Sensitivity analysis of geostatistical approach to recover pollution source release history in groundwater

Y. Q. Long[1], T. T. Cui[1], W. Li[1], Z. P. Yang[2], Y. W. Gai[3]

[1]Nanjing Hydraulic Research Institute, Nanjing, China
[2]College of civil engineering, Chongqing University, Chongqing, China
[3]Water Resources Service Center of Jiangsu Province, Nanjing, China

Abstract: The geostatistical approach has been studied for many year to identify the pollution source re-lease history in groundwater. We focus on the influence of observation error and hydraulic parameters on the groundwater pollution identification (PSI) result in the paper. Numerical experiment and sensitivity analysis are carried out to find the influence of observation point configuration, error and hydraulic parameters on the PSI result in a 1D homogeneous aquifer. It has been found out that if concentration observation data could accurately describe the characteristics of the real concentration plume at the observed time point, a nice identification of the pollution release process could be obtained. If the calculated pollution discharge process has good similarity with the real discharge process, the order of the observation error fell within 10-6 and 10-3.5, the dispersion coefficient varies fells within -10% and 5%, and the actual mean velocity fell within ±2%. The actual mean velocity is the most sensitive parameter of the geostatistical approach in this case.

1 Introduction

In the recent twenty years, groundwater quality in some places of China is getting worse because of the industrial and living sewage, pesticide, leakage of petroleum tank and landfill yard. According to the survey of shallow groundwater in 118 cities in China, 97.5% cities' groundwater was polluted, and 40% was serious polluted[5]. It is very important to protect groundwater so as to assure the sustainable development and survival safety. However, groundwater pollution is difficult to be perceived, and the pollution source is hard to be identified. Pollution source identification (PSI) refers to reconstructing the pollution source locations and releasing histories from observed concentration records [12]. As one of the first steps in environmental remediation project, PSI can be classified into three typical types [8]: namely, finding the release history of a source, finding the location of a source, and recovering the initial distribution of a contaminant plume. The PSI is helpful to making a cost-effective remediation strategy, partitioning the cleanup cost among liable parties [10].

The mathematical and simulation approaches of pollution source identification has been extensively investigated in the past thirty years. Atmadja & Bagtzoglou [1] have subdivided the existing mathematical methods into four major groups, namely optimization, analytical and direct methods as well as probabilistic and geostatistical approaches. Snodgrass & Kitanidis [11] used a probabilistic approach combining Bayesian theory and geostatistical techniques to estimate the pollution source function. The method is an improvement from some other methods in that the solutions are more general and make no blind assumptions about the nature and structure of the unknown source function. Limitation to this approach is that the location of the potential source must be known a priori. Butera &

Tanda [2] use the method to find the source function in a 2D problem. Michalak & Kitanidis [7] combine the method with the adjoint state method to identify the source function in a 3D problem. Butera et al. [3] extend the method to find both the source function and location. Though this method is give extensively studied, most of these researches give their attention on the theory, the influence of observation error and hydraulic parameters on the PSI result are seldom discussed.

In this paper, we give our attention on the influence of observation error and hydraulic parameters on the PSI result when the geostatistical approaches is used to find the source release history in a 1D homogeneous aquifer. Numerical experiment and sensitivity analysis are carried out to find the influence of observation point configuration, error and hydraulic parameters on the PSI result.

2 Theory of Geostatistical Approach

Snodgrass & Kitanidis [11] used the geostatistical approach to estimate the pollution source release history in a simple 1D homogeneous aquifer. The pollution source release history was taken as an unknown function which is represented as a random process because there was uncertainty associated with the function and its true value may never be found. The set of all possible functions that fit the data were consistent with additional information. Each of these function was assigned a probability that it was the solution. The expected value of this set was sought as a best estimate along with its covariance as a measure of the estimation uncertainty.

2.1 *Geostatistical model*

The estimation problem could be expressed as:

$$\mathbf{z} = \mathbf{Hs} + \mathbf{v} \tag{1}$$

Where **z** is an $m\times1$ vector of observations. **H** is a known sensitivity matrix assembled by transfer function. **s** is an $n\times1$ "state vector" obtained from the discretization of the unknown function that we wish to estimate. The measurement error is represented by the vector **v** which is assumed to have zero mean and known covariance matrix **R**. The expected value and covariance of **s** could be expressed as equation (2) and (3).

$$E\left[\mathbf{s}\right] = \mathbf{X\beta} \tag{2}$$

$$\mathbf{Q}(\theta) = E\left[(\mathbf{s}-\mathbf{X\beta})(\mathbf{s}-\mathbf{X\beta})^T\right] \tag{3}$$

Where **X** is a know $n\times p$ matrix and $\boldsymbol{\beta}$ are p unknown drift coefficients. $\mathbf{Q}(\theta)$ is a Gaussian function of unknown parameters θ.

