An Artificial Intelligence Based Self-Adaptive Dynamic Process Control System for Enhancing In-Situ Bioremediation of Benzene-Contaminated Groundwater

Li He[1], Xu Duan[1], Chenyang Li[1], Yang Xu[1], and Mengxi He[1]

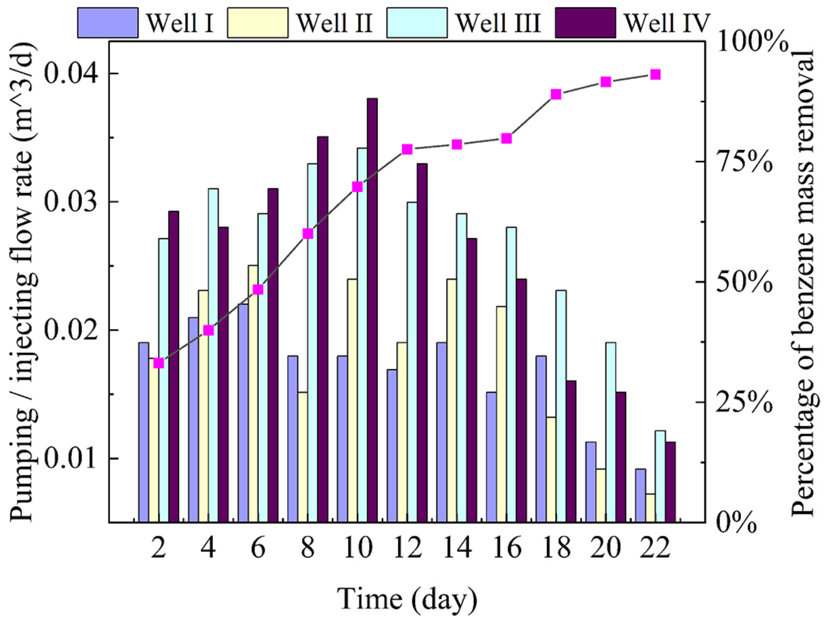
[1]State Key Laboratory of Hydraulic Engineering Simulation and Safety at Tianjin University

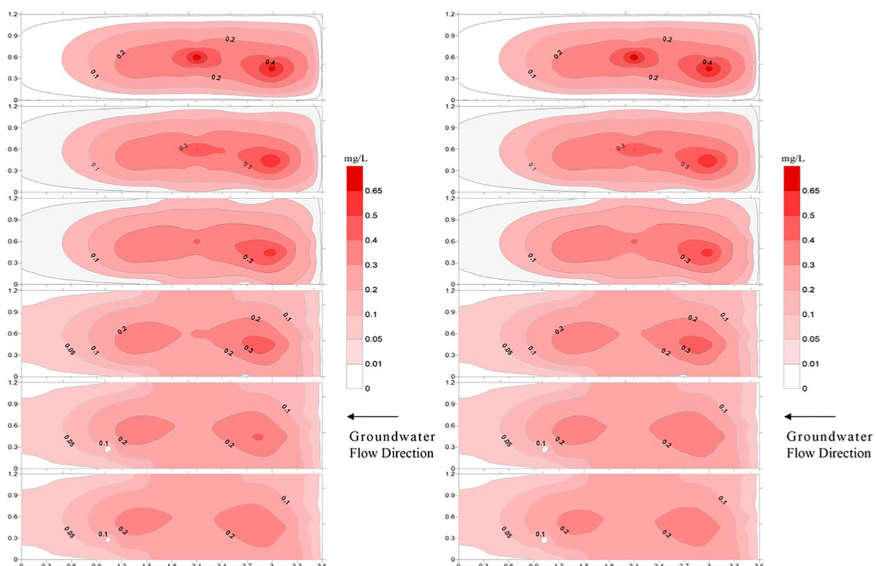
December 17, 2022

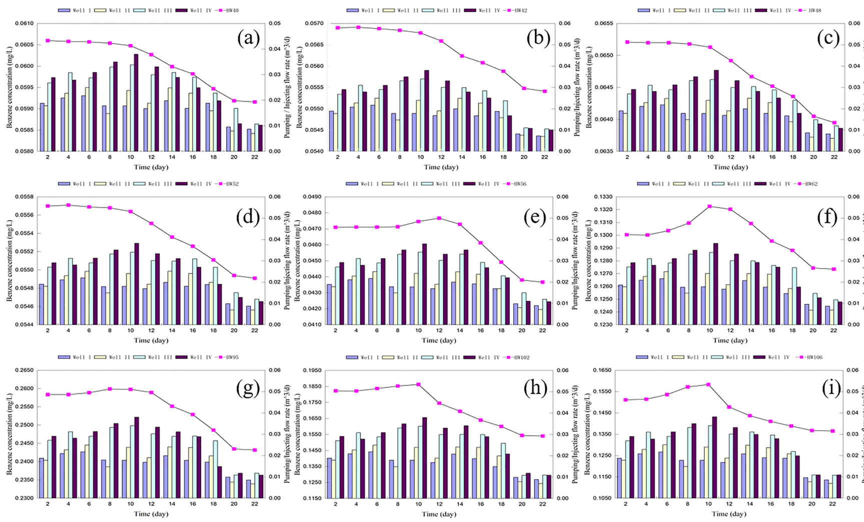
Abstract

This study develops a new AI-based Self-Adaptive DPC (SADPC) system based on stepwise inference combing with genetic algorithm optimization technologies, including a filtered-clustering inference prediction model (FCI simulator), a stepwise inference controller (SI emulator), a model predictive control controller (MPC controller), a 1st-stage optimizer, and a 2nd-stage optimizer. This system effectively reflects the dynamics and complexity of the biodegradation process and realizes the control for the remediation system based on the feedback information. To achieve this goal, a statistical model for simulating the bioremediation process through the FCI simulator is proposed, which can predict the resulting contamination situation based on the previous contamination situation and control action. Then a bridge between control actions and contamination situations is established through the SI emulator, which can generate a control action based on a given contamination situation. Through running the SADPC system, the desired control action can be identified. Results show that The SADPC system increases the removal rate of benzene and arrives at the remediation goal earlier than other systems. This suggested decision makers that guidelines and policies on remediation-oriented SADPC systems could be tentatively investigated, developed, and applied in the future effort.

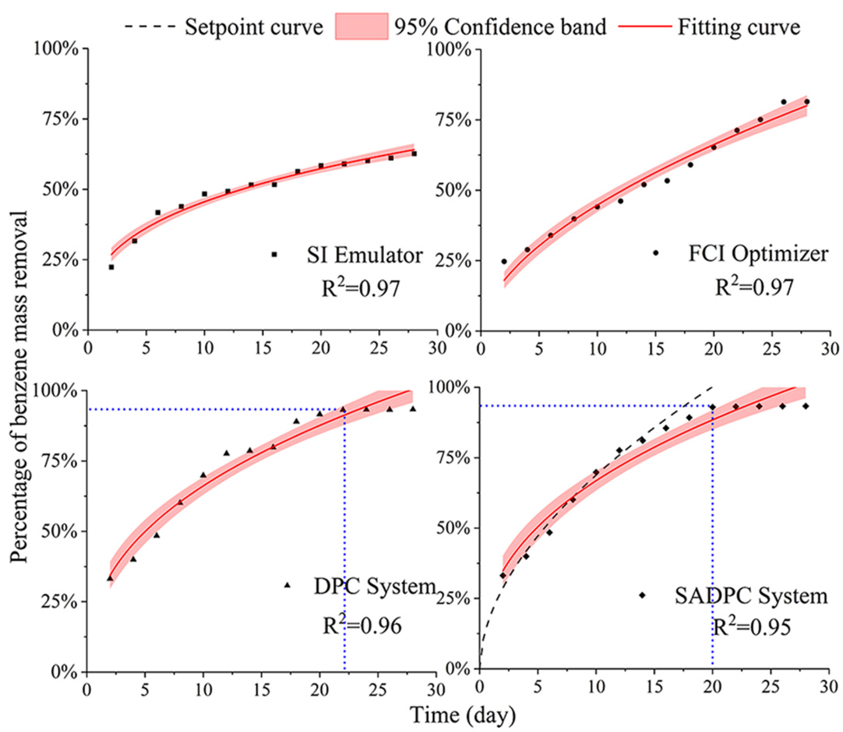
Hosted file


Well I
Well II
Well III
Well IV
Pumping / injecting flow rate (m^3/d)
0.04
0.03
0.02
0.01
100%
75%
50%
25%
0%
Percentage of benzene mass removal
2
4
6
8
10
12
14
16
18
20
22
Time (day)

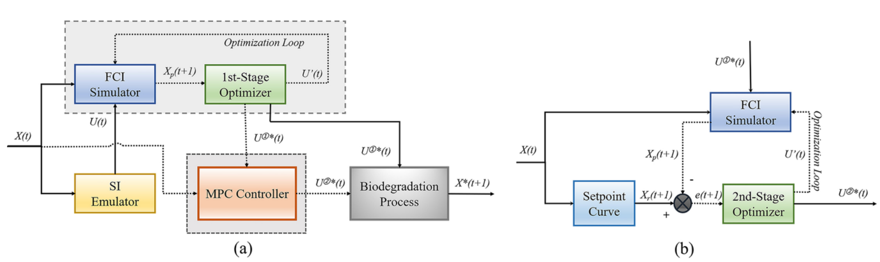

mg/L
0.65
0.5
0.4
0.3
0.2
0.1
0.05
0.01
0
Groundwater
Flow Direction
mg/L
0.65
0.5
0.4
0.3
0.2
0.1
0.05
0.01
0
Groundwater
Flow Direction



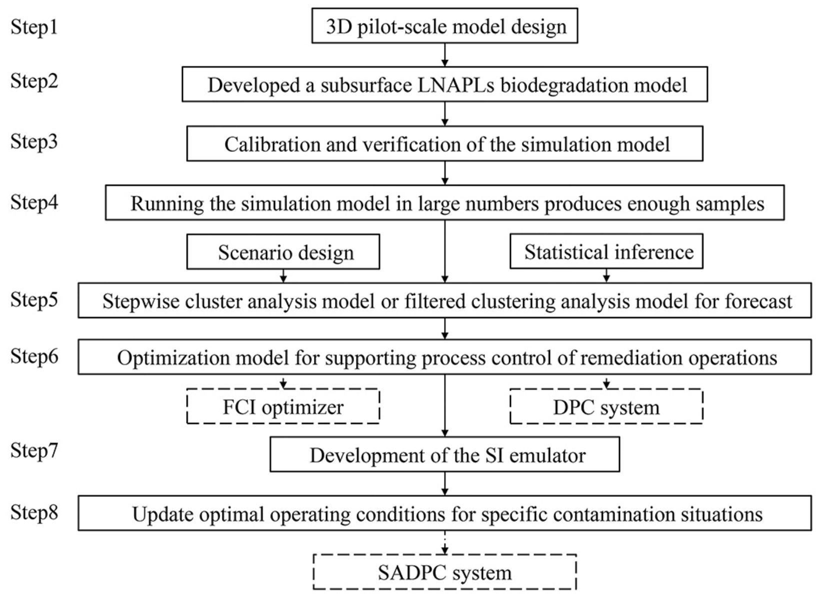
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
Time (day)



Setpoint curve
95% Confidence band
Fitting curve
Percentage of benzene mass removal
SI Emulator
R^2=0.97
FCI Optimizer
R^2=0.97
DPC System
R^2=0.96
SADPC System
R^2=0.95
Time (day)



Optimization Loop
FCI Simulator
1st-Stage Optimizer
SI Emulator
MPC Controller
Biodegradation Process
(a)
FCI Simulator
Setpoint Curve
2nd-Stage Optimizer
Optimization Loop
(b)


Step1
3D pilot-scale model design
Step2
Developed a subsurface LNAPLs biodegradation model
Step3
Calibration and verification of the simulation model
Step4
Running the simulation model in large numbers produces enough samples
Scenario design
Statistical inference
Step5
Stepwise cluster analysis model or filtered clustering analysis model for forecast
Step6
Optimization model for supporting process control of remediation operations
FCI optimizer
DPC system
Step7
Development of the SI emulator
Step8
Update optimal operating conditions for specific contamination situations
SADPC system

An Artificial Intelligence Based Self-Adaptive Dynamic Process Control System for Enhancing In-Situ Bioremediation of Benzene-Contaminated Groundwater

Enter authors here: L. He[1*†], X. Duan[1†], C. Li[1], Y. Xu[1], and M. He[1]

[1]State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Tianjin 300350, China

[*]Corresponding author: Li He (helix111@tju.edu.cn)

[†]L. He and X. Duan contributed equally to this manuscript

Key Points:

- Perform a simulation of enhanced remediation process shown by a contaminant fate and transport model
- Develop a self-adaptive dynamic process control (SADPC) system and provide suggestion for guidelines and policies on SADPC systems
- Improve the removal rate of benzene in groundwater during the bioremediation process

Abstract

This study develops a new AI-based Self-Adaptive DPC (SADPC) system based on stepwise inference combing with genetic algorithm optimization technologies, including a filtered-clustering inference prediction model (FCI simulator), a stepwise inference controller (SI emulator), a model predictive control controller (MPC controller), a 1st-stage optimizer, and a 2nd-stage optimizer. This system effectively reflects the dynamics and complexity of the biodegradation process and realizes the control for the remediation system based on the feedback information. To achieve this goal, a statistical model for simulating the bioremediation process through the FCI simulator is proposed, which can predict the resulting contamination situation based on the previous contamination situation and control action. Then a bridge between control actions and contamination situations is established through the SI emulator, which can generate a control action based on a given contamination situation. Through running the SADPC system, the desired control action can be identified. Results show that The SADPC system increases the removal rate of benzene and arrives at the remediation goal earlier than other systems. This suggested decision makers that guidelines and policies on remediation-oriented SADPC systems could be tentatively investigated, developed, and applied in the future effort.

Keywords

Self-adaptive dynamic process control, in-situ bioremediation, contaminant fate and transport modeling, physically groundwater simulation

1 Introduction

In-situ bioremediation (ISB) techniques (Albers et al., 2015; Zhang et al., 2020) aim to enhance the biodegradation of organic constituents in the subsurface by encouraging the growth and reproduction of indigenous microorganisms. The ISB technique involves a mechanism for stimulating and maintaining the activity of intrinsic bioremediation processes, by which indigenous microbes convert contaminants to innocuous end products via electron acceptor and/or inorganic nutrient amendments. The normal operation of ISB consists of routine checking of operation and maintenance of equipment, groundwater levels, extraction and injection rates, groundwater electron acceptor concentrations, nutrient levels, pH, and conductivity.

System optimization approaches have been demonstrated to be useful in enhancing remediation efficiency and reduce remediation cost during water treatment and remediation (Chiandussi et al., 2012; He et al., 2008a; He et al., 2008b; He et al., 2008c; Passino, 2002; Sun et al., 2020; Wang et al., 2020). Compared with system optimization approaches, dynamic process control (DPC) could be a better way in fulfilling real-time system optimization by temporally regulating a set of operating conditions such as additions of electron acceptors and nutrients, groundwater extraction and injection rates, remedial cleanup time, etc. Various studies have been undertaken on development and applications of DPC techniques (Ahmed & Rodriguez, 2020; Bashivan et al., 2019; Bechet et al., 2016; Diangelakis et al., 2016; Liu et al., 2016; Mayne, 2014; Miller et al., 2016; Stentoft et al., 2021; Zeng & Liu, 2015). For example, Stentoft et al. (2021) proposed a general model predictive control algorithm to achieve the optimal operating conditions by controlling the effluent concentrations, total costs, and other management objectives. This approach allows the water resource recovery facilities to quickly accommodate new control requirements. Liu et al. (2016) presented an Event-driven Model Predictive Control (EMPC) method to ensure that the flows of sewage streams containing the dosed chemical are reasonably distributed throughout the sewer networks. The EMPC strategy substantially enhanced the performance of sulfide mitigation when dealing with the corrosion and odor problems. Ahmed & Rodriguez (2020) demonstrated a non-linear model predictive control (NMPC) system to optimize the automatic start-up of anaerobic digesters, which achieved a higher target methane production rate and superior control variables set-point tracking error performance.

However, three challenges of the conventional DPC techniques lead to the difficulty in applying to a general ISB system. First, conventional DPC depends on the use of a set of nonlinear state equations group (or prediction model) representing the input-output relations. While there is difficulty in analytically or numerically solving the equations group, the DPC would fail to work because of extremely low solution efficiency. Because the ISB prediction model is computationally costly, proxy modeling may be a good means of solving this challenge, i.e., to produce a set of proxy models to replace initial ones through statistical or artificial intelligence (AI) methods. Usually, proxy models have the advantages of computation-rapid, result-stable, and error-tolerable (Gopalakrishnan et al., 2011; Gorelick and Zheng 2015; He et al., 2008a; He et al., 2008b;Meray et al., 2022; Siade et al., 2020; Stramer et al., 2010). Second, generation of optimal operating conditions within a given time period by conventional DPC relies on the difference (or error) between the predicted remediation performance and pre-determined level. For facilitating computation, a prediction model implied in the DPC framework is generally assumed to maintain static (without any variation) during the entire remediation process. This assumption may not be suitable particularly where complex hydrogeological conditions and biochemical process exist in the groundwater. A feasible approach is to introduce self-adaptive prediction or proxy models that can be dynamically trained and improved subject to external environmental variations. Third, it tends to make predictions that cannot meet expectations when dealing with the complicated situation of contaminant degradation, since the conventional DPC prediction model remains static during the entire remediation process. To alleviate the problem of falling into the local optimality dilemma triggered by the DPC static prediction model, a near-ideal biodegradation process can be obtained by conducting a second-stage optimization and developing the predicted trajectory (setpoint curve) based on the entire process prediction of the DPC system.

Therefore, this paper aims to present a new AI-based self-adaptive DPC system (SADPC) for enhancing in-site bioremediation of benzene-contaminated groundwater due to non-aqueous phase liquids (NAPLs) leakage from underground storage tank. The DPC system includes a filtered-clustering inference prediction model (also called FCI simulator) (Gerber & Horenko, 2015; Pizzagalli et al., 2019; Zhan et al., 2018), a stepwise inference controller (SI emulator), an 1st-stage optimizer, a model predictive control controller (MPC controller), and a biodegradation process. The MPC controller includes an FCI simulator, an 2nd-stage optimizer, and an error regulator.

This task entails: 1) developing a statistical model for simulating the bioremediation process through the FCI simulator, which can predict the resulting contamination situation based on the previous contamination situation and control action; 2) establishing a bridge between control actions and contamination situations through the SI emulator, which can generate a control action based on a given contamination situation; 3) running the SADPC system to identify the desired control action.

2 Materials and Methods

2.1 Development of the pilot-scale reactor

A pilot-scale reactor was developed and used to physically simulate the flow and transport of benzene (gasoline) in the groundwater (He, 2008; He et al., 2008a). It also facilitated the implementation of enhanced in-situ biodegradation and the relevant simulation efforts. The reactor (Figure S1) is of cuboid shape with an interior dimension of Length × Width × Height = 3.6 × 1.2 × 1.4 m3. It was composed of four sections, each of which contained a supporting part, a loading manhole, and two observation windows. More details regarding the reactor were shown in the supporting information.

For the simulation of hydrocarbon leakage, 12 liters of gasoline were injected into the bottom of the second soil layer at an upper stream location during a 1.5-day period. At the same time, tap water from a water container was pumped into the system as groundwater inflow at a rate of 20 L/day (through a peristaltic pump). The water level in the upstream gauge was 55 cm high and that in the downstream one was 45 cm high. After the leakage period, such flow conditions were maintained for 40 days to simulate the process of natural attenuation in the subsurface. The enhanced in-situ biodegradation process was then started right after this 40-day period. The experiment of flow and transport lasted 40 days after the gasoline leakage, followed by a 22-day enhanced in-situ biodegradation action. Environmental managers are more concerned with benzene than toluene, ethylbenzene, and xylenes (TEX) due to the fact that benzene is highly toxic and carcinogenic. In addition, during the remediation process, concentrations of TEX would become much lower than the respective environmental criteria as long as the benzene concentration is lower than the regulated criterion. Therefore, only benzene concentrations were analyzed in this study. The set-up of the reactor and the detailed analysis of the pilot-scale experimentation can be seen in Sections S1 and S2 of the supporting information (McDonald & Harbaugh, 1988; Jimenez et al., 2006; Zhang et al., 2008; Wolicka et al., 2009; Liang et al., 2013; Niswonger & Prudic, 2013; Xin et al., 2013; Yang et al., 2019; Hu et al., 2021; Umar et al., 2021).

2.2 Contaminant fate and transport modeling in the groundwater

A critical step in understanding the impact of a subsurface release of NAPL is a modeling analysis of the NAPL flow and transport and fate of its crucial constituents. The 3D multiphase and multicomponent (3DMM) model is used to simulate contaminant fate and transport in the groundwater. The basic mass conservation equation for components in the subsurface can be written as follows (Li et al., 2007; Schaerlaekens et al., 2005):

$$\frac{\partial}{\partial t}(\phi \tilde{C}_k \rho_k) + \vec{\nabla} \cdot [\sum_{l=1}^{n_p} \rho_k (C_{kl} \vec{u}_l - \phi S_l \vec{\vec{D}}_{kl} \cdot \vec{\nabla} C_{kl})] = R_k \quad (1)$$

where k is the component index; l is the phase index; ϕ is the soil porosity; $\tilde{C}_k$ is the overall concentration of component k (volume fraction); ρ_k is the density of component k [ML^{-3}]; n_p is the number of phases; C_{kl} is the concentration of component k in phase l (volume fraction); $\vec{u}_l$ is the Darcy velocity of phase l [LT^{-1}]; S_l is the saturation of phase l; R_k is the total source/sink term for component k (volume of the component k per unit volume of porous media per unit time); $\vec{\vec{D}}_{kl}$ is the dispersion tensor. The overall concentration ($\tilde{C}_k$) denotes the volume of the component k summed over all phases. Formulas, solution methods and other details are given in the Section S3 in the supporting information (Bear, 1979; Faust et al., 1989; Delshad et al., 1996).

The model can be solved numerically through the block-centered finite difference method, and it is possible to obtain the concentration of specified component k in phase l (C_{kl}) at a certain time. In addition, the biodegradation model with single substrate, single electron acceptor and single biological species should be required for a system (de Blanc, 1998; Huang et al., 2006). The solution to the flow equations was used as the initial conditions for the biodegradation reactions. By incorporating the component concentrations obtained through the pollutant migration model (C_{kl}) into the biodegradation modeling of contaminants in the groundwater, the substrate degradation rate during this time period can be computed. Details regarding the biodegradation modeling of contaminants in the groundwater are shown in Sections S4 and S5 (Rittmann et al., 1991; Chang & Alvarez-Cohen, 1995; de Blanc, 1998; Chang & Alvarez-Cohen, 2010).

2.3 Framework of the Study Method

In this paper, an artificial intelligence-based self-adaptive dynamic process control (SADPC) system for enhancing in-situ bioremediation of benzene-contaminated groundwater is established. SADPC is used to temporarily adjust a set of operating conditions in the aforementioned biodegradation process to achieve real-time system optimization. For the realization of in-situ bioremediation of benzene-contaminated groundwater, four groundwater control models (i.e., SI emulator, FCI optimizer, DPC system and SADPC system) were simultaneously used to predict and adjust operating conditions for efficient pollutant degradation.

Considering the high complexities and dynamics of the bioremediation system, it is inevitable that some important information might be missed/ignored when establishing a biodegradation model since almost all the models are a selective, dynamic abstraction of reality. In some situations, if there are a large number of experimental data, a statistical relationship can be developed to substitute the general simplified model (Huang et al., 2006).

In this study, a set of surrogate simulators can be established to quantify the relationship between pumping/injecting flow rate and benzene concentration by employing a stepwise cluster analysis (SCA) method, detailed descriptions have been shown in He et al (He, 2008; He et al., 2008a; He et al., 2008b). More information can be seen in the supporting information (Section S6) (Rao, 1952). To determine the optimal repair strategy, an FCI simulator was presented based on the SCA method (Zou et al., 2009).

