Identification of groundwater pollution sources based on a modified plume comparison method

Wenlong Gu, Wenxi Lu, Ying Zhao, Qi OuYang and Chuanning Xiao

ABSTRACT

The identification of groundwater contaminant sources is a primary step in designing and implementing a remediation strategy. The work presented here was undertaken to develop an efficient strategy that addresses the unknown multiple contaminant sources problem, and that could identify the number, location and magnitude of the groundwater contaminant sources and select optimal sampling locations. A Monte Carlo approach was used first to obtain the statistical characteristics of groundwater flow and transport model. Then the linear Kalman filter and a modified comparison method were utilized to update the estimation of concentration values and source weights, which represent the similarity between the estimated composite plume and each individual plume. Moreover, an optimization method was employed to identify the magnitude of contamination and the optimal sampling location. All of these steps were repeated until the weights stabilized and converged. A synthetic example was used to test the strategy and a further uncertainty analysis was conducted. The evaluation demonstrated that the strategy effectively addresses unknown multiple-source problems, under the condition that the error of concentration measurement value was controlled to less than 10%, and the time error was controlled to less than 6%.

Key words | groundwater, inverse problem, plume comparison, pollution sources

Wenlong Gu
Wenxi Lu (corresponding author)
Ying Zhao
Qi OuYang
Chuanning Xiao
Key Laboratory of Groundwater Resources and Environment,
Ministry of Education, Jilin University,
Changchun 130021,
China
and
College of Environment and Resources,
Jilin University,
Changchun 130021,
China
E-mail: *luwenxi@jlu.edu.cn*

INTRODUCTION

Identifying the characteristics of groundwater contaminant sources is not only a major task before adopting a feasible remediation strategy, but also a general basis for how to apportion the remediation duties and costs from a legal and regulatory point of view (Butera *et al.* 2013). Due to the complexities of any groundwater system, it is always difficult to obtain all the useful information and utilize it to solve the inverse problem perfectly.

Characteristics of contaminant sources mainly include the number, location, magnitude and contaminant release history. For single-source inverse problems, researchers usually focus on identifying the location and release history of the contaminant source, and have achieved fruitful results. Skaggs & Kabala (1994) used the Tikhonov regularization method to recover the release history of a known groundwater contaminant source. Neupauer *et al.* (2000) used both Tikhonov regularization and the minimum relative entropy method to reconstruct the release history of a known source. Michalak & Kitanidis (2004) employed the adjoint state method to identify the source location. Wang & Jin (2013) adopted the Bayesian method to infer the location and magnitude of contaminant sources in three dimensions. Jha & Datta (2015) identified the location, starting release time and active duration of a source based on the dynamic time warping method. Nevertheless, the most common problems encountered in practice are those in which the number of contaminant sources is unknown. So a method tailored to complex and challenging multiple-source problems would be of considerable practical significance. Singh *et al.* (2004) used artificial neural networks to

doi: 10.2166/ws.2016.122

identify the location and release history of contaminant sources based on two synthetic examples. Jha & Datta (2013) identified the locations and release history of contaminant sources in a three-dimensional case based on an adaptive simulated annealing algorithm. Gurarslan & Karahan (2015) proposed a differential evolution algorithm to characterize the locations and release history of groundwater contaminant sources.

However, when solving multiple-source problems, these optimal methods coupled with a heuristic algorithm and the methods based on artificial neural networks may take too much time, or may not be stable. Dokou & Pinder (2009, 2011) proposed a new approach for identifying groundwater contaminant sources based on the Kalman filter and fuzzy set theory. The algorithm could search for a reasonable solution under a single contaminant source case, but was unsuitable for the multiple-source case because the final weights of the true sources could not converge to 1 simultaneously (one weight was 1.0 and the other was 0.8, although both sources were 'true' sources). The work presented here was conducted to extend and improve the earlier algorithm developed by Dokou & Pinder (2009, 2011). A modified comparison method for fuzzy sets and a means of describing an 'equivalent source' was developed that would be more efficient and robust for solving the unknown source identification problems, whether for a single-source case or a multiple-source case.

METHODOLOGY

A basic assumption of this research is that a solute is conservative. That is to say, the relation between the source's solute input and the observed concentration can be regarded as approximately linear (Snodgrass & Kitanidis 1997; Butera *et al.* 2013). In this research, several tools and methods were employed. The main process and solution steps are shown in Figure 1.

Monte Carlo approach

Considering the uncertainty of a groundwater system, which is due to the heterogeneity of hydraulic conductivity, a Monte Carlo approach was used to statistically describe the distribution of contaminant concentration. This approach assumes that the hydraulic conductivity follows a log-normal distribution. First, 100 hydraulic conductivity fields were generated (Zhang & Pinder 2003), and each of them was used to calculate the concentration field for each potential source. Then the mean value of 100 concentration fields of each potential source was identified (called the 'individual plume'), and used to calculate the weighted average concentration field (called the 'composite plume') and the initial variance–covariance matrix.