2.2 *Estimation procedure*

The estimation procedure is divided into two parts. First the optimal structural parameters θ are found, and then the unknown function **s** is estimated. The structural parameters θ are estimated by maximizing the probability of the measurements given θ:

$$p(z\,|\,\theta) \propto |\boldsymbol{\Sigma}|^{-1/2}\left|\mathbf{X}^T\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{HX}\right|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{z}^T\boldsymbol{\Xi}^{-1}\mathbf{z}\right] \tag{4}$$

$$\boldsymbol{\Sigma} = \mathbf{HQH}^T + \mathbf{R} \tag{5}$$

$$\boldsymbol{\Xi} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{HX}\left(\mathbf{X}^T\mathbf{H}^T\boldsymbol{\Sigma}^{-1}\mathbf{HX}\right)^{-1}\mathbf{X}^T\mathbf{H}^T\boldsymbol{\Sigma}^{-1} \tag{6}$$

Maximizing $p(\mathbf{z}|\theta)$ is equivalent to minimizing

$$L(\theta)=\frac{1}{2}\ln|\mathbf{\Sigma}|+\frac{1}{2}\ln\left|\mathbf{X}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{HX}\right|+\frac{1}{2}\mathbf{z}^T\mathbf{\Xi}^{-1}\mathbf{z} \tag{7}$$

The minimization can be achieved by taking derivatives of $L(\theta)$ with respect to θ and setting them to zero. Gauss-Newton iterations is used to find the minimization. When the iteration converge, $\mathbf{Q}(\theta)$ is known and solve the system

$$\begin{bmatrix}\mathbf{\Sigma} & \mathbf{HX}\\ (\mathbf{HX})^T & 0\end{bmatrix}\begin{bmatrix}\mathbf{\Lambda}^T\\ \mathbf{M}\end{bmatrix}=\begin{bmatrix}\mathbf{HQ}\\ \mathbf{X}^T\end{bmatrix} \tag{8}$$

Where $\mathbf{\Lambda}$ is a m×n matrix of coefficients and $\mathbf{M}$ is p×n matrix of multipliers. The best estimates of the function $\mathbf{s}$ and its covariance are

$$\hat{\mathbf{s}}=\mathbf{\Lambda z} \tag{9}$$

$$\mathbf{V}=-\mathbf{XM}+\mathbf{Q}-\mathbf{QH}^T\mathbf{\Lambda}^T \tag{10}$$

2.3 *Nonnegative constrain*

The method does not enforce the nonnegativity of concentration. A transformation of the concentration is used to assure the nonnegativity of concentration. Define

$$\tilde{\mathbf{s}}=\alpha\left(\mathbf{s}^{1/\alpha}-1\right) \tag{11}$$

The equation (1) in the transformed space becomes

$$\mathbf{z}=\mathbf{h}\left[\left((\tilde{\mathbf{s}}+\alpha)/\alpha\right)^{\alpha}\right]+\mathbf{v}=\tilde{\mathbf{h}}(\tilde{\mathbf{s}})+\mathbf{v} \tag{12}$$

Then the transfer function $\tilde{\mathbf{h}}(\tilde{\mathbf{s}})$ is not linear with respect to the transformed unknown $\tilde{\mathbf{s}}$. The best estimate of $\mathbf{s}$ can be found by the quasi-linear procedure [6, 11] and could be expressed as

$$\hat{\mathbf{s}}=\left(\frac{\tilde{\mathbf{s}}_l+\alpha}{\alpha}\right)^{\alpha} \tag{13}$$

3 Numerical Case and Experiment Plan

3.1 *Numerical case*

The advective and dispersive transport of a conservative solute in a 1D homogeneous aquifer [10] is taken as an example problem to discuss the influence of observation point configuration, error and hydraulic parameters on the PSI result. The problem could be expressed as

$$\frac{\partial C}{\partial t}=D\frac{\partial^2 C}{\partial x^2}-v\frac{\partial C}{\partial x} \tag{14}$$

$$C(x,T)=g(x) \tag{15}$$

$$C(0,t)=0 \quad 0\le t\le T \tag{16}$$

$$C(\infty,t)=0 \quad 0\le t\le T \tag{17}$$

Where C is the pollutant concentration, D is the dispersion coefficient (D=1), v is the actual mean velocity (v=1), x is the transport distance ($x\in[0,300]$), t is time.