In the FCI simulator, the relationships between contaminant concentrations and remediation operating conditions can be established through the filtered-clustering inference method based on a number of simulation runs. Given the pollution situation, the optimal operating conditions of the FCI simulator can be obtained under the constraints of the optimization objective. Based on the FCI simulator, the FCI optimizer was developed to optimize the biodegradation process (the framework is shown in Figure S2). Sections S7 and S8 of the supporting information details the procedures of the FCI simulator and the optimization model for the FCI optimizer (Maybeck, 1979; Jacobs, 1993).

After the optimal operation conditions for each scenario are determined, the SI emulator is developed through the obtained knowledge base. For the SI emulator, the corresponding operating conditions can be obtained by a given benzene concentration, and the benzene concentration of the next stage can be obtained through the biodegradation process. The framework of the SI emulator is shown in Figure S3. The operating conditions coming from the SI emulator cannot be considered the optimal one because there are no standard optimization curves for the biodegradation process as a reference for the experiments. Therefore, operating conditions must be optimized before it is applied to the real bioremediation process, which can be realized through the adjustment of the ranges of control conditions.

At the same time, since the operating conditions obtained by the FCI simulator are optimized for cost minimization without taking remediation efficiency into consideration, the actual bioremediation often does not achieve ideal results and cannot be used as an actual optimization curve to guide in situ bioremediation of groundwater system. Therefore, the minimized operation cost and maximized degradation efficiency should be considered. For the operating conditions, decisions of oxygen and nutrient injection rates and groundwater extraction rates directly affect the operating cost. The lower the injection or extraction rate, the lower the cost and contaminant removal rate. According to the content above, the optimization model for the DPC system in this study is given in the Section S9 in the supporting information (Huang et al., 2008).

For the poor performance of traditional DPC technology applied to general ISB systems, this study proposes a SADPC technology to reduce the impact of the defects of traditional DPC technology itself. According to this improvement, the reference trajectory (i.e., prediction curve/setpoint curve) can be obtained based on the prediction results from the entire DPC process. Then rolling optimization is performed to update the prediction curve in the next time period. The optimization model for the SADPC system in this study is given in Section S10. GA is used to solve all the developed discrete and nonlinear model. More information on GA can be seen in the supporting information (Section S11) (Holland, 1975; Kuo et al., 2006; Matott et al., 2006; Stramer et al., 2010; Opher & Ostfeld, 2011; Greenland et al., 2016; Hou et al., 2017; Liu et al., 2017; Shen et al., 2018; Liao et al., 2020).

The framework of the DPC system is presented in Figure 1 (a). Based on the DPC system, the SADPC system can be optimized by adding an MPC controller (Figure 1 (b)). The major components include a SI emulator, an FCI simulator, an MPC controller, and an optimization procedure1. In Figure 1, $X(t)$ is the input for the SI emulator, the FCI simulator and the MPC controller; $X_P(t+1)$ is the output of the FCI simulator; $X_r(t+1)$ is the setpoint; $e(t+1)$ is the error between $X_P(t+1)$ and $X_r(t+1)$; $X^*(t+1)$ is the optimal contamination situation after system operation; $U(t)$ is the control action coming from the SI emulator; $U'(t)$ is the tentative control signal; $U^{①*}(t)$ is the optimal control action after the 1st-stage optimization procedure; $U^{②*}(t)$ is the optimal control action after the 2nd-stage optimization procedure. The specific operation program of the MPC controller can be seen in Section S12 in the supporting information.

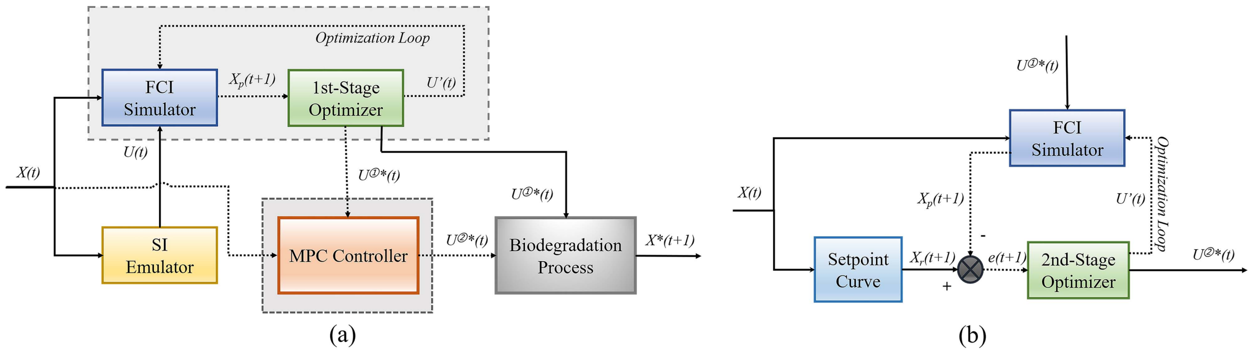


Figure 1. Framework of the SADPC system(a) and MPC controller (b)

Figure 2 shows the framework of the study method. The general procedure of developing a process control system for enhanced in-situ biodegradation consists of eight steps. The specific steps can be seen in Section S13 in the supporting information. Given the same initial benzene concentration, the predicted optimal degradation strategies of these four developed groundwater control models are various. By comparing the restoration processes and results, the optimal groundwater control models and operation strategies can be decided through the comparison of their degradation processes and the removal results.

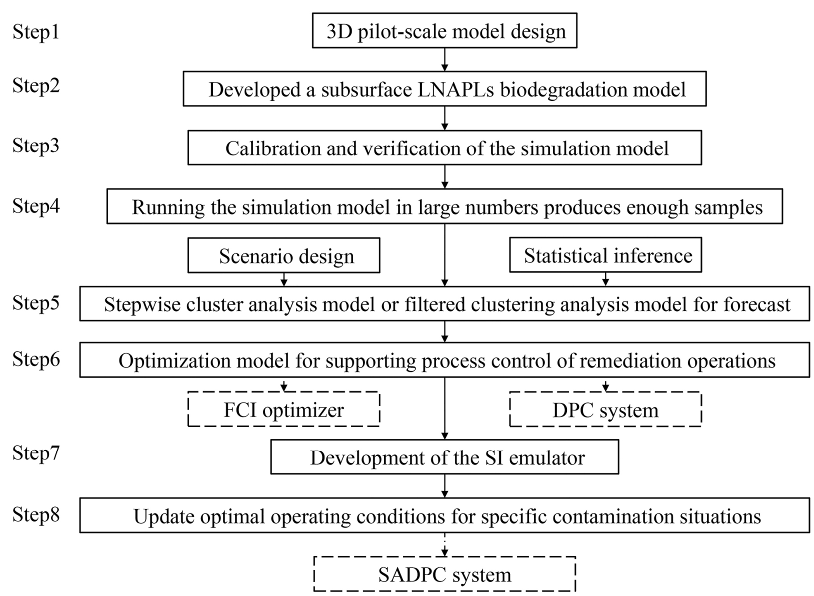


Figure 2. Flowchart of the solution method

3 Results

3.1 Clustering analysis

The developed NAPLs biodegradation model can then be used for simulating the system's responses under various operating conditions. However, it will bring high complexities and computation requirements if the developed simulation model is directly incorporated into the optimization framework. The filtered clustering method can be used to establish the relationship between the remediation efforts (i.e., pumping/injecting rates of selected wells) and the system's responses (i.e., benzene concentrations). The developed simulation model was used to generate a large number of inputs and outputs for supporting the establishment of such a relationship.

According to the characteristics of the soil profile, the NAPLs fate and transport and the contaminant plume movement, benzene concentrations in six wells were used as the representatives of the contamination situation in the groundwater. These included wells 5, 7, 8, 10, 11 and 12 (Benzene concentrations in these wells were denoted as $x_1^0, x_2^0, x_3^0, x_4^0, x_5^0$, and x_6^0). In order to reflect as many contamination situations as possible, a large range of the benzene concentration levels was considered. The maximum benzene concentration was 30 mg/L, and the minimum was 0 mg/L. Within this range, 50 concentration levels were generated randomly for each concerned well such that 50 contamination situations were produced (Table S5) (Lenczewski et al., 2003; Zhang et al., 2020).

A groundwater pumping system was used to circulate nutrients and oxygen through the contaminated aquifer. The process involves (a) the introduction of aerated and nutrient- and biomass-enriched water into the contaminated zone through two injection wells, and (b) the recovery of the down-gradient water through two extraction wells. The amendments were circulated through the contaminated zone to provide mixing and intimate contacts among the oxygen, nutrients, contaminant, and microorganisms. Therefore, the pumping/injecting rates directly affected the contaminant removal efficiency and system operation cost. In this study, pumping/injecting rates of selected wells were identified as the main control conditions.

The ranges of pumping/injecting rates were determined by considering the soil porosities and permeabilities in the pilot system and testifying them through the developed biodegradation model. The maximum flow rate was set as 40 L/day while the minimum was 10 L/day. The biomass/oxygen/nutrient concentrations were 20/8/1500 (mg/L) in the injecting fluid, respectively. Totally 50 scenarios of the operating conditions were randomly generated (Table S6 in the supporting information). The relevant control variables were denoted as u_1 (injection rate for well I, L/d), u_2 (injection rate for well II, L/d), u_3 (extraction rate in well III, L/d), and u_4 (extraction rate in well IV, L/d).

The combination of the 50 contamination-level scenarios and the 50 operating-condition scenarios led to 2500 scenarios. Correspondingly, 2500 input files were produced for the developed NAPLs biodegradation model. The experimental results indicated that benzene concentrations in the groundwater reduced significantly 18 days after the remediation started. Therefore, a 22-day duration was set, which was divided into 11 2-day periods. For each contamination-level scenario ($x_1^0, x_2^0, x_3^0, x_4^0, x_5^0$, and x_6^0), 50 sets of data about (1) the respondent percentage of benzene mass removal (η) and (2) the operating conditions of enhanced in-situ biodegradation (u_1, u_2, u_3, and u_4) can be obtained from simulation runs. Basing on the 300 cluster trees obtained through the filtered clustering analysis in total, the value of η can be predicted given the inputs of operating conditions.

The relationship between the process operating conditions and respondent value of η under 50 contamination-level scenarios were established through the filtered clustering analysis based on a large number of simulation runs under 50 operating-condition scenarios. For each contamination-level scenario (with the initial benzene concentrations of ($x_1^0, x_2^0, x_3^0, x_4^0, x_5^0$, and x_6^0), the resulting cluster tree system can be incorporated into a discrete and nonlinear optimization model. Genetic algorithm was used to solve the developed discrete and nonlinear model under each contamination-level scenario[38-41]. The number of generations was set as 200; the crossover rate (R^{CRO}) was 0.6; the mutation rate (R^{MUT}) was 0.003; and the number of the initial population was 70.

Over-parameterization is described as the scenario where the number of parameters of the model is redundant compared to the training dataset. Its high power consumption and memory occupation can degrade the performance of the model and make prediction action worse. This research is not over-parameterized because the network pruning method (Akyol, 2020; Hao & Chiang, 2006; Itoh & Adachi, 2017; Sun et al., 2015) was used to alleviate the problems and achieve the best results by evaluating the importance of the parameters based on the absolute values and removing the unimportant parameters. Then the regularizer can be added to the loss in order to make the weights sparse in the training process. During the experiment, a number of attempts was made to make sure that the model performs well under these specific model parameters.

3.2 Process control action analysis

Process control is used for operating the ISB system based on the SI emulator, FCI simulator and GA-based optimizer. In the DPC system, firstly, benzene concentrations at the concerned wells at the beginning of time period t were monitored. Then the highest contaminant concentration anywhere in the mesh ($Ƀ^{\mathrm{MAX}}$) and the percentage of benzene mass removal (η) can be used as inputs for the SI emulator to generate the optimal operating schemes correspondingly for the period t. the outputs are pumping/injecting rates of wells I, II, III, and IV. For the FCI simulator, the inputs include operating conditions of selected wells and the corresponding benzene concentrations; the outputs are the highest contaminant concentration anywhere in the mesh ($Ƀ^{\mathrm{MAX}}$) and the percentage of benzene mass removal (η) (detailed in Table S7 in the supporting information). It means that the pumping/injecting rates depends on the biodegradation process of benzene at time period t. Next, benzene concentrations in the concerned wells were monitored at the end of period t; then they can be regarded as new initial states for the next time period. The entire biodegradation process in benzene-contaminated groundwater could be controlled with cost-effective operational decisions step by step.

According to the degradation situation and operation process, a second-stage rolling optimization model (SADPC system) was used to meet the further expectation. An ideal setpoint curve can be produced based on the control process and benzene removal information of DPC system and new setpoint curves should be updated in the next optimization period. Therefore, a SADPC system can improve the biodegradation performance in the DPC system.

The contaminant concentration distribution of Day 57 was used as initial conditions for the SADPC system, and the pumping/injecting rates were assumed to be adjusted every two days. Figure 3 presents the 11 optimal operating conditions for the 11 2-day periods of the entire remediation duration in the DPC system. Results indicate that the operating conditions of 4 wells varied significantly under different instructions. Both highest injection rates for well I (u1) and well II (u2) are found on Day 6 at 0.022 and 0.025 m^3/d respectively, and the extraction rates for well III (u3) and IV (u4) reach the highest on Day 10 with 0.034 and 0.038 m^3/d respectively.

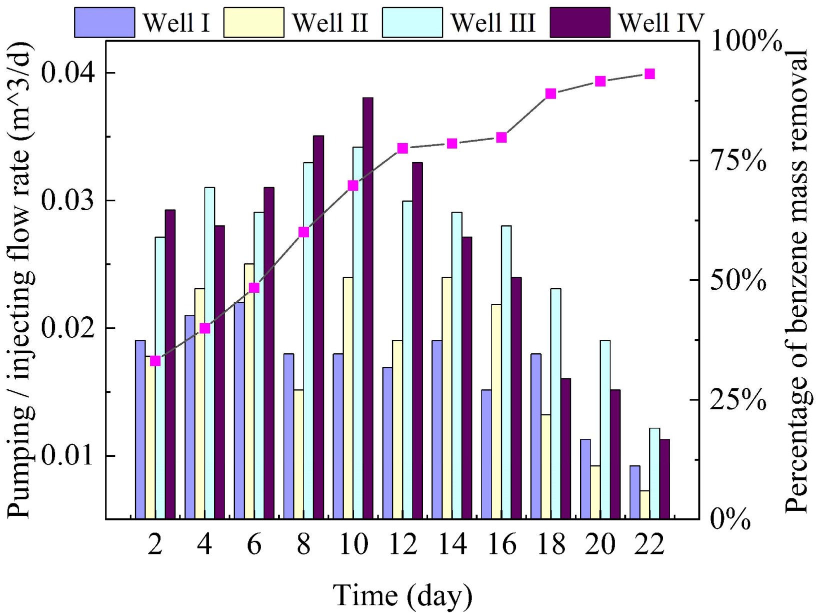


Figure 3. Optimal pumping/injecting rates of Well I to IV for the remediation duration and the corresponding percentage of benzene mass removal

Figure 4 presents the predicted remediation results of the DPC and SADPC systems from Day 2 to Day 22 after the leakage. It is shown that the contamination level has been reduced significantly through both systems. The benzene concentrations are over 0.5 mg/L at the initial stage and then transport and decrease gradually with groundwater flow. Therefore, the peak benzene concentrations at the upstream are gradually getting decreased over time to only 0.2 mg/L on day 22 in the DPC system, which has been removed around 60% of the initial contamination. The SADPC system focus on the biodegradation process from Day 12 to Day 22, so Figure 4 also shows the different levels between the DPC system and the SADPC system. The levels of benzene dispersion through the SADPC system are relatively lower compared with the DPC system since Day 14, and reach the degradation goal on Day 20. It indicates that the SADPC system has outstanding performance at benzene removal in the groundwater system.

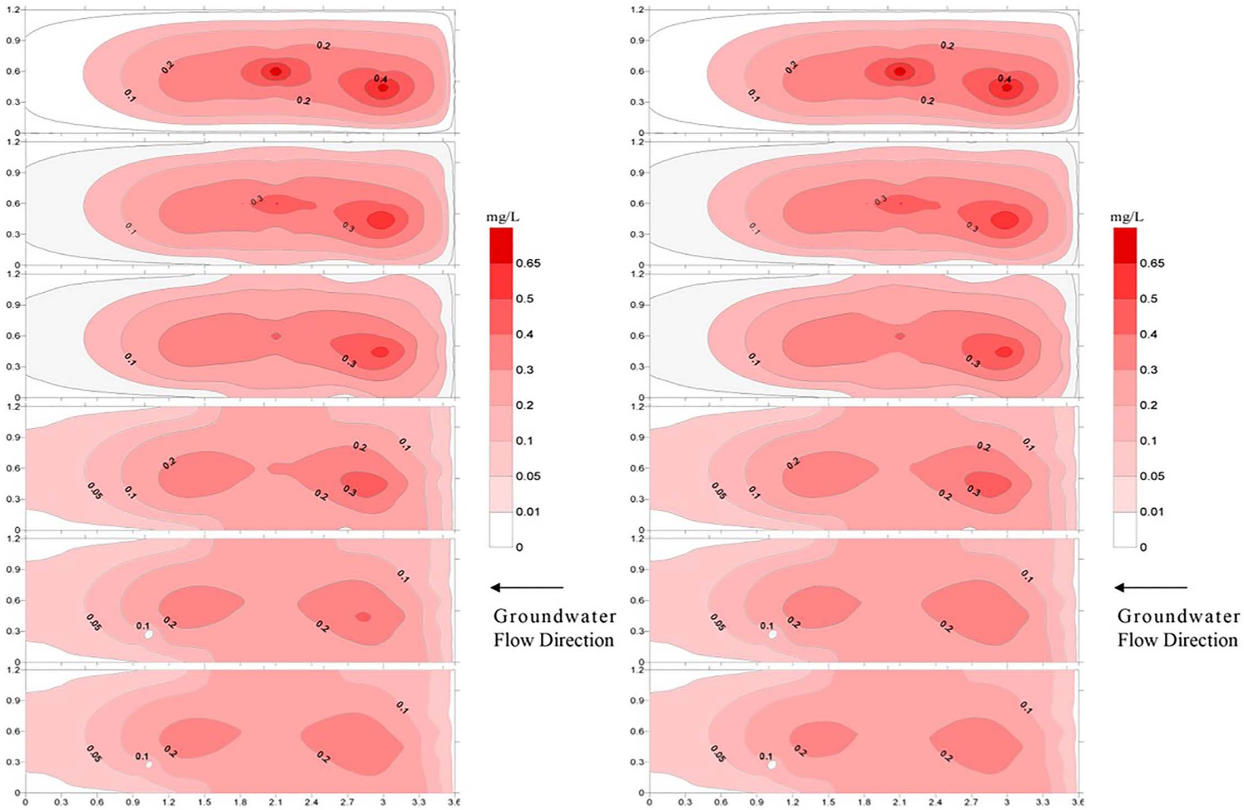


Figure 4. Benzene concentrations on Days 2, 6, 10, 14, 20 and 22 of the DPC system(a) and SADPC system(b)

3.3 Comparison analysis

To examine the remediation efficiency by the DPC system and the SADPC system, nine hypothetical wells (HWs) are selected from the simulation domain and the locations of these HWs can be seen in Figure S4. Figure 5 presents the benzene concentrations of the SADPC system at the nine HWs from Day 2 to Day 22 (Benzene concentrations of the DPC system is shown in Figure S5). In the DPC system, the analysis of the predicted data of benzene concentrations at the nine HWs indicates that the benzene concentrations during the first ten days of remediation decrease slowly or even increase at some locations, and the pumping/injecting rates has been increased to a certain degree accordingly. The signal of an increasing concentration of contaminants triggers the necessary adjustment of the operation. The predicted data also indicates decreases in the contaminant concentrations at almost all locations after ten days of operation, especially at HW-56, HW-102 and HW-106, which show sharp decreases on Day 10 or Day 14. The distances between HW-102/HW-106 and the contaminant source are the same, which can be the reason why these two sites show the similar trend during contaminant degradation. After second-stage rolling optimization, it can be seen that the benzene concentrations at the nine HWs decrease in a faster rate in the SADPC system since Day 12, and reach balances 2 days faster than those in the DPC system. The pumping/injecting rates has been decreased from Day 12 according to the instructions from both systems. Based on the entire degradation process, the benzene concentrations in HW-40, HW-42, HW-48 and HW-52 decline gently, which may due to the fact that these wells are relatively far from the contaminant source. Therefore, the benzene can be removed much easier than those sites that are located near the contaminant source.

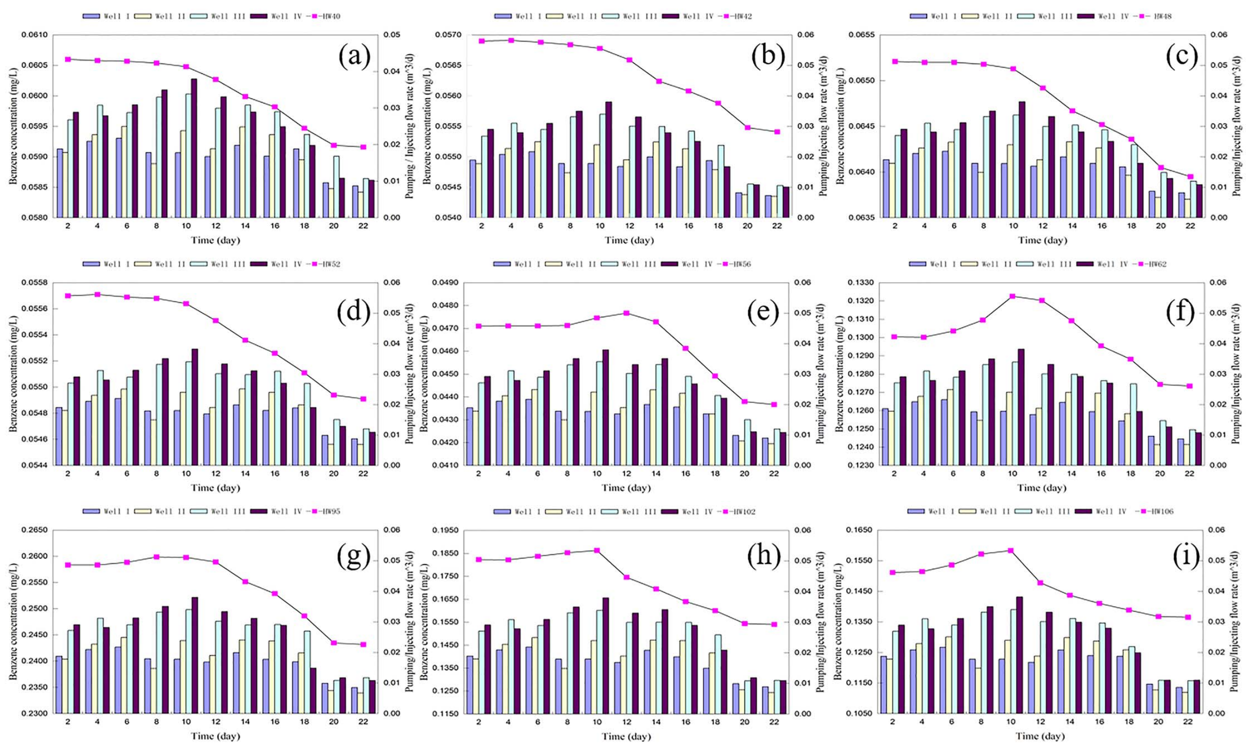


Figure 5. Benzene concentrations of the SADPC system from Day 2 to Day 22, where Figs. (a) to (i) represents the concentrations at HW-40, HW-42, HW-48, HW-52, HW-56, HW-62, HW-95, HW-102, and HW-106

The values of the percentage of benzene mass removal during the remediation process in the SADPC system can be seen in Figure 6. To compare the degradation effect in a straight way, benzene mass removal situations in the SI emulator, FCI optimizer and DPC system were also simulated. Each subgraph constructs an exponentially fitting curve with a 95% confidence band. The coefficient of determination values (R^2) show that all of the degrees of fitting of these models are satisfactory. The removal rate in the SI emulator is the minimum at the beginning of the remediation process, and then the percentage of benzene mass removal rises slowly with fluctuations until the remediation ends on Day 28. The FCI optimizer deals with benzene in groundwater at a relatively steady rate especially from Day 2 to Day 10, which is different from other three methods. Besides, it reaches the equilibrium stage with the removal rate around 82% on Day 26, only faster than the SI emulator. For the DPC system, it is found there is a "remediation plateau period" during the whole process (Day 12 to Day 16), while the percentage of benzene mass removal keeps growing and reaches the cleanup goal on Day 22. About 93% of the benzene mass has been removed by using the DPC system, which means the highest remaining benzene concentration anywhere in the simulation domain has been reduced to below 300 μg/L. Because the benzene removal rates of DPC system cannot meet the expectation from Day 11, an ideal setpoint curve is produced based on the data of the first six points in the DPC system. It should be noted that the SADPC method is applied from Day 14 to Day 28, and the setpoint curve of the SADPC system in Figure 6 is only used for the optimization of the seventh point. The SADPC system improves the biodegradation performance in the DPC system by using the rolling optimization model, which increases the removal rate during the "remediation plateau period" and arrives the cleanup goal on Day 20.

