Sampling Network Design for Transport Parameter Identification

By Theodore G. Cleveland,[1] Associate Member, ASCE, and William W-G. Yeh,[2] Member, ASCE

Abstract: An optimal experimental design algorithm is developed to facilitate the planning and the optimal configuration and scheduling of a ground-water-tracer test whose data are to be used to estimate model parameters. A maximal information criterion is used to select among competing designs. The proposed criterion is equivalent to a weighted sum of squared sensitivities, employing the observation that parameters are most accurately estimated at points with high sensitivity to the parameter but that the relative magnitudes of sensitivities to different parameters are different. The fundamental advantage of this criterion is that it is comparatively simple. The influence-coefficient method is used to compute the sensitivity coefficient. A zero-one integer hueristic is used to solve a simplified example for experiment configurations under a given experimental duration. The design considers the installation cost, which is a function of location and depth of the observation well and of the samples themselves. The resulting designs are intuitively reasonable. It was found that a dramatic increase in information can be obtained with an experimental budget increase in a heterogeneous example case.

Introduction

Awareness of environmental quality and the desire to control pollution have resulted in the need to predict the movement of ground water and pollutants carried by the water. Mathematical models are used to predict the response of a ground-water system to various management policies. Estimation of the parameters used in these models is a critical step in their operation, and ultimately in the management decisions that are made.

Aquifer-response prediction is quite complex. The physical structure is unobservable. Since the structure cannot be observed directly, it must be inferred by the response at a few observation points. The observations are used to estimate parameters imbedded in the governing equations that describe the behavior of the system. This is the inverse problem of parameter identification, which has been studied by many authors. A state-of-the-art review of inverse procedures is given by Yeh (1986).

The inverse procedure is applied after a monitoring network has been installed and a sampling schedule proposed. The purpose of this paper is to examine a technique to configure a monitoring network whose data will be used to estimate parameters for a ground-water-contamination model.

Literature Review

The basic experimental design problem consists of determining the observation locations and sampling times so that the data obtained yield the

[1]Postdoctoral Scholar, Dept. of Civ. Engrg., Univ. of California, Los Angeles, Los Angeles, CA 90024.

[2]Prof., Dept. of Civ. Engrg., Univ. of California, Los Angeles, Los Angeles, CA.

Note. Discussion open until April 1, 1991. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on September 13, 1989. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 116, No. 6, November/December, 1990. ©ASCE, ISSN 0733-9496/90/0006-0764/$1.00 + $.15 per page. Paper No. 25265.

most information for estimating the parameters for a given cost. A design algorithm considers the trade-off between the economy of single site sampling with many samples versus the marginal decrease in information at a site as an experiment proceeds. Experimental design procedures developed in the field of statistics are used in medicine, automatic control, physics, and other areas, and can be applied to ground-water systems as well.

St. John and Draper (1979) reviewed the D-optimal design criterion and presented algorithms for constructing such designs. A D-optimal design is a design that satisfies all constraints and minimizes the determinant of the estimates' covariance matrix. Qureshi et al. (1980) applied D-optimality to identify locations of sensors for two problems: a heat-diffusion process and vibrating string. Periodic boundary conditions were used to eliminate computational difficulties.

Steinberg and Hunter (1984) reviewed the history of experimental design schemes and compiled the more successful schemes. Among these, A-optimal and D-optimal designs are discussed. An A-optimal design satisfies all constraints as well as minimizing the trace of the estimated parameters' covariance matrix.

It was not until recently that the problem of experimental design in ground water received its attention. Yeh and Sun (1984) developed an extended identifiability criterion, called the δ-identifiability, which can be used for the design of a pumping test. A δ-identifiable pumping test is an experiment that produces sufficient data to guarantee that the parameter estimates of the simulation model yield predictions that are sufficiently accurate for the overall management objective. McCarthy and Yeh (1989) used this approach to obtain minimum-cost pumping tests for a hypothetical aquifer where the uncertain parameter is transmissivity.

Carrera et al. (1984) proposed a scheme to locate observation sites for sampling of fluoride concentrations in an aquifer by minimizing estimation variance of the average fluoride concentrations.

Hsu and Yeh (1989) formulated an experimental design problem of a ground-water flow system. The objective was to minimize test cost subject to parameter-reliability constraints and some institutional constraints. They used the A-optimal reliability criterion to solve several parameter-identification problems in ground-water hydraulics, where the unknown parameter is transmissivity.

Nishikawa and Yeh (1989) used D-optimality as a reliability criterion to generate optimal pumping tests for a hypothetical aquifer where the uncertain parameter is transmissivity. The objective was to minimize test cost subject to parameter-reliability constraints.

In mass transport, the statistical implications for the transport parameters' identification were studied by Wagner and Gorelick (1987). They found that parameters are more reliably estimated if sampling is distributed in both space and time. They did not study the effect of structural inhomogeneity, but stated such studies might be useful.

Knopman and Voss (1987) studied the behavior of sensitivities in one-dimensional solute-transport equations and the implications for parameter estimation and sampling design. They found that parameters are most accurately estimated at points with high sensitivity to the parameter, but designs that minimize the variance of one parameter may not simultaneously minimize the variance of others. They reported that maximum sensitivity to velocity is a function of spatial location and hence experiment duration, while maximum sensitivity to dispersion is a function of sample frequency. They studied the effect of various experimental designs on the determinant of the estimated parameter's covariance matrix and found that the design with the smallest determinant gave the most reliable estimates with regard to estimate variance.

The importance of experimental design in connection with the planning of monitoring networks has been recognized for some time. Recent papers include Kaunas and Haimes (1985), Strecker et al. (1985), Meyer and Brill (1988), Knopman and Voss (1988), and Loaiciga (1989).

Model Aquifer

The experimental design algorithm is applied to a model aquifer that describes the features of the aquifer shown in Fig. 1. The aquifer is confined, piecewise homogeneous, and isotropic. The injection location and extraction location are assumed known. The experimental design problem seeks to determine the placement of multilevel sampling wells and the sampling frequency. It is also assumed that flow field can be approximated by a two-dimensional model.

