APPLICATION OF KALMAN FILTER FOR ANALYSIS OF CONTAMINANT TRANSPORT IN GROUND WATER ORIGINATING FROM POINT SOURCES

B.K. Panigrahi[1] and J.C. Hwang[2]

[1]Senior Geohydrologist, ECKENFELDER INC.
Mahwah, New Jersey

[2]National Expert, UST Program, USEPA-Region III
Philadelphia, Pennsylvania

ABSTRACT

A sequential stochastic model is developed to identify the locations and strengths of pollution sources in a two-dimensional uniform ground water flow field. The model also estimates the concentration distribution in the flow field due to the presence of pollution sources and the error covariances of the estimation process. The present study focuses on incorporating the measurement errors and modeling errors. The findings presented herein are considered only a first step toward introducing the potential application of system theory to contaminant transport problems in the ground water environment. This paper presents details on the model development and the physical significance of various steps associated with its development. In addition, the following items are discussed in detail: the criteria for selecting initial source covariances, the role of measurement error covariances, the criteria for handling measurement errors when there is insufficient information to define those errors, the functional relationship between covariances and Kalman gain matrices, the model's sensitivity to initial assumptions of covariances, and the limitations of the model developed during the present study. A series of numerical experiments was conducted to support the discussions of the above listed issues for a case with a single point source and a constant injection rate.

INTRODUCTION

Ground water pollution sources can be identified by the conventional method of plotting concentration contours (Roux and Althoff, 1980), the non-sequential approach (DiStefano and Rath, 1982; Gorelick, 1983; Gorelick et. al., 1983), or the sequential approach (Hwang and

Koerner, 1983). These methods are based on deterministic approaches. The present study follows the sequential approach but, unlike the other studies, develops a stochastic linear dynamic system model. Although the present methodology is intended for stochastic simulation purposes (i.e., the direct problem), it serves as a prelude to the inverse problem within the framework of extended Kalman filtering, in which concentration measurements vary considerably in space and time. The primary utility of a stochastic model is its capability in formulating the system error, if desired. The system error may stem from incorrect choices of parameters (such as dispersivity, velocity, etc), measurement errors (such as in concentration), modeling errors (such as in physical processes, model discretization, source definition, etc.), or any combination thereof. The present study focuses on incorporating measurement and modeling errors related to pollution source definition. The findings presented herein are considered only a first step toward the application of system theory to contaminant transport problems in the ground water environment.

MODEL DEVELOPMENT

Mathematical Model

The equation governing transport of a dissolved solute in a two-dimensional uniform ground water flow field can be written as:

$$\begin{aligned} R(\partial C/\partial t) &= \nabla D\nabla C - \nabla(VC) - \lambda RC + S \qquad (x,y) \in \Omega \\ a_1 C \cdot n + a_2 C &= a_3 \qquad (x,y) \in \partial\Omega \end{aligned} \tag{1}$$

where $C(x,y)$ is the concentration of the dissolved solute $[M/L^3]$; $D(x,y)$ is the coefficient of hydrodynamic dispersion $[L^2/T]$; $V(x,y)$ is the seepage or pore velocity $[L/T]$; λ is the first-order decay constant of the solute $[1/T]$; R is the retardation factor; S is the material flux of the source/sink $[M/L^3T]$; a_1, a_2, and a_3 are constants whose values determine the type of boundary conditions; n is the outward normal vector to the boundary at a given point; Ω is the flow domain; $\partial\Omega$ is the domain boundary; and ∂ and ∇ are the differential operators.

Linear Dynamic System Model

The first step in the process of model development is to transform the governing non-linear partial differential equation, eq. (1), into a form of linear dynamic model. This is achieved by spatial discretization of the flow domain, eq. (1), using a numerical method. There are a number of numerical methods, such as finite analytic, finite difference, or finite element, which can be used for this purpose. Finite analytic method is used for the present study. The end result of spatial discretization of eq. (1) is a linear dynamic system given by:

$$\partial C/\partial t = A'C + B'S \qquad \text{(system model)} \tag{2a}$$

$$C^* = H\,C \qquad \text{(measurement model)} \tag{2b}$$

where A' is the state (concentration) coefficient matrix which contains information relating concentration vector and aquifer parameters, such as

D, V, R, and λ; B' is a diagonal control (source) matrix expressing relationship between the source and the concentration vectors; and H is the observation coefficient matrix which describes the relationship between the measurements C^* and the concentration C.

