A Multi-chain Surrogate-assisted Hybrid Optimization Framework for Joint Identification of Groundwater Contaminant Sources and Hydrogeological Parameters

Mengtian Wu[1,2,3], Xuan Huang[4], Pengcheng Xu[5], Han Chen[6], Yang Xu[6], Jin Xu[6], Qingyun Duan[1,2,3]

[1] National Key Laboratory of Water Disaster Prevention, Hohai University, Nanjing, China
[2] College of Hydrology and Water Resources, Hohai University, Nanjing, China
[3] China Meteorological Administration Hydro-Meteorology Key Laboratory, Hohai University, Nanjing, China
[4] Key Laboratory of Taihu Basin Water Resources Research and Management of Ministry of Water Resources, Nanjing Hydraulic Research Institute, Nanjing, China
[5] Macau Environmental Research Institute, Faculty of Innovation Engineering, Macau University of Science and Technology, Macau, China
[6] State Key Laboratory of Water Cycle and Water Security, College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing, China

Correspondence to: Qingyun Duan (qyduan@hhu.edu.cn)

Abstract. Rapid and accurate identification of groundwater contaminant information and hydrogeological parameters is crucial for effective groundwater remediation and risk management. Within a simulation-optimization framework, this task is inherently posed as a mixed-variable optimization problem involving discrete parameters (e.g., source locations) and continuous ones (e.g., hydraulic heads, conductivities, and release fluxes). However, several challenges arise in this context. First, conventional optimization algorithms often exhibit slow convergence and unstable performance. Second, they typically require thousands of simulations to adequately explore the complex parameter space, resulting in prohibitive computational costs. To address these issues, this study develops a surrogate-assisted hybrid algorithm that integrates the Cooperative Search Algorithm (CSA) and Tabu Search (TS) within a synergistic multi-chain optimization framework, termed SA-CSA-TS. In each iteration, individual chains first perform independent CSA-based optimization to promote broad global exploration, after which they collaboratively refine source locations through a neighbourhood search guided by a shared tabu list. In addition, surrogate models equipped with a reconstruction strategy partially replace groundwater simulations, thereby substantially reducing the computational burden. Case studies reveal that the Radial Basis Function (RBF) outperforms other mainstream surrogate models in both accuracy and stability. Furthermore, comparative experiments confirm that the proposed SA-CSA-TS framework not only achieves higher solution accuracy but also significantly reduces computational demand, demonstrating strong potential for efficient groundwater contamination diagnosis.

https://doi.org/10.5194/egusphere-2025-6140
Preprint. Discussion started: 18 December 2025

1 Introduction

Groundwater contamination has become an increasingly critical issue, posing significant risks to environmental safety and public health (Gorelick and Zheng, 2015; Li et al., 2021a). Effective groundwater remediation requires rapid and accurate identification of contaminant source parameters (Bai and Tahmasebi, 2022; Mahar and Datta, 2001; Zhao et al., 2016). However, due to the invisibility of groundwater systems and sparse monitoring (Mirghani et al., 2009), source information cannot always be obtained directly. Instead, it must be inferred from observations, typically within a simulation-optimization (S-O) framework (Singh, 2015).

Within the S-O framework, simulation models such as MODFLOW, MT3DMS, and FEFLOW are employed to describe the spatial and temporal evolution of contaminant plumes (Delshad et al., 1996; Harbaugh, 2005; Zheng and Wang, 1999). The quality of candidate parameter sets is evaluated through performance metrics (e.g., NSE and RMSE) that measure the discrepancy between simulated and observed data. Optimization algorithms then iteratively adjust these parameters to minimize the selected metrics, thereby identifying the most probable parameter values. Common algorithms, including Genetic Algorithm (GA) (Ayvaz and Elci, 2018; Singh and Datta, 2006), Particle Swarm Optimization (PSO) (Meenal and Eldho, 2012; Pan et al., 2023), and Simulated Annealing (Jha and Datta, 2013), have demonstrated considerable success in groundwater contamination source identification (GCSI) (Swetha et al., 2025). Consequently, the S-O framework incorporating these groundwater models and algorithms has been widely adopted in groundwater contamination studies (Guneshwor et al., 2018).

Despite these advantages, the S-O framework still faces several challenges that hinder its accuracy and computational efficiency (Wu et al., 2022b). For instance, real-world GCSI often requires identifying source locations, which inherently transforms the task into a mixed-variable problem (Li et al., 2023). Such problems involve the simultaneous estimation of both discrete parameters (e.g., source locations) and continuous parameters (e.g., time-dependent contaminant release rates) (Wang et al., 2024). However, many existing optimization algorithms handle discrete variables through simple conversion techniques, such as binary encoding, grid-based discretization, or rounding schemes. These treatments can introduce approximation errors or impose artificial constraints, ultimately reducing solution quality. In addition, the mixed-variable structure produces highly complex, discontinuous, and multimodal objective landscapes. As a result, algorithms are more likely to converge prematurely to local optima (Chang et al., 2021).

For these reasons, some studies have introduced new or hybrid algorithms. For instance, Flying Foxes Optimization (FFO) has demonstrated superior search efficiency and accuracy in groundwater problems (Li et al., 2023). Similarly, the hybrid GA-PSO algorithm (Wang et al., 2015) improves performance by combining the global exploration capabilities of GA with the fast convergence of PSO, while Li et al. (2021b) also propose a Hybrid Homotopy-Genetic Algorithm. However, most of these approaches adopt a simultaneous optimization strategy that treats source locations and release rates as equivalent variables. In practice, this assumption oversimplifies the physical reality of groundwater transport. Source locations typically exert a dominant influence because they determine the transport pathways and the geometry of the plume.

In contrast, release rates and hydrogeological parameters mainly scale the concentration magnitudes. This sensitivity disparity creates a multimodal response surface, where multiple location combinations can reproduce sparse field observations with similar accuracy. This characteristic significantly increases the risk of premature convergence and may lead to the misidentification of critical source information.

The computational burden associated with GCSI cannot be ignored, as optimization algorithms often require thousands of simulations to adequately explore the parameter space (Asher et al., 2015; Ouyang et al., 2017). This intensive demand severely limits practical applications, particularly for complex or large-scale groundwater simulation models (Song et al., 2019). In this context, surrogate modelling, as a data-driven technique, has become a widely adopted choice (Song et al., 2018). By approximating the behaviour of high-fidelity groundwater models, surrogate models can enable more efficient and feasible source identification. Common surrogate models include Kriging, Gaussian Process (GP) (Rasmussen and Williams, 2006), Support Vector Regression (SVR) (Chang and Lin, 2011), Radial Basis Function (RBF) (Broomhead and Lowe, 1988), and ensembles of these models (Xing et al., 2019; Yin and Tsai, 2020; Zhu et al., 2024). However, most existing studies still select surrogate models based primarily on empirical preference, and few have systematically evaluated or compared their performance and suitability for groundwater systems (Hou and Lu, 2018; Wu et al., 2022a; Luo et al., 2025). To address this gap, the present study conducts a comprehensive comparison of mainstream surrogate models and identifies the most effective one for GCSI.

Overall, this study proposes a multi-chain surrogate-assisted hybrid optimization framework, termed SA-CSA-TS. The framework adopts a multi-chain structure operating across two distinct optimization stages. In the first stage, individual chains execute CSA-based optimization to enhance global exploration, with well-trained surrogate models replacing time-consuming groundwater simulations. In the second stage, chains collaboratively refine source locations through a neighbourhood search guided by a shared tabu list. This cooperative strategy enables efficient identification of source positions that control the contaminant plume distribution. To support the framework, several surrogate models are evaluated, and the RBF model is found to provide the most accurate approximation for groundwater applications. Case studies show that SA-CSA-TS can reduce computational cost by up to 85-88% while achieving higher identification accuracy than conventional algorithms. These results demonstrate the efficiency and reliability of the proposed framework and offer valuable insights for groundwater contamination remediation.

2 Methodology

For GCSI, this study adopts a simulation-optimization framework (Mahar and Datta, 2001). In this framework, the simulation model describes the spatial and temporal evolution of contaminants under specified input parameters, while the optimization algorithm iteratively refines these inputs to minimize the discrepancy between observed and simulated concentrations. Building on this concept, we develop a multi-chain surrogate-based optimization algorithm, SA-CSA-TS, which synergizes the Cooperative Search Algorithm (CSA) and Tabu Search (TS). The following sections first introduce the groundwater model used in this study and then describe the implementation and components of SA-CSA-TS.

2.1 Simulation

There are various effective simulation techniques available for groundwater modelling. In this study, MODFLOW 6 (Hughes et al., 2017) is adopted to simulate groundwater flow and pollutant transport, facilitated by the Python package FloPy (Bakker et al., 2016), which provides a convenient and flexible interface for model construction and execution. The governing partial differential equation for transient flow in a two-dimensional aquifer system can be given as follows:

$$\frac{\partial}{\partial x_i}\left(K_{i,j}\frac{\partial h}{\partial x_j}\right) + W = S_s\frac{\partial h}{\partial t} \tag{1}$$

where $K_{i,j}$ denotes the hydraulic conductivity, $m \cdot d^{-1}$; h denotes the hydraulic head, m; x_i and x_j are the coordinates along the axis, m; S_s is the specific storage of the porous material; and W is the volumetric flux per unit area.

