Quantitative methods to direct exploration based on hydrogeologic information

Andrew J. Graettinger, Jejung Lee, Howard W. Reeves and Deepu Dethan

ABSTRACT

Quantitatively Directed Exploration (QDE) approaches based on information such as model sensitivity, input data covariance and model output covariance are presented. Seven approaches for directing exploration are developed, applied, and evaluated on a synthetic hydrogeologic site. The QDE approaches evaluate input information uncertainty, subsurface model sensitivity and, most importantly, output covariance to identify the next location to sample. Spatial input parameter values and covariances are calculated with the multivariate conditional probability calculation from a limited number of samples. A variogram structure is used during data extrapolation to describe the spatial continuity, or correlation, of subsurface information. Model sensitivity can be determined by perturbing input data and evaluating output response or, as in this work, sensitivities can be programmed directly into an analysis model. Output covariance is calculated by the First-Order Second Moment (FOSM) method, which combines the covariance of input information with model sensitivity. A groundwater flow example, modeled in MODFLOW-2000, is chosen to demonstrate the seven QDE approaches. MODFLOW-2000 is used to obtain the piezometric head and the model sensitivity simultaneously. The seven QDE approaches are evaluated based on the accuracy of the modeled piezometric head after information from a QDE sample is added. For the synthetic site used in this study, the QDE approach that identifies the location of hydraulic conductivity that contributes the most to the overall piezometric head variance proved to be the best method to quantitatively direct exploration.

Key words | directed exploration, FOSM, MODFLOW, Taylor series, uncertainty analysis

Andrew J. Graettinger (corresponding author)
Department of Civil and Environmental Engineering,
University of Alabama,
Tuscaloosa, AL 35487,
USA
E-mail: *andrewg@coe.eng.ua.edu*

Jejung Lee
Department of Geosciences,
University of Missouri – Kansas City,
Kansas City MO 64110,
USA
E-mail: *leej@umkc.edu*

Howard W. Reeves
US Geological Survey,
USGS Michigan Water Science Center,
6520 Mercantile Way, Suite 5, Lansing,
MI 48911,
USA
E-mail: *hwreeves@usgs.gov*

Deepu Dethan
Environmental Resources Management,
1110 Montlimar Drive, Suite 150, Mobile,
AL 36609,
USA
E-mail: *Deepu.Dethan@erm.com*

INTRODUCTION

A method to quantitatively direct exploration rather than subjectively judge site characterization has been desired for decades (Casagrande 1965; Dowding 1978; National Research Council 1995; Shackelford *et al.* 1996). Because of the limited number of monitoring wells and soil borings employed during subsurface characterization, inherent uncertainty in site information exists. This uncertainty in turn produces uncertainty in the design of geotechnical or geoenvironmental systems. To reduce uncertainty in the design process, site exploration techniques may be formulated to reduce the uncertainty in site characterization that most efficiently improves the design. Input information uncertainty, model sensitivity and modeled output uncertainty can be employed to guide site exploration. Quantitatively Direct Exploration (QDE), proposed by Graettinger & Dowding (1999), is an efficient method for selecting sampling locations based solely on output uncertainty. This original QDE approach directed exploration to the location of greatest variance in calculated model output, where model results are most uncertain. Six additional methods for quantitatively directing exploration are presented and evaluated based on the improvement in the modeled results for a synthetic case study. The synthetic case study models piezometric head calculated from

doi: 10.2166/hydro.2006.006

uncertain hydraulic conductivity. MODFLOW-2000 is employed as the hydrogeologic computational model.

In terms of directing exploration, the QDE approaches employ three main components of modeling – input data uncertainty, model sensitivity and output data uncertainty. Intuitively, by sampling and improving model input information, the model output will improve and be less uncertain. Directing exploration based solely on improving uncertain input information has the tendency to direct exploration away from past sampling, which typically is near the area of interest, and directs exploration toward areas that have not been sampled, which typically are at the edges of a site.

Spatial uncertainty of hydrogeologic parameters, resulting from the scarcity of sampling, has been quantified based on the concept of geostatistics (Matheron 1963; Davis 1986; Isaaks & Srivastava 1989). Using the spatial continuity in geologic structures, extrapolation methods such as ordinary kriging and cokriging have been developed. These methods provide extrapolated values and an estimate of error in that value throughout space. To improve upon the estimate of error, and to provide a covariance between input parameters, a multivariate conditional probability approach to extrapolate input data (hydraulic conductivity) is employed (Gelman *et al.* 1995).

Input parameters and uncertainty in input parameters can be modeled as a single "lumped" value or as a set of discrete, spatially varying values. Lumped parameters employ one value to describe a site or portion of a site, such as a single value of hydraulic conductivity for a geologic layer. To capture the spatial variability of a lumped parameter, a single value of uncertainty, typically the standard deviation, is assigned to the lumped parameter value.

In this work, a set of discrete, spatially varying values of hydraulic conductivity and uncertainty in hydraulic conductivity are employed. Therefore, each cell in the model has a unique, but spatially correlated, value of hydraulic conductivity and uncertainty in hydraulic conductivity. By including spatial information in the model through a set of spatially varying input data, sampling locations down to the cell resolution can be identified.

Model sensitivity is the change in model output per change in model input, which for the example presented is the derivative of piezometric head with respect to hydraulic conductivity. By solely employing model sensitivity to direct exploration, input parameters that most influence the model output can be identified. This technique is often used during model calibration. Making sensitive input parameters less uncertain improves model output. Unfortunately, the most sensitive parameter remains the most sensitive even after it has been sampled. Therefore, only employing model sensitivity continually directs exploration back to the same location.

Model sensitivity can be calculated by parameter perturbation or direct derivative coding (DDC) (Graettinger *et al.* 2002). In this work, MODFLOW-2000 is used which has sensitivities directly coded into the model (Hill *et al.* 2000). By using MODFLOW-2000, all sensitivities are calculated with a single model run rather than multiple runs required by perturbation.

Combining input covariance and model sensitivity provides an estimation of output covariance. A first-order Taylor series expansion, commonly referred to as First-Order Second Moment (FOSM), is employed to perform this calculation. Although higher-order and higher-moment calculations can be made, to simplify the discussion herein, only the FOSM calculation is presented. Directing exploration based on covariance in modeled output equally weights model sensitivity (greatest near a pumping well) and input data uncertainty (greatest at distances from sample locations). This original QDE method (Graettinger & Dowding 1999) identifies the location where output variance (variance in modeled piezometric head) is greatest and samples the input parameter (hydraulic conductivity) at that location.