Two-dimensional confined flow in porous media is described by Bear (1972, 1979).

$$S\frac{\partial H}{\partial t} = \frac{\partial}{\partial x}\left(K_{xx}\frac{\partial H}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_{yy}\frac{\partial H}{\partial y}\right) + M \quad \ldots\ldots\ldots\ldots \quad (1)$$

subject to the following initial and boundary conditions:

$$H(x,y,0) = \text{known } (x,y) \in \Omega \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (2a)$$

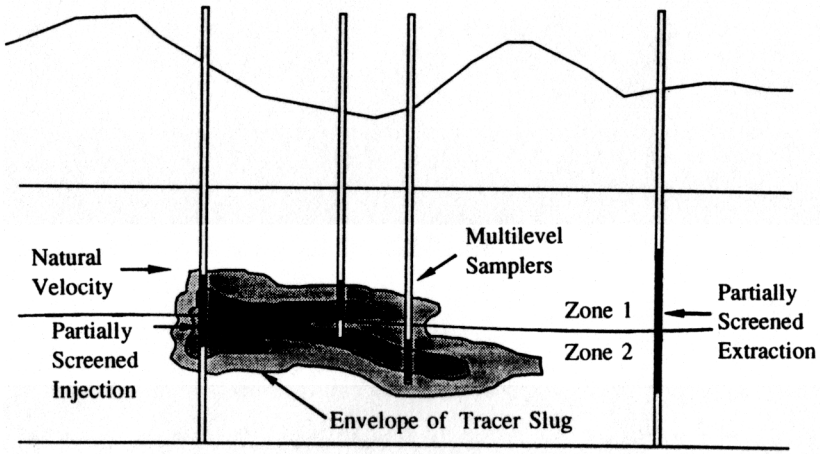

FIG. 1. Conceptual Aquifer

$$H(x,y,t) = \text{known } (x,y) \in d\Omega_1 \quad (2b)$$

$$\left(K_{xx}\frac{\partial H}{\partial x}\right)\frac{\partial x}{\partial n} + \left(K_{yy}\frac{\partial H}{\partial y}\right)\frac{\partial y}{\partial n} = \text{known } (x,y) \in d\Omega_2 \quad (2c)$$

where H = hydraulic head (L); K_{xx} = hydraulic conductivity (x-plane, x-direction, L/T); K_{yy} = hydraulic conductivity (y-plane, y-direction, L/T); S = storage coefficient; M = net injection-extraction rate (L/T); Ω = flow region; $d\Omega$ = boundary of flow region ($d\Omega_1 \cup d\Omega_2 = d\Omega$); and $\partial/\partial n$ = normal derivative to boundary.

The governing equation of the solute-transport process used in the present study is given by Bear (1972, 1979)

$$R\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D_{xx}\frac{\partial C}{\partial x} + D_{xy}\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial y}\left(D_{xy}\frac{\partial C}{\partial x} + D_{yy}\frac{\partial C}{\partial y}\right)$$

$$- V_x\frac{\partial C}{\partial x} - V_y\frac{\partial C}{\partial y} + M \quad (3)$$

subject to the following initial and boundary conditions:

$$C(x,y,0) = \text{known } (x,y) \in \Omega \quad (4a)$$

$$C(x,y,t) = \text{known } (x,y) \in d\Omega_1 \quad (4b)$$

$$\left(C\frac{V_x}{R} - \frac{D_{xx}}{R}\frac{\partial C}{\partial x} - \frac{D_{xy}}{R}\frac{\partial C}{\partial y}\right)\frac{\partial x}{\partial n} + \left(C\frac{V_y}{R} - \frac{D_{yx}}{R}\frac{\partial C}{\partial x} - \frac{D_{yy}}{R}\frac{\partial C}{\partial y}\right)\frac{\partial y}{\partial n}$$

$$= \text{known } (x,y) \in d\Omega_2 \quad (4c)$$

where C = mass of solute per volume of medium (M/L^3); D_{xx}, D_{xy} ... = components of hydrodynamic dispersion tensor; R = retardation factor; V_x = average fluid velocity in x-direction (L/T); V_y = average fluid velocity in y-direction (L/T); M = net mass injection-extraction rate (M/T); Ω = flow region; and $d\Omega$ = boundary of flow region ($d\Omega_1 \cup d\Omega_2 = d\Omega$).

Eq. 3 reflects certain implicit assumptions. First, there is no generation or decay of the solute. Also, adsorption is described by a linear equilibrium isotherm, hence the use of a retardation factor (Bear and Verruijt (1987). Included are terms for sources and sinks. Imbedded in all terms is the porosity, which permits the transformation of mass per fluid volume and mass per medium volume. The two-dimensional approximation of the flow field ignores head and concentration variations in the z-direction, which is horizontal in this case. This approximation is appropriate for a line of closely spaced parallel injection and extraction wells or for two parallel ditches. For a single pair of injection and extraction wells, radial flow would be significant near them and the formulation would be different. However, this two-dimensional approximation is adopted for demonstrating the proposed sampling network design. The hydrodynamic-dispersion coefficients for an isotropic porous medium are expressed by Bear (1972, 1979):

$$D_{xx} = (\alpha_L - \alpha_T)\frac{V_x^2}{V} + \alpha_T V + D^* \quad (5)$$

$$D_{xy} = (\alpha_L - \alpha_T)\frac{V_x V_y}{V} \quad (6)$$

$$D_{yy} = (\alpha_L - \alpha_T)\frac{V_y^2}{V} + \alpha_T V + D^* \quad (7)$$

where $V = (V_x^2 + V_y^2)^{1/2}$; α_L = longitudinal dispersivity; α_T = transverse dispersivity; and D^* = molecular diffusivity. The distribution of average fluid velocities is computed using Darcy's Law, written

$$V_x = -\frac{K_{xx}}{n}\frac{\partial H}{\partial x} \quad (8)$$

$$V_y = -\frac{K_{yy}}{n}\frac{\partial H}{\partial y} \quad (9)$$

where n = average porosity of the porous medium. In this study, molecular diffusion is ignored.

The transport equation is coupled to the flow equation through the average fluid velocities, this approximation is suitable for regional, shallow flow regime (two dimensional). If the solute significantly alters the density of the solvent (water) as a function of concentration, the flow model must be altered; the two equations coupled and solved simultaneously with the aid of equation of state. For simplicity it is assumed that this is not the case, and the flow equation can be solved independently.

Various finite difference, finite element, and other methods have been proposed for the numerical solution of these partial-differential equations. For simplicity, an explicit finite difference scheme is employed to solve these equations. An upwind formulation is used for the velocity terms of the transport equation to alleviate overshoot and undershoot associated with the numerical solution when advection dominates. An upwind formulation models an advective velocity front by averaging velocities at the node of interest and the node directly upward (or upgradient in this paper).

Solutions to Eqs. 1 and 3 were obtained using a forward Euler scheme. Centered-difference approximations were used for the spatial-discretization equation (Eq. 1). The stability criterion for the flow model is the smaller of

$$\Delta t \le \frac{1}{2} K_{xx}\Delta x^2 \quad (10a)$$

or

$$\Delta t \le \frac{1}{2} K_{yy}\Delta y^2 \quad (10b)$$

where Δ_x = characteristic length of discretized domain in x-direction; and Δ_y = characteristic length of discretized domain in y-direction.

