Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2018.22271

125 (2018) 193–197
September

Analysis and discussion of groundwater pollution source based on iterative local update set smoothing algorithm

Xinhao Ji[a,*], Lei Qin[b]

[a]*Informatization Office, Zhejiang Business College, Hangzhou, China, email: jxh@zjbc.edu.cn (X. Ji)*
[b]*School of Economics and Management, Zhejiang Business College, Hangzhou, China, email: qlei@zjbc.edu.cn (L. Qin)*

Received 7 March 2018; Accepted 26 March 2018

ABSTRACT

To better manage groundwater and evaluate environmental risks of groundwater pollution, the fate of pollutants was accurately predicted, with the help of the numerical simulation. A set of iterative updating local smoothing algorithm was put forward. In the process of implementing the algorithms, each sample in the set was not directly updated, but local sample set of each sample was updated to fully explore the possible multi-peak distribution. In order to verify the effectiveness of ILUES algorithm, five numerical examples were verified, taking into account different prior parameters, such as parameter prior multi modal, parameter posterior multi modal and parameter high dimension. These example results showed the effect of the ILUES algorithm in the parameter inversion of the complex model. To sum up, compared with the common MCMC algorithm, the ILUES algorithm has a significant advantage in computational complexity.

Keywords: Groundwater pollution; ILUES; MCMC

1. Introduction

Compared with surface water, groundwater has the advantages of stable water quantity, good water quality, low seasonal variability and even distribution. Therefore, groundwater is regarded as one of the most important water sources in human production and life. However, due to the influence of human activities, the groundwater is suffering from excessive exploitation and water pollution. In particular, the increasingly serious groundwater pollution poses a very serious threat to human health [1].

In order to better manage groundwater and evaluate the environmental risk of groundwater pollution, we need to use numerical models to analyse and predict the fate of pollutants in underground water. However, due to the difficulty in direct observation of groundwater system, people can only get sparse, indirect and error observation data, which brings great challenges to the accurate description of the whole groundwater system. Based on the understanding of physical, chemical and biological principle, we use mathematical methods and the groundwater model to describe the specific groundwater system, so as to obtain the quantitative causality of groundwater system. We also apply observation data to reduce the uncertainty of underground water system, so as to provide important technical support for effective groundwater resources management [2].

At present, the most widely used standard model in the simulation of saturated groundwater movement is MOD FLOW. It is a three-dimensional groundwater horizontal type developed by the United States Geological Survey (USGS) based on the finite difference method. The earliest version of MOD FLOW appeared in 1984, and there are now published versions of MOD FLOW-88, MOD FLOW-96, MOD FLOW-2000, and MODFLOW-2005. Although the original idea of MOD FLOW is only to simulate the water movement of groundwater, the modular structure of MOD FLOW provides a powerful framework for the simulation of other processes. Nowadays, MOD FLOW series and related program already has been able to achieve the simulation of the groundwater / surface water system (GS FLOW), particle tracking (MOD PATH), solute transport (MT3 DMS), variable density and non-saturated

*Corresponding author.

zone groundwater flow (SEA WAT), aquifer contraction and ground subsidence (SUB), groundwater model parameter estimation (MOD FLOWP), zone water balance (ZONE BUDGET) and groundwater management (GWM) [3,4]. In addition, in order to improve the simulation accuracy, MOD FLOW-LGR can also be used to locally refine the grid of water movement model of groundwater. In the unsaturated groundwater simulation, HYDRUS model developed by Simunek and others has been widely used. It can be used to simulate the migration and transformation of water, heat and solute in one-dimensional to three-dimensional unsaturated groundwater, and can be coupled with MOD FLOW to study the interaction between the vadose zone and saturated underground water. In order to facilitate construction and display the results of the above model, some people developed some auxiliary tools and friendly user interface (including some commercial software). For instance, Model Muse and Visual MOD FLOW can provide a friendly user interface of models like MOD FLOW, MOD FLOW-LGR, MT3DMS, MOD PATH and ZONE BUDGET l; Flopy can create and operate MOD FLOW related models with the help of Python language, as well as provide post-processing toolkit [5,6].