The analytical solution of the 1D problem is

$$C(x,T)=\int_0^T s(\tau)\,f(x,T-\tau)\,d\tau \tag{18}$$

$$f(x,T-\tau)=\frac{x}{2\sqrt{\pi D(T-\tau)^3}}\exp\left[-\frac{\left(x-v(T-\tau)\right)^2}{4D(T-\tau)}\right] \tag{19}$$

Equation 20 shows the true release history (Fig. 1). There are 25 observation points in the *x* direction and the curve of observed concentration **z** at *t*=300 is shown in Figure 2. The covariance of the measurement errors is expressed as $\mathbf{R}=\sigma^2_R\mathbf{I}$ ($\sigma^2_R=1\times10^{-12}$). The **Q** is expressed as equation (21).

$$s(t)=\exp\left[-\frac{(t-130)^2}{50}\right]+0.3\exp\left[-\frac{(t-150)^2}{200}\right]+0.5\exp\left[-\frac{(t-190)^2}{98}\right] \tag{20}$$

$$Q(t_i,t_j|\theta)=\sigma^2\exp\left[-\frac{(t_i-t_j)^2}{l^2}\right] \tag{21}$$

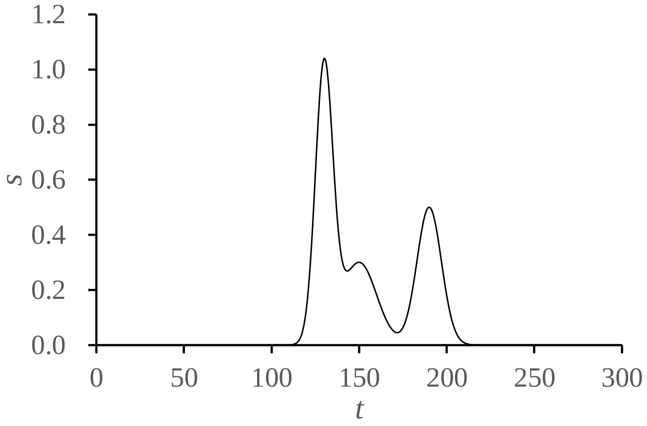


Figure 1 Pollution source release curve

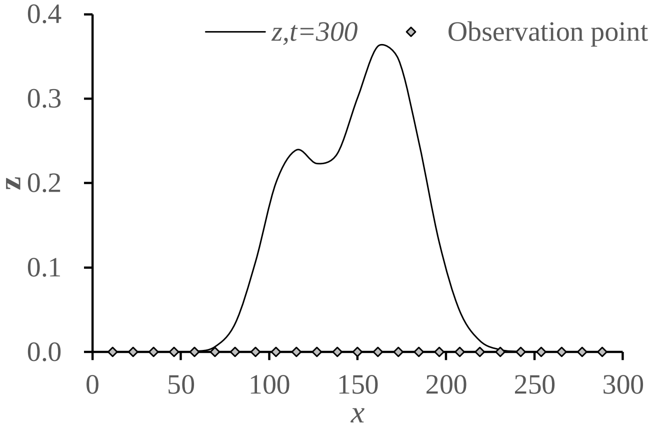


Figure 2 observation location and observed concentration at t=300

3.2 *Numerical experiment plan*

(1) The influence of observation point configuration on PSI

The complexity of PSI problems depends on the amount of observations available and the number of system inputs that must be determined [9]. The amount and configuration of observation points decided by the capital and location are the base to obtain the available observation date for PSI. It is helpful for designing observation point to analyze the influence of amount and configuration. To find the influence of observation amount on PSI, different amount observation points are set in the *x* direction within (0, 300]. Then the size of observation zone and observation point location are discussed by setting the same amount of observation points in different size of observation zone, and moving the observation zone in the *x* direction within (0, 300].

(2) The sensitivity of observation error

As a kind of ill-posed problem, the solution of PSI may do not satisfy the general conditions of existence, uniqueness, or stability. The ill-posed problems are extremely sensitive to errors in data, so small errors in the measurement of the existing plume may drastically change the calculated plume history [10]. It is helpful to analyze the sensitivity of observation error for finding the proper error scope which could not affect the estimate of release history extremely. Snodgrass & Kitanidis [10] set the covariance of the measurement errors as $\mathbf{R}=\sigma^2_{\mathrm{R}}\mathbf{I}$ ($\sigma^2_{\mathrm{R}}=1\times10^{-12}$). We discuss the sensitivity of observation error by increasing σ^2_{R} from 10^{-12} to 10^{-1}.

