Subsurface characterization using a D-optimality based pilot point method
Yong Jung, Ranji S. Ranjithan and G. Mahinthakumar

ABSTRACT
Detailed hydraulic conductivity estimation is a difficult problem as the number of direct measurements available at a typical field site is relatively few and sparse. A common approach to estimate hydraulic conductivity is to combine direct hydraulic conductivity measurements with secondary measurements such as hydraulic head and tracer concentrations in an inverse modeling approach. Even with secondary measurements this may constitute an underdetermined (or over-parameterized) inverse problem giving rise to ‘non-unique’ and incorrect estimates. One approach to reduce over-parameterization is to estimate hydraulic conductivity at a few carefully chosen points called ‘pilot points’ (i.e. reduction in parameter space). This paper develops a D-optimality based criterion method (DBM) for pilot point selection and tests its effectiveness for estimating hydraulic conductivity fields using several synthetic cases. Results show that the selected pilot points using this approach lead to a more accurate hydraulic conductivity characterization than either random or sequential pilot point location selection methods.

Key words | characterization, groundwater inverse modeling, hydraulic conductivity, pilot point method

INTRODUCTION
Inadequate representation of hydraulic conductivity or permeability of the subsurface is one of the greatest sources of uncertainty in developing good groundwater flow and transport models for field applications. Unfortunately, very few direct hydraulic conductivity measurements are available for most real field cases. Given this reality, considerable research has been expended during the last few decades in developing methods that can improve hydraulic conductivity estimates from secondary measurements such as hydraulic head or tracer concentrations. These methods fall under the category of inverse modeling methods as the unknown true hydraulic conductivity field is an input to a forward groundwater and transport models while the known hydraulic heads and/or tracer concentrations are outputs from the forward model. Given the potentially large parameter space and lack of adequate and accurate secondary measurements (observation error), combined with the uncertainties in prior information and forward models (model error), these inverse problems are generally ill-posed suffering from non-uniqueness and instability. Reducing the parameter space by estimating only a few unknown hydraulic conductivities can alleviate the ill-posedness of the inverse problem. For example, Stallman (1956) introduced zonal representation of a hydraulic conductivity field, which has been applied by several others (e.g. Carrera & Neuman 1986a, b). An alternate approach is to use a geostatistical representation of the hydraulic conductivity field (Kitanidis & Vomvoris 1983; de Marsily et al. 1984; Certes & de Marsily 1991; Sun et al. 1995). In addition, regularization can be employed to the objective function to further reduce non-uniqueness (Emsellem & de Marsily 1971; Carrera & Neuman 1986a, b). A number of texts and publications including Yeh (1986), Carrera & Neuman (1986b), Dietrich & Newsam (1990), and McLaughlin & Townley (1996) provide excellent reviews on this topic.
Pilot point methods (PPM) offer a promising alternative to zonal models and an improvement over traditional geostatistical models. This method was first proposed by de Marsily et al. (1984) where a reduced set of calibration points called ‘pilot points’ are selected from the model domain at which the hydraulic head is produced from estimated hydraulic conductivity to match the head measurements. A number of researchers have proposed extensions to the basic PPM and applied it to different situations in terms of reducing instability (i.e. large variance of estimated parameter values) that results from over-parameterization of the hydraulic conductivity field (Certes & de Marsily 1991; RamaRao et al. 1995; Cooley 2000; LaVenue & de Marsily 2001; Doherty 2003; Alcolea et al. 2006; Singh et al. 2008; Castagna & Bellin 2009; Le Ravalec-Dupin 2010). Comparative studies of PPM with other nonlinear inverse approaches can be found in Keidser & Rosbjerg (1991), McLaughlin & Townley (1996), Zimmerman et al. (1998), Floris et al. (2001) and Carrera et al. (2005).

The choice of the number of pilot points and their locations are major factors in the effectiveness of PPM. In general, the number of pilot points should be less than or equal to the number of secondary measurements available to prevent over-parameterization and non-uniqueness. Beyond this constraint, there is no fixed rule on the optimal number of pilot points for a given scenario. One may fix the number and location of pilot points during the whole procedure (de Marsily et al. 1984; Fasanino et al. 1986; Certes & de Marsily 1991; Romero & Carter 2001; Doherty 2005) or iteratively add the pilot points while searching for their locations (LaVenue & Pickens 1992; LaVenue et al. 1995; RamaRao et al. 1995; LaVenue & de Marsily 2001). Both approaches have been widely used. RamRao et al. (1995) suggested that sequential addition is superior to simultaneous selection. For the number of pilot points, RamRao et al. (1995) and LaVenue et al. (1995) suggested that a lower number of pilot points results in better performance, since it is less likely to give large fluctuations in the hydraulic conductivity field. Furthermore, de Marsily (1978) suggested that the number of pilot points should be less than the number of measured hydraulic conductivity values. However, Capilla et al. (1997) and Doherty (2003) pointed out that more pilot points might yield a more realistic hydraulic conductivity field. Thus, as suggested by Chavent & Bissell (1998), there is usually an optimal number of pilot points, as too few pilot points may dishonor the secondary measurements and too large a number may cause over-parameterization. In our opinion, the optimal number of pilot points would be the minimal number required to honor the secondary measurements such as head and concentrations. This would minimize the chance of having too many correlated points (i.e. points that can invoke similar changes in secondary measurements, thus leading to potential non-uniqueness by redundancy). Having a minimal number of highly sensitive points is also less likely to destroy the original covariance structure of the measured hydraulic conductivities. Preserving the covariance structure might be an important consideration in cases where there are a significant number of measured hydraulic conductivity values.