At present, in the saturated groundwater solute migration, the most widely used model was MT3DMS developed by Zheng Chunmiao and so on. It can be coupled with MOD FLOW, used to simulate transport and transformation of various pollutants, including convection, diffusion, source mixing and chemical reaction process. The first version of MT3DMS was MT3DPS developed in 1990. Compared with earlier versions, MT3DMS can simulate multi-group component biological and geochemical reactions, and has higher numerical accuracy and stability. It can also simulate non-equilibrium adsorption and two-zone convection dispersion [7].

The groundwater model based on finite element can deal with more complex flow fields and boundary conditions, but it has more complex solution process and slower solution speed. In the past few decades, significant increase in computing power makes it possible to construct a more complex model. The development of sensor technology makes the measured data obtained more easily and the deepening understanding of groundwater system various processes makes the construction of conceptual model more accurate [8].

2. Iterative local update set smoothing algorithm

For the convenience of illustration, we use the following formula to represent an arbitrary physical model:

$$d = f(m) + \varepsilon \tag{1}$$

In the above formula, d is a vector of N_d*1, which represents the observed value. $f()$ is a system model, m is a vector of N_m * 1, which shows the unknown model parameter, and ε is a vector of N_d * 1, indicating the observation error. According to the observational value d with error, we can use ES to update the unknown parameters:

$$m_j^a = m_j^f + C_{MD}^f \left(C_{DD}^f + C_D\right)^{-1}\left[d_j - f\left(m_j^f\right)\right] \tag{2}$$

In the above formula, $M^f = \left[m_1^f, \ldots, \mathrm{m}_{N_e}^f\right]$ is a set consisting of N_e prior parameters samples; $M^a = \left[m_1^a, \ldots, \mathrm{m}_{N_e}^a\right]$ is the sample set after updating; $D^f = \left[f\left(m_1^f\right), \ldots, f\left(m_{N_e}^f\right)\right]$ refers to the covariance matrix whose dimension is N_m * N_d; C_{DD}^f is the D^f covariance matrix whose dimension is N_d * N_d [9]; C_D is the observation error covariance matrix whose dimension is the same as that of C_{DD}^f; d_j means the sample $d_j = d + \varepsilon_j$ after the measured value is added with random disturbance.

If the prior or posterior of the parameter are multi modal, the error results will be obtained by using the formula (2) to update the parameters directly. In order to accurately explore the # peak distribution, we do not update the N_e samples directly, instead, we update the local set of the N_e samples. For the sample $M_j^f\left(j = 1, \ldots N_e\right)$, we define its local set by the following index:

$$J(m) = J_1(\mathrm{m})/J_1^{\max} + J_2(\mathrm{m})/J_2^{\max} \tag{3}$$

$J_1(\mathrm{m}) = \left[f(m) - d\right]^T C_D^{-1}\left[f(m) - d\right]$ is the distance between the model output $f(m)$ and observed value d; $J_2(\mathrm{m}) = \left(m - m_j^f\right)^T C_{MM}^{-1}\left[m - m_j^f\right]$ suggests the distance between the model parameter m and sample m_j^f; C_{mm} indicates the covariance matrix of the model parameter, with dimension of $N_m \times N_m$; $J_1^{\max}$ and $J_2^{\max}$ are the maximum values of $J_1(m)$ and $J_2(m)$. Then, the local set of sample m_j^f is $N_l = \alpha N_e\left(\alpha \in [0,1]\right)$ samples with the minimum J value, namely $M_j^{l,f} = \left[m_{j,1}^f, \ldots, m_{j,N_l}^f\right]$. Next, we can update the samples in the local set:

$$m_{j,i}^a = m_{j,i}^f + C_{MD}^{l,f}\left(C_{DD}^{l,f} + C_D\right)^{-1}\left[d_j - f\left(m_{j,i}^f\right)\right] \tag{4}$$