In Eq. 3, the space discretization of the second-order partials is the same, leading to an analogous stability criterion, with D_{xx}, D_{yy} replacing the conductivities in Eq. 10. The first-order partials are approximated using forward or backward differences, depending on the local velocity direction (upwind formulation) when the local Peclet number exceeds 2.0. The Peclet P_e number in the x-direction is computed from

$$P_{e_x} = \frac{|V_x|\Delta x}{|\mathbf{D}|} \quad (11)$$

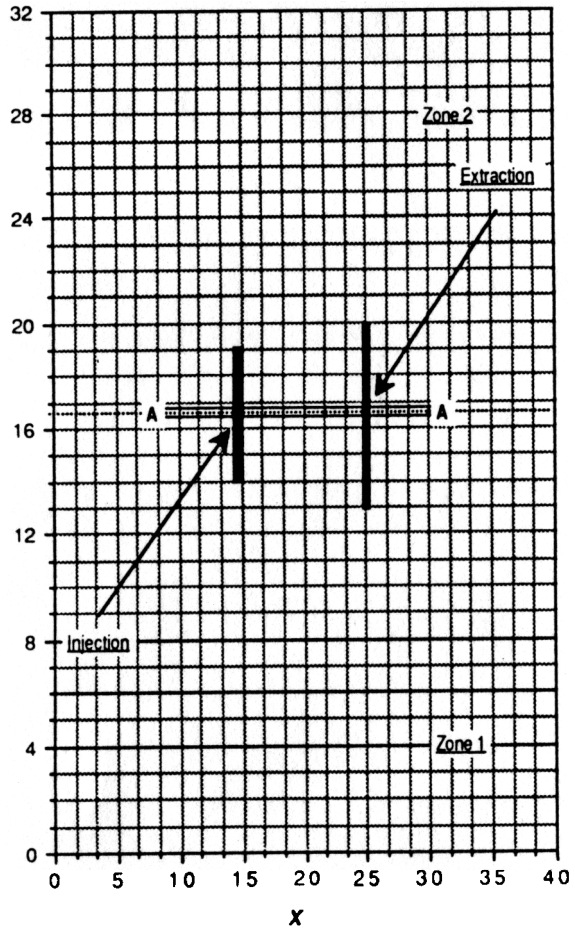

FIG. 2. Computational Domain

where $|\mathbf{D}|$ = some norm of the dispersion tensor. In this study, $|\mathbf{D}| = (D_{xx}^2 + D_{yy}^2)^{1/2}$. When velocity is significant, an additional stability criterion must be met (the Courant criterion):

$$\Delta t \leq \frac{\Delta x}{V_x} \quad \ldots\ldots (12)$$

Analogously, the Peclet number and Courant criteria are computed in the y-direction, and the smallest Δt and appropriate first-order partial approximations are used.

For this study, stability was determined by trial and error, Δt being adjusted until the scheme remained stable for the duration of the simulation. Fig. 2 shows the finite difference approximation grid used to simulate the aquifer in Fig. 1. The space discretization used is $\Delta x = \Delta y = 0.5$ m.

The optimal design selection is based on an information measure constructed from model sensitivities. Model sensitivities are the partial derivative of the state variable with respect to each of the model parameters. The method of computing sensitivities used here is the influence coefficient method described by Yeh (1986).

These sensitivities are then assembled into the Jacobian matrix, for instance, suppose a two-point design is used (i.e. $\{(x_1, t_1); (x_2, t_2)\}$), then the Jacobian would have the form

$$\mathbf{J} = \begin{bmatrix} (\partial C/\partial\theta_1) & (x_1, t_1, \theta) & (\partial C/\partial\theta_2) & (x_1, t_1, \theta) & \ldots & (\partial C/\partial\theta_k) & (x_1, t_1, \theta) \\ (\partial C/\partial\theta_2) & (x_2, t_2, \theta) & (\partial C/\partial\theta_2) & (x_2, t_2, \theta) & \ldots & (\partial C/\partial\theta_k) & (x_2, t_2, \theta) \end{bmatrix} \quad \ldots\ldots (13)$$

where $\boldsymbol{\theta}$ = parameter vector, which contains $(\theta_1, \theta_2, \ldots, \theta_k)$.

Methodology and Assumptions

The optimal design algorithm uses the numerical simulation model to locate as well as schedule the multilevel sampling sites. The performance criterion used to evaluate a particular design is based on the estimate's information matrix, assuming a least-squares-parameter-identification scheme is used. The approach is different from others [e.g., Hsu and Yeh (1989)] in that the selection criterion is used as the evaluator rather than a reliability constraint. The performance measure is constructed from the sensitivity information. For an additive-error model

$$C_{\text{Observed}} = C_{\text{Model}}\boldsymbol{\theta} + \text{Error} \quad \ldots\ldots (14)$$

where

$$E(\text{Error}) = 0 \quad \ldots\ldots (15)$$

$$V(\text{Error}) = \Sigma \quad \ldots\ldots (16)$$

in which Σ = common covariance matrix of the error. The symbols E and V denote mathematical expectation and variance-covariance operator. Assuming a least-squares-parameter-identification scheme is used, the linear approximation to the estimates' covariance matrix is (Yeh and Yoon 1981):

$$\mathbf{V\theta} = \mathbf{M\Sigma M}^T \quad \ldots\ldots (17)$$

where

$$\mathbf{M} = (\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T \quad \ldots\ldots (18)$$

In the special case where the errors are not correlated and have equal variance

$$\Sigma = \sigma^2\mathbf{I} \quad \ldots\ldots (19)$$

where σ^2 = variance; and $\mathbf{I}$ = identity matrix. Then, the approximation of the estimate's covariance matrix is

$$\mathbf{V\theta} = \sigma^2(\mathbf{J}^T\mathbf{J})^{-1} \quad \ldots\ldots (20)$$

The information matrix for the special case in Eq. 17 is

$$\mathbf{I\theta} \quad \frac{1}{\sigma^2}(\mathbf{J}^T\mathbf{J}) \quad (21)$$

To ensure small variance of the estimates, it is desired that the information matrix be large in some sense. Since different designs only change the Jacobian, the $1/\sigma^2$ term is ignored. The proposed measure is the trace of the weighted information matrix. This is equivalent to a weighted sum of squared sensitivities, employing the observation that parameters are most accurately estimated at points with high sensitivity to the parameter but that the relative magnitudes of sensitivities to different parameters are different. The fundamental advantage of this criterion over that of *D*- or *A*-optimality is that it is computationally simple. The $\mathbf{J}^T\mathbf{J}$ matrix never has to be completely constructed or inverted. The information matrix is evaluated at the prior estimates of parameters, which are assumed to be available before performing the experiments.