Solute transport can be described by the following advection-dispersion-reaction equation under known hydrogeological conditions:

$$\frac{\partial}{\partial x_i}\left(\theta D_{ij}\frac{\partial c^k}{\partial x_j}\right) - \frac{\partial}{\partial x_i}(\theta v_i C^k) + q_s C_k^s + \sum R_n = \frac{\partial(\theta c^k)}{\partial t} \tag{2}$$

where θ is effective porosity; C^k is the dissolved concentration of the species k, $mg \cdot L^{-1}$; $D_{i,j}$ is the dispersion coefficient tensor, $m^2 \cdot d^{-1}$; v_i is the linear pore water velocity, $m \cdot d^{-1}$; q_s is the volumetric flow rate per unit volume, representing sources or sinks; C_k^s is the source or sink concentration of species k, $mg \cdot L^{-1}$; R_n is the chemical reaction term, $mg \cdot L^{-1} \cdot d^{-1}$.

It is worth noting that although MODFLOW 6 is used here, the simulation-optimization framework is versatile and applicable to other groundwater models.

2.2 Optimization

2.2.1 UQPyL

UQPyL[1] is a Python package developed by our team to support uncertainty quantification and optimization in computational modelling. The package integrates a comprehensive set of tools, including sampling techniques, surrogate modelling, parameter analysis methods, and global as well as hybrid optimization algorithms. Its modular and extensible design enables users to flexibly combine different components, facilitating rapid prototyping and testing of new algorithms. Moreover, UQPyL includes a default interface to couple external numerical simulators, making it suitable for computationally intensive applications such as groundwater modelling. In this study, UQPyL provides the foundation for

[1] www.uq-pyl.com

implementing the proposed SA-CSA-TS algorithm, conducting surrogate-model comparison experiments, and ensuring a consistent environment for benchmarking different optimization methods.

2.2.2 Cooperative Search Algorithm

The Cooperative Search Algorithm (CSA) is a population-based optimization method inspired by cooperative behaviours observed in social systems (Feng et al., 2021). CSA emphasizes team communication, reflective learning and internal competition among individuals. These mechanisms enable the algorithm to maintain diversity and accelerate convergence, making it suitable for solving high-dimensional, nonlinear, and multimodal problems. In CSA, a population of candidate solutions $\{x_i\}_{i=1}^{N}$ is initially generated. During the optimization process, individuals improve their positions by learning from others within the population. For example, at iteration t, the update of the i-th individual typically follows a team communication rule:

$$u_i^{t+1} = x_i^t + A_i^t + B_i^t + C_i^t \tag{3}$$

$$A_i^t = log\left(1/\phi(0,1)\right) \cdot (g_{ind}^t - x_i^t)$$

$$B_i^t = \alpha \cdot \phi(0,1) \cdot (gm^t - x_i^t)$$

$$C_i^t = \beta \cdot \phi(0,1) \cdot (pm^t - x_i^t)$$

where A_i^t, B_i^t and C_i^t denote the knowledge components from the chairman, board of directors, and board of supervisors, respectively. g_{ind}^t is the ind-th global best individual at iteration t. The gm^t represents the mean position of the top M global best individuals. The pm^t is the mean position of the ith personal best individual.

In addition, the individual is also updated by summing its own experience in its opposite direction, which can be expressed as follows:

$$v_{i,j}^{t+1} = \begin{cases} r_{i,j}^{t+1} \; if\left(u_{i,j}^{k+1} \geq c_j\right) \\ p_{i,j}^{t+1} \; if\left(u_{i,j}^{k+1} < c_j\right) \end{cases}$$

$$r_{i,j}^{t+1} = \begin{cases} \phi\left(ub_j + lb_j - u_{i,j}^{t+1}, c_j\right) & if(\left|u_{i,j}^{t+1} - c_j\right| < \phi(0,1) \cdot \left|ub_j - lb_j\right|) \\ \phi\left(lb_j, ub_j + lb_j - u_{i,j}^{t+1}\right) & otherwise \end{cases} \tag{4}$$

$$p_{i,j}^{t+1} = \begin{cases} \phi\left(c_j, ub_j + lb_j - u_{i,j}^{t+1}\right) & if(\left|u_{i,j}^{t+1} - c_j\right| < \phi(0,1) \cdot \left|ub_j - lb_j\right|) \\ \phi\left(\, ub_j + lb_j - u_{i,j}^{t+1}, ub_j\right) & otherwise \end{cases}$$

$$c_j = 0.5 \cdot \left(ub_j + lb_j\right)$$

where ub_j, lb_j denotes the jth value of the upper and lower bounds, respectively.

After generating candidate solutions, an elitist selection mechanism is employed to retain the best-performing individuals. Through repeated cooperative interactions and adaptive learning, the population collectively explores the search space and converges toward the global optimum.

Overall, CSA is a flexible and efficient optimization method that has been implemented within UQPyL and serves as an essential component of the hybrid optimization framework developed in this study.

2.2.3 Tabu search

Tabu Search (TS) is an optimization technique focused on local search, incorporating adaptive memory and strategic exploration. The core of TS is the tabu list, which stores recently visited positions and prevents their immediate reconsideration, thereby avoiding cycling and enabling systematic exploration of new regions. To mitigate excessive restrictions, TS includes aspiration criteria, allowing a tabu move if it improves upon any previously found solution. This balance of prohibition and relaxation helps TS avoid local optima and explore the search space more effectively.

In the context of GCSI, TS enhances the SA-CSA-TS by enabling structured exploration of discrete source-location configurations, thereby improving its ability to escape local traps and identify globally competitive solutions.

2.3 Surrogate models

To reduce the computational cost of repeated numerical simulations, surrogate models are incorporated into the optimization process (Razavi et al., 2012a). In groundwater applications, each candidate solution requires running simulations and post-processing to compute metrics such as RMSE and NSE, which is time-consuming. Surrogate models are therefore trained to approximate the relationship between model inputs and these metrics, enabling rapid evaluations and substantially accelerating the optimization.

Over the past two decades, surrogate models such as Kriging, Gaussian Process (GP) model, Support Vector Machine (SVR), and Radial Basis Function (RBF) have been widely used (Razavi et al., 2012b). All four surrogate models are implemented in UQPyL, and their performance differences are examined in this study.

2.3.1 Kriging

Kriging is a type of interpolation model originally developed in geostatistics. In UQPyL, Kriging is implemented based on the DACE toolbox within the Python environment. The detailed mathematical derivation of Kriging can be found in Lophaven et al. (2002). In brief, the predicted value $\tilde{f}(\boldsymbol{x})$ of an unknown input $\boldsymbol{x}$ is given by:

$$\tilde{f}(\boldsymbol{x}):\ \hat{y}(\boldsymbol{x}) = f(\boldsymbol{x})^T\boldsymbol{\beta}^* + r(\boldsymbol{x},\bar{\boldsymbol{x}})^T r^*$$

$$\boldsymbol{\beta}^* = (\mathbf{F}^T\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{R}^{-1}\mathbf{Y} \tag{5}$$

$$\boldsymbol{R}r^* = \mathbf{Y} - \mathbf{F}\beta^*$$

$$\mathbf{F} = \begin{bmatrix} f(\bar{\boldsymbol{x}}_1) \\ \vdots \\ f(\bar{\boldsymbol{x}}_{N_t}) \end{bmatrix} \quad \mathbf{R} = \begin{bmatrix} r(\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_1) & \dots & r(\bar{\boldsymbol{x}}_1,\bar{\boldsymbol{x}}_{N_t}) \\ \vdots & \ddots & \vdots \\ r(\bar{\boldsymbol{x}}_{N_t},\bar{\boldsymbol{x}}_1) & \dots & r(\bar{\boldsymbol{x}}_{N_t},\bar{\boldsymbol{x}}_{N_t}) \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_{N_t} \end{bmatrix}$$

where $f(\cdot)$ denotes the original function; $r(\cdot,\cdot)$ denotes the correlation function. In this study, the Gaussian correlation function is adopted.

2.3.2 Gaussian Process

A Gaussian Process (GP) can be viewed as a distribution over functions whose finite-dimensional realizations follow multivariate normal distributions. A detailed introduction to GP models is provided in Rasmussen and Williams (2006). Here, we present the prediction formula of an unknown input $\boldsymbol{x}$ in a GP model:

$$\tilde{f}(\boldsymbol{x}) : \hat{y}(\boldsymbol{x}) = \mathbf{K}_* \mathbf{K}^{-1} \mathbf{Y}$$

$$\mathbf{K}_* = \begin{bmatrix} k(\boldsymbol{x}, \boldsymbol{x}_1) \\ \vdots \\ k(\boldsymbol{x}, \boldsymbol{x}_{N_t}) \end{bmatrix} \quad \mathbf{K} = \begin{bmatrix} k(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_1) & \dots & k(\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_{N_t}) \\ \vdots & \ddots & \vdots \\ k(\bar{\boldsymbol{x}}_{N_t}, \bar{\boldsymbol{x}}_1) & \dots & k(\bar{\boldsymbol{x}}_{N_t}, \bar{\boldsymbol{x}}_{N_t}) \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} \bar{y}_1 \\ \vdots \\ \bar{y}_{N_t} \end{bmatrix} \tag{6}$$

where $k(\cdot,\cdot)$ denotes the kernel (correlation) function.