Linear Kalman filter

The linear Kalman filter was first proposed for addressing prediction problems in communication and control (Kalman 1960). When employed in contaminant source identification (Dokou & Pinder 2009, 2011), the main steps that were used to update the state variable and error covariance matrix are as follows.

Compute Kalman gain:

$$K = P^{-}H^{T}(HP^{-}H^{T} + v)^{-1} \tag{1}$$

Update the estimated concentration using measurement z:

$$\tilde{c}^{+} = \tilde{c}^{-} + K(z - H\tilde{c}^{-}) \tag{2}$$

Update the error covariance matrix:

$$P^{+} = (I - KH)P^{-} \tag{3}$$

In Equations (1)–(3), K is the Kalman gain matrix; P is the error covariance matrix; H is the sampling matrix with dimension $l \times n$, whose element is 1 when a specific position is taken as a sampling point, otherwise the elements are zero; v is the covariance of measurement error; $\tilde{c}$ is the estimate of contaminant concentration; and z is a vector of l noise-corrupted measurements. In this analysis, $l = 1$ because just one sampling data point was used at one time, and '$-$' denotes a prior estimate and '$+$' denotes a posterior estimate.

For the initial error covariance matrix $P0$, the element in the ith row and jth column can be calculated using

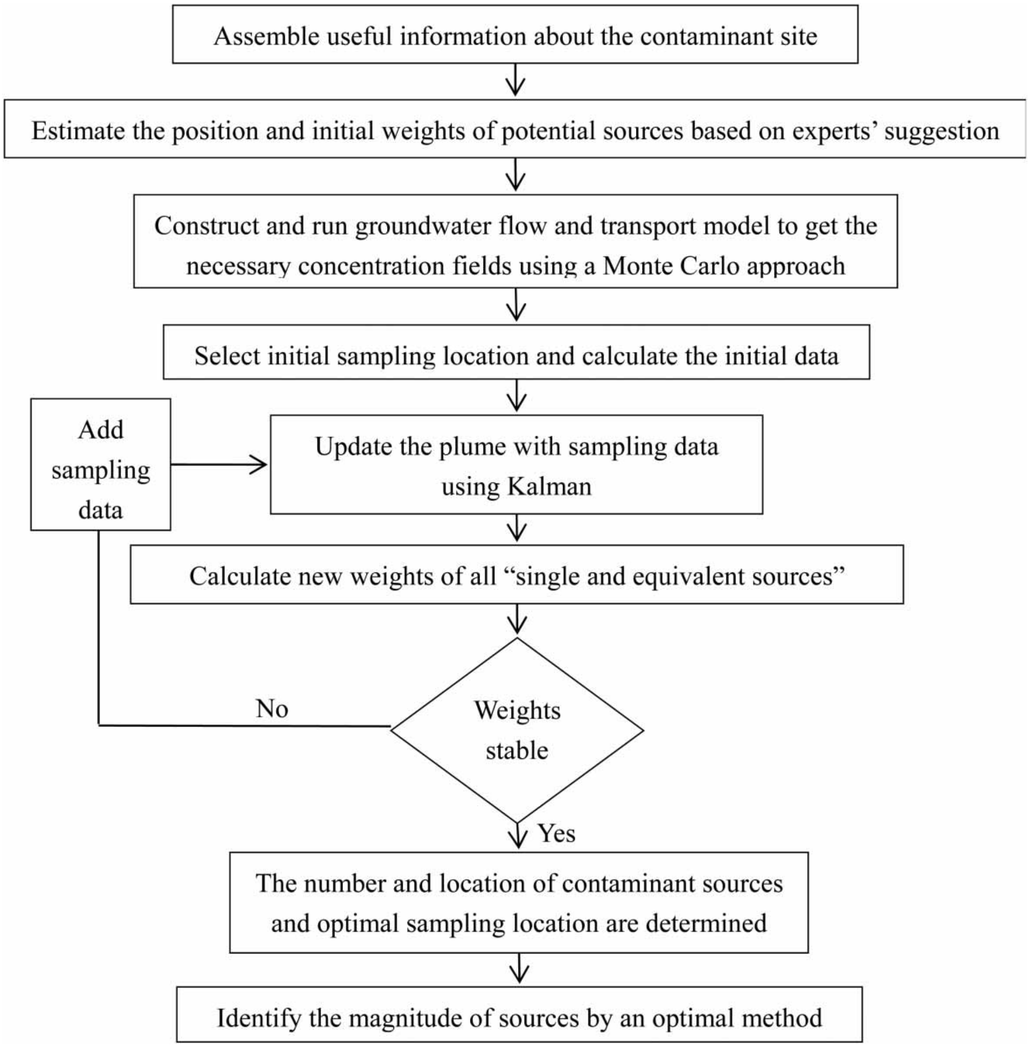
Assemble useful information about the contaminant site
Estimate the position and initial weights of potential sources based on experts' suggestion
Construct and run groundwater flow and transport model to get the necessary concentration fields using a Monte Carlo approach
Select initial sampling location and calculate the initial data
Add sampling data
Update the plume with sampling data using Kalman
Calculate new weights of all "single and equivalent sources"
No
Weights stable
Yes
The number and location of contaminant sources and optimal sampling location are determined
Identify the magnitude of sources by an optimal method

Figure 1 | Flow chart of methodology.

