Genetic algorithm hyper-parameter optimization using Taguchi design for groundwater pollution source identification

Xuemin Xia, Simin Jiang, Nianqing Zhou, Xianwen Li and Lichun Wang

ABSTRACT

Groundwater pollution has been a major concern for human beings, since it is inherently related to people's health and fitness and the ecological environment. To improve the identification of groundwater pollution, many optimization approaches have been developed. Among them, the genetic algorithm (GA) is widely used with its performance depending on the hyper-parameters. In this study, a simulation–optimization approach, i.e., a transport simulation model with a genetic optimization algorithm, was utilized to determine the pollutant source fluxes. We proposed a robust method for tuning the hyper-parameters based on Taguchi experimental design to optimize the performance of the GA. The effectiveness of the method was tested on an irregular geometry and heterogeneous porous media considering steady-state flow and transient transport conditions. Compared with traditional GA with default hyper-parameters, our proposed hyper-parameter tuning method is able to provide appropriate parameters for running the GA, and can more efficiently identify groundwater pollution.

Key words | genetic algorithm, groundwater pollution, hyper-parameters, pollution source identification, Taguchi experimental design

Xuemin Xia
Simin Jiang (corresponding author)
Nianqing Zhou
Department of Hydraulic Engineering,
Tongji University,
Shanghai 200092,
China
E-mail: *jiangsimin@tongji.edu.cn*

Xianwen Li
College of Water Resources and Architectural Engineering,
Northwest A&F University,
Yangling 712100,
China

Lichun Wang
Department of Geological Sciences,
The University of Texas at Austin,
Texas, 78712,
USA

INTRODUCTION

With the gradual aggravation of the global water pollution problem, research on various types of water pollution is rapidly increasing (Ayvaz 2010; Parsaie & Haghiabi 2017a; Parsaie & Haghiabi 2017b). Groundwater pollution, which can cross boundaries and affect large areas, has become the focus of water pollution research. In consideration of the great difficulties and high cost of groundwater pollution remediation, the most effective solution is to limit the damage at source, so information concerning pollution sources, such as source location, release magnitude and release period, needs to be collected before taking actions to reverse the trend (Yeh *et al.* 2007). Pollution source identification aims at recognizing the source information from sparsely available observation wells, paving the way for effective management and restoration strategies to achieve a better subsurface environment (Mirghani *et al.* 2009).

The simulation–optimization method, i.e., in which the physically based simulator is externally linked to the optimization model, is one of the widely used methods for identifying contaminant sources (Singh & Datta 2006; Ayvaz 2010). Specifically, groundwater flow and contaminant transport processes are performed first. Afterwards, the optimization model is used to identify the pollution source as an optimization problem (Datta *et al.* 2011). Many optimization approaches, such as the simulated annealing (SA) method (Jha & Datta 2013), pattern search (PS) method (Davey 2008), classified optimization-based method (Datta *et al.* 2011), have been developed to minimize the objective

doi: 10.2166/ws.2018.059

function of deviations between measured and simulated concentrations.

Genetic algorithm (GA) is one of the optimization methods that has received increasing attention in recent decades (Mitra *et al.* 1998; Kalayci *et al.* 2016). Its popularity is attributed to the independence of functional derivatives, and the capability to solve both discrete and continuous optimization problems and to avoid getting trapped in the local optima that inevitably appear in many practical optimization problems (Javadi *et al.* 2005). GA has been widely applied in many groundwater and surface water inverse problems (Huang & Lei 2010; Ayvaz 2016). For example, Zou *et al.* (2007) carried out a search for near-optimal solutions of a complicated water quality model. Singh *et al.* (2008) adopted an interactive multi-objective GA to improve the plausibility of the estimated hydraulic conductivity fields for the heterogeneous groundwater model. Jin *et al.* (2009) successfully identified a groundwater contaminant source at the Borden emplacement site using GA coupled with a local search approach.

To enhance GA optimization performance, a set of appropriate hyper-parameters like mutation function, population size, elite count, crossover fraction, fitness scaling function, and so on, need to be determined. The typical methods of determining these hyper-parameters are based on user-experience and/or trial-and-error methods (Erickson *et al.* 2002; Fazal *et al.* 2005). However, there are no physical explanations about how to determine the optimal hyper-parameter set in previous studies (Islam *et al.* 2015; Velayutham *et al.* 2016).

Recent years have witnessed the increasing availability of combined methods for selecting the GA hyper-parameter set, according to which some parameters are selected by design of experiment (DoE), some are obtained through user-experience/trial-and-error methods, some are chosen from published papers and others are set by default (Bhunia & Sahoo 2011; Arabali *et al.* 2013). Some default parameters and trial-and-error-based methods are used (Giacobbo *et al.* 2002), while some parameters are chosen from existing publications and others are set by experience when GA is applied to groundwater inverse modeling problems (Bastani *et al.* 2010). In addition, default parameters, user-experience-based methods as well as those adopted from other papers are combined to get a suitable set of GA parameters for removing heavy metal pollutants from groundwater (Awad *et al.* 2013). Finally, some parameters from the DoE method, some by user-experience-based methods and the rest by default are synthesized to solve groundwater management problems (Moharram *et al.* 2012).

