IDENTIFIABILITY VERSUS HETEROGENEITY IN GROUNDWATER MODELING SYSTEMS

Received 15/12/2001– Accepted 05/03/2003

Abstract

Review of history matching of reservoirs parameters in groundwater flow raises the problem of identifiability of aquifer systems. Lack of identifiability means that there exists parameters to which the heads are insensitive. From the guidelines of the study of the homogeneous case, we inspect the identifiability of the distributed transmissivity field of heterogeneous groundwater aquifers. These are derived from multiple realizations of a random function $Y = \log T$ whose probability distribution function is normal.

We follow the identifiability of the autocorrelated block transmissivities through the measure of the sensitivity of the local derivatives $D_T \boldsymbol{h} = (\partial h_i / \partial T_j)$ computed for each sample of a population $N(0; \sigma_Y, \alpha_Y)$. Results obtained from an analysis of Monte Carlo type suggest that the more a system is heterogeneous, the less it is identifiable.

***Keywords**: groundwater modeling, inverse problem, identification, optimisation, heterogeneity.*

Résumé

Le calage des paramètres des modèles numériques de nappes pose le problème de l'identifiabilité des systèmes aquifères. Le défaut d'identifiabilité signifie qu'il existe des paramètres dont les variations n'affectent pas sensiblement les charges hydrauliques h.

Des enseignements tirés de l'étude du cas homogène, l'identifiabilité des systèmes aquifères hétérogènes est envisagée en générant des champs de transmissivités T à partir de réalisations de la fonction aléatoire $Y = \log T$.

Cette question est traitée sur des exemples synthétiques en utilisant une analyse de type Monte Carlo. L'identifiabilité des transmissivités des blocs aquifères auto-corrélés est inspectée en évaluant la sensitivité des dérivées $D_T \boldsymbol{h} = (\partial h_i / \partial T_j)$ pour chaque échantillon extrait de la population $N(0; \sigma_Y, \alpha_Y)$. Nous établissons que, de fait, un système est d'autant moins identifiable qu'il est hétérogène.

***Mots clés**: hydrogéologie, problème inverse, identification, optimisation, hétérogénéité.*

A.M. BENALI
Laboratoire Eau & Environnement
Université d'Oran
BP 1524, Es-Sénia, Algérie

L.W. de BACKER
Unité de Génie Rural
Université de Louvain-la-Neuve
B1348, Belgique

ملخص

إن توابث النماذج الرقمية للغطاءات المائية يطرح مشكل تعيين وتعريف الأنظمة المائية. والخلل في ذلك يدل على وجود عناصر ذات تغيرات لا تمس الحمولات المائية h.

إن دلالات أستنتجت من خلال دراسة حالة متجانسة أثبتت أنه يمكن التعرف عى الأنظمة المائية عن طريق خلق حقول منقولية T ومن خلال إنشاء وتحقيق دالة آنية Y = log T.

هذه المسألة عولجت بأمثلة تركيبية وطلك باستعمال تحليل على شاكلة مونت اكارلو. إن تعيين منقوليات المجسمات المائية المقارنة ذاتيا فتشت بتقييم محسوسية المشتقات $D_T \boldsymbol{h} = (\partial h_i / \partial T_j)$ لكل عينة من Y مستخرجة من التوزيع الأم (السكان الأصلي) $N(0; \sigma_Y, \alpha_Y)$.

يمكننا إذن ترسيخ أن نظاما ما هو أقل تتعرفا لكونه غير متجانس.

***الكلمات المفتاحية:** هيدروجيولوجيا، مشكل عكسي، تعيين، الأفضلي، عدم تجانس.*

The increasing development of computer groundwater modeling still have to face to the problem of estimating the aquifer parameters. This decisive stage is still conducted in the traditional manual way of trial -and - error procedure. Although very instructive, this approach is large time consuming. Several attempts to perform automatically the model calibration, i.e. history matching, actually failed because of ill posedness. Behind this question point out insensitivity, nonunicity and instability.

Insensitivity means that we do not provide enough information to support estimation of the parameters. Nonuniqueness appears when match to the observations may be realized with several combinations of parameters. Instability indicates that slight changes in parameters values sharply affect model outputs.

Now, however, a large body of the works on the subject defines the conditions under which identifiability of aquifer systems is possible.