Equation (4):

$$R(C_i, C_j) = \frac{1}{q-1}\sum_{k=1}^{q}(C_{i,k} - \bar{C}_i)\cdot(C_{j,k} - \bar{C}_j) \tag{4}$$

where q is the total number of hydraulic conductivity realizations (equal to 100), $C_{i,k}$ and $C_{j,k}$ denote the ith and jth individual plume concentration generated by the kth hydraulic conductivity realization, and $\bar{C}_i$ and $\bar{C}_j$ denote the ith and jth composite plume concentrations, respectively.

In the analysis, a means of identifying an 'equivalent source' was proposed. For example, assume there were three potential sources named S_1, S_2 and S_3. Besides the three single individual contaminant plumes, all combinations of the three potential sources should also be calculated. In other words, for a scenario consisting of three potential sources, three 'single sources' and four 'equivalent sources' should be characterized.

Fuzzy mathematics and implementation

A modified comparison technique based on fuzzy mathematics was proposed to determine the weights of all 'single sources' and 'equivalent sources' during the Kalman filter process. First, all concentration fields were represented as a fuzzy set, whose elements were defined as the normalized concentration values. For example, if there was a concentration field such as c, the normalized c becomes c' as shown below.

$$c = \begin{bmatrix} 2 & 3 & 2 \\ 3 & 5 & 4 \end{bmatrix} \quad c' = \begin{bmatrix} 0.4 & 0.6 & 0.4 \\ 0.6 & 1.0 & 0.8 \end{bmatrix}$$

$$c'' = \begin{bmatrix} \mathrm{E} & 0.6 & \mathrm{E} \\ 0.6 & \mathrm{E} & \mathrm{E} \end{bmatrix}$$

Then a λ-cut method was used to compare the fuzzy sets and get the new weights. The original λ-cut was a crisp set that contains all the elements of a fuzzy set, whose membership degrees were greater than or equal to the λ. However, in unknown contaminant source problems where more than one true source exists, the 'equivalent individual plume' that emanates from the sources must be larger than the 'single individual plumes'. When comparing the common area with the 'composite plume' by the original λ-cut method, the resulting 'equivalent individual plume' would be dominant and distorted. To solve this problem, a modified λ-cut method was proposed, in which λ is no longer a specific value, but rather a specific interval. If set $\lambda = [0.5, 0.6]$ (for which '[ ]' denotes a closed interval), the matrix c' after being operated becomes c'', as illustrated above in which 'E' denotes an empty set.

The degree of similarity between each individual plume and composite plume would be assigned as the new weights in the next phase of the updating process after being normalized. Several λ ranges (if the number is m) were used to calculate the corresponding degree of similarity Si. The final global degree g could be obtained using Equation (5), which provides a measure of how similar the two plumes were, as well as ensuring that higher concentration values were weighted more than lower concentration values.

$$g = \sum_i \lambda_i S_i \quad i = 1, 2, \cdots, m \tag{5}$$

Optimal method for magnitude

The optimal problem can be described using Equations (6) and (7):

$$min \sum_i |C_i - Z_i| \quad i = 1, 2, \cdots, n \tag{6}$$

$$0 \le S_j \le 100 \quad j = 1, 2, \cdots, k \tag{7}$$

where C_i is the modelled concentration value at sampling position i, Z_i is the measured concentration value at the ith position, S_j is the magnitude of the jth source, n is the total number of all sampling positions, and k is the total number of potential sources.

For conservative solute transport, the concentration value at each sampling location can be expressed as Equation (8):

$$C_s = \omega_j C_{0j,s} \tag{8}$$

in which $C_{0j,s}$ is the concentration at sampling location S under a unity injection for source j, and ω_j is a coefficient.

Incorporating Equation (6) into Equation (8) yields the best-fitted values of ω. The magnitude of a contaminant source is then equal to the product of ω and the unity injection value.