In (4), $i = 1, \ldots, N_l$. Here $C_{MD}^{l,f}$ is the cross-covariance matrix of $M_j^{l,f}$ and $D_j^{l,f} = \left[f\left(m_{j,1}^f\right), \ldots, f\left(m_{j,\mathrm{N_l}}^f\right)\right]$ and its dimension is $N_m \times N_d$; $C_{DD}^{l,f}$ indicates the self-covariance matrix of $D_j^{l,f}$ and its dimension is $N_d \times N_d$; $d_i = d + \varepsilon_i$ suggests the sample after the observed value is added with random interference. From the local set $M_j^{l,\mathrm{a}} = \left[m_{j,1}^a, \ldots, m_{j,\mathrm{N_l}}^a\right]$ after updating, we can randomly select a sample $M_j^{l,\mathrm{a}}$ recorded as the updating sample of $M_j^f\left(j = 1, \ldots, N_e\right)$. In this way, we can obtain the updated overall set $M^a = \left[m_1^{l,a}, \ldots, m_{\mathrm{N_e}}^{l,a}\right]$.

For nonlinear strong problems, we use a simple iterative process, that is, in each iteration step, we make $M^f = M^a$, and re-implement the above local update algorithm and get a new updated set. We stop the iterative process when the difference between the sample sets of the two adjacent iterations is small enough, or when the maximum number of iterations is reached. Similar iterative processes have been used in the set Kalman filter and set smoother to solve nonlinear and strong parameter problems.

3. Results and discussion

3.1. Case 1

In iterated local update ensemble smoothing (ILUES) algorithm, from the updated local sample set, we randomly choose a sample $M_j^{l,a}$, used as the update sample of M_j^f, which is called random selection. However, it seems to be a good choice for choosing the sample with the smallest J value (formula 3) from the updated local sample set as an update sample of M_j^f, which is called "optimal choice". From Fig. 1, we can see that the model fitting result of "optimal choice" is often better than that of "random selection", and the y axis represents the Log RMSE value between the final model set output corresponding to the sample set and the observed value. However, in some cases, the "optimal choice" may lead to a large deviation in the parameter inversion result. We will explain it in the next example. From this figure, we can also see that the selection of smaller α values can usually get better fitting results. However, when the N_e value is relatively small, the smaller α value will result in very poor results (for example, when N_e = 200, α = 0.01 and it is random selection). This is because, when the sample is very few in the local set, we cannot update the parameters accurately by formula (4).

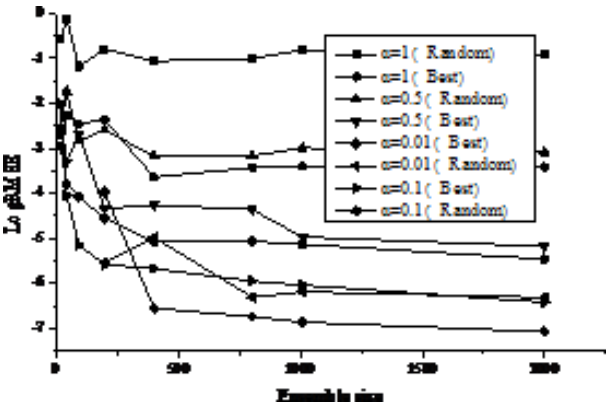

Fig. 1. Natural logarithm (Log RMSE) of root mean square error between the model output and the observed values when the number of set samples N_e and coefficient α of different values are given.