2.3.3 Support Vector Regression

This study adopts the ε-Support Vector Regression (ε-SVR) model. The prediction function is given by:

$$\tilde{f}(\boldsymbol{x}) = g(\boldsymbol{x}, \boldsymbol{w}^*) = \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{\varphi}(\boldsymbol{x}) + b \tag{7}$$

where $\boldsymbol{w}$ are the weights for each feature term; b denotes the constant term; $\varphi(\cdot)$ denotes the feature mapping. For convenience, we define $\boldsymbol{w}^* = [\boldsymbol{w}; b]$ and $\boldsymbol{x}^* = [\boldsymbol{x}; 1]$.

The optimal coefficients $\boldsymbol{w}^*$ are obtained by solving:

$$\begin{gathered} \boldsymbol{w}^* = argmin_{\boldsymbol{w}^*}(\frac{1}{2}\|w^*\|^2 + C\sum_{i=1}^{N_t}(\xi_i + \xi_i')) \\ s.t. f_i - (\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\varphi}(\boldsymbol{x}) + b) \leq \varepsilon + \xi_i \\ (\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\varphi}(\boldsymbol{x}) + b) - f_i \leq \varepsilon + \xi_i' \\ \xi_i, \xi_i' \geq 0 \end{gathered} \tag{8}$$

where C is the regularization factor; ξ_i and ξ_i' are slack variables; ε is the tolerance of prediction error. Both C and ε are user-defined parameters.

2.3.4 Radial Basis Function

Radial Basis Function (RBF) approximates a given data set using a weighted sum of radial basis functions. Given a dataset consisting of N_t training points $\bar{\boldsymbol{x}}$ and their corresponding response values, the potential function $\tilde{f}(x)$ can be expressed as:

$$\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^{N_t} \lambda_i \cdot \varphi(\boldsymbol{x}, \bar{\boldsymbol{x}}_i) + p(\bar{\boldsymbol{x}}) \tag{9}$$

where $\boldsymbol{\lambda}$ denotes the weights of the radial basis functions; $\varphi(\cdot)$ is the radial basis function (also known as the kernel function); $\bar{\boldsymbol{x}}$ represents the training data; and $p(\cdot)$ can be a polynomial function, a constant value, or no additional term, depending on the choice of radial basis function φ. In this study, the cubic radial basis function is used.

3 The SA-CSA-TS

3.1 Overview

This study develops a surrogate-assisted hybrid optimization algorithm, SA-CSA-TS, built upon a multi-chain framework in which each chain iteratively performs a two-stage search. Global exploration is conducted using the Cooperative Search Algorithm (CSA), followed by local refinement using Tabu Search (TS). To reduce dependence on computationally expensive groundwater simulations, surrogate models with dynamic reconstruction are embedded into both stages. In addition, designed inter-chain communication enables the exchange of evaluated samples, enhancing data diversity and improving surrogate accuracy.

Figure 1 illustrates the overall workflow. The process begins with initial sampling, and the groundwater model is used to evaluate these samples to initialize the chain archive D. After that, the algorithm enters the multi-chain optimization phase.

During each iteration, surrogate models are at first constructed. The key feature is synergistic learning, where each chain builds its surrogates not only from its own history but also from the evaluated solutions shared by other chains (see the red arrows in Fig. 1). In the first stage, each chain independently performs CSA under the guidance of surrogates to explore the global search space. The best individual from each chain is then evaluated using the groundwater simulator and used to update *D*. Local refinement is performed in the second stage. Before activating TS, the surrogate models are reconstructed using all newly obtained evaluations. TS subsequently explores neighbourhood solutions through multiple-move operators, and cooperation among chains is realized via a shared tabu list, which prevents redundant searches and promotes effective diversification. Surrogates continue to pre-screen candidate solutions, and only the most promising candidate from each chain is evaluated with the groundwater model. This iterative process continues until the predefined maximum evaluations of groundwater model FE_{max} is reached.

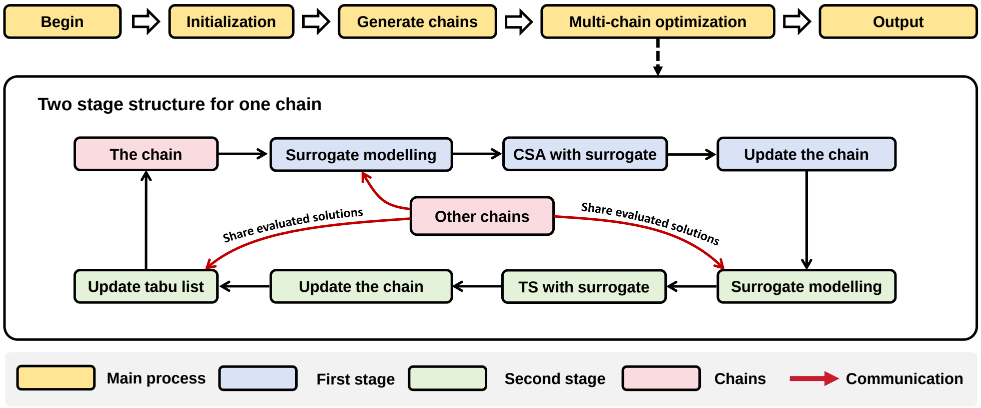


Figure 1: Overall framework of SA-CSA-TS

In summary, SA-CSA-TS enhances GCSI efficiency through three integrated mechanisms. First, the multi-chain framework enables synergistic learning by sharing evaluated information across chains. Second, the sequential deployment of CSA and TS provides a strong balance between global exploration and local intensification. Finally, surrogate models with dynamic reconstruction reduce computational burden while preserving high-fidelity prediction accuracy to guide the search effectively.

For clarity, the pseudocode of SA-CSA-TS is also provided:

Algorithm 1: SA-CSA-TS

Input: The maximum number of high-fidelity evaluations FE_{max}; The number of initial samples N_I; The number of chains K.

Output: The best optimal solution $best$.

```
Initialize $FE \leftarrow 0$; $T \leftarrow \emptyset$; $D \leftarrow \emptyset$;  /* T: tabu list, D: chain archive */
Initialize $N_I$ by Latin Hypercube Sampling
$S \leftarrow$ evaluated by the groundwater model and distributed evenly into $D$
$FE \leftarrow FE + N_I$
while $FE < FE_{max}$:
    Construct the surrogate model $\tilde{F}$ with $D$
    For each chain $k = 1$ to $K$:  /*Enter the first stage*/
        perform the cooperative search optimization using $\tilde{F}$ as the evaluator
    End For
    $S_1 \leftarrow$ Collect the best individual from each chain
    Evaluate $S_1$ using the groundwater model; $D \leftarrow D \cup S_1$
    Retrain the surrogate model $\tilde{F}$ with $D$
    For each chain $k = 1$ to $K$:  /* Enter the second stage */
        TS operator using $\tilde{F}$ as the evaluator
    End For
    $S_2 \leftarrow$ Collect the best individual from each chain
    Evaluate $S_2$ using the groundwater model; $D \leftarrow D \cup S_2$
    $T \leftarrow$ update the tabu list based on $S_2$
    $FE \leftarrow FE + 2 * K$
End While
  $best \leftarrow$ update the best optimal solution from $D$
Return $best$
```

3.2 Surrogate modelling

Within the optimization-simulation framework, the quality of candidate parameters is evaluated by the groundwater model (see the dashed line of Fig. 2). However, the entire optimization process typically requires thousands of forward simulations. To alleviate the computational demand, the SA-CSA-TS incorporates a surrogate modelling technique (see the solid line of Fig. 2).

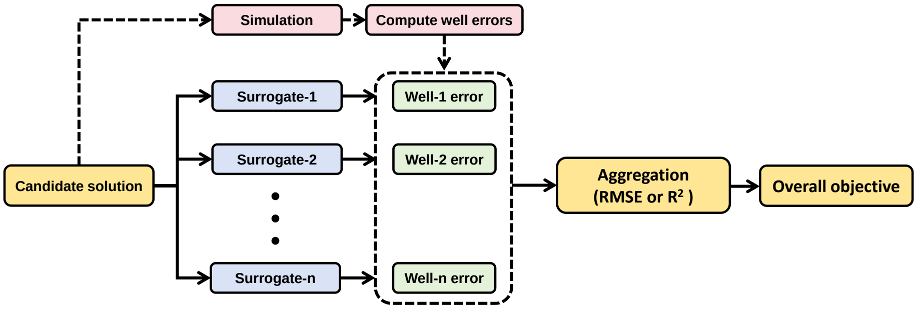


Figure 2: Workflow of solution evaluation with simulation or surrogate models

As illustrated in Fig. 2, a set of surrogate models is constructed to estimate the discrepancy (e.g., $RMSE$ or R^2) between simulated and observed concentrations at each monitoring well. Therefore, the number of surrogate models equals the number of observation wells. During optimization, these surrogates substitute for repeated groundwater simulations and provide rapid approximations of the error. The predicted discrepancies across all wells are then aggregated, and their sum is adopted as the overall objective function, guiding the evaluation of candidate parameters and the subsequent optimization.

