# A least squares method for identification of unknown groundwater pollution source

Zhukun He, Rui Zuo, Dan Zhang, Pengcheng Ni, Kexue Han, Zhenkun Xue, Jinsheng Wang and Donghui Xu

## ABSTRACT

The identification of unknown groundwater pollution sources is one of the most important premises in groundwater pollution prevention and remediation. In this paper, an exploratory application of a least squares method to identify the unknown groundwater pollution source is conducted. Supported by a small amount of observation data and the analytical solutions of the pollutant transport model, the initial concentration, the leakage location and the pollutant mass are identified by using the least squares method under a sand tank experiment and a gas station area. In the sand tank experiment, it is found that the fitting errors of three cross-sections are within 6%. In the gas station area, it is found that the results are nearly consistent with the site investigation information. The results indicate that the least squares method has considerable application values in the identification of groundwater pollution sources.

**Key words** | groundwater, identification, least squares method, multivariate nonlinear regression, pollution source

**Zhukun He**
**Rui Zuo** (corresponding author)
**Pengcheng Ni**
**Kexue Han**
**Zhenkun Xue**
**Jinsheng Wang**
**Donghui Xu**
College of Water Sciences,
Beijing Normal University,
Beijing 100875,
China
and
Engineering Research Center of Groundwater Pollution Control and Remediation,
Ministry of Education,
Beijing 100875,
China
E-mail: *zr@bnu.edu.cn*

**Dan Zhang**
Beijing Municipal Research Institute of Environmental Protection,
National Engineering Research Center of Urban Environmental Pollution Control,
Beijing Key Laboratory for Risk Modeling and Remediation of Contaminated Sites,
Beijing 100037,
China

## HIGHLIGHTS

- A least squares method is conducted in a sand tank experiment and a gas station area.
- The initial concentration, the leakage location and the pollutant mass are identified.
- The relative errors between observed values and fitting values are within 6%.
- Provide a new approach to identify unknown groundwater pollution sources.
- The release history of the tank is speculated.

## INTRODUCTION

Groundwater pollution could cause harm to human health, ecosystem and other environment, which has resulted in great urgency for protecting groundwater environments. The investigation of groundwater pollution sources is the premise of groundwater pollution remediation. However, the cost of acquiring the information on pollutants through observation is relatively high, and the observed data is usually not adequate to determine the location of pollution sources and the range of pollution plumes (Ayvaz 2016). In general, groundwater pollution source identification includes the identification of location, type, concentration and the diffusion range of the pollutant (Gu *et al.* 2017), which can make up for the above shortcomings effectively. Therefore, the identification of groundwater pollution sources becomes one of the most important premise tasks in groundwater



doi: 10.2166/nh.2021.088

pollution prevention, remediation and health risk assessment (Chang & Kashani 2009).

There are many methods to identify pollution sources in groundwater, which can be summarized by the following methods: the geochemical footprint method, the stochastic method and the optimization method (Mahar & Datta 2000). The geochemical footprint method refers to the method to obtain the parameters of pollution sources by means of isotope or chemical fingerprinting. For example, N-15 and O-18 were used as tracers to separate nitrate pollution plumes in underground aquifers and to identify the source of nitrate (Aravena *et al.* 1993; Li *et al.* 2019), isotopic ratios of N-15 and N-14 were used to identify the nitrate sources (Skute *et al.* 2016) and carbon stable isotopes were used to identify deposition location and time in oil drilling (Skaare *et al.* 2009). The geochemical footprint method is easy to understand and widely used in practice, but it has quantitative limitations due to its strict requirements on experimental conditions and standards (Mansuy *et al.* 1997).

The probability of each parameter of the groundwater pollution source is taken as the resulting output by the stochastic method. It can analyze the uncertainty of the pollution source identification, but the calculation load is large when there are many unknowns. The common methods in a random method are the Bayesian inference and the adjoint state method. Bayesian inference considers the ill-posed inverse problem of pollution source identification as a well-posed problem in an extended random space. It can calculate the posterior probability distribution of the inversion variables and solve the non-uniqueness solution problem caused by observation data noise. However, its expression is extremely complex, and it needs to sample the posterior parameter probability distribution to estimate the parameters (Wang & Zabaras 2006; Rojas *et al.* 2008; Koch & Nowak 2016). The adjoint state method is a method based on the idea of geological statistics, which obtains the sensitivity matrix through a specified function and the calculation of the probability of the occurrence to pollution sources at various positions (Dimov *et al.* 1998; Michalak & Kitanidis 2004).

By solving the pollutant transport model of groundwater, comparing the simulation results with the observation results and optimizing the model continuously, the pollution source parameters can eventually be obtained. The optimization method is one of the most widely used methods in the study of groundwater inverse problems. In recent years, some optimization methods are applied by more and more researchers. Yeh applied a hybrid heuristic approach for identifying contaminant source information in transient groundwater flow systems (Yeh *et al.* 2014). Huang applied a simulation-optimization model by integrating MODFLOW and MT3DMS into a shuffled complex evolution optimization algorithm, providing an approach to solve the source identification problems under the complex conditions (Huang *et al.* 2018). Borah proved that the numerical- and ANN-based simulation-optimization models have the potential for real-world field applications (Borah & Bhattacharjya 2016). In addition, the Bayesian global optimization algorithm (Pirot *et al.* 2019) and the alternative model method (Zhao *et al.* 2016; Xia *et al.* 2019) are also applied to identify contaminant source localization.

At present, there are few studies on the identification of groundwater pollution sources by the least squares method. The least squares method is a kind of mathematical optimization technology. It is widely used to estimate the numerical values of the parameters by fitting a function to a set of measured data (See *et al.* 2018). In this paper, the least squares method is applied to the laboratory and the site area to evaluate the identification results, and the limitations in the identification of groundwater pollution source are discussed.