In this context, Kitamura [1], Chavent [2] and Carrera *et al.*, [3] have pointed out some basic guidelines where identifiability of parameters can be realized. They confirmed the orientations of the previous works seeking for preliminary data, setting upper and lower bounds, attempting to reduce errors of measure and trying judicious variable transformations to moderate non linearities.

Actual modeling packages take advantage of these features to

propose successful yet powerful inverse codes [4,5] which agree the task mentioned in the preface of [6]. Moreover, recent trends in the discipline reveal a tendency to get out the single valued predictions and, further, to produce confidence bounds. The previous evolution let the parameters not transgressing their allowed domains as dictated by the ad hoc field through the definition of its probability distribution function (pdf) and the characteristics of its spatial variability.

Concerning the log - transmissivity field Y, its pdf is normally distributed with a mean m_Y and a standard deviation σ_Y. Its spatial variability is described through a covariance function expressing the way values of Y conciliate. The autocorrelation parameter α_Y is a characteristic of this covariance function whereas Y stands up for a well-known transformation variable. Henceforth, boldface will refer to matrices.

Thus, our point will be this: According to the definition, now largely agreed, that lack of identifiability means that there exists some parameters to which the heads $\boldsymbol{h} = (h_i)$ are insensitive, we look into this insensitivity with respect to the heterogeneity field of the transmissivities $\boldsymbol{T} = (T_i)$. Through the measure of the local derivatives $D_{\boldsymbol{T}}\boldsymbol{h} = (\partial h_i / \partial T_j)$, realized on multiple replications of a single field log T, we follow the evolution of $D_{\boldsymbol{T}}\boldsymbol{h}$ for different values of the parameters σ_Y and α_Y of the heterogeneous field. These realizations of the random function Y are obtained from a generator of auto-correlated log-normal values of transmissivities. These are extracted from the population N (0; σ_Y, α_Y). The simulations performed on these synthetic realizations are of Monte Carlo type. They provide both a measure of the sensitivity of $\boldsymbol{h}$ to $\boldsymbol{T}$ and of $\boldsymbol{T}$ to $\boldsymbol{h}$ parameterized on σ_Y and α_Y. As one cannot intuitively guess it, identifiability of aquifer systems is not significantly affected by the degree of variability of Y. Only the estimates are perturbated.

HISTORY MATCHING

Modeling a groundwater system is indeed a three stages process where (1) characterization points out the right model, (2) history matching estimates parameters and (3) simulation performs prediction on the heads evolution. Practically, we do not deal with such a linear scheme. In fact, stages 2 and 3 alternate in a cyclic way up to conform the estimates to the data sets.

Automatic history matching by linear estimators seek to minimize a performance criterion in view to estimate the flow field parameters. The most often used belongs either to the class of the least squares or to the bayesian and the maximum likelihood estimators. Let us examine the principle of the more elementary of them: The least squares (LS) estimator.

An LS estimator seeks for a distributed values, say $\{T_i\}$ to minimise a performance criterion Ψ :

$$Min_{T} \Psi = \left\| h(T) - h^* \right\|^2 \qquad (1)$$

$\boldsymbol{h}^*$ are the m measured heads and $\boldsymbol{h}$ the computed head values obtained by solving the analogue form :

$$\boldsymbol{B}(\boldsymbol{T})\,\boldsymbol{h} = \boldsymbol{q} \qquad (2)$$

of the equation which describes the 2-D steady state groundwater flow in the unit square domain :

$$\text{div}\,(T\,\text{grad}\,h) = q \qquad (3a)$$

subjected to the boundary conditions :

$$\begin{aligned} h(x,0) &= h_N \\ h(x,1) &= h_S \\ h(0,y) &= h(1,y) = h_L \end{aligned} \qquad (3b)$$

$\boldsymbol{B}$ is the flow matrix, $\boldsymbol{q}$ a column matrix including the source terms q and the boundary contributions whereas $T = T(x, y)$ is the parameter of the field namely known as the transmissivity and $h = h(x, y)$ is the usual hydraulic head staying for the unknown function of the partial differential equation (PDE).