Pilot point locations can be selected based on a number of criteria including empirical and sensitivity considerations. Several authors including de Marsily et al. (1984), Fasanino et al. (1986), Certes & de Marsily (1991), Doherty (2003) and Hernandez et al. (2003) used empirical considerations such as (1) density of uniformly distributed points and (2) geological indication of large amounts of heterogeneity (e.g. high gradient of measured hydraulic heads). An efficient sensitivity-based approach to select pilot point locations was introduced by LaVenue & Pickens (1992). They employed adjoint sensitivity techniques that calculate sensitivities of hydraulic heads with respect to hydraulic conductivities at each potential pilot point location. These sensitivities are ranked and the locations with the highest sensitivities are successively added to the pool of pilot points until no further improvement is possible. The resulting method is known as the sequential search method (SSM). Slight variations of the SSM method have also been used in other studies (LaVenue et al. 1995; RamaRao et al. 1995; Hendricks-Franssen 2000).

The remainder of this paper is organized as follows. The next section describes the pilot point method developed in this paper. The third section provides a description of the computational components, including genetic algorithm, geostatistics, and the groundwater flow model. The fourth describes the two test problems along with the results and discussion for each case. Finally, major conclusions are summarized in the final section.
PILOT POINT METHODOLOGY

The basic approach of the PPM proposed by de Marsily et al. (1984) is as follows: (1) initialize the hydraulic conductivity field using a geostatistical method (ordinary/universal kriging) based on measured hydraulic conductivity values, (2) select pilot point locations where no values of measured hydraulic conductivity are available, (3) iterate to find optimal values of hydraulic conductivity at selected pilot point locations by minimizing an objective function (generally, the sum of squared differences between observed hydraulic heads and calculated hydraulic heads obtained using a forward model based on the kriged hydraulic conductivity field) and (4) obtain the final hydraulic conductivity field when stopping criteria are met. Based on the given approach, the PPM developed in this paper involves two steps: (1) selecting the pilot point locations (next subsection) and (2) determining hydraulic conductivities at the pilot point locations (the second subsection). Prior to performing these two steps, a synthetic hydraulic conductivity field (‘true $K$ values’) and an ‘initial hydraulic conductivity field’ (prior information) are generated for testing purposes. In this study the true $K$ field is generated by kriging randomly generated hydraulic conductivity values at selected points. The initial hydraulic conductivity field (i.e. observed $K$ field or prior $K$ field) is generated by excluding a subset of the original hydraulic conductivity measurements and then kriging these values. The hydraulic head measurements at the observation locations are generated by performing a forward groundwater flow simulation using the true $K$ field. In step 1 of the PPM algorithm, starting with the initial hydraulic conductivity field, a genetic algorithm is used to search for the locations of pilot points that give the maximum D-optimality value (next subsection). In step 2, the hydraulic conductivities at the selected pilot point locations are perturbed to minimize the difference between the observed and calculated hydraulic head values (the second subsection). The calculated head values are based on the kriged hydraulic conductivity field that combines the original hydraulic conductivity measurements with the pilot point perturbed hydraulic conductivity values. A schematic of this procedure with step 1 on the left-hand side and step 2 on the right-hand side is illustrated in Figure 1. As shown in Figure 1, the two separate procedures are sequentially carried out.

Pilot point location procedure – D-optimality criterion

It is desirable to select the most sensitive locations with least correlation as pilot points because this will maximize the data worth of the hydraulic head measurements. In other words, these would be the most unique set of points (i.e. minimal redundancy) that require minimal perturbations of hydraulic conductivities to match hydraulic head observations. A criterion that satisfies these conditions is the D-optimality criterion mathematically expressed as (Kiefer & Wolfowitz 1959; Fedorov & Pazman 1968; Knopman & Voss 1987):

$$\text{MAXIMIZE} \quad obj\_1 = \det[X^TX]$$

where
- $obj\_1$: objective function in Step 1
- $X$: sensitivity matrix
- $h$: hydraulic head observation
- $n$: number of hydraulic head observations
- $K$: hydraulic conductivity values at potential pilot point locations
- $m$: number of selected pilot points.