Selection of sampling positions

A proper monitoring network design can provide much more useful information at a lower cost than can a poorly designed network. Based on previous research (Herrera & Pinder 2005), a statistical framework method was used to search the optimal sampling positions. If some initial measured data and a set of weights were determined, such as α, β and γ (for three potential sources), the weighted mean of the individual plume with the jth hydraulic conductivity field C_j and the composite plume $\tilde{C}$ could be expressed as Equations (9) and (10), respectively.

$$C_j = \alpha C_{1,j} + \beta C_{2,j} + \gamma C_{3,j} \tag{9}$$

$$\tilde{c} = \frac{1}{n} \sum_j C_j \, j = 1, 2, \cdots, n \tag{10}$$

In Equations (9) and (10), $C_{1,j}$, $C_{2,j}$ and $C_{3,j}$ denote the three individual plumes that emanate from each potential source with the jth hydraulic conductivity random field, where n is the number of the hydraulic conductivity random field.

The covariance between each C_j and $\tilde{c}$ under the current weighted value was the basic criterion for determining the optimal sampling position because it could reflect the uncertainty of the groundwater system. To simplify the calculations, the variance of C_j was used instead of the covariance between each C_j and $\tilde{c}$. The position with the highest value of variance was chosen as the optimal sampling position.

Case study

To demonstrate the flexibility of the proposed strategy, a synthetic example was developed. The example and data used were as realistic as possible and based on practical experience. The task in the hypothetical scenario was to identify the number, location and magnitude of the true contaminant source (or sources).

A 2-D steady-state flow system was considered, as illustrated in Figure 2, which had a heterogeneous unconfined aquifer. The aquifer is 800 m long and 500 m wide with linearly varying heads forming boundaries along the left and right sides of the aquifer and confining boundaries along the other sides. The active period of the true sources of contamination is known to be 4 years, and the pollutant is injected through wells. The parameters of the aquifer are summarized in Appendix A (available with the online version of this paper). The true source locations are shown in Figure 2, and the simulated contaminant concentrations with some noise (the absolute value of which is less than 3%) were treated as the sampling data.

In this scenario, three potential source locations were determined as the result of a primary investigation. Each hydraulic conductivity random field was run for all sources individually. Because the synthetic example represented a typical real-world situation in which there was no exact information about initial weights, an assumption was made that all initial weights were equal to 0.5. The left boundary of λ used to calculate the degree of similarity of two plumes is equal to the number ranges from 0.1 to 0.8, with an interval of 0.05. Each right boundary is equal to the left boundary plus a value of 0.1. For example, the first three ranges were [0.1, 0.2], [0.15, 0.25] and [0.2, 0.3].

RESULTS AND DISCUSSION

Results

The main updating process of the strategy is shown as Appendix B (available with the online version of this paper), and a more

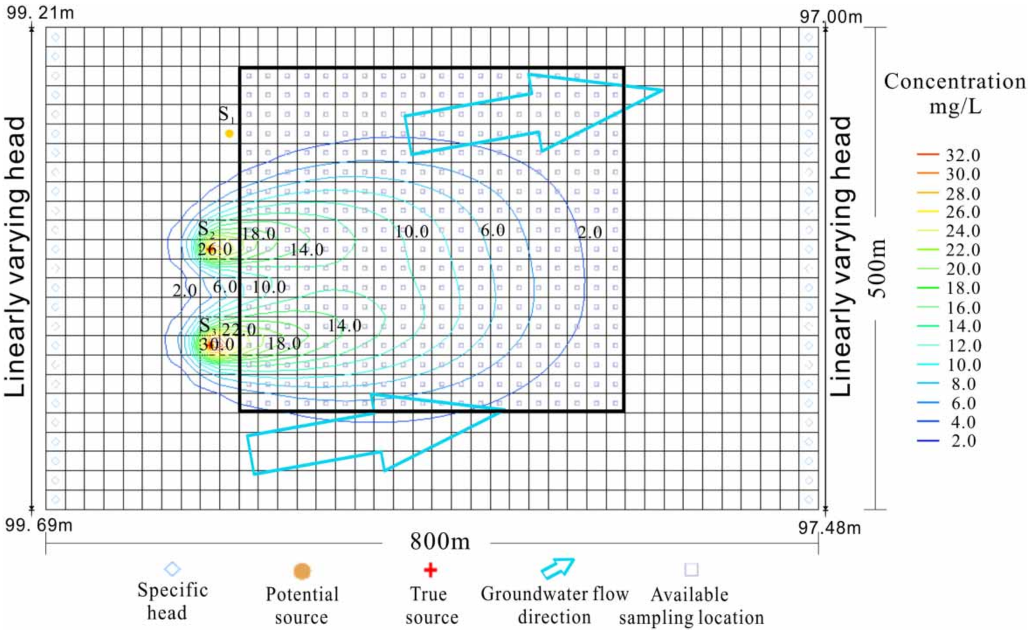

Figure 2 | Conditions of the synthetic aquifer.