Nevertheless, the above methods are all problem-related methods, and the performance may not be maximized when dealing with other problems. Also, the mixing of default parameters, parameters determined by other publications and trial-and-error/user-experience-based methods weaken those combined methods, making them unable to comprehensively reflect the effect of parameter selection on the performance. Therefore, a systematic approach is required to select the optimal parameter set to overcome these limitations.

Our study proposes to use the Taguchi design (TD) for tuning hyper-parameters to enhance the performance of GA. The accuracy of TD has been verified when solving a two-dimensional and four large-scale mathematical problems (Dao *et al.* 2016). Moreover, the solutions obtained by GA with optimized hyper-parameters are much better than those obtained by other methods, including PS, SDA (spiral dynamic algorithm) and BFA (bacterial foraging algorithm) (Yang & El-Haik 2003).

The objective of this study is to develop a comprehensive and systematic method, i.e., TD-based GA (TD_GA), to identify the pollution sources in a two-dimensional aquifer. Our paper consists of four parts. Firstly, the hyper-parameters were initialized to generate TD and a series of linked simulation–optimization model experiments were conducted. Then, ANOVA (analysis of variance) analysis of the experimental data proceeded for determining the optimal hyper-parameters. Subsequently, the determined hyper-parameters were used to identify the characteristics of groundwater pollution sources. Finally, assessment was done to demonstrate the efficiency and accuracy of the proposed method.

GROUNDWATER SIMULATION–OPTIMIZATION MODEL

Formulation of the groundwater simulation model

Contaminant transport may potentially include advection, dispersion, diffusion, adsorption, and biodegradation processes. Prior to the numerical simulation of contaminant transport,

the groundwater flow field was implemented by solving the steady-state flow in a two-dimensional aquifer system:

$$\frac{\partial}{\partial x_i}\left(T_{ij}\frac{\partial h}{\partial x_j}\right)+W=0 \;\; i,j=1,2 \tag{1}$$

where T_{ij} is the transmissivity, h is the hydraulic head, W is the volumetric flux per unit volume (positive for inflow and negative for outflow), and x are the Cartesian coordinates. The head distribution can be used to determine υ_i according to Darcy's law:

$$\upsilon_i=-\frac{K_{ij}}{\theta}\frac{\partial h}{\partial x_j} \;\; i,j=1,2 \tag{2}$$

where K_{ij} is the hydraulic conductivity, and υ_i is the average linear velocity of groundwater flow.

The two-dimensional contaminant transport in groundwater is given as:

$$\frac{\partial(\theta C)}{\partial t}+\frac{\partial}{\partial x_i}(\theta C\upsilon_i)-\frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial C}{\partial x_{ij}}\right)-\frac{C_s W}{\theta}=0 \;\; i,j=1,2 \tag{3}$$

where θ is the porosity, C is the contaminant concentration, D_{ij} is the dispersion coefficient (a second-order tensor), and C_s is the source term for the concerned contaminant.

The temporal and spatial concentration distribution of contaminants being released at a specified point can be simulated by Equations (1)–(3). In this study, MODFLOW and MT3DMS were used to simulate the groundwater flow and contaminant transport processes, respectively.

Formulation of the groundwater pollution optimization model

The optimization model determines the unknown source characteristics by minimizing the difference error between the observed and estimated concentration. We used the root mean square error (*RMSE*) as the metric to assess the goodness of the optimization process:

$$RMSE=\sqrt{\frac{\sum_{i=1}^{Nt}\sum_{k=1}^{Nk}\left(Cobs_i^k-Cest_i^k\right)^2}{Nt}} \tag{4}$$

subject to:

$$C=f(q) \tag{5}$$

$$C^L\le C\le C^U \tag{6}$$

$$q^L\le q\le q^U \tag{7}$$

where $Cobs_i^j$ is the observed concentration at the observation location i, in the stress period k, $Cest_i^j$ is the estimated concentration at the observation location i in the stress period k, Nt is the number of observation locations, Nk is the number of observation stress periods, q is the vector of source fluxes, and C is the vector of estimated concentrations.

In Equation (5), $f(q)$ represents the function that transforms the contaminant source fluxes into simulated concentrations via the physically based model. It is linked to the optimization model. Equations (6) and (7) present the constraints of the concentrations and source fluxes, respectively.

The following will discuss the applicability of the GA optimization approach with optimal hyper-parameters for solving groundwater pollution source identification problems.

METHODOLOGY

There are seven parameters that can significantly affect the outcomes of the GA optimization method for groundwater source identification: migration direction, population size, fitness scaling function, selection function, elite count, crossover fraction, and crossover function. Most of them are discrete parameters, while crossover fraction is a continuous parameter, and population size and elite count are integer numbers. However, both integer and continuous parameters are treated as being continuous for simplicity (Dao *et al.* 2016). The TD_GA that was developed here contains four steps for getting the optimal set of hyper-parameters, i.e., the input for the GA optimization method.

Generating Taguchi experimental design

Taguchi experimental design is a comprehensive quality strategy that builds robustness into the process at its

design state when a large number of parameters are considered (Yang & El-Haik 2003). Determined by the number of parameters in consideration and parameter levels available in the optimization approach, the Taguchi Orthogonal Array Design L32(2^1,4^7) was selected (Table 1). In this method, each parameter (migration direction) has two levels, and each subsequent parameter has four levels in the whole 32 experiments (this is demonstrated by the experimental layout in Table 2). The details of the experimental design are shown in Tables 1 and 2.