Stochastic System Model

The second step in the process of model development is to obtain a discrete-time stochastic system model. Mathematically, identifying the pollution source location and strength is the same as identifying the vector S. This forms the basis for a stochastic measurement model as well as for a deterministic simulation model. Assuming the true state vector to be a random process, the continuous-time stochastic system model can be expressed as (McLaughlin, 1978):

$$\partial C/\partial t = A'C + B'S + p \qquad \text{(system model)} \qquad (3a)$$

$$C^* = H\,C + W \qquad \text{(measurement model)} \qquad (3b)$$

with error statistics given by:

$$E[C(t_o)] = C_o \;;\; E[C(t_o)\,C^T(t_o)] = P_o \qquad (3c)$$

$$E[p(t)] = 0 \;;\; E[p(t)\,p^T(t+\tau)] = Q\delta(t-\tau) \qquad \text{for all } t \qquad (3d)$$

$$E[W(t)] = 0 \;;\; E[W(t)\,W^T(t+\tau)] = R\delta(t-\tau) \qquad \text{for all } t \qquad (3e)$$

where p is a vector of process noise (model error) terms applied at the spatial grid points; $E[\bullet]$ is an expectation operator; δ is the Dirac delta function; superscript T denotes a transpose operation; and P, Q, and R are the error covariances of state, process noise, and measurement errors, respectively.

The uncertainty of the model can be attributed to uncertainties in the aquifer properties, solute characteristics, concentration measurements, pollution sources, numerical methods, etc. Assuming that the solute characteristics and aquifer properties do not change in time; that the spatial discretization of the domain does not introduce significant errors; that the numerical method accurately simulates the flow field; and that the pollution sources (control parameter) and the aquifer properties are deterministic in nature, the process noise can be eliminated and the discrete-time stochastic system model can be given by the following discrete-time recurrent equations (Chen, 1983):

$$C_{k+1} = A_k C_k + B_k S_k \qquad \text{(dynamic model)} \qquad (4a)$$

$$C^*_{k+1} = H_k C_k + W_k \qquad \text{(measurement model)} \qquad (4b)$$

$$S_{k+1} = S_k \qquad \text{(constant input)} \qquad (4c)$$

$$E[W_k] = 0 \quad \text{and} \quad E[W_k W_l^T] = R_k \delta_{kl} \qquad \text{(error statistics)} \qquad (4d)$$

where subscript k is the time step indicator; A is an n x n state transition matrix; B is an n x q control transition matrix; S is the source vector of dimension q x 1; C is the concentration vector of

dimension n x 1; H is the measurement coefficient matrix of dimension m x n; W is the measurement error vector of dimension m x 1; $m \leq n$; and $n \geq q$.

The transition matrices, A_k and B_k, are computed by (Chen, 1983):

$$A_k = \exp(A'.\Delta t) = I + (A'.\Delta t) + (A'.\Delta t)^2/2! + (A'.\Delta t)^3/3! + \ldots \quad (5a)$$

$$B_k = [\exp(A'.\Delta t) - I]\,[A']^{-1}B' \quad (5b)$$

Where Δt = time step interval = $t_k - t_{k-1}$ and ! represents the factorial operation.

Solution Scheme

The next step involves application of system theory or estimation theory to develop a recursive algorithm to solve the linear dynamic system model. It may be noted that the control vector S_k contains the location and strength of the source(s) at time step k. Therefore, to identify the pollution source location and strength is to identify the deterministic control vector S_k. However, C_k is not a deterministic vector and is simulated simultaneously with S_k to determine the concentration distribution over the entire domain. The simultaneous estimation of both these parameters is achieved by adjoining C_k to S_k and defining the adjoint state vector (Z_k) as:

$$Z_k = [C_k \; S_k]^T \qquad \varepsilon \; \Omega^{n+q} \quad (6)$$

The estimation of Z_k is a state estimation problem, the solution of which can be obtained from the classical recursive estimation algorithm developed by Kalman and Bucy (1961). The significant component of the computational effort and accuracy of this solution scheme depends on solving the matrix Riccatti equation for the covariance matrix of the adjoint state estimate. Therefore, the required amount of memory and the round-off error are increased as the dimension of space coordinates becomes large. This is resolved through redefining the optimal estimate of the concentration vector as a linear process by (Panigrahi et. al., 1984):