## METHODS

### Least squares method in groundwater pollution sources identification

During the migration of pollutant in porous media, processes such as convection, dispersion and adsorption, as well as chemical processes such as oxidation–reduction, ion exchange, dissolution and sedimentation, and biodegradation, often occur. And these processes are also influenced by hydrogeological conditions such as formation lithology, aquifer thickness and saturation. Therefore, it is generally considered that the concentration distribution of pollutants in aquifers is a multiple nonlinear mathematical model. When using the least squares method to identify

groundwater pollution sources, it can be considered as a multiple nonlinear regression process using the least squares method.

After obtaining a small amount of measured data and the hydrogeological conditions of the site, the identification of groundwater pollution sources is a process that uses certain mathematical methods to invert the relevant information of pollution sources. The information of pollution sources usually includes sources' amount, location, intensity, pollution events and the spatiotemporal distribution of pollutants.

In a certain pollution situation, there is a batch of pollutant concentration monitoring data with time change on the section with the distance of $X$ from the pollution source $\{(t_i, C_i)\ (i = 0, 1, 2, \ldots, m)\}$. If the pollutant concentration distribution conforms to $C_i = C(t_i)\ (i = 0, 1, \ldots, m)$ and the minimum sum of the square of the error $\epsilon$ is obtained by fitting an approximate function of concentration distribution with the given concentration monitoring data, as shown in Equation (1), then the obtained $C*(t_i)$ can be regarded as the pollutant concentration distribution function under the pollution scenario, and the coefficients in the function can be viewed as the other groundwater pollution source information.

$$\|\varepsilon^2\| = \sum_{i=0}^{m} \varepsilon^2 = \sum_{i=0}^{m} [C*(t_i) - C_i]^2 = \min \sum_{i=0}^{m} [C(t_i) - C_i]^2 \quad (1)$$

### Groundwater flow and transport simulation

The Darcy's Law is the basic law to describe the groundwater flow in the saturated porous media. It can be written as Equation (2), where $V_i$ is a vector of the average linear velocity of groundwater flow ($LT^{-1}$), $n_e$ is the effective porosity (dimensionless), $K_{ij}$ is the hydraulic conductivity tensor of the porous media ($LT^{-1}$), $h$ is the hydraulic head (L) and $x_i$ are the Cartesian coordinates (Yeh *et al.* 2014).

$$V_i = -\frac{K_{ij}}{n_e}\frac{\partial h}{\partial x_i} \quad (2)$$

Based on three assumptions, the solute transport in the saturated and homogeneous aquifer, the groundwater flow conforms to Darcy's Law and the solute is chemically conservative, the solute transport models in groundwater can be described as the advection–dispersion equation (ADE), and it can be written as Equation (3), where $C$ is the concentration of solute ($ML^{-3}$), $D_{ii}$ is the dispersion coefficient along with different axis ($LT^{-1}$) and $u_i$ is the Darcy velocity ($LT^{-1}$).

$$\frac{\partial C}{\partial t} = \frac{\partial}{\partial x}\left(D_{xx}\frac{\partial C}{\partial x}\right) + \frac{\partial}{\partial x}\left(D_{yy}\frac{\partial C}{\partial y}\right) + \frac{\partial}{\partial z}\left(D_{zz}\frac{\partial C}{\partial z}\right) - \frac{\partial (u_x C)}{\partial x} - \frac{\partial (u_y C)}{\partial y} - \frac{\partial (u_z C)}{\partial z} \quad (3)$$

## STUDY AREA AND DATA

### Case 1: sand tank experiment

In this paper, the least squares method is used to identify groundwater pollution sources based on the total petroleum hydrocarbon (TPH) migration test data of indoor sand tanks. Figure 1 shows the illustration of the sand tank. The length, width and height of the sand tank are 160, 30 and 80 cm, respectively. And it is made of transparent organic glass. The tank is of the width of 10 cm in the left side and right side, functioning as the groundwater inlet area and outlet areas, respectively. There are 20 sampling holes on the side of the sand tank. Firstly, the petroleum pollutant is injected into the water surface 20 cm below the water surface by a peristaltic pump at a distance of about 25 cm from the water inlet area, then the TPH concentration of the

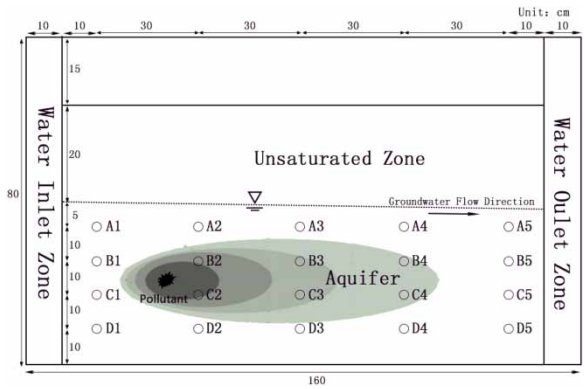


**Figure 1** | Sand tank illustration.

sample is detected by an infrared spectrophotometer at a certain time interval.

## Case 2: gas station area

There was a gas station with an area of about 8,700 $m^2$ in the study area, which was discarded in the 1970s. Due to the vulnerable structure and redox processes over many years, the storage tanks (including six diesel tanks and four gasoline tanks) were corroded, and the leakage has polluted the aquifer. Figure 2 shows the plan view of the old gas station site. The terrain is flat, with an average elevation of 56–57 m, inclined from north to south. The flow direction of groundwater is roughly from north to south, the same as that of elevation decline. The content of petroleum hydrocarbon in the soil of the study area was in the range of 3.17–16.47 mg $kg^{-1}$, the concentration reaches the maximum at 10 m below the surface.

The contaminated site is located in the alluvial proluvial area. The vadose zone is mainly composed of fine sand, gravel, coarse sand and other coarse-grained media. The aquifer is mainly located in the stratum with a depth of 25–40 m. The upper aquifer is a phreatic aquifer with a thickness of about 3 m; there is a clay layer with a thickness of about 4 m in the middle; the lower confined aquifer has a buried depth of about 32 m and a thickness of 8 m. Both aquifers are composed of gravel and coarse sand with good water yield. In addition, bedrock was found at a depth of 40 m.