Conducting the experiments

The hyper-parameters of GA are set according to the obtained experimental layout, based on the well-defined objective function and constraints. To be a fair comparison, the calculation time for each experiment should be exactly the same. Each experiment should be repeated several times (i.e., six times) to improve the consistency of the experimental response.

Analysis of the experimental data

We used the SPSS ANOVA analysis to evaluate the effects of the hyper-parameters on GA optimization performance. As can be seen in the results of the ANOVA analysis (Table 3), the F value, compared with the sum of squares, seems to be a better measure of the relative importance of an effect for the experimental response. The larger the F ratio is, the more important the hyper-parameter is. At the same time, the p value was used to determine whether a parameter is statistically significant for the experimental response. As p is less than 0.05, the effect will be considered significant.

Selection of the hyper-parameters

After completing the ANOVA analysis, the parameters of GA optimization method can be divided into two groups: one group has significant effects on results, while the other does not.

The method of selecting the optimal hyper-parameter values are the same for significant and insignificant parameters to some extent, except for the continuous parameters. For the significant discrete parameters, the rule for selecting the hyper-parameter level is based on the main-effect chart generated by SPSS. That is, the level connected with the highest fitness values should be selected. But for the significant continuous parameters, the gradient-based technique can be used to find the optimal values for a further tuning process if necessary, or as mentioned before, the optimal parameter level should be selected with the highest fitness values. Moreover, for insignificant parameters, the means of selecting the optimal parameter level is exactly the same as for the significant discrete parameters.

CASE STUDY

A hypothetical site based on the case in Jiang *et al.* (2013) was used to assess the TD_GA in terms of searching for the optimal hyper-parameters of GA. In this illustrative

Table 1 | GA hyper-parameters and corresponding experimental levels

			Level			
No.	**Parameter**	**Code**	**1**	**2**	**3**	**4**
1	Migration direction	A	Both	Forward		
2	Population size	B	50	100	150	200
3	Fitness scaling function	C	Proportional	Shift linear	Top	Rank
4	Selection function	D	Tournament	Uniform	Roulette	Stochastic uniform
5	Elite count	E	1	5	10	15
6	Crossover fraction	F	0.3	0.5	0.7	0.9
7	Crossover function	G	Scattered	Tow point	Arithmetic	Single point

Table 2 | Experiment layout and data

	Parameter of TD_GA								Fitness Value					
Experiment	A	B	C	D	E	F	G	Computing Time (s)	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
1	1	3	2	3	1	1	4	14,400	0.2234	0.8170	0.3475	0.5009	0.2368	0.3510
2	2	2	1	2	3	2	4	14,400	0.2353	0.2401	0.2412	0.3080	0.2194	0.2577
3	1	3	3	4	3	1	3	14,400	0.2008	0.2787	0.2549	0.2760	0.2195	0.2605
4	1	2	4	3	3	3	1	14,400	0.1949	0.1972	0.4273	0.5429	0.1912	0.3995
5	2	4	3	3	1	2	2	14,400	0.3206	0.6718	0.4779	0.4906	0.3216	0.4715
6	1	2	3	1	4	4	1	14,400	0.9058	0.9775	0.4841	0.4887	0.9907	0.4706
7	1	1	1	1	1	1	1	14,400	2.7073	3.1215	0.9311	3.1215	2.5704	0.8340
8	2	2	4	1	1	2	3	14,400	0.8543	1.4907	1.1829	3.0398	0.8347	1.1869
9	2	1	4	4	1	4	4	14,400	0.1798	0.7349	0.4917	0.5172	0.1814	0.5196
10	1	1	2	3	2	2	1	14,400	0.3349	0.3503	0.2020	0.3168	0.3307	0.2039
11	2	1	2	1	4	3	3	14,400	0.6815	0.6817	0.5120	0.5130	0.7077	0.4956
12	2	4	1	2	4	1	1	14,400	0.2652	0.3923	0.2989	0.7356	0.2587	0.2721
13	1	3	1	1	2	2	4	14,400	0.8405	0.8561	0.3258	0.3263	0.7741	0.3281
14	1	4	1	4	2	4	3	14,400	0.3420	0.3499	0.4920	0.4969	0.3396	0.4941
15	1	4	3	1	3	3	4	14,400	0.4282	1.3938	0.5197	1.1288	0.4735	0.5362
16	2	3	1	3	4	3	2	14,400	0.4796	0.5207	0.2868	0.3220	0.4847	0.2932
17	1	1	3	4	4	2	2	14,400	0.1929	0.1963	0.4818	0.4838	0.1899	0.4551
18	1	4	4	3	4	4	4	14,400	0.3977	0.4063	0.5168	0.5552	0.3886	0.5420
19	1	2	1	4	1	3	2	14,400	0.5921	0.5944	0.3253	0.3289	0.6083	0.3422
20	2	3	4	4	2	3	1	14,400	0.6852	1.8180	0.3886	0.4496	0.6821	0.3741
21	2	1	3	2	2	3	4	14,400	0.1417	0.1840	0.7222	0.8839	0.1511	0.7033
22	1	4	2	2	1	3	3	14,400	0.1054	0.3030	0.0767	0.0947	0.0960	0.0745
23	2	4	4	1	2	1	2	14,400	0.7307	2.3096	0.5659	4.9578	0.7510	0.5891
24	2	1	1	3	3	4	3	14,400	0.4841	0.4844	0.8385	0.8385	0.4832	0.9394
25	1	3	4	2	4	2	3	14,400	0.0702	0.2547	0.2521	0.5957	0.0711	0.2613
26	2	2	2	4	4	1	4	14,400	0.1835	0.2673	0.2724	0.3748	0.1864	0.2733
27	2	4	2	4	3	2	1	14,400	0.1856	0.4006	0.4117	0.9047	0.1811	0.4268
28	1	1	4	2	3	1	2	14,400	0.1272	0.2557	0.1923	0.2034	0.1287	0.1855
29	2	3	3	2	1	4	1	14,400	0.7926	1.0035	0.6051	0.6393	0.8221	0.5985
30	2	2	3	3	2	1	3	14,400	0.1909	0.2003	0.4489	0.5079	0.1769	0.4800
31	2	3	2	1	3	4	2	14,400	0.3782	0.5556	0.3741	0.4115	0.4069	0.3551
32	1	2	2	2	2	4	2	14,400	0.5640	0.5804	0.8076	0.8202	0.5313	0.8722