3.3 Surrogate-assisted CSA

In SA-CSA-TS, the first stage focuses on global exploration, where each chain independently executes the Cooperative Search Algorithm (CSA) with the support of trained surrogate models. Fig. 3 illustrates the workflow of this surrogate-assisted CSA. Unlike the original CSA, the evaluation of individuals is performed using surrogate predictions instead of running computationally expensive groundwater simulations (the red dashed line of Fig. 3). Only the superior solutions produced in this stage are then used to update the chain's current position (Line 08 in Algorithm 1). Furthermore, this surrogate-assisted CSA module is implemented as a standalone benchmark algorithm, referred to as SA-CSA, enabling a direct comparison with the complete SA-CSA-TS to show the specific contributions of the multi-chain architecture and the Tabu Search.

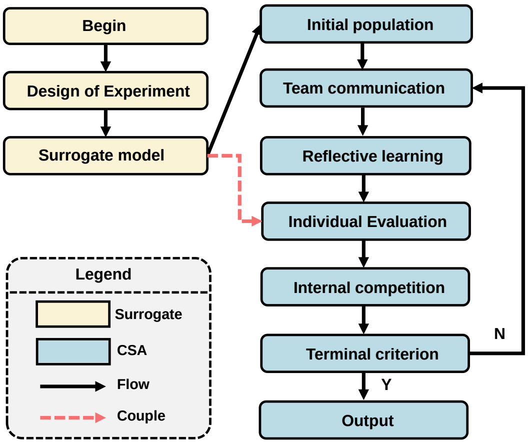


Figure 3: Workflow of surrogate-assisted CSA

3.4 Surrogate-assisted TS

Unlike the previous stage, where CSA operates independently in each chain, the Tabu Search (TS) stage is executed under a coordinated multi-chain framework. In this design, all chains share a common tabu list, which serves as a collective memory to prevent any chain from revisiting previously explored regions. The corresponding search mechanism is illustrated in Fig. 4. Guided by the retrained surrogate model, each chain explores its neighbourhood to identify promising candidates. As shown, the search trajectories are strictly constrained by the shared history, enabling the algorithm to better navigate multi-modal landscapes. For example, moves that enter tabu-listed areas (highlighted by red arrows) are prohibited. After selecting

https://doi.org/10.5194/egusphere-2025-6140
Preprint. Discussion started: 18 December 2025

the most promising solutions, the algorithm performs simulation-based evaluations and subsequently updates the shared tabu list, thereby allowing dynamic information exchange among all chains.

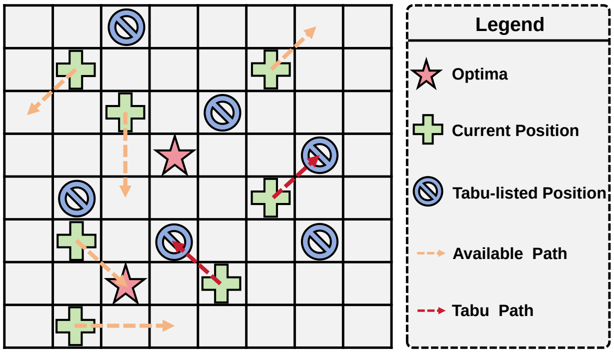


Figure 4: Diagram of multi-chain Tabu Search

We describe the rule for updating the tabu list. Let x_i and f_i denote the current solution and its objective value of the i-th chain, respectively, and let f_{best}^{i} represent the historical best objective value recorded by that chain. The update mechanism consists of the following three cases:

If $f_i > f_{best}^{i}$, the discrete component of x_i, denoted x_i^d, is added to the tabu list T, preventing the algorithm from revisiting this configuration in subsequent iterations.

If $f_i < f_{best}^{i}$, and $x_i^d \in T$, the tabu status of x_i^d is removed, allowing the algorithm to reconsider this configuration since a better solution has been found.

If $f_i < f_{best}^{i}$ and $x_i^d \notin T$, both the best solution x_{best} and the best objective f_{best}^{i} are updated accordingly.

4 Case studies

To comprehensively evaluate the performance of the proposed SA-CSA-TS algorithm, three case studies are conducted. Cases 1 and 2 are hypothetical scenarios designed to compare the effectiveness of different surrogate models and to enable an in-depth examination of the internal behaviour of SA-CSA-TS. Case 3 involves a practical engineering problem, suitable for validating the applicability and robustness of SA-CSA-TS in real-world conditions.

4.1 Case 1

The study area is a two-dimensional, homogeneous, anisotropic confined aquifer ($800\ m \times 1200\ m$), as illustrated in Fig. 5. The left and right boundaries are assigned constant hydraulic heads, and the remaining boundaries are treated as no-flow. For simulation, the domain is discretized into a grid of 16×24 cells, with a uniform cell size of $50\ m$. The basic hydrogeological parameters used in this case are summarized in Table 1.

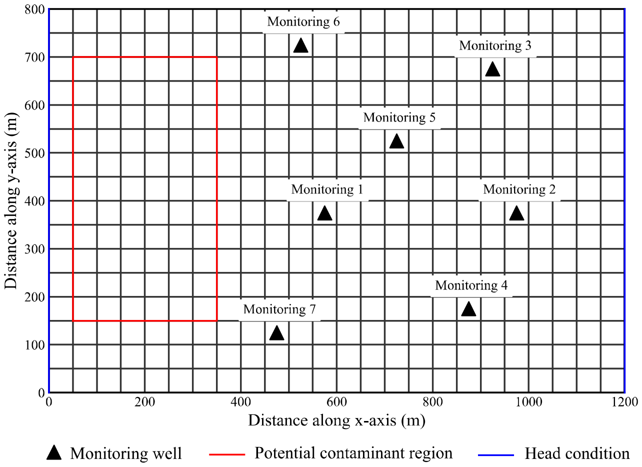


Figure 5: Schematic diagram in Case 1

Table 1: Basic values and ranges of hydrogeological parameters in Case 1

Name	Value or range
Hydraulic conductivity, K, m/day	15.0 – 35.0
Porosity, θ	0.25
Longitudinal dispersity, α_L, m	40.0
Transverse dispersity, α_T, m	15.0
Saturated thickness, b, m	20.0
Hydraulic head of the left boundary, H_1, m	40.0 – 50.0
Hydraulic head of the right boundary, H_2, m	30.0 – 40.0

The potential contamination source zone, also shown in Fig. 5, represents an industrial area with intensive activities, where contaminants may be intermittently released into the aquifer. Within this zone, one or more contamination sources may exist. To capture solute transport behaviour and provide data for the inverse analysis, seven monitoring wells are distributed across the study area (the triangle in Fig. 5).

In Case 1, a single contaminant source is considered. The total simulation time is 40 months, divided into 20 stress periods (SPs), with the source releasing contaminants only during the first five SPs. The true source location and its release fluxes for these five SPs are listed in Table S1. The contaminant plume distributions at the 5th and 10th SPs are shown in Fig. 6.

For this case, the parameters to be identified include:

(a) Hydrogeological parameters: The hydraulic conductivity (K) and the boundary head (H_1 and H_2). Their ranges are listed in Table 1;

(b) Source-related parameters: The source locations (SI and SJ, where SI denotes the grid index in the x-direction and SJ denotes the grid index in the y-direction, respectively) and their time-varying release fluxes (S_iP_t, where i denotes the index of the source, $i = 1$; and t denotes the index of the stress period, $t = 1$ to 5), with the value of each flux bounded between 0 and 100 kg/day.

https://doi.org/10.5194/egusphere-2025-6140
Preprint. Discussion started: 18 December 2025

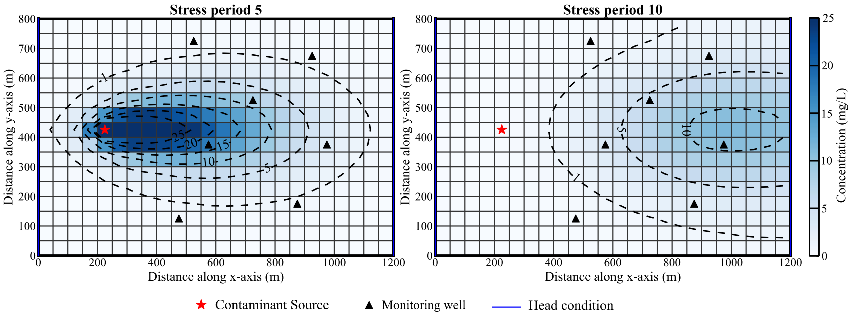


Figure 6: Distribution of contaminant plume in the 5th and 10th SPs of Case 1

4.2 Case 2

Case 2 adopts the same hydrogeological setting and numerical configuration as Case 1, but involves a more complex contamination scenario. In this case, three independent contaminant sources are introduced within the potential source zone. Their true locations and time-varying release fluxes are summarized in Table S2. The contaminant plume distribution at the 5th and 10th SPs is illustrated in Fig. 7.

