), that can be used to

quantify the accuracy and uncertainty of simulated results. They are defined as follows:

$$\text{AAE}(x) = \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{1}{N_r} \sum_{r=1}^{N_r} |x_{i,r} - x_{\text{ref},i}|$$
$$\text{AES}(x) = \frac{1}{N_n} \sum_{i=1}^{N} \sigma_{x_i}^2 \tag{6}$$

where $x_{i,r}$ is conductivity at location $i$ for realization $r$, $x_{\text{ref},i}$ is the reference conductivity value at

location $i$, $N_n$ is the number of nodes, $N_r$ is the number of realizations in the ensemble and $\sigma_{x_i}^2$ is the

ensemble variance at location $i$.

3. Calculate the ensemble connectivity of conductivity. Here, the code CONNEC3D developed by Pardo-

Igúzquiza and Dowd [60] is used to compute the connectivity. In this approach, the conductivity values

are first converted to indicators and then the connectivity of indicators can be calculated, dependent on

the user defined spatial template or connectivity direction. The definition and computation of connec-

tivity is referred the study by Pardo-Igúzquiza and Dowd [60]. In our case, horizontal connectivity (i.e.,

east-west direction) is calculated because the reference conductivity field shows dominant connectivity

in that direction.

*3.3. Results*

*3.3.1. Hydraulic conductivity characterization*

Using the EnPAT data integration algorithm, an ensemble of 500 conductivity realizations are obtained

for each conditioning time step. Here, all the evaluation is based on the indicator transform of conductivity,

i.e., high-conductivity has an indicator value of 1 and low-conductivity shale of 0. Figure 5 shows the

ensemble mean and variance of conductivity at three time steps. Figure 6 displays two randomly selected

realizations before and after conditioning to the dynamic data. Figure 7 shows the spatial distribution of

pilot points and the simulated values at those locations in two realizations. The evolution of the calculated

$AAE$ and $AES$ with time is shown in Figure 8. Figure 9 displays the connectivity calculated for all members of the ensemble as well as the ensemble average before and after data conditioning.

Analyzing the evolution of ensemble mean and variance maps with time, it is evident that (a) without any conditioning data, the mean and variance of the initial ensemble is spatially constant because the channels are randomly distributed in space; (b) the main pattern of channels gradually converges to the reference, and also the uncertainty of ensemble conductivity is reduced. From a quantitative point of view, Figure 8 also shows that the $AAE$ is reduced by about 30% and $AES$ by about 60% after conditioning to all available head data. Both $AAE$ and $AES$ attain a plateau after time step 4 implying that further conditioning to head data will not improve the characterization of the conductivity field. Other information such as solute concentration or a denser network of wells would be needed to further resolve the conductivity field. The connectivity statistics in Figure 9 also demonstrate an extreme reduction in uncertainty after conditioning to head data. More specifically, the ensemble mean of connectivity is close to the reference and the spread of connectivity over the ensemble realizations is reduced after conditioning to the head data. The two individual realizations in Figure 6 depict the channels before and after data conditioning. We can conclude that the channels in the initial model deviate considerably from the reference while the updated model exhibits similar spatial distribution of channels as in the reference. Figure 7 shows the pilot point locations used for two of the realizations. Judging by the conductivity values at the pilot point locations, we can conclude that the spatial distribution of channels has already been broadly represented in the pilot point values. This further supports our implementation of conditioning to head and conductivity patterns at pilot point locations and subsequent spatial interpolation of only conductivity values away from pilot point locations.

Based on the results above it can be conjectured that the EnPAT algorithm can be efficiently used to integrate head data into the simulation of conductivity fields. Note that the results further indicate that in the absence of hard conductivity data here, features such as channels can be simulated based on head data and based on learning the relationship between conductivity and the dynamic response.

*3.3.2. Flow predictions*

The ultimate objective of the dynamic data integration process is to make better prediction of future performance of the aquifer flow process. The ensemble of geologic models were rerun from time zero using the updated conductivity after conditioning to all the head data.

Figure 10 shows the variation in predicted head values over the ensemble and the corresponding average over the ensemble for the 4 wells. The head data for the first 8.6 *days* was used for conditioning and the data for the remaining period is used to assess the prediction accuracy of the models. We can see that without

13

any conditioning, the spread in head values is large and the average deviates significantly from the reference. After conditioning to the head data, the ensemble average of predicted head is close to the reference and the variation in predicted values (uncertainty) is also reduced.

## 4. Sensitivity to the number of pilot points

One key issue with this modified EnPAT algorithm is the optimization of the number and location of the pilot points. Figure 11 displays the evolution of the ensemble mean of updated conductivity values for the first time step using different number of pilot points within EnPAT (i.e., $Np = 20, 100, 500$ and $1500$). We can see that the pattern exhibited by the ensemble mean is best reproduced with 500 or more pilot points. Clearly, if there are no pilot points then no conditioning to head data is achieved and in the absence of any hard conductivity data, the ensemble mean map would be constant as shown in Figure 5 ($nT = 0$). On the other hand, if there are a larger number of pilot points, the computational expense is high and also the simulated images show artifacts due to the strong constraints enforced during the pattern search process (see Figure 4 of paper by Zhou et al. [47] taking all the gridblocks as pilot points). The speed up ($S_p$) has been calculated to evaluate the computational gain, i.e.,

$$S_p = \frac{T}{T_p} \tag{7}$$

where $T$ is the CPU time consumed under the original EnPAT implementation (i.e., all the nodes are pilot points and needed to be simulated with constraints of both conductivity and piezometric head), $T_p$ is the CPU time using $p$ pilot points. Figure 12 shows the speedup with the number of pilot points. As expected, speed up is reduced as more pilot points are used.

Choosing a reasonable number of pilot points not only reduces the computational cost, but also improves the quality of the updated conductivity fields. In the current implementation, no further postprocess procedure is required while in the original EnPAT algorithm proposed by Zhou et al. [47], the postprocess is needed to remove the artifacts which are not exhibited in the training images. In this example, using about one fifth of the total number of nodes, the speed up is about 2.5 times, and the main spatial pattern of the channels is exhibited in the ensemble mean map, similar to the original implementation of EnPAT (see Figure 11 and 12). A practical guideline for selecting the number of pilot points could be such that the average distance between pilot points is smaller than the channel width in order to capture the details of the curvilinear structure.

14

## 5. Discussion

The EnPAT algorithm shares some features with the EnKF [21]: (1) an ensemble of realizations is used, and thus uncertainty in flow and transport predictions can be assessed; (2) a full-physical simulator is integrated as a black-box in both methods; (3) the correlation between parameter and state is explicitly estimated through the ensemble realizations. However, a significant difference exists in the way the parameters are updated. In the EnKF, the parameter is updated as the initial value plus the Kalman Gain weighted dynamic data mismatch. That is, the complex non-linear relationship between the model parameters and the dynamic response is approximated by the Kalman Gain term that is computed based on the covariance, and thus the updating procedure is optimal only if the parameter and state variables are jointly multiGaussian. In the EnPAT algorithm, on the other hand, the model parameters are simulated using the pattern-search method, in which the correlation between parameter and the state variables (dynamic response) is described in terms of a spatial pattern (i.e. beyond two-point covariance), and thus the parameter and state do not necessarily follow a multiGaussian distribution. Furthermore the curvilinear/complex structure of the aquifer parameter is explicitly preserved through the multipoint statistics.