In this paper, seven QDE approaches are presented to direct subsurface exploration for the calculation of piezometric head produced by MODFLOW-2000. QDE Approach 1, "Largest Reduction in Input Uncertainty", identifies the location that most reduces overall input uncertainty. Approach 1 only evaluates input uncertainty information and does not consider model sensitivity or output uncertainty. QDE Approach 2, "Most Sensitive Output", identifies the location where the model output is most sensitive to changes in model input, while QDE Approach 3, "Most Important Input", identifies the location where a change in model input has the largest affect on model output. Approaches 2 and 3 only analyze the spatial sensitivity matrix to direct exploration and do not consider

input or output uncertainty. QDE Approach 4, "Largest Correlated Variance", identifies the location where output uncertainty from correlated input data is largest. QDE Approach 5, "Largest Uncorrelated Variance", identifies the location where output uncertainty from uncorrelated input data is largest. Approaches 4 and 5 are the published QDE methods (Graettinger & Dowding 1999) and are presented herein for comparison of the new QDE approaches. QDE Approach 6, "Most Contributing Correlated Input to Output Variance", identifies the location where correlated input uncertainty contributes the most to output uncertainty. QDE Approach 7, "Most Contributing Uncorrelated Input to Output Variance", identifies the location where uncorrelated input uncertainty contributes the most to output uncertainty. Approaches 6 and 7 both analyze the output variance matrix to find the location that contributes the most to overall variance in piezometric head summed from across a site.

Details of each of the QDE approaches along with a discussion of the FOSM calculation are presented in the methodology section. This is followed by a case study applying and evaluating the seven QDE approaches on a synthetic site. By employing a synthetic site a comparison between the "true" site data (hydraulic conductivity and piezometric head) and the QDE modeled site data can be made for evaluation purposes. Although the QDE approaches are demonstrated on a ground water model, this is not a limitation of these approaches as they can be employed for any analysis where model results are calculated from input data from throughout a site.

METHODOLOGY

Input data extrapolation and covariance

The proposed FOSM framework, which is the basis for the seven QDE approaches, utilizes a multivariate normally distributed conditional probability calculation to extrapolate input information. This calculation generates a regularized grid of input data values and associated spatially distributed uncertainty (covariance) that is assumed to be normally distributed about the mean value of the input data at each grid location. For the example presented in the case study, hydraulic conductivity is the uncertain spatially distributed input data. During extrapolation of hydraulic conductivity, the actual hydraulic conductivities from sampled locations are stored in a vector, $\boldsymbol{V}$. Next, a vector, $E(\boldsymbol{U})$, is produced that contains a prior estimate of mean hydraulic conductivity at each grid point. Finally an informed prior covariance matrix, $\mathrm{Cov}(\boldsymbol{U})$, is calculated by employing a variogram function that represents the covariance relationships between the known hydraulic conductivities at sampled locations. From these matrices, the prior estimates, $E(\boldsymbol{U})$ and $\mathrm{Cov}(\boldsymbol{U})$, are updated to posteriors, $E(\boldsymbol{U}|\boldsymbol{V})$ and $\mathrm{Cov}(\boldsymbol{U}|\boldsymbol{V})$, using data from the sampled points, through the following equations (Gelman *et al.* 1995):

$$E(U|V) = E(U) + \mathrm{Cov}(U,V)\,\mathrm{Cov}(V)^{-1}(V - E(V)) \quad (1)$$

$$\mathrm{Cov}(U|V) = \mathrm{Cov}(U) - \mathrm{Cov}(U,V)\,\mathrm{Cov}(V)^{-1}\mathrm{Cov}(V,U). \quad (2)$$

$E(\boldsymbol{U}|\boldsymbol{V})$ is the updated vector of hydraulic conductivities given the known hydraulic conductivities at the sampled locations. $\mathrm{Cov}(\boldsymbol{U},\boldsymbol{V})$ is a subset of the full covariance matrix, that stores the covariance between the hydraulic conductivities being estimated and the known hydraulic conductivities at sampled locations. $\mathrm{Cov}(\boldsymbol{V})^{-1}$ is again a subset of the full covariance matrix, and is the inverse of the covariance between the known hydraulic conductivities. $E(\boldsymbol{V})$ is a subset of $E(\boldsymbol{U})$ and holds the prior hydraulic conductivities estimates at the sampled locations. $\mathrm{Cov}(\boldsymbol{U}|\boldsymbol{V})$ is the updated covariance matrix and $\mathrm{Cov}(\boldsymbol{U})$ is the prior covariance matrix, calculated by the covariance function. A FORTRAN program was written to extrapolate the hydraulic conductivity data and its associated uncertainty using conditional probability calculations shown in Equations (1) and (2). Generated output files were then used as input for MODFLOW-2000.

Performance model and sensitivity

As stated previously, the QDE approaches can be used with any model that employs spatially distributed information to calculate a model result. For this discussion, MODFLOW-2000 is used to calculate site performance, which is the piezometric head across a site. The extrapolated hydrogeologic information from Equation (1) is used as input to

MODFLOW-2000, which for the case study presented is run under steady state conditions.

An advantage of using MODFLOW-2000 is that sensitivity routines are included in the code. Sensitivities are calculated for each specified input parameter for the entire model grid (Hill *et al.* 2000). Sensitivity represents the change in model output per change in an input parameter. Specifically for this work, sensitivity is the change in piezometric head per change in hydraulic conductivity. Because the head at any node is affected by a change in hydraulic conductivity at any node, a grid of sensitivities for the site is produced for each hydraulic conductivity input parameter.

MODFLOW-2000 outputs sensitivities as 1% scaled sensitivities: each sensitivity is multiplied by the corresponding input parameter and divided by 100. These normalized sensitivities are "un-normalized" for use in the FOSM calculation. This is done by multiplying the sensitivities by 100 and dividing by the input parameter value.

FOSM calculation

In the QDE framework, data uncertainty and model sensitivity are combined through a first-order Taylor series expansion to produce the variance in computed head. Calculation of piezometric head covariances from correlated input data is shown by Equation (3) (Harr 1996):

$$\mathrm{Cov}[Y_l, Y_k] \approx \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial f_l}{\partial x_i}\right)\left(\frac{\partial f_k}{\partial x_j}\right)\mathrm{Cov}[x_i, x_j]. \quad (3)$$

Here $\mathrm{Cov}[Y_l,Y_k]$ is the covariance of computed head between node l and node k. $\left(\frac{\partial f_k}{\partial x_i}\right)$ is the sensitivity of the head at node k to a change in the input at node i. $\mathrm{Cov}[x_i,x_j]$ is the covariance between input at nodes i and j.

Using the extrapolated data and covariance files produced by Equations (1) and (2), and the sensitivity file from MODFLOW-2000, variance in the head is calculated for the entire model using Equation (3).

Seven QDE approaches

All QDE approaches are based on three matrices: 1) uncertainty in input information, 2) sensitivity of model results and 3) variance in calculated results. Each of the data elements in these matrices are related to specific locations throughout a site. By analyzing the data in these three matrices, seven approaches to Quantitatively Directed Exploration are developed and evaluated.