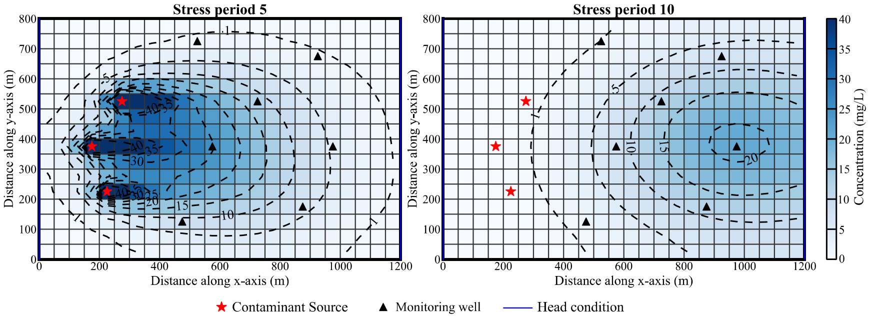


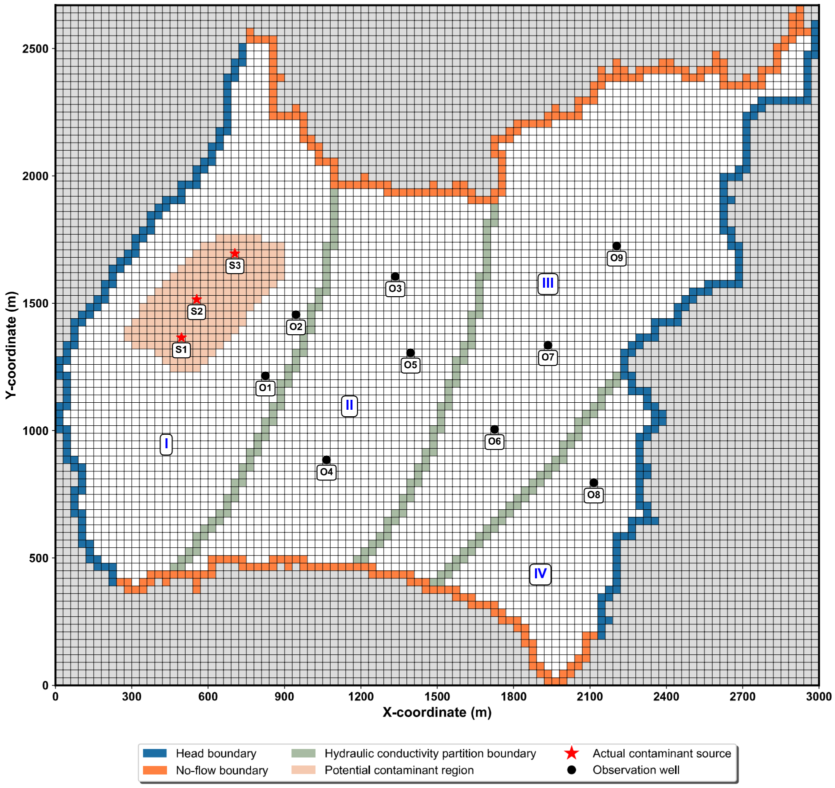
Figure 7: Distribution of contaminant plume in the 5th and 10th SPs of Case 2

Compared with Case 1, Case 2 presents a significantly higher level of complexity for surrogate modelling and optimization. The number of discrete variables associated with source locations increases from 2 to 6, and the total number of unknown parameters rises from 10 to 24 due to the introduction of additional sources and their time-varying release fluxes.

4.3 Case 3

This case study is designed as a realistic numerical experiment based on the hydrogeological conditions of a mining area in Henan Province, China. As shown in Fig. 8, the study area covers approximately 2.67 × 3 km. According to field investigations, the rivers on the western and eastern sides are conceptualized as constant-head boundaries, while the northern and southern edges, dominated by low-permeability granite with negligible recharge or discharge, are represented as no-flow boundaries. The aquifer is further divided into four hydraulic conductivity zones, denoted as Zones I, II, III, and IV. The basic hydrogeological parameters of the aquifer are summarized in Table 2.

https://doi.org/10.5194/egusphere-2025-6140
Preprint. Discussion started: 18 December 2025



325 **Figure 8: Overview of the research region in Case 3**

A potential contaminant source region is delineated, as highlighted in pink in Fig. 8. Field investigations identify three waste-ore deposits (S_1, S_2, S_3) within this region. These sources continuously release contaminants into the groundwater during the first five stress periods (out of a total of ten). Nine observation wells are distributed across the study area to monitor contaminant migration, and Fig. 9 illustrates the observed temporal concentration variations over the stress periods.

330 **Table 2: Basic settings of Case 3**

Name	Value or range
Hydraulic conductivity of Zone I, K_I, m/day	$15.0 - 35.0$
Hydraulic conductivity of Zone II, K_{II}, m/day	$15.0 - 35.0$
Hydraulic conductivity of Zone III, K_{III}, m/day	$15.0 - 35.0$
Hydraulic conductivity of Zone IV, K_{IV}, m/day	$15.0 - 35.0$
Porosity, θ	0.3
Longitudinal dispersity, α_L, m	40.0
Transverse dispersity, α_T, m	11.0
Saturated thickness, b, m	30.0
Hydraulic head of the left boundary, H_1, m	97.4
Hydraulic head of the right boundary, H_2, m	83.1

In summary, the parameters to be identified include: (a) Hydrogeological parameters: The hydraulic conductivity (K_1, K_2, K_3, K_4); (b) Source locations (SI_i and SJ_i, $i = 1, 2, 3$) and their release fluxes (S_iP_t, $i = 1, 2, 3$ and $t = 1\ to\ 5$), with the value of each flux bounded between 0 and 100 kg/day. Their actual values are listed in Table S3.

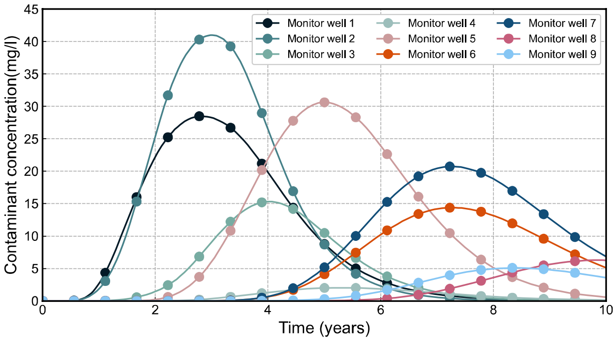


Figure 9: Observed concentrations at monitor wells in Case 3

5 Comparison of surrogate models

5.1 Experiment setup

This study employs four commonly used surrogate models to investigate their performance in predicting the discrepancy between observed and simulated data for a given set of solutions: a. Kriging; b. Gaussian Process (GP); c. Support Vector Regression (SVR); d. Radial Basis Function (RBF).

To ensure a fair comparison, all surrogate models are constructed using UQPyL on a computer equipped with 12th Gen Intel(R) Core (TM) i5-12490F CPU, and 32.0 GB of RAM. For model configurations, the Kriging model uses a Gaussian kernel, whereas GP and RBF adopt radial and cubic kernels, respectively. All remaining hyperparameters are kept at their default settings in UQPyL.

For sample generation, Latin Hypercube Sampling (LHS) is used in Cases 1-3 to produce a set of parameter samples, which are subsequently input into the groundwater models to obtain contaminant concentrations. For each sample, the RMSE between the simulated and observed concentrations at all monitoring wells is calculated. RMSE is selected here because it provides a steeper and more informative gradient, which is advantageous for optimization. The generated parameter sets and their corresponding RMSE values constitute the full input-output datasets.

To evaluate model performance, four training datasets, denoted as DS1-DS4 with sample sizes of 100, 200, 300, and 500, respectively, are constructed. An independent set of 50 samples is generated for testing.

5.2 Evaluation of surrogate models

As described earlier, SA-CSA-TS constructs individual surrogate models for each monitoring well, and the corresponding outputs are summed to derive an ensemble objective value for optimization. To evaluate the effectiveness of this approach, we first examine the ensemble prediction performance of four surrogate models across Cases 1-3, based on the coefficient of determination (R^2). The results are summarized in Table 3.

Table 3: Ensemble prediction performance of four surrogate models

Case	Surrogate	Dataset			
		DS1	DS2	DS3	DS4
Case 1	KRG	0.73	0.80	0.87	0.89
	GP	0.68	0.78	0.90	0.91
	SVR	0.46	0.54	0.72	0.75
	RBF	0.81	0.88	0.95	0.95
Case 2	KRG	0.60	0.71	0.83	0.83
	GP	0.55	0.74	0.84	0.85
	SVR	0.35	0.47	0.62	0.64
	RBF	0.71	0.85	0.91	0.91
Case 3	KRG	0.53	0.68	0.77	0.79
	GP	0.45	0.65	0.80	0.81
	SVR	0.30	0.37	0.46	0.48
	RBF	0.68	0.83	0.88	0.90

Across all datasets (DS1-DS4) and all three cases, RBF clearly delivers the most stable and accurate ensemble predictions. KRG and GP achieve acceptable accuracy, whereas SVR consistently performs the weakest. All models benefit from increasing training data. In comparison, RBF demonstrates a superior sensitivity to data enrichment, aligning well with the iterative reconstruction strategy of SA-CSA-TS. In Case 3, the prediction task becomes significantly more challenging due to more complex hydrogeological conditions, leading to lower R^2 values for all models. However, RBF still maintains robust predictive capability.