The method of choosing weights is to compute the sensitivities for each parameter at each time. These sensitivities are then integrated with respect to time. These "aggregate" sensitivities reflect the sensitivity of a parameter over the entire computation domain. The weights are factors that are required to make each aggregate sensitivity equal to unity. This way, the effect of different order of magnitude sensitivities dominating the solution is reduced. The criterion (the trace of the weighted information matrix, in which $\mathbf{I}\theta = \mathbf{J}^T\mathbf{J}$) at some point in space and time is written

$$I_{i,l} = \text{tr}(\mathbf{J}_{i,l}^T\mathbf{J}_{i,l}\mathbf{W}) = \sum_{j=1}^{k} w_j\left(\frac{\partial C_{i,l}}{\partial \theta_j}\right)^2 \quad \ldots\ldots (22)$$

where $I_{i,l}$ = information at ith location lth time; $\mathbf{J}_{i,l}$ = Jacobian matrix at ith location lth time; w_j = weight on jth parameter; $\mathbf{W}$ = diagonal matrix of weights with determinant one; and k = number of parameters. Allowing $\mathbf{W}$ to be any nonnegative matrix generates different criteria, some of which may include cross sensitivities. Here, only diagonal matrices are used, which is a special class of more general criteria discussed by Sacks and Ylvisaker (1968).

Using this measure of information and some assumptions about costs, the optimization model is a finite-dimension integer-programming problem. It is assumed that the overall experiment duration is known. It is also assumed (for computational simplicity) that once sampling has begun at a site, it continues until the end of the experiment. With these assumptions, the objective function is constructed as the sum of time-integrated information numbers at each site. For instance, if site (x_i, y_i) is considered, and sampling is begun at time (t_l) and the experiment ends at time (t_{NT}), then the total information available at that site is:

$$\text{Total Information } (x_i, y_i) = \sum_{l=J}^{NT} I(x_i, y_i, t_l, \theta) \quad \ldots\ldots (23)$$

where NT = number of possible sampling times; and $I(x_i, y_i, t_l, \theta)$ = unit information number at site (x_i, y_i) and time (t_l) for the parameter set θ.

It is assumed that the design parameters concerning the installation, development, and operation of the injection and extraction system are known and money has been budgeted. The remaining discretionary budget is allocated to sample-site installation, sample collection, and analysis. Installation cost is a function of location and depth as well as the cost of the sampler (pump) itself. In a manner analogous to the compution of total information,

the cost at a site is the time-integrated cost of the installation and the consequent sampling costs.

$$\hat{C}(x_i, y_i)_j = \bar{\bar{C}}(x_i, y_i) + \sum_{l=j}^{NT} \bar{C}(x_i, y_i, t_l) \quad \ldots\ldots (24)$$

where $\bar{\bar{C}}(x_i, y_i)$ = installation cost at site (x_i, y_i); and $\bar{C}(x_i, y_i, t_l)$ = sample cost at site (x_i, y_i) time t_l. These costs are computed as a combined function $\hat{C}(x_i, y_i)$, which is the cost of installation plus sample collection from time j until NT.

A zero-one indicator variable is used to identify which sites are selected. A value of one means a particular site is chosen. The zero-one variable is double-subscripted. The first subscript is location, the second is sample-initiation time. For instance, $z_{2,7} = 1$ means site 2 (with locations x_2, y_2) with sampling starting at time (t_7) is indicated as a selected point.

The optimal design problem can be written

$$\max_{z_{ij}} \sum_{i \epsilon S} Z_{i,1}\left[\sum_{j=1}^{NT} I(x_i, y_i, t_j, \theta)\right] + \ldots + z_{i,NT}\left[\sum_{j=NT}^{NT} I(x_i, y_i, t_j, \theta)\right] \quad (25)$$

subject to

$$\sum_{i \epsilon S} z_{i,1}\,\hat{c}(x_i, y_i)_1 + \quad + z_{i,NT}\hat{c}(x_i, y_i)_{NT} \leq \text{budget} \quad (26a)$$

$$z_{i,1} + \ldots + z_{i,NT} \leq 1, \forall i \epsilon \bar{S} \quad \ldots\ldots (26b)$$

$$z_{i,NT} = 0 \quad \text{or} \quad 1, \forall i \epsilon \bar{S} \quad \ldots\ldots (26c)$$

where n_x = number of x-locations; n_y = number of y-locations; $\bar{S} = (1, 2, \ldots, n_x \cdot n_y)$; and $I(x_i, y_i, t_i, \theta)$ = unit information at (x_i, y_i) and time t.

The solution of this multidimensional 0-1 integer-programming problem defines the optimal configuration and schedule for the sampling-network design.

The multidimensional 0-1 integer-programming problem described in Eqs. 25 and 26 is approximated by only considering the first column of the optimization problem. That sampling is started at time 1 and continued throughout the experiment. This leads to the following unidimensional 0-1 integer-programming problem:

$$\max \sum_{i \epsilon S} z_{il}\left[\sum_{j=1}^{NT} I(x_i, y_i, t_j,)\right] \quad \ldots\ldots (27)$$

subject to

$$\sum_{i \epsilon S} z_{il}\,c(x_i, y_i)_1 \leq \text{budget} \quad \ldots\ldots (28a)$$

$$z_{il} = 0 \quad \text{or} \quad 1, \forall_{i \epsilon S} \quad \ldots\ldots (28b)$$

This problem is solved using a polynomial in a time-approximation scheme described in detail by Papdimitriov and Steiglitz (1982). Using this approximation ignores sample initiation time, thus it is a configuration algorithm based on time-integrated information at each location. It is felt that this approximation reasonably describes a field study where to ensure nothing is

missed everything is turned on sampling is performed without regard to incoming information.

Application

Example 1

An example of the methodology is presented. The aquifer of Fig. 1 is first assumed homogeneous, i.e., zones 1 and 2 have the same properties. The installation cost of any sampling site is 10.0 units, while sampling and analysis is 1.0 units. Fig. 2 shows the computational grid used as well as the relative locations of the injection and extraction sites. The distance between the two sites is 5 m. A uniform discretization of $\Delta x = \Delta y = 0.5$ m is used in the numerical model. In Fig. 2, and the rest of the figures, distances in the x- and y-directions are represented in terms of the number of increments in Δx and Δy from the origin. The approximate hydraulic travel time is 2.75 days at the average hydraulic velocity at steady state. The mass is injected for 0.40 days; sampling is at every 0.40 days for a total of 4.0 days. The injection and extraction rates are 0.83 gpm and 2.5 gpm, respectively, while the injected concentration is 1 mg/kg. Table 1 shows the prior estimates of the model parameters. It is important to note at this point that the results of experimental design are predicted on the prior estimates. Column 1 of Table 1 lists the parameters of interest, Column 2 gives the values assumed in zone 1 as shown in Fig. 2, and column 3 gives the values assumed in zone 2. In this example, the values for each parameter in each zone are the same.