QDE Approach 1. Largest Reduction in Input Uncertainty: the location that produces the largest reduction in overall input uncertainty

The spatial continuity of input information also represents the spatially continuity of uncertainty in input data. Sampling reduces the uncertainty at a sampled location and also reduces the uncertainty to some degree around that location. As the distance from a sample increases, so does the uncertainty. The goal of Approach 1 is to find the location that, if sampled, produces the largest reduction in the summation of input uncertainty from across a site. This reduction is determined by calculating the volume under the input uncertainty surface as shown in Equation (4):

$$V = \sum_{i=1}^{n}\frac{\sum_{j=1}^{4}\sigma_{ij}}{4}x^2. \quad (4)$$

Here, V is the volume under the uncertainty surface, σ_{ij} is the standard deviation of input information at location j in cell i, where j is between 1 and 4 and represents the four corners of a cell, and x^2 is the plan area of a cell. Figure 1(a) shows an example of input uncertainty for a 10-cell by 10-cell model that is sampled at five locations indicated by the arrows. At the sampled locations, the surface drops to zero and increases as the distance from a sample point increases. This increase with distance matches the spatial continuity of the data described by the variogram, seen in Figure 1(b). By calculating the estimated reduction in the volume under the uncertainty surface caused by sampling each location, one location at a time, the location that caused the greatest reduction in input uncertainty is determined.

QDE Approach 2. Most Sensitive Output: the location of the most sensitive output due to a change in input information at all locations

The sensitivity matrix is information about how a model result will change due to an increase in an input parameter. Because piezometric head (model output) and hydraulic

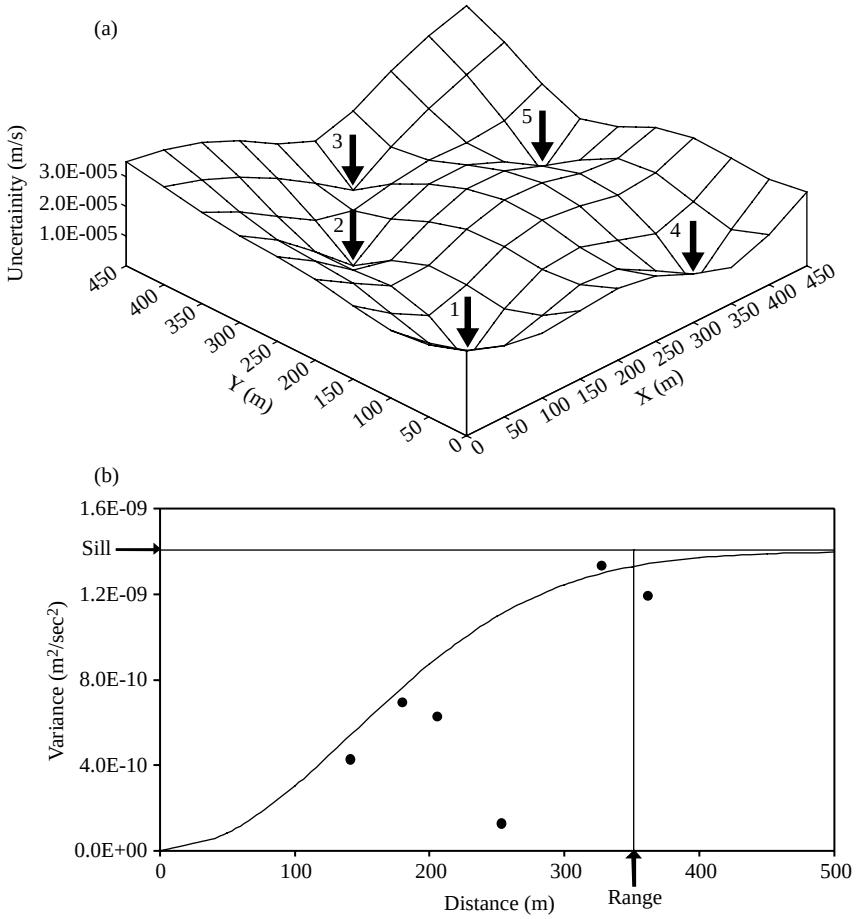

Figure 1 | (a) Input data uncertainty surface for a 10 × 10 cell model, indicating location of sample points with arrows. (b) Variogram model employed to generate the input covariance matrix employed in data extrapolation.

conductivity (model input) are both spatially distributed across the modeling domain, the dimensions of the sensitivity matrix are number of model outputs by number of model inputs, which is $(n \times n)$, where n is the number of cells in a model. The sensitivity matrix can be summed with respect to output (piezometric head) as in QDE Approach 2, or summed with respect to input (hydraulic conductivity) as in QDE Approach 3.

For Approach 2, the sensitivity of output due to a change in input everywhere in the domain is obtained by summing each row of the sensitivity matrix as shown in Equation (5):

$$\mathrm{Sen}(h_j) = \sum_{i=1}^{n} \frac{\partial h_j}{\partial HK_i}. \tag{5}$$

Here, $\mathrm{Sen}(h_j)$ is the sensitivity of the piezometric head at cell j due to a change in hydraulic conductivity in all cells. This is shown graphically by summing a row in Figure 2. If the purpose of exploration is to find the most sensitive output location, then the maximum $\mathrm{Sen}(h_j)$ is the next sampling location.

QDE Approach 3. Most Important Input: the location of the input information that produces the largest change in output

If the elements of the sensitivity matrix are summed based on input information, then the summed value represents how much the entire output domain will change (piezometric head across the site) for a change in one specific input value (hydraulic conductivity in one cell). This summation is shown in Equation (6):

$$\mathrm{Sen}(HK_i) = \sum_{j=1}^{n} \frac{\partial h_j}{\partial HK_i}. \tag{6}$$

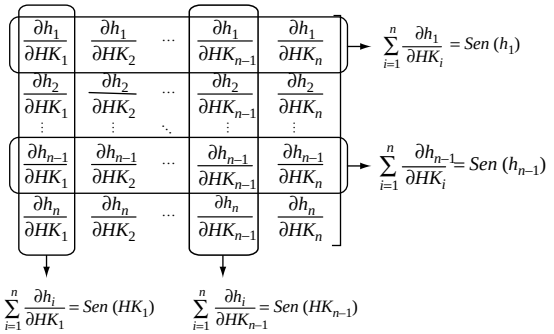

Figure 2 | Sensitivity matrix showing that the summation of each row is the sensitivity of h at location i due to a change in HK at every location and summation of each column is the sensitivity of h everywhere due to a change in HK at location i.

Here, Sen(HK_i) represents how much a change in HK_i contributes to a change of h everywhere. This is shown graphically by summing a column in Figure 2. The maximum Sen(HK_i) identifies the input information that has the greatest effect on model output for the entire domain.