In the gas station area, toluene in the phreatic aquifer is recognized as the characteristic pollutant. Toluene pollution is mainly from gasoline, while the content of diesel is very small. From July 2017 to September 2018, three wells in the study area were continuously monitored every 2 months. The release history of methylbenzene at 3# is shown in Figure 3.

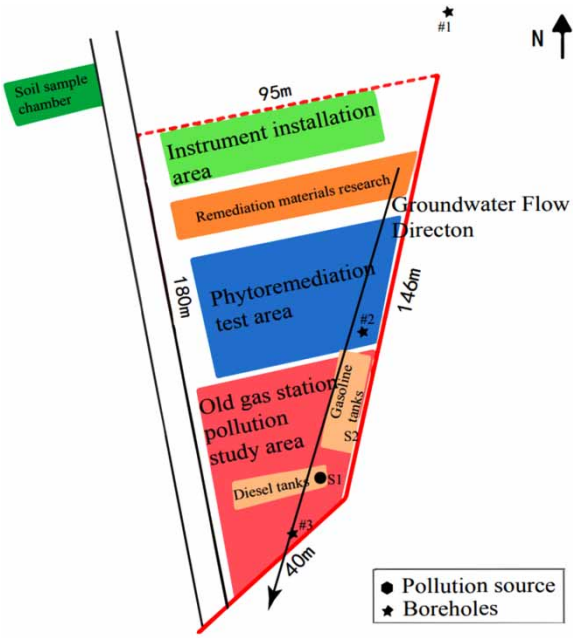


**Figure 2** | Plan view of the old gas station.

## Variation of components in the sand tank experiment

In order to investigate TPH's behavior in aquifer, the TPH concentration, pH and some concentrations of conventional ions including $NO_3^-$ and $SO_4^{2-}$ are monitored. Because the main groundwater flow direction is lined with B2 to B5, we analyzed this cross-sectional data partially. The profile of pH, $NO_3^-$ and $SO_4^{2-}$ can be seen in Figure 4, while the TPH concentration profile in the sand tank experiment can be seen in Figure 6.

pH at B5 showed a trend of first decreasing and then increasing. The decrease is because the respiration of microorganisms produces a certain amount of $CO_2$, which is slightly soluble in water and turns into $H_2CO_3$, moreover, the degradation of petroleum pollutants is a process involving acid production, as can be seen in Equation (4). When the pollutants had decreased to a certain degree, microbial degradation abated and water–rock interactions led to pH increasing (Qian *et al.* 2018). In addition, due to