(3) The sensitivity of dispersion coefficient

Contaminants in groundwater are transported by the following three processes: advection, mechanical dispersion, and molecular diffusion. Mechanical dispersion and molecular diffusion collectively are referred to as hydrodynamic dispersion which could be obtained by in situ experiment [4]. Many researches obtain the dispersion coefficient by numerical model calibration [13]. Both in situ experiment and numerical model calibration cannot avoid the error or uncertainty which might bring disturbance in the PSI result. In the numerical case of Snodgrass & Kitanidis [11], the dispersion coefficient *D* is assigned the value 1. We analyze the sensitivity of dispersion coefficient by changing *D* within ±15%.

(4) The sensitivity of actual mean velocity

Advection is a result of the large-scale gradients in fluid head and it is most significant mass transport process. The velocity of groundwater is described by the actual mean velocity of the water movement through the pores of the soil. The observation error and model uncertainty could be brought in the PSI result, when the experimental or numerical method is carried out to estimate the actual mean velocity. Snodgrass & Kitanidis [11] assigned *v* the value 1. We analyze the sensitivity of *v* by changing it within ±15%.

3.3 *Index for evaluating the sensitivity*

The linear correlation coefficient and width of confidence interval are taken as the index for evaluating the sensitivity of observation error, dispersion coefficient and actual mean velocity. The linear correlation coefficient *r* calculated as equation (22) shows the similarity of the calculated release history *cal* and the real release history *Rea*. When the *r* approximates 1, the *cal* is more similar with the *Rea*. The width of confidence interval shows the uncertainty of calculated release history based on the observation data and model parameters. The Euclidean distance *de* between the up and low bound (*σu* and *σl*) of the 95% confidence interval is used to evaluate the confidence interval (equation 23). The sensitivity β_i of the *i*th parameter a_i on the index *I* is expressed as equation (24), Δa_i is the change of the parameter a_i.

$$r=\frac{\sum_{i=1}^{n}\left(cal_i-\overline{Rea}\right)\left(Rea_i-\overline{Rea}\right)}{\sqrt{\sum_{i=1}^{n}\left(cal_i-\overline{Rea}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Rea_i-\overline{Rea}\right)^2}} \tag{22}$$

$$de = \sqrt{\sum_{i=1}^{n} \left(\sigma u_i - \sigma l_i\right)^2} \tag{23}$$

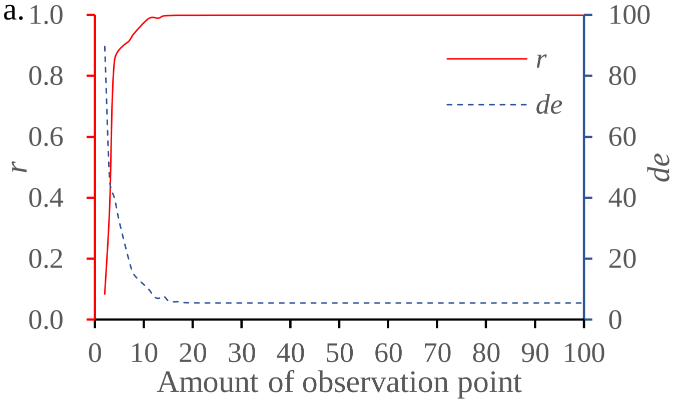
$$\beta_i = \frac{\left(I\left(a_i + \Delta a_i\right) - I\left(a_i\right)\right) / I\left(a_i\right)}{\Delta a_i / a_i} \tag{24}$$

4 Discussion

4.1 *Configuration of observation points*

Throughout the numerical case, we find that when the amount of observation points increase, *r* approximates 1 and *de* approximates 5.44 (Fig. 3a). The concentration information provided by less observation points cannot express the accurate pollution plume at *t*=300 (Fig. 4), so the PSI result depended on these information cannot describe the real release history. When the amount of observation points increase, the calculated history approximates the real history. When the number of observation points equal 25, the concentration information provided by observation points can accurately express the pollution plume at *t*=300. Then it is helpless to improve the PSI result by increasing observation points. It means 25 observation points could provide enough concentration information for identifying the release history in the case.

We set 25 observation points in different observation zones of different width, and each observation zone moves from *x*=0 to *x*=300. The PSI results of different zones are shown in Figure3b. To keep the figure clear, we draw the curve of *r* and *de* corresponding to the width of 25, 100, 150, 200 and 300. Figure 3b shows that when the width of observation zone is narrow, the concentration information provided by observation points cannot express the accurate pollution plume at *t*=300. The narrower the width is, the bigger the deviation between the calculated and the real release history is. The peak and valley of the curve are the important information to decide the characteristic of the pollution plume at *t*=300. If the narrow zone can provide the concentration information of these peak and valley, it could obtain a better PSI result than the zone lost these peak and valley does. So the accuracy of a recovered plume history strongly depends on the accuracy of the characterization of the current plume(Skaggs & Kabala 1994).