In this paper, we used a synthetic 2D example to demonstrate the robustness and effectiveness of proposed method. In real case, the aquifer could be 3D in nature and the simulation of channeled aquifer will be more CPU demanding and time consuming in 3D than in 2D but, in principle, the method is general in its formulation and is applicable to 3D domains. It is however possible that the flow response in 3D systems are not as seriously impacted by complex reservoir heterogeneity which in turn can impact the quality and variability of the inverted permeability fields [61]. This issue will be investigated in a subsequent paper.

## 6. Conclusion

Inverse modeling is helpful for modeling and prediction of subsurface flow and transport processes. Most of the current approaches are based on an optimization framework. Recently, the EnPAT algorithm [47, 48] was developed to condition non-Gaussian conductivity fields to dynamic data. Unlike current approaches, the EnPAT is based on pattern search and simulation and does not explicitly involve minimization of an objective function. In this approach, the joint pattern of both conductivity and piezometric head data is considered, and the direct sampling multiple-point statistics method [5] is applied to find the matching pattern from an ensemble of aquifer models. Consequently, the updated conductivity realizations not only are conditioned to the available piezometric head data, but also preserve complex curvilinear structure as described by the spatial pattern.

In this paper, we integrate the pilot point concept [16, 19] into the EnPAT method. Specifically, the conductivity at the pilot point location is simulated by searching for the pattern composed of both conductivity and piezometric head. Subsequently, the conductivity at the remaining locations is simulated using the original Direct Sampling algorithm. The main difference being that the matching pattern is searched over an ensemble of training images instead of a single training image.

We demonstrate the proposed method using a synthetic 2D transient flow example. Our results indicate that: (1) the computational cost of EnPAT is reduced by using pilot points while at the same time the quality of updated conductivity realizations is improved; (2) indirect data such as piezometric head have information that helps us identify the complex spatial characteristics of the conductivity field even when measured conductivity data may not be available.

# References

[1] Journel A. Geostatistics for conditional simulation of ore bodies. Economic Geology 1974;69(5):673–87.

[2] Deutsch CV, Journel AG. GSLIB, Geostatistical Software Library and User's Guide. New York: Oxford University Press; 1992.

[3] Gómez-Hernández JJ, Journel AG. Joint sequential simulation of multigaussian fields. In: Geostatistics Troia92. Springer; 1993, p. 85–94.

[4] Strebelle S. Conditional simulation of complex geological structures using multiple-point statistics. Mathematical Geology 2002;34(1):1–21.

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

[6] Huang Y, Srinivasan S. Efficient conditional simulation of spatial patterns using a pattern-growth algorithm. Geostatistics Oslo 2012 2012;:209–20.

[7] Yoram R, et al. Applied stochastic hydrogeology. Oxford University Press; 2003.

[8] Dagan G, Neuman SP. Subsurface flow and transport: a stochastic approach. Cambridge University Press; 2005.

[9] Gómez-Hernández JJ. Complexity. Ground Water 2006;44(6):782–5.

[10] Sun N. Inverse problems in groundwater modeling. Kluwer Academic Durdrecht; 1994.

[11] Zimmerman D, De Marsily G, Gotway C, Marietta M, Axness C, Beauheim R, et al. A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resources Research 1998;34(6):1373–413.

[12] Carrera J, Alcolea A, Medina A, Hidalgo J, Slooten L. Inverse problem in hydrogeology. Hydrogeology Journal 2005;13(1):206–22.

[13] Hendricks Franssen H, Alcolea A, Riva M, Bakr M, van der Wiel N, Stauffer F, et al. A comparison of seven methods for the inverse modelling of groundwater flow. application to the characterisation of well catchments. Advances in Water Resources 2009;32(6):851–72.

[14] Oliver D, Chen Y. Recent progress on reservoir history matching: a review. Computational Geosciences 2011;15(1):185–221.

[15] Zhou H, Gómez-Hernández J, Li L. Inverse methods in hydrogeology: evolution and recent trends. Advances in Water Resources 2013;Under review.

[16] De Marsily G, Lavedan G, Boucher M, Fasanino G. Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model. Geostatistics for Natural Resources Characterization, Part 1984;2:831–49.

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

[18] Kitanidis P. Quasi-linear geostatistical theory for inversing. Water Resources Research 1995;31(10):2411–9.

[19] Gómez-Hernández JJ, Sahuquillo A, Capilla JE. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 1, Theory. Journal of Hydrology 1997;203(1–4):162–74.

17

[20] Hu LY. Gradual deformation and iterative calibration of gaussian-related stochastic models. Math Geology 2000;32(1):87–108.

[21] Evensen G. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics 2003;53(4):343–67.

[22] Capilla J, Jaime Gómez-Hernández J, Sahuquillo A. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data 2. demonstration on a synthetic aquifer. Journal of Hydrology 1997;203(1):175–88.

[23] Capilla J, Jaime Gómez-Hernández J, Sahuquillo A. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric head data 3. application to the culebra formation at the waste isolation pilot plan (wipp), New Mexico, USA. Journal of Hydrology 1998;207(3):254–69.

[24] Hendricks Franssen H, Gómez-Hernández J, Sahuquillo A. Coupled inverse modelling of groundwater flow and mass transport and the worth of concentration data. Journal of Hydrology 2003;281(4):281–95.

[25] LaVenue A, Pickens J. Application of a coupled adjoint sensitivity and kriging approach to calibrate a groundwater flow model. Water Resources Research 1992;28(6):1543–69.

[26] Wen X, Deutsch C, Cullick A. High-resolution reservoir models integrating multiple-well production data. SPE Journal 1998;3(4):344–55.

[27] Wen X, Lee S, Yu T. Simultaneous integration of pressure, water cut, 1 and 4-d seismic data in geostatistical reservoir modeling. Mathematical geology 2006;38(3):301–25.

[28] Wen X, Capilla J, Deutsch C, Gómez-Hernández J, Cullick A. A program to create permeability fields that honor single-phase flow rate and pressure data. Computers & Geosciences 1999;25(3):217–30.

[29] Heidari L, Gervais V, Ravalec ML, Wackernagel H. History matching of petroleum reservoir models by the ensemble kalman filter and parameterization methods. Computers & Geosciences 2013;55:84–95.

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

[31] Wen X, Gómez-Hernández J. Numerical modeling of macrodispersion in heterogeneous media: a comparison of multi-gaussian and non-multi-gaussian models. Journal of Contaminant Hydrology 1998;30(1-2):129–56.

[32] Zinn B, Harvey CF. When good statistical models of aquifer heterogeneity go bad: A comparison of flow, dispersion, and mass transfer in connected and multivariate gaussian hydraulic conductivity fields. Water Resources Research 2003;39(3).

[33] Kerrou J, Renard P, Hendricks Franssen H, Lunati I. Issues in characterizing heterogeneity and connectivity in non-multigaussian media. Advances in Water Resources 2008;31(1):147–59.