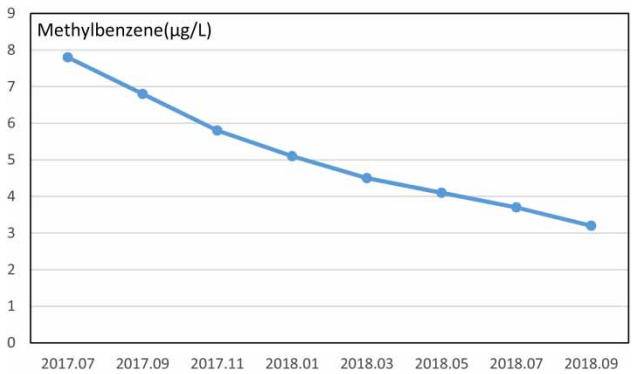


**Figure 3** | Release history of methylbenzene at 3#.

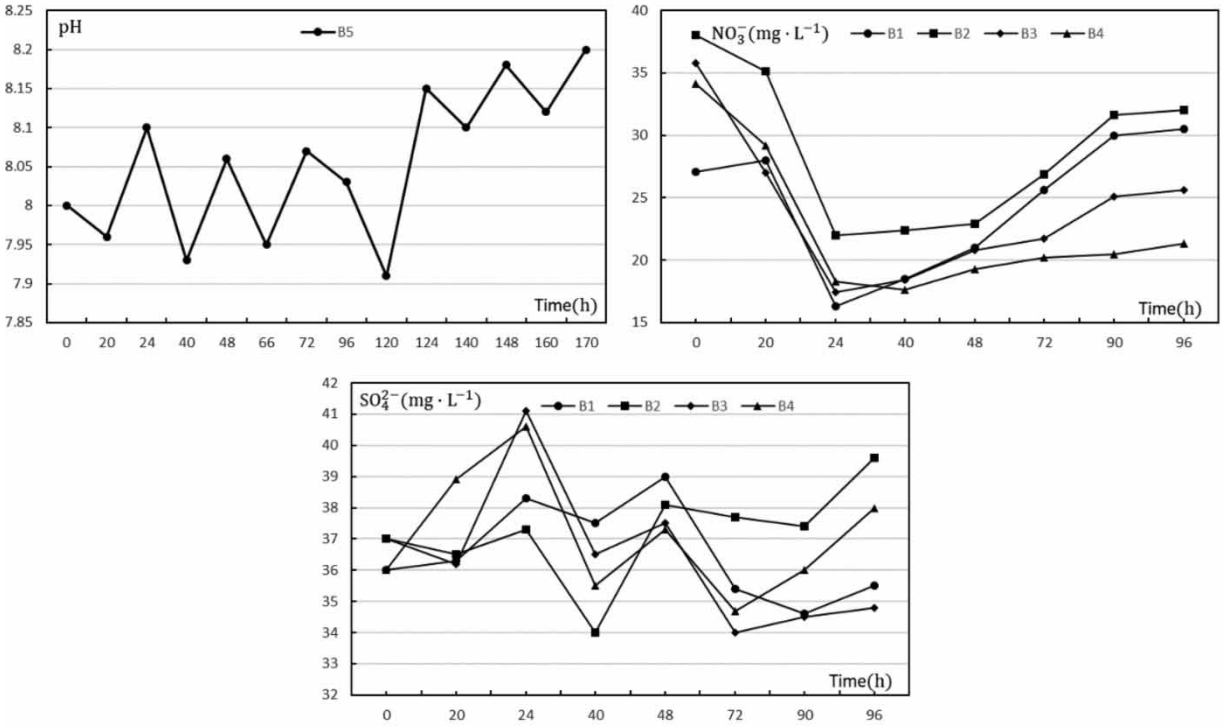


**Figure 4** | Profiles of pH, $NO_3^-$ and $SO_4^{2-}$ in the sand tank experiment.

the low concentration of petroleum pollutants, the microbial degradation is weak, resulting in a minor variation of pH. $NO_3^-$ tends to decrease first and then increase. Since $NO_2^-$ has been detected in some samples, it can be concluded that the decrease of $NO_3^-$ is due to the microbial denitrification in Equation (5), in which $NO_3^-$ was an electron acceptor to be consumed. Generally, the priority of the electron acceptor being consumed is $DO > NO_3^- > Fe^{3+} > SO_4^{2-}$, a small variation of $SO_4^{2-}$ indicated that $NO_3^-$ and $Fe^{3+}$ are not run out. Therefore, the above results show that the petroleum pollutants in sand tank and gas station are affected by a weak microbial degradation. The results can provide a reliable basis for the application of the least squares method identification under a sand tank experiment and a gas station area.

$$\begin{aligned} R - CH_2CH_3 + O_2 &\rightarrow R - CH_2CH_3OH \\ &\rightarrow R - CH_2CHO - R \\ &\rightarrow R - CH_2COOH \end{aligned} \quad (4)$$

$$CH_2O + 2\,NO_3^- \rightarrow H_2O + CO_2 + 2\,NO_2^- \quad (5)$$

## Mathematical model

Based on the hydrogeological conditions of the contaminated site and the pollution history, it is found that the gas station has been discarded since the 1970s. It is concluded that the oil pollution source has stopped leaking. The reason why pollutants can be detected in groundwater at present is that the terrain of the site is flat, the migration speed of groundwater is slow and the pollutants leaked in the early stage are not completely degraded and diffused. The leakage process can be generalized into two stages: in the early stage, pollutants continue to leak from the storage tank; when all the pollutants in the tank leak out, the pollutants migrate and diffuse in the way of point source under the effect of groundwater flow.

Generally, the pollutant transporting through the aquifer is affected by diffusion, advection, dispersion, absorption and degradation. However, this paper focuses on the migration and diffusion process based on two main reasons: As for absorption, it is found in relevant research that the absorbance of nonpolar organic matter is minor in a low-organic-carbon and saturated aquifer

(Perlinger *et al.* 1990). Therefore, we dismissed this process. As for degradation, based on the minor variation of pH, $NO_3^-$ and $SO_4^{2-}$ in the sand tank experiment and a low level of pollutant concentration, we concluded that the degradation is not the main process that influenced pollutants' transport. As a result, the transport of pollutant in sand tank and gas stations can be simplified into a one-dimensional point source with instantaneous leakage and it is mainly affected by advection and dispersion.

Therefore, when the main direction of groundwater flow is set as the positive direction of $x$, the mathematical model can be expressed as Equation (6). Where $C(x, t)$ is the concentration of petroleum pollutant ($ML^{-3}$) at the time of $t$ (T) from the distance $x$ (L), $D_L$ is the dispersion coefficient ($LT^{-1}$), $n$ is the effective porosity, $s$ is the cross-sectional area of the groundwater flow ($L^2$), $M$ is the mass of pollutant (M) and $V$ is the Darcy velocity ($LT^{-1}$).

$$\begin{cases} \dfrac{\partial C}{\partial t} = D_L \dfrac{\partial^2 C}{\partial x^2} - v\dfrac{\partial C}{\partial x} \\ f_{-\infty}^{+\infty} nsCdx = M \\ C(x,t)_{|t=o} = 0 \\ C(x,t)_{|x\to\infty} = 0 \end{cases} \tag{6}$$

The analytical solution of the TPH concentration at $t$ from the distance $x$ can be written as follows:

$$C(x,t) = \frac{M/S}{2n\sqrt{\pi D_L t}} \exp\left[-\frac{(x-vt)^2}{4D_L t}\right] \tag{7}$$

## Hydrogeological parameters values

According to the investigation results, the contaminated aquifer consists mainly of gravel and coarse sand with good water yield. Since there is not any stratigraphic sand lens according to the boreholes data, and the hydrogeological cross-section figure and the landform of the gas station are plain, the surface fluctuation of the aquifer is minor, it is concluded that the groundwater flow's velocity and direction will remain a steady state at least in the study area, and the porous media in the target aquifer can be simplified as homogenous media.

The values of hydrogeological parameters involved in this calculation example are shown in Table 1. For parameters of the sand tank experiment, the hydraulic gradient, the cross-sectional area, the permeability coefficient and the flow rate are all obtained through the test. The porosity and the dispersion coefficient are both obtained from the empirical value associated with the hydrogeological conditions of the site. Moreover, the dispersion coefficient is calculated by the method given by relevant researches (Neuman & Zhang 1990; Gelhar *et al.* 1992).

For parameters of the gas station, as the medium of the phreatic aquifer is large grain gravel, and it is mixed with these coarse sands, the permeability coefficient and porosity are given according to the empirical value. According to the average hydraulic gradient of the site of 1‰ and Darcy's law, the Darcy velocity is calculated. The value of $a_L$ is 5.3 m, and the calculation method of its dispersion coefficient is the same as that in the laboratory. As the average thickness of the phreatic aquifer is 3 m and the width is about 40 m, the cross-sectional area is calculated as 120 $m^2$. Table 1 also shows that the leakage mass $M$, the leakage position $X_0$ and the initial leakage time $T_0$ are the parameters to be inverted by the least squares method.