Although estimating the heads from equation (2) is not a linear problem in T, we are still using linear estimators. We only need to take care to linearize h according to some classical techniques as the Gauss-Newton -Kantorovich one [6] :

$$\boldsymbol{h} = \boldsymbol{h}(\check{\boldsymbol{T}}) + D_T\,\boldsymbol{h}(\boldsymbol{T}\text{-}\check{\boldsymbol{T}}) + \boldsymbol{\varepsilon}(\boldsymbol{\Delta}^2) \qquad (4)$$

The last term indicates that the local truncation error for this formula is in the order of $\boldsymbol{\varepsilon}(\boldsymbol{\Delta}^2)$. This term is not used directly in the application of the formula but only as an indicator of the accuracy of the approximation. Then, the performance criterion of equation (1) can be written near an estimate $\check{\boldsymbol{T}}$ of $\boldsymbol{T}$:

$$\Psi(\boldsymbol{T}) = \left\| \boldsymbol{h} - [\,\boldsymbol{h}(\check{\boldsymbol{T}}) + D_T\,\boldsymbol{h}(\boldsymbol{T}\text{-}\check{\boldsymbol{T}})\,] \right\|^2 \qquad (5)$$

The classical solution of which is known. It is given by :

$$\hat{\boldsymbol{T}} = [\,(D_{\boldsymbol{T}}\boldsymbol{h})^t\,D_{\boldsymbol{T}}\boldsymbol{h}\,]^{-1}\,(D_{\boldsymbol{T}}\boldsymbol{h})^t\,[\,\boldsymbol{h} - \boldsymbol{h}(\check{\boldsymbol{T}})\,] + \check{\boldsymbol{T}} \qquad (6)$$

where $D_{\boldsymbol{T}}\boldsymbol{h} = (\partial h_i / \partial T_j)$ is the local derivative of $\boldsymbol{h}$ with respect to $\boldsymbol{T}$. It stands for the sensitivity of $\boldsymbol{h}$ to perturbations on $\boldsymbol{T}$. The superscript "t" means matrix transpose operation. The linearized equations (6) are often presented under the normal form :

$$\boldsymbol{u} = [\,(D_{\boldsymbol{T}}\boldsymbol{h})^t\,D_{\boldsymbol{T}}\boldsymbol{h}\,]^{-1}\,(D_{\boldsymbol{T}}\boldsymbol{h})^t\,[\,\boldsymbol{h} - \boldsymbol{h}(\check{\boldsymbol{T}})\,] \qquad (7)$$

where $\boldsymbol{u} = \hat{\boldsymbol{T}} - \check{\boldsymbol{T}}$ is an upgraded parameter which can be achieved through an iterative process via the Gauss-Marquard-Levenberg's method.

SENSITIVITY OF THE LEAST SQUARES ESTIMATOR

The solution given by equation (4) is an estimation performed with the LS estimator. It is the best linear unbiaised estimator of $\boldsymbol{T}$ appearing in equation (1). If σ_h^2 stands for the variance of the actual observations $\boldsymbol{h}^*$ then it can be evaluated by :

$$\sigma_h^2 = \frac{\Psi}{m-n}$$

where (m-n) indicates the difference between the number of observations and the number of parameters to be estimated i.e the degree of freedom.

To examine the sensitivity of LS to uncertainty on the head $\boldsymbol{h}$, we may differentiate equation (4). Then, we obtain :

$$D_{\boldsymbol{h}}\,\hat{\boldsymbol{T}} = [\,(D_{\boldsymbol{T}}\boldsymbol{h})^t\,(D_{\boldsymbol{T}}\boldsymbol{h})\,]^{-1}\,(D_{\boldsymbol{T}}\boldsymbol{h})^t \qquad (8)$$

It is further useful to compute its covariance which is expressed by :

$$\boldsymbol{C}_{\boldsymbol{T}} = \sigma_h^2\,(D_{\boldsymbol{h}}\,\hat{\boldsymbol{T}})\,(D_{\boldsymbol{h}}\,\hat{\boldsymbol{T}})^t \qquad (9)$$

According to the previous development concerning the local derivatives of $\boldsymbol{T}$ i.e. $D_{\boldsymbol{h}}\,\boldsymbol{T} = (\partial h_i / \partial T_j)$, we should notice that the elements of this sensitivity matrix are local sensitivity coefficients which depend on the value of $\boldsymbol{T}$. They can be used only in the neighbourhood of the optimum of the LS estimator. Accuracy in derivatives computation is fundamental to the success of the optimisation process.