[34] Renard P, Allard D. Connectivity metrics for subsurface flow and transport. Advances in Water Resources 2011;51:168–96.

[35] Guardiano F, Srivastava R. Multivariate geostatistics: beyond bivariate moments. In: Soares A, editor. Geostatistics-Troia. Dordrecht: Kluwer Academic Publ; 1993, p. 133–44.

[36] Hu LY, Chugunova T. Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review. Water Resources Research 2008;44(11).

[37] Haslauer C, Guthke P, Bárdossy A, Sudicky E. Effects of non-gaussian copula-based hydraulic conductivity fields on macrodispersion. Water Resources Research 2012;48(7).

[38] Caers J. Geostatistical history matching under training-image based geological model constraints. SPE Annual Technical Conference and Exhibition, 2002;:SPE 77429.

[39] Alcolea A, Renard P. Blocking Moving Window algorithm: Conditioning multiple-point simulations to hydrogeological data. Water Resources Research 2010;46.

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

[41] Sun AY, Morris AP, Mohanty S. Sequential updating of multimodal hydrogeologic parameter fields using localization and clustering techniques. Water Resources Research 2009;45(7).

[42] Sarma P, Chen W. Generalization of the ensemble kalman filter using kernels for nongaussian random fields. In: SPE Reservoir Simulation Symposium. 2009,.

[43] Zhou H, Gómez-Hernández J, Hendricks Franssen H, Li L. An approach to handling non-gaussianity of parameters and state variables in ensemble kalman filtering. Advances in Water Resources 2011;34(7):844–64.

[44] Jafarpour B, Khodabakhshi M. A probability conditioning method (PCM) for nonlinear flow data integration into multipoint statistical facies simulation. Mathematical Geosciences 2011;43(2):133–64.

[45] Hu LY, Zhao Y, Liu Y, Scheepens C, Bouchard A. Updating multipoint simulatings using the ensemble Kalman filter. Computers & Geosciences 2013;51:7–15.

[46] Lange K, Frydendall J, Cordua KS, Hansen TM, Melnikova Y, Mosegaard K. A frequency matching method: solving inverse problems by use of geologically realistic prior information. Mathematical Geosciences 2012;44(7):783–803.

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

[48] Li L, Srinivasan S, Zhou H, Gómez-Hernández JJ. Simultaneous estimation of both geologic and reservoir state variables within an ensemble-based multiple-point statistic framework. Mathematical Geosciences 2012;under review.

[49] Fu J, Gómez-Hernández JJ. A Blocking Markov Chain Monte Carlo Method for Inverse Stochastic Hydrogeological Modeling. Mathematical Geosciences 2009;41(2):105–28.

[50] Wen XH, Chen W. Real-time reservoir model updating using ensemble Kalman filter. In: SPE reservoir simulation symposium. 2005,.

[51] Gu Y, Oliver D. The ensemble Kalman filter for continuous updating of reservoir simulation models. Journal of Energy Resources Technology 2006;128:79–87.

[52] Chen Y, Zhang D. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Advances in Water Resources 2006;29(8):1107–22.

[53] Aanonsen S, Nævdal G, Oliver D, Reynolds A, Valles B. The ensemble Kalman filter in reservoir engineering–a review. Spe Journal 2009;14(3):393–412.

[54] Nowak W. Best unbiased ensemble linearization and the quasi-linear Kalman ensemble generator. Water Resources Research 2009;45(4).

[55] Li L, Zhou H, Hendricks Franssen HJ, Gómez-Hernández JJ. Modeling transient groundwater flow by coupling ensemble Kalman filtering and upscaling. Water Resources Research 2012;48(1).

[56] Zhou H, Li L, Franssen HJH, Gómez-Hernández JJ. Pattern recognition in a bimodal aquifer using the normal-score ensemble kalman filter. Mathematical Geosciences 2012;44(2):169–85.

[57] Schöniger A, Nowak W, Hendricks Franssen HJ. Parameter estimation by ensemble Kalman filters with transformed data: Approach and application to hydraulic tomography. Water Resources Research 2012;48(4).

[58] Harbaugh AW, Banta ER, Hill MC, McDonald MG. MODFLOW-2000, the U.S. Geological Survey modular ground-water model. Reston, VA, Denver, CO: U.S. Geological Survey, Branch of Information Services; 2000.

[59] Meerschman E, Pirot G, Mariethoz G, Straubhaar J, Meirvenne M, Renard P. A practical guide to performing multiple-point statistical simulations with the direct sampling algorithm. Computers & Geosciences 2013;52:307–24.

[60] Pardo-Igúzquiza E, Dowd P. CONNEC3D: a computer program for connectivity analysis of 3D random set models. Computers & geosciences 2003;29(6):775–85.