case, irregular geometry and non-uniform media were considered. The released pollutant was conservative, and the steady-state flow and transient transport conditions were all studied for the groundwater flow system. Moreover, the initial conditions, the boundary conditions, and the length of release periods were known as priors.

Case description

We focused on a two-dimensional, heterogeneous, isotropic confined aquifer (Figure 1). The flow domain is 25 km in the x-direction and 10 km in the y-direction. The upper-left (AB) and right (CD) sites are constant hydraulic head boundaries

Table 3 | ANOVA analysis

Source	DF	Seq SS	Adj SS	Adj MS	F	p
A	1	0.23	0.23	0.23	1.027	0.312
B	3	0.61	0.61	0.20	0.892	0.447
C	3	3.73	3.73	1.24	5.452	0.001
D	3	20.78	20.78	6.93	30.341	0.000
E	3	6.87	6.87	2.29	10.025	0.000
F	3	2.37	2.37	0.79	3.457	0.018
G	3	3.28	3.28	1.09	4.782	0.003
Error	172	39.28	39.28	0.23		
Total	192	147.93				

with hydraulic heads of 75.0 m and 50.0 m, respectively, while the other two sides were set to be no-flow boundaries. There were two active pollution sources (S1, S2), and six observation wells (O1–O6) for collecting the pollution concentrations at each stress period. There were four different hydraulic conductivity zones in the aquifer domain and the hydraulic conductivities (K) were uniform and isotropic in each zone, the values for zones 1–4 being $K_1 = 20$ m/d, $K_2 = 10$ m/d, $K_3 = 5$ m/d, and $K_4 = 30$ m/d, respectively.

The targeted simulation time duration was 6,000 days long and was divided into 60 equal stress periods, so that each stress period was 100 days. It was assumed that pollutant sources were released in the first four stress periods. The source fluxes were unknown and needed to be determined by the observational data. It can be inferred that there were eight unknown variables, which contained source fluxes of two sources in four stress periods (2×4). The aquifer parameters and source characteristics are listed in Tables 4 and 5.

Applicability of the proposed method

The first step was constructing the experiments to generate the necessary data for the subsequent ANOVA analysis. Hyper-parameters of GA in every experiment were set according to Table 2 to achieve the optimized model. Meanwhile, we ran all 32 designed experiments with the same computing time to estimate the fitness values of the objective function (Table 2), from which we picked out the optimal parameters in the ANOVA analysis based on the minimum value.

The second step is to analyze the resultant experimental data. The ANOVA table and main-effect chart generated by SPSS are presented in Table 3 and Figure 2. The primary concern is the magnitude of p in the ANOVA analysis, which can be further used to evaluate the significance of the hyper-parameters for GA optimization performance.