QDE Approach 4. Largest Correlated Variance: the location of the largest output uncertainty produced from correlated input information

This original QDE approach proposed by Graettinger & Dowding (1999) directs exploration to the location where the output variance is the largest. Figure 3 shows the structure of the output covariance matrix calculated by Equation (3) (left) and a plan view of a 10 × 10 grid domain (right). The diagonal elements in Figure 3 are the output variances that are analyzed to identify the most uncertain location which, for QDE Approach 4, is the next sampling location. The variances in Figure 3 are related to a single cell at a site, while the covariances, off-diagonal terms, are related to two cells at a site. Each variance or covariance in Figure 3 is composed of 100 × 100 piecewise elements, as indicated by the smaller squares seen in the figure. A piecewise element (small square) is one term of the first-order Taylor series multiplication, and all piecewise elements related to a specific location are summed together to produce the variances and the covariances. It can be seen in the inset of Figure 3 that columns are

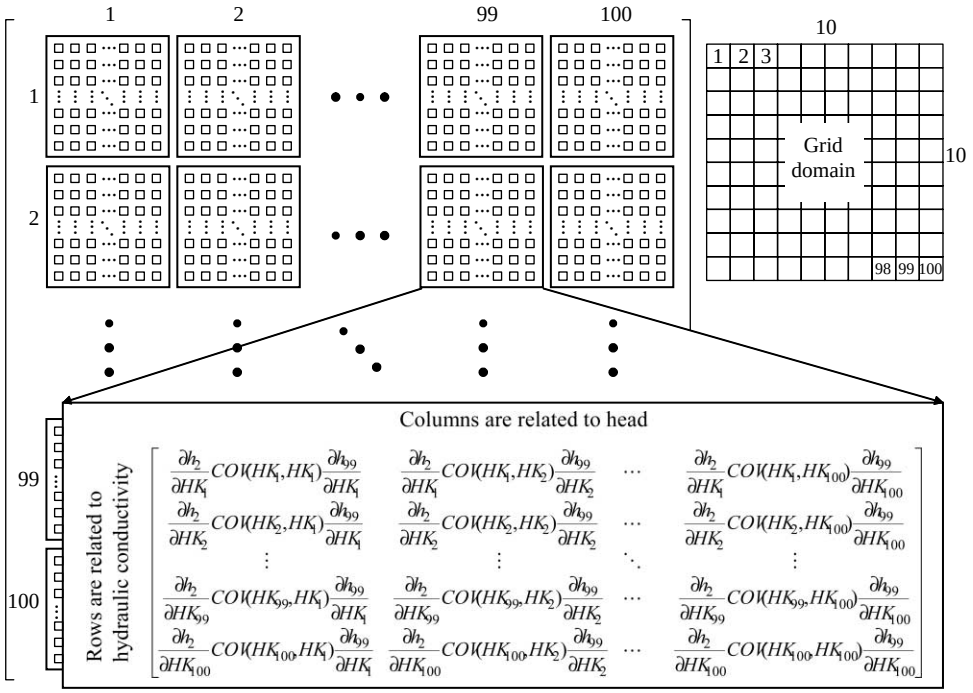

Figure 3 | Schematic representation of the First-Order Second-Moment (FOSM) calculation of variance–covariance matrix of the output variable (head). Large blocks along the diagonal of the matrix (1,1; 2,2; 3,3; etc.) are the estimated variance in the head, which is computed by summing small squares within each block. Each small square is one term of the large block, as shown in the inset.

related to head at specific locations while rows are related to hydraulic conductivity at specific locations.

Although all large squares (covariance and variances) can be calculated by the FOSM approach, only the variance terms on the main diagonal are employed herein. Figure 4 shows a graphical representation of the output covariance matrix (left) along with a plan view of a 10 × 10 site (right). It can be seen that the variance in the upper-left corner of the site comes from summing the first group of terms in the output covariance matrix. Each group of terms along the main diagonal of the covariance matrix are mapped to a specific location at a site. In addition, each group of terms along the main diagonal are related to a single piezometric head at a specific location.

Output variance at a location is the sum of input information uncertainty and model sensitivity from across a site. QDE Approach 4 directs exploration to the location of greatest output variance. A situation can arise where the location of greatest output variance has already been sampled. In this case, the input uncertainty at that location is zero, but because the output uncertainty is a sum of information from across a site, that location still has the maximum uncertainty. Resampling at that location adds no new information to the analysis; therefore, Approach 4 must identify the location of maximum output variance that has not been previously sampled.

QDE Approach 5. Largest Uncorrelated Variance: the location of the largest output uncertainty produced from uncorrelated input information

The FOSM calculation can be in an uncorrelated form for cases where the input parameters are unrelated. Although this is not the case for hydraulic conductivity or most information at a site, QDE Approach 5, which uses the uncorrelated FOSM calculation, has advantages. The uncorrelated form of the FOSM calculation is shown in Equation (7):

$$\mathrm{Var}[h_k] \approx \sum_{i=1}^{n} \left(\frac{\partial h_k}{\partial HK_i}\right)^2 \mathrm{Var}[HK_i]. \tag{7}$$

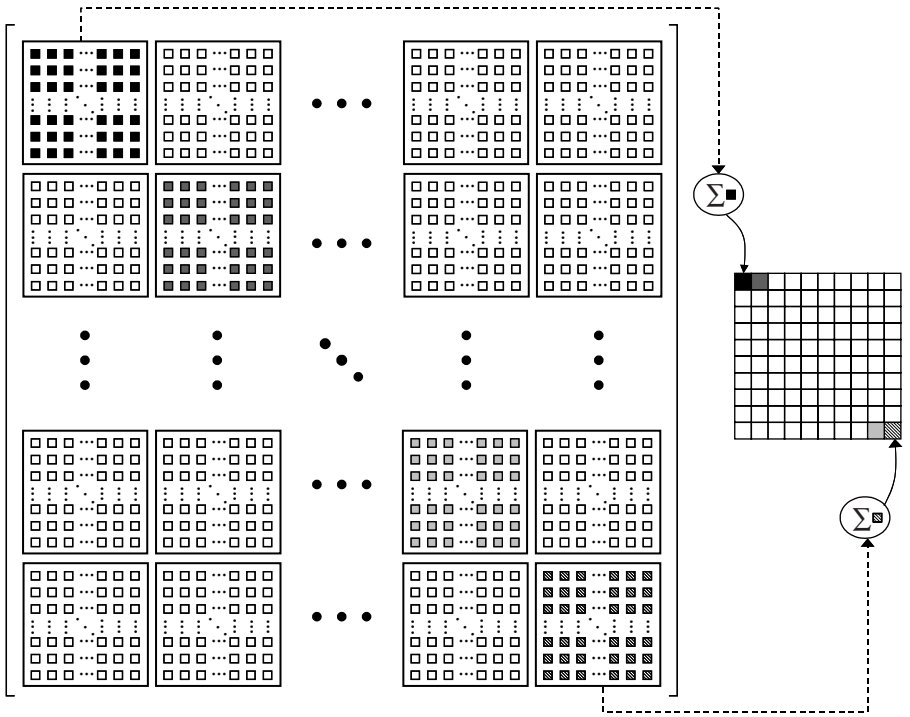

Figure 4 | Schematic of output variance calculation. Linkage of the operation performed on the large diagonal blocks from FOSM variance–covariance matrix, which is defined in Figure 3, to the corresponding 100 spatial locations (cells) shown in the map on the right. The operation required for correlated input is the summation of the entire 100 × 100 (10 000) terms in the large block. Thus, the shaded cells on the map contain the variation in head at locations 1, 2, 99 and 100 contributed by uncertain input variable at all locations. Exploration involves sampling at the location of maximum estimated variance in the head.