[61] Fogg GE. Groundwater flow and sand body interconnectedness in a thick, multiple-aquifer system. Water Resources Research 1986;22(5):679–94.
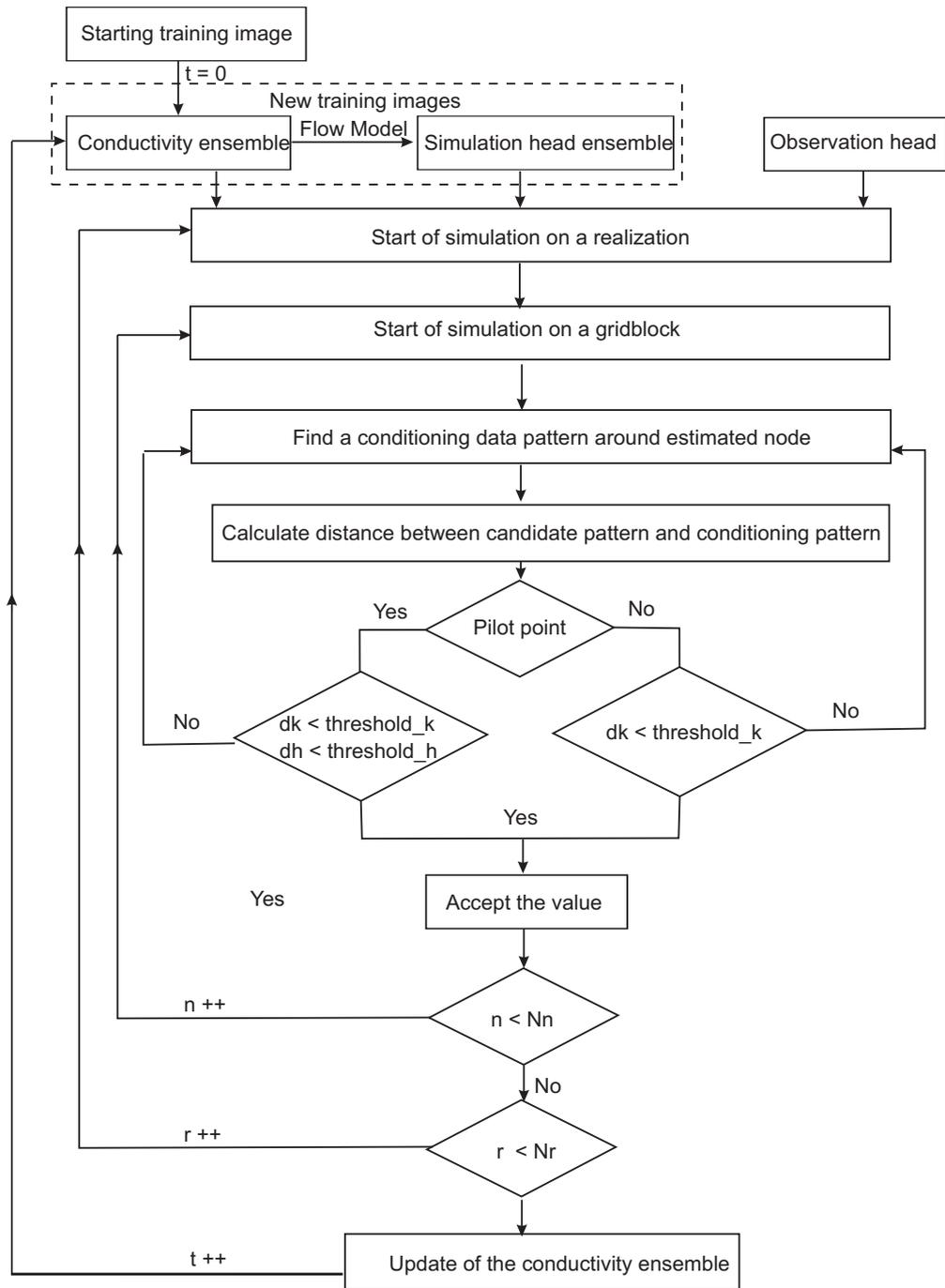
Figure 1: Flowchart of the algorithm.
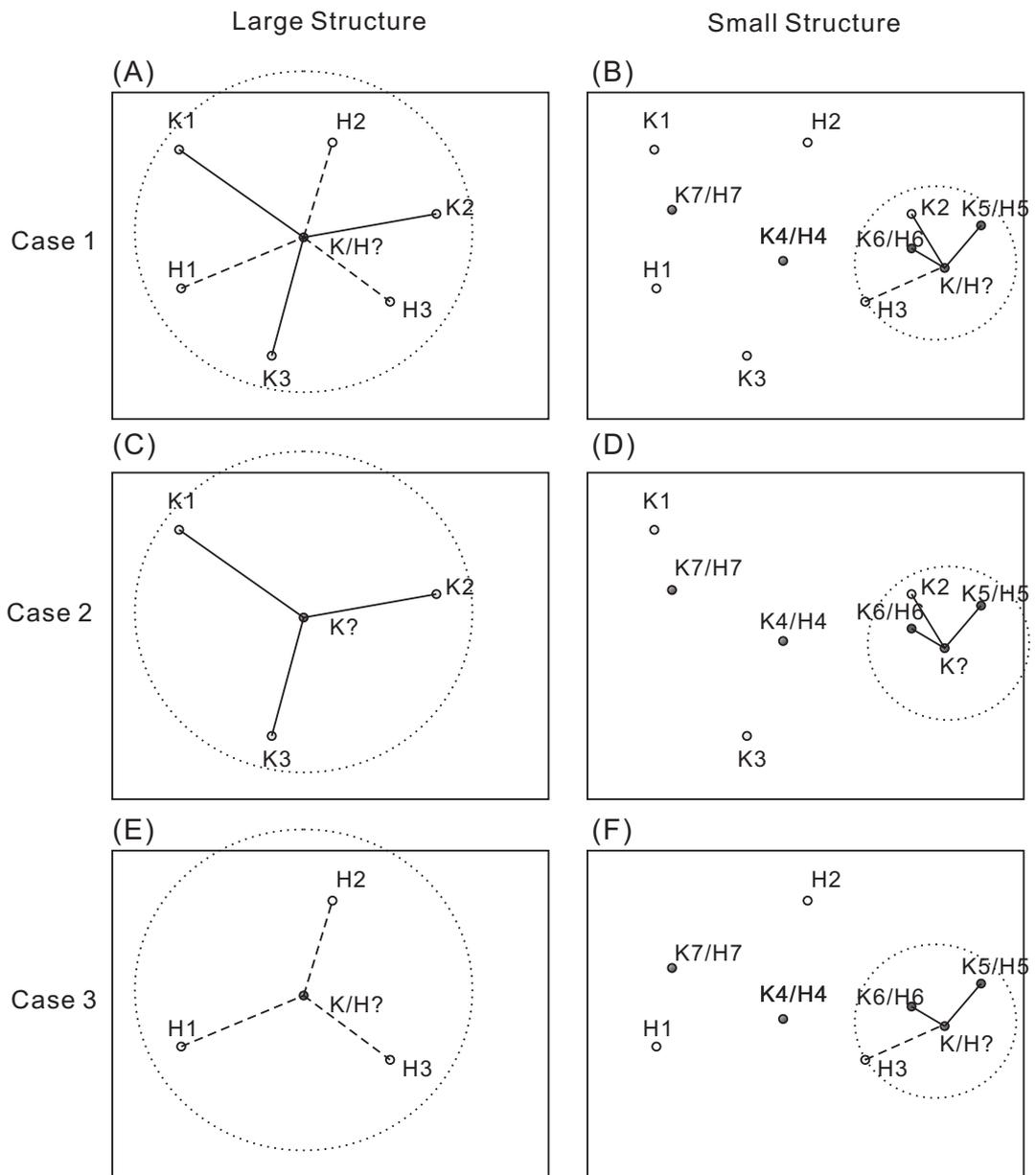
Figure 2: Spatial evolution of the pattern in the EnPAT algorithms. Case 