Based on the ensemble results, Cases 2 and 3 under dataset DS3 are selected for detailed surrogate evaluation at the individual monitoring wells. These two cases represent more challenging prediction scenarios. In addition, DS3 provides a sufficiently informative training set, yielding a clear performance improvement over DS2, while the additional gain from DS3 to DS4 is marginal. Figure 10 illustrates the prediction performance for Case 2 using DS3. Accuracy varies substantially across monitoring wells, primarily due to the spatial distribution of the contaminant plume. Wells 1, 2, and 5 are located within the main plume body, where steep and highly nonlinear concentration gradients dominate. Consequently, all surrogate models except RBF show marked reductions in R^2 at these locations. In contrast, Wells 6 and 7 lie far from the plume centre, where concentration gradients are smooth, enabling all models to reach their highest performance. A similar trend is observed in Case 3 (see Fig. 11). Wells situated in high-gradient zones (e.g., Wells 1, 2, 3, and 5) pose greater challenges, leading to noticeable performance declines for SVR, GP, and KRG. In contrast, RBF consistently maintains strong performance across all monitoring wells.

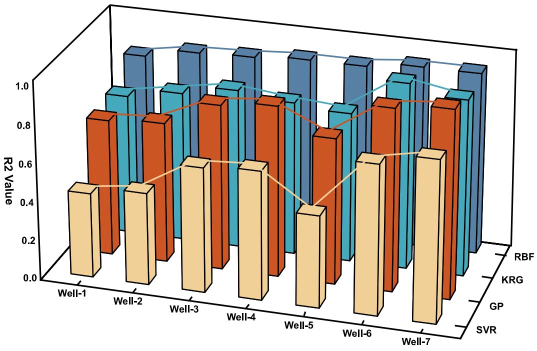


Figure 10: Prediction performance of surrogate models for Case 2 under the dataset DS3

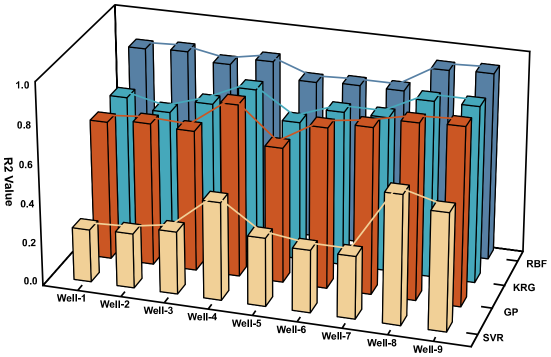


Figure 11: Prediction performance of surrogate models for Case 3 under the dataset DS3

Figures 12 and 13 present the sample-wise predicted values at representative locations: Well 1 for Case 2 and Well 5 for Case 3. In both scenarios, RBF achieves the highest R² and lowest RMSE, followed by KRG, GP, and SVR. In Case 3, SVR fails to capture the nonlinearity of contaminant concentrations, with its predictions collapsing into a narrow range. For optimization applications, high fidelity in the low-value region of the response is particularly important, as deviations in this domain can significantly affect the quality of the optimal solution. RBF provides more stable and accurate predictions in these low-value zones, further reinforcing its reliability as a surrogate model for optimization.

In addition to prediction accuracy, the computational cost of training is a critical consideration for SA-CSA-TS, which involves iterative surrogate reconstruction. Theoretically, GP and KRG are computationally intensive with a complexity of $O(k \cdot N^3)$, where k denotes the number of iterations required by the construction algorithm. In contrast, RBF and SVR offer higher computational efficiency, with complexities of $O(N^3)$ and $O(N^2) \sim O(N^3)$, respectively. This theoretical advantage is further supported by empirical results obtained using UQPyL on dataset DS4. In terms of actual training time, GP and KRG require approximately 1 s and 4 s, whereas RBF and SVR significantly reduce the cost to 0.22 s and 0.01 s.

In summary, RBF overcomes the precision limitations of SVR while avoiding the computational inefficiencies associated with KRG and GP. It thus provides the best balance between accuracy and efficiency, making it the most suitable surrogate model for the proposed SA-CSA-TS framework.

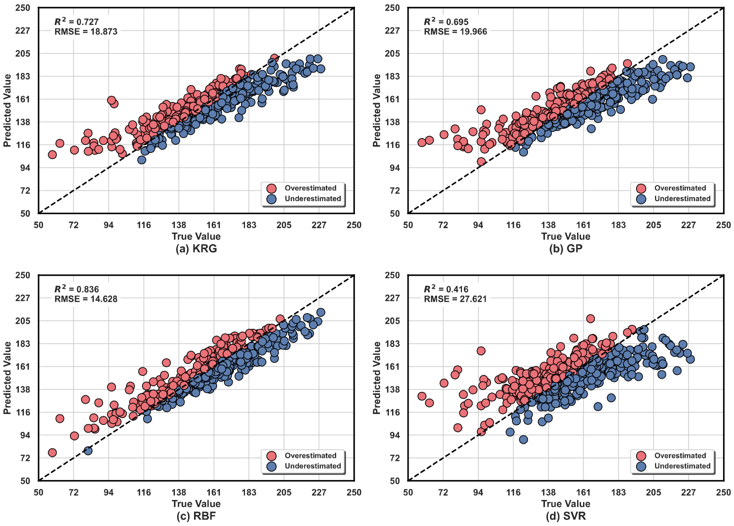


Figure 12: Sample-wise predicted values of surrogate models in Case 2, Well-1

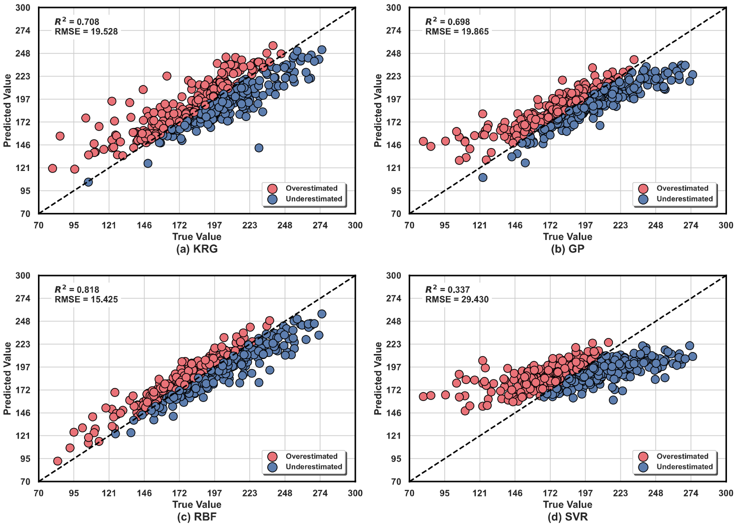


Figure. 13 Sample-wise predicted values of surrogate models in Case 3, Well-5

6 Optimization

6.1 Experiment setup

This section aims to investigate the performance of SA-CSA-TS in GCSI. For comparison, three additional optimization algorithms are considered: Genetic Algorithm (GA), Cooperative Search Algorithm (CSA) and SA-CSA. GA is widely used as a benchmark, whereas CSA represents a state-of-the-art method in recent years. SA-CSA is included to isolate and assess the contributions of the multi-chain framework and the Tabu Search. All algorithms are implemented within UQPyL to ensure a consistent and fair computational environment.

For the standard evolutionary algorithms (GA and CSA), the maximum number of simulations (FE_{max}) and the population size N_p are set to 20,000 and 100. For GA, the user-defined parameters p_c, η_c, p_m, η_m are set to 1, 20, $1/D$, 20, respectively, where D denotes the dimensionality of the problem. For CSA, the parameters are set as $\alpha = 0.10$, $\beta = 0.15$, and $M = 3$.

For surrogate-assisted algorithms (SA-CSA-TS and SA-CSA), the RBF model is employed. The FE_{max} is reduced to 2,000, as surrogate models enable efficient optimization with substantially fewer exact evaluations. Based on the results summarized in Table 3, the number of initial samples N_I for surrogate construction is set to 300. For SA-CSA-TS, the number of chains is set to $K = 10$.

For Cases 1-3, the optimization problem is formulated as:

$$minimize: f = \sum_{m=1}^{M}\left(\sqrt{\frac{\sum_{t=1}^{T}\left(S_m^t - O_m^t\right)^2}{T}}\right) \tag{10}$$

$$subject\ to: LB \leq \{H, K, SI, SJ, SP\} \leq UB$$

where S_m^t and O_m^t represent the simulated and observed concentrations at the m-th monitoring well in stress period t, respectively. LB and UB are the lower and upper bounds of parameters to be estimated.

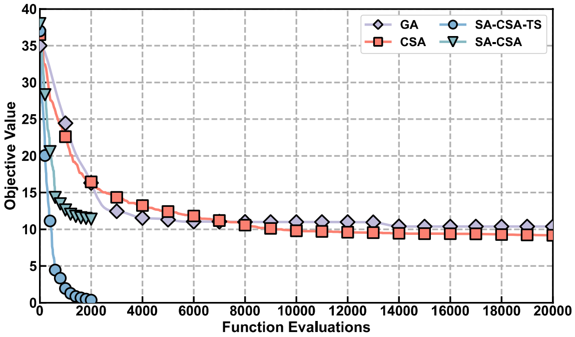


Figure 14: Convergence curves of all algorithms in Case 1

6.2 Optimization Results

6.2.1 Case 1

Figure 14 presents the convergence curves of the four algorithms in Case 1. SA-CSA-TS achieves the best objective value (0.35) within only 2,000 simulation runs, outperforming CSA, GA, and SA-CSA. As listed in Table 4, while all algorithms achieve satisfactory calibration for hydrogeological parameters, SA-CSA-TS is the only algorithm that consistently identifies the true source location $(5, 9)$ and release fluxes. This discrepancy highlights that the primary bottleneck lies in the discrete source search, where the proposed two-stage framework with Tabu Search effectively prevents the search chains from becoming trapped in local basins. Moreover, relative to conventional optimization approaches, the surrogate-assisted framework significantly reduces computational cost while maintaining high-quality solutions.