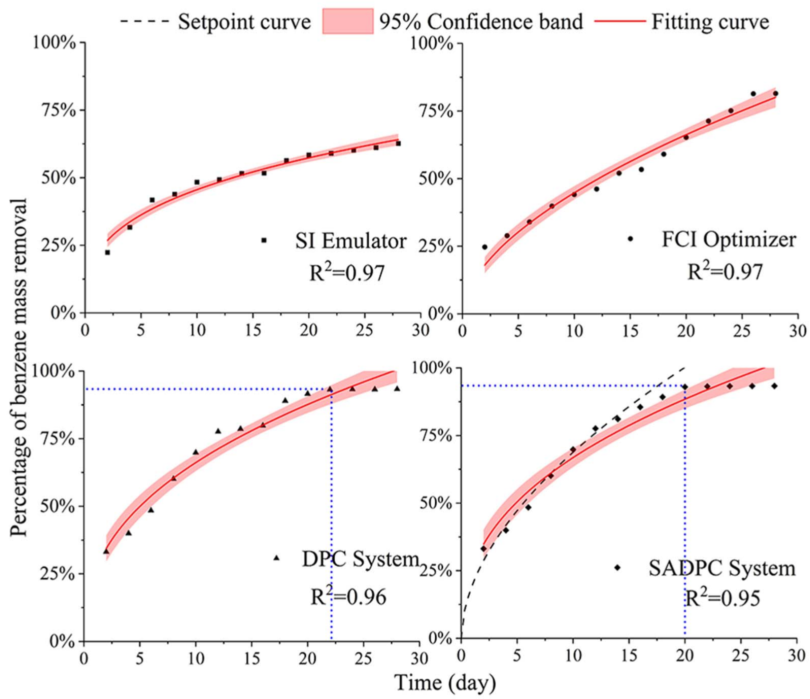


Figure 6. Comparison of predicted benzene removal efficiency by four control systems

The SI emulator only focuses on contamination degradation without considering the cost during the degradation process. For example, the SI emulator improves the pumping/injecting rate when the benzene concentration within the high level, so the removal rate is relatively high at the initial stage. However, the pumping/injecting rate can be reduced as the contaminant concentration deceases. Only 62.6% of the contaminant in the groundwater system has been removed by using the SI emulator which shows that it can hardly remove the benzene effectively. Besides, the whole degradation process in the SI emulator continues for 28 days, which imposes a heavy economical burden on actual stakeholders when applied to practical projects in the future. The optimization objective of the FCI optimizer mainly concentrates on the cost of the operation, which restricts the performance of the benzene removal. Although DPC system improves the degradation rate to some degree, it cannot always meet the expectations of people. In general, the remediation effect of the SADPC system is regarded as the best on the basis of the DPC system, because it fulfils the removal goal within a relatively short period and takes both cost and efficiency into consideration at the same time.

3.4 Policy implication

Bioremediation has demonstrated to be one of the most cost-effective technologies in organic-contaminated groundwater remediation. It can also be combined with other in-situ (or ex-situ) physical or chemical technologies to enhance remediation efficiency, shorten remediation duration, and reduce remediation cost. Historically, numerous studies have been undertaken by concentrating on advancement of new remediation technologies (e.g., adopting highly-efficient microbes and agents, designing new process flows, or optimizing operating conditions) just for improving remediation performance. Unfortunately, few of them attempted to use dynamic process approaches from the perspective of whole process to address the challenge. This study has suggested that a well-designed DPC system could be an easy-to-implement and strong-to-generalize approach compared to those conventional efforts. While this study focused on bioremediation process of benzene-contaminated groundwater, the developed SADPC system can be conveniently extended to many other cases no matter what one will need to challenge: organic or inorganic, physicochemical or biological, and water or soil. The major effort that needs to be accomplished is to construct a set of equations capable of capturing the relationships (frequently called proxy equations) between operating conditions and remediation performance. Nonetheless, as this study is a first attempt, much improvement will be desired, for example, simplifying the system framework to alleviate computational effort, strengthening the error information feedback to shorten decision duration, and introducing stochastic analysis to further mitigate the uncertainty impact.

AI plays an important role in running this SADPC system, which includes stepwise inference, stepwise filtered-clustering inference, genetic algorithm, etc. As conventional physically based models can hardly be directly used by SADPC considering the independence of prior assumptions for model forms (He et al., 2008), these machinery-learning similar inference methods are introduced to create a set of computation-fast and accuracy-reliable proxy equations to replace the conventional physical model. Note that a physical model is not a must by all the cases (this study uses the physical model aiming to generate a substantial number of statistical samples to obtain proxy equations). This implies that AI techniques have high potential to be used in remediation studies and practices because of their strong capabilities of convenient modeling (particularly in modeling highly nonlinear input-output relations), automatic learning, self-adaptation, and reliable generalization. It is desired that more state-of-the-art AI techniques be introduced in the SADPC system particularly including those intelligent denoising, error correction, and decision-making. This will much increase the performance of dynamic process control and enrich available control approaches.

An obvious knowledge gap implied in this study is the lack of related guidelines when designing a SADPC system. Historically, various guidelines on groundwater and soil remediation have been proposed in these past years at the national, provincial (or state-), and municipal levels. In terms of these guidelines, one can easily know what technologies can be used, how the remediation wells should be configured, which criteria should be satisfied after remediation, etc. Without the guidelines, there will be a difficulty in guaranteeing the stability, maturity, and reliability of a newly designed SADPC system, probably leading to extraordinary carefulness or even refuse of the potential users. This suggested that guidelines on remediation-oriented SADPC systems could be tentatively investigated, developed, and applied in future effort. This work is significant for offering users a set of principles or rules to follow when designing a SADPC system and helping them clarify the problems such as what procedures should be implemented when designing and running a SADPC system, how the control errors should be guaranteed during the whole remediation process, and what criteria could be adopted to evaluate the control performance.

4 Conclusions

A system was developed to improve the removal rate of benzene for enhancing in-situ bioremediation, which can effectively reflect the dynamics and complexity of the biodegradation process and realize the control for the remediation system based on the feedback information. The insights from this study can suggest to decision makers that guidelines and policies on remediation-oriented SADPC systems could be tentatively investigated, developed, and applied in future effort.

- Results from the error analysis of the contaminant fate and transport model show that the model agrees well with the data obtained from the pilot experiments, so it can be used for developing the SADPC system.
- The SADPC system is consist of an FCI simulator, a SI emulator, an MPC controller, a 1st-stage optimizer, and a 2nd-stage optimizer.
- The SADPC system improves the biodegradation performance in the DPC system by using the rolling optimization model, which increases the removal rate during the "remediation plateau period" and arrives at the cleanup goal earlier than the DPC system, as well as the SI emulator and the FCI optimizer.

Acknowledgments

The authors acknowledge funding from National Key Research and Development Program of China (Grant No. 2020YFC1807904).

References

Ahmed, W., & Rodriguez, J. (2020). A model predictive optimal control system for the practical automatic start-up of anaerobic digesters. *Water Research*, 174. 115599. https://doi.org/10.1016/j.watres.2020.115599.

Akyol, K. (2020). Comparing of deep neural networks and extreme learning machines based on growing and pruning approach. *Expert Systems with Application*, 140(Feb.),112875.1-112875.7. https://doi.org/10.1016/j.eswa.2019.112875.

Albers, C. N., Feld, L., Ellegaard-Jensen, L., & Aamand, J. (2015). Degradation of trace concentrations of the persistent groundwater pollutant 2,6-dichlorobenzamide (BAM) in bioaugmented rapid sand filters. *Water Research*, 83, 61-70. https://doi.org/10.1016/j.watres.2015.06.023.

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, 364(6439), 453-+. https://doi.org/10.1126/science.aav9436.

Bear, J. (1979). *Hydraulics of Ground Water*. Newyork: Dover Publications.

Bechet, Q., Shilton, A., & Guieysse, B. (2016), Maximizing Productivity and Reducing Environmental Impacts of Full-Scale Algal Production through Optimization of Open Pond Depth and Hydraulic Retention Time. *Environmental Science Technology*, 50(7), 4102-4110. https://doi.org/10.1021/acs.est.5b05412.

Chang, H. L., & Alvarez-Cohen, L. (1995). Model for the cometabolic biodegradation of chlorinated organics. *Environmental Science & Technology*, 29(9), 2357. https://doi.org/10.1021/es00009a031.

Chang, H. L., & Alvarez-Cohen, L. (2010). Transformation capacities of chlorinated organics by mixed cultures enriched on methane, propane, toluene, or phenol. *Biotechnology & Bioengineering*, 45(5), 440-449. https://doi.org/10.1002/bit.260450509.

Chiandussi, G., Codegone, M., Ferrero, S., & Varesio, F. E. (2012). Comparison of multi-objective optimization methodologies for engineering applications. Computers & Mathematics with *Applications*, 63(5), 912-942. https://doi.org/10.1016/j.camwa.2011.11.057.

de Blanc, P. C. (1998). Development and demonstration of a biodegradation model for non-aqueous-phase liquids in groundwater, (Doctoral dissertation). Austin, Texas: The University of Texas at Austin.

Delshad, M., Pope, G.A., & Sepehrnoori, K. (1996). A compositional simulator for modeling surfactant enhanced aquifer remediation, 1 formulation. *Journal of Contaminant Hydrology*, 23(4), 303-327. https://doi.org/10.1016/0169-7722(95)00106-9.

Diangelakis, N. A., Avraamidou, S., & Pistikopoulos, E. N. (2016). Decentralized Multiparametric Model Predictive Control for Domestic Combined Heat and Power Systems. *Industrial & Engineering Chemistry Research*, 55(12), 3313-3326. https://doi.org/10.1021/acs.iecr.5b03335.

Faust, C. R., Guswa, J. H., & Mercer, J. W. (1989). Simulation of three-dimensional flow of immiscible fluids within and below the unsaturated zone. *Water Resources Research*, 25(12), 2449-2464. https://doi.org/10.1029/WR025i012p02449.

Gerber, S., & Horenko, I. (2015). Improving clustering by imposing network information. *Science Advances*, 1(7). https://doi.org/10.1126/sciadv.1500163.

Gopalakrishnan, G., Minsker, B. S., & Valocchi, A. J. (2011). Monitoring Network Design for Phytoremediation Systems Using Primary and Secondary Data Sources. *Environmental Science & Technology*, 45(11), 4846-4853. https://doi.org/10.1021/es1042657.

Gorelick, S. M., & Zheng, C. M. (2015). Global change and the groundwater management challenge. *Water Resources Research*, 51(5), 3031-3051. https://doi.org/10.1002/2014WR016825.

Greenland, S., Daniel, R., & Pearce, N. (2016). Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. *International Journal of Epidemiology*, 45(2), 565-575. https://doi.org/10.1093/ije/dyw040.

Hao, P. Y., & Chiang, J. H. (2006). Pruning and model-selecting algorithms in the RBF frameworks constructed by support vector learning. *International Journal of Neural Systems*, 16(4), 283-293. https://doi.org/10.1142/S0129065706000688.

He, L., 2008. Development of integrated simulation and optimization models for petroleum-contaminated groundwater remediation management under various uncertainties, (Doctoral dissertation). Regina, Saskatchewan: The University of Regina.

He, L., Huang, G. H., & Lu, H. W. (2008a). Health-Risk-Based Groundwater Remediation System Optimization through Clusterwise Linear Regression. *Environmental Science & Technology*, 42(24), 9237-9243. https://doi.org/10.1021/es800834x.

He, L., Huang, G. H., Lu, H. W., & Zeng, G.M. (2008b). Optimization of surfactant-enhanced aquifer remediation for a laboratory BTEX system under parameter uncertainty.

Environmental Science & Technology, 42(6), 2009-2014. https://doi.org/10.1021/es071106y.

He, L., Huang, G. H., Zeng, G. M., Lu, & H. W. (2008c). An integrated simulation, inference, and optimization method for identifying groundwater remediation strategies at petroleum-contaminated aquifers in western Canada. *Water Research*, 42(10-11), 2629-2639. https://doi.org/10.1016/j.watres.2008.01.012.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* Ann Arbor: University of Michigan Press.

Hou, Z., Chi, R., & Gao, H. (2017). An Overview of Dynamic-Linearization-Based Data-Driven Control and Applications. *Ieee Transactions on Industrial Electronics*, 64(5), 4076-4090. https://doi.org/10.1109/TIE.2016.2636126.

Hu, Y., Patmonoaji, A., Xu, H., Kaito, K., & Suekane, T. (2021). Pore-scale investigation on nonaqueous phase liquid dissolution and mass transfer in 2d and 3d porous media. *International Journal of Heat and Mass Transfer*, 169, 120901. https://doi.org/10.1016/j.ijheatmasstransfer.2021.120901.

Huang, G. H., Huang, Y. F., Wang, G. Q., & Xiao, H.N. (2006). Development of a forecasting system for supporting remediation design and process control based on NAPL-biodegradation simulation and stepwise-cluster analysis. *Water Resources Research*, 42(6). https://doi.org/10.1029/2005WR004006.

Huang, Y. F., Wang, G. Q., Huang, G. H., Xiao, H. N., & Chakma, A. (2008). IPCS: An integrated process control system for enhanced in-situ bioremediation. *Environmental pollution*, 151(3), 460-469. https://doi.org/10.1016/j.envpol.2007.04.010.

Itoh, Y., & Adachi, M. (2017). *Reconstruction of bifurcation diagrams using an extreme learning machine with a pruning algorithm*, Paper presented at International Joint Conference on Neural Networks (IJCNN), Institute of Electrical and Electronics Engineers Inc., Anchorage, AK, United states.

Jacobs, O. L. R. (1993). *Introduction to Control Theory*. Oxford: Oxford University Press,.

Jimenez, N., Vinas, M., Sabate, J., Diez, S., & Bayona, J. M. (2006). The prestige oil spill. 2. enhanced biodegradation of a heavy fuel oil under field conditions by the use of an oleophilic fertilizer. *Environmental Science & Technology*, 40(8), 2578-2585. https://doi.org/10.1021/es052370z.

Kuo, J. T., Wang, Y. Y., & Lung, W. S. (2006). A hybrid neural–genetic algorithm for reservoir water quality management. *Water research*, 40(7), 1367-1376. https://doi.org/10.1016/j.watres.2006.01.046.

Lenczewski, M., Jardine, P., McKay, L., & Layton, A. (2003). Natural attenuation of trichloroethylene in fractured shale bedrock. *Journal of Contaminant Hydrology*, 64(3-4), 151-168. https://doi.org/10.1016/S0169-7722(02)00090-6.

Li, Y. S., Abriola, L. M., Phelan, T. J., Ramsburg, C.A., & Pennell, K. D. (2007). Experimental and numerical validation of the total trapping number for prediction of DNAPL mobilization. *Environmental Science & Technology*, 41(23), 8135-8141. https://doi.org/10.1021/es070834i.

Liang X., Devine C. E., Nelson J., Lollar, B. S., Zinder, S., & Edwards, E. A. (2013). Anaerobic conversion of chlorobenzene and benzene to CH4 and CO2 in bioaugmented microcosms. *Environmental Science & Technology*, 47(5): 2378-2385. https://doi.org/10.1021/es3043092.

Liao, R., Jin, Z., Chen, M., & Li, S. (2020). An integrated approach for enhancing the overall performance of constructed wetlands in urban areas. *Water Research*, 187, 116443. https://doi.org/10.1016/j.watres.2020.116443.

Liu, L., Zhou, Q., Liang, H. J., & Wang, L. J. (2017). Stability and Stabilization of Nonlinear Switched Systems Under Average Dwell Time. *Applied Mathematics and Computation*, 298, 77-94. https://doi.org/10.1016/j.amc.2016.11.006.

Liu, Y. Q., Ganigue, R., Sharma, K., & Yuan, Z. G. (2016). Event-driven model predictive control of sewage pumping stations for sulfide mitigation in sewer networks. *Water Research*, 98, 376-383. https://doi.org/10.1016/j.watres.2016.04.039.

Matott, L. S., Bartelt-Hunt, S. L., Rabideau, A. J., & Fowler, K. (2006). Application of heuristic optimization techniques and algorithm tuning to multilayered sorptive barrier design. *Environmental Science & Technology*, 40(20), 6354-6360. https://doi.org/10.1021/es052560+.

Maybeck, P. S. (1979). *Stochastic models, estimation and control.* (Vol. 1). London: Academic Press, INC.

Mayne, D. Q. (2014). Model predictive control: Recent developments and future promise. *Automatica* 50(12), 2967-2986. https://doi.org/10.1016/j.automatica.2014.10.128.

McDonald, M. G., & Harbaugh, A. W. (1988). *A modular three-dimensional finite-difference groundwater flow model*. Washington, DC: United States Government Printing Office (USGPO).

Meray, A. O., Sturla, S., Siddiquee, M. R., Serata, R., Uhlemann, S., Gonzalez-Raymat, H., Denham, M., Upadhyay, H., Lagos, L. E., Eddy-Dilek, C., & Wainwright, H. M. (2022). PyLEnM: A Machine Learning Framework for Long-Term Groundwater Contamination Monitoring Strategies. *Environmental Science & Technology*, 56(9), 5973-5983. https://doi.org/10.1021/acs.est.1c07440.

Miller, T. H., Baz-Lomba, J. A., Harman, C., Reid, M. J., Owen, S. F., Bury, N. R., Thomas, K. V., & Barron, L. P. (2016). The First Attempt at Non-Linear in Silico Prediction of Sampling Rates for Polar Organic Chemical Integrative Samplers (POCIS). *Environmental Science & Technology*, 50(15), 7973-7981. https://doi.org/10.1021/acs.est.6b01407.

Niswonger, R. G., & Prudic, D. E. (2013). *Modeling variably saturated flow using kinematic waves in MODFLOW*. American Geophysical Union.

Opher, T., & Ostfeld, A. (2011). A coupled model tree (MT) genetic algorithm (GA) scheme for biofouling assessment in pipelines. *Water research*, 45(18), 6277-6288. https://doi.org/10.1016/j.watres.2011.09.037.

Passino, K. M. (2002). Biomimicry of bacterial foraging for distributed optimization and control. *Ieee Control Systems Magazine* 22(3), 52-67. https://doi.org/10.1109/mcs.2002.1004010.

Pizzagalli, D. U., Gonzalez, S. F., & Krause, R. (2019). A trainable clustering algorithm based on shortest paths from density peaks. *Science Advances*, 5(10). https://doi.org/10.1126/sciadv.aax3770.

Rao, C. R. (1952). Advanced statistical methods in biometric research. *Biometrics*, 28(1), 253. https://doi.org/10.2307/2528984.

Rittmann, B. E., Henry, B., Odencrantz, J. E., & Sutfin, J. A. (1991). Biological fate of a polydisperse acrylate polymer in anaerobic sand-medium transport. *Biodegradation*, 2(3), 171-179. https://doi.org/10.1007/BF00124491.

Schaerlaekens, J., Carmeliet, J., & Feyen, J. (2005). Multi-objective optimization of the setup of a surfactant-enhanced DNAPL remediation. *Environmental Science & Technology*, 39(7), 2327-2333. https://doi.org/10.1021/es049148z.

Shen, H., Li, F., Wu, Z. G., Park, J. H., & Sreeram, V. (2018). Fuzzy-Model-Based Nonfragile Control for Nonlinear Singularly Perturbed Systems With Semi-Markov Jump Parameters. *Ieee Transactions on Fuzzy Systems*, 26(6), 3428-3439. https://doi.org/10.1109/TFUZZ.2018.2832614.

Siade, A. J., Cui, T., Karelse, R. N., & Hampton, C. (2020). Reduced-Dimensional Gaussian Process Machine Learning for Groundwater Allocation Planning Using Swarm Theory. *Water Resources Research*, 56(3). https://doi.org/10.1029/2019wr026061.

Stentoft, P. A., Munk-Nielsen, T., Moller, J. K., Madsen, H., Valverde-Perez, B., Mikkelsen, P. S., & Vezzaro, L. (2021). Prioritize effluent quality, operational costs or global warming? - Using predictive control of wastewater aeration for flexible management of objectives in WRRFs. *Water Research*, 196. https://doi.org/10.1016/j.watres.2021.116960.

Stramer, Y., Brenner, A., Cohen, S. B., & Oron, G. (2010). Selection of a Multi-Stage System for Biosolids Management Applying Genetic Algorithm. *Environmental Science & Technology*, 44(14), 5503-5508. https://doi.org/10.1021/es902981t.

Sun, J., Miao, Z., Gong, D. W., Zeng, X. J., Li, J. Q., & Wang, G. G. (2020). Interval Multiobjective Optimization With Memetic Algorithms. *Ieee Transactions on Cybernetics*, 50(8), 3444-3457. https://doi.org/10.1109/tcyb.2019.2908485.

Sun, K., Yu, Y., & Huang, Z. (2015). *A Generalized Pruning Algorithm for Extreme Learning Machine*. Paper presented at IEEE International Conference on Information and Automation 2015, Lijiang, PEOPLES R CHINA.

Umar, M. F., Rafatullah, M., Abbas, S. Z., Ibrahim, M., & Ismail, N. (2021). Enhanced benzene bioremediation and power generation by double chamber benthic microbial fuel cells fed with sugarcane waste as a substrate. *Journal of Cleaner Production*, 310(4), 127583. https://doi.org/10.1016/j.jclepro.2021.127583.

Wang, W. J., Tian, G. D., Chen, M. N., Tao, F., & Zhang, C. Y., Ai-Ahmari, A., Li, Z. W., & Jiang, Z. G. (2020). Dual-objective program and improved artificial bee colony for the optimization of energy-conscious milling parameters subject to multiple constraints. *Journal of Cleaner Production*, 245. https://doi.org/10.1016/j.jclepro.2019.118714.

Wolicka, D., Suszek, A., Borkowski, A., & Bielecka, A. (2009). Application of aerobic microorganisms in bioremediation in situ of soil contaminated by petroleum products. *Bioresource Technology*, 99(13), 3221-3227. https://doi.org/10.1016/j.biortech.2009.02.020.

Xin, B. P., Wu, C. H., Wu, C. H., & Lin, C. W. (2013). Bioaugmented remediation of high concentration btex-contaminated groundwater by permeable reactive barrier with immobilized bead. *Journal of Hazardous Materials*, 244-245(JAN.15), 765-772. https://doi.org/10.1016/j.jhazmat.2012.11.007.

Yang, C., Liu, S., Su, Y., Chen, Y., Lin, C., & Lin, K. (2019). Bioremediation capability evaluation of benzene and sulfolane contaminated groundwater: determination of bioremediation parameters. *The Science of the Total Environment*, 648, 811-818. https://doi.org/10.1016/j.scitotenv.2018.08.208.

Zeng, J., & Liu, J. F. (2015). Economic Model Predictive Control of Wastewater Treatment Processes. *Industrial & Engineering Chemistry Research*, 54(21), 5710-5721. https://doi.org/10.1021/ie504995n.