1 denotes the pattern at the pilot points locations; Case 2 is one at the non-pilot point locations and Case 3 corresponds to the pattern in case there is only the measured dynamic data.
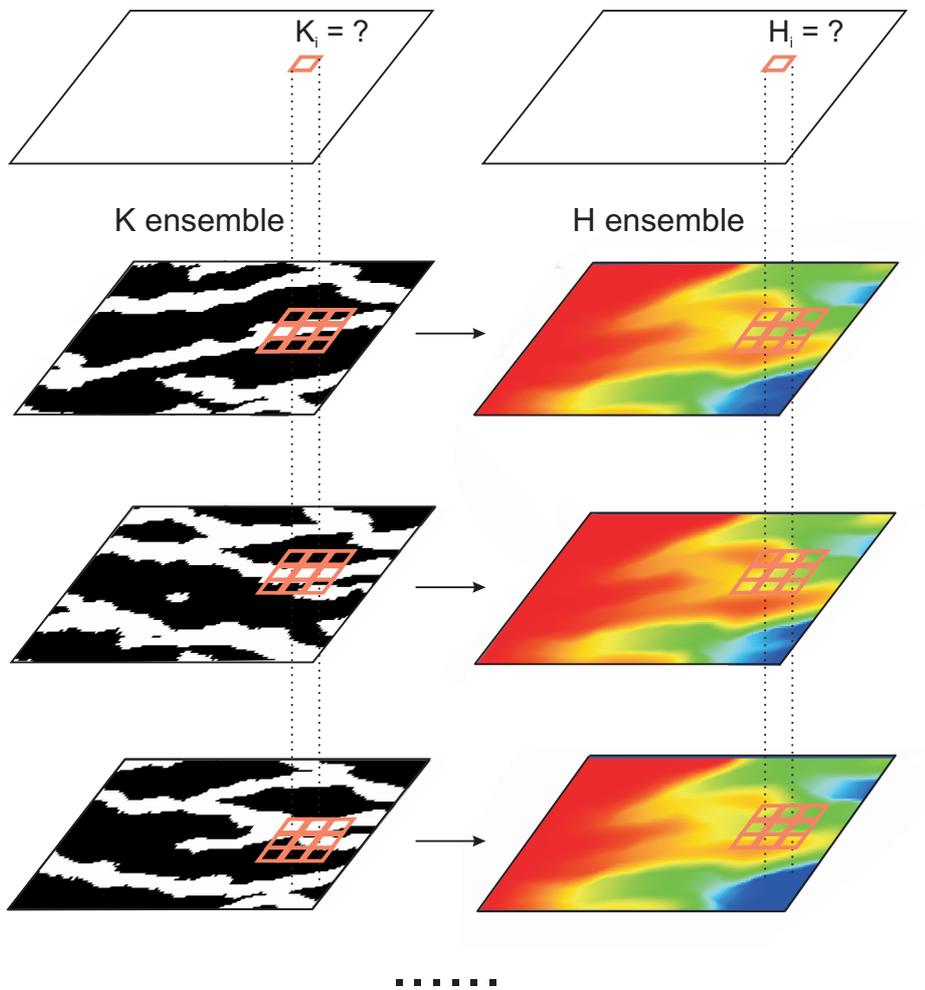
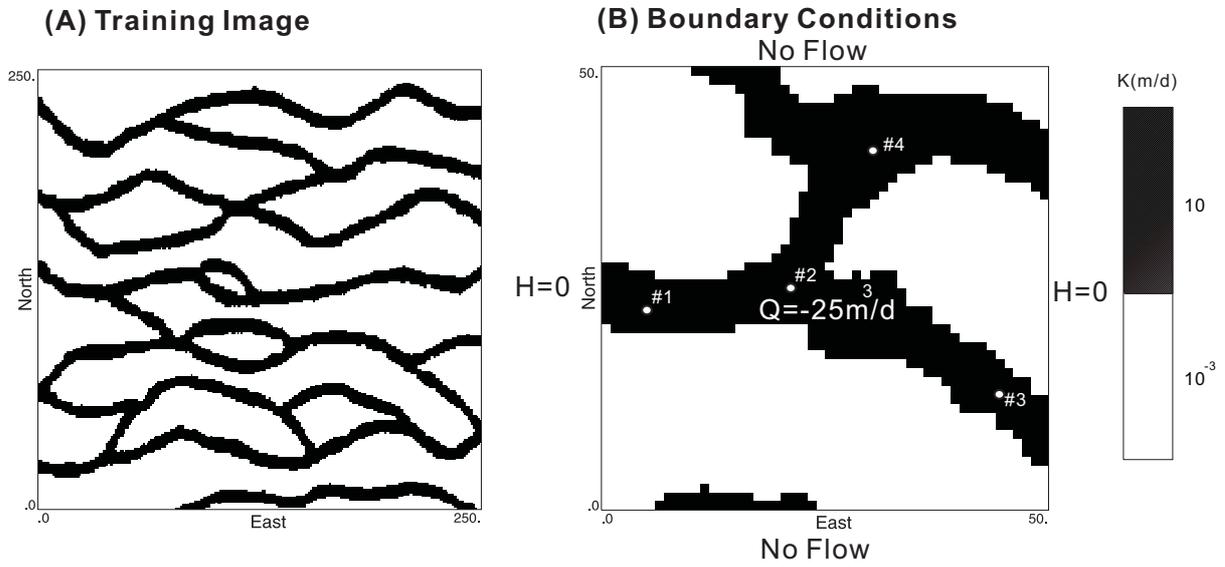Figure 3: Pattern search with ensemble training images.