Table 4: Optimization results of all algorithms in Case 1

Algorithms	Location	Hydrogeological parameters			Release fluxes (kg/day)					Objective
	(SI, SJ)	H_1	H_2	K	S_1P_1	S_1P_2	S_1P_3	S_1P_4	S_1P_5	value
SA-CSA-TS	（5, 9）	42.3 (0.9%)	35.1 (0.5%)	18.3 (1.1%)	20.7 (3.3%)	51.7 (1.0%)	13.1 (3.0%)	41.6 (2.2%)	23.8 (3.9%)	0.35
SA-CSA	(4, 10)	43.2 (1.2%)	35.7 (1.1%)	17.8 (1.7%)	19.1 (8.9%)	49.6 (4.7%)	12.2 (6.4%)	43.5 (8.8%)	21.6 (2.1%)	11.38
GA	（6, 8）	42.4 (0.7%)	34.7 (1.7%)	18.5 (1.7%)	19.6 (6.8%)	50.4 (3.0%)	11.7 (9.8%)	36.7 (8.2%)	19.0 (13.6%)	10.37
CSA	（3, 7）	43.7 (2.3%)	35.9 (1.7%)	18.2 (0.6%)	19.9 (5.3%)	52.3 (0.7%)	12.5 (3.5%)	42.3 (5.8%)	20.9 (5.1%)	9.19

6.2.2 Case 2

Compared to Case 1, Case 2 involves three contaminant sources and therefore requires more parameters to be identified. Figure 15 presents the convergence behaviour of all algorithms. SA-CSA-TS achieves the best objective value (1.29), followed by CSA (18.23), SA-CSA (21.38) and GA (22.85). SA-CSA-TS also converges much more rapidly, stabilizing within the first 1,500 simulation runs.

Figure 16 compares the optimal solutions obtained by all algorithms. Higher radial values indicate more accurate estimates, with 100% denoting a perfect match to the true values. SA-CSA-TS encloses the largest area in the radar chart, indicating the highest overall estimation accuracy. While all algorithms provide satisfactory estimates of hydrogeological parameters, only SA-CSA-TS correctly identifies the three contaminant source locations (highlighted in red in Fig. 16). Other algorithms exhibit noticeable deviations. Moreover, these incorrect source locations are accompanied by inaccurate release rates, suggesting that location errors are compensated by adjustments to other parameters, leading the search into local optima. Overall, with the assistance of surrogate models and Tabu Search, SA-CSA-TS demonstrates a strong ability to avoid such local traps and to accurately resolve the multi-source identification problem under this more complex scenario.

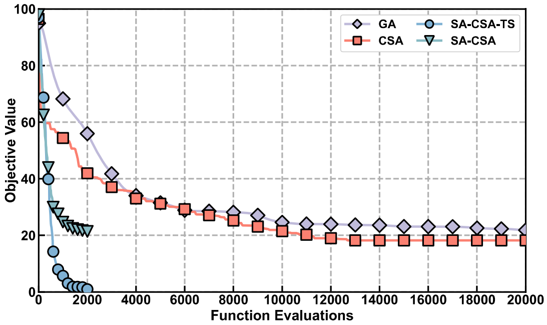
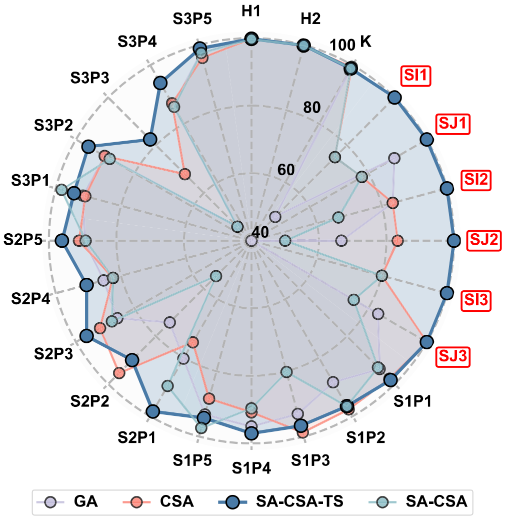


Figure 15: Convergence curves of all algorithms in Case 2



445

Figure 16: Radar chart comparing the optimal solutions obtained by all algorithms in Case 2

6.2.3 Case 3

Case 3 presents the most challenging optimization landscape due to the increased number of parameters and scenario complexity. As illustrated in Fig. 17, the surrogate-assisted algorithms maintain a distinct efficiency advantage. In particular, SA-CSA-TS rapidly converges to the best solution within 2,000 simulations, whereas GA and CSA stagnate at significantly higher objective values. 450

Figure 18 details the identification accuracy for specific parameters. Consistent with previous cases, all algorithms estimate the hydrogeological parameters (K_1-K_4) with acceptable accuracy. However, a sharp performance divergence is observed in the source-related parameters: only SA-CSA-TS maintains high accuracy for the location variables (SI, SJ), 455 while other algorithms exhibit substantial deviations. This failure to pinpoint source locations explains the stagnation observed in the other methods. Overall, Case 3 confirms that surrogate models effectively reduce computational cost, and that the multi-chain framework is indispensable for ensuring robustness and avoiding local optima in practical problems.

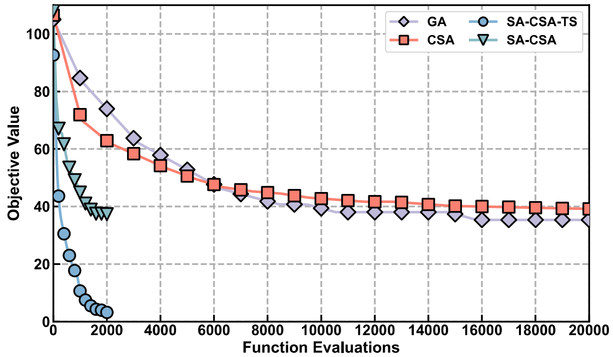


Figure 17: Convergence curves of all algorithms in Case 3

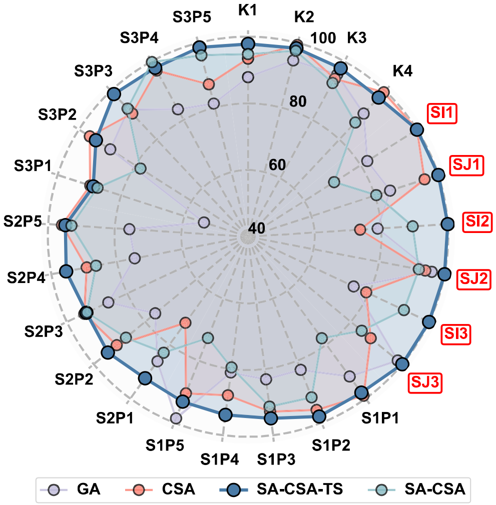


Figure 18: Radar chart comparing the optimal solutions obtained by all algorithms in Case 3

7 Discussion

7.1 Effects of surrogate models

Surrogate models are incorporated into SA-CSA-TS to alleviate the computational burden of high-fidelity simulations. Figure 19 breaks down the runtime of all algorithms across the three cases into simulation time (blue) and algorithm time (red). It is evident that the simulation cost overwhelmingly dominates the total runtime. Although surrogate-assisted methods introduce a slight overhead for model construction and updating, this cost is negligible compared to the time savings achieved by reducing high-fidelity evaluations. Specifically, in three case studies, SA-CSA-TS reduces the total runtime by approximately 85-88%, compared to the GA and CSA. This result confirms that the efficiency advantage of the surrogate-assisted framework becomes increasingly pronounced as the problem complexity grows.