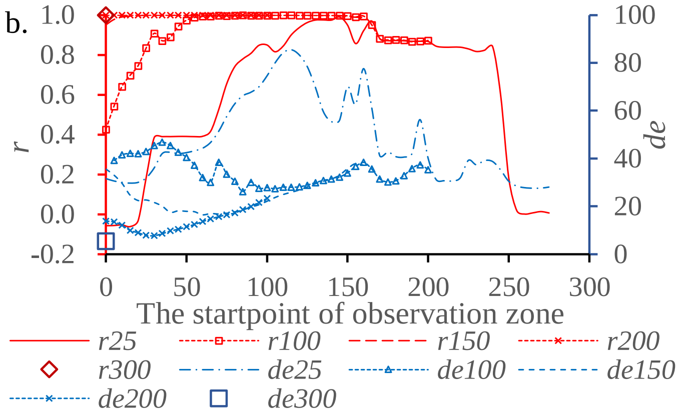


Figure 3 The influence of configuration of observation point on the *r* and *de*

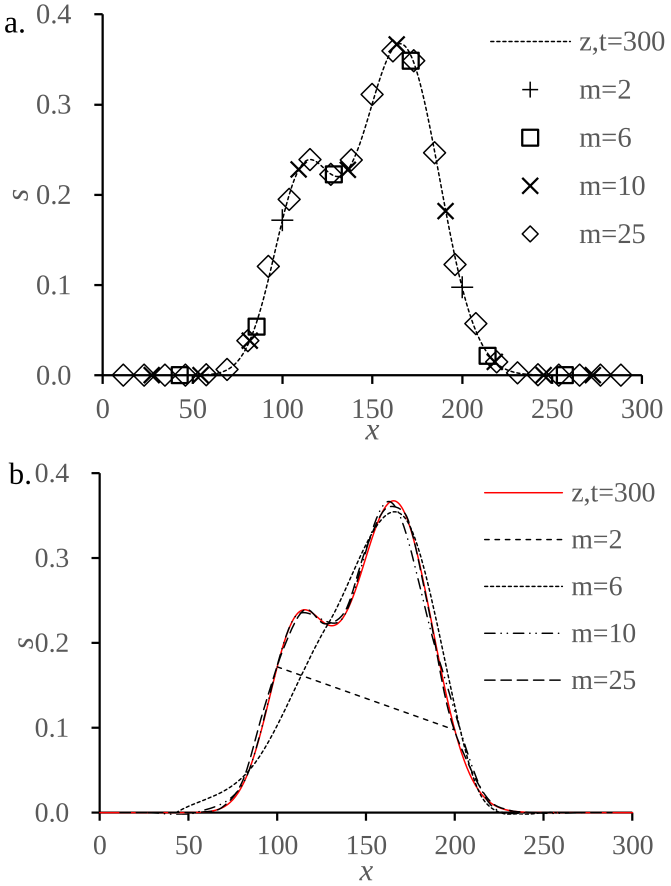


Figure 4 The location of different set of observation points and the interpolated concentration curves

4.2 *Configuration of observation points*

Snodgrass & Kitanidis [11] set σ^2_R equals 1×10^{-12}, and obtained a perfect PSI result. We take 1×10^{-12} as a datum to evaluate its influence on PSI. Figure 5 shows that when σ^2_R increases, the PSI result gets worse. If the σ^2_R is bigger than 10^{-8}, the PSI result get worse rapidly, and the peak of calculated history become lower.

The order of the detection limit is not less than 10^{-7} usually, so the order bigger than 10^{-7} makes sense. σ^2_R controls the covariance of the measurement errors. When σ^2_R equals 10^{-12}, the order of **z** is 10^{-6}. It means a perfect PSI result could be obtain when the order of observation error is 10^{-6}. If the σ^2_R increases to 10^{-7}, the observation error equals to 10$^{-3.5}$. The *r* drops down -0.28%, and the *de* increases 10.24%. Then the calculated history gets a moderate similarity with real history (Fig. 6). When σ^2_R

equals 10^{-6}, the order of the error of **z** is 10^{-3}, the *r* drops down -0.90%, and the *de* increases 30.77%. However, the peak during the concentration curve is different from the real history.