Fig. 3 shows the optimal design for three budget levels. The notation of Fig. 3 is the following: The vertical line labelled *a* represents design A in column 2 of Table 2. This line represents sampling at six locations whose x-coordinate is fixed. The actual locations are indicated by dots on the graph. The vertical lines labelled *b* (which include line *a*, for a total of 12 locations) represent design B in column 2 of Table 2. The line *c* (along with lines *a* and *b*) represents design C in column 2 of Table 2, for a total of 18 locations. Table 2 lists the information available with each design at each budget level. As budget is increased (column 1, Table 2), the designs change and the information obtained (column 3) changes. The last column (column 4), shows the incremental increase in information for each successive budget increase. The location of the single point (design A) is exactly at the location where one would expect the peak concentration to have arrived at one-half the experiment duration time.

TABLE 1. Parameters for Example 1

Parameter (1)	Value in zone 1 (2)	Value in zone 2 (3)
K_{xx}	3.0	3.0
K_{yy}	3.0	3.0
S	0.1	0.1
n	0.5	0.5
α_L	0.1	0.1
α_T	0.1	0.1
R	1.0	1.0

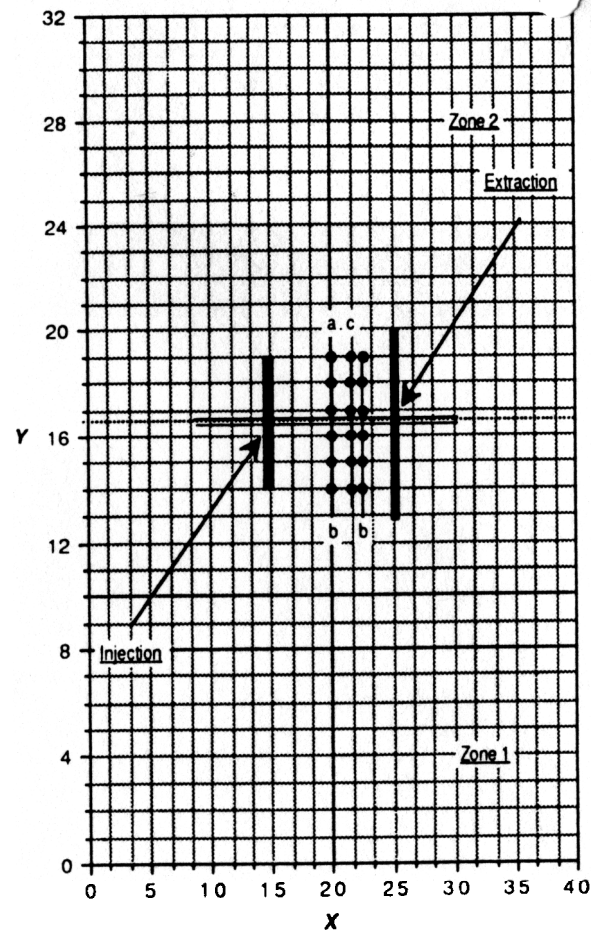

FIG. 3. Optimal Designs for Example 1

Fig. 4 shows concentration profiles in the cross section A-A aligned with the axis of symmetry between the injection and extraction points (dominant streamline). Fig. 5 shows the information measure at the same three times. Observe that the information moves in space with time as well as decays over time.

TABLE 2. Tradeoff Table for Example 1

Budget (1)	Design (2)	Information (3)	Δ Information (4)
22.0	A	12.18	—
44.0	B	24.41	12.23
66.0	C	36.42	12.01

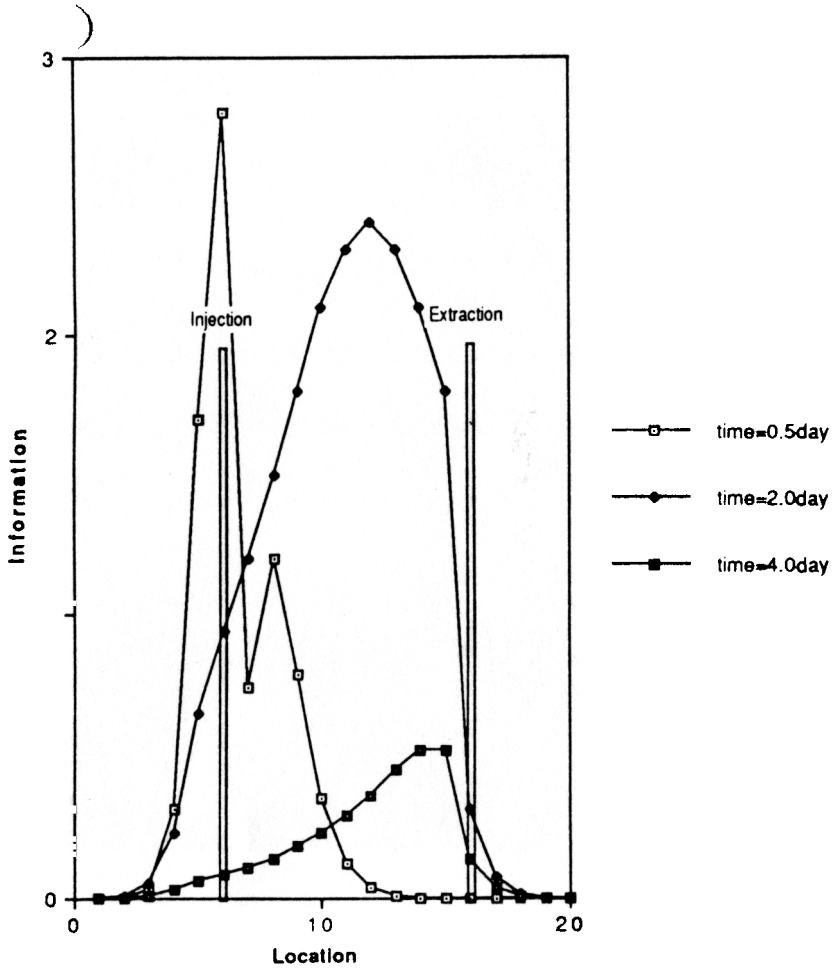

FIG. 4. Concentration Profiles through Section A-A

From Table 2, it is observed that as budget increases linearly the information increase is less than linear; this reinforces the observation by Strecker et al. (1985) that the incremental improvement of a sampling network as observations are added diminishes over time.