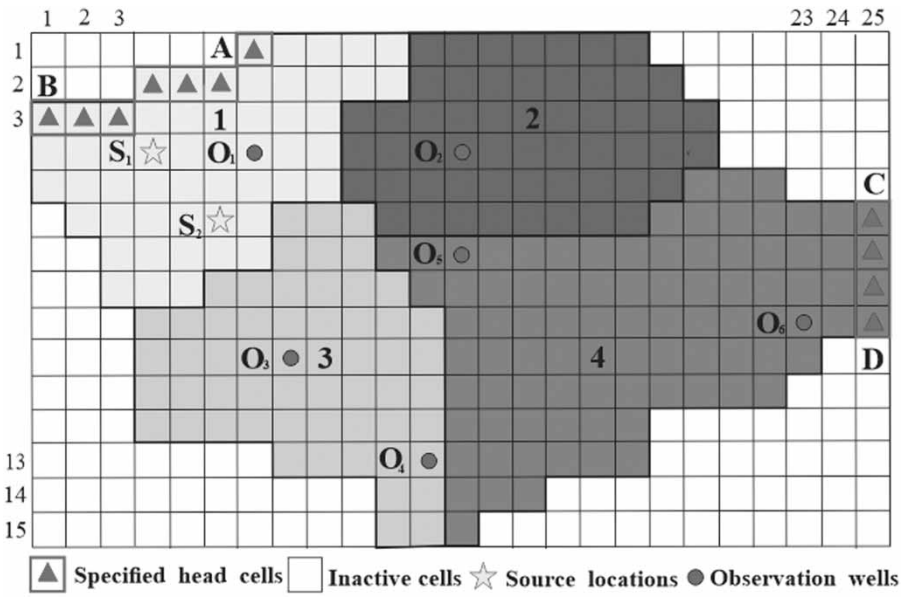

Figure 1 | Plan view of the hypothetical aquifer model.

Table 4 | Aquifer parameters for hypothetical aquifer model

Parameters	Values
Longitudinal dispersivity α_L (m)	25
Transverse dispersivity α_T (m)	2.5
Effective porosity θ	0.35
Saturated thickness b (m)	30
Initial concentration C_0 (mg/L)	0
Grid spacing in x-direction Δx (m)	100
Grid spacing in y-direction Δy (m)	100
Length of the stress periods Δt (day)	100

Table 5 | Actual values of the source fluxes in first four stress periods

	Source fluxes in each stress period(g/s)			
Pollution sources	SP1	SP2	SP3	SP4
S_1	48	10	24.5	34.5
S_2	24	50	20	45

That is, if the p value of the parameter is less than 0.05, the parameter is widely considered to be significant for the relevant problem.

The third step is to select the optimal GA hyper-parameter set for solving the case study problem. As mentioned before, the level of parameter with the highest fitness value is selected to be the optimal hyper-parameter set of the GA in the optimization program (Figure 2).

RESULTS AND DISCUSSION

In order to identify the two-dimensional groundwater pollution sources, the TD_GA was used to minimize discrepancies between the simulated concentrations and actual observations. As expected, the larger the fitness value is, the better the performance of the GA optimization approach is.