Here, Var[h_k] is the uncorrelated output variance at location k and Var[HK_i] is the input variance at location i. As with Approach 4, the maximum location of Var[h_k] is the next sampling point. The advantage of QDE Approach 5 for directing exploration is that without input covariance terms the input variance becomes more important. At a sampled location, the input variance is assumed zero; therefore, there is a smaller chance of directing exploration to a point that has already been sampled. If the assumption can be made that the input parameters are not correlated, this approach is simpler than Approach 4 and tends to avoid the resampling problem. Figure 5 shows the piecewise elements that are summed to calculate the uncorrelated output variance.

QDE Approach 6. Most Contributing Correlated Input to Output Variance: the location where correlated input information uncertainty contributes the most to output uncertainty

Output variance is the multiplication of input covariance and model sensitivity from across a site, as shown in Equation (3). As discussed in Approach 4, it is possible to have the maximum output variance be at a location that has been previously sampled. Sampling at the largest output uncertainty (QDE Approach 4), reduces the input information uncertainty at that location and typically directs exploration to a new unsampled location. In some cases, the location of largest output uncertainty is at a previously sampled location. Resampling at that location will not improve the model or reduce uncertainty. Therefore, QDE Approach 6 is developed to find the location of the input information that most contributes to output variance across a site.

The FOSM equation, Equation (3), is rearranged to sum output variance terms related to individual input parameters rather than output. Equation (8) represents the contribution from each input parameter to the output uncertainty across a site:

$$\mathrm{cont}(i) = \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n} \frac{\partial h_j}{\partial HK_i} \mathrm{Cov}(HK_i, HK_k) \frac{\partial h_j}{\partial HK_k} \right\}. \tag{8}$$

Here, cont(i) is the contribution to output variance from each input parameter. Figure 6 shows a graphical representation of the sum of the terms from the output covariance matrix.

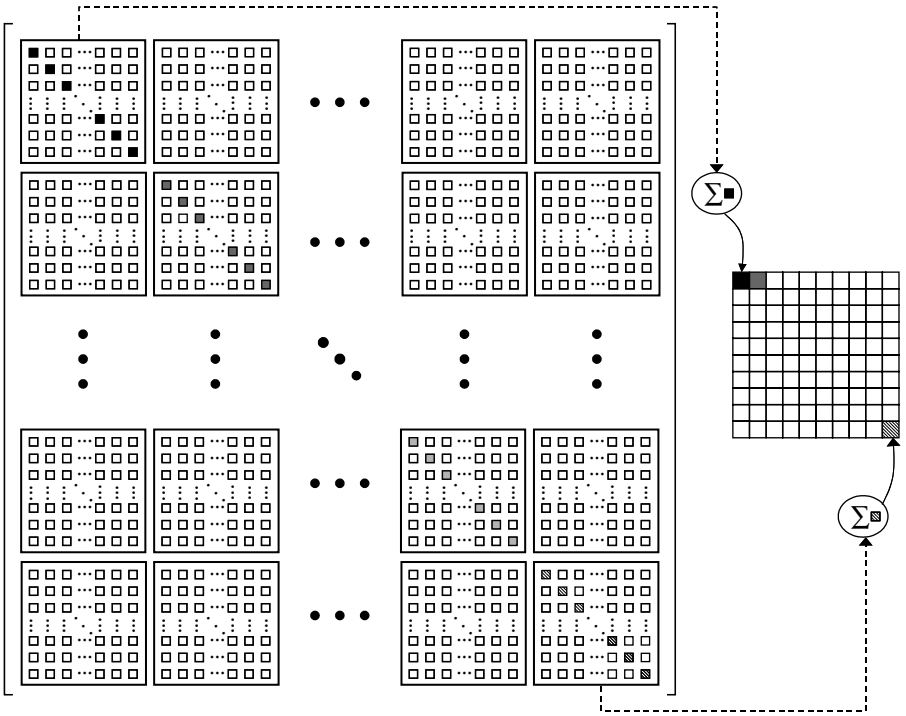

Figure 5 | In contrast to the correlated case described in Figure 4, the operation required for uncorrelated input is the summation of only the 100 diagonal terms of the matrix in the large block. The shaded cells and method of exploration are the same as given in Figure 4.

In Figure 6, the small black squares represent the terms of the FOSM related to input parameter HK_1. The sum of the small black squares is the portion of the total output uncertainty that is related to input parameter HK_1. For a 10×10 cell site, Equation (8) can be written as Equation (9) for HK_1:

$$\text{cont}(1) = \sum_{j=1}^{100} \left\{ \sum_{k=1}^{100} \frac{\partial h_j}{\partial HK_1} \text{Cov}(HK_1, HK_k) \frac{\partial h_j}{\partial HK_k} \right\}. \quad (9)$$

By comparing cont(i) from $i = 1$ to 100 for a given 10×10 model, the maximum cont(i) is the location of the input parameter information that contributes the most to total output variance.

QDE Approach 7. Most Contributing Uncorrelated Input to Output Variance: the location where uncorrelated input information uncertainty contributes the most to output uncertainty

QDE Approach 7 is the same as QDE Approach 6 except it identifies the most contributing uncorrelated input parameter. By removing the covariance terms from Equation (8), QDE Approach 7 considers only input information variance as shown in Equation (10):

$$\text{ucont}(i) = \sum_{j=1}^{n} \left\{ \sum_{k=1}^{n} \frac{\partial h_j}{\partial HK_i} \text{Var}(HK_i) \frac{\partial h_j}{\partial HK_i} \right\}. \quad (10)$$

This approach is used to find the most contributing input parameter when input parameters are not correlated or can be assumed not correlated. The graphical representation of Approach 7 is shown in Figure 7.

CASE STUDY

Synthetic site description

The seven QDE approaches are applied and evaluated on a hydrogeologic synthetic model. The synthetic model is a steady-state ground water model in a confined aquifer where hydraulic conductivity is the uncertain input information and piezometric head is the calculated model result.