A key quantity in this procedure is the sensitivity matrix $(X)$ shown in Equation (2) that measures the sensitivity of
hydraulic heads with respect to hydraulic conductivities at potential pilot point locations. As indicated in Figure 1, in every step of the algorithm, the initial $K$ field (kriged field from the prior $K$ observations) is perturbed to calculate the sensitivity values in $X$. In this paper, the sensitivity matrix $X$ is calculated using central finite differences using a small $K$ perturbation ($\Delta K$). Since the calculation of D-optimality does not involve the inversion of the sensitivity matrix, finite-differencing, though approximate, is generally sufficient for this purpose. The value of $\Delta K$ was chosen as unity after a detailed sensitivity study involving a wide range of $\Delta K$ values.

Equation (1) implies that the determinant of the Fisher information matrix ($X^TX$), which is known as D-optimality (Knopman & Voss 1987) is being maximized. Instead of maximizing $\text{obj}_2$ in Equation (1), one may choose to minimize $-\log|\det(X^TX)|$, especially if the resulting values are extremely small. The rationale for D-optimality is as follows. If the Fisher information matrix is $A$, then the variance–covariance matrix is given by $B = A^{-1}$, and the correlation matrix $C$ is given by matrix $B$ scaled both column-wise and row-wise by the square root of its respective diagonal entries (Hill & Osterby 2003). Thus the Fisher information matrix $A$ is proportional to the inverse of correlation matrix $C$. Also, maximization of the determinant of the Fisher information matrix is mathematically equivalent to the problem of minimizing the norm of the variance–covariance matrix $B$ (Knopman & Voss 1987). The set of pilot points chosen in this manner implies that these are the least correlated set of points in terms of hydraulic head response at the head observation locations to hydraulic conductivity changes. In other words, changes in hydraulic conductivity at any two points from this set will invoke dissimilar changes in head values at the head observation points, thereby reducing any potential redundancy among pilot points. In an optimization sense, these sets of points are less prone to non-uniqueness as we are more likely to find a unique set of hydraulic conductivity values at these points to match the head observations. As opposed to minimizing the determinant or norm of the covariance matrix, the D-optimality criterion (i.e. maximizing the determinant of the Fisher information matrix) has two additional advantages. First, it does not require calculation of the correlation matrix, thus resulting in significant computational savings and smaller impact of round-off errors. Second, D-optimality also ensures that the pilot points selected are highly sensitive to the head observations. Higher sensitivity means that smaller perturbations are required at the pilot points to match the head observations, thus maximizing the data worth of the head observations and minimizing the impact of observation errors. Smaller perturbations in $K$ values at the pilot points also attempt to preserve the original covariance structure of the prior $K$ field. It is worth noting that alternative approaches exist for minimizing the deviation from the prior $K$ field including regularization and incorporation of this as a separate objective in a multi-objective problem (e.g. Singh et al. 2008).

Similar to the D-optimality criterion, the sequential search method (SSM) employed by LaVenue & Pickens (1992) and RamaRao et al. (1995) using adjoint sensitivity also ensures that the points selected are maximally sensitive. However, unlike D-optimality, SSM does not ensure that the combined set of pilot points is least correlated, as described earlier. Due to these advantages, D-optimality has been used for a number of applications including sampling design for parameter estimation in groundwater (Knopman & Voss 1987) and characterization of water distribution systems (Bush & Uber 1998). To our knowledge, it has not been used in pilot point selection.

To evaluate the D-optimality criterion, the number of pilot points ($m$ in Equation (2)), need to be known a priori. In the absence of good prior information, the number of pilot points should be less than or equal to the number of head measurements. Using a larger number would lead to redundant pilot points and the matrix $X$ will be rank-deficient (number of unknowns $m >$ number of rows $n$). This would likely make the Fisher information matrix ($X^TX$) to be linearly dependent, forcing its determinant (D-optimality) to be essentially zero. A number $m$ that is too small will not make efficient use of the observed head values. Based on these observations and preliminary tests conducted with different numbers of pilot points, in this study we chose the number of pilot points to be less than the total number of head measurements (see the fourth section).

A Genetic Algorithm (GA) is used to search for the pilot point locations. For a given trial set of pilot points (all points in a set are considered simultaneously), $\text{obj}_1$ is calculated and GA is employed to find a set that maximizes this value.

**Determination of hydraulic conductivity values**

Once the pilot point locations are determined, the hydraulic conductivity values need to be determined at these locations.
from the available hydraulic head measurements. This inverse problem can be formulated as an optimization problem. The objective is to minimize the difference between the observed and the simulated hydraulic head values by perturbing the hydraulic conductivities at the pilot point locations. This difference can be quantified using a sum of squared residuals and the simulated hydraulic head values by perturbing the objective is to minimize the difference between the observed

\[
\text{MINIMIZE } \text{obj} \_2 = \sum_{i=1}^{n} \left( h_{i}^{\text{observed}} - h_{i}^{\text{calculated}} \right)^{2} \tag{3}
\]

where

\begin{align*}
\text{obj} \_2 & \text{: objective function in Step 2} \\
 h_{i}^{\text{observed}} & \text{: observed hydraulic head value at observation point } i \\
 h_{i}^{\text{calculated}} & \text{: calculated hydraulic head values at observation point } i.
\end{align*}