$$c_k = \langle c_k \rangle + d_k = \langle c_k \rangle + V_k s_k \quad (7)$$

where c is the optimal estimation of the concentration C; $\langle c \rangle$ is the source-free estimate of the concentration being computed as if no source were present; d is the contribution of concentration due to the presence of source(s); s is the optimal estimate of the source S (or, source strength); and V is a matrix containing the ratio of the covariance of $\langle c \rangle$ and s to the variance of s. Physically speaking, $\langle c \rangle$ is the background concentration of a chemical species in the aquifer. The background concentration in an aquifer may or may not remain constant with time. The matrix V physically defines and determines the proportion of the source strength that is contributed towards the concentration distribution in the flow domain. This transformation process allows the computation of a source-free estimate of concentration independent of the source strength estimations, and then addition of a correction due to the

presence of sources. After a sequence of mathematical manipulations, the final solution scheme can be given by the following recursive estimation equations (Friedland, 1969):

$$\langle c_k \rangle = A_{k-1} \langle c_{k-1} \rangle = K_{ck}[C^*_k - H_k A_{k-1} \langle c_{k-1} \rangle] \quad (8)$$

$$s_k = s_{k-1} + K_{sk}[C^*_k - H_k A_{k-1} \langle c_{k-1} \rangle - Y_k s_{k-1}] \quad (9)$$

$$V_k = U_k - K_{ck} Y_k \quad \text{(auxiliary matrix)} \quad (10)$$

$$U_{k+1} = A_k V_k + B_k \quad \text{(auxiliary matrix)} \quad (11)$$

$$Y_k = H_k U_k \quad \text{(auxiliary matrix)} \quad (12)$$

$$K_{ck} = P_k H_k^T [H_k P_k H_k^T + R_k]^{-1} \quad \text{(conc. gain matrix)} \quad (13)$$

$$K_{sk} = G_{k+1} V_k^T H_k^T R_k^{-1} \quad \text{(source gain matrix)} \quad (14)$$

$$G_{k+1} = G_k - G_k Y_k^T [H_k P_k H_k^T + R_k + Y_k G_k Y_k^T]^{-1} \quad \text{(source covar. matrix)} \quad (15)$$

$$P_{k+1} = A_k [I - K_{ck} H_k] P_k A_k^T \quad \text{(conc. covar. matrix)} \quad (16)$$

where subscripts c and s refer to the concentration and source terms, respectively.

DISCUSSION OF INITIAL CONDITIONS

The gain matrix K_s is directly proportional to the source covariance matrix G and indirectly proportional to the measurement error covariance matrix R. Initial conditions are important in obtaining correct solutions from a model. In general, assigning a correct initial source covariance (G_o) is unrealistic since the source vector (s) is unknown, and the measurements obtained at selected points in the flow domain are relatively more reliable than those at the non-measurement locations. Based on this information, two functional terms are defined as follows (Panigrahi, 1985):

The reliability factor (RF) is defined as the ratio of source covariances at non-observed points to those of the observation or measured locations. The noise ratio (NR) is defined as the ratio of source variance to the measurement error covariance at a location. Assigning certain fixed values to the non-observed locations (μ_n) and measured locations (μ_m) of source covariances, the reliability factor is expressed as RF = (μ_n/μ_m). Similarly, assigning a certain value to all the measurement locations of measurement error covariance (β), the noise ratio is expressed as NR = (μ_m/β).

The initial values of RF indicate the relative uncertainty of the initial information at the non-monitoring locations. In other words, this indicates the relative accuracy of the initial source vector s_k. Thus, a significantly large RF value (several orders of magnitude) is assigned

initially indicating that the initial conditions at the non-observed locations are less reliable.

The convergence of the solution to an accurate estimate of source location and strength also depends on the initial specification of NR, since the gain matrix is directly proportional to NR. Therefore, assigning an initial NR value of greater than unity indicate that the initial source assumptions are less reliable than the measurements. This also allows sufficient flexibility for the gain matrix to reflect the transitional changes and provide stable final solutions.