One can get a fast idea about the quality of these estimates either through the examen of the correlation matrix $\boldsymbol{\rho} = [\ \sigma_{ij} / (\sigma_{ii}\ \sigma_{jj})^{1/2}\ ]$ or the diagonal of $\boldsymbol{C_T}$. A quick appreciation of the global variance of these estimates is given by the trace of the covariance matrix $\boldsymbol{C_T}$. In this case, we can get evaluation of Tr($\boldsymbol{C_T}$) through the sum of the eigenvalues λ_i of $\boldsymbol{C_T}$:

$$\mathrm{Tr}\,(\boldsymbol{C_T}) \;=\; \sigma_h^2 \sum_i \lambda_i^{-1} \qquad (10)$$

in order to evaluate the accuracy of estimates $\check{\boldsymbol{T}}$.

IDENTIFIABILITY

Mathematics of history matching is still plagued by problems of identifiability [7]. This notion due to Kubrusly [8] is strongly connected to the class of inverse problems whose solutions are instable. Ill posedness arises from their direct formulation which does not account for the non-satisfaction of the so-called Lipschitz condition [3,9]. Many attempts to stabilize these solutions have been tried in view to get well posed problems. Following Kitamura *et al*. [1], history matching is an inverse problem studied in the framework of the theory of partial differential equations. The case where identifiability imply uniqueness requires exact values of h and drastic limitations on both the boundary conditions and the parameters .

The situation is somewhat different in [2] and more crudely in [3] since incorporating prior information on the parameters allow one to deal with positive definite Hessian matrix i.e. a sufficient condition for uniqueness which has to be confirmed via sensivity analysis.

So indeed, identifiability allowed to circumscribe the conditions to be verified to obtain either continuous stable solutions in the LS sense as in [3] or to assure unicity as in [1].

So far, identifiability actually temperate the scepticism due to the sensitivity analysis concerning history matching.

Such is the framework in which we will analyse the sensitivity of the LS estimator and the sensitivity of its estimates to perturbations.

The next section documents on Monte Carlo technique used to enhance this analysis.

MONTE CARLO SIMULATIONS

According to the Monte Carlo approach, let R_{ij} a realization of a porous medium $M_{ij}\,(\omega)$ which stands for an underlying stochastic process.

Given a set Ω of elementary events ω, to M_j is associated a random function $Y = \log T$:

$$Y_j \;=\; Y_j\ (x,y;\,\omega), \quad \omega \in \Omega \qquad (11)$$

and a matrix equation :

$$\boldsymbol{B}\,(\omega)\,\boldsymbol{h} \;=\; \boldsymbol{q} \qquad (12)$$

which is the discrete stochastic analogue of (2).

In a similar way, to each R_{ij} is associated a realization ;

$$Y_{ij} \;=\; Y_j\,(\,x,y\,;\,\omega_i) = \;Y_j(x,y)$$

and a discrete deterministic analog :

$$\boldsymbol{B}\,\boldsymbol{h} \;=\; \boldsymbol{q} \qquad (13)$$

Thus far according to (11) and (12) repeating realizations of the random function Y_j will allow the definition of a family of flow matrix $\boldsymbol{B}$:

$$\boldsymbol{\beta} \;=\; \{\ \boldsymbol{B}\,(\omega_i)\ \} \qquad i = 1,2,\ \ldots$$

Monte Carlo simulations operate successively on equation (13) with $\boldsymbol{B}$ of the set $\boldsymbol{\beta}$. Through this way we are able to produce the first two moments of the solution $\boldsymbol{h}$, the jacobians $(\partial h_i / \partial T_j)$ and $(\partial T_i / \partial h_j)$. This process can be repeated for each porous medium M_j i.e. for each random function Y_j .

NUMERICAL EXPERIMENTS

The model we simulate is described by a dimensionless form of equation (2) subject to conditions (3). The presence of a pumping well in the center of the flow domain (Fig. 1) allows us to consider the problem as well posed in that the distributed parameters T of the aquifer system are identifiable according to [1]. Inputs are produced by an adequate numerical generator of autocorrelated log T values [10]. Each trial is performed with 30 realizations.