3.2. Case 2

In this case, we test the ability of the ILUES algorithm to deal with multi-peak prior problems. Here, we consider in a more practical example and this example is the rainfall-runoff model based on Boyle development, HYMOD. As shown in Fig. 2, the HYMOD model takes account of the following process: after a rainfall, the watershed passes through 3 high flow reservoirs and 1 low flow reservoir, and then generates runoff. The HYMOD model has 5 unknown parameters in total, including the maximum storage capacity C_{max}, storage capacity space change index b_{exp} [–], distribution coefficient β [–] between high flow reservoir and low flow reservoir, duration time for low flow reservoir R_s [T] and duration time for high flow reservoir R_q [T]. Among them, the priori of C_{max} and b_{exp} is multi modal and can be represented by the formulas (5) and (6) Gauss mixture model.

$$p(C_{\max}) = \frac{1}{3}N\left(100,20^2\right) + \frac{1}{3}N\left(250,20^2\right) + \frac{1}{3}N\left(400,20^2\right) \quad (5)$$

$$p\left(b_{exp}\right) = \frac{1}{3}N\left(0.5,0.1^2\right) + \frac{1}{3}N\left(1,0.1^2\right) + \frac{1}{3}N\left(1.5,0.1^2\right) \quad (6)$$

The prior distribution of the remaining 3 parameters is uniformly distributed, and their prior range is shown in Table 1.

By setting up N_e = 200 and α = 0.1, we can accurately estimate the parameters of the unknown model with the ILUES algorithm. Compared with the first case, although this case considers more model parameters, because the number of peaks in the parameter distribution is relatively small, we can use the smaller N_e value to accurately estimate the parameters.

Table 1
The priori range and true value of model parameters in case 2

Parameters	Range	True value
C_{max} (L)	[1500]	409.1018
b_{exp}	[0.12]	1.5430
β	[0.1, 0.99]	0.8998
R_S (L)	[0, 0.1]	0.0233
R_q (L)	[0.1, 0.99]	0.7232

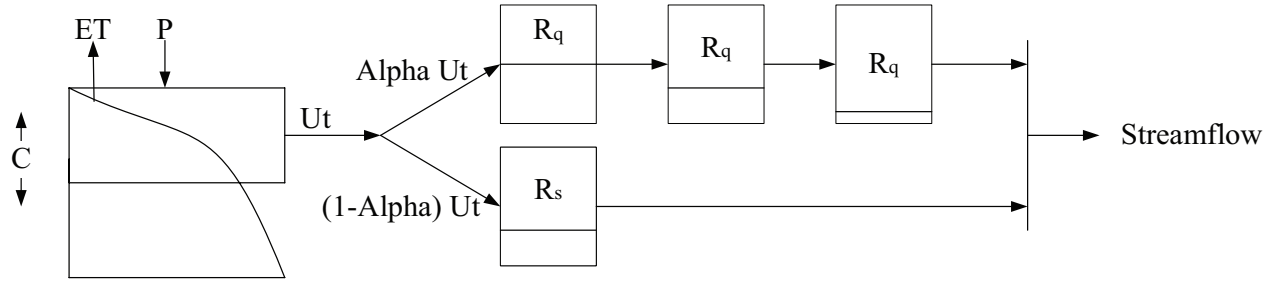

Fig. 2. A schematic diagram of the HYMOD model.

3.3. Case 3

Here we consider an example of the analysis of the source of pollution in the steady flow of groundwater. As shown in Fig. 3, the flow field has the upper and lower boundaries of the impermeable water and the left and right boundary of the fixed water head. Here, we assume that the permeability coefficient and porosity of the flow field are homogeneous and known, and their values are $K = 8$ [LT^{-1}] and $\theta = 0.25$ [–]. At the upstream of the flow field located at (x_s, y_s) [L], the pollution source, from the moment of t_{on} [T], begins to release pollutants to the downstream with constant intensity S_s [MT^{-1}], and it stops releasing at t_{off} [T] time.

The prior of the five pollution source parameters are evenly distributed, and their range is shown in Table 2. To estimate these 5 parameters, at the moments of t = [6, 8, 10, 12, 14] [T], we obtained concentration observations from a sampling location (the circle of Fig. 3), where the observation error accords with $\varepsilon = N(0, 0.01^2)$.