Zhan, K., Zhang, C. Q., Guan, J. P., & Wang, J. S. (2018). Graph Learning for Multiview Clustering. *Ieee Transactions on Cybernetics*, 48(10), 2887-2895. https://doi.org/10.1109/tcyb.2017.2751646.

Zhang, C., Wu, D. J., & Ren, H. X. (2020). Bioremediation of oil contaminated soil using agricultural wastes via microbial consortium. *Scientific Reports*, 10(1). https://doi.org/10.1038/s41598-020-66169-5.

Zhang, C., Yoon, H., Werth, C. J., Valocchi, A. J., Basu, N. B., & Jawitz, J. W. (2008). Evaluation of simplified mass transfer models to simulate the impacts of source zone architecture on nonaqueous phase liquid dissolution in heterogeneous porous media. *Journal of Contaminant Hydrology*, 102(1-2), 49-60. https://doi.org/10.1016/j.jconhyd.2008.05.007.

Zou, Y., Huang, G. H., & Nie, X. H. (2009). Filtered Stepwise Clustering Method for Predicting Fate of Contaminants in Groundwater Remediation Systems: A Case Study in Western Canada. *Water Air and Soil Pollution*, 199(1-4), 389-405. https://doi.org/0.1007/s11270-008-9887-5.

Figure1.


Optimization Loop
FCI Simulator
$X_p(t+1)$
1st-Stage Optimizer
$U'(t)$
$U(t)$
$X(t)$
$U^{①*}(t)$
$U^{①*}(t)$
SI Emulator
MPC Controller
$U^{②*}(t)$
Biodegradation Process
$X^*(t+1)$
$U^{①*}(t)$
FCI Simulator
Optimization Loop
$X(t)$
$X_p(t+1)$
$U'(t)$
-
Setpoint Curve
$X_r(t+1)$
+
$e(t+1)$
2nd-Stage Optimizer
$U^{②*}(t)$

(a) (b)

Figure2.

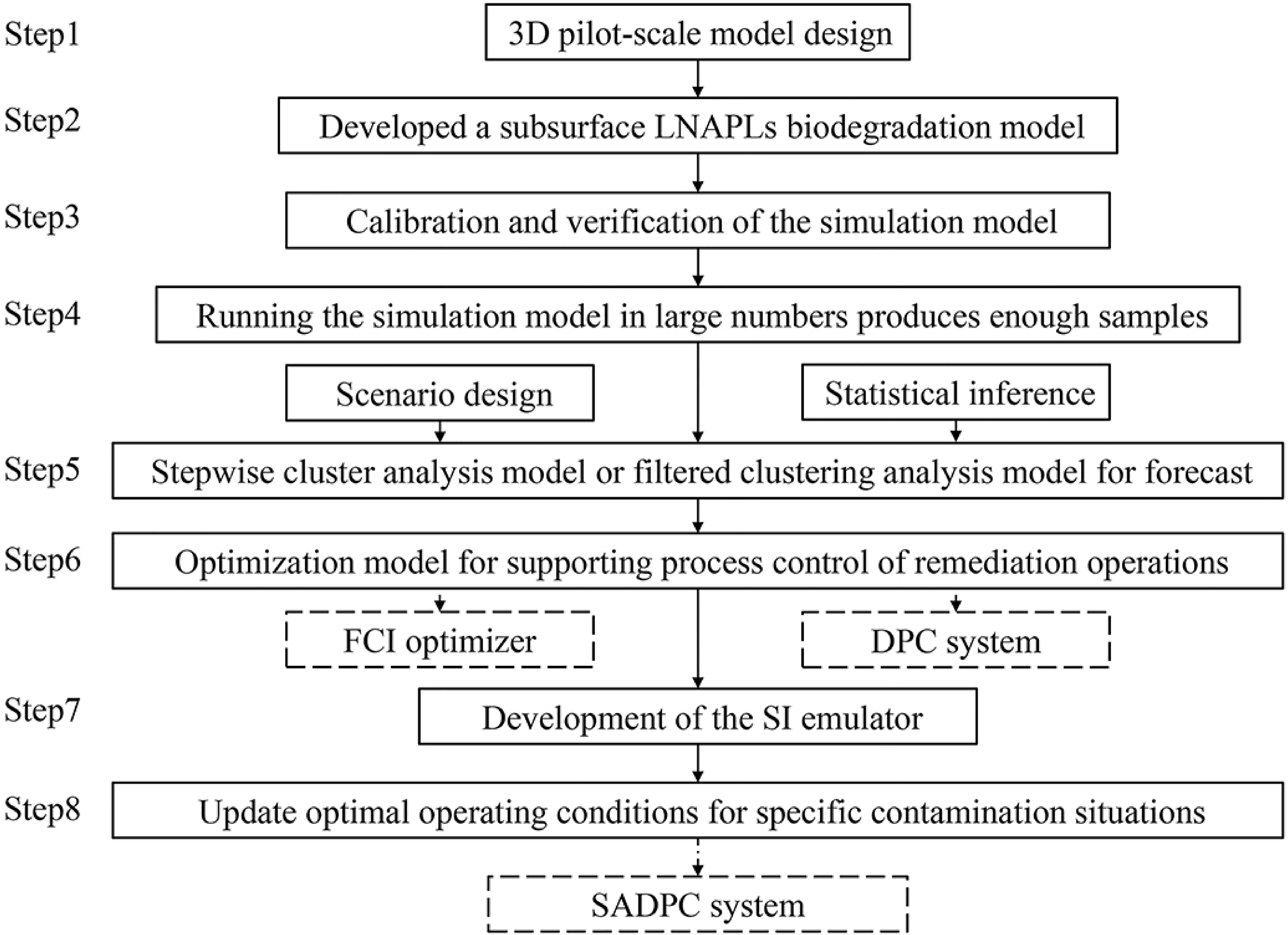
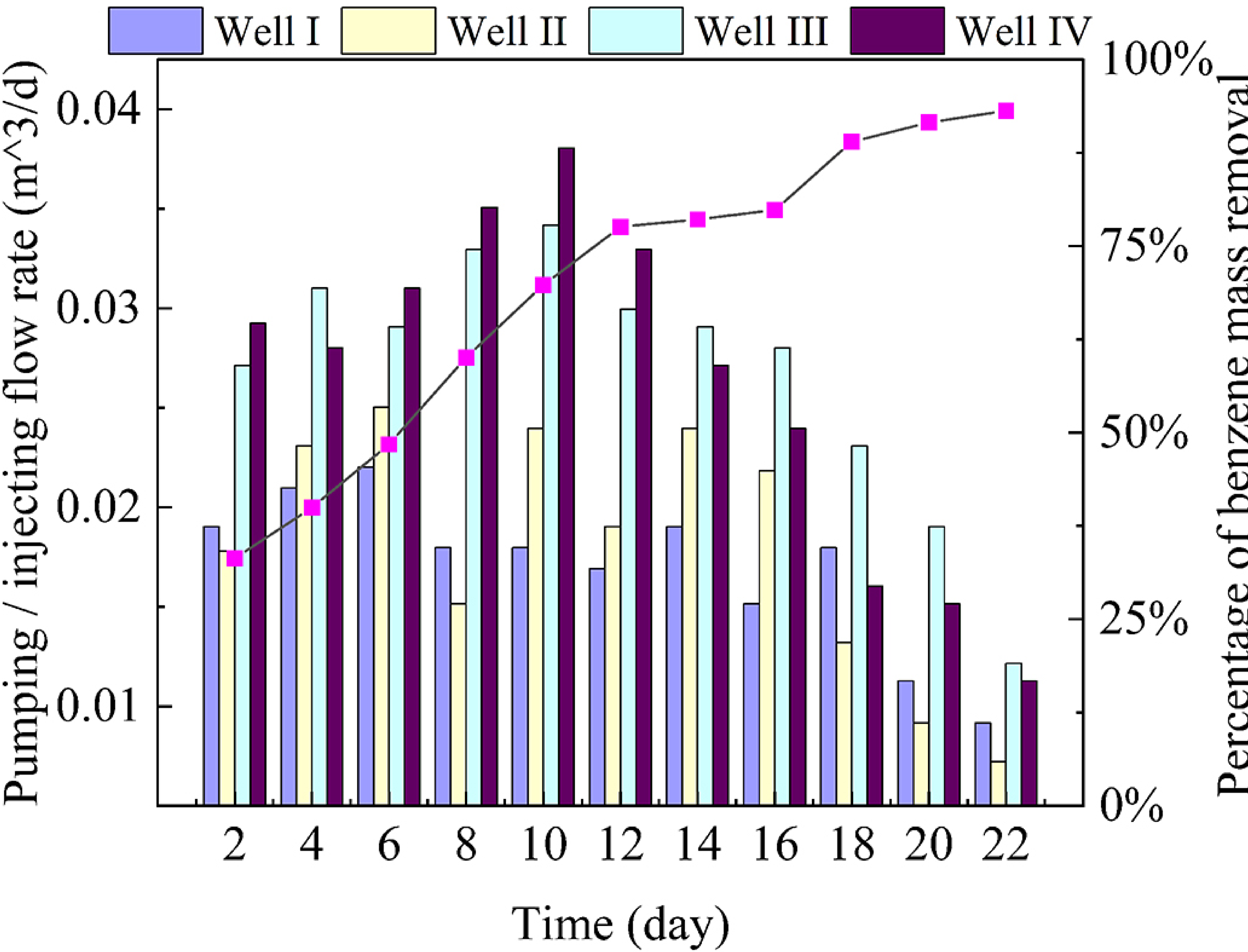

Step1
3D pilot-scale model design
Step2
Developed a subsurface LNAPLs biodegradation model
Step3
Calibration and verification of the simulation model
Step4
Running the simulation model in large numbers produces enough samples
Scenario design
Statistical inference
Step5
Stepwise cluster analysis model or filtered clustering analysis model for forecast
Step6
Optimization model for supporting process control of remediation operations
FCI optimizer
DPC system
Step7
Development of the SI emulator
Step8
Update optimal operating conditions for specific contamination situations
SADPC system

Figure3.

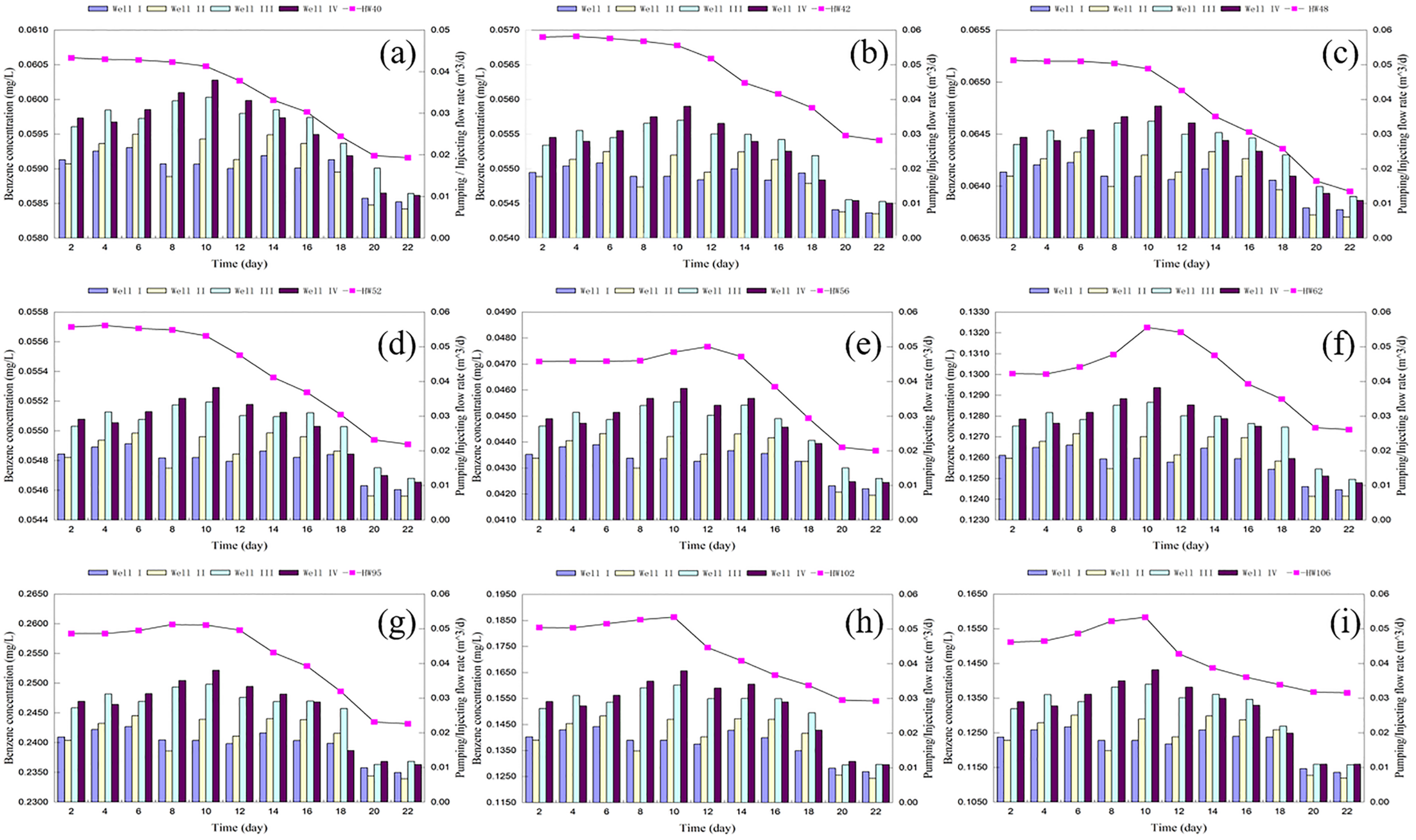


Well I
Well II
Well III
Well IV
Pumping / injecting flow rate (m^3/d)
0.01
0.02
0.03
0.04
Percentage of benzene mass removal
0%
25%
50%
75%
100%
2
4
6
8
10
12
14
16
18
20
22
Time (day)

Figure4.

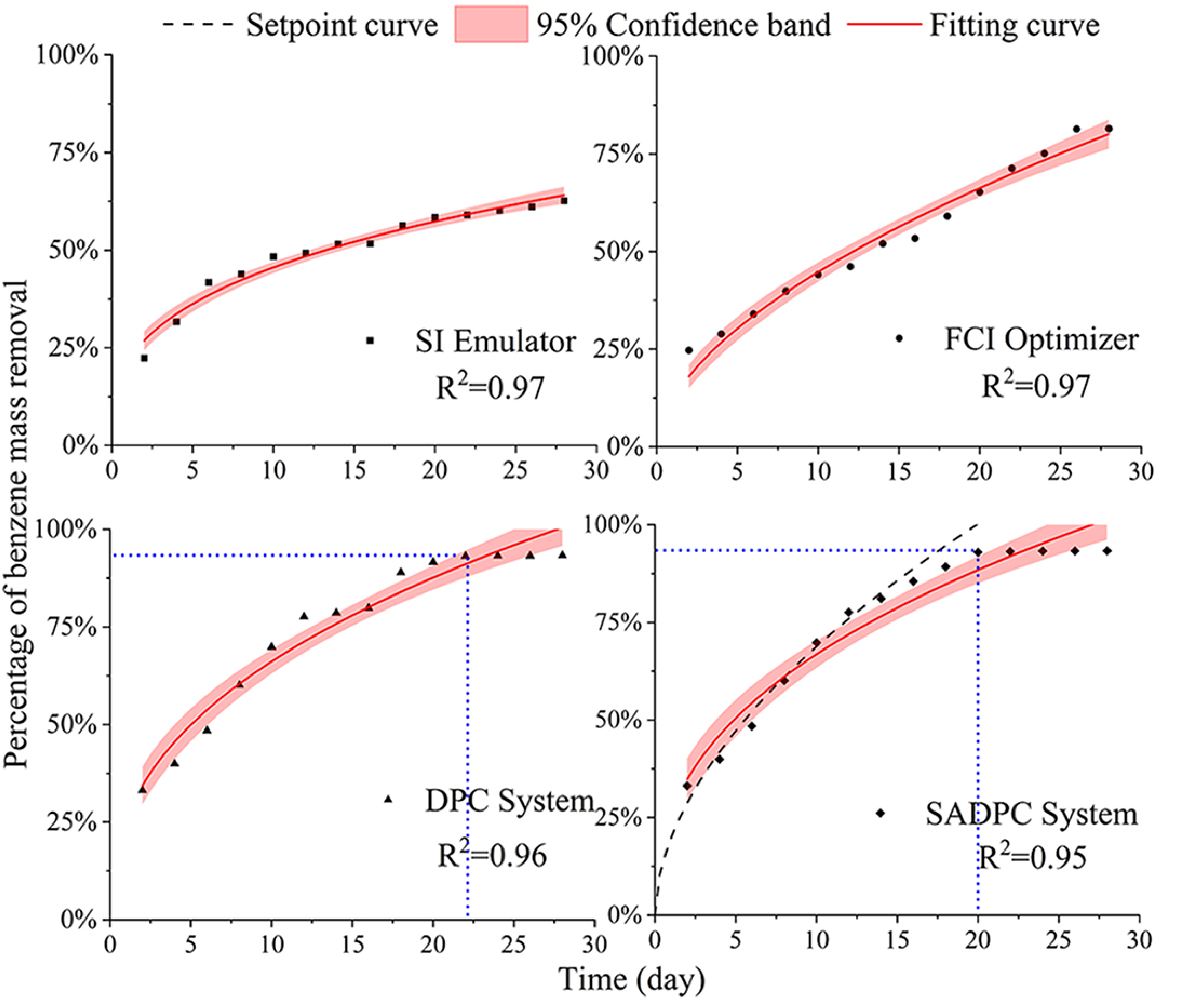

mg/L
0.65
0.5
0.4
0.3
0.2
0.1
0.05
0.01
0
Groundwater Flow Direction
mg/L
0.65
0.5
0.4
0.3
0.2
0.1
0.05
0.01
0
Groundwater Flow Direction

Figure5.



Well I
Well II
Well III
Well IV
HW40
(a)
HW42
(b)
HW48
(c)
HW52
(d)
HW56
(e)
HW62
(f)
HW95
(g)
HW102
(h)
HW106
(i)
Benzene concentration (mg/L)
Pumping / Injecting flow rate (m^3/d)
Pumping/Injecting flow rate (m^3/d)
Time (day)

Figure6.



Setpoint curve
95% Confidence band
Fitting curve
Percentage of benzene mass removal
Time (day)
SI Emulator
R^2=0.97
FCI Optimizer
R^2=0.97
DPC System
R^2=0.96
SADPC System
R^2=0.95
100%
75%
50%
25%
0%
0
5
10
15
20
25
30

Water Resources Research

Supporting Information for

An Artificial Intelligence Based Self-Adaptive Dynamic Process Control System for Enhancing In-Situ Bioremediation of Benzene-Contaminated Groundwater

Li He[1*†], Xu Duan[1†], Chenyang Li[1], Xu Yang[1], Mengxi He[1]

[1]State Key Laboratory of Hydraulic Engineering Simulation and Safety, Tianjin University, Yaguan Road 135, Haihe Education Park, Jinnan District, Tianjin, 300350, China

*Corresponding author: Li He (helix111@tju.edu.cn)

Contents

Text S1. Set-up of the pilot-scale reactor

The modeling domain is defined as three-dimensional (3-D), with the contaminated zone around the groundwater table considered as the major pollution source. The area of the simulation domain is 3.6 × 1.2 m^2. Vertically, the simulation domain is discretized into four grid blocks corresponding to four simulation layers; each layer is located in the middle of the grid block that facilitates the application of a block-centered finite difference scheme. In the horizontal plain, each layer is discretized into 24 × 8 grids. Each grid has dimensions of 0.15, 0.15, and 0.30 m in x, y, and z directions, respectively. The total number of grids in this 3-D computational system is 768 (24 × 8 × 4). Layers 3 and 4 are located in the saturated zone, while layers 1 and 2 are situated in the unsaturated zone. The detailed views of the pilot-scale reactor can be seen in Figure S1(a)-(d) of the supporting information.

Three soil distributions in the four layers are shown in Figure S1(f). The monitoring wells were used to obtain the subsurface hydrogeology "representative" view (well locations are shown in Figure S1(e)) The LNAPLs (Light Non-Aqueous Phase Liquids) initially occupied a contaminated area in layers 3 and 4 around the groundwater table. The zero-flow boundary conditions were enforced at the top and bottom of the modeling domain, as well as at the sides parallel to the x-axis. Constant hydraulic heads were employed at the left and right boundaries, allowing continuous water flow in the aquifer. Benzene concentrations in the system can be forecasted through the developed simulator. Table S1 presents the input parameters for the simulation model.

Water-table level gauges were installed in the first and fourth sections to monitor water depths inside the reactor. Observation windows were built on the front side of each section while another on the top. The side windows were used to observe the subsurface conditions, and the top ones to observe the soil surface. The four sections were connected to each other with flanges, each of which had 44 bolts. Gaskets made of anti-organic solvent and anti-high temperature rubber and silicone pastern were placed between the flanges to prevent the leakages.

For soils loading, the pre-selected clay was from the construction site of the Saskatchewan Indian Federated College, Regina, Saskatchewan, at depths of 2 to 6 m from the ground surface. The clayey till and fine sand were provided by the Waxy's Bobcat and Landscaping Ltd. The initial properties of soils can be seen in Table S2. The value of soil organic carbon (SOC) in the soils is measured at around 1.14%. Some activities around the concentration site accelerate the decomposition of the organic carbon, and there is little vegetation and insufficient organic carbon input, resulting in the particularly low concentration. Therefore, the transport of organic carbon is not considered in this model, which is assumed to migrate with the movement of the organic phase. We thus ignored the effect of SOCs in this study, also considering the high concentration of benzene contamination, as treated by many existing studies (Jimenez et al., 2006; Wolicka et al., 2009; Xin et al., 2013; Umar et al., 2021; Yang et al., 2019). For instance, Liang et al (2013) mentioned that benzene can be used as the sole source of organic carbons when discussing the chlorobenzene and benzene degradation in the groundwater. Besides we did not directly consider the effect of the grain size, instead of using porosity and intrinsic permeability when describing the transport of the target contaminant in the mathematical model. We did it by hypothesizing the gran size can affect soil porosity and intrinsic permeability and have a further impact on the fate and transport of benzene. This method is generally used in many groundwater models such as MODFLOW and NAPL3D (McDonald & Harbaugh, 1988; Zhang et al., 2008;

Niswonger & Prudic, 2013; Hu et al., 2021). Detailed information and guides can be seen in USGS website (https://www.usgs.gov/mission-areas/water-resources/science/modflow-and-related-programs) and EPA website (https://www.epa.gov/water-research/non-aqueous-phase-liquid-napl-simulator).

The inner surface of the reactor wall was labeled and divided into grids where different types of soils were loaded. Clay and clay-till were sieved by 1/4 inch sieving meshwork before loading. Sand was firstly loaded followed by clay-till and then clay. Upon the completion of every 100 mm depth in every grid, tap water was spread on the surface, and the soil in the grid was vibrated by concrete vibrator and pressed by impinging a hammer on a wood board that directly contacted the soil surface to ensure homogeneity and non-fracture structure. Upon the completion of each layer, more water was spread and the layer was left overnight to settle. After the soil loading process was completed, the system was then left still for three months with a water flow of 10 L/day through the loaded soils. No noticeable further settlement or consolidation was observed afterwards.

A thermostatic room, in which the pilot-scale system and the accessorial equipment were assembled, was built to realize various temperatures by an air conditioner. Water and drainage containers were each connected to the upstream inlet and downstream outlet, respectively. Water level gauges were used to show the depth of water table in the reactor. Tap water in a water container was pumped into the reactor through six water inlets on the inlet-end board as upstream groundwater inflow through a peristaltic pump. Before the start of the experiments, water in the container was kept still overnight to reach the room temperature. A 7-day buffering time preceded all experiments in the pilot model so that the temperature at every location in the reactor could reach equilibrium. The upstream water was kept flowing through system to acquire the desired soil temperature.