## Application steps

Figure 5 shows the steps of groundwater pollution sources identification by the least squares method. Firstly, based on the monitoring data of pollutant concentration, it is necessary to determine the mathematical model that the

**Table 1** | Hydrogeological parameters involved in the experimental calculation

| Hydrogeological parameters | Experiment value | Site value |
|---|---|---|
| Hydraulic conductivity (m/d) | 9.78 | 45 |
| Porosity | 0.3 | 0.25 |
| Darcy velocity (m/d) | 0.2139 | 0.045 |
| Dispersion coefficient ($m^2$/d) | 0.5 | 0.2385 |
| Cross-sectional area ($m^2$) | 0.12 | 120 |
| Leak position | Unknown | Unknown |
| Leak mass | Unknown | Unknown |
| Initial leak time | Unknown | Unknown |

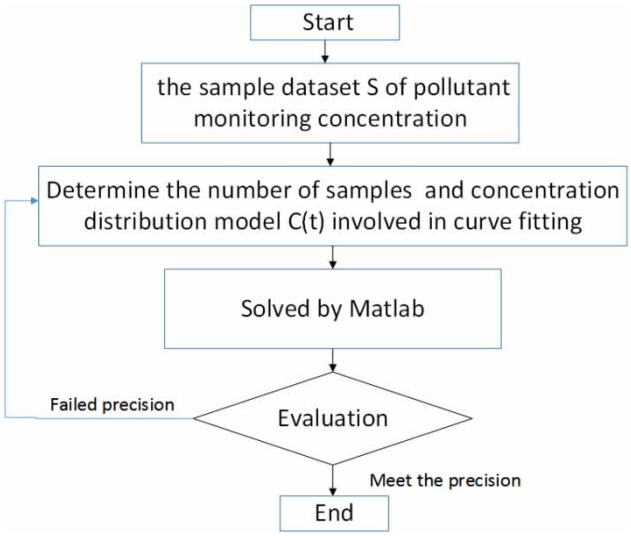


**Figure 5** | Steps of groundwater pollution sources identification by the least squares method.

distribution of pollutant concentration conforms to. According to the hydrogeological conditions of the study area, we should determine the boundary conditions of pollutant and chemical processes occurring during transport. Then, if the fitting error is minor enough, the least squares solution $f(x)$ is obtained with the coefficients related to the pollution source, and the identification process of the pollution source is completed.

Lsqcurvefit function is essential to solve optimization problems, and it is a common tool in nonlinear least squares fitting (Lopez-Luna *et al.* 2019). By using lsqcurvefit function of MATLAB R2018b, the coefficient matrix satisfying the function of one-dimensional point source solute distribution model is obtained. The application steps are as follows:

A. In Equation (7), TPH concentration is taken as the dependent variable, $t$ as the independent variable, TPH mass $M$ and leakage position $X_0$ are taken as the coefficients to be solved, and function files are generated in Matlab;
B. Choose the section at $X = 0.15$ m, and give the function initial value $m$ for 0.2 kg, $t$ for 0.1 d, $X_0$ for 1.00 m;
C. Use the lsqcurvefit function of Matlab, calculate the sum of squared residuals and the fitting results are obtained;
D. Change the data of different section in 0.45, 0.75 and 1.05 m, respectively, repeat the Step C;
E. Compare the calculated value of $M$, $X_0$ and $T_0$ with the observed values of $M$, $X_0$ and $T_0$, evaluate the identification results.

## RESULTS AND DISCUSSION

### Case 1: sand tank experiment

Figure 6 shows the least squares fitting curve in the sand tank experiment. The initial concentration of all sections falls evenly on both sides of the fitted concentration curve. Table 2 shows the fitting results of the least squares in the sand tank experiment. For the 0.45 m section, the leakage position $X_0$ is 0.4610 m, the relative error is 2.44%, the pollutant mass $M$ is 0.6303 kg, the relative error is 5.05%, the initial leakage time $T_0$ is 0.1053 d and the relative error is 5.30%; for the 0.75 m section, the retrieved leakage position $X_0$ is 0.7625 m, the relative error is 1.67%, pollutant mass $M$ is 0.6272 kg, the relative error is 4.53%, the initial leakage time $T_0$ is 0.1025 d and the relative error is 0.62%; for the section at 1.05 m, the calculated leakage position $X_0$ is 1.0565 m, the relative error is 4.40%, the pollutant mass $M$ is 0.6264 kg, the relative error is 5.05%, the initial leakage time $T_0$ is 0.1064 d and the relative error is 6.40%. It seems that better identification results will be obtained when the distance to pollutant is further, which can contribute to the vibration of data when the groundwater flows under a transient condition at the very first pollutant injection time. Overall, the fitting errors of each section are within 6%, which is reasonably minor for the distance, initial leak time and mass. Therefore, it can be concluded that the least squares method has a good effect in identifying the initial leakage location, the mass of pollutants and the time of initial leakage of groundwater pollution sources.

### Case 2: gas station area

The pollution source parameters $M$, $T_0$ and $X_0$ calculated by the least squares method are 11.0 kg, 4,282 d and 35.5 m, respectively. Figure 7 shows the least squares fitting curve in the gas station area. The initial concentration of all sections falls evenly on both sides of the fitted concentration