By setting up $N_e = 300$ and $\alpha = 0.1$, we can accurately inverse the parameters of the unknown model with the ILUES algorithm. The posteriori of y_s is in double-peak distribution. In order to verify the accuracy of the parameter inversion results obtained by the ILUES algorithm, we use the DREAM algorithm to retrieve the 5 pollution sources parameters again. In the DREAM algorithm, we use 8 parallel chains and each chain length is 2000. We also choose the Gauss form likelihood ratio. The model parameter obtained by DREAM algorithm is very consistent with that of the ILUES algorithm, and the computational complexity required by the ILUES algorithm is far lower than that of the DREAM algorithm.

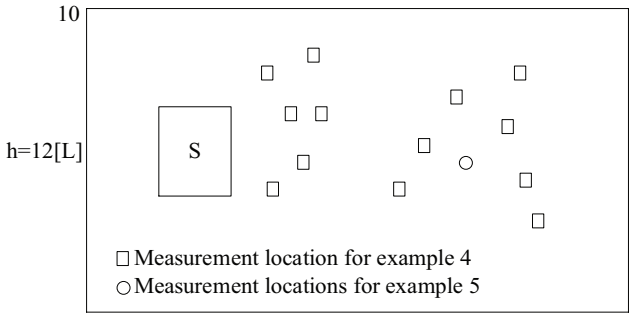

Fig. 3. The flow field and sampling position in case 3.

Table 2
The priori range and true value of model parameters in case 3

Parameters	Range	True value
x_s (L)	[3,5]	3.8537
y_s (L)	[3,7]	5.9994
S_s (MT^{-1})	[10,13]	11.0442
t_{on} (T)	[3,5]	4.8966
t_{off} (T)	[9,11]	9.0745

3.4. Case 4

Here we consider a case of high dimensional pollution source analysis. Unlike case 3, the intensity of the pollution source varies with time and it is described by 6 parameters, namely S_{si} [MT^{-1}] and $t_i = i{:}i{+}1$[T] where i = 1,..., 6. In this way, in addition to the pollution source location parameter (x_s, y_s), we need 8 parameters in total to describe the source of pollution. The prior of the 8 pollution source parameters conforms to uniform distribution, and their range is shown in Table 3. In this case, we use 100 KL expansion to express the permeability coefficient field, and after taking the natural logarithm, the spatial correlation formula of the field accords with Eq. (7). In addition, the mean value of the field is 2, the variance is 1, and the correlation length of x and y directions is 10 [L] and 5 [L], respectively.

$$C_Y(x, y, x', y') = \sigma_Y^2 \exp\left(-\frac{|x - x'|}{\lambda_x} - \frac{|y - y'|}{\lambda_y}\right) \tag{7}$$

Therefore, there are 108 parameters to be estimated in this case, that is, 8 pollution source parameters and 100 permeability coefficient parameters (that is, 100 KL terms). In order to estimate these 108 unknown parameters, we obtained water head and concentration observations at 15 sampling locations at t = [4, 5, 6, 7, 8, 9, 10, 11 and 12] [T] moments, and the observed errors accord with $\varepsilon = N(0, 0.005^2)$. Because this case contains more unknown parameters, we choose a larger collection of samples $N_e = 3000$ in the ILUES algorithm. The ILUES algorithm can accurately estimate the pollution source parameters within 5 iterations. In addition, we show the real Y field, the posterior sample of 3 Y fields, the mean value of the Y field estimation, and the variance of the Y field estimate. The ILUES algorithm can also accurately estimate the Y field. This shows the applicability of the ILUES algorithm in the inverse problem of high dimensional parameters.