An observation that can be made from the results of this design algorithm are that the optimal single point is the location where the hydraulic gradient would be expected to carry the mass in half the experiment duration. An explanation of this is that maximum sensitivities to different parameters occur at different times during the passage of the concentration pulse, but sampling at the one-half distance allows sampling to occur before, during, and after a peak pulse has passed, thereby allowing for more information to be gained for all parameters.

Example 2

The experimental design algorithm is repeated assuming the aquifer of

TABLE 3. Parameters for Example 2

Parameter (1)	Value in Zone 1 (2)	Value in Zone 2 (3)
K_{xx}	1.5	3.0
K_{yy}	1.5	3.0
S	0.1	0.1
n	0.5	0.5
α_L	0.1	0.1
α_T	0.1	0.1
R	1.0	1.0

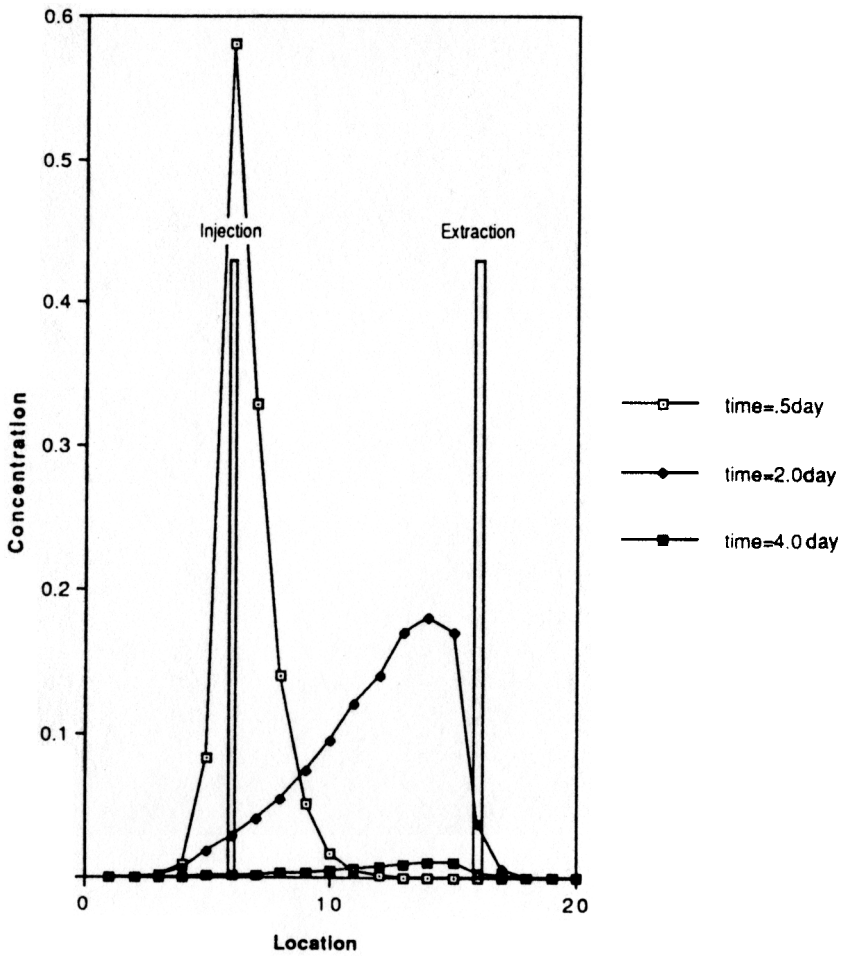

FIG. 5. Information Measure through Section A-A

TABLE 4. Tradeoff Table for Example 2

Budget (1)	Design (2)	Information (3)	Δ Information (4)
22.0	A	21.14	—
44.0	B	21.87	0.73
66.0	C	43.01	21.14
88.0	D	43.03	0.02

Fig. 1 is piecewise inhomogeneous with a hydraulic conductivity contrast of 2:1. Table 3 shows the model parameters (prior estimates) used. Column 1 of Table 3 lists the parameters of interest, column 2 gives the values assumed in zone 1 of Fig. 2, and column 3 gives the values assumed in zone 2. Table 4 lists the information available with each budget level in a fashion analogous to Table 2.

Fig. 6 shows the four optimal designs. The vertical line labelled *a* corresponds to design A in Table 4, with a total of six locations. The two lines labelled *B* correspond to design B in Table 4, for a total of six locations.

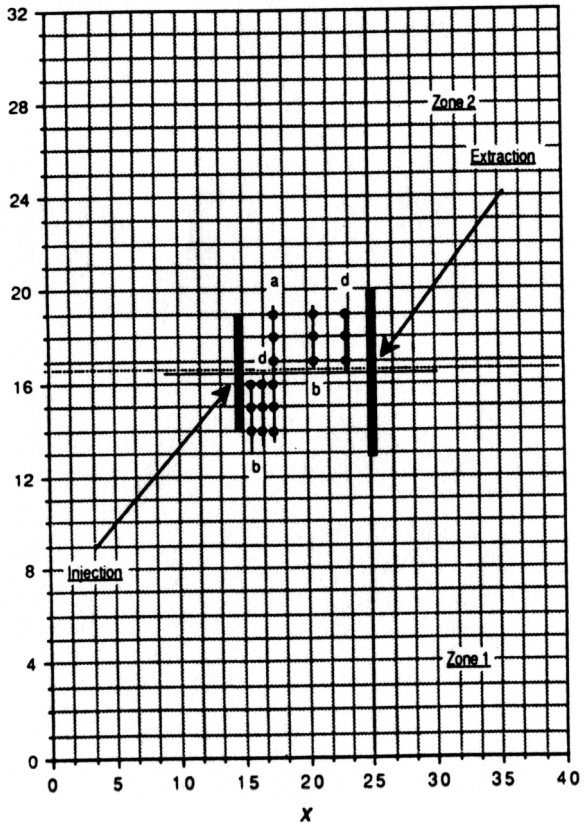

FIG. 6. Optimal Designs for Example 2 (Inhomogeneous)

In this example, we use the multilevel behavior of the design algorithm where the locations are now different for each zone. The lines labelled *b* and the one labelled *a* (or two, one in each zone) taken together are design C in Table 4, with a total of 12 locations. In design D, the lines are labelled *b* and *d*, with a total of 18 locations.