**(A) Training Image**

**(B) Boundary Conditions**



Figure 4: (A)Training image (B)Reference conductivity, boundary conditions of flow model and observation wells.

Table 1: Parameters used in EnPAT.

| | |
|---|---|
| Search radius for conductivity | 25 m |
| Search radius for head | 25 m |
| Max. number of element in the pattern for conductivity | 10 |
| Max. number of element in the pattern for head | 10 |
| Distance function for conductivity | Manhattan |
| Distance function for head | Weighted Euclidian |
| Distance tolerance for conductivity | 0.0 |
| Distance tolerance for head | 0.0 |
| Number of pilot point | 500 |
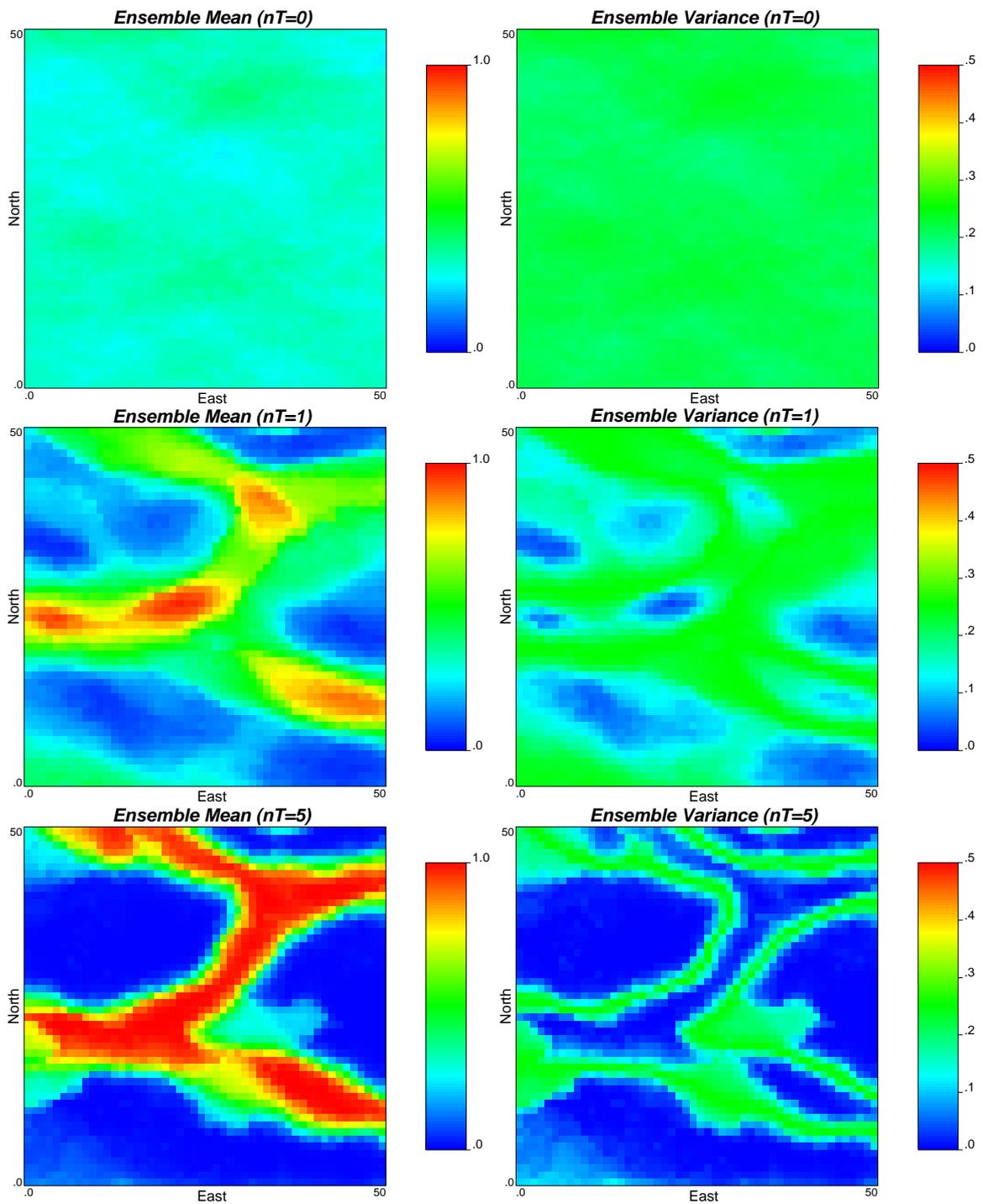
Figure 5: The ensemble mean and variance of conductivity indicator values (i.e., shale with 0 and sand with 1)
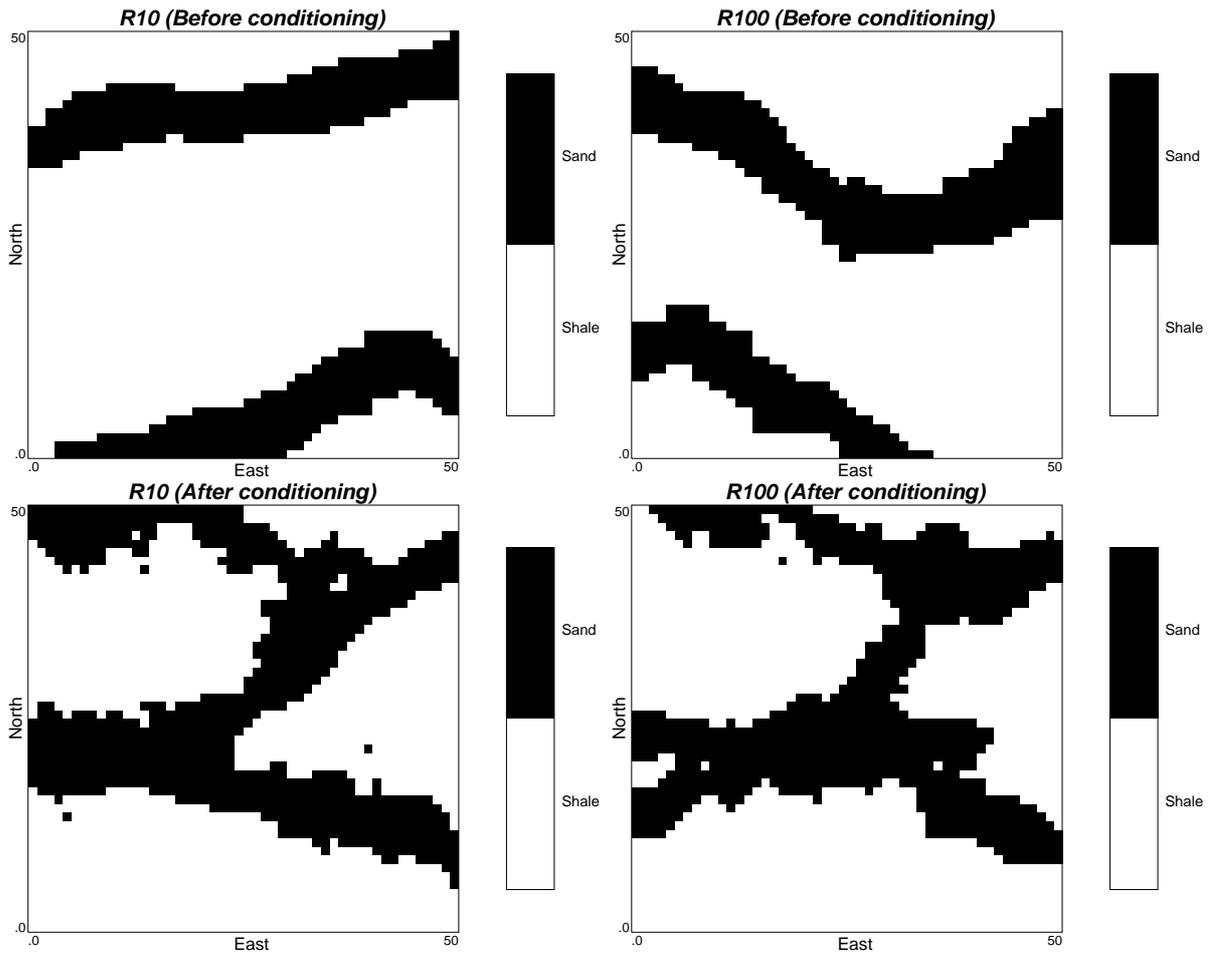
Figure 6: Randomly selected two realizations before and after conditioning
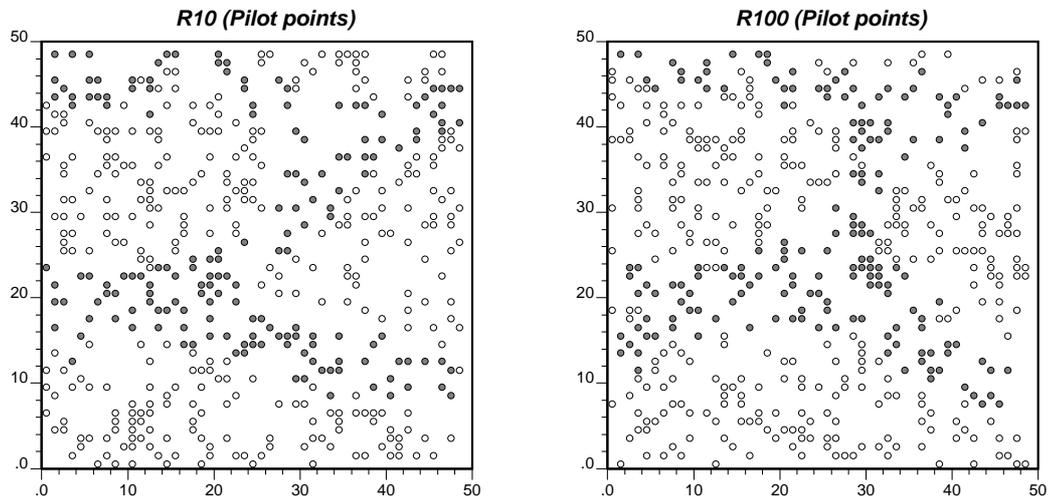