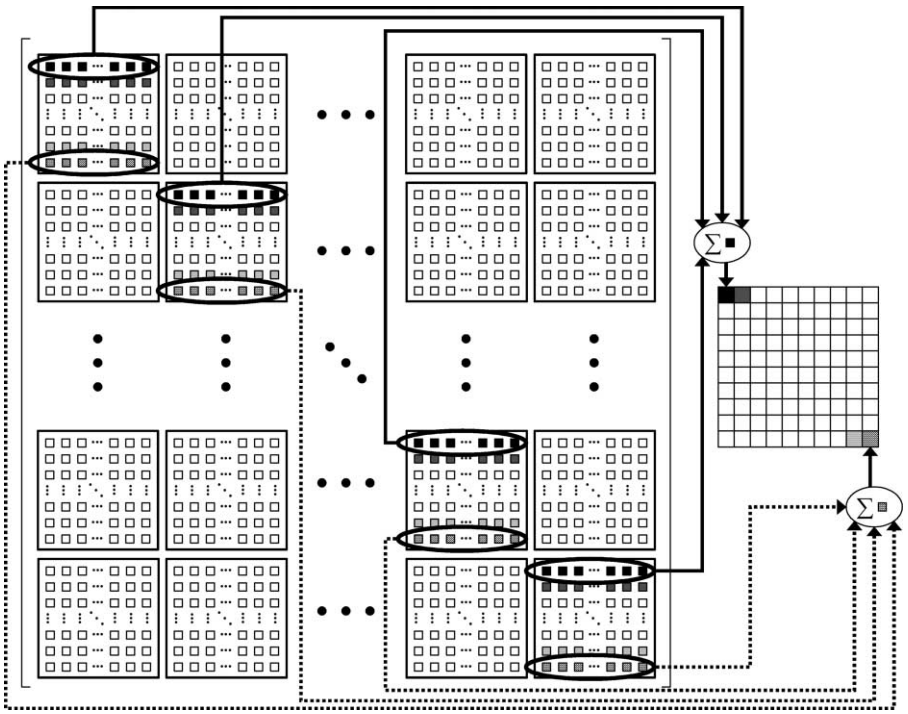

Figure 6 | Schematic of the contributing variance calculation. In this correlated case, operation on the FOSM matrix is the summation of rows of terms from individual large blocks corresponding to the 100 locations. Thus, the shaded cells on the map contain the variation in head at all locations contributed by the terms at locations 1, 2, 99 and 100. Exploration involves sampling at the location that gives the maximum contribution to the estimated variance in all heads.

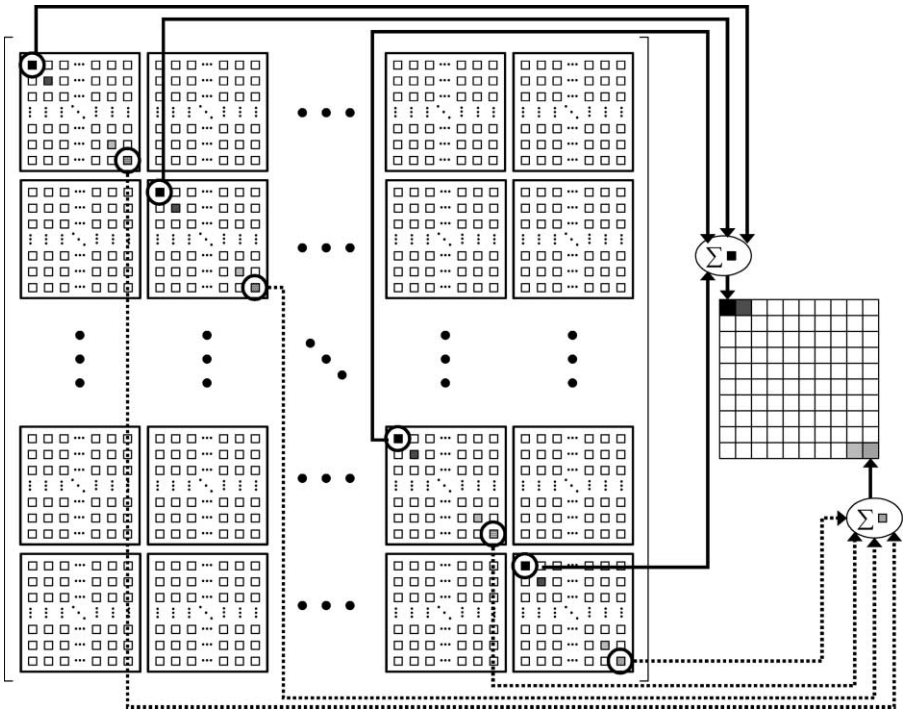

Figure 7 | In contrast to the correlated case described in Figure 6, the operation required for uncorrelated input is the summation of only the 100 corresponding diagonal terms from each large block in the matrix. The shaded cells and method of exploration are the same as given in Figure 6.

Figure 8 is a plan view of the site showing contour lines representing the "true" hydraulic conductivity in the field, which is highly spatially correlated. The hydraulic conductivity (*HK*) is measured at five existing sample locations (numbered black dots). It is assumed that hydraulic conductivity is measured at only these locations and no other information is used for the seven QDE approaches. Table 1 shows the measured hydraulic conductivity at the original five sampe locations. The site is divided into 100 50 m × 50 m cells, making a 10 × 10 grid. Piezometric heads at the site are fixed as: 5 m on the left boundary and 0 m on the right boundary, and there is a no-flow condition at $Y = 0$ m and $Y = 450$ m. A pumping well (triangle) at the center of the domain is pumping at a rate of 0.003 m^3/s. Because this is a synthetic site, the "true" head in the field is known and is shown in Figure 9. This "true" piezometric surface was calculated by MODFLOW-2000 given every "true" hydraulic conductivity shown in Figure 10.

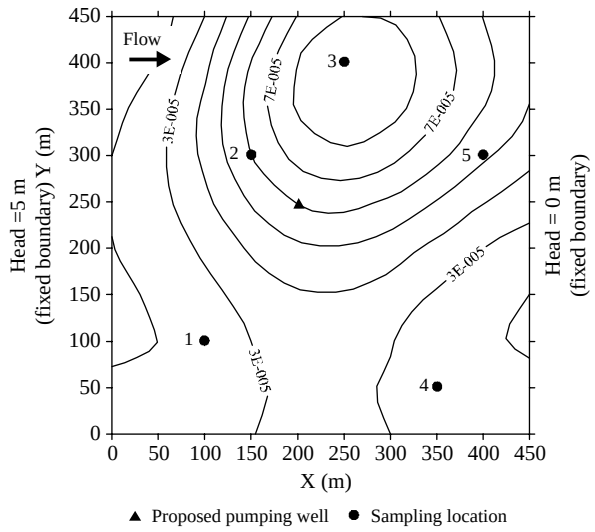

Figure 8 | Plan view of a synthetic site with contour lines of the "true" hydraulic conductivity.