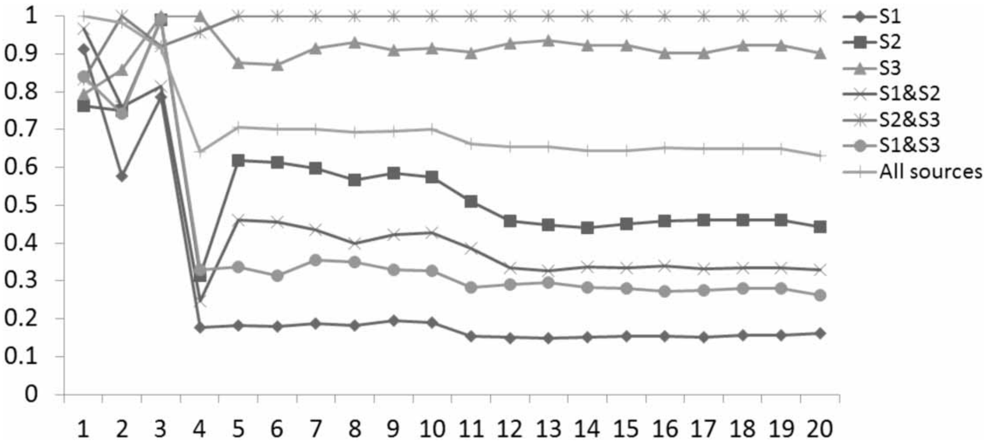

Figure 3 | Updated source weights as a function of different numbers of samples and two true contaminant sources (S2&S3).

focused result is shown in Figure 3. The source weights were interlaced and irregular during the first five samples, and changed slowly in the sixth to 12th samples. After the 13th sample was used, the weights became stable and converged. However, the weight of source S3 (remaining at approximately 0.9) obviously was not reduced by the algorithm, in contrast to other source weights. The main cause of this phenomenon may have two aspects: one is that the updated concentration field was not as accurate as the model solution (comparing Figure 2 with Figure B1(g), Appendix B). The other is that the plumes emanating from S3 and S2&S3 are somewhat similar, which makes an accurate comparison difficult. In spite of this imperfection, the source locations could be determined because the weighting curve was not irregular.

Then an optimal method as described under 'Selection of sampling positions' was used to determine the best estimated value of the magnitude. The mean relative error (MRE) of the estimated magnitude was 9.33%, indicating that the solution was not ideal. To analyse the reason for the poor solution, we calculated the objective function value using information for the true sources based on the forward model, which is greater than the objective function value obtained based on the estimated magnitude. This shows that the primary error is not caused by the algorithm, but rather because the linear assumption is not accurate enough for this example.

A simpler example with only one true contaminant source (S2, Figure 2) was also tested. The composite plume converged when the 13th samples were taken, and the MRE of the estimated magnitude was 5.83%. The variation of the weights of different sources as a function of the number of samples is shown in Figure 4.

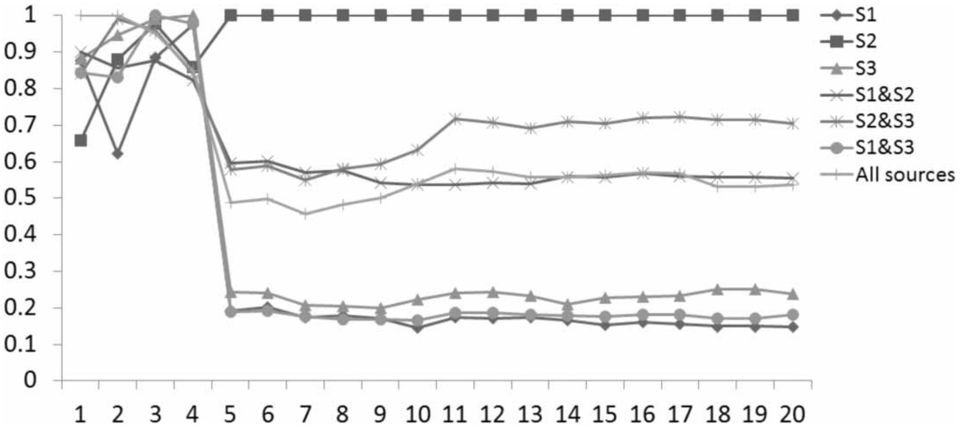

Figure 4 | Updated source weights as a function of different numbers of samples and one true contaminant source (S2).

Uncertainty analysis

For inverse problems such as identifying groundwater contaminant sources, the uncertainty of the solution mainly derives from two aspects: one is randomness and the other is fuzziness. Randomness is typically quantified by the measurement error among various terms, such as hydrogeological parameters and concentration values. In contrast, fuzziness always occurs in the conceptualization or decision process, which involves some anthropic factors. In most instances, randomness and fuzziness cannot be clearly divided.