For steady-state problems with no source/sink terms, hydraulic heads are primarily sensitive to spatial variations of hydraulic conductivity rather than their absolute magnitudes (i.e. the hydraulic conductivity field can be rescaled by a constant value without any change in hydraulic head values). Thus it is customary in these cases to constrain the magnitude of hydraulic conductivity to realistic ranges based on prior information (e.g. LaVenue & Pickens 1992). Incorporating prior information can also alleviate the over-parameterization problem to an extent (i.e. the number of pilot points can be greater than the number of head observations Alcolea et al. 2006). In this paper, we use the geometric mean of measured hydraulic conductivity values to restrict the hydraulic conductivity search space. To accomplish this, we add a penalty to the objective in Equation (3) if the geometric mean of the hydraulic conductivity values at the pilot points \( k_{gm} \) deviates from the geometric mean of the measured hydraulic conductivity values \( k_{gm} \) by more than a prescribed relative value \( \nu \). Application of this penalty can be expressed as

\[
\delta = \frac{|k_{gm} - k_{gp}|}{k_{gm}} \\
\text{if } \delta > \nu \rightarrow \text{obj} \_2 = \text{obj} \_2 + \delta \tag{4}
\]

where \( k_{gm} \): geometric mean of measured hydraulic conductivities \( k_{gm} = \left( \prod_{i=1}^{nk} k_{i}^{p} \right)^{1/nk} \)

\( k_{gp} \): geometric mean of estimated hydraulic conductivities \( k_{gm} = \left( \prod_{j=1}^{m} k_{j}^{p} \right)^{1/m} \)

\( \nu \): measure of allowable deviation from measured hydraulic conductivity

\( nk \): number of measured hydraulic conductivity values

\( k_{i}^{p} \): measured hydraulic conductivity value at point \( i \)

\( k_{j}^{p} \): estimated hydraulic conductivity at pilot point \( j \)

A larger value of \( \nu \) would give less importance to prior information. In this study \( \nu \) is assumed to be 0.75 (i.e. no more than a 75% deviation from the prior mean is allowed). This value was chosen after a preliminary study involving a range of \( \nu \) values. Again, GA is used to search for the hydraulic conductivity values according to Equation (3) or (5). Instead of using the penalty approach described by Equation (4), one might consider using a regularization term in the objective function (4) to incorporate the prior information. In this case, Equation (5) would be modified as

\[
\text{obj} \_2 = \sum_{i=1}^{n} \left( h_{i}^{\text{observed}} - h_{i}^{\text{calculated}} \right)^{2} + \alpha \sum_{j=1}^{m} \left( k_{j}^{p} - k_{gm} \right)^{2} \tag{5}
\]

where \( \alpha \) is a regularization coefficient that weights the importance to the prior information. Unless normalization is used, \( \alpha \) is subjective to the relative magnitudes of \( h \) and \( k \) values. Also, since the performance of this approach is highly sensitive to the parameter \( \alpha \) (more so than \( \nu \) in Equation (4)), even a slightly larger value of \( \alpha \) may devalue the worth of information of the head observations which we have already maximized with the D-optimality criterion.

**COMPUTATIONAL COMPONENTS**

The pilot point method developed in this paper uses three main computational components: (a) a genetic-algorithm-based search method for locating pilot points and determining hydraulic conductivity values, (b) ordinary kriging for interpolation of \( K \) values and (c) a forward groundwater flow
model for determining hydraulic head values from the K field. For simplicity, these computational components are implemented in a MATLAB environment.

**Genetic Algorithm (GA)**

In estimating the hydraulic conductivity, many applications of PPM have used gradient-based search methods (de Marsily et al. 1984; Fasanino et al. 1986; LaVenue et al. 1995; RamaRao et al. 1995; Wen et al. 1998b). A potential shortcoming of gradient-based search methods is that they are local search methods; this might cause it to converge to a local minimum. Gradient-based search methods are either the pilot point locations (integer node indices in the grid) or hydraulic conductivity values. For simplicity, these computational components are usually encoded as real or binary strings. In this paper, the decision variables are either the pilot point locations (integer node indices in the grid) or hydraulic conductivity values. For simplicity, a real representation is used for these decision variables. The built-in genetic algorithm function in the MATLAB GA toolbox (now part of the Global Optimization Toolbox available online at http://www.mathworks.com/products/global-optimization/) is used in this paper.

Heuristic unconstrained search optimization approaches such as GA frequently enforce constraints by the use of penalty functions. These are simply penalty values added to the objective function to encourage the search towards a feasible space. Since the version of the MATLAB GA toolbox used in this paper (i.e. the 2007a version) did not support constraints, penalties were used to enforce constraints. In the pilot point location search procedure, bound constraints limiting the pilot points to nodes in the computational domain are enforced by adding a penalty wherever this is violated. The measurement locations were avoided in the search by encoding the decision variables (pilot point locations) by excluding the measurement locations. The D-optimality criterion automatically ensures that no two pilot points are the same since the value of D-optimality in this case would be zero.