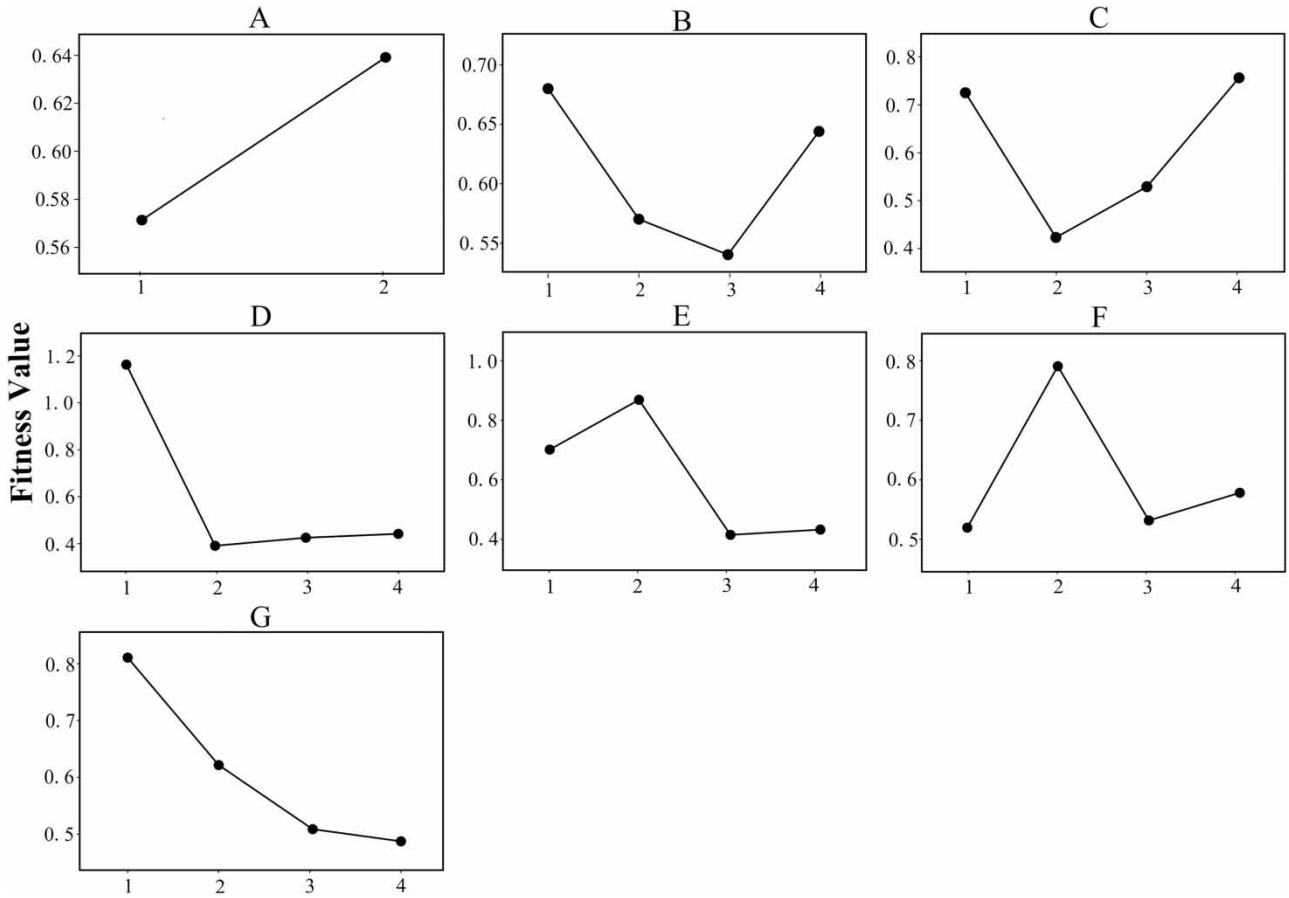

Figure 2 | Main-effect chart.

The p values of C, D, E, F and G are all less than 0.05 (Table 3), which means these parameters are statistically of pronounced significance for GA optimization performance. Among these parameters, E and F, which represent elite count and crossover fraction, were both treated as continuous. The highest fitness values appear at level 2 of both elite count and crossover fraction in the main-effect chart (Figure 2) and not the boundary levels like those of level 1 and 4, so it is unnecessary to tune these two hyper-parameters. The values of elite count and crossover fraction were set to be 5 and 0.5 in the optimal hyper-parameter set. Regarding C, D and G parameters, their highest fitness values were reflected at level 4, level 1 and level 1, respectively, in sequence from the main-effect chart in Figure 2. Combined with Table 1, the fitness scaling (C), selection (D) and crossover function (G) were set as rank, tournament and scattered crossover respectively.

The selected levels of migration direction (A) and population size (B), the two insignificant parameters, may not much affect the performance of GA when dealing with inverse problems. Their levels were selected based on the main-effect chart (Figure 2) generated by SPSS, where the levels associated with the highest fitness values should be chosen. Therefore, the migration direction (A) was selected as forward (level 2), while the population size (B) was set to 50 (level 1) in order to reduce computation burden in this process.

To check the effectiveness of TD_GA, the simulation models were incorporated into four other optimization methods, including PS, SA, GA (GA1), and GA (GA2). The former three are set with default parameters, while GA2 is constrained with a population size of 200 and the other parameters set by default.

Table 6 summarizes the identification results of the TD_GA as well as the PSO, SA, GA1, and GA2. It shows the statistical analysis of the TD_GA (among six trials) of the estimated pollution source; the results of the other four methods (among six trials) are also listed. We used a normalized error (NE) of source fluxes as a metric to assess the goodness of the optimization process (Datta *et al.* 2009). The formula of NE in percent can be defined as:

$$NE = \frac{\sum_{t=1}^{N_p}\sum_{i=1}^{N_s}|\tilde{q}_i(t) - \hat{q}(t)|}{\sum_{t=1}^{N_p}\sum_{i=1}^{N_s}|\hat{q}(t)|} \times 100 \quad (8)$$

where $\tilde{q}_i(t)$ is the estimated source flux in stress period t for source i, and $\hat{q}(t)$ is the actual source flux in stress period t for source i. N_p is the number of stress periods, N_s is the number of pollution sources. In general, if the value of NE is close to zero, the estimated source fluxes are approximately equal to their true values.

Figure 3 shows the performance of the proposed TD_GA in this study. The deviation between the estimated source fluxes and the actual source fluxes for the hypothetical case is trivial. Specifically, the RE value for each identified source flux is less when parameters are fitted by the TD_GA than that obtained by other methods

Table 6 | Comparison of the estimated results with other methods

Source	Source period	Actual source fluxes (g/s)	TD_GA (g/s)			PS (g/s)		SA (g/s)		GA1 (g/s)		GA2 (g/s)	
			Mean	*RE*[a]	*SD*[b]	*Best*	*RE*[a]	*Best*	*RE*[a]	*Best*	*RE*[a]	*Best*	*RE*[a]
S_1	1	48.00	47.64	0.76	2.76	46.20	3.76	33.11	31.03	48.84	1.76	30.63	36.19
	2	10.00	11.47	14.66	0.86	15.72	57.20	34.50	244.98	16.99	69.92	43.86	338.63
	3	24.50	22.76	7.11	1.06	18.53	24.36	19.38	20.90	7.09	71.05	31.74	29.56
	4	34.50	35.21	2.05	3.33	36.61	6.13	28.15	18.40	44.34	28.52	17.17	50.23
S_2	1	24.00	24.66	2.75	1.61	27.28	13.65	58.85	145.19	24.75	3.12	16.79	30.04
	2	50.00	47.73	4.55	3.65	41.02	17.96	6.40	87.20	32.86	34.27	23.32	53.36
	3	20.00	22.49	12.43	0.86	28.11	40.56	13.96	30.18	51.83	159.13	50.00	150.00
	4	45.00	44.13	1.94	3.70	36.61	18.63	59.88	33.07	29.38	34.71	32.57	27.62
NE (%)			4.13			17.33		58.68		39.22		59.43	

[a]RE = Relative error in percent.
[b]SD = Standard deviation.