APPLICATIONS

The solutions [eqs. (5a), (5b), and (7) through (16)] are applied to a hypothetical aquifer. The aquifer simulation has the following characteristics:

- a rectangular area of 1,200 ft x 600 ft (365.76 m x 182.88 m);
- average flow velocities of 4 ft/d (1.219 m/d) and 1 ft/d (0.305 m/d) in the x- and y-directions, respectively;
- average dispersion coefficients of 80 ft^2/d (7.432 m^2/d) and 10 ft^2/d (0.929 m^2/d) in the x- and y-directions, respectively;
- a constant point source with a concentration of 1,000 ppb located at the center of the aquifer;
- initial conditions of zero concentration in the aquifer;
- Neumann boundary conditions along the boundaries (concentration gradient = 0);
- spatial grid intervals of 200 ft (60.96 m) and 100 ft (30.48 m) in the x- and y-directions, respectively; and
- a simulation period of 200 days with a time interval of 5 days.

The discretization of the hypothetical aquifer is shown in Figure 1. As shown in Figure 1, the source is located at node 25 and the monitoring wells are assumed to be present downgradient at nodes 37 through 41.

Based on the above information, the model simulation was completed using the finite analytic method (Panigrahi, 1985); however, other numerical methods could be used instead. The simulated concentration distributions were perturbed by a set of random errors with a mean and standard deviation of 0 and 10^{-4}, respectively. The perturbed data at the assumed monitoring locations were treated as the measurements and the errors in them as the measurement errors. The simulated data at locations other than the monitoring points were discarded and thus created a scenario where the source location and its strength are unknown.

In order to start the algorithm, it is necessary to assign initial values to the source vector, the measurement error covariance matrix, and the source covariance matrix. The initial source vector was assumed to be zero, indicating that no information is available regarding the location and strength of the point source and that the likelihood of any grid point to be identified as a pollution source location is the same throughout the flow domain. The initial values of the covariance matrices were assigned by assuming certain values of RF and NR as described earlier. The following cases were simulated to demonstrate the application of the filter theory:

- Source Identification: RF = 10^{8} and NR = 1
- Effect of Reliability Factor: RF = 10^{-4} to 10^{17} and NR = 1
- Effect of Noise Ratio: RF = 10^{8} and NR = 10^{-10} to 10^{10}

DISCUSSION OF RESULTS

The source estimation and the sensitivity analyses are performed by solving eqs. (3a), (3b), and (5) through (14) using the initial and boundary conditions specified in the previous section for the hypothetical aquifer. For a predefined set of NR and RF, the solution scheme calculates the location and strength of the source, concentration distribution in the flow field, and the covariance matrix.

Source Estimation

Figures 2 (a) through 2(d) present the time evolution of correct source location and strength estimations at selected time steps (after 15, 20, 35, and 200 days since the simulation began). The difference between the estimated source strength (1,048 ppb) and the actual source strength (1,000 ppb) is approximately five percent. Figure 3 presents a perspective view of the estimated concentration distribution in the flow field at the end of the simulation (after 200 days), which closely agrees with the actual (simulated) concentration distribution in the flow field. The results indicate that nodal point 25 has the highest strength and the highest concentration. This infers that node 25 is the most likely pollution source location which, of course, is the actual case. The correct convergence of the solution scheme can also be verified from inspection of the final covariance matrix. A correct convergence would indicate that the final covariance at the source location is much smaller than the initial guess and often in the same order of magnitude as the measurement locations. This is illustrated in Figure 4, which presents perspective views of the initial and final source covariances.

Effect of Reliability Factor (RF)

A numerical experiment was conducted to study the effect of RF on the convergence of the solution scheme. The value of RF was varied from 10^{-4} to 10^{17} while the value of NR was maintained constant at unity (NR = 1). The experiment (RF = 10^{-4} to 10^{17}) was repeated for three sets of R or β (= 10^{0}, 10^{-4}, 10^{-8}) while maintaining NR = 1. The results are presented in Figure 5, which illustrates the effect of the initial assumptions of source covariances on the correct estimation of the source strength and

location. A review of this figure indicates that a lower limit of RF must be maintained in order to achieve an accurate and stable solution from the algorithm. The lower limit is equal to the inverse of the variance of the actual measurements, which is 10^8 for the present study. For actual field applications, a sensitivity analysis for a range of RF values with NR = 1 should be performed to obtain the lower limit of RF, which in turn will indicate the potential standard deviation of the measurements. Furthermore, RF is fairly robust with respect to the time step size (time step was varied from one day to five days), the source location, and the noise ratio. For brevity, these results are not repeated in this report.