The monitoring wells are for facilitating access to the groundwater so that a "representative" view of the subsurface hydrogeology can be obtained, either through the collection of water samples or the measurement of physical and hydraulic parameters. In this study, a few monitoring wells were also used for pumping and injecting purposes during the remediation processes. Locations of the wells are presented in Figure S1(e). There are 25 wells allocated in four sections of the pilot system. Soil in the system was stratified into four layers, with the third and fourth layers being saturated with water. Each layer is 30 cm deep. Among the wells, 13 of them (with PVC pipes) were installed to reach the third soil layer; the other 12 wells could reach the fourth layer (Figure S1(f)). Small holes were uniformly made around the bottom sections of the pipes. Screens were used to wrap the pipes to prevent from soil clogging. Soil particles were prevented from moving into the wells while the groundwater could infiltrate into them. The wells were sealed by rubber caps at the tops. For each well, a hose was installed that passed through the caps and reached its bottom. The outside of the hose was clamped by a clip so that air and groundwater in the well were isolated from the atmosphere.

Text S2. Pilot-scale experimentation analysis

Benzene concentrations monitored in both natural-attenuation and bioremediation phases are listed in Table S3 and Figure S6(a-b) of the Supporting Information. The highest benzene concentrations were encountered in well 6 during the entire period of the natural-attenuation phase. This was due to the fact that well 6 was close to the leakage source. High concentrations were also observed in wells 3 and 10, which were placed in the third layer. The contaminant can easily reach these wells along with groundwater flow since the leakage occurred at the top of the third layer. Moreover, much of the contaminant transported within the third layer and did not migrate to the fourth one since gasoline is lighter than water. In comparison, the highest concentrations in the fourth layer were observed in well 5, which was installed in the sand zone near the leakage source. Due to the low porosity and permeability of silty and clayey soils, benzene was not observed in the down gradient domain of the pilot system until day 32 (the contaminant reached well 16 on day 32).

On day 40, enhanced in-situ biodegradation action was undertaken. It is shown from Table S3 of the Supporting Information that the benzene concentrations vary greatly due to flow-condition changes resulting from the pumping and injecting actions. The location of the peak concentration moved towards the downstream. The benzene concentrations in the groundwater also decreased greatly compared with those in the earlier periods. The peak benzene concentration decreased from 7.34 mg/L at the beginning of the remediation program to 0.633 mg/L on day 17 after the remediation action started. It is indicated that the enhanced in-situ bioremediation had efficiency in removing benzene from the groundwater.

The experimental results indicated that the developed pilot-scale reactor can effectively facilitate the simulation of both natural attenuation and enhanced remediation processes. The experimental results can be used for validating, calibrating and verifying the developed numerical model under different site conditions.

Calibration and verification of the developed biodegradation model were undertaken using data obtained from the pilot experiments. The results of the error analysis are provided in Table S4 of the Supporting Information. The absolute errors between the simulated and observed concentrations of 12 wells range from 0.00 to 0.40 mg/L with a mean of 0.21 mg/L. The root-mean-square error is 0.27 mg/L, and the correlation coefficient is 0.93. According to the error analysis, the biodegradation simulation result has been proved to be within a relatively reasonable range. Figure S7 shows the verification results for Day 57, which is the end of phase I bioremediation. The concentration distributions of benzene are generally the same based on observed data and simulated data. The highest concentration levels are obtained at the bottom right corner of the experiment region, and the coordinates of these points are around (2.1, 0.6) and (3, 0.45). The verification results for the temporal variations of benzene concentrations in well 5 and well 6 are shown in Figure S8, which indicates that this model can simulate the actual degradation process of benzene properly. After the calibration and verification, the simulation model can be used for investigating the effects of different bioremediation strategies on benzene concentrations.

Text S3. Contaminant fate and transport modeling in the groundwater

A critical step in understanding the impact of a subsurface release of NAPL is a modeling analysis of the NAPL flow and transport and fate of its crucial constituents. A complete description of multiphase flow and transport in subsurface must include flow of the fluid phases (water, gas, NAPL, etc.), mass transfer of species between these phases, and transport of species in each phase. A three-dimensional multiphase multicomponent SEAR model is recognized as an effective tool in investigating complex physical processes involved in NAPLs flow and transport. Several models have been developed to simulate the flow of multiple fluid phases in subsurface during recent years. All these models included simplifying assumptions with respect to phase presence and dimensionality, NAPL contaminant mass balance and water and air in subsurface. Important assumptions used in the development of mass conservation equations are: 1) the solid phase is immobile; 2) soil and fluids are slightly compressible; 3) dispersion is of Fickian form; 4) components mix ideally; 5) Darcy's law applies in the calculation of phase velocities. The basic mass conservation equation for components in subsurface can be written as (Delshad et al., 1996):

$$\frac{\partial}{\partial t}(\phi \widetilde{C}_k \rho_k) + \vec{\nabla} \cdot [\sum_{l=1}^{n_p} \rho_k (C_{kl} \vec{u}_l - \phi S_l \vec{\vec{D}}_{kl} \cdot \vec{\nabla} C_{kl})] = R_k \tag{S1}$$

where k is the component index; l is the phase index; ϕ is the soil porosity; $\tilde{C}_k$ is the overall concentration of component k (volume fraction); ρ_k is the density of component k [ML^{-3}]; n_p is the number of phases; C_{kl} is the concentration of component k in phase l (volume fraction); $\vec{u}_l$ is the Darcy velocity of phase l [LT^{-1}]; S_l is the saturation of phase l; R_k is the total source/sink term for component k (volume of the component k per unit volume of porous media per unit time); $\vec{\vec{D}}_{kl}$ is the dispersion tensor. The overall concentration ($\tilde{C}_k$) denotes the volume of the component k summed over all phases.

The dispersion tensor ($\vec{\vec{D}}_{kl}$) can be expressed as follows (Bear, 1979):

$$\vec{\vec{D}}_{klij} = \frac{D_{m,kl}}{\tau}\delta_{ij} + \frac{\alpha_{Tl}}{\phi S_l}|\vec{u}_l|\delta_{ij} + \frac{(\alpha_{Ll} - \alpha_{Tl})}{\phi S_l}\frac{u_{li}u_{lj}}{|\vec{u}_l|} \tag{S2}$$

where τ is tortuosity (defined as a value greater than 1); $D_{m,kl}$ is molecular diffusion coefficient of component k in phase l [L^2T^{-1}]; δ_{ij} is Kronecker delta function; α_{Ll} and α_{Tl} are longitudinal and transverse dispersivities of phase l, respectively [L]; u_{li} and u_{lj} are Darcy velocities of phase l in directions i and j, respectively [LT^{-1}]; $|\vec{u}_l|$ is magnitude of the vector flux for phase l [LT^{-1}]. The phase flux can be calculated from the multiphase form of the Darcy's law (Faust et al., 1989):

$$\vec{u}_l = -\frac{k_{rl}\vec{\vec{K}}}{\mu_l} \cdot (\vec{\nabla} P_l - \rho_l g \vec{\nabla} z) \tag{S3}$$

where k_{rl} is relative permeability of porous medium to phase l; $\vec{\vec{K}}$ is intrinsic permeability tensor [L^2]; μ_l is viscosity of phase l [$ML^{-2}T^{-1}$]; ρ_l is density of phase l [ML^{-3}]; g is acceleration of gravity [LT^{-2}]; z is vertical distance which is defined as positive downward [L]; P_l is pressure of phase l [$ML^{-1}T^{-2}$].

Grumberg and Nissan's correlation is used to calculate the NAPL viscosity as a function or organic species concentration:

$$\ln \mu_{2l} = \sum_{k=1}^{n_o} x_{kl}^o \ln \mu_k^o \tag{S4}$$

where μ_{2l} is the organic mixture viscosity; x_{kl}^o is molar fraction of each organic component in phase l (water, NAPL, etc.); μ_k^o is the viscosity of single organic component; and n_o is the number of organic components in NAPL. For a NAPL mixture, the overall organic hydrostatic pressure gradient is obtained by assuming ideal mixing:

$$C_{2l}\gamma_{2l} = \sum_{k=1}^{n_o} C_{kl}^o \gamma_{kl}^o \tag{S5}$$

where C_{2l} is concentration of NAPL mixture in phase l (water, NAPL, etc.); γ_{2l} is density of NAPL mixture in phase l; C_{kl}^o is concentration of each organic component in phase l; γ_{kl}^o is density of single organic component in phase l; and n_o is the number of organic components in NAPL.

The aquifer boundaries are modeled as either constant potential surfaces or closed surfaces. The model can be solved numerically through the block-centered finite difference method. The solution method for the contaminant-transport model is the implicit pressure-explicit saturation method. The only unknown in the pressure equation is the pressure of water phase.

Text S4. Biodegradation modeling of contaminants in the groundwater

Generally, the biodegradation model involves simulation of substrate competition, nutrient limitation, product toxic inhibition, and aerobic cometabolism. The basic structure of the biodegradation model for a system with single substrate, single electron acceptor and single biological species can be characterized as follows (de Blanc, 1998):

$$\frac{dC^{APS}}{dt} = -\frac{A^{SM}k^{SMT}\overline{C^{AAB}}}{m^{CSM}}\left(C^{APS} - \overline{C^{ABS}}\right) - \frac{\mu_{max}C^{AUB}}{k^{Y}}\left(\frac{C^{APS}}{k^{SHS} + C^{APS}}\right)\left(\frac{C^{APE}}{k^{EHS} + C^{APE}}\right) - k^{FRR}C^{APS} \tag{S6}$$

$$\frac{d\overline{C^{ABS}}}{dt} = \frac{A^{SM}k^{SMT}}{V^{SM}}(C^{APS} - \overline{C^{ABS}}) - \frac{\mu_{max}\rho_{X}}{k^{Y}}\left(\frac{\overline{C^{ABS}}}{k^{SHS} + \overline{C^{ABS}}}\right)\left(\frac{\overline{C^{ABE}}}{k^{EHS} + \overline{C^{ABE}}}\right) - k^{FRR}\overline{C^{ABS}} \tag{S7}$$

$$\frac{dC^{APE}}{dt} = -\frac{A^{SM}k^{EMT}\overline{C^{AAB}}}{m^{CSM}}(C^{APE} - \overline{C^{ABE}}) - \frac{\mu_{max}C^{AUB}m^{EAC}}{k^{Y}}\left(\frac{C^{APS}}{k^{SHS} + C^{APS}}\right)\left(\frac{C^{APE}}{k^{EHS} + C^{APE}}\right) \tag{S8}$$

$$\frac{d\overline{C^{ABE}}}{dt} = \frac{A^{SM}k^{EMT}}{V^{SM}}(C^{APE} - \overline{C^{ABE}}) - \frac{\mu_{max}\rho_{X}m^{EAC}}{k^{Y}}\left(\frac{\overline{C^{ABS}}}{k^{SHS} + \overline{C^{ABS}}}\right)\left(\frac{\overline{C^{ABE}}}{k^{EHS} + \overline{C^{ABE}}}\right) \tag{S9}$$

$$\frac{dC^{AUB}}{dt} = \mu_{max}C^{AUB}\left(\frac{C^{APS}}{k^{SHS} + C^{APS}}\right)\left(\frac{C^{APE}}{k^{EHS} + C^{APE}}\right) - k^{ED}C^{AUB} \tag{S10}$$

$$\frac{d\overline{C^{AAB}}}{dt} = \mu_{max}\overline{C^{AAB}}\left(\frac{\overline{C^{ABS}}}{k^{SHS} + \overline{C^{ABS}}}\right)\left(\frac{\overline{C^{ABE}}}{k^{EHS} + \overline{C^{ABE}}}\right) - k^{ED}\overline{C^{AAB}} \tag{S11}$$

where C^{APS} is the aqueous phase substrate concentration (substrate mass per unit volume of aqueous phase); $\overline{C^{ABS}}$ is the substrate concentration in attached biomass (mass of substrate per unit volume of biomass); C^{APE} is the aqueous phase electron acceptor concentration (mass of electron acceptor per unit volume of aqueous phase); $\overline{C^{ABE}}$ is the electron acceptor concentration in attached biomass (mass of electron acceptor per unit volume of biomass); C^{AUB} is the aqueous phase concentration of unattached biomass (mass of unattached cells per unit volume of aqueous phase); $\overline{C^{AAB}}$ is the attached biomass concentration (mass of attached cells per volume of aqueous phase); A^{SM} is the surface area of a single microcolony (L^2); k^{EMT} is the electron acceptor mass transfer coefficient (LT^{-1}); k^{SMT} is the substrate mass transfer coefficient (LT^{-1}); μ_{max} is the maximum specific growth rate (T^{-1}); m^{CSM} is the mass of cells in a single microcolony, $m^{\mathrm{CSM}} = \rho V^{\mathrm{SM}}$ (M); m^{EAC} is the mass of electron acceptor consumed per mass of substrate biodegraded; ρ is the biomass density (mass of cells per volume of biomass); V^{SM} is the volume of a single microcolony (L^3); k^{Y} is the yield coefficient (mass of cells per volume of biomass); k^{SHS} is the substrate half-saturation coefficient (ML^{-3}); k^{EHS} is the electron acceptor half-saturation coefficient (ML^{-3}); k^{FRR} is the first-order reaction rate coefficient (for abiotic decay reactions, T^{-1}); k^{ED} is the endogenous decay coefficient (T^{-1}); and t is the time (T).

Reduction of contaminants in the aqueous phase in Equation (S6) results from three mechanisms. The first term accounts for diffusion of contaminants from liquid phase across a stagnant liquid film into attached biomass. The second one indicates the reduction of contaminants by unattached microorganisms in the bulk liquid. The reduction rate is affected by concentrations of contaminant and electron acceptor through the Monod kinetic. Substrate competition, nutrient limitations, inhibition, and reducing power limitations can also be incorporated within the second tern as described in the following sections. The third term accounts for abiotic loss of contaminants through first-order reactions. One equation of the same form as Equation (S6) will be used for each substrate.

Equation (S7) describes the loss of substrate within attached biomass. It describes processes of substrate diffusion into attached biomass, biodegradation within the biomass, and abiotic decay. Substrate competition, nutrient limitations, inhibition, and reducing power limitations can also be incorporated into this term for biodegradation of the substrate. Equations (S8) and (S9) describe the loss of the electron acceptor, which are of the same form as Equations (S6) and (S7). Equations (S8) and (S9) simulate the growth and decay of unattached and attached biomass, respectively.

The attached biomass concentration ($\overline{C^{\mathrm{AAB}}}$) is dependent upon the biomass density, microcolony volume and microcolony mass (de Blanc, 1998):

$$\overline{C^{\mathrm{AAB}}} = \frac{N^{\mathrm{CPS}} D^{\mathrm{PM}} m^{\mathrm{CSM}}}{N^{\mathrm{CPM}} \phi} \tag{S12}$$

where N^{CPS} is the number of cells per mass of solid; D^{PM} is the bulk density of the porous medium; N^{CPM} is the number of cells per microcolony (a constant); and ϕ is the porosity.

Since the biomass density, number of cells, mass of one microcolony, and medium porosity are assumed to be constant, $\overline{C^{\mathrm{AAB}}}$ is proportional to N^{CPS} or, alternately, to ($N^{\mathrm{CPS}} / N^{\mathrm{CPM}}$), (the number of microcolonies). Moreover, the area available for transport of species from the aqueous phase to the biomass is directly proportional to $\overline{C^{\mathrm{AAB}}}$, because the surface area per microcolony is assumed constant.

<u>Multiplicative Monod kinetics</u>

For multiplicative Monod kinetics, it is assumed that other limiting nutrients are also limiting microbial growth besides substrates and electron acceptors. When other chemical species or nutrients such as nitrogen or phosphorous are limiting factors, the substrate utilization term can be modified correspondingly in order to account for these additional limitations (Rittmann et al., 1991):

$$\mu'_{max} = \mu_{max} \cdot \frac{C^{LN}}{k^{LNH} + C^{LN}} \quad \text{(S13)}$$

where C^{LN} is concentration of a limiting nutrient (ML^{-3}); and k^{LNH} is limiting nutrient half-saturation coefficient concentration (ML^{-3}).

Biomass growth

The basic biomass growth expression of equations (S10) and (S11) contains an additional term to limit the volume of the biomass. With this limitation, the general form of the biomass growth expression is (de Blanc, 1998):

$$\frac{dC^{\text{AUB}}}{dt} = \mu_{\max} C^{\text{AUB}} \left(\frac{C^{\text{APS}}}{k^{\text{SHS}} + C^{\text{APS}}}\right)\left(\frac{A^{\text{SM}}}{K_A + A^{\text{SM}}}\right)\left(1 - \frac{C^{\text{AUB}}}{0.9\rho}\right) - k^{\text{ED}} C^{\text{AUB}} \quad \text{(S14)}$$

The linear biomass growth expression limits the total volume of biomass to 90% of the aqueous phase volume. At low biomass concentrations, such limits have negligible effects on biomass growth and substrate utilization because the biomass occupies a small volume of the total pore space.

When the biomass concentration begins to occupy a significant fraction of the pore volume, as might be expected near in-situ bioremediation injection wells, the key modeling assumption that biofilms in the pore space are thin and can be fully penetrated will likely be violated. The reduction (or near cessation) of biomass growth becomes less important than biofilm mass transport effects that are not considered in the model. Thus, through using the linear growth limitation expression, the model can only crudely approximate biological growth in grid blocks occupied by a substantial volume of biomass. At low biomass concentrations, the term has an insignificant effect.

The total biomass in the aquifer consists of the attached biomass and the unattached biomass is:

$$B^{\text{T}} = B^{\text{AP}} + \overline{B^{\text{A}}} \quad \text{(S15)}$$

where B^{T} is the total biomass, B^{AP} is the aqueous phase biomass, and $\overline{B^{\text{A}}}$ is the attached biomass. The attached biomass is composed of the minimum biomass population ($\overline{B^{\text{A}}}_{\text{min}}$, which does not partition between the solid and the aqueous phase) and the biomass in equilibrium with the aqueous phase biomass:

$$\overline{B^{\text{A}}} = \overline{B^{\text{A}}}_{\text{min}} + \kappa B^{\text{AP}} \quad \text{(S16)}$$

Substituting the equilibrium relationship of equation (S15) into mass balance (S16) results in the following equilibrium concentration of aqueous phase biomass:

$$X = \frac{X_T - \overline{X}_{\min}}{\kappa + 1} \quad \text{(S17)}$$

The attached biomass concentration is then calculated from equation (S16). The κ of infinity would mean that all of the biomass is attached, while the κ of 0 would mean that all of the biomass, except $\overline{B^{\mathrm{A}}}_{\min}$, would exist in the aqueous phase.

Substrate competition

When two substrates (substrates 1 and 2) compete for the same enzyme, it reduces the rate of biodegradation. The half-saturation coefficient of each substrate in Monod term is suggested, and thus, the Monod terms for the two substrates would become (Chang & Alvarez-Cohen, 1995):

Substrate 1:

$$\frac{C_{\mathrm{S1}}}{k_{\mathrm{S1}}^{\mathrm{HS}}(1+\frac{C_{\mathrm{S2}}}{k_{\mathrm{S2}}^{\mathrm{HS}}})+C_{\mathrm{S1}}} \tag{S18}$$

Substrate 2:

$$\frac{C_{\mathrm{S2}}}{k_{\mathrm{S2}}^{\mathrm{HS}}(1+\frac{C_{\mathrm{S1}}}{k_{\mathrm{S1}}^{\mathrm{HS}}})+C_{\mathrm{S2}}} \tag{S19}$$

where C_{S1}, C_{S2} are concentrations of substrates 1 and 2, respectively ($\mathrm{ML^{-3}}$); $k_{\mathrm{S1}}{}^{\mathrm{HS}}$, $k_{\mathrm{S2}}{}^{\mathrm{HS}}$ are half-saturation coefficients of substrates 1 and 2, respectively ($\mathrm{ML^{-3}}$).

Inhibition

Inhibition effects can be addressed through multiplying the substrate biodegradation rate term by an inhibition factor (Chang &Alvarez-Cohen, 2010):

$$\left(\frac{I_{ih}}{I_{ih}+C_{ih}}\right) \tag{S20}$$

where I_{ih} is an experimentally determined inhibition constant for species ih. Inhibition can be used to simulate the sequential use of electron acceptors or the reduction of biodegradation rates due to the presence of a toxic or inhibitory compound. The term for substrate utilization and biomass growth can be calibrated by using one inhibition factor for each inhibiting substance.

Aerobic cometabolism

To describe the loss of cometabolite and attached biomass growth in aerobic cometabolic reactions, the following equations can be used (in the case of no mass transfer resistance, no inhibition, and no substrate competition) (de Blanc, 1998):

$$\frac{dC^{\mathrm{APC}}}{dt}=-R^{\mathrm{SCB}}C^{\mathrm{AUB}}(\frac{C^{\mathrm{APC}}}{k^{\mathrm{CHS}}+C^{\mathrm{APC}}})(\frac{C^{\mathrm{APE}}}{k^{\mathrm{EHS}}+C^{\mathrm{APE}}})(\frac{C^{\mathrm{RP}}}{K^{\mathrm{RHS}}+C^{\mathrm{RP}}}) \tag{S21}$$

$$\frac{dC^{\mathrm{AUB}}}{dt} = \frac{\mu_{\max} C^{\mathrm{AUB}}}{k^{\mathrm{Y}}} \left(\frac{C^{\mathrm{APS}}}{k^{\mathrm{SHS}} + C^{\mathrm{APS}}}\right)\left(\frac{C^{\mathrm{APE}}}{k^{\mathrm{EHS}} + C^{\mathrm{APE}}}\right)\left(\frac{C^{\mathrm{RP}}}{K^{\mathrm{RHS}} + C^{\mathrm{RP}}}\right)\left[\frac{0.9(1 - C^{\mathrm{AUB}})}{\rho}\right]$$
$$- \frac{R^{\mathrm{SCB}} C^{\mathrm{AUB}}}{k^{\mathrm{TC}}} \left(\frac{C^{\mathrm{APC}}}{k^{\mathrm{CHS}} + C^{\mathrm{APC}}}\right)\left(\frac{C^{\mathrm{APE}}}{k^{\mathrm{EHS}} + C^{\mathrm{APE}}}\right)\left(\frac{C^{\mathrm{RP}}}{K^{\mathrm{RHS}} + C^{\mathrm{RP}}}\right) - k^{\mathrm{ED}} C^{\mathrm{AUB}} \quad \text{(S22)}$$

where R^{SCB} is maximum specific cometabolite biodegradation rate ($ML^{-3}T^{-1}$); C^{APC} is aqueous phase cometabolite concentration (ML^{-3}); C^{RP} is reducing power (NAD(P)H) concentration within the cells (mMOL e-/mass biomass); K^{RHS} is NAD(P)H half-saturation constant (mMOL e^-/mass biomass); k^{CHS} is cometabolite half-saturation coefficient (ML^{-3}); $\mu_{\max}$ is maximum specific growth rate on growth substrate (T^{-1}); and k^{TC} is transformation capacity (mass cells deactivated/mass cometabolite biodegraded).

Text S5. Procedures for solving the coupled flow and transport problem

The solution procedures are as follows:

Step 1. Solve the pressure equation implicitly using a Jacobi conjugate gradient solver to yield water phase pressure in all grid blocks;

Step 2. Capillary pressures from previous time step are used to determine the pressure of other phases in each grid block once the water phase pressure is known;

Step 3. The Darcy's law is used to determine the phase velocities;

Step 4. Mass conservation equations are solved explicitly to yield concentration of each component in each grid block;

Step 5. Phase concentrations and saturations are determined through flash calculations;

Step 6. The concentration of the components calculated by the pollutant migration model was used as the initial condition of the biodegradation model to obtain the pollutant degradation rate for this time step.

Step 7. New capillary pressures are determined from the new saturations;

Step 8. Repeat the procedures for each time step until simulation ends.