In this case, although no posterior parameter is obviously a multi-peak distribution, we can still use the same setting as the multi-peak case. From this point of view, the ILUES algorithm is better than the clustering-based algorithm, because people need to set the number of clustering in the algorithm based on clustering analysis. In addition, the computational complexity of the ILUES algorithm is lower than that of the MCMC algorithm, especially in the inverse problem of high dimensional parameters. For example, in this 108 dimensional case, even if we use highly

Table 3
The priori range and true value of model parameters in case 4

Parameters	Range	True value
x_s [L]	[3,5]	3.5196
y_s [L]	[4,6]	4.4366
S_{s1} [MT^{-1}]	[0,8]	5.6916
S_{s2} [MT^{-1}]	[0,8]	7.8833
S_{s3} [MT^{-1}]	[0,8]	6.3064
S_{s4} [MT^{-1}]	[0,8]	1.4852
S_{s5} [MT^{-1}]	[0,8]	6.8717
S_{s6} [MT^{-1}]	[0,8]	5.5517

efficient DREAM algorithm, we need to call original model for exceeding hundreds of thousands of times, while ILUES algorithm only takes 18000 times. Moreover, ILUES algorithm can make full use of parallel computing. If we can run ILUES algorithm on multi-core workstations, then the time needed for simulation will be greatly reduced.

4. Conclusion

We propose an algorithm called iterative locally updated ensemble smoothing (ILUES), which is used to solve the problem of parameter inversion in high dimensional Gauss case. Without clustering analysis, the ILUES algorithm can accurately identify the multi-peak distribution of the parameters. Compared with the MCMC algorithm, the ILUES algorithm has a significant advantage in the computational complexity, especially for the high dimensional problem.

In order to verify the effect of the ILUES algorithm, we tested 5 numerical examples. The results showed that the ILUES algorithm could effectively deal with the problem of parameter inversion of high dimensional multi-peak. Then, we tested 3 hydrological models, taking into account the prior multiple peaks of the parameters, the posterior multi-peak parameters and the high dimension parameters 3 different scenes. These examples show the effect of the ILUES algorithm in the parameter inversion of the complex model.

References

[1] M. Arauzo, J.J. Martínez-Bastida, Environmental factors affecting diffuse nitrate pollution in the major aquifers of central Spain: groundwater vulnerability vs. groundwater pollution. Environ. Earth Sci., 73 (12) (2015) 1–16.

[2] M. Sakizadeh, E. Ahmadpour, Geological impacts on groundwater pollution: a case study in Khuzestan Province. Environ. Earth Sci., 75 (1) (2016) 1–12.

[3] K. Ostad-Ali-Askari, M. Shayannejad, H. Ghorbanizadeh-Kharazi, Artificial neural network for modeling nitrate pollution of groundwater in marginal area of Zayandeh-rood River, Isfahan, Iran. KSCE J. Civil Eng., 21(1) (2016) 1–7.

[4] M. Lasagna, D.A.D Luca, E. Franchino, Nitrate contamination of groundwater in the western Po Plain (Italy): the effects of groundwater and surface water interactions. Environ. Earth Sci., 75(3) (2016) 1–16.

[5] S. Venkatramanan, S.Y. Chung, T. Ramkumar, G. Gnanachandrasamy, S. Vasudevan, S.Y. Lee, Application of GIS and hydrogeo chemistry of groundwater pollution status of Nagapattinam district of Tamil Nadu, India, Environ. Earth Sci., 73(8) (2015) 4429–4442.

[6] J. Li, X. Li, N. Lv, Y. Yang, B. Xi, M. Li, S. Bai, D. Liu, Quantitative assessment of groundwater pollution intensity on typical contaminated sites in China using grey relational analysis and numerical simulation, Environ. Earth Sci., 74(5) (2015) 3955–3968.

[7] L. Duarte, A.C. Teodoro, J.A. Gonçalves, A.J. Guerner Dias, E. Marques, A dynamic map application for the assessment of groundwater vulnerability to pollution, Environ. Earth Sci., 74(3) (2015) 2315–2327.

[8] B. Zhang, G. Li, P. Cheng, T.J. Yeh, M. Hong, Landfill risk assessment on groundwater based on vulnerability and pollution index, Water Resour. Manage., 30(4) (2016) 1465–1480.

[9] E. Martínez, S. Singh, J.L. Hueso, D.K. Gupta, Local convergence of a family of iterative methods for Hammerstein equations, J. Math. Chem., 54(7) (2016) 1370–1386.