**Kriging**

Kriging is an interpolation tool used to obtain unbiased predicted values of hydraulic conductivities over a set of grid points based on observed correlated values at certain points. In this study, “ordinary kriging” is employed to interpolate the K field across the entire domain (i.e. all grid points) from measured and/or pilot point estimated values. By using ordinary kriging we assume intrinsic stationarity. Intrinsic stationarity implies that the mean of the process is constant and the variance of the difference of two response values depends only on the distance between the locations (Schabenberger & Gotway 2005). More specifically, if \( Z(s) \) represents hydraulic conductivity at location \( s \) in domain \( D \), then \( \{Z(s), s \in D\} \) is said to be intrinsically stationary, if (i) \( \mathbb{E}[Z(s)] = \mu, \forall s \in D \) and (ii) \( \text{Var}[Z(s_1) - Z(s_2)] = \gamma(s_1 - s_2), \forall s_1, s_2 \in D \).

For this study, the exponential model is applied for the semi-variogram \( \gamma(h) \):

\[
\gamma(h) = c_0 + c_1 \left(1 - e^{-\frac{h}{a}}\right)
\]

where the sill, \( c_1 \), represents the maximum value of semi-variogram; \( h \) is the separation distance given by \( |s_1 - s_2| \); the range, \( a \), denotes the correlation scale of observations: where beyond this distance the observations are least correlated; and \( c_0 \) is the nugget. We assume that the observed hydraulic conductivities are log-normally distributed. The initial hydraulic conductivity field is generated by kriging the observed K-field data transformed in the log-normally distributed space. Then, the hydraulic conductivities are perturbed at the pilot point locations to honor the observed hydraulic head values. Kriging is reapplied using the perturbed values at the pilot points and measured conductivity values to produce the final estimate of hydraulic conductivity values. The kriging model used in this study was implemented.
in MATLAB and validated with the ‘R’ package (Venables et al. 2004). Figure 2 shows a sample semi-variogram and kriged K field used for Model Problem II (see the fourth section) employed in this study. For this case, \( c_0 = 0, c_1 = 0.862 \) and \( a = 3.402 \).

**Forward groundwater flow model**

The forward groundwater flow model is based on a central difference approximation of the two-dimensional steady-state groundwater flow equation:

\[
\frac{\partial}{\partial x} \left( K b \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left( K b \frac{\partial h}{\partial y} \right) + q = 0 \tag{7}
\]

where
- \( K(x,y) \): isotropic, heterogeneous, conductivity field (m/d)
- \( h(x,y) \): head field (m), \( q \): specific discharge (source is positive and sink is negative) (m/d)
- \( x,y \): spatial locations (m)
- \( b \): constant aquifer thickness.

This forward model is used to obtain the head field \( h(x,y) \) for a given hydraulic conductivity field \( K(x,y) \). The groundwater flow model is implemented in MATLAB and validated with MODFLOW solutions involving a variety of boundary conditions.

**TEST PROBLEMS AND RESULTS**

Two model problems are tested. The first problem involves a smaller domain with no source/sink terms. The second problem involves a larger domain with a single extraction well.

**Model problem I**

Model problem I consists of a two-dimensional rectangular domain of size 700 m by 1000 m with uniform grid blocks of size 100 m by 100 m. The aquifer thickness is assumed to be 10 m (\( b \) in (7)). The steady state groundwater flow field is based on a confined, isotropic and heterogeneous aquifer. The true \( K \) field is generated by interpolating (kriging) nine randomly generated hydraulic conductivities at random points. Based on these hydraulic conductivity measurements, the hydraulic head measurements are generated at four arbitrary observation locations using the groundwater flow model. The nine hydraulic conductivity measurement points (squares) and the four hydraulic head observation points (triangles) are shown in Figure 3(a). The contour map of the true \( K \) field is shown in Figure 3(b).

To generate different initial \( K \)-field scenarios for the numerical experiments, four hydraulic conductivities are selected from the original nine measurements by excluding five measurements. Two different synthetic \( K \) fields (Figure 4) with two different flow conditions (Figure 5) constituting a total of four scenarios are used in the experimental study. Scenarios 1 and 2 correspond to the case with mean uniform flow and scenarios 3 and 4 correspond to the case with mean non-uniform flow. For scenarios 1 and 2 the flow field is obtained by applying constant head boundary conditions at the left (upstream) and right (downstream) boundaries and no-flow boundary conditions at the other sides. For scenarios 3 and 4 the flow field is obtained by applying constant head boundary conditions at the left (upstream) and bottom (downstream) boundaries and no-flow boundary conditions at the other sides. The upstream and downstream boundaries have 20 m and 15 m constant hydraulic heads, respectively.

\[\text{Figure 2} \quad \text{An example of kriging results in this study. (a) semi-variogram graph with sill and range (x axis: distance (m), y axis: semi-variogram (m/d)^2). (b) Surface plot of kriged hydraulic conductivity field (x and y axis: grid points (100 m x 100 m blocks), z axis: hydraulic conductivity values (m/d)).}\]
Of the four scenarios, the initial $K$ field shared by scenarios 2 and 4 is significantly different from the true $K$ field when compared to the one used by scenarios 1 and 3.