The random hydraulic conductivity field and a Monte Carlo approach were employed to reduce the randomness of hydrological parameters. Thus, in our analysis, the noise of concentration measurement and the estimation of activity periods of the sources were focused. The example we next describe is the multiple contaminant sources scenario presented previously.

Concentration measurement noise

The noise of concentration measurement was assumed to be a random number whose absolute value was less than 3% (included), as described above in the 'Case study' section. In the analysis, four additional criteria (i.e. noise levels) were used to test the strategy (5%, 10%, 15% and 20%), as shown in Figure 5.

Based on Figure 5, when the absolute value of noise was less than 5%, the updated weights stabilized and converged, and the estimated contaminant source location could be determined as S2&S3. When the noise criterion became 10%, the true sources could also be identified to be S2&S3. However, when the noise index rose to 15%, the location of the sources could not be identified confidently, and when the absolute value of noise extended to 20%, the identification of sources was irregular and could not provide any useful identification information. The accuracy of the estimated magnitude values changed little. All of the MRE for these estimates were approximately 9% to 10%, indicating that the observed data were not so sensitive to the injection rate because the active period of the sources was 4 years.

Noise of active period

In fact, when a parameter such as the active period is known, a source's location would also be known (in theory). However, it is difficult to determine this term accurately at first from a primary investigation. Thus, testing the strategy using different durations of noise during the active period is necessary to illustrate the strategy's practical applicability.

Four different durations of noise (30 d, 61 d, 92 d and 183 d) were used to examine the effect of this variable on source weights (i.e. source identification), and the results are shown in Figure 6. A similar pattern was observed in these tests as with tests of different measurement errors. When the error duration was 30 d and 61 d (Figure 6(a) and 6(b)), the updated source weights stabilized after several samples were used. When the time error increased to 92 d, some overlap in parts of the weighting curves for S3 and for S2&S3 began to appear, and when the error duration was 183 d, the procedure could hardly differentiate among sources accurately. In the estimation of magnitude, the MRE ranged from 8% to 10% and exhibited little fluctuation as well.

Based on the analysis above, for the examples tested, the proposed strategy can identify the true sources of contamination within an acceptable margin of error. For the concentration measurement, the absolute value of the error should be controlled at 10% or less, and the time error, which affected the results to a greater extent, should be controlled to less than 6% (i.e. 3 months within 4 years). However, the MRE of the estimated magnitude was affected little, no matter what and how large the error was.

CONCLUSIONS

(1) The proposed modified strategy for identifying unknown sources of groundwater contamination can successfully identify the sources no matter how many sources exist or where in a study area they are located.
(2) To guarantee the reliability of a solution, the error of concentration measurement should be controlled at 10% or less, and the time error, which affected the results more