In this example, the tradeoff between budget and information is more pronounced. Using four times the budget, from 22.0 to 88.0, only gives a little more than twice the information. Obviously, if suboptimal design points are chosen this would not be the case. An important implication is that a well-selected point (given that the governing parameters are somewhat known) gives a lot of information. Design C is probably the most informative per unit of budget that the two zones are significantly different.

To summarize the results of each example, first, a useful amount of information is available at the half-duration location, although in the heterogeneous case multilevel sampling gives a dramatic information increase. Second, both examples illustrate the diminishing return in terms of information with increased experimental budget. More examples can be found in Cleveland (1989).

Monte Carlo Analysis

The two examples presented assume that prior estimates of the parameters are available. It is safe to say that some estimates of the parameters can be made prior to performing an experiment (one can always guess). A sequential design approach can be used to update estimates. The fundamental difficulty is that the parameters are required in order to identify an optimal design to obtain the data to estimate them, but they are unknown beforehand. In practice, initial estimates of parameters are determined from prior information. Based upon the initial estimates, experimental design is carried out and data collected. The inverse problem of parameter identification is solved to update parameter estimates and the design of the experiment is repeated. The concept of sequential design and its convergence property are reported by Hsu and Yeh (1989) and Nishikawa and Yeh (1989). In this study, sequential design is not performed; however, the range of parameters for which a design remains optimal is explored.

Two approaches that are used are to design an experiment that performs well in the average sense or to design an experiment that performs well in the min-max sense. D-optimality is a common criterion for these approaches, although in general the approaches generate different designs (Walter and Pronzato 1987). Although these approaches identify robust designs, they are not independent of the underlying parameters that are unknown and thus dependent on the estimates (which are used to approximate the covariance matrix). To gain insight in this respect, the present study uses a Monte Carlo approach to identify parameter ranges for which a design remains optimal. Since it has been demonstrated that the parameters of interest change the objective surface in a complex fashion, this trial-and-error approach should yield some insight that can be used to improve the robustness of a design.

The approach is the following: Parameters are chosen from a uniform distribution with known upper and lower bounds. Several thousand realizations are performed for some budget level, and the frequency of selected points is saved. With parameters at some setting and small ranges, a unique design

TABLE 5. Parameter Ranges

Parameter (1)	Zone 1 (2)	Zone 2 (3)
K_{xx}	1.0–2.0	2.5–3.5
K_{yy}	1.0–2.0	2.5–3.5
S	0.9–1.1	0.9–1.1
α_L	0.25–0.75	0.25–0.75
α_T	0.25–0.75	0.25–0.75
R	0.75–1.25	0.75–1.25

should appear with frequency 1.0. The ranges are adjusted until this frequency changes. This is then said to be the optimal design for all parameters in the stated ranges. Table 5 lists the parameter ranges used. Table 6 lists the results of the Monte Carlo simulation. It is observed that the stable designs are fairly robust for the hydraulic conductivity and retardation, but very sensitive to storage coefficient and dispersivities. It is noted here that the storage coefficients used are abnormally large (compared to observed values for real aquifers). This has the effect of making the model describe transient behavior longer than would be encountered using more realistic parameters.

For parameters outside these ranges the marginal decrease in information by switching from the optimal design to an average design can be approximated by comparing the information in the average design with the information in the optimal design. On the average, the information loss is dramatic, about 70%. This is alarming in that a one-shot approach will have serious difficulties if the prior information is poor.

The marginal decrease in information is calculated by taking the difference in information (in a 32 point design) between the optimal design (for the out-of-range parameter set) and the average design. The difference is divided by the information in the optimal design to compute the proportion that the difference represents. The implication of such a dramatic difference is that the sequential design may be doomed from the start if a design is not robust enough for the prior information.

For instance, suppose the prior estimate is terrible with respect to the true parameter. This estimate is used to specify a design and an experiment is performed. Since the estimate is terrible, the design will surely be suboptimal. Suppose first the suboptimal design is adequate. The estimate will be updated, and a new design specified. On the other hand, suppose the information in the suboptimal design is inadequate so that the update is still

TABLE 6. Parameter Ranges for Stable Design

Parameter (1)	Zone 1 (2)	Zone 2 (3)
K_{xx}	1.26–1.98	2.57–3.45
K_{yy}	1.08–1.93	2.55–3.37
S	0.9–1.0	0.9–1.0
α_L	0.27–0.55	0.38–0.69
α_T	0.44–0.49	0.44–0.72
R	0.75–1.15	0.78–1.24

in the "robust" range of the initial suboptimal design. If this is the case, then the design will not change. However, the design won't change if it is optimal in the first place, either.

In the mathematical abstraction of the system, this should not occur in high-dispersion cases, since mass (albeit very small) will move instantaneously throughout the system. This is explained because the model is a parabolic partial-differential equation. In a low-dispersion case, when the equation becomes hyperbolic the mass will have finite velocities, and there are locations in the model where mass never appears. These locations will change as the parameters are changed. If the prior estimate causes the algorithm to choose a design that is a zero-information design for the true parameters, then the update will be worthless. But there is no way to know this beforehand, and although a zero-information design is useful in some sense (it tells where not to sample) it makes the next design difficult to specify. Clearly, an important avenue for more study is how to ensure that the first design gives some information, even with terrible prior information.

In the physical system, this behavior has been observed. Often in the hydraulics of wells, a radius of influence is used beyond which a particular well has no apparent effect on the system (at a particular pumping rate). If parameters concerning that well are to be estimated, clearly a design must take samples somewhere within the radius of influence, even though the mathematical abstraction may indicate information available beyond the physical radius of influence.

Conclusions

A maximal information criterion was applied to an experimental design problem of ground-water hydrology. The goal was to identify reasonable sampling configurations under a given experimental duration for obtaining data to estimate model parameters. The criterion has computational advantages over criteria, and identifies reasonable designs.

Practical implications illustrated by the method are that a desirable sampling location is at the one-half experiment duration. Also, the marginal increase in information as experimental budget is increased is not linear, that is, doubling the budget does not double the information.

Finally, tradeoffs must be evaluated to determine an appropriate budget, as illustrated by the last example. The smallest budget may give the most information per unit of budget, but if it is known that the sample domain contains heterogenities, a budget increase may give a dramatic increase in information.

A future direction of this research is to pursue the joint configuration and scheduling problem (i.e. the full-optimization model). Since in some sense the joint approach is a capacitated-expansion problem a dynamic programming approach like that of Becker and Yeh (1974) may be promising. The alternative is to solve the multidimensional-knapsack problem directly, which is computationally difficult as the design space grows larger.