The hydraulic conductivity error metric ($K_{error}$), as defined by Equation (8), is used to evaluate the efficacy of each method in matching the true hydraulic conductivity field:

$$K_{error} = \frac{1}{N_t} \sqrt{\sum_{i=1}^{N_t} (K_{ci} - K_{ri})^2}$$

where
- $N_t$: total number of grid points
- $K_{ci}$: calculated hydraulic conductivity values at grid point $i$
- $K_{ri}$: true hydraulic conductivity value at grid point $i$.

The ‘ga’ function in the MATLAB GA toolbox is used to search for the pilot point locations (Step 1) and hydraulic conductivity values at the selected pilot point locations (Step 2). Real encoding is used for representing the populations in these experiments. The GA parameters used in this study are listed in Table 1. Computations are carried out on a PC with a Pentium 4 CPU, 3.0 GHz clock speed and 1.0 GB RAM. The typical simulation time for each trial for this problem is about 2 h (approximately 110 000 groundwater flow simulations).

**Pilot point locations**

Based on rank discussed earlier in the second section, the number of pilot points was fixed at three (one less than the total number of head measurements) throughout the search process. Fixing this number significantly simplifies the search process for the pilot point locations. The locations of these three pilot points are then searched using a GA based on the D-optimality criterion (Equation (1)). The measurement points are excluded from the search as described earlier in
this section. The final selected pilot point locations (displayed as circles) are presented in Figure 6 for the four scenarios. The final pilot point locations are reasonably well distributed in the given domains for all four scenarios. This is probably due to the fact that the D-optimality criterion automatically ensures good coverage (or distribution), as pilot points that are close together would likely be correlated thus resulting in smaller D-optimality values.

The final D-optimality values for the pilot point sets from the GA search are presented in Table 2 for each scenario. One might notice that these values are extremely small. This is plausible since hydraulic heads are known to have very low sensitivities to the magnitude of hydraulic conductivities for steady state problems in the absence of source/sink terms as all $K$ values can be rescaled by a constant with little or no change in head values. However, hydraulic heads are sensitive to the variability of hydraulic conductivity (i.e. relative differences in $K$ values) that is exploited when the pilot points are selected in a collective fashion.

To evaluate the effectiveness of GA in searching pilot point locations with high D-optimality values, we calculated D-optimality values of 30 random pilot point sets (three points in each set) in each scenario and compared it to the values obtained by DBM with different starting random seeds for GA search. For all four scenarios, the D-optimality values for the pilot points searched by GA are one or two orders of magnitude greater than those for a random set of pilot points. This indicates that GA is effective in searching for pilot points with high D-optimality values.

To evaluate the effectiveness of the D-optimality metric in DBM, two additional metrics are calculated: (1) Sum-Corr = sum of correlations between individual pilot point locations (sum of the absolute values of the off-diagonal entries of the correlation matrix $C$ described in the second section) and (2) NormSens = norm of sensitivity matrix. The SumCorr metric measures the correlation among the selected set of pilot points. A low SumCorr value is better for estimation as this will minimize non-uniqueness. The NormSens metric measures the sensitivity of the pilot points to the measured hydraulic heads. A high NormSens metric is better for estimation as this will reduce the amount of perturbation required to match the measured head values. Figure 7 compares these metrics for DBM pilot points and random pilot points for the four scenarios for 30 random trials. In most cases, the pilot points from DBM are less correlated with significantly larger sensitivities. Occasionally randomly selected pilot point locations provide less correlated locations but with insignificant sensitivity values. In other cases, the random pilot point locations are highly sensitive to the

### Table 1 | Chosen genetic algorithm options in MATLAB for Model problem I

<table>
<thead>
<tr>
<th></th>
<th>First procedure (location search)</th>
<th>Second procedure (value search)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>Double vector</td>
<td>Double vector</td>
</tr>
<tr>
<td>Selection</td>
<td>Tournament</td>
<td>Stochastic uniform</td>
</tr>
<tr>
<td>Cross over</td>
<td>Scattered (0.6)</td>
<td>Scattered (0.4)</td>
</tr>
<tr>
<td>Mutation</td>
<td>Uniform (0.3)</td>
<td>Uniform (0.3)</td>
</tr>
<tr>
<td>Population size</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>Number of generations</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
hydraulic head measurements, but these locations are strongly correlated with each other. In other words, the DBM-based pilot points provide a good compromise between these two metrics.