Measured hydraulic conductivities at the five original locations are analyzed to estimate their spatial continuity using a variogram function. Since the multivariate conditional calculation employs a variogram, estimating variogram parameters such as sill and range is the first step to extrapolate spatial input parameters. These estimates are, however, performed in a subjective manner. The subjective estimates affect extrapolated input information and associated uncertainty and therefore may cause errors in

Table 1 | Hydraulic conductivity (*HK*) at five original sample locations

Number of wells	*X*(m)	*Y*(m)	*HK*(m/s)
1	100	100	2.45E-05
2	150	300	5.98E-05
3	250	400	8.90E-05
4	350	50	2.56E-05
5	400	300	5.18E-05

estimating the true output and output uncertainty (Dethan 2003). The effect of the variogram parameters has been examined by others (Lee 2001; Supriyasilp *et al.* 2003). It was shown that the range, a measure of spatial continuity, influences the smoothness of the extrapolated input information and uncertainty. The sill, a measure of the variance in the data set, affects the magnitude of the extrapolated uncertainty. For this example a high-sill ($1.4 \times 10^{-9}\,m^2/s^2$) and high-range (350 m) condition is employed, as shown in Figure 1(b). This produces a larger spatial continuity of hydraulic conductivity and larger input and output variance.

Application and result of the seven QDE approaches

QDE Approaches 1–7 are applied to a hydrogeologic synthetic site given only information from the five original

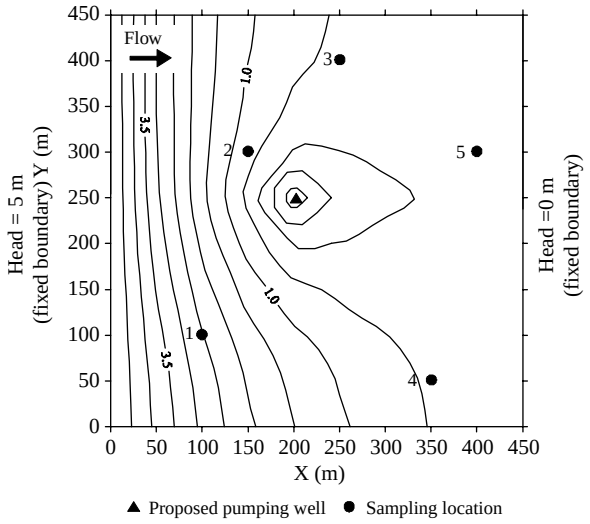

Figure 9 | Plan view of a synthetic site with contour lines represent the "true" piezometric head.

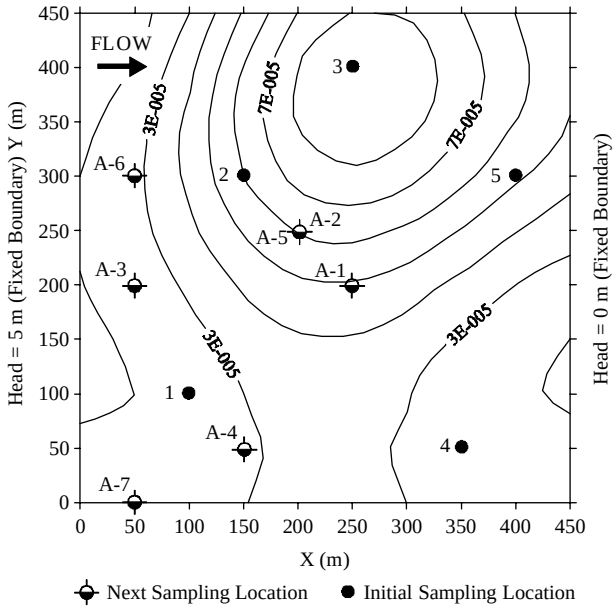

Figure 10 | Plan view of a synthetic site showing hydraulic conductivity contour lines, location of the original five samples (triangles) and QDE location for the next sampling point from each approach (labeled stars).

sample locations. Each QDE approach identified a new location to sample at the site, as shown in Figure 10. Table 2 lists the coordinates of the new sample points identified by each approach. With each additional sample, the new hydraulic conductivity information from the sixth sample is added to the model by rerunning the extrapolation calculation. This new extrapolated hydraulic conductivity information is used in the next model run to calculate the piezometric head and estimate the head variance through the FOSM calculation.

Each approach is evaluated based on the accuracy of the modeled head from the six borings compared to the "true" head. For every QDE approach, the additional sample improved the modeled head, moving it closer to the "true" head. In addition, each QDE approach reduced the overall standard deviation in the modeled head. Table 2 presents the mean absolute error for the five original sampling points (top row) and each of the seven QDE approaches with the information from the sixth location added to the analysis. In addition, Table 2 also shows the percent change in standard deviation of the piezometric head from across the site. Equation (11) calculates the percent difference between the volume under the standard

Table 2 | Head improvement and reduction in site uncertainty for the seven QDE approaches

QDE approach	Added point	X(m)	Y(m)	Mean absolute residual (m)	Piezometric head improvement	Δ in output standard deviation
Original 5 points	–			0.49	–	–
1	A1	250	200	0.43	12%	9%
2	A2	200	250	0.41	14%	6%
3	A3	50	200	0.39	20%	9%
4	A4	150	50	0.46	6%	7%
5	A5	200	250	0.41	14%	6%
6	A6	50	300	0.23	53%	8%
7	A7	50	0	0.45	8%	10%

deviation surface before and after the addition of data from the sixth sample location:

$$\Delta V(\%) = \frac{100}{n}\sum_{i=1}^{n}\frac{(\sigma_{b_i} - \sigma_{a_i})}{\sigma_{b_i}}. \quad (11)$$

Here, $\Delta V(\%)$ is the percentage volume change under the output uncertainty surface after a new sample is included in the model, σ_{b_i} is the standard deviation of output data at in cell *i before* new data is added and σ_{a_i} is the standard deviation of output data in cell *i after* new data is added.

Each of the seven QDE approaches developed and presented herein has advantages and disadvantages. QDE Approach 1 focuses only on the uncertainty in input information which, through the FOSM calculation, affects output uncertainty. Approach 1 is the best approach in terms of reducing input uncertainty across a site, but it should be noted that this approach may direct exploration away from areas of interest and to areas of largest input uncertainty. As shown in Figure 10, location A-1 is at the center of the site and at the largest distance from the five previously sampled locations. Therefore, sampling at A-1 will most reduce the volume under the input information uncertainty surface, which is shown in Figure 1(a). Approaches 2 and 3, exploring based solely on the sensitivity matrix, decrease the overall head uncertainty and move the modeled head closer to the "true" head. These two approaches are beneficial when the goal of exploration is to improve model calibration. QDE Approach 2 identifies which head location is most sensitive to a change in hydraulic conductivity, while QDE Approach 3 identifies the location of hydraulic conductivity that most affects the head across the site. As seen in Figure 10, A-2 is located at the pumping well where the head is very sensitive to changes in hydraulic conductivity. A-3 is up-gradient of the pumping well; therefore, changing hydraulic conductivity information at this location affects the head the most across the entire site.