It would also be desirable to compare results using a field test of D-optimal versus A-optimal versus the approach described here.

Acknowledgments

The writers wish to acknowledge the support of Prof. Douglas A. Mackay of the University of California, Los Angeles, whose field studies inspired

this work and confirmed the cost ratios used. The contents of this paper were developed under a grant from the Department of Interior, U.S. Geological Survey (Award No. 14-08-0001-G1499); however, the contents do not necessarily represent the policy of the agency, and endorsement by the federal government should not be assumed. The writers would like to thank the three anonymous reviewers for their constructive comments and suggestions.

Appendix I. References

Bear, J. (1972). *Dynamics of fluids in porous media*. American Elsevier, New York, N.Y.

Bear, J. (1979). *Hydraulics of groundwater*. McGraw-Hill Book Co., Inc., New York, N.Y.

Bear, J., and Verruijt, A. (1987). *Modelling groundwater flow and pollution*. Reidel, Boston, Mass.

Becker, L., and Yeh, W. W-G. (1974). "Optimal timing, sequencing, and sizing of multiple reservoir surface water supply facilities." *Water Resour. Res.*, 10(1), 57–62.

Carrera, J., Usnoff, E., and Szidarovsky, F. (1984). "A method for optimal observation network design for groundwater management." *J. Hydrol.*, 73, 147–163.

Cleveland, T. G. (1989). "Sampling strategies for transport parameter identification," thesis presented to the University of California, Los Angeles, at Los Angeles, Calif., in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Hsu, N. S., and Yeh, W. W-G. (1989). "Optimum experimental design for parameter identification in groundwater hydrology." *Water Resour. Res.*, 25(5), 1025–1040.

Kaunas, J. R., and Haimes, Y. Y. (1985). "Risk management of groundwater contamination in a multiobjective framework." *Water Resour. Res.*, 22(11), 1721–1730.

Knopman, D. S., and Voss, C. I. (1987). "Behavior of sensitivities in the one-dimensional advection-dispersion equation: Implications for parameter estimation and optimal design." *Water Resour. Res.*, 23(2), 253–272.

Knopman, D. S., and Voss, C. I. (1988). "Discrimination among one-dimensional models of solute transport in porous media: Implications for sampling design." *Water Resour. Res.*, 24(11), 1859–1876.

Loaiciga, H. A. (1989). "An optimization approach for groundwater quality monitoring network design." *Water Resour. Res.*, 25(8), 1771–1782.

McCarthy, J. M., and Yeh, W. W-G. (1990). "Optimal pumping test design for parameter estimation and prediction in groundwater hydrology." *Water Resour. Res.*, 26(4), 779–791.

Meyer, P. D., and Brill, E. D. Jr. (1988). "A method for locating wells in a groundwater monitoring network under conditions of uncertainty." *Water Resour. Res.*, 24(8), 1277–1282.

Nishikawa, T., and Yeh, W. W-G. (1989). "Optimal pumping test design for the parameter identification of groundwater systems." *Water Resour. Res.*, 25(7), 1737–1747.

Papadimitriou, C., and Steiglitz, K. (1982). *Combinatorial optimization*. Prentice Hall, Englewood Cliffs, N.J.

Qureshi, Z. M., Ng, T. S., and Goodwin, G. C. (1980). "Optimum experimental design for identification of distributed parameter systems." *Int. J. Control*, 31(1), 21–29.

Sacks, J., and Ylvisaker, D. (1968). "Designs for regression problems with correlated errors. Many parameters." *The Annals of Mathematical Statistics*, 39(1), 49–69.

St. John, R. C., and Draper, N. R. (1979). "D-optimality for regression designs: A review." *Technometrics*, 17(1), 15–23.

Steinberg, D. M., and Hunter, W. G. (1984). "Experimental design: Review and comment." *Technometrics*, 26(2), 71–97.

Strecker, E. W., Chu, W.-S., and Lettenmaier, D. P. (1985). "Evaluation of data requirements for groundwater contaminant transport modelling." *Tech. Report #94*, University of Washington, Seattle, Wash.

Wagner, B. J., and Gorelick, S. M. (1987). "A statistical methodology for estimating transport parameters: Theory and applications to one-dimensional advective dispersive systems." *Water Resour. Res.*, 23(7), 1162–1174.

Walter, E., and Pronzato, L. (1978). "Robust experiment design: Between qualitative and quantitative identifiabilities." *Identifiability of parametric models*. E. Walter, ed., Pergamon Press, New York, N.Y.

Yeh, W. W-G., and Sun, N. Z. (1984). "An extended identifiability in aquifer parameter identification and optimal pumping test design." *Water Resour. Res.*, 20(12), 1837–1847.

Yeh, W. W-G. (1986). "Review of parameter identification procedures in groundwater hydrology: The inverse problem." *Water Resour. Res.*, 22(2), 95–108.

Yeh, W. W-G., and Yoon, Y. S. (1981). "Aquifer parameter identification with optimum dimension in parameterization." *Water Resour. Res.*, 17(3), 664–672.

Appendix. Notation

The following symbols are used in this paper:

C = concentration;
$\bar{C}(x,y)$ = installation cost;
$\tilde{C}(x,y,t)$ = sample cost;
$\hat{C}(x_i,y_i)_j$ = combined cost;
$\mathbf{D}_{xx}, \mathbf{D}_{xy}, \mathbf{D}_{yy}$ = hydrodynamic dispersion tensor;
D^* = molecular diffusivity;
$d\Omega$ = boundary of flow region;
$E(\ )$ = expectation;
H = hydraulic head;
$\mathbf{I}(\ )$ = information matrix;
$I(x_i,y_i,t,\theta)$ = total information at x_i, y_i, t, θ;
$\mathbf{J}$ = Jacobian matrix;
K_{xx}, K_{yy} = hydraulic conductivity;
M = injection-extraction rate;
n = average porosity;
n_x = number of x-locations;
n_y = number of y-locations;
P_e = Peclet number;
R = retardation factor;
S = storage coefficient;
$\bar{S}$ = index set;
V_x, V_y = average fluid velocity;
$\mathbf{V}(\ )$ = variance, covariance;
$\mathbf{W}$ = weight matrix;
W_j = parameter weight;
$Z_{i,l}$ = information number at i, l; binary indicator variable;
$z(x,y,t,\theta)$ = unit information number;
α_L = longitudinal dispersivity;
α_T = transverse dispersivity;
Δ_t = time discretization;

Δ_x = space discretization x-direction;
Δ_y = space discretization y-direction;
$\partial/\partial n$ = normal derivative;
$\boldsymbol{\theta}$ = parameter vector;
σ^2 = variance; and
Ω = flow region.