Text S6. Stepwise cluster analysis (SCA)

In the stepwise-cluster analysis, the solutions of the numerical model (benzene concentrations at concerned locations) are considered as dependent variables; the operating conditions are independent variables. If the developed simulation model is run under n scenarios of system conditions, there will then be n sets of such independent and dependent variables (e.g., if the model is run 50 times under various system conditions, then n = 50). Assume that there are m independent variables [e.g., four process control variables, denoted as $x = (x_1, x_2, \ldots, x_m)$, where m = 4], and p dependent variables [e.g., benzene concentrations at six concerned locations, denoted as $y = (y_1, y_2, \ldots, y_p)$, where p = 6]. Thus, all data can be given by matrixes $X = (x_{tr})_{n\times m}$ and $Y = (y_{tr})_{n\times p}$, where $r = 1, 2, \ldots, m$, and $i = 1, 2, \ldots, p$.

The first step is to determine the clustering principles for the patterns. In SCA, patterns of responses will be cut or merged into a number of sets, and explanatory variables will be the references in judging which pattern set in the parent set should enter. After completion of cutting and merging processes, cluster trees could be produced and further used for predicting responses according to new explanatory values. The essence of this method is, based on a given criteria, to cut one pattern set of responses into two, and to merge two sets into one, step by step, in order to classify sets and sieve variables. Let cluster h, which contains n_h patterns, be cut into two sub-clusters e and f, containing n_e and n_f patterns, respectively (i.e., $n_e + n_f = n_h$). According to Wilks' likelihood-ratio criterion, if the cutting point is optimal, the value of Wilks' Λ (Λ =|W|/|T|) should be minimum (Wilks, 1960; 1962; 1963; Kennedy and Gentle, 1981), where T and W are total-sample sum of the squares and cross products (SSCP) matrix $\{t_{ij}\}$ and within-groups SSCP matrix $\{w_{ij}\}$, respectively, and T and W mean determinants of matrixes $\{t_{ij}\}$ and $\{w_{ij}\}$, respectively. When the Λ value is very large, clusters e and f cannot be cut, but must be merged into greater cluster h. By Rao's F-approximation (R-Statistic), we have:

$$R = \frac{1-\Lambda^{1/S}}{\Lambda^{1/S}} \cdot \frac{Z \cdot S - P \cdot (K-1)/2 + 1}{P \cdot (K-1)} \tag{S23}$$

$$Z = n_h - 1 - (P + K)/2 \tag{S24}$$

$$S = \frac{P^2 \cdot (K-1)^2 - 4}{P^2 + (K-1)^2 - 5} \tag{S25}$$

where statistic R is distributed approximately as an F-value with $v_1 = P \cdot (K-1)$ and $v_2 = P \cdot (K-1)/2 + 1$ degrees of freedom, K is number of groups, and P is number of responses. The R - statistics will reduce to an exact F-value when P = 1 or 2, or when K = 2 or 3. Since the number of groups is two (K = 2 for system operating conditions and benzene concentrations at concerned locations) in this study, an exact F-test is possible based on Wilks' Λ criterion. Thus, we have:

$$F(P, n_h - P - 1) = \frac{1-\Lambda}{\Lambda} \cdot \frac{n_h - P - 1}{P} \tag{S26}$$

Therefore, the criteria of cutting and merging clusters become to make a number of F-tests (Rao, 1952).

The second step is to test optimal cutting points, for which n_h patterns in cluster h are sequenced according to the value of $x_{r,k}^{(h)}$ in $\{x_r\}$, i.e., $x_{r,1^r}^{(h)} \le x_{r,2^r}^{(h)} \le \cdots \le x_{r,n_h^r}^{(h)}$. Then the total-pattern SSCP matrix and within-groups SSCP matrix of responses y are calculated based on the sequence statistic $\{K^r\}$:

$$b_{ij}(K^r, n_h^r) = \frac{n_h^r K^r \cdot \left\{\left[B_i^{(h)}(K^r) - B_i^{(h)}(n_h^r)\right] \cdot \left[B_j^{(h)}(K^r) - B_j^{(h)}(n_h^r)\right]\right\}}{n_h^r - K^r} \quad \text{(S27)}$$

$$t_{ij}(n_h^r) = A_{ij}^{(h)}(n_h^r) - n_h^r B_i^h(n_h^r) B_j^h(n_h^r) \quad \text{(S28)}$$

$$w_{ij}(K^r, n_h^r) = t_{ij}(n_h^r) - b_{ij}(K^r, n_h^r) \quad \text{(S29)}$$

where:

$$B_{i\ or\ j}^{(h)}(u) = \frac{1}{u} \sum_{k=1}^{u} y_{i\ or\ j,k}^{(h)} \quad \text{(S30)}$$

$$A_{ij}^{(h)}(u) = \sum_{k=1}^{u} y_{i,k}^{(h)} y_{j,k}^{(h)} \quad \text{(S31)}$$

$$k^r = 1^r, 2^r, \cdots, (n_h^r - 1), \forall r$$

$$i, j = 1,2, \cdots, p, \text{ and } r = 1,2, \cdots, m$$

For each x_r, a cutting point k^{*r} is derived, which satisfies:

$$\Lambda(k^{*r}, n_h^r) = \min_{k^r=1^r}^{(n_h^r-1)} \{\Lambda(k^r, n_h^r)\} \quad \text{(S32)}$$

For each explanatory variable, the index of response that will be used for cutting judgments (denoted as r^*) is derived, which satisfies:

$$\Lambda\left(k^{*r^*}, n_h^r\right) = \min_{r=1}^{m} \{\Lambda(k^r, n_h^r)\} \quad \text{(S33)}$$

Thus, the optimal cutting point of cluster h is k^{*r^*}, and the relevant value of explanatory variable (i.e., the reference for new pattern prediction) is $x_{r^*,k^{*r^*}}^{(h)}$.Then a F-test can be undertaken.

If

$$F(P', n_h^{r*} - P' - 1) = \frac{1 - \Lambda(k^{*r^*}, n_h^{r*})}{\Lambda(k^{*r^*}, n_h^{r*})} \frac{n_h^{r*} - P'}{P'} \ge F_1 \quad \text{(S34)}$$

is satisfied, cluster h can be cut into two sub-clusters according to the distribution of x_{r^*}: (a) data in explanatory sets with $k^{r^*} \le k^{*r^*}$ are allocated into sub-cluster e ($< f$); (b) data in explanatory sets with $k^{r^*} > k^{*r^*}$ are allocated into sub-cluster f, where P' is number of responses under consideration. Among explanatory variables, x_{r^*} is the most important one affecting the response. If equation (S12) is not satisfied, cluster h cannot be cut. Then the other clusters will be tested to decide whether to cut or not, i.e., to test $h = 1, 2, \ldots, H$ (H is total number of clusters at the current stage). When no cluster can be cut, the next step is to merge the clusters.

The third step is the mergence of clusters. To test the mergence of clusters e and f for existing clusters, the total-sample SSCP matrix and within-groups SSCP matrix should be calculated firstly:

$$t_{ij}(n_e,n_f) = A_{ij}^{(e)}(n_e) + A_{ij}^{(f)}(n_f) - \left[n_e B_i^{(e)}(n_e) + n_f B_i^{(f)}(n_f)\right] \cdot \left[n_e B_j^{(e)}(n_e) + n_f B_j^{(f)}(n_f)\right] / (n_e + n_f) \quad \text{(S35)}$$

$$b_{ij}(n_e,n_f) = \frac{n_e n_f\left[B_i^{(e)}(n_e) - B_i^{(f)}(n_f)\right] \cdot \left[B_j^{(e)}(n_e) - B_j^{(f)}(n_f)\right]}{(n_e+n_f)} \quad \text{(S36)}$$

$$w_{ij}(n_e,n_f) = t_{ij}(n_e,n_f) - b_{ij}(n_e,n_f) \quad \text{(S37)}$$

where A_{ij} and $B_{i\ or\ j}$ have the same formulation as equations (S30) and (S31); $i, j = 1, 2, \ldots, p$. Then a F-test can be undertaken. If

$$F(P', n_e + n_f - P' - 1) = \frac{1-\Lambda(n_e,n_f)}{\Lambda(n_e,n_f)} \frac{(n_e+n_f)-P'-1}{P'} < F_2 \quad \text{(S38)}$$

is satisfied, clusters e and f can be merged into a new cluster h. Otherwise, it should be similarly tested whether other clusters can be merged for $e = 1, 2, \ldots, (H\text{-}1)$ and $f = 2, 3, \ldots, H$.

The final step is the prediction of the response according to new explanatory variables. After all calculations and tests have been completed (i.e., all hypotheses of further cutting or mergence are rejected), a cluster tree can be derived for each response. Each cutting point, which leads to two branches, corresponds to the value ($x_{r^*,k^*r^*}^{(h)}$) of an explanatory variable. When a new pattern set of explanatory variables $\{x_r\}$ is examined, its x_{r^*} value can be compared with $x_{r^*,k^*r^*}^{(h)}$ at the cutting point, and classified into relevant branches. Step-by-step, the pattern will finally enter a tip cluster which cannot be either cut or merged further. The criterion to classify a new sample to relevant branches is that, (a) sample data with $x_{r^*} \leq x_{r^*,k^*R^*}^{(h)}$ are merged into cluster e ($<f$) and (b) sample data with $x_{r^*} > x_{r^*,k^*R^*}^{(h)}$ are merged into cluster f. Let e' be the tip cluster where the new sample enters. Then the predicted dependent variable $\{y_i\}$ is:

$$y_i = y_i^{(e')} \pm R_i^{(e')} \quad \text{(S39)}$$

where $y_i^{e'}$ is mean of dependent variable i in sub-cluster e', and $R_i^{e'}$ is radius of y_i in cluster e':

$$y_i^{(e')} = \left\{\max_{k=1}^{n_{e'}}(y_{i,k}^{(e')} + \min_{k=1}^{n_{e'}}(y_{i,k}^{(e')})\right\}/2, \forall i \quad \text{(S40)}$$

$$R_i^{(e')} = \left\{\max_{k=1}^{n_{e'}}(y_{i,k}^{(e')} - \min_{k=1}^{n_{e'}}(y_{i,k}^{(e')})\right\}/2, \forall i \quad \text{(S41)}$$

Text S7. Filtering Process Model

After the clustering process, a number of leaf clusters are produced. Each leaf cluster contains a group of modeling outputs with similar statistical attributes; these modeling outputs provide an output value range for the leaf cluster. The purpose of filtering is to calculate an optimal estimate for each leaf cluster; this estimate can be used as an optimal output value for the leaf cluster. The set of leaf clusters for all well patterns thus can be regarded as all possible results for the remediation design.

Among various filtering methods, the well-known Kalman filter has been recognized as a powerful tool in supporting estimations of past, present, and future states. In this study, a filtering process model based on the Kalman filter method was developed to calculate the optimal estimate for each leaf cluster.

Generally, the Kalman filter addresses the problem of estimating the state of a discrete-time controlled process, z ($z \in R^f$), that is governed by the following linear stochastic difference equation:

$$z_k = Az_{k-1} + Bu_{k-1} + w_{k-1} \tag{S42}$$

with a measurement ($q \in R^g$) as follows:

$$q_k = Hz_k + v_k \tag{S43}$$

where u_k is the optional control input ($u \in R^l$); w_k and v_k represent the process and measurement noise (random variables), respectively. They are assumed to be independent (of each other), white, and with normal probability distributions $p(w) \sim N(0, Q^{PNC})$ and $p(v) \sim N(0, R^{MNC})$, respectively. A white noise process is defined as a random process of random variables that are uncorrelated, have mean zero, and a finite variance. The process noise covariance Q^{PNC} and measurement noise covariance R^{MNC} matrices are assumed to be constant.

In equation (S42), the $f \times f$ matrix A relates the state at the previous time step (k−1) to the state at the current step k, in the absence of a process noise. The $f \times l$ matrix B relates the optional control input u to the state z. The $g \times f$ matrix H in equation (S21) relates the state to the measurement q. Matrices A and H are assumed to be constants.

It is defined that $\hat{z}_{\bar{k}}$ ($\hat{z}_{\bar{k}} \in R^f$) is a priori-state estimate at step k given knowledge of the process prior to step k, and $\hat{z}_k$ ($\hat{z}_k \in R^f$) to be a posteriori-state estimate at step k given measurement q_k. It is then defined a priori-estimate error and a posteriori-estimate error as $e_{\bar{k}} \equiv z_k - \hat{z}_{\bar{k}}$ and $e_k \equiv z_k - \hat{z}_k$, respectively. Thus, the priori-estimate error covariance can be written as $P_{\bar{k}} = E[e_{\bar{k}} {e_{\bar{k}}}^T]$ and the posteriori-estimate error covariance as $P_k = E[e_k e_k^T]$.

The posteriori-state estimate ($\hat{z}_k$) can be calculated as:

$$\hat{z}_k = \hat{z}_{\bar{k}} + K(q_k - H\hat{z}_{\bar{k}}) \tag{S44}$$

The difference between the actual measurement (q_k) and the measurement prediction, ($q_k - H\hat{z}_{\bar{k}}$), in equation (S44) is called the residual, which reflects the discrepancy between the predicted measurement and the actual measurement.

The $f \times g$ matrix K in equation (S44) is Kalman gain, which is chosen to minimize the posteriori error covariance. The Kalman gain K_k can be given as follows (Maybeck, 1979; Jacobs, 1993):

$$K_k = P_{\bar{k}} H^T (H P_{\bar{k}} H^T + R)^{-1} \tag{S45}$$

As the R^{MNC} approaches zero, the gain K weights the residual more heavily. Specifically, $\lim_{R_k \to 0} K_k = H^{-1}$. On the other hand, as the priori-estimate error covariance $P_{\bar{k}}$ approaches zero, the gain K weights the residual less heavily. Specifically, $\lim_{P_{\bar{k}} \to 0} K_k = 0$.

The Kalman filter consists of time-update equations and measurement-update equations. The discrete time-update equations are written as:

$$\hat{z}_{\bar{k}} = A\hat{z}_{k-1} + Bu_{k-1} \tag{S46}$$

$$P_{\bar{k}} = AP_{k-1}A^T + Q^{PNC} \tag{S47}$$

The time-update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain the priori-estimates for the next time step. The discrete measurement-update equations are given as:

$$K_k = P_{\bar{k}} H^T (H P_{\bar{k}} H^T + R^{MNC})^{-1} \tag{S48}$$

$$\hat{z}_k = \hat{z}_{\bar{k}} + K_k(q_k - H\hat{z}_{\bar{k}}) \tag{S49}$$

$$P_k = (1 - K_k H)P_{\bar{k}} \tag{S50}$$

The measurement-update equations are responsible for the feedback—i.e., for incorporating a new measurement into the priori-estimate to obtain an improved posteriori-estimate.

The operation of the filter is shown below. The first step is to compute the Kalman gain, K_k. The next step is to actually measure the process to obtain q_k and then to generate a posteriori-state estimate by incorporating the measurement as in equation (S49). The final step is to obtain a posteriori error covariance estimate via equation (S50). After each iteration of time update and measurement update, the process is repeated with the previous posteriori-estimates used to predict the new priori-estimates. This recursive nature is one of the very appealing features of the Kalman filter. For example, compared with the implementation of a Wiener filter, which operates on all the data directly for each estimate, the implementation of the Kalman filter is much more feasible.

In this study, the modeling outputs (samples) in each leaf cluster can be regarded as measurements. For any leaf cluster, the time update equations were written as:

$$\hat{z}_{\bar{k}} = \hat{z}_{k-1} \tag{S51}$$

$$P_{\bar{k}} = P_{k-1} + Q^{PNC} \tag{S52}$$

where A=1 (the state did not change from step to step), and u=0 (there was no control input). The measurement update equations were given as:

$$K_k = P_{\bar{k}}(P_{\bar{k}} + R^{MNC})^{-1} \tag{S53}$$

$$\hat{z}_k = \hat{z}_{\bar{k}} + K_k(q_k - \hat{z}_{\bar{k}}) \tag{S54}$$

$$P_k = (1 - K_k)P_{\bar{k}} \tag{S55}$$

where H=1 (the noisy measurement is of the state directly); k denotes the number of samples (modeling outputs) in each leaf cluster.

After the clustering and filtering, an optimal estimate can be obtained for each leaf cluster. A new sample can be grouped into a corresponding leaf cluster by comparing the values of x_{tr} with those of $x_{r^*}^{\alpha}(h^*)$. The corresponding output variable can be predicted as $y_i = \hat{z}_{k,i}$, $\forall i$.

Text S8. Nonlinear Optimization model of the FCI optimizer

The Nonlinear Optimization model of the FCI optimizer can be formulated as follows (to identify the optimum control conditions):

$$\text{Min } Z = \sum_{i=1}^{I} U_i^{In} + \sum_{j=1}^{J} U_j^{Ex} \tag{S56}$$

subject to:

$$X_{kt}\left(U_i^{In}, U_j^{Ex}\right) \leq X_{max} \text{ for all k=1,2, …, K} \tag{S57}$$

$$0 \leq U_i^{In} \leq U_{i,max}^{In} \tag{S58}$$

$$0 \leq U_j^{Ex} \leq U_{j,max}^{Ex} \tag{S59}$$

$$\sum_{i=1}^{I} U_i^{In} = \sum_{j=1}^{J} U_j^{Ex} \tag{S60}$$

where Z is the total pumping rate for all injection and extraction wells; U_i^{In} and U_j^{Ex} are pumping rates for the *ith* injection well and the *jth* extraction well after a period of remediation; $U_{i,max}^{In}$ and $U_{j,max}^{Ex}$ are maximum pumping rates for the *ith* injection well and the *jth* extraction well; X_{max} is environmental standard; I, J, K are numbers of injection well, extraction well, and monitoring well, respectively; X_{kt} is predicted benzene concentration at t. Constraint (S60) indicates that all the extracted water will be injected into the aquifer. This constraint is emphasized to ensure such a stable hydraulic gradient that the groundwater can flow directed toward the plume interior.

Text S9. Nonlinear Optimization model of the DPC system

The Nonlinear Optimization model of the DPC system can be formulated as follows (to identify the optimum control conditions):

$$\text{Min } Z = w_1(X)(S(X) - H)^2 + w_2(U)U \quad \text{(S61)}$$

subject to:

$$S(X) = (X - X^0) / X^0 \quad \text{(S62)}$$

$$X = F(U) \quad \text{(S63)}$$

$$U_L \le U \le U_U \quad \text{(S64)}$$

where *Z* is the optimization objective, representing the system cost; U_L and U_U are the lower and upper bounds of *U*, respectively; w_1 and w_2 are the weights to reflect different priorities for the remediation efficiency and cost. In this optimization model, *S(X)* is within the range of 0 to 1; therefore, the injection and extraction rates (*U*) are normalized to fit it. *H* is a constant greater than or equal to 1 which is the highest contaminant removal rate. In this optimization model, a pseudo-equation X=F(U) is used to describe the relationship between X and U.

Text S10. Nonlinear Optimization model of the SADPC system

The Nonlinear Optimization model of the SADPC system can be formulated as follows (to update the optimum control conditions):

$$\text{Min} \quad J = \left\{\sum_{i=1}^{p} \omega_i\,(X)\big(X_r(t+i) - X_p(t+i)\big)^2 + \sum_{i=1}^{p} \omega_i(U)U(t+i-1)\right\} \quad \text{(S65)}$$

subject to:

$$X = F(U) \quad \text{(S66)}$$

$$0 \leq X_r(t+i) \leq X_{max} \quad \text{(S67)}$$

$$0 \leq X_p(t+i) \leq X_{max} \quad \text{(S68)}$$

$$U_L \leq U \leq U_U \quad \text{(S69)}$$

where J is the optimization objective, representing the system cost; P is the prediction horizon; $w_i(X)$ and $w_i(u)$ are the weights to reflect different priorities for the remediation efficiency and cost. $X_r(t+i)$ and $X_P(t+i)$ are setpoint and predicted value, respectively; X_{max} is environmental standard. U is the operating condition; U_L and U_U are the lower and upper bounds of U, respectively. In this optimization model, a pseudo-equation $X = F(U)$ is used to describe the relationship between X and U.

Text S11. Genetic algorithms (GA)

GAs are heuristic search procedures based on the mechanisms of genetics and Darwin's natural selection principles, combining an artificial survival of the fittest with genetic operators abstracted from nature (Holland, 1975).

An initial random population of genomes within the search space is generated. Each genome represents a possible solution to the search/optimization problem and is represented by a string of values (genes), one per each search variable. Survival of the fittest is accomplished by evaluating each genome's fitness through an appropriate objective function and a biased random selection procedure of individuals for "reproduction", where higher rated genomes are more likely to be selected. Generation of a new population is achieved by means of crossover (partial exchange of information between pairs of strings) and mutation (a random change in a random location within the string). The fittest individuals are transferred unchanged to the next generation, an approach known as "elitism". Every new generation of genomes is expected to be more closely concentrated in the vicinity of the optimal solution. The process is repeated until a convergence criterion is met or a pre-set maximum number of generations reached. GA input parameters include: population size, number of generations, range limits of each gene, crossover and mutation rates and a fitness function for genome evaluation.

In this study, GA is used to solve the developed discrete and nonlinear model. A set of parameters are needed to be predefined for guiding the genetic algorithm (Kuo et al., 2006; Matott et al., 2006; Stramer et al., 2010; Opher and Ostfeld, 2011; Liao et al., 2020), including: (1) chromosome length LCHR which is the product of the number of decision variables (n) and the length of a string (k); (2) population size M which is usually within the range of 30 to 200; (3) crossover rate RCRO which is usually within the range of 0.6 to 0.95; (4) mutation rate RMUT which is usually within the range of 0.001 to 0.05; and (5) convergence criterion which is used to judge whether stop the search process. Normally the process is stopped after a predetermined generation number NG is reached or when there are no significant differences among the best solutions.

Text S12. MPC control module procedure

The running procedures are as follows:

Step 1. Set the prediction time domain P and the weighting coefficients ω_i;

Step 2. Use the expected output sequence $x_r(t)$ in the future, and the reference trajectory comes from the first-order exponential form fitting the actual output value of the DPC system;

Step 3. The control amount obtained in this sampling time period from the DPC system is brought into the biodegradation process to obtain the actual system output $x(t)$;

Step 4. Use the DPC system to obtain the model output $x_m(t)$ of the current sampling time period and the predicted output $x_m(t+i)$ of the future time period, and obtain the system predicted output value $x_p(t+i)$ after feedback correction;

Feedback correction:

$$x_p(t+i) = x_m(t+i) + he(t) \quad \text{(S70)}$$

$$e(t) = x(t) - x_m(t) \quad \text{(S71)}$$

h is the compensation coefficient;

Step 5. The optimization algorithm is used to solve the rolling optimization, and the optimal sequence $U(t+i-1)$ is obtained;

Step 6. Apply the first control variable $U(t)$ of the optimal sequence to the system, and then return to step 2.

Text S13. General procedure for developing a process control system for enhanced in situ biodegradation

Step 1. A 3D pilot-scale model is designed for supporting the operation of enhanced in-situ biodegradation.

Step 2. After the occurrence of a hydrocarbon spill, an enhanced in-situ biodegradation process is to be undertaken. A subsurface LNAPLs biodegradation model is then developed to reflect the in-situ LNAPL biodegradation process.

Step 3. After calibration and verification, the interactions between contaminant concentrations and operating conditions are simulated through the subsurface model.

Step 4. Considering high complexities and computational requirements in incorporating numerical simulation model directly into optimization frameworks, coupled with the inability to obtain enough samples due to the high cost of sampling, a statistical relationship between remediation system performance and operating condition will be developed based on a large number of runs for the developed simulation model under various system conditions. Different scenarios of contamination situations and operating conditions are considered for the simulation. Under each contamination situation, the effects of various operation conditions on contaminant concentrations at concerned locations are examined.

Step 5. The stepwise cluster analysis method or the filtered clustering analysis method is used to develop to reflect the effects of variations of operating-condition on contaminant concentrations. Thus, a bridge between the subsurface model and the operating decision is established for further determining the desired operating conditions.

Step 6. Based on the established statistical relationships, a corresponding nonlinear discrete optimization model for groundwater control is established to determine optimal operating conditions corresponding to specific contamination situations. The GA technique is used to solve the developed optimization model.