1	4	7
2	5	8
3	6	9

Figure 1: Zonation of the flow domain.

$H = h_N = 0$

	1	2	3	4	5	6	7	8
	9							16
	17							24
$h = h_L$	25							32
	33							40
	41							48
	49							56
	57	58	59	60	61	62	63	64

$h = h_S = 0$

Figure 2: Discretization of the flow domain.

Influence of the variability on the jacobian

The starting point is the evaluation of the jacobian $(\partial h_i / \partial T_j)$ in the context of different heterogeneity fields.

Firstly, the results obtained do confirm that uncertainty on head increases as long as the standard deviation σ_Y and the autocorrelation parameter α_Y increase.

Secondly, perturbating the log$T$ field do not affect significantly the flow field. The values of the elements of the jacobian are at most of the order of unity. These results suggest an insensivity of h to perturbations of T. Yakowitz *et al*. [9] and McElwee [11] have already reported them.

In figure 1, we depict the zonation of the flow domain adopted to reduce the magnitude of the components of the jacobian.

On table 1, the values $\partial h_i / \partial T_j$ of the jacobian are given. Only the elements of the greatest module are reported for each j. These computations are repeated for all the thirty realizations. This approach of a Monte Carlo type provides values of $\partial h_i / \partial T_j$ close to three.

Thirdly, this insensitivity persists whatever the values of the parameter α_Y as we can see it in table 2.

0.21	0.20	0.21
0.20	0.17	0.23
0.24	0.32	0.23

Table 1: Greatest modulus components of the sensitivity matrix extracted from N(0; 0.3; 0.5) for i =1, $1 \le j \le 9$.

$(\sigma_Y ; \alpha_Y)$	ZONES								
	1	2	3	4	5	6	7	8	9
(0. ; 0.)	0.4	0.4	0.4	1.2	0.8	1.2	0.4	1.8	0.4
(0.1;0.5)	2.9	2.7	2.7	8.9	7.3	7.8	7.5	16.7	12.0
(0.2;0.5)	3.0	2.6	3.3	8.9	8.4	8.7	19.7	19.0	12.9
(0.3;0.5)	3.2	2.8	4.2	9.0	9.7	10.5	32.9	24.2	13.8

Table 2: Jacobian sup $\partial h_i / \partial T_j$ (x10) for different heterogeneous log T fields ($1 \le i \le 36$; $1 \le j \le 9$) averaged on 36 realizations.

Influence of the variability on LS estimates

We present the results obtained in the same fashion as we did for the sensitivity matrix $D_T \boldsymbol{h}$.

In table 3, the greatest modulus of the derivatives $\partial T_i / \partial h_j$ are depicted for each row i whereas in table 4 their mean values are reported.

9.09	7.92	7.74
8.00	10.10	5.08
38.90	64.80	29.90

Table 3: Greatest modulus components of the sensitivity matrix D_h T extracted from N(0; 0.3; 0.5) for i =1, $1 \le j \le 9$.

$(\sigma_Y ; \alpha_Y)$	ZONES								
	1	2	3	4	5	6	7	8	9
(0. ; 0.)	7.5	7.2	6.7	13.0	3.5	12.0	30.0	2.4	29.0
(0.1;0.5)	26.8	18.3	20.1	22.6	12.8	8.5	41.7	11.3	23.8
(0.2;0.5)	14.	13.	15.	6.	10.	7.	38.	7.	23.
(0.3;0.5)	15.	13.	14.	8.	10.	7.	39.	5.	26.

Table 4: Jacobian sup $\partial h_j / \partial T_i$ for different heterogeneous log T fields ($1 \le i \le 36$; $1 \le j \le 9$) averaged on 36 realizations.

The usual interpretation of these sensitivity coefficients is to consider them as errors on the estimates due to LS and following an error of unity on the heads. The order of these values let us think that LS is indeed a rough estimator at present. It is so because of the nonlinear character of $\boldsymbol{h}$ with respect to $\boldsymbol{T}$ in equation (2). The attempt to linearize it seems to fail when increasing σ_Y. We may confirm this untractable non-linearity through the examination of the covariance matrix $\boldsymbol{C_T}$ reported in table 4. The deviation of the estimates through the valued trace of $\boldsymbol{C_T}$ for different log T fields is consigned in table 5.