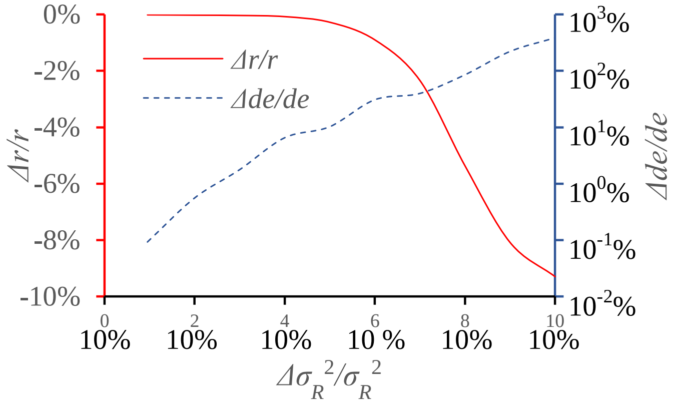


Figure 5 The sensitivity of observation on *r* and *de*

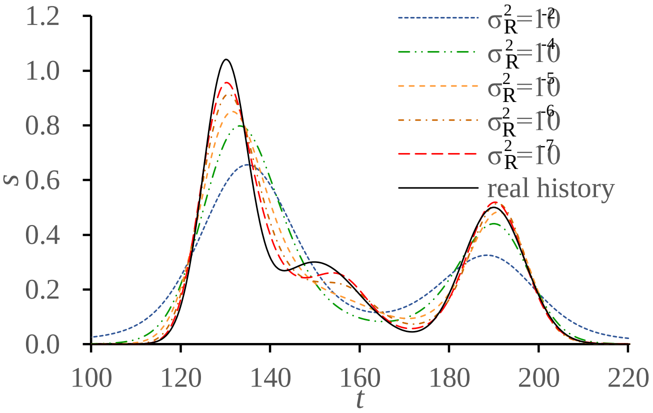


Figure 6 The calculated under different σ2R and the real release history

4.3 *Dispersion coefficient*

When the *D* increases from -15% to 15%, the *r* changes from -35% to 5%, and the *de* changes from -15% to 65% (Fig. 7). Then at the most of time, the *r* is bigger than 0.9, and the *de* is smaller than 7 (Fig. 8). However, the peak at *t*=150 disappears, and a strange peak appears at *t*=240. When the *D* increases, the concentration curve peak increases. When the *D* decreases, the concentration curve peak decreases too. Small *de* tells that the uncertainty decreases based on the known pollution concentration. However, small uncertainty does not mean the calculated release history approximate the real history, because the model parameter might be not accurate.

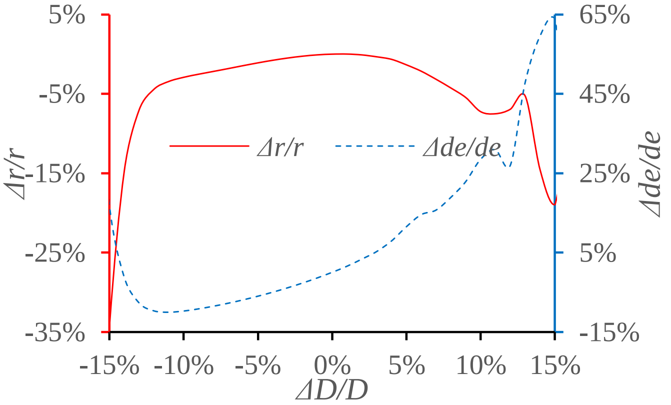


Figure 7 The sensitivity of *r* and *de* to *D*

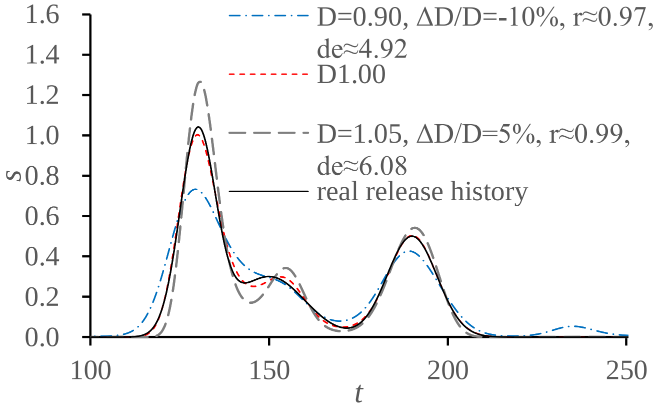


Figure8 The calculated under different *D* and the real release history

4.4 *Dispersion coefficient*

When the *v* increases from -5% to 5%, the *r* changes from -40% to 0%, and the *de* changes from -10% to 10% (Fig. 9). If the *v* is smaller than the real actual mean velocity, the peak and valley of the calculated release history curve appear earlier than the real release history. If the *v* is bigger than the real actual mean velocity, the peak and valley of the calculated release history curve appear later than the real release history (Fig. 10). If the *v* changes within ±2%, the calculated history is similar with the real history, but if the *v* fell out of this scope, the similarity becomes worse because the peak and valley might be far different from the real history. When the *v* changes within ±5%, the *de* decreases to 5, while the *v* increases. Small *de* tells that the uncertainty decreases based on the known pollution concentration. However, small uncertainty does not mean the calculated release history approximate the real history, because the *v* might be not accurate.