Effect of Noise Ratio (NR)

A second set of numerical experiment was conducted to study the effect of NR on the correct estimation of source strength. Maintaining RF as a constant at 10^8, the value of NR was varied from 10^{-10} to 10^{10}. The results are presented in Figure 6. A review of this figure indicates that convergence of the solution scheme to the correct estimate of source strength is unaffected by values of $NR \geq 1$. A value of NR significantly less than unity still results in a converged solution, but yields to an incorrect estimate of source strength. The significance of NR is directly related to build-up of the gain matrix K_s. For cases with a small initial assumption of NR, the gain matrix K_s does not accumulate sufficient gains for proper convergence of the solution scheme; and thus, the solutions deviate from correct source estimates. Physically speaking, this indicates that the initial source assumptions are less reliable than the measurements.

SUMMARY AND CONCLUSIONS

The location and strength of a pollution source in a two-dimensional uniform ground water flow field are identified. The concentration distribution in the flow field and the associated error covariances of the estimation process are also quantified. A numerical sequential system model is developed to solve the governing stochastic partial differential equation via Friedland's approach and the extended Kalman filter. The model implementation is demonstrated with a test case for a hypothetical aquifer. The reliability factor (RF) and the noise ratio (NR) are used to define relative initial conditions of the error covariances. RF is fairly robust with respect to the time step size, the source location, and the noise ratio. Convergence of the solution scheme to the correct estimate of source strength is unaffected for $NR \geq 1$.

ACKNOWLEDGMENT

The research leading to this report was supported by the Department of Civil Engineering at Drexel University, Philadelphia. The results and opinions presented in this report are solely the views of the authors.

REFERENCES

Chen, C.T., 1983. Linear System Theory and Design, CBS College Publishing, New York.

Distefano, N. and A. Rath, 1982. An Identification Approach to Subsurface Hydrological Systems, Water Resources Research, Vol. 18, No. 3, pp. 597-662.

Friedland, B., 1969. Treatment of Bias in Recursive Filtering, IEEE Transactions on Automatic Control, Vol. AC-14, No. 4, August 1969, pp. 359-367.

Gorelick, S.M., 1983. A Review of Distributed Parameter Ground Water Management Modeling Methods, Water Resources Research, Vol. 19, No. 2, pp. 305-319.

Gorelick, S.M., B. Evans and I. Remson, 1983. Identifying Sources of Ground Water Pollution: An Optimization Approach, Water Resources Research, Vol. 19, No. 3, pp. 779-790.

Hwang, J.C. and R.M. Koerner, 1983. Ground Water Pollution Source Identification from Limited Monitoring Well Data: Part 1 - Theory and Feasibility, Journal of Hazardous Materials, 8, pp. 105-119.

McLaughlin, D.B., (Editor - C.L. Chiu), 1978. Potential Applications of Kalman Filtering Concepts to Ground Water Basin Management, Proc. of AGU Chapman Conference, Pittsburgh, USA, May 22-24, 1978.

Kalman, R.E. and R.S. Bucy, 1961. New Results in Linear Filtering and Prediction Theory, Trans. ASME, Jour. Basic Engrg, Ser. D, Vol. 83, pp. 95-108, March 1961.

Panigrahi, B.K., 1985. A Stochastic Approach to Estimate the Locations and Strengths of Ground Water Pollution Sources, Ph.D. Thesis, Drexel University.

Panigrahi, B.K., J.C. Hwang, and A. Yousuff, 1984. Identification of Locations and Strengths of Ground Water Pollution Sources Using a Numerical Sequential Model with Kalman Filter, Proceed. Int. AMSE Conf. "Modeling & Simulation", Minneapolis, USA, August 13-17, 1984, Vol. 4, pp. 59-70.

Roux, R.H. and W.F. Althoff, 1980. Investigation of Organic Contamination of Ground Water in South Brunswick Township, New Jersey, Ground Water, Vol. 18, No. 5.