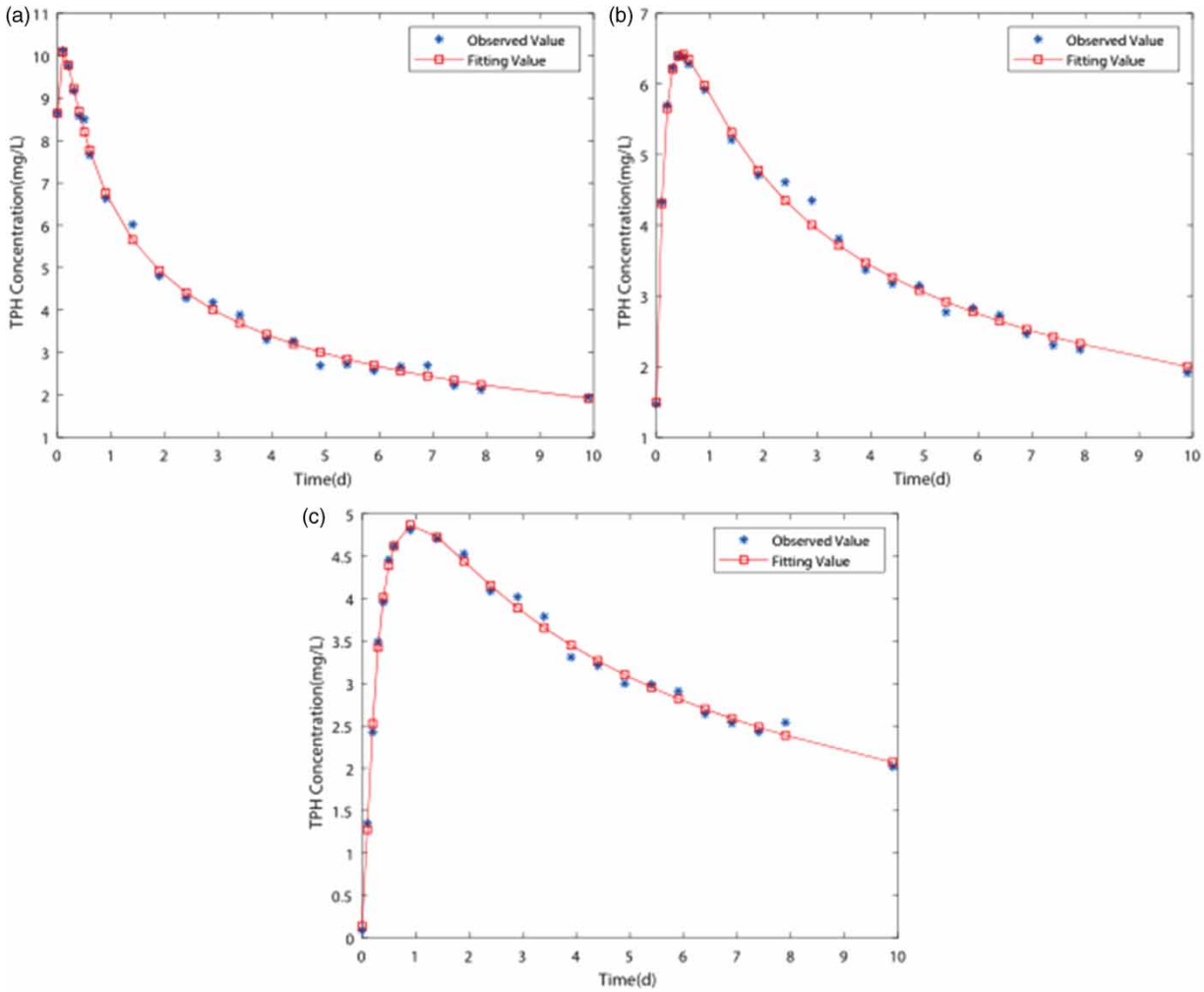


**Figure 6** | Least squares fitting curve in the sand tank experiment. (a) 0.45 m from pollutant; (b) 0.75 m from pollutant and (c) 1.05 m from pollutant.

**Table 2** | Fitting results of the least squares in the sand tank experiment

| Distance to pollutant $X$(m) | Leak position $X_0$(m) | Relative error (%) | Initial leak time $T_0$(d) | Relative error (%) | TPH emission mass $M$(kg) | Relative error (%) | Residual sum of squares res |
|---|---|---|---|---|---|---|---|
| 0.45 | 0.4610 | 2.44 | 0.1053 | 5.30 | 0.6303 | 5.05 | 0.6243 |
| 0.75 | 0.7625 | 1.67 | 0.1025 | 2.50 | 0.6272 | 4.53 | 0.3175 |
| 1.05 | 1.0565 | 0.62 | 0.1064 | 6.40 | 0.6264 | 4.40 | 0.1565 |

curve. According to the early data from the gas station area survey, the identification results of pollution sources are verified and analyzed.

Firstly, the calculation result of pollution source location $X_0$ is 36.35 m, which means that the location of pollution source is 36.35 m away from the upstream of observation data. Taking the groundwater flow direction as the positive $x$ direction, the coordinates of the contaminated site are established, and it can be found that the upstream 36.35 m is just located in the gasoline irrigation area.

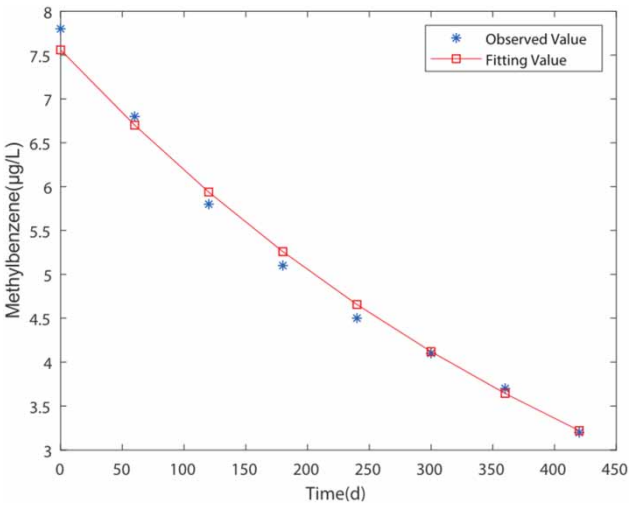


**Figure 7** | Least squares fitting curve in a gas station area.

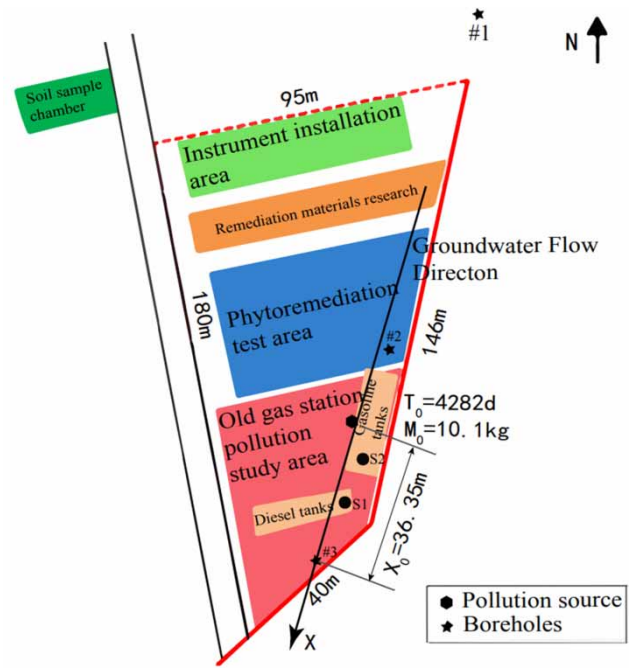


**Figure 8** | Illustration of the identification results in a gas station area.

Figure 8 shows the illustration of the identification results in the gas station area. Therefore, it is believed that $X_0$ is consistent with the observed values.