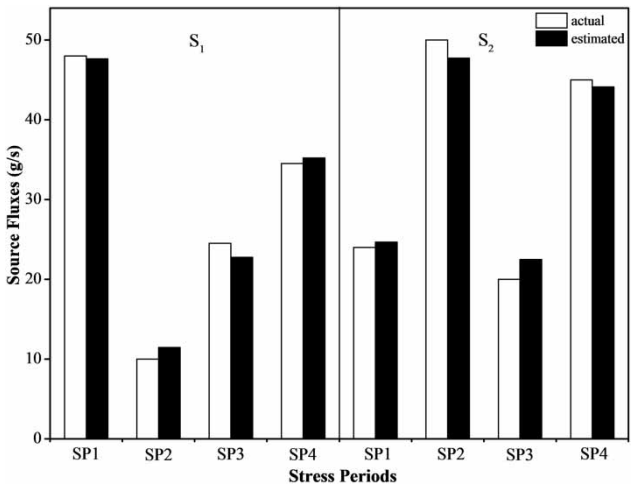

Figure 3 | Comparison of identified source fluxes with actual fluxes by TD_GA.

(Table 6), implying that TD_GA performs better than the other methods. The *RE* value can assess the deviation of a particular variable, but the *NE* value can comprehensively assess the overall performance of these methods. The *NE* values are 17.33%, 58.68%, 39.22% and 59.43% for PS, SA, GA1 and GA2, respectively. The *NE* value of the proposed TD_GA decreased to 4.13%, which suggests the advantage of TD_GA in identifying the optimal parameters for the optimization process.

To further test the effectiveness of the proposed TD_GA, three representative observation wells O_1, O_3 and O_5 were chosen, where the estimated concentration profiles based on the TD_GA and the actual concentrations are depicted in Figure 4. The predicted concentration in each stress period is fairly consistent with the actual concentration for all representative observation wells. This further confirms the reliability of the proposed TD_GA method to identify the characteristics of the pollution source, and thus paves a way for solving groundwater inverse problems.

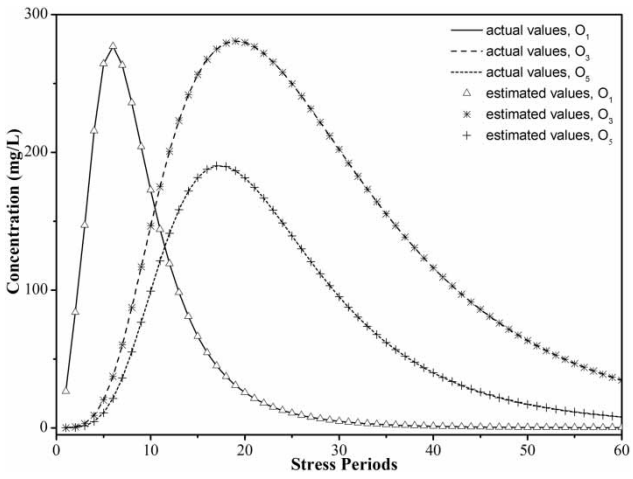

Figure 4 | Actual and estimated concentration profiles at O_1, O_3 and O_5.

CONCLUSIONS

1. The proposed TD_GA method uses Taguchi experimental design for tuning the hyper-parameters of the GA method, while the classical GA algorithm uses trial-and-error or user-experience-based methods to obtain the hyper-parameter set.
2. We applied the proposed TD_GA method for identifying the unknown groundwater pollution sources where the transport equation was solved for a hypothetical aquifer. Compared with four other methods, the TD_GA method performs better in terms of the *NE* and *RE* values.
3. Taguchi experimental design in the TD_GA method can be potentially applied to solve real site problems, indicating that this systematic approach of selecting hyper-parameters can not only be used in GA of groundwater inverse problems, but also can optimize the performance of optimization algorithms.

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (KZ0023020171537), the Joint Foundation of Key Laboratory of Institute of Hydrogeology and Environmental Geology CAGS (KF201611) and the National Natural Science Foundation of China (41502225). The authors would like to thank the Editor and anonymous reviewers for their constructive and valuable comments and suggestions, which significantly improve the quality of this work.

REFERENCES

Arabali, A., Ghofrani, M., Etezadi-Amoli, M., Fadali, M. S. & Baghzouz, Y. 2013 Genetic-algorithm-based optimization

approach for energy management. *IEEE Transactions on Power Delivery* **28** (1), 162–170.

Awad, A. R., Poser, I. V. & Aboul-Ela, M. T. 2013 Optimal removal of heavy metals pollutants from groundwater using a real genetic algorithm and finite-difference method. *Journal of Computing in Civil Engineering* **27** (5), 522–533.

Ayvaz, M. T. 2010 A linked simulation–optimization model for solving the unknown groundwater pollution source identification problems. *Journal of Contaminant Hydrology* **117** (1–4), 46–59.

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

Bastani, M., Kholghi, M. & Rakhshandehroo, G. R. 2010 Inverse modeling of variable-density groundwater flow in a semi-arid area in Iran using a genetic algorithm. *Hydrogeology Journal* **18** (5), 1191–1203.

Bhunia, A. K. & Sahoo, L. 2011 Genetic algorithm based reliability optimization in interval environment. *Computers & Industrial Engineering* **62** (1), 152–160.