QDE Approach 4 samples at the location of largest correlated output variance. As shown in Table 2, the estimated head is relatively unchanged, only a 6% improvement, and the overall uncertainty improvement is 7%. Approach 4 is the most justifiable approach based on mathematics. This approach uses the first-order Taylor series to estimate the variance in modeled output. From a physical standpoint, A-4 in Figure 10 is at the location of largest output uncertainty. Sampling at this point will reduce input uncertainty, which in turn will reduce output uncertainty. In some cases, the location of largest output uncertainty may already have been sampled. This is the main disadvantage with Approach 4.

Approach 5 is identical to Approach 4 except the covariance terms are dropped from the calculation of output variance because the input data are assumed to be uncorrelated. This assumption is not true for most subsurface data, but by making this assumption, the likelihood of directing

exploration back to an already sampled point is reduced. Location A-5 shown in Figure 10 is at the pumping well which is the location of largest sensitivity. By removing the covariance terms in Approach 5, sensitivity becomes more important and exploration is directed to the pumping well.

QDE Approach 6 identifies the input information that contributes the most to overall output variance. This approach produced the best results among all seven QDE approaches. Approach 6 most reduced the mean absolute residual between the modeled and "true" piezometric head, with an improvement of 53%, and produced a decrease in the overall output standard deviation of 8% as shown by Table 2. The location of A-6 in Figure 10 is up-gradient of the pumping well and the majority of the site. Improving hydraulic conductivity information at this location has the largest effect on model results.

Finally, QDE Approach 7, which is the same as Approach 6 except it is based only on uncorrelated input information, shows only a small improvement in the modeled head. The location of A-7 in Figure 10 is at high head elevation. Improving hydraulic information at this location will affect the modeled head down-gradient.

As can be seen in Figure 10, all QDE identified locations are on the up-gradient half of the site. Locations A-3, A-6 and A-7 are located at $X = 50$ m, which is the first column of cells that does not have a fixed head elevation. Locations A-2 and A-5 are at the pumping well which is the most sensitive location at this site, while location A-1 is at the most uncertain point, which is at the greatest distance from existing samples.

CONCLUSION

The seven QDE approaches employ a FOSM calculation to combine the uncertainty in extrapolated geologic data and the sensitivity of a hydrogeologic model to calculate the variance in modeled output. While the performance model, which was MODFLOW-2000 for this example, computed the peizometric head, the generality of the QDE computation makes the seven QDE approaches applicable for directing site exploration for a variety of site analyses.

Based on this work and the synthetic case study presented, the following conclusions are advanced:

(1) Seven Quantitatively Direct Exploration (QDE) approaches for evaluating the input uncertainty matrix, model sensitivity matrix and output variance matrix are developed and applied to a synthetic model. For the given synthetic model, the QDE approach that identified the location of the input information that most contributes to the overall output variance performed the best in terms of reducing the mean absolute residual between the modeled and "true" piezometric head.

(2) Both uncertainty in the geologic input information and sensitivity of the groundwater model must be combined to quantitatively determine the model result variance. The example demonstrated the FOSM method of calculating variance in the piezometric head generated from a MODFLOW-2000 groundwater model.

(3) Only the initial hydrogeologic information and a spatial correlation structure that respects this information is required to begin the QDE process. Extrapolation of hydraulic conductivities from known values at sampled locations through the multivariate conditional probability calculation allows for an estimation of hydraulic conductivity and uncertainty in hydraulic conductivity at unsampled locations. In addition, the conditional probability calculation provides the full covariance matrix that completely describes the probabilistic subsurface, which is necessary for the correlated version of the FOSM calculation.

(4) Because MODFLOW-2000 contains sensitivity calculations that are directly coded into the program, all sensitivities were computed with a single model run. MODFLOW-2000 has the capability of producing separate, spatially located sensitivities for each input parameter.

ACKNOWLEDGEMENTS

This work is a result of a combination of research projects directed by AJG (University of Alabama) and HWR (Northwestern University). Drs C H Dowding (NU) and T Igusa (Johns Hopkins University) greatly contributed to this effort. Support at UA was provided by the Alabama

Department of Public Health through the Alabama Legacy for Environmental Research Trust Grant Program. The work at NU was supported in part by the US Environmental Protection Agency STAR Program through grant R 827126-01-0. The work was performed prior to HWR joining the US Geological Survey and has not been subject to USGS review and Director's approval. It also has not been subjected to EPA review and, therefore, does not necessarily reflect the views of either USGS or EPA, and no official endorsement should be inferred.

REFERENCES

Casagrande, A. 1965 Role of 'calculated risk' in earthwork and foundation engineering. *J. Soil Mech. Found. Div., ASCE* **91** (SM4), 1–40.

Davis, C. D. 1986 *Statistics and Data Analysis in Geology*. John Wiley and Sons, Inc., New York.

Dethan, D. 2003 Quantitatively directing exploration for hazardous waste site characterization using MODFLOW-2000. *MS thesis*. University of Alabama, Tuscaloosa, AL.

Dowding C. H. (Ed.) 1978 *Site Characterization and Exploration. Proc. Geotech. Div.* ASCE, New York.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. 1995 *Bayesian Data Analysis*, Chapman and Hall, London, pp. 478–479.

Graettinger, A. J. & Dowding, C. H. 1999 Directing exploration with 3-D FEM sensitivity and data uncertainty. *J. Geotech. Geoenviron. Engng., ASCE* **125** (11), 959–967.

Graettinger, A. J., Lee, J. & Reeves, H. W. 2002 Efficient conditional modeling for geotechnical uncertainty evaluation. *Int. J. Numer. Analyt. Meth. Geomech.* **26**, 163–179.

Harr, M. E. 1996 *Reliability-based Design in Civil Engineering*, Dover Publications, New York, pp. 186–202.

Hill, M. C., Banta, E. R., Harbaugh, A. W. & Anderson, E. R. 2000 *MODFLOW-2000, The US Geological Survey Modular Ground-water Model – User Guide to the Observation, Sensitivity, and Parameter-estimation Process and Three Post-processor Programs*. Report 00-184. US Geological Survey.

Isaaks, E. H. & Srivastava, R. M. 1989 *An Introduction to Applied Geostatistics*, Oxford University Press, New York, pp. 278–322.

Lee, J. 2001 Reliability-based approach to groundwater remediation design. *PhD thesis*. Northwestern University, IL.

Matheron, G. 1963 Principals of geostatistics. *Econ. Geol.* **58**, 1246–1266.

National Research Council 1995 *Probabilistic Methods in Geotechnical Engineering*. Board on Energy and Environmental Systems, National Research Council, Washington, DC.

Shackelford C. D., Nelson P. P. & Roth M. J. S. (Eds.) 1996 *Uncertainty in the Geologic Environment: From Theory to Practice, Geotech. Special Publ. No. 58.* ASCE, New York.

Supriyasilp, T., Graettinger, A. J. & Durrans, S. R. 2003 Quantitatively directed sampling for main channel and hyporheic zone water quality modeling. *Adv. Wat. Res.* **26** (9), 1029–1037.