Step 7. After the optimal operation conditions for each scenario are determined, the SI emulator is developed through the obtained knowledge base.

Step 8. A new nonlinear discrete optimization model is formulated by using the part that meets the expectation and its epitaxial as the setpoint curve for the contamination situations that do not meet the expectation in each scenario. Rolling optimization determines the optimal operating conditions corresponding to a specific contamination situation. The GA technique is used to solve the newly developed optimization model, and the optimal operating conditions of each scene are updated.

Text S14. Collinearity test of the independent variables

The presence of high collinearity in an FCI simulator implies that the conclusions of the analysis can be questioned. For example, the accuracy of estimations cannot be guaranteed due to high variances of the estimators. Thus, detection of collinearity should be a compulsory first step in every correlation analysis. Collinearity measures have been widely applied to examine if there are any co-relations among the independent variables. Variance Inflator Factor (VIF)is commonly used to evaluate the level of collinearity, which can be calculated as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{S72}$$

$$\mathbf{X} = \left(\mathrm{u1}, \mathrm{u2}, \mathrm{u3}, \mathrm{u4}\right) \tag{S73}$$

$$\mathrm{VIF}_i = \tfrac{1}{1\text{-}R_i^2}, i = 1, 2, 3, 4 \tag{S74}$$

where $\boldsymbol{\varepsilon}$ represents the random disturbance and R_i is the negative correlation coefficient of the independent variable for the regression analysis of the remaining independent variables.

If the data matrix has no full column rank that can be considered "severe multicollinearity", e.g., an independent variable can be expressed linearly by other independent variables. The closer the VIF value near 1, the lower the collinearity level is. The threshold value is usually 10. In this study, we selected groundwater injection rates of oxygen and nutrient in Well I and Well II (u1 and u2), and groundwater extraction rates in Well III and Well IV (u3 and u4) as the independent variables (called control variables in this paper) (Table S6). Results show that the corresponding VIF values of all the independent variables are much less than the threshold value (10), indicating that the variables are independent and do not have the multicollinearity (Table S8).

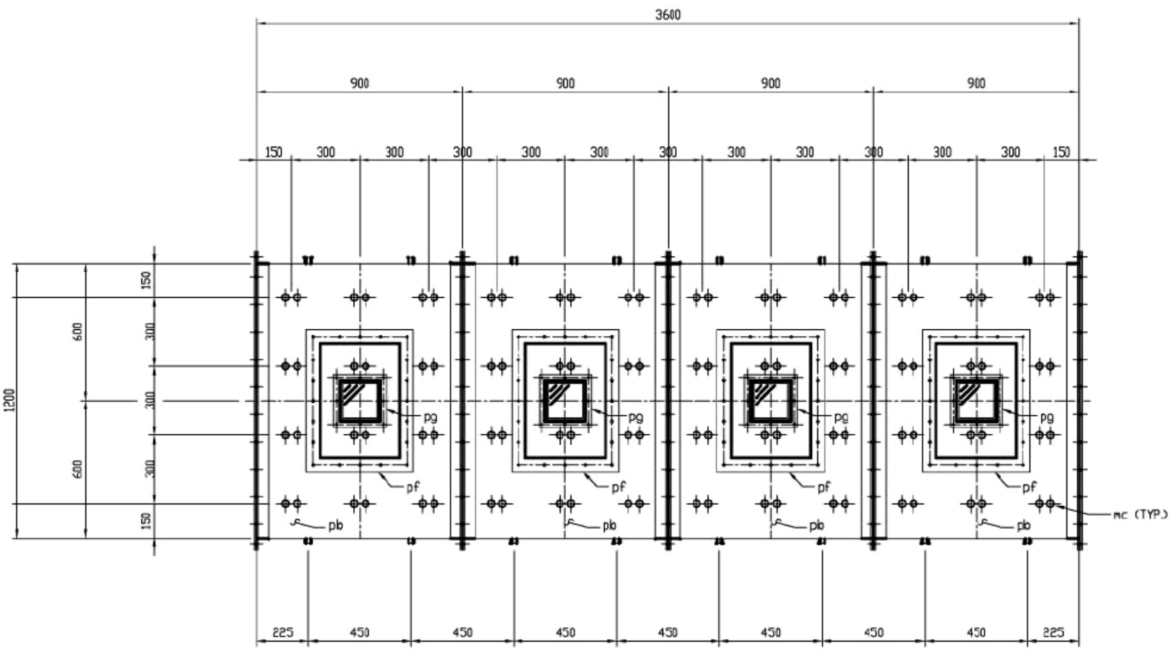


Figure S1(a). Plan view of the pilot scale system

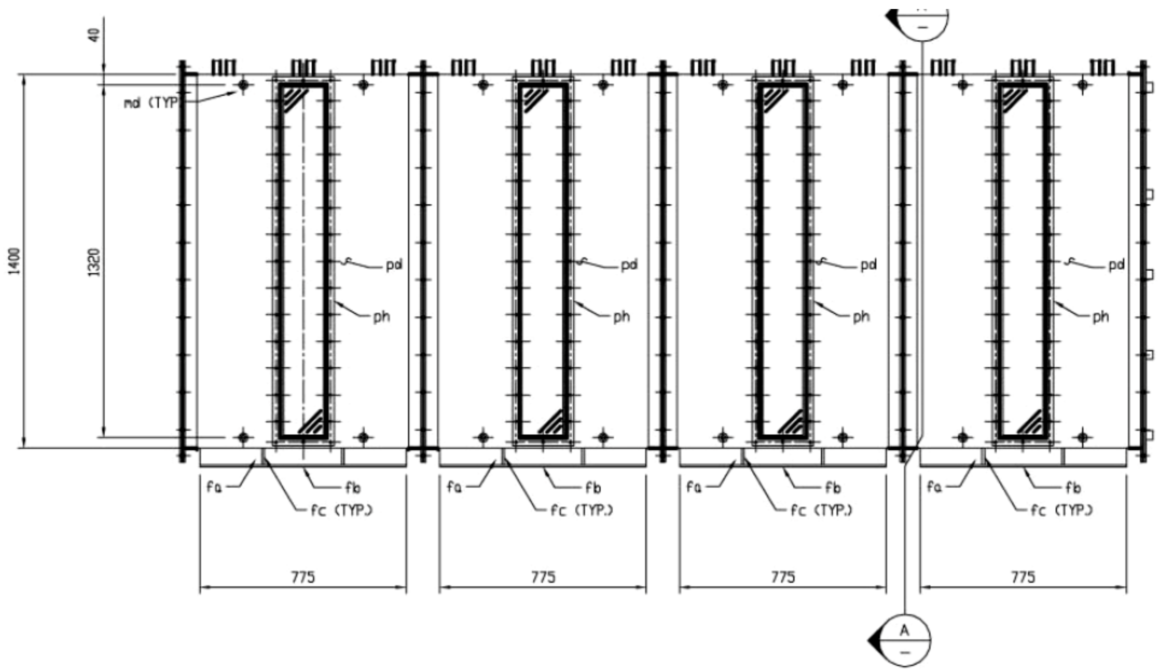


Figure S1(b). Front view of the pilot scale system

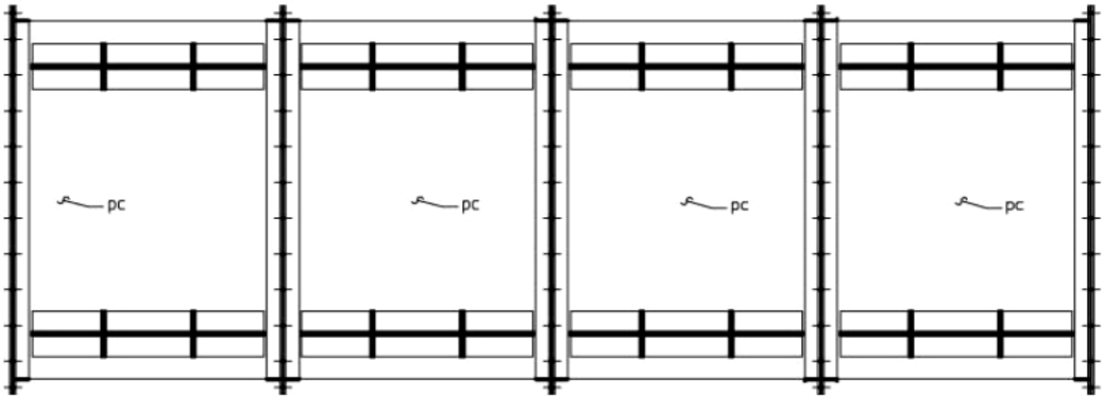


Figure S1(c). Bottom view of the pilot scale system

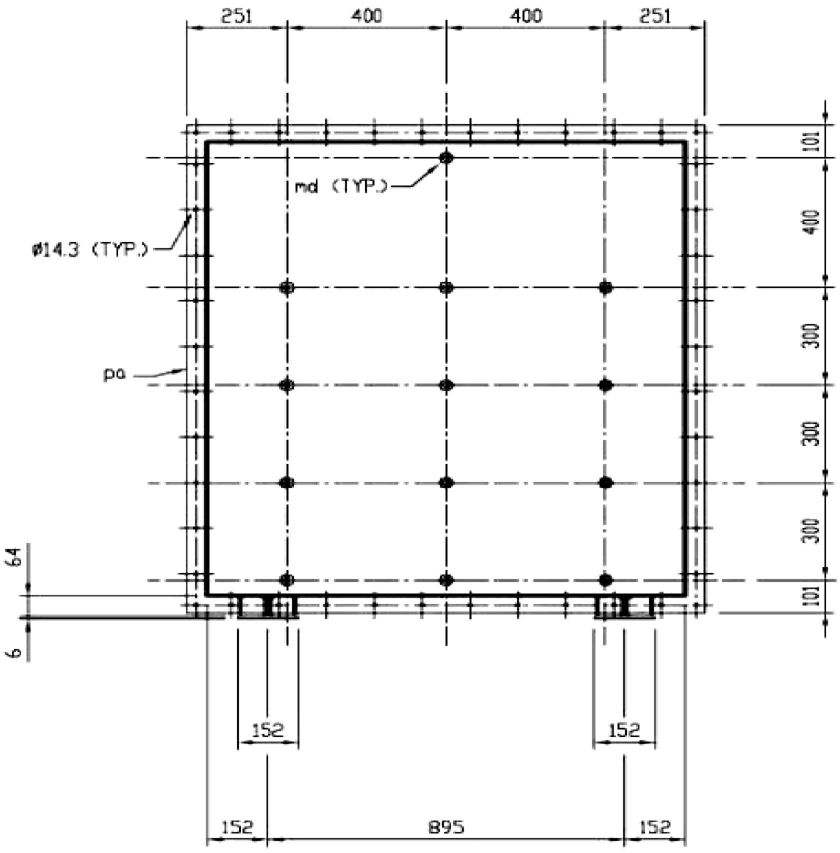


Figure S1(d). End elevation of the pilot scale system

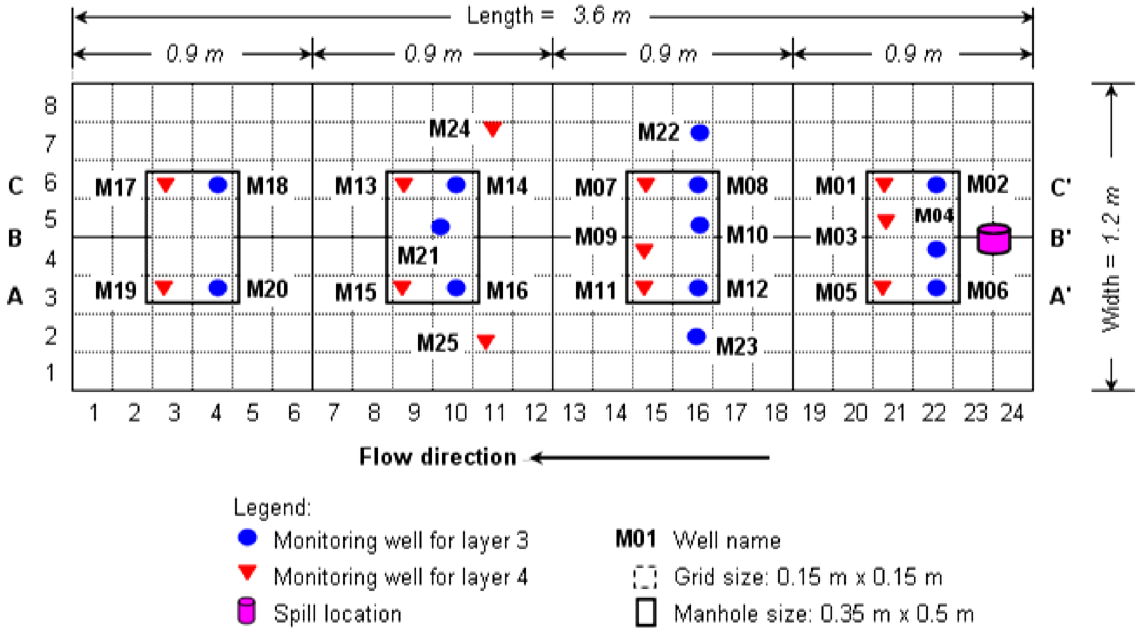


Figure S1(e). Well locations (plan view)

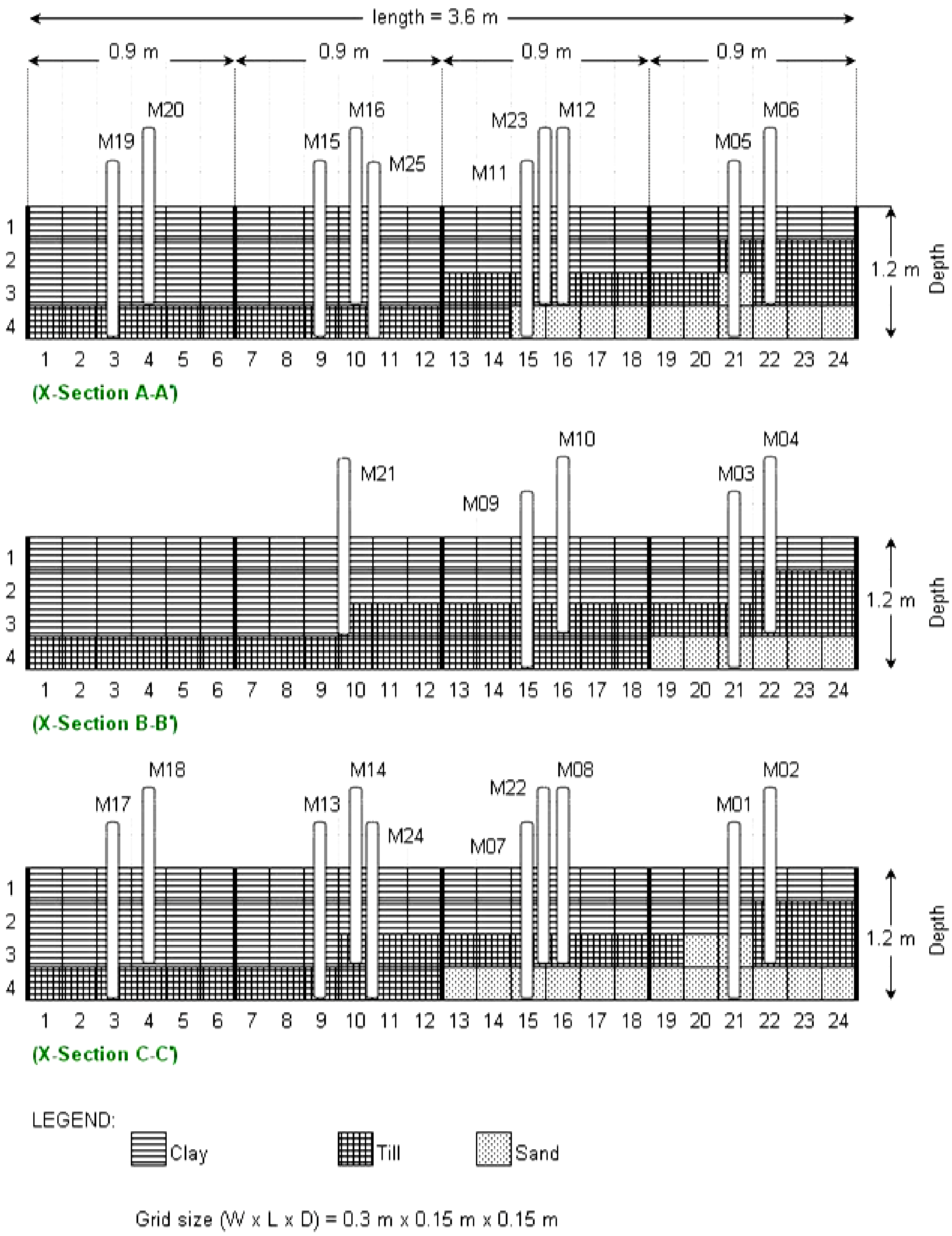


Figure S1(f). Well locations and soil types (section view)

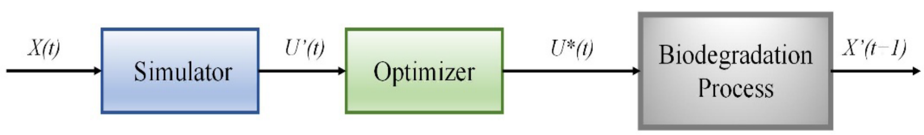


Figure S2. Framework of the FCI Optimizer

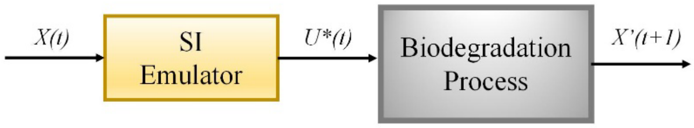


Figure S3. Framework of the SI Emulator

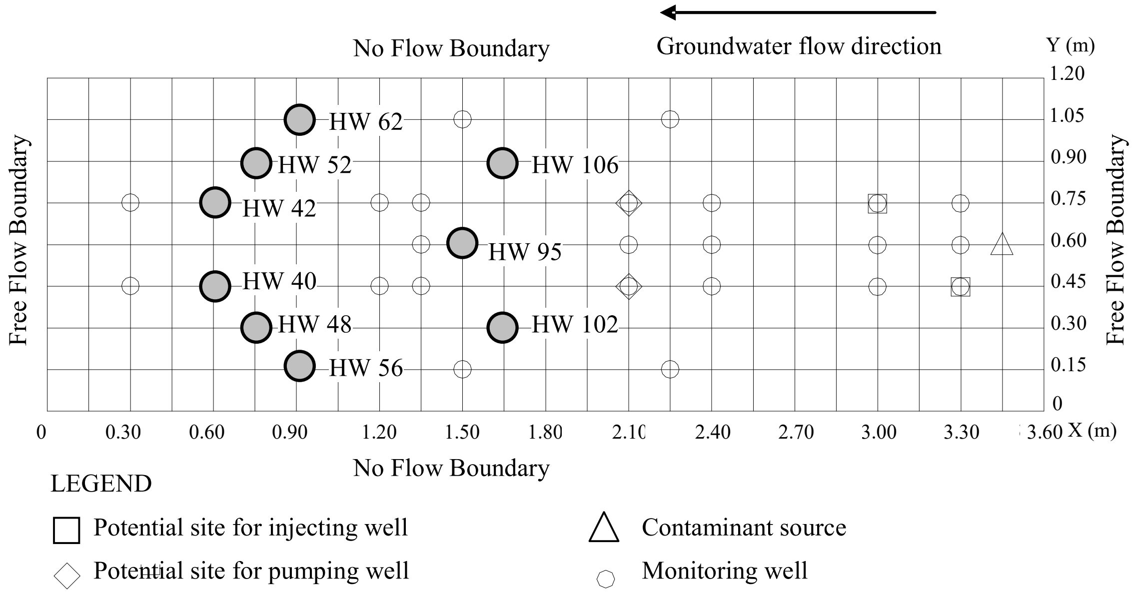


Figure S4. Locations of the hypothetical wells

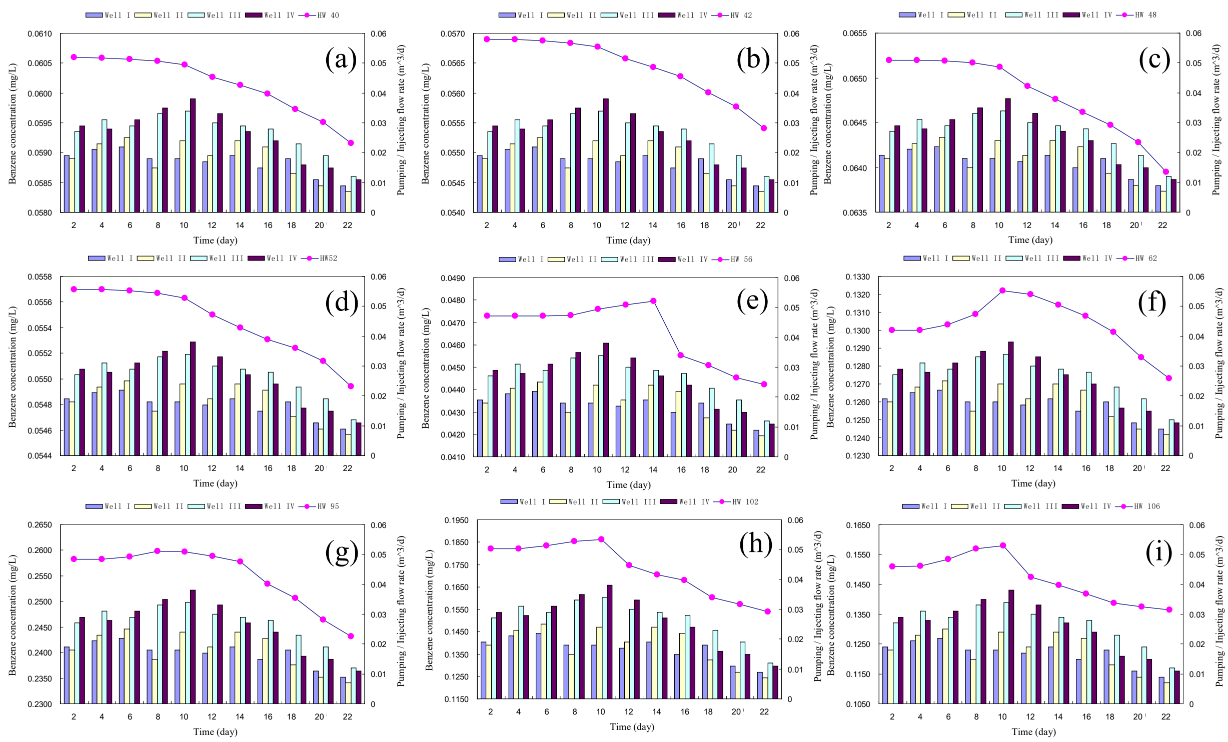


Figure S5. Benzene concentrations of the DPC system from Day 2 to Day 22, where Figs. (a) to (i) represents the concentrations at HW-40, HW-42, HW-48, HW-52, HW-56, HW-62, HW-95, HW-102, and HW-106