Dao, S. D., Abhary, K. & Marian, R. 2016 Maximising performance of genetic algorithm solver in Matlab. *Engineering Letters* **24** (1), 75–83.

Datta, B., Chakrabarty, D. & Dhar, A. 2009 Optimal dynamic monitoring network design and identification of unknown groundwater pollution sources. *Water Resources Management* **23** (10), 2031–2049.

Datta, B., Chakrabarty, D. & Dhar, A. 2011 Identification of unknown groundwater pollution sources using classical optimization with linked simulation. *Journal of Hydro-Environment Research* **5** (1), 25–36.

Davey, K. R. 2008 Latin hypercube sampling and pattern search in magnetic field optimization problems. *IEEE Transactions on Magnetics* **44** (6), 974–977.

Erickson, M., Mayer, A. & Horn, J. 2002 Multi-objective optimal design of groundwater remediation systems: application of the niched Pareto genetic algorithm (NPGA). *Advances in Water Resources* **25** (1), 51–65.

Fazal, M. A., Imaizumi, M., Ishida, S., Kawachi, T. & Tsuchihara, T. 2005 Estimating groundwater recharge using the SMAR conceptual model calibrated by genetic algorithm. *Journal of Hydrology* **303** (1–4), 56–78.

Giacobbo, F., Marseguerra, M. & Zio, E. 2002 Solving the inverse problem of parameter estimation by genetic algorithms: the case of a groundwater contaminant transport model. *Annals of Nuclear Energy* **29** (8), 967–981.

Huang, Y. T. & Lei, L. 2010 Multiobjective water quality model calibration using a hybrid genetic algorithm and neural network-based approach. *Journal of Environmental Engineering* **136** (10), 1020–1031.

Islam, M., Buijk, A., Rais-Rohani, M. & Motoyama, K. 2015 Process parameter optimization of lap joint fillet weld based on FEM–RSM–GA integration technique. *Advances in Engineering Software* **79** (C), 127–136.

Javadi, A. A., Farmani, R. & Tan, T. P. 2005 A hybrid intelligent genetic algorithm. *Advanced Engineering Informatics* **19** (4), 255–262.

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

Jiang, S., Zhang, Y., Wang, P. & Zheng, M. 2013 An almost-parameter-free harmony search algorithm for groundwater pollution source identification. *Water Science and Technology* **68** (11), 2359–2366.

Jin, X., Mahinthakumar, G., Zechman, E. M. & Ranjithan, R. S. 2009 A genetic algorithm-based procedure for 3D source identification at the Borden emplacement site. *Journal of Hydroinformatics* **11** (1), 51–64.

Kalayci, C. B., Polat, O. & Gupta, S. M. 2016 A hybrid genetic algorithm for sequence-dependent disassembly line balancing problem. *Annals of Operations Research* **242** (2), 321–354.

Mirghani, B. Y., Mahinthakumar, K. G., Tryby, M. E., Ranjithan, R. S. & Zechman, E. M. 2009 A parallel evolutionary strategy based simulation–optimization approach for solving groundwater source identification problems. *Advances in Water Resources* **32** (9), 1373–1385.

Mitra, K., Deb, K. & Gupta, S. K. 1998 Multiobjective dynamic optimization of an industrial nylon 6 semibatch reactor using genetic algorithm. *Journal of Applied Polymer Science* **69** (1), 69–87.

Moharram, S. H., Gad, M. I., Saafan, T. A. & Allah, S. K. 2012 Optimal groundwater management using genetic algorithm in El-Farafra Oasis, western desert, Egypt. *Water Resources Management* **26** (4), 927–948.

Parsaie, A. & Haghiabi, A. H. 2017a Numerical routing of tracer concentrations in rivers with stagnant zones. *Water Science and Technology: Water Supply* **17** (3), 825–834.

Parsaie, A. & Haghiabi, A. H. 2017b Computational modeling of pollution transmission in rivers. *Applied Water Science* **7** (3), 1213–1222.

Singh, R. M. & Datta, B. 2006 Identification of groundwater pollution sources using GA-based linked simulation optimization model. *Journal of Hydrologic Engineering* **11** (2), 1216–1227.

Singh, A., Minsker, B. S. & Valocchi, A. J. 2008 An interactive multi-objective optimization framework for groundwater inverse modeling. *Advances in Water Resources* **31** (10), 1269–1283.

Velayutham, K., Venkadeshwaran, K. & Selvakumar, G. 2016 Process parameter optimization of laser forming based on FEM-RSM-GA integration technique. *Applied Mechanics & Materials* **852**, 236–240.

Yang, K. & El-Haik, B. 2003 *Design for Six Sigma: A Roadmap for Product Development*. McGraw-Hill, New York, USA.

Yeh, H. D., Chang, T. H. & Lin, Y. C. 2007 Groundwater contaminant source identification by a hybrid heuristic approach. *Water Resources Research* **43** (9), W09420.

Zou, R., Lung, W. S. & Wu, J. 2007 An adaptive neural network embedded genetic algorithm approach for inverse water quality modeling. *Water Resources Research* **43** (8), W08427.

First received 11 August 2017; accepted in revised form 3 March 2018. Available online 16 March 2018