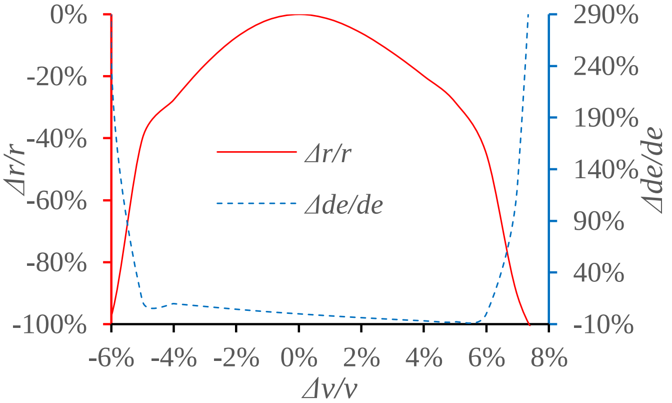


Figure 9 The sensitivity of *r* and *de* to *v*

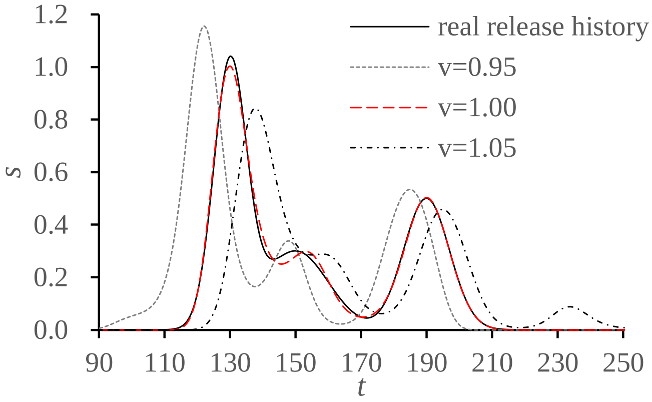


Figure 10 The calculated under different *v* and the real release history

4.5 *Sensitivity of parameters*

The σ^2_{R} related to the observation of the pollutant concentration controls the covariance of the measurement errors, it is the external cause which could influence the PSI effect of the geostatistical approach. The *D* and *v* are the intrinsic factors of the equation that describes the pollutant transportation. Figure 11 shows the sensitivity of *r* and *de* to σ^2_{R}, *D* and *v*. The sensitivity could be negative and the orders are so different, we have to use a semilogarithmic coordinate like Figure 11 to show the sensitivity of these parameter clearly. When these parameters change within ±6%, the *v* is the most sensitive parameter for *r*. When these parameters change within ±5%, *de* is sensitive to *v* and *D*. If the change goes beyond this scope, the sensitivity of *de* to *v* increases rapidly. In the well-posed transportation calculation, if the convection calculation can describe the real transportation well, the dispersion could be minority [13], that means the *v* is very important part of pollutant transportation calculation. As a typical ill-posed problem, tiny turbulent could lead to a notable diversity of PSI result, so both the *r* and *de* are sensitive to *v*.