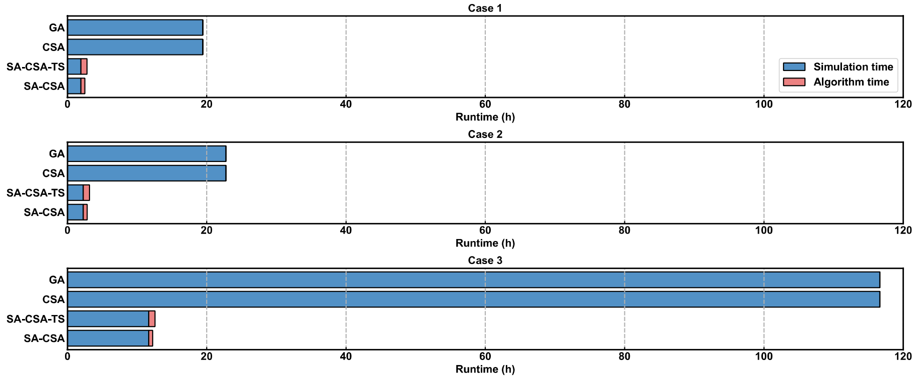


Figure 19: Runtime breakdown of all algorithms across three cases

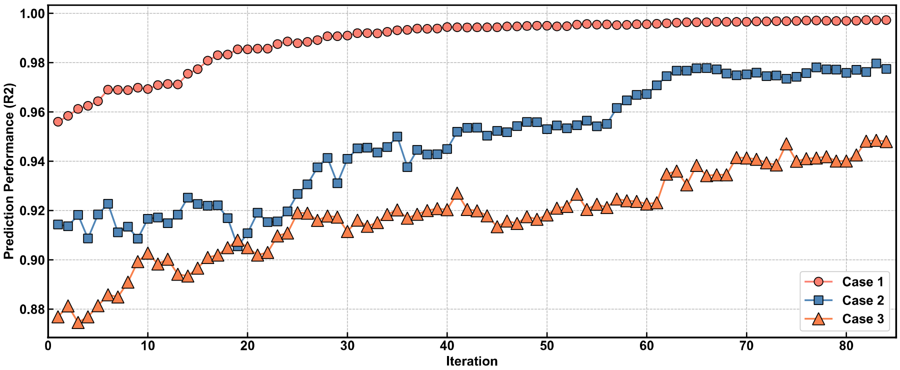


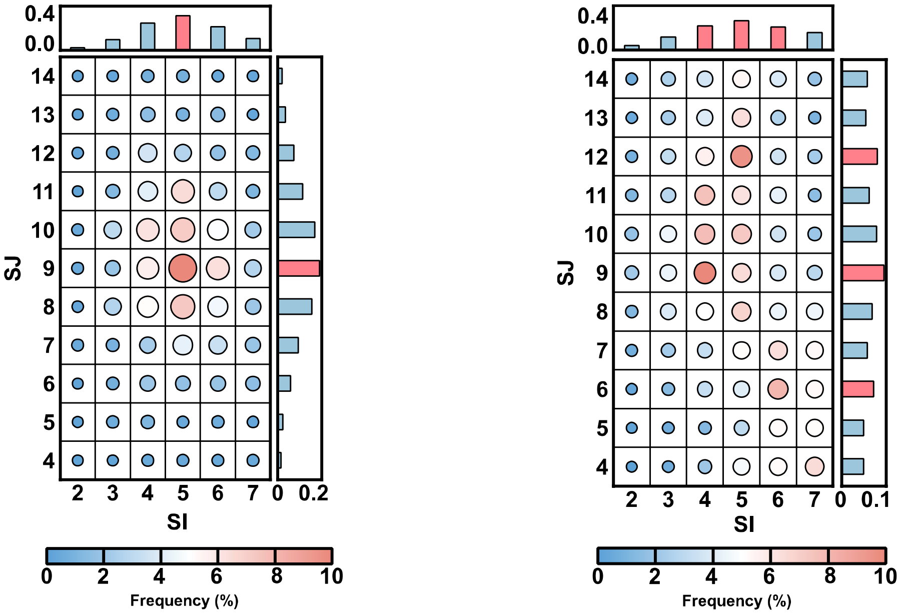
Figure 20: Evolution of the prediction accuracy of the RBF model on the validation set during the optimization process

Given the negligible overhead of surrogate modelling, the effect of the iterative reconstruction strategy is further examined. Figure 20 tracks the evolution of prediction accuracy (R^2) of surrogate models on a validation set during optimization. In Case 1, the accuracy remains high and stable. In contrast, Cases 2 and 3 exhibit noticeable fluctuations. These oscillations are not indicators of failure but rather reflect the algorithm's active exploration of underrepresented regions. Driven by the Tabu Search mechanism, the optimizer periodically escapes local basins and enters unexplored areas where the surrogate model initially has lower accuracy. However, the subsequent recovery of R^2 values confirms that the surrogate model successfully adapts to these new regions. Crucially, this dynamic updating process prevents the convergence stagnation observed in the other algorithms, ensuring that the search remains robust even in complex landscapes.

7.2 Effects of the multi-chain framework

Groundwater contaminant source identification is an inherently multi-modal optimization problem, where inaccurate location estimates may easily trap algorithms in inferior local solutions. As observed in Figs. 14, 15, and 17, GA, CSA and SA-CSA frequently exhibited instability and stagnation. This failure is largely attributed to their reliance on a single search population or trajectory, which lacks the mechanism to escape local basins. In contrast, SA-CSA-TS successfully identified the source information in all three cases. To understand this mechanism, we examine the behaviour of the proposed multi-chain framework.

Figure 21 depicts the search-frequency maps of candidate source locations by SA-CSA-TS for Case 1 and Case 2. In both scenarios, the true source locations (marked by red bars) correspond to the highest visit frequencies (red circles), indicating that the majority of chains consistently converge toward the correct region. Notably, the surrounding cells also exhibit high visit frequencies. This phenomenon confirms the parameter-compensation effect, where spatial inaccuracies are temporarily balanced by adjustments in release fluxes or hydraulic conductivity. This "equifinality" trap explains why conventional algorithms often stagnate near, but not exactly at, the true source. Furthermore, Case 2 displays more dispersed secondary hotspots than Case 1, reflecting a more rugged landscape with stronger compensability. Despite this complexity, the proposed framework successfully concentrates the search effort on the true location, demonstrating robust global convergence.



(a)

(b)

Figure 21: Search-frequency maps of candidate source locations obtained by the multi-chain framework in (a) Case 1 and (b) Case 2. The red bars indicate the true source locations.

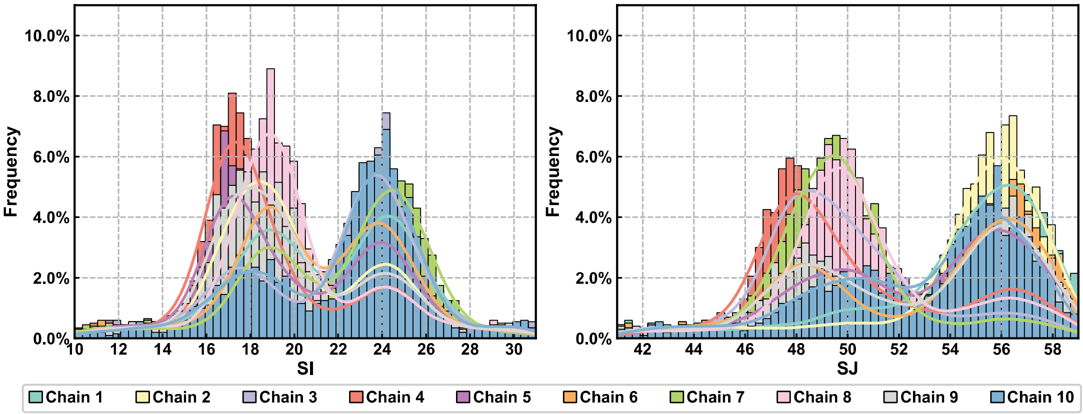


Figure 22: Distribution of search trajectories across ten chains for source coordinates in Case 3. The fitted curves highlight the multi-modal nature of the search landscape.

Figure 22 provides a deeper insight by analysing the distribution of search trajectories across ten independent chains in Case 3. The histograms for the source coordinates (*SI* and *SJ*) reveal a distinct multi-modal distribution, confirming the existence of multiple local optima. While the majority of chains converge to the primary peak (the true source), a few chains (e.g., Chains 1, 2, and 10) are entrapped in secondary peaks. Therefore, if a single-chain method (like standard GA or CSA) is used, and it happens to follow the trajectory of Chain 1, the identification would fail entirely. However, the multi-chain framework mitigates this risk by exploring multiple basins simultaneously. This collective intelligence allows the algorithm to filter out local optima and stabilize estimates around the true global solution, effectively overcoming the equifinality and multi-modality challenges that hinder conventional single-population methods.

7.3 Robustness analysis

To evaluate the robustness of SA-CSA-TS under data uncertainty, additional experiments are conducted based on three case studies. Random Gaussian noise with varying levels (0.5%, 1%, and 2%) is superimposed on the noise-free observation data, following the equation:

$$C_{obs}^{*} = C_{true} \cdot (1 + \delta \cdot \xi)$$

where C_{obs}^{*} and C_{true} denote the noisy and noise-free observations, respectively; δ denotes the noise level; and ξ is a random number following the standard normal distribution $N(0, 1)$.

Figure 23 illustrates the Average Relative Errors (ARE) for the three cases under these noise levels. A clear trend is observed where the identification error increases marginally with the noise intensity. Specifically, for Case 1, the ARE rises from 1.59% (noise-free) to 3.09% (2% noise). For the more complex scenarios in Cases 2 and 3, the errors start at approximately 3.7–3.8% and increase to roughly 4.5% under the maximum noise level. Despite these increases, the average errors for all cases consistently remain below 5%, indicating that the proposed method maintains high performance without significant degradation when observation data is subject to measurement noise.

Tables S4-S6 provide the specific identification results in three cases. Notably, the discrete source locations match the true values exactly across all noise levels. As for continuous variables, the hydrogeological parameters show only slight fluctuations. In comparison, the source release parameters exhibit relatively larger variations. This phenomenon is largely attributed to the complementary effects between different stress periods or among multiple sources, where slight deviations in one parameter may compensate for another. Despite this, the overall errors remain within an acceptable range, confirming the robustness of SA-CSA-TS against data uncertainty.