Figure 7: Spatial distribution of pilot points and their estimated values for two realizations (white color means the shale and black color corresponds to sand )
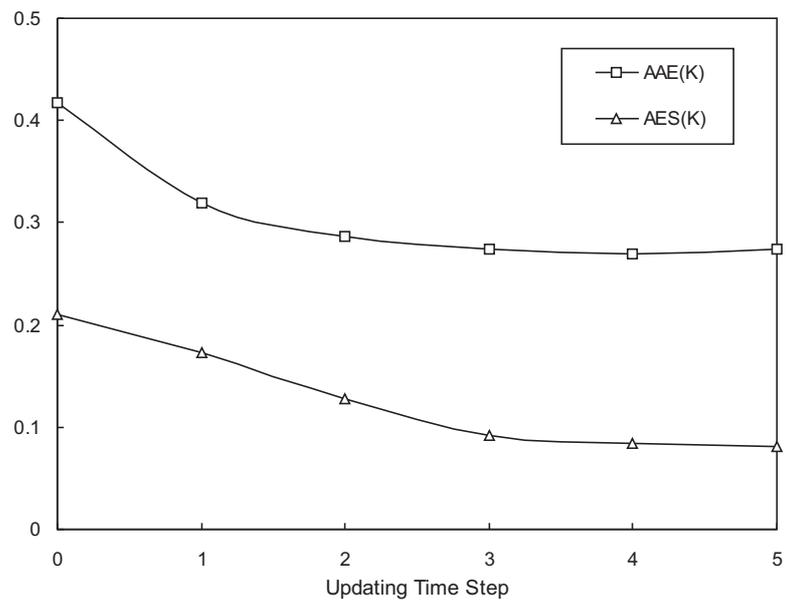
27

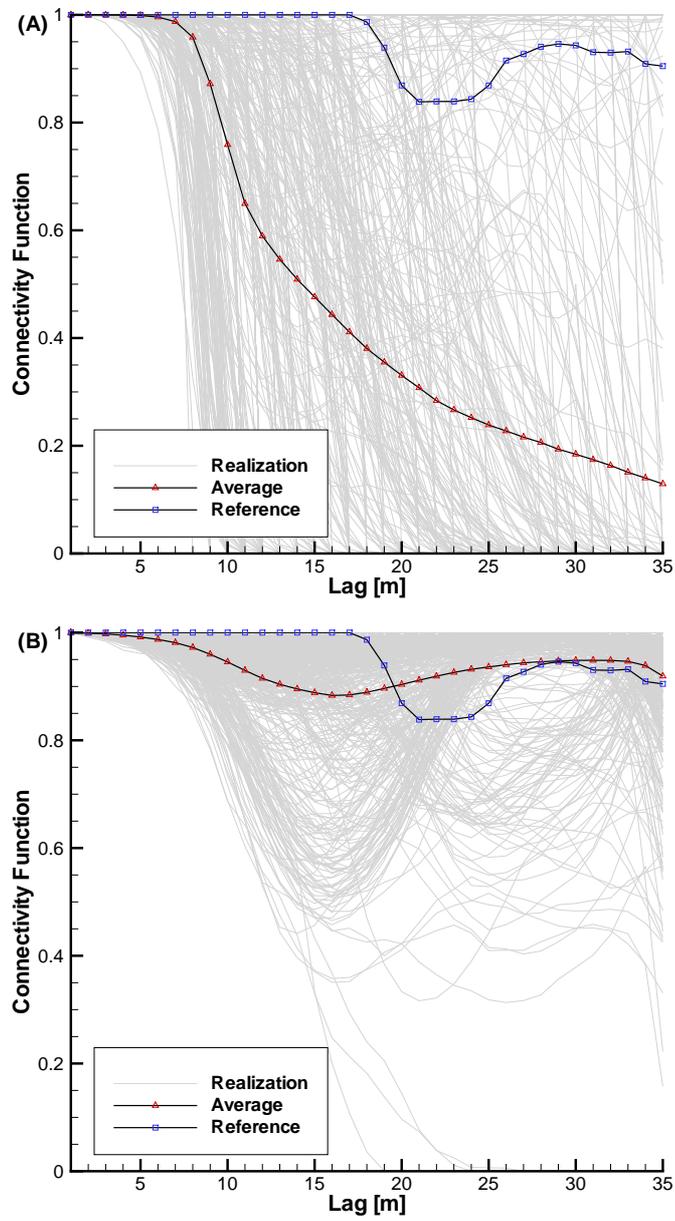Figure 8: The evolution of conductivity AAE and AES with the updating time steps