The fitting result of $T_0$ is 4,282 days. The time interval between pollutant leakage and the first sampling test (July 2017) is 4,282 days (about 12 years); therefore, it can be speculated that the time of leakage is about 2005. The gas station was discarded in the 1970s and the old underground tanks in China used to be made of steel, and the standard specification thickness of the steel is usually 6 mm. Due to its long-term contact with air and water, the thickness of the steel will become thinner due to the corrosion, which will eventually lead to the leakage of oil pollutants to varying degrees. In this paper, the corrosion rate is assumed to be 0.4 mm/year, i.e. the damage to the oil tank takes 15 years. This assumption is also in line with the actual condition. Therefore, the oil tank was in the stage of corrosion damage from 1970 to 1985, and the tank continued to leak for 20 years from 1985 to 2005. In 2005, when all the pollutants in the tank were leaked, the pollutants transported to the downstream. The release history of the tank is shown in Figure 9.

The calculation result of leakage mass of pollutants $M$ is 10.1 kg, which indicates that about 10.1 kg of toluene enters into the aquifer, so the leakage strength of toluene is $1.33\ \mathrm{g\,d^{-1}}$. Therefore, when the steel plate of the oil storage tank is not completely corroded in the early stage, the pollutant leakage intensity is 0; after all the pollutants in the oil storage tank leak out after 2005, the pollution source intensity is 0.

## Application conditions

From the steps of solving the problem, it can be seen that there are three applicable conditions for using the least squares method to identify groundwater pollution sources: (1) the migration model of pollutants in groundwater flow field is obtained accurately. When determining the function type of pollutant migration, it is necessary to determine the mathematical model of pollutant migration in combination with the hydrogeological conditions of the site, so that the

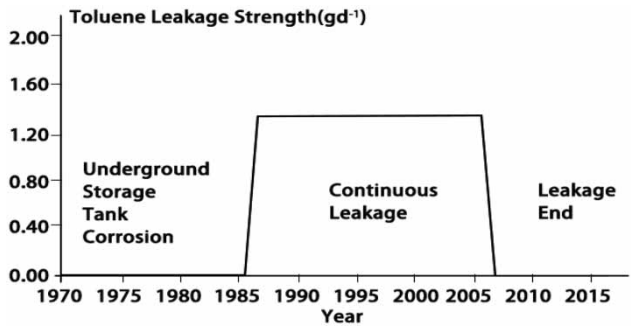


**Figure 9** | Reconstructing release history of the tank in a gas station area.

boundary conditions and the convection dispersion, biodegradation and adsorption process during migration conform to the corresponding mathematical model; (2) the pollutant transport model must have analytical solutions. Because the specific form of the objective function needs to be given in the application of the least squares method, the analytical solution of the model is required, but in practice, there is little idealized migration; (3) the main direction of the pollution source distance section is known. The calculated leakage location is a scalar value, which only represents the distance between the pollution source and a monitoring section. Moreover, the least squares method focuses on the identification of groundwater pollution sources in the main direction of groundwater flow. Although the sand tank is a two-dimensional setup and the site area is a three-dimensional setup, the main flow direction is unique. Therefore, it is often necessary to determine the direction of the pollution source according to the investigation data of the pollution site or judging the groundwater flow.

In addition, the parameters of pollution sources are limited. The mathematical form of the analytic solution function directly determines that the relevant parameters of the pollution source are limited. In the calculation example, the parameters such as the initial time of pollution source leakage, the distance between the leakage point and the monitoring section, and the initial leakage concentration are obtained. As for the flux of pollution source in a certain section, the leakage mode and the type of pollutants still need to be determined by other means.

The requirements of the migration model, the idealized assumption of the analytical solution, the determination of the distance between the pollution sources and the limited parameters of the solution show the limitations of the least squares method in the identification of groundwater pollution sources. In practical application, information such as groundwater flow direction, general direction of pollutants and possible types of pollutants should be quickly determined by means of data collection, field survey and other investigation means. Then, combined with the concentration distribution data of the monitoring section, the mathematical model of pollutant migration should be analyzed and the analytical solution of the corresponding model should be solved. Finally, the least squares method can be applied to identify pollution source-related parameters. For parameters that are not considered in the mathematical model, they need to be confirmed in combination with other pollution source identification methods.

## CONCLUSIONS

This paper explores the application of the least squares method in the identification of groundwater pollution sources. Based on the theory of groundwater solute transport, the process of using the least squares method to identify groundwater pollution sources is about finding out the relevant parameters of pollution sources combined with the mathematical model of groundwater solute transport multivariate nonlinear regression.

In the laboratory test of TPH, different cross-section observation data are selected to solve the parameters of initial leakage time, leakage location and the mass of TPH pollution source. It is found that the least squares method has good fitting effect in identifying pollutant mass, initial leakage location and initial leakage time of groundwater pollution source, and the fitting errors of each cross-section are within 6%. Furthermore, the identification results of toluene in an old gas station aquifer show that the initial time of leakage, the location of leakage and the mass of pollutants were nearly consistent with the observed values, which indicates that the least squares method has application values to the identification of groundwater pollution sources.

The use of this method covers under the pollution scenario of instantaneous leakage of one-dimensional point source solute transport process without considering further processes of adsorption and biodegradation during the transport process. The applicable conditions are more ideal and have certain limitations. In the future, we can try to discuss the application of the least squares method in two-dimensional or three-dimensional groundwater solute transport, or compare the results with other groundwater pollution source identification methods.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

## REFERENCES


Aravena, R., Evans, M. L. & Cherry, J. A. 1993 Stable isotopes of oxygen and nitrogen in source identification of nitrate from septic systems. *Ground Water* **31** (2), 180–186.