**Values at pilot point locations**

Once the pilot point locations are identified, the values of hydraulic conductivity at these points need to be estimated. As discussed in the second section, hydraulic conductivities at the selected pilot points are perturbed to minimize the differences in observed and calculated hydraulic heads (Equation (3)) at observation points. Observed heads correspond to the true $K$ field and the calculated heads correspond to the iteratively estimated hydraulic conductivity distribution in the GA search. Ignoring non-uniqueness effects, the estimated conductivity is close to the true $K$ field when the differences between the head values are small.

**Table 2**  D-optimality values of final selected pilot point locations for D-optimality based criterion method (DBM)

<table>
<thead>
<tr>
<th>Scenario</th>
<th>D-optimality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>$2.18 \times 10^{-10}$</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>$1.00 \times 10^{-14}$</td>
</tr>
<tr>
<td>Scenario 3</td>
<td>$1.17 \times 10^{-9}$</td>
</tr>
<tr>
<td>Scenario 4</td>
<td>$1.26 \times 10^{-14}$</td>
</tr>
</tbody>
</table>

*Figure 6*  | Final selections of pilot points in four scenarios. Axis labels indicate grid indices and each grid block is of size $100 \text{ m} \times 100 \text{ m}$. 
Figure 8 compares the objective function values (Equation (3)) and hydraulic conductivity differences (Equation (8)) for the random pilot points and D-optimality based pilot points for all four scenarios for the 30 trials carried out in the previous subsection. For fair comparison, the same GA parameters are used for the hydraulic conductivity search at the D-optimality based pilot points and random pilot points. Overall, the D-optimality based pilot points yield several orders of magnitude better objective function of heads than random pilot points. However, occasionally, randomly selected pilot point locations give better fitness values. For example, in scenario 3, approximately one-third of the 30 trials result in slightly better performance in terms of objective value for the random case. However, in $K$ error comparison DBM shows better results at the same locations. When the initial hydraulic conductivity distribution is significantly different from the true $K$ field (scenarios 2 and 4), DBM provides more accurate hydraulic conductivity estimation with minimum fluctuations. When comparing uniform flow cases (scenarios 1 and 2) with non-uniform flow cases (scenarios 3 and 4), non-uniform cases result in slightly larger fluctuations in hydraulic conductivity estimates and larger hydraulic head errors for both DBM and Random. Other than that, different flow conditions do not significantly influence hydraulic conductivity estimation for these problem sets. From these results, we can conclude that, in general, the D-optimality based pilot point selection is superior to random selection and is a viable and promising alternative to other methods of pilot point selection.
Influence of magnitude constraint for $K$

The hydraulic conductivity contour maps of the true $K$ field, initial $K$ field and the solutions obtained with and without the magnitude constraint for $K$ (Equation (4)) are presented in Figure 9 for scenario 2. The regularization approach presented by Equation (5) is not evaluated in this paper but it is expected to perform similarly. For this illustration, scenario 2 is chosen over other scenarios since its initial $K$ field is very different from the true $K$ field, thus presenting a more difficult problem. A comparison of these figures indicates that both the constrained and unconstrained searches show a dramatic improvement over the initial $K$ field in predicting the true $K$ field. Also, both methods identify the pattern of the $K$ variation extremely well, particularly on the left side. The head error values of both these cases are three orders of magnitude better (unconstrained: $2.05 \times 10^{-4}$, constrained: $4.17 \times 10^{-4}$) than the value corresponding to the initial $K$ field ($1.50 \times 10^{-1}$). However, average error in the estimated $K$ values is higher for the unconstrained case ($0.7345$) than the constrained case ($0.4605$). Also, the variance of $K$ field between different search trials is also significantly reduced for the constrained case (data not shown). This shows that using prior information to constrain the $K$ values will improve the estimate of the $K$ field. In highly heterogeneous aquifers, however, using this constraint can lead to erroneous estimates of the $K$ field. In these cases, every effort should be made to obtain tracer or flow rate measurements to augment the head measurements in the $K$-field estimation process.
Comparison with sequential search method

Here we compare the effectiveness of the D-optimality based PPM method with the sequential search method (SSM) discussed earlier in the first section (LaVenue & Pickens 1992; RamaRao et al. 1995). This method sequentially repeats the process of searching for points with the highest sensitivity and then adjusting the conductivity value at these locations until no further improvement is possible. The results are presented for scenario 2 in Table 3 and Figures 10 and 11. From Table 3, we see that, while the fitness value (head error) is improving with increasing pilot points, even after four pilot points the fitness value is still three orders of magnitude larger than the D-optimality-based method. Also, there is no significant improvement in solution error ($K$ error) for this example. Figure 10 shows that the pilot points selected using SSM have less spatial coverage than those selected by DBM. The resulting hydraulic conductivity distribution shown in Figure 11 is also significantly different from the true $K$ field. This shows that SSM is less effective than DBM for pilot point selection.