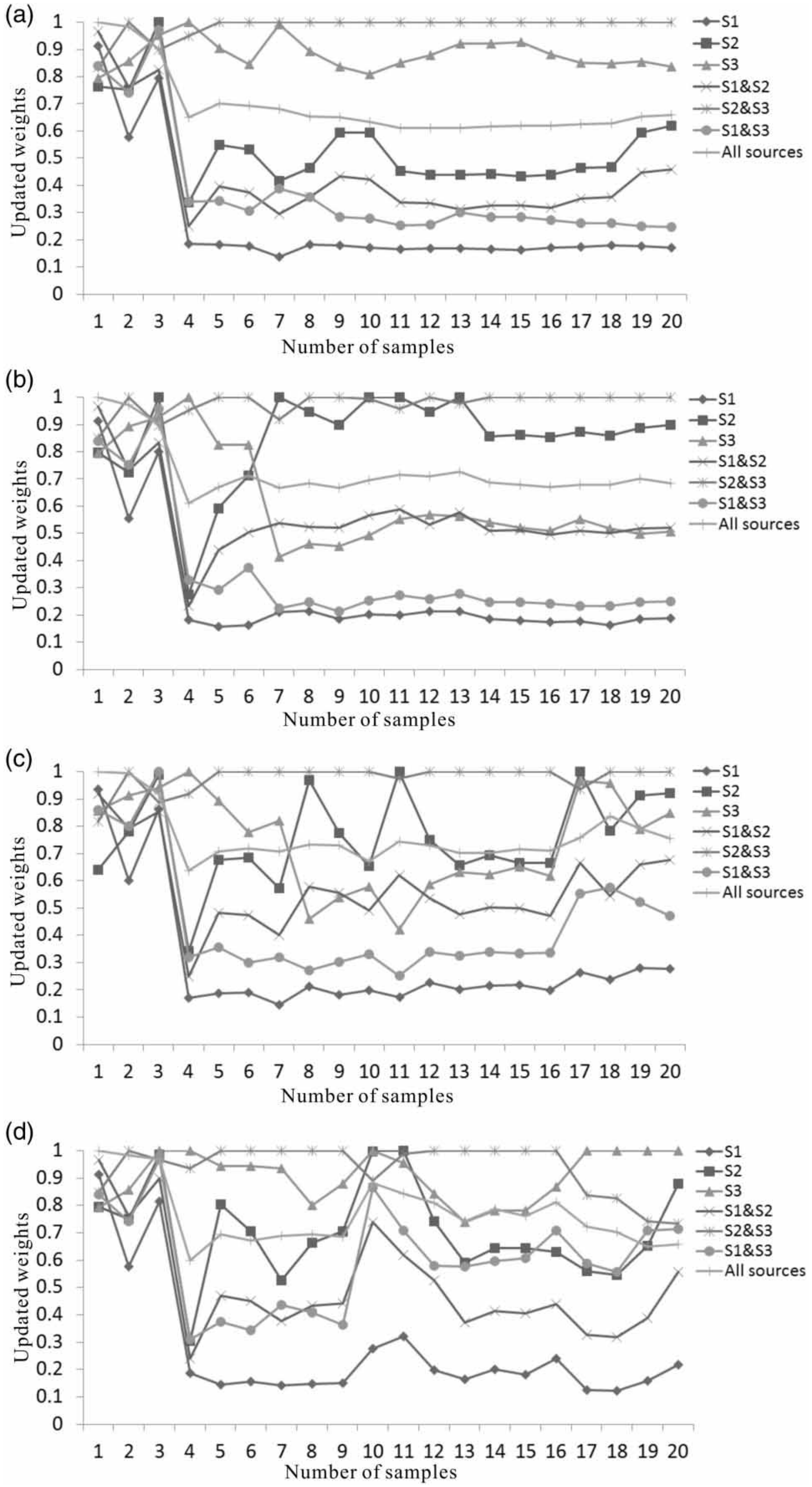

Figure 5 | Updated source weights as a function of number of samples with (a) less than 5% noise, (b) less than 10% noise, (c) less than 15% noise, and (d) less than 20% noise. There were two true contaminant sources (S2 and S3).