Ayvaz, M. T. 2016 A hybrid simulation–optimization approach for solving the areal groundwater pollution source identification problems. *Journal of Hydrology* **538**, 161–176.

Borah, T. & Bhattacharjya, R. K. 2016 Development of an improved pollution source identification model using numerical and ANN based simulation-optimization model. *Water Resources Management* **30** (14), 5163–5176.

Chang, S. Y. & Kashani, F. R. 2009 Linear programming method for investigating the disposal histories and locations of pollutant sources in an aquifer. *Journal of Environmental Informatics* **13** (1), 1–11.

Dimov, I., Jaekel, U., Vereecken, H. & Wendt, D. 1998 A numerical approach for determination of sources in reactive transport equations. *Computers & Mathematics with Applications* **32** (5), 31–42.

Gelhar, L. W., Welty, C. & Rehfeldt, K. R. 1992 A critical review of data on field-scale dispersion in aquifers. *Water Resources Research* **28** (7), 1955–1974.

Gu, W. L., Lu, W. X., Zhao, Y., Quyang, Q. & Xiao, C. N. 2017 Identification of groundwater pollution sources based on a modified plume comparison method. *Water Science and Technology - Water Supply* **17** (1), 188–197.

Huang, L. X., Wang, L. C., Zhang, Y. Y., Xing, L. T., Hao, Q. C., Xiao, Y., Yang, L. Z. & Zhu, H. H. 2018 Identification of groundwater pollution sources by a SCE-UA algorithm-based simulation/Optimization model. *Water* **10** (2), 193.

Koch, J. & Nowak, W. 2016 Identification of contaminant source architectures – a statistical inversion that emulates multiphase physics in a computationally practicable manner. *Water Resources Research* **52** (2), 1009–1025.

Li, C., Li, S. L., Yue, F. J., Liu, J., Zhong, J., Yan, Z. F., Zhang, R. C., Wang, Z. J. & Xu, S. 2019 Identification of sources and transformations of nitrate in the Xijiang River using nitrate isotopes and Bayesian model. *Science of the Total Environment* **646**, 801–810.

Lopez-Luna, J., Ramirez-Montes, L. E., Martinez-Vargas, S., Martinez, A. I., Mijangos-Ricardez, O. F., Gonzalez-Chavez, M. D. A., Carrillo-Gonzalez, R., Solis-Dominguez, F. A., Cuevas-Diaz, M. D. & Vazquez-Hipolito, V. 2019 Linear and nonlinear kinetic and isotherm adsorption models for arsenic removal by manganese ferrite nanoparticles. *SN Applied Sciences* **1** (8), UNSP950.

Mahar, P. S. & Datta, B. 2000 Identification of pollution sources in transient groundwater systems. *Water Resources Management* **14** (3), 209–227.

Mansuy, L., Philp, R. P. & Allen, J. 1997 Source identification of oil spills based on the isotopic composition of individual components in weathered oil samples. *Environmental Science & Technology* **31** (12), 3417–3425.

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** (8), W08302.

Neuman, S. P. & Zhang, Y. K. 1990 A quasi-linear theory of non-Fickian and Fickian subsurface dispersion: 1. Theoretical analysis with application to isotropic media. *Water Resources Research* **26** (5), 887–902.

Perlinger, J. A., Eisenreich, S. J., Capel, P. D., Carr, P. W. & Park, J. H. 1990 Adsorption of a homologous series of alkylbenzenes to mineral oxides at low organic carbon content using headspace analysis. *Water Science & Technology* **22** (6), 7–14.

Pirot, G., Krityakierne, T., Ginsbourger, D. & Renard, P. 2019 Contaminant source localization via Bayesian global optimization. *Hydrology & Earth System Sciences* **23** (1), 351–369.

Qian, H., Zhang, Y. L., Wang, J. L., Si, C. Q. & Chen, Z. X. 2018 Characteristics of petroleum-contaminated groundwater during natural attenuation: a case study in northeast China. *Environmental Monitoring & Assessment* **190** (2), 80.

Rojas, R., Feyen, L. & Dassargues, A. 2008 Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research* **44** (12), W12418.

See, J. J., Jamaian, S. S., Salleh, R. M., Nor, M. E. & Aman, F. 2018 Parameter estimation of Monod model by the Least Squares method for microalgae, *Botryococcus Braunii* sp. *Journal of Physics: Conference Series* **995**, UNSP012026.

Skaare, B. B., Schaanning, M. & Morkved, P. T. 2009 Source identification for oil-based drill cuttings on the seabed based on stable carbon isotopes. *Environmental Chemistry Letters* **7** (2), 183–189.

Skute, A., Osipovs, S. & Vardanjans, D. 2016 Source identification of nitrate by means of stable nitrogen isotopes in the river Daugava and loads of nitrogen to the Gulf of Riga. *International Journal of Environmental and Analytical Chemistry* **96** (1), 1–14.

Wang, J. B. & Zabaras, N. 2006 A Markov random field model of contamination source identification in porous media flow. *International Journal of Heat and Mass Transfer* **49** (5–6), 939–950.

Xia, X. M., Zhou, N. Q., Wang, L. C., Li, X. W. & Jiang, S. M. 2019 Identification of transient contaminant sources in aquifers through a surrogate model based on a modified self-organizing-maps algorithm. *Hydrogeology Journal* **27** (7), 2535–2550.

Yeh, H. D., Lin, C. C. & Yang, B. J. 2014 Applying hybrid heuristic approach to identify contaminant source information in transient groundwater flow systems. *Mathematical Problems in Engineering* **2014** (5), 1–13.

Zhao, Y., Lu, W. X. & Xiao, C. N. 2016 A Kriging surrogate model coupled in simulation–optimization approach for identifying release history of groundwater sources. *Journal of Contaminant Hydrology* **185**, 51–60.


First received 17 June 2020; accepted in revised form 10 December 2020. Available online 12 January 2021