Model problem II

Model problem II involves a larger domain, $5000 \text{ m} \times 2500 \text{ m}$ in a $50 \times 25$ grid arrangement, as shown in
Figure 12(a). The 20 hydraulic conductivity values shown in Figure 12(a) (red squares) is used to generate the true \( K \) field. The flow model consists of two no-flow boundaries at the left and the bottom sides and a constant head boundary at the right side of the domain. The top boundary has a constant water inflow (1.0 m\(^3\)/d) and an extraction well with 1200 m\(^3\)/d is placed near the center (2400 m and 1000 m from the bottom left corner). Using the above conditions we obtained 20 synthetic hydraulic head measurements (green triangles in Figure 12(a)) that are used for estimating hydraulic conductivity. The contour map of the true \( K \) field and the resulting head field are shown in Figure 12(b) and 12(c). The initial \( K \) field is generated by using 10 of the 20 original \( K \)-field values.

Pilot point locations and \( K \) values

To evaluate D-optimality based PPM for this larger problem, we compared this with uniform pilot point selection. The number of pilot point locations was fixed at 12 (significantly less than the 20 head measurements). Genetic algorithm options are the same as those used for Model problem I (Table 1), with the exception of increased mutation size (0.5) and population size (1000) in the pilot point location search procedure. The pilot point selection from DBM and uniform cases are shown in Figure 13. For this larger case, DBM provides a well-distributed set of pilot point locations as in Model problem I. Computation time for this larger problem is approximately 10 h (approximately 200 000 groundwater flow simulations).

From Step 2 for searching hydraulic conductivity values at the identified pilot points, we finally generate \( K \) distributions for DBM and Uniform pilot points. Figure 14 shows the results of DBM and Uniform pilot points. At a cursory glance it may appear that the final \( K \) distributions from uniformly selected pilot point locations are closer to the true \( K \) field than the DBM \( K \) field. However, upon closer examination, DBM has a smaller error (\( K_{\text{error}}: 6.6338 \)) than Uniform (\( K_{\text{error}}: 8.0516 \)). Also, hydraulic head contours in Figure 15 support that DBM results are closer to reality as the \( K \) distribution from Uniform pilot points gives a much higher hydraulic head drawdown around the extraction well location.

<table>
<thead>
<tr>
<th>No.</th>
<th>X</th>
<th>Y</th>
<th>Hydraulic conductivity (K)</th>
<th>Head error</th>
<th>Average error of actual K</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>4</td>
<td>49.245</td>
<td>0.112</td>
<td>1.881</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>2</td>
<td>61.857</td>
<td>0.112</td>
<td>1.896</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>84.744</td>
<td>0.104</td>
<td>2.002</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>4</td>
<td>71.586</td>
<td>0.092</td>
<td>2.010</td>
</tr>
</tbody>
</table>
CONCLUSION

A D-optimality based pilot point method is developed in this study for hydraulic conductivity estimation from existing hydraulic conductivity and head measurements. The D-optimality based criterion satisfies the requirements of maximizing sensitivity with minimum redundancy among pilot points. Since maximally sensitive points require minimal perturbation to match head measurements, the covariance structure of the original measured hydraulic conductivities is well preserved and the impact of head measurement errors is reduced. Also, since the points selected using this criterion are minimally correlated, redundancy among pilot points are reduced, thus leading to a more unique estimation. Two problems with different size domains and flow conditions are examined. Case I involves a smaller problem (10 × 7 grid) with specified head conditions and Case II involves
a larger problem (50 × 25 grid) with specified head and specified flow conditions. For Case I, four scenarios with different initial $K$-field distributions and different flow conditions are examined. In all four scenarios, D-optimality based PPM provides better results than a random set of pilot points and different flow conditions have minimal influences on the results. Furthermore, DBM is shown to perform much better than a previously developed sequential search method. For Case II also, DBM is shown to produce a better $K$ field than a uniform set of pilot points. Domain size does not have significant impact on the capability of DBM.

There are some limitations of this study that could be explored in the future:

- Comparison with another method such as simultaneously searching for pilot point locations and their hydraulic conductivity values to match the head measurements should be investigated.
- The effect of incorporating tracer concentration measurements in the pilot point selection and hydraulic conductivity estimation need to be studied.
- Effectiveness of this method for different test problems, especially those with larger mesh sizes, need to be investigated.

Figure 13 | Pilot point locations for DBM (top) and Uniform (bottom). Axis labels indicate grid indices and each grid block is of size 100 m × 100 m.
Computational improvement using parallel computing and/or surrogate modeling should be considered to reduce the computational burden of the large number of intensive forward model calculations involved in this method.

**ACKNOWLEDGEMENTS**

This work was partially supported by the National Science Foundation (NSF) under grant no. BES-0238623. Any
opinions, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

REFERENCES


Capilla, J. E., Gomez–Hernandez, J., & Sahuquillo, A. 1997 Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data demonstration on synthetic aquifer. J. Hydrol. 203, 175–188.


First received 16 March 2009; accepted in revised form 30 March 2010. Available online 28 October 2010