Figure 9: The ensemble connectivity of conductivity before (A) and after (B) data conditioning
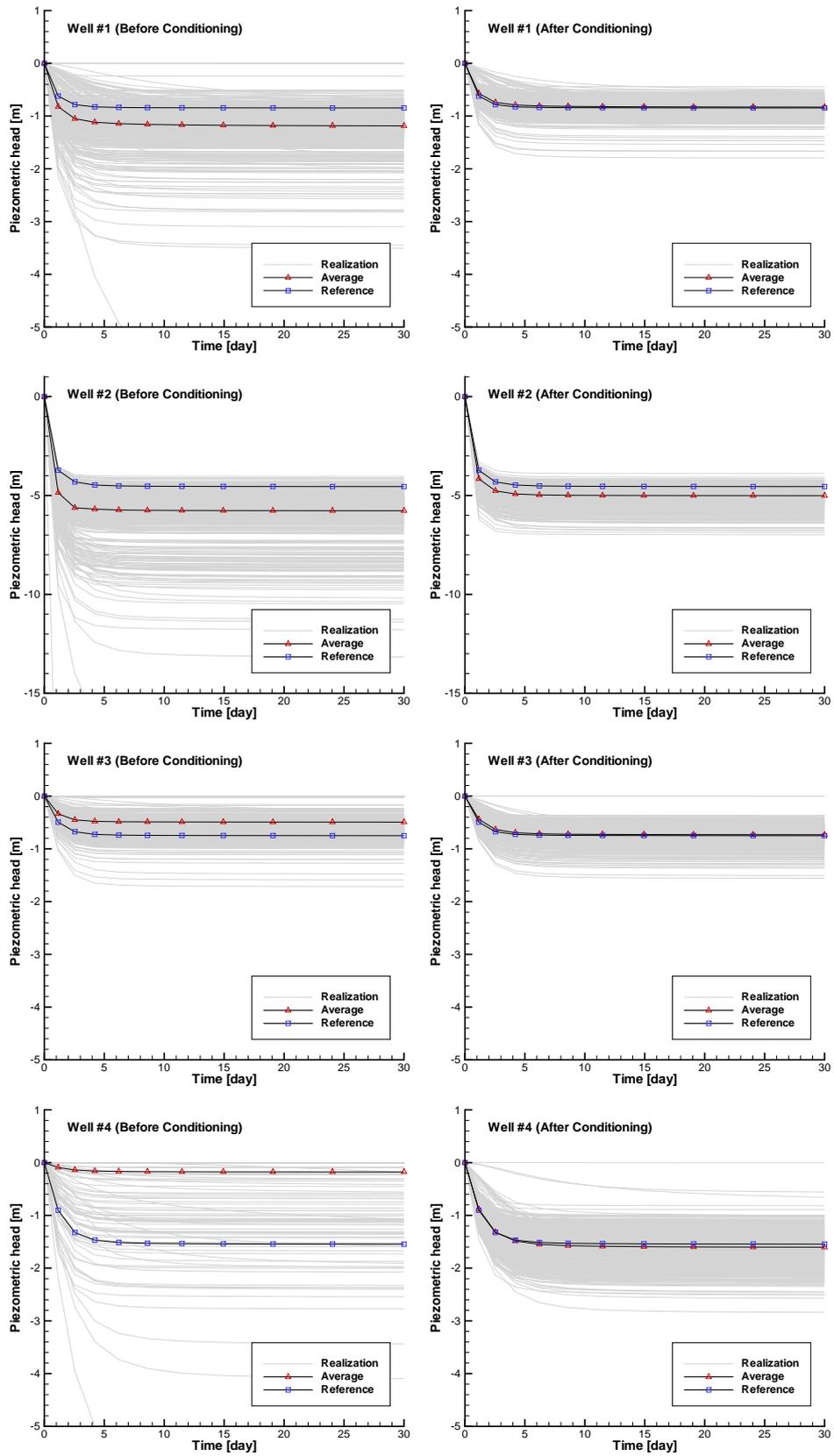
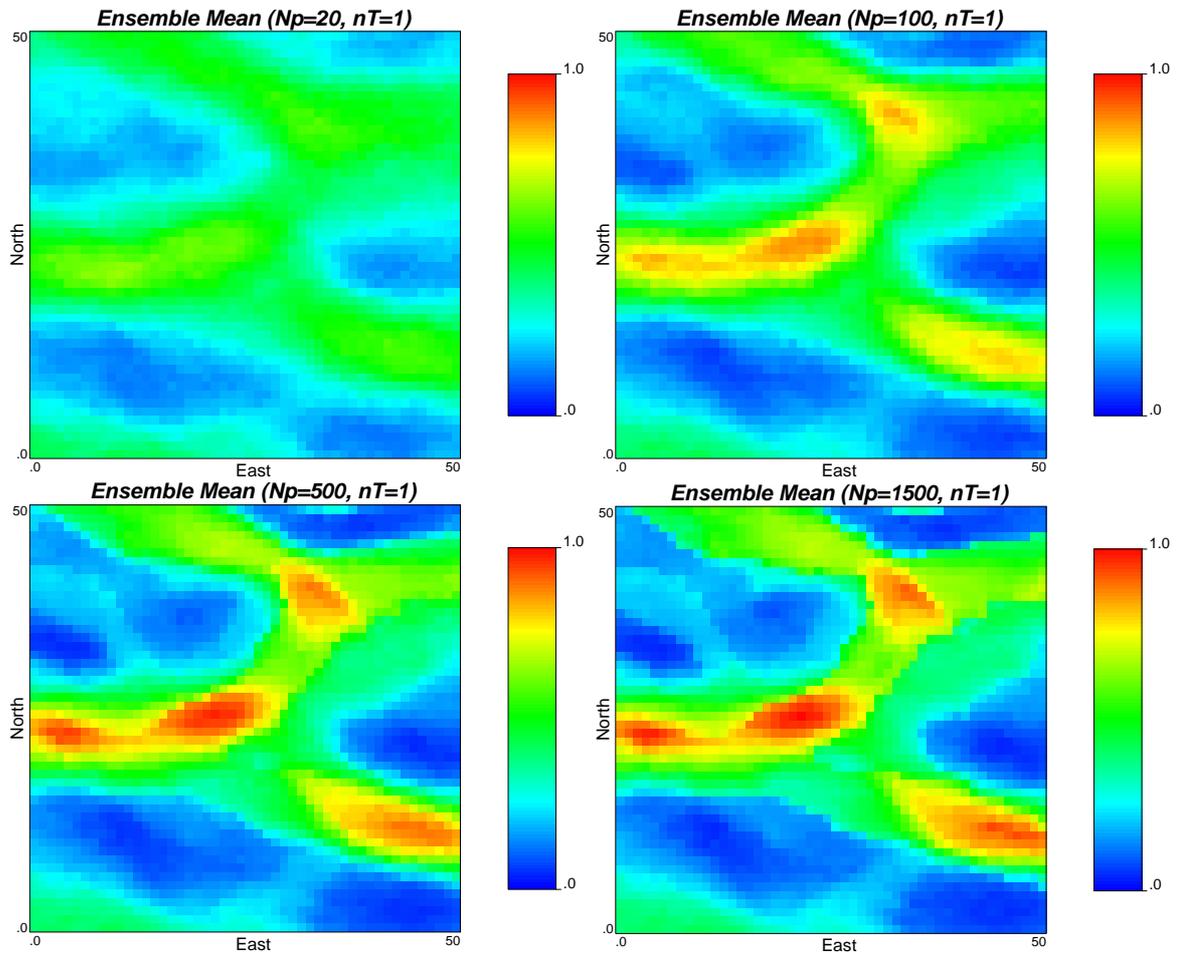Figure 10: The simulate head before and after data conditioning at 4 wells

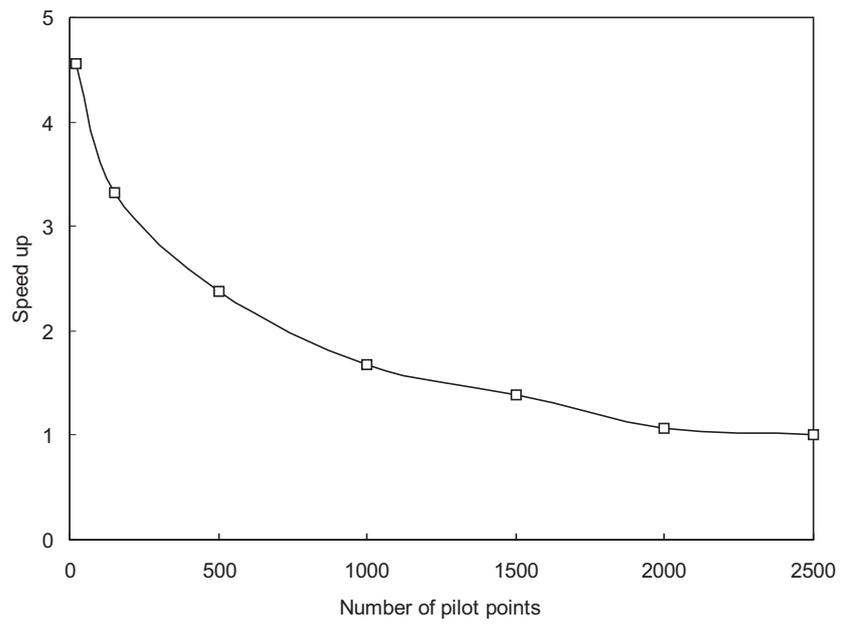Figure 11: The ensemble mean of conductivity using different number of pilot points

Figure 12: The speed up of the EnPAT algorithm using different number of pilot points