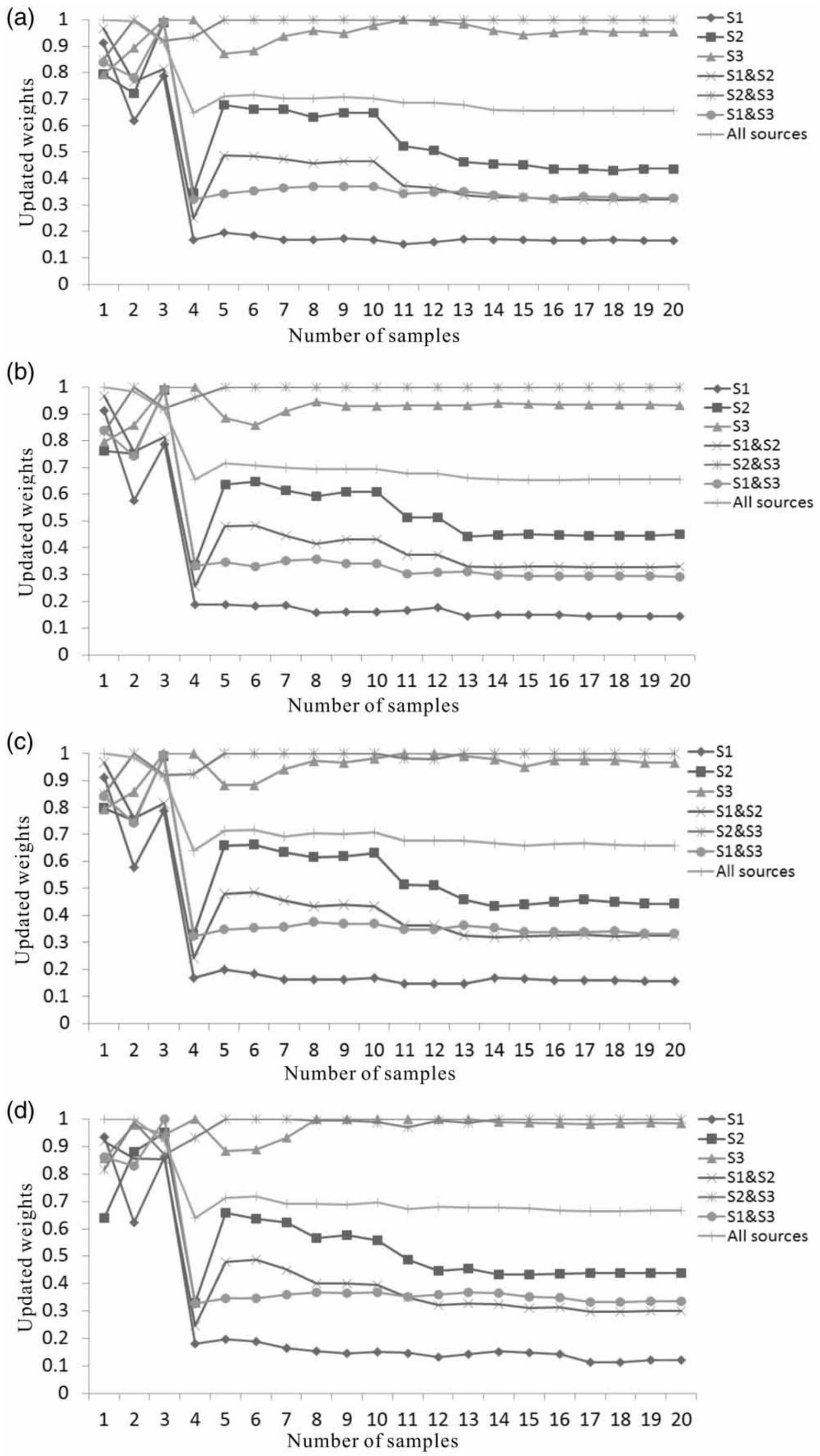

Figure 6 | Updated source weights with a time error of (a) 30 days, (b) 61 days, (c) 92 days, and (d) 183 days.

than did the concentration measurement error, should be controlled at less than 6%.

(3) With a MRE of 5.83%, the solution of a single contaminant source case was more accurate than the solution for multiple sources (MRE = 9.33%), and the MRE values changed little within an acceptable error range.

The strategy presented here has two major advantages over existing analytical techniques. One is that it includes a novel means of describing an 'equivalent source'. This feature makes it possible to identify the location of 'true' contaminant sources using the final source weights directly, rather than having to base identification on a series of slightly lower weights. The second advantage is that the strategy includes a modified plume comparison method to calculate the degree of similarity between two contaminant plumes. This modified method is a more consistent approach to solving the practical problem of identifying unknown contaminant sources than are existing techniques.

However, the linear assumption used in the modified strategy perhaps is not accurate enough for estimating the magnitude of contamination. This potential shortcoming could be the focus of future research in a follow-up study to improve the strategy.

ACKNOWLEDGEMENTS

This research was supported by the National Natural Science Foundation of China (41372237) and the Program of the Environmental Protection Department of Jilin Province (2015-11). We thank the editor and reviewers for their insightful comments and suggestions.

REFERENCES

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 Assessment* **27** (5), 1269–1280.

Dokou, Z. & Pinder, G. F. 2009 Optimal search strategy for the definition of a DNAPL source. *Journal of Hydrology* **376** (3–4), 542–556.

Dokou, Z. & Pinder, G. F. 2011 Extension and field application of an integrated DNAPL source identification algorithm that utilizes stochastic modeling and a Kalman filter. *Journal of Hydrology* **398** (3–4), 277–291.

Gurarslan, G. & Karahan, H. 2015 Solving inverse problems of groundwater-pollution-source identification using a differential evolution algorithm. *Hydrogeology Journal* **23** (6), 1109–1119.

Herrera, G. S. & Pinder, G. F. 2005 Space-time optimization of groundwater quality sampling networks. *Water Resources Research* **41** (12), W12407.

Jha, M. & Datta, B. 2013 Three-dimensional groundwater contamination source identification using adaptive simulated annealing. *Journal of Hydrologic Engineering* **18** (3), 307–317.

Jha, M. & Datta, B. 2015 Application of dedicated monitoring–network design for unknown pollutant-source identification based on dynamic time warping. *Journal of Water Resources Planning and Management* **141** (11), 04015022-1-13.

Kalman, R. E. 1960 A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering* **82** (1), 35–45.

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(1-14).

Neupauer, R. M., Borchers, B. & Wilson, J. L. 2000 Comparison of inverse methods for reconstructing the release history of a groundwater contamination source. *Water Resources Research* **36** (9), 2469–2475.

Singh, R. M., Datta, B. & Jain, A. 2004 Identification of unknown groundwater pollution sources using artificial neural networks. *Journal of Water Resources Planning and Management* **130** (6), 506–514.

Skaggs, T. H. & Kabala, Z. J. 1994 Recovering the release history of a groundwater contaminant. *Water Resources Research* **30** (1), 71–79.

Snodgrass, M. F. & Kitanidis, P. K. 1997 A geostatistical approach to contaminant source identification. *Water Resources Research* **33** (4), 537–546.

Wang, H. & Jin, X. 2013 Characterization of groundwater contaminant source using Bayesian method. *Stochastic Environmental Research and Risk Assessment* **27** (4), 867–876.

Zhang, Y. & Pinder, G. F. 2003 Latin hypercube lattice sample selection strategy for correlated random hydraulic conductivity fields. *Water Resources Research* **39** (8), 1226.

First received 7 March 2016; accepted in revised form 11 July 2016. Available online 27 July 2016