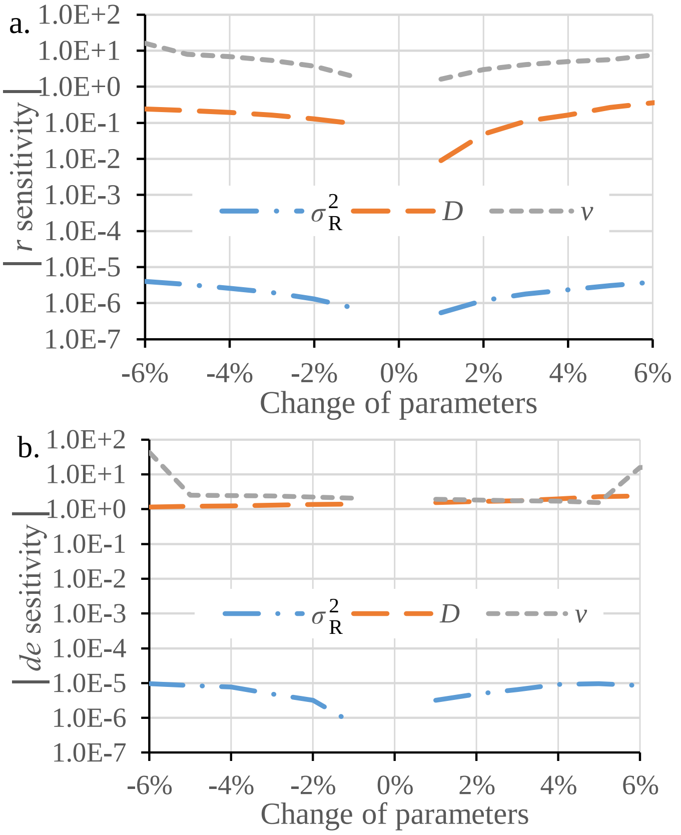


Figure 11 The comparison of sensitivity of *r* and de to σ^2_{R}, *D* and *v*

5 Conclusion

(1) The accuracy of a recovered plume history strongly depends on the accuracy of the characterization of the concentration distribution at the final time spot. The aim of setting the number and configuration of observation points is obtaining the accuracy concentration distribution at the final time spot. (2) When the σ^2_{R} changes within 10^{-12} and 10^{-7}, the order of observation error changes within 10^{-6} and $10^{-3.5}$, the calculated history gets a moderate similarity with real history. (3) When the *D* changes within -10% and 5%, the calculated release history approximate the real history. (4) When the *v* changes within ±2%, the calculated history has good similarity with the real history. (5) The σ^2_{R} is the external cause of PSI problem, while the *D* and *v* are the intrinsic factors, the *v* is the most sensitive parameter of the PSI problem.

Acknowledgement

This study is financially supported by the National Natural Science Foundation of China (Grants No. 51409161, No. 51509157), the Natural Science Foundation of Jiangsu Province (Grants No. BK20140080), the Technical demonstration project of Ministry of Water Resources (SF-201706), The National Key Research and Development Program of China (No. 2016YFC0402808 and 2016YFC040281002).

References

[1] Atmadja, J. & Bagtzoglou, A.C. 2001. State of the Art Report on Mathematical Methods for Groundwater Pollution Source Identification, *Environmental Forensics* 2(3): 205-214.

[2] Butera, I. & Tanda, M.G. 2003. A geostatistical approach to recover the release history of groundwater pollutants. *Water Resources Research* 39(12): 1372, doi: 10.1029/2003WR002314.

[3] Butera, I., Tanda, M.G., Zanini, A. 2013. Simultaneous identification of the pollutant release history and the source location in groundwater by means of a geostatistical approach, *Stochastic Environmental Research and Risk Assessm*. 27(5): 1269-1280.
[4] Fetter, C.W. 2008. *Contaminant Hydrogeology*. Waveland Pr Inc: Illinois.
[5] Jiang, J.J. 2007. Groundwater pollution and prevention of China, *Environmental Protection* 38(19): 16-17.
[6] Kitanidis, P.K. 1995. Quasi-linear geostatistical theory for inversing, *Water Resources Research* 31(10): 2411-2419.
[7] Michalak, A.M. & Kitanidis, P.K. 2004. Estimation of historical groundwater contaminant distribution using the adjoint state method applied to geostatistical inverse modeling, *Water Resources Research* 40, W08302, doi: 10.1029/2004WR003214.
[8] Milnes, E. & Perrochet, P. 2007. Simultaneous identification of a single pollution point-source location and contamination time under known flow field conditions, *Advances in Water Resources* 30(12): 2439-2446.
[9] Mirghani, B.Y., Mahinthakumar, K.G., Michael, E.T., et al. 2009. A parallel evolutionary strategy based simulation–optimization approach for solving groundwater source identification problems, *Advances in Water Resources* 32(9): 1373-1385.
[10] Skaggs, T.H. & Kabala, Z.J. 1994. Recovering the release history of a groundwater contaminant, *Water Resources Research* 30(1): 71-79.
[11] Snodgrass, M.F. & Kitanidis, P.K. 1997. A geostatistical approach to contaminant source identification, *Water Resources Research* 33(4): 537-546.
[12] Sun, A.Y., Painter, S.L., Wittmeyer, G.W., 2006. A constrained robust least squares approach for contaminant source release history identification, *Water Rescources Research* 42, W04414 doi 10.1029/2005WR004312.
[13] Zheng, C.M. & Bennett, G.D. 2002. *Applied Contaminant Transport Modeling*. Wiley-Interscience: New York.