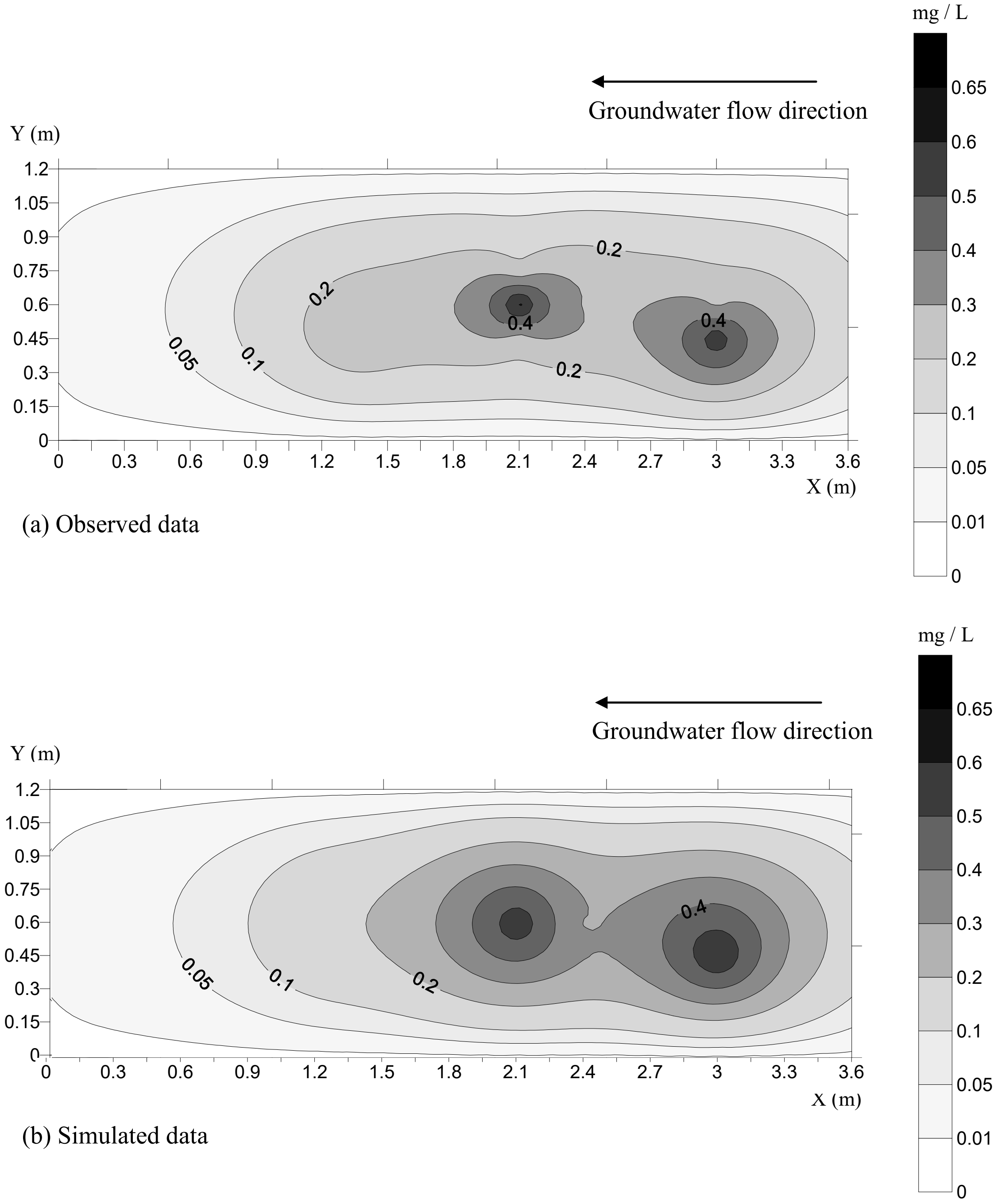


Figure S6. The concentration distribution of Benzene on Day 57 of the experiment

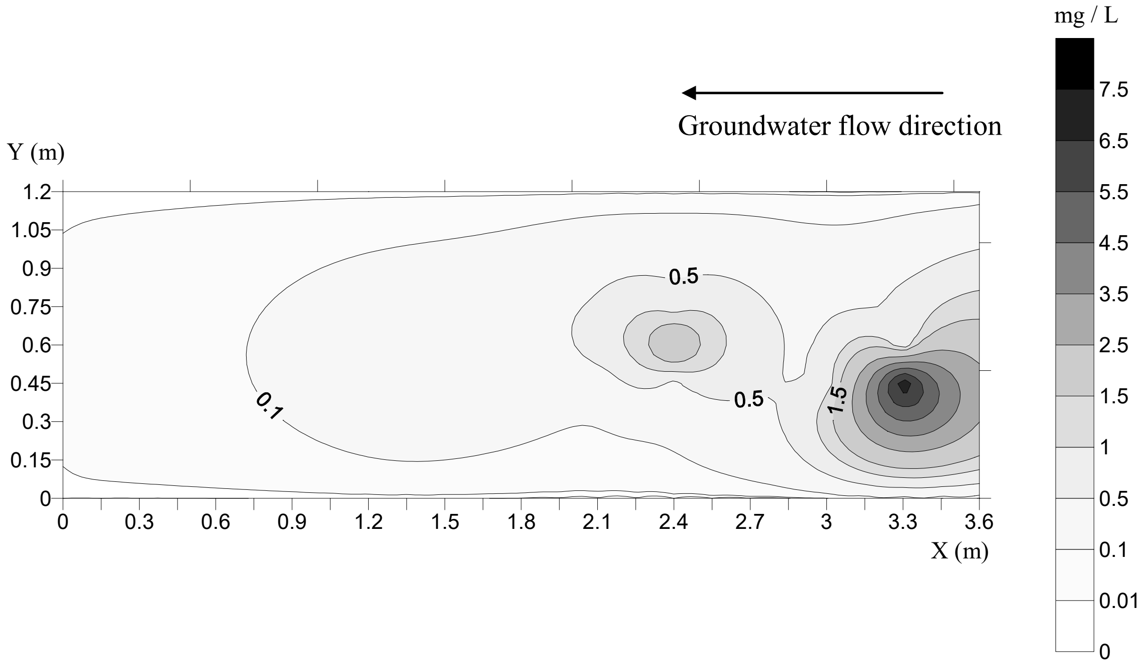


Figure S7(a). The concentration distribution of Benzene on Day 40

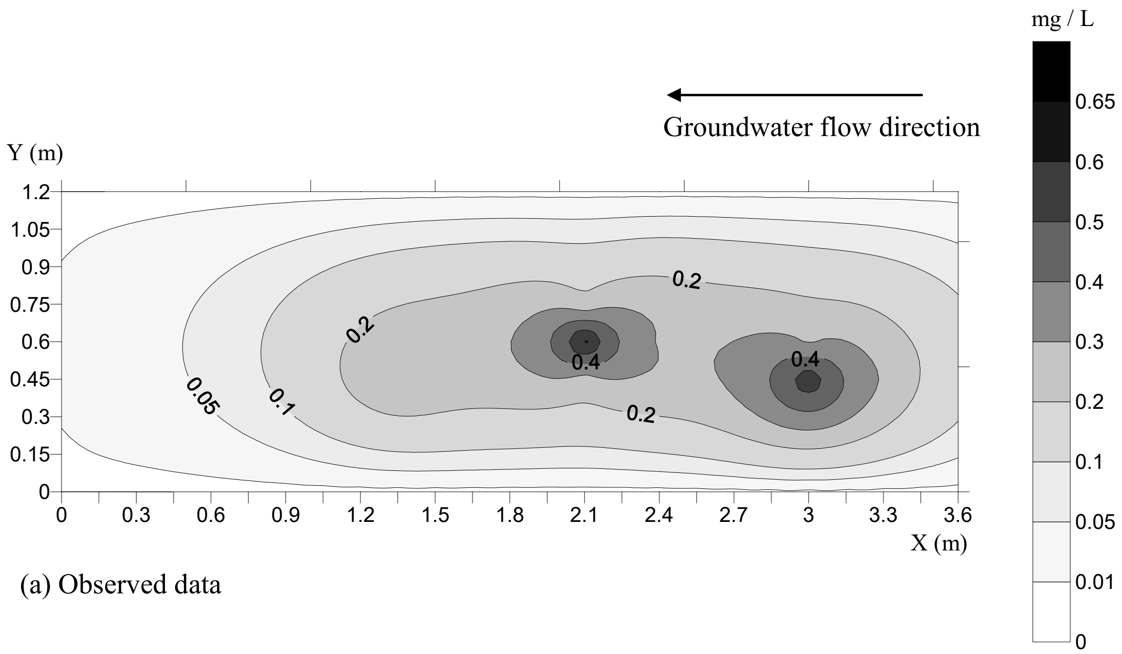


Figure S7(b). The concentration distribution of Benzene on Day 57

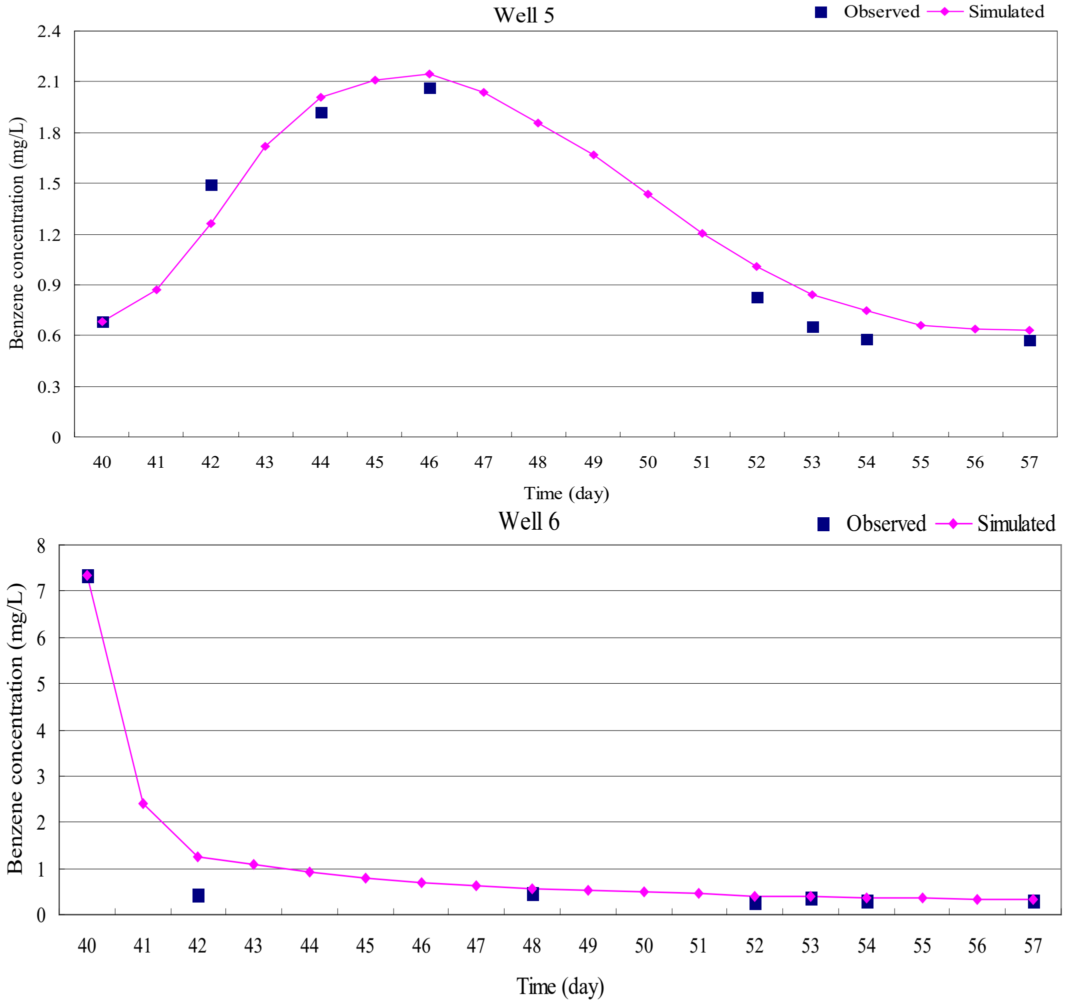


Figure S8. Verification results for well 5 and well 6

Table S1. Input parameters for contaminant transport simulation

Parameter	Value
Flow and transport simulation parameters	
Hydraulic conductivity of sand/ till/ clay	10 / 5 / 2.5 m/d
Permeability of sand/ till/ clay	1500/430/ 890 MD
Porosity of sand/ till/ clay	0.35 / 0.30 / 0.45
Longitudinal dispersivity of sand/ till/ clay	0.1 / 0.1 / 0.1 m
Transverse dispersivity of sand/ till/ clay	0.01 / 0.01 / 0.01 m
Van Genuchten's alpha of sand/ till/ clay	10 m^{-1}
Van Genuchten's n of sand/ till/ clay	6.8
First-order reaction rate coefficient of benzene	0.21 /d
Endogenous decay coefficient	0.2544 /d
Residual water saturation	0.01
Water dynamic viscosity	1.0 cp
Water interfacial tension	45 Dynes/cm
Benzene density	0.713 g/cm^3
Hydraulic gradient	0.03 m/m
Water partition coefficient of benzene	0.00203
Benzene solubility	1750 mg/L
Aquifer thickness	1.2 m
Time step	0.101 day
Maximum time step size	10 day
Tolerance for concentration change	0.001
Enhanced biodegradation simulation parameters	
Water injecting rate	20 L/d
NH_4NO_3 nutrient injecting rate	1750 mg/L
NH_4HPO_4 nutrient injecting rate	1100 mg/L
Heterotrophs microorganism injecting rate	20 mg/L
Oxygen injecting rate	8 mg/L
Water pumping rate	30 L/d
Microorganisms maximum specific growth rate	4.2 per day
Biomass density	0.09 g/cm^3
Yield coefficient (g cell/g benzene)	1.0 cells/g soil
Half-saturation coefficient	0.77 mg/L
Bulk density of porous medium	1.64 g/cm^3
Simulation period	12 day

Table S2. Initial Geochemical and Microbial Properties of the Soil

Parameter	Value
Soil classification	Silty clay, sand, and clay matrix till
Hydraulic conductivity	In the range of 10^{-7} to 10^{-5} (m/s)
Moisture content	7.5-32.5% (by volume)
Porosity	30-53.1%
Na	436-548 mg/L
K	16-19.7 mg/L
Ca	562-629 mg/L
Mg	338-407 mg/L
Fe	0.12-1.04 mg/L
Cl	10-79 mg/L
N, NO_2^-, NO_3^-	31-115 mg/L
Soil organic carbon	1.14%
Dissolved oxygen concentration	<1.0 mg/L to 1.5 mg/L
Initial microbial species	*Pseudomonas* sp. Strain CFS-215, *Geobacter* sp., and *Rhodocuccus* sp. Strain 33

Table S3. Observed benzene concentrations (mg/L)

Well	Day 13	Day 15	Day 17	Day 19	Day 21	Day 24	Day 26	Day 28	Day 32	Day 34	Day 36	Day 38	Day 40
1		0.032			1.073	0.033		0.174	0.696	0.837	0.738	0.462	0.439
2	0.484	0.564	0.262	0.360	0.672	0.978	0.732	0.699	1.054	1.252	0.682	0.542	0.702
3	0.643	0.734	3.169	2.391	3.408	1.777	2.137	1.858	1.834	1.897	1.077	0.712	0.606
4	0.236				0.245			0.090	0.663	0.827	0.671	0.409	0.415
5	1.347	2.074	1.362	0.888	0.842	0.204	0.601	0.745	0.825	0.974	0.993	0.578	0.685
6	8.131	7.482	7.795	5.530	7.438	7.068	8.696	5.716	4.080	4.887	6.337	3.519	7.340
7	0.279							0.392	0.875	1.370	0.756	0.761	0.566
8	0.296			0.357				0.175	0.851	1.117	0.738	0.594	0.720
9					1.198	0.186		0.300	0.884	1.067	0.724	0.637	0.741
10	1.359	1.218	0.498		1.284	4.698	1.029	1.278	1.384	1.890	1.663	0.546	2.488
11			0.507					0.851	0.508	0.213	0.474	0	0.203
12		0.502	0.808					0.843	0.578	0.288	0.502	0.036	0.352
13													
14													
15													
16									0.485	0.211			0.324

Table S3. (continued)

Well	Day 42	Day 44	Day 46	Day 48	Day 52	Day 53	Day 54	Day 57
1	0.304							
2								
3	0.526	1.593	1.429	0.508	0.501	0.733	0.386	0.285
4			0.090	0.400	0.285	0.293	0.292	0.245
5	1.497	1.920	2.070		0.824	0.651	0.581	0.575
6	0.444			0.472	0.265	0.361	0.284	0.291
7	0.385	0.733	1.227	0.686	0.300	0.268	0.241	0.224

8	0.524	0.703	0.366	0.527	0.359	0.316	0.310	0.273
9	1.070	0.918	0.698	1.357	0.831	0.874	0.397	0.633
10	1.366	1.628	0.947	1.590	2.055	1.292	1.162	0.292
11	0.349	0.710	0.163	0.554	0.363	0.280	0.288	0.249
12	0.417	0.741	0.166	0.443	0.322	0.267	0.263	0.285
13			0.079	0.376				
14				0.376	0.231			
15	0.512		0.090	0.398	0.261	0.248	0.237	0.235
16	0.334	0.373	0.453	0.776	0.563	0.507	0.296	0.296

Table S4. Error analysis for the biodegradation simulation results

Well number	Observed concentration (mg/L)	Simulated concentration (mg/L)	Absolute Error (mg/L)
3	0.00	0.03	0.03
4	0.00	0.05	0.05
5	0.51	0.25	0.26
6	0.40	0.35	0.05
7	0.80	0.80	0.00
8	0.47	0.16	0.31
9	0.69	0.49	0.20
10	0.53	0.78	0.25
11	1.36	1.70	0.34
12	2.00	2.40	0.40
15	0.41	0.80	0.19
16	0.44	0.20	0.24
Mean absolute error		0.21	
Root mean square error		0.27	
Correlation coefficient		0.93	

Table S5. Fifty levels of contamination situation (mg/L)

No.	M 5 x_1	M 7 x_2	M 8 x_3	M 10 x_4	M 11 x_5	M 12 x_6	No.	M 5 x_1	M 7 x_2	M 8 x_3	M 10 x_4	M 11 x_5	M 12 x_6
1	1.68	22.07	5.63	14.10	2.88	1.62	26	6.53	25.45	3.99	2.86	19.79	1.63
2	7.67	4.59	7.35	22.43	22.71	2.05	27	2.34	19.26	20.82	2.65	1.62	1.84
3	8.71	16.50	24.76	4.73	2.57	26.12	28	19.72	5.72	3.73	1.72	3.64	2.33
4	7.53	1.79	12.45	2.26	16.31	20.51	29	9.30	1.93	2.93	4.55	10.89	22.91
5	3.63	4.24	17.74	3.50	2.12	27.21	30	10.25	23.75	24.24	4.31	16.00	2.81
6	14.59	7.04	14.73	10.55	23.17	2.34	31	1.63	1.64	2.38	2.57	5.30	6.46
7	5.27	5.12	6.76	10.95	4.93	20.10	32	5.95	10.31	3.53	3.53	2.17	19.80
8	1.63	16.66	19.79	10.25	16.10	6.22	33	3.71	16.88	14.63	6.29	4.52	3.29
9	3.72	1.89	4.63	13.64	23.44	3.41	34	2.61	14.90	20.27	21.74	2.42	1.92
10	2.17	1.73	6.10	4.88	14.01	5.45	35	6.79	18.64	1.62	12.37	2.39	10.58
11	4.22	1.73	6.59	4.60	2.49	2.37	36	19.48	2.16	27.50	15.27	20.50	3.13
12	10.04	24.28	18.27	10.83	12.57	9.21	37	17.92	2.97	1.64	25.34	2.14	2.30
13	4.99	6.13	25.44	2.06	18.22	2.28	38	4.85	21.68	7.44	26.07	16.37	2.87
14	1.76	21.19	2.29	2.09	2.04	1.72	39	9.17	4.35	3.61	19.46	2.53	23.10
15	2.62	15.15	27.64	3.16	4.30	2.92	40	3.93	19.13	13.06	4.08	23.73	5.44
16	14.17	2.00	15.40	7.13	9.85	13.62	41	14.76	20.89	2.40	19.80	14.06	2.47
17	8.58	27.04	6.77	13.23	15.66	2.10	42	5.04	1.65	21.25	17.10	6.10	17.77
18	15.63	18.38	13.20	6.81	1.68	25.45	43	2.60	4.41	25.02	2.85	1.62	3.82
19	2.22	4.55	7.90	17.96	2.70	2.71	44	4.25	1.73	17.42	1.63	17.27	1.62
20	2.60	15.15	18.04	16.46	25.59	4.85	45	1.63	15.60	22.53	15.09	1.65	23.74
21	2.50	6.36	9.61	4.93	11.25	15.38	46	8.25	1.74	24.30	22.94	1.63	2.68
22	5.41	15.46	5.02	9.83	3.44	22.02	47	3.28	4.63	20.23	4.99	20.29	15.02
23	2.40	5.94	2.23	2.01	12.77	2.09	48	2.42	2.76	1.62	3.05	4.78	6.56
24	4.23	20.07	3.26	16.44	1.83	1.68	49	2.91	1.77	3.79	14.65	8.37	4.08
25	2.21	4.79	2.15	14.61	9.72	2.07	50	19.04	5.04	2.13	1.91	23.67	19.58

Table S6. Fifty scenarios of operating conditions

No.	u_1 (L/d)	u_2 (L/d)	u_3 (L/d)	u_4 (L/d)	No.	u_1 (L/d)	u_2 (L/d)	u_3 (L/d)	u_4 (L/d)
1	37.62	19.42	21.12	13.24	26	10.12	14.16	37.08	10.02
2	15.7	34.46	36.14	30.24	27	11.48	15.66	17.82	27.96
3	34.42	10.84	25.2	11.16	28	13.16	19.38	34.78	21.78
4	14.16	11.48	12.92	38.64	29	20.1	11.08	29.9	18.21
5	32.78	30.18	20.34	12.74	30	23.2	25.1	32.2	12.62
6	27.92	13	11.36	29.72	31	20.5	38.76	10.68	22.76
7	16.64	18.46	32.78	31.4	32	14.6	13.34	19.05	18.76
8	13.08	23.5	30.16	10.08	33	19.8	31.98	26.36	31.88
9	15.6	30.4	13.92	14.92	34	14.92	17.24	38.68	37.88
10	15.6	13.5	29.8	12.78	35	30	27.5	13.6	15.56
11	29.7	26.88	12.6	24.9	36	21.3	12.32	23.36	29.28
12	28.56	29.3	13.26	25.98	37	18.38	16.08	38.64	10.14
13	12.96	15.24	25.28	13.9	38	38	23.6	30.42	32.76
14	35.4	26.68	11.6	13.6	39	17.04	33.84	11.3	14.78
15	26.14	26.98	39.82	20.66	40	19.6	8.42	32.14	38.56
16	27.7	39.78	13.92	11.88	41	29.46	24.22	15.74	23.58
17	28.7	34.9	39.52	12.42	42	16.12	15.06	36.32	33.98
18	34.54	39.78	35.78	22.16	43	16.64	10.66	17.42	34.34
19	14.1	37.58	23.2	13.28	44	33.4	24.4	17.02	17.04
20	24.4	17.66	18.02	38	45	23.32	31.86	28.54	12.72
21	12.12	16.9	20.24	17.32	46	19.06	18.24	11.92	22.44
22	21.34	14.3	35.96	15.78	47	33.96	13.1	32.56	15.09
23	33.42	13.48	23.92	32	48	13.98	13.8	39.72	39.6
24	22.5	36.2	13	22.64	49	20.06	34.96	16.54	38.44
25	20.2	27.94	34.72	25.84	50	39.8	26.42	38.82	21.52

Table S7. Input and output variables for SI emulator and FCI simulator

SI Emulator	Input (I) or Output (O)	Symbol	FCI Simulator	Input (I) or Output (O)	Symbol
Highest contaminant concentration anywhere in the mesh	I	$\mathcal{Z}^{MAX}$	Highest contaminant concentration anywhere in the mesh	I	$\mathcal{Z}^{MAX}$
Percentage of benzene mass removal	I	η	Percentage of benzene mass removal	I	η
Injecting rate of well I	O	u_1	Injecting rate of well I	I	u_1
Injecting rate of well II	O	u_2	Injecting rate of well II	I	u_2
Pumping rate of well III	O	u_3	Pumping rate of well III	I	u_3
Pumping rate of well IV	O	u_4	Pumping rate of well IV	I	u_4
			Highest contaminant concentration anywhere in the mesh	O	$\mathcal{Z}^{MAX}$
			Percentage of benzene mass removal	O	η

Table S8. The result of the collinearity test

Variable	VIF	Tolerance
u1	1.068	0.936
u2	1.128	0.886
u3	1.043	0.959
u4	1.067	0.937