1.50								
0.16	0.76							
0.64	0.12	0.99						
1.50	0.57	0.95	3.30					
0.90	0.39	0.68	1.30	0.90				
0.41	0.17	0.34	0.68	0.37	0.31			
1.90	0.68	1.10	4.50	1.50	0.81	6.60		
0.65	0.31	0.47	1.00	0.62	0.30	1.20	0.49	
0.50	0.23	0.83	0.98	0.40	0.70	1.30	0.29	3.20

Table 5: Values of the covariance matrix (x10^3), N(0;0.3;0.5).

HETEREGENEOUS FIELDS INDEXED ON $(\sigma_Y ; \alpha_Y)$			
(0.0 ; 0.0)	(0.1 ; 0.5)	(0.2 ; 0.5)	(0.3 ; 0.5)
0.38	13.73	13.26	18.05

Table 6: Values of the trace of the covariance matrix (x10^4).

So all these observations do confirm that LS estimation of T is highly uncertain. It would require a precision on $\boldsymbol{h}$ values far from the common ones. According to [8] incorporation of prior information in the expression of Ψ improves scarcely the performance of LS. Furthermore, the spectrum of the eigenvalues of $\boldsymbol{C_T}$ appears as more extended as the system is heterogeneous. This is a result which is positively correlated with the one concerning the eigenvalues of the flow matrix [12]. Some elements of the spectrum of the covariance matrix are even associated with directions of relative insensitivity [13].

CONCLUSIONS

Automatic history matching with LS is assurely an hypothetic estimator of the distributed values of the log T field. Incorporating prior information, indeed, could improve it; although a choice of a more performing estimator is highly suggested. Moreover, sensitivities of the heads are not particularly affected by the magnitude of the variability of the log T field. They only affect the estimates once the propense instability of the identification problem avoided [9,11].

REFERENCES

[1]- Kitamura S. and Nagakiri S., "Identifiability of spatially varying and constant parameters in Distributed Systems of Parabolic type", SIAM J. Contr. Optimiz., 15 (5), (1977).

[2]- Chaven G., Dupuy M. and Lemmonier P., "History matching by use of optimal theory", *Soc of Petrol. Eng. Journal*, (1975), pp. 74-86.

[3]- Carrera J.C., Neuman S.P., "Estimation of aquifers parameters under transient and steady state conditions", Water res. research, (1989).

[4]- Hill M.C., "Methods and guidelines for effective model calibration with application to Ucode and Modflowp", US Geological Water Survey, Water Resources Investigations, report 98-4005, Denver-Colorado, (1998), p. 90.

[5]- Doherty J., Brebber L. and White L., "PEST a model independent parameter estimation", Watermark computing, (1994), p. 146.

[6]- Angel E. and Bellman R., "Dynamic programming and partial differential equation", Vol.88, in Mathematics in Science and Engineering, Academic press, New York, (1972), p. 202.

[7]- Tarantola A., "Inverse problem theory Methods for data fitting and model parameter estimation", Elsevier, New York, (1987), p. 613.

[8]- Kubrusly C.S., "Distributed parameter system : a survey", *International Journal of Control*, 26 (4), (1977), pp. 509-535.

[9]- Yakowitz S. and Duckstein L., "Instability in aquifer parameter identification : Theory and case studies", Water Res. Research, 16 (6), (1980), pp. 1045-1064.

[10]- Mantoglu A. and Wilson J.L., "Simulation of random fields with the turning bands method", report n° 264, Ralph M. Parsons Lab., dept of civil eng., MIT, (1981).

[11]- McElwee C.D., "Sensitivity analysis and the groundwater inverse problems", Groundwater 20 (6), (1982), pp. 723-735.

[12]- Benali A.M., "Correlation links between the diagonal elements of the flow matrix : Preliminary results", MNEM'95, V^{ème} Colloque Maghrébin sur les Méthodes Numériques de l'Ingénieur, Rabat 21-23 nov., (1995).

[13]- Shah P.C, Gavalas G.R. and Seinfeld J.H., 'Errors analysis in history matching. The optimum level of parametrization", *Soc of Pet. Eng. Jour*., (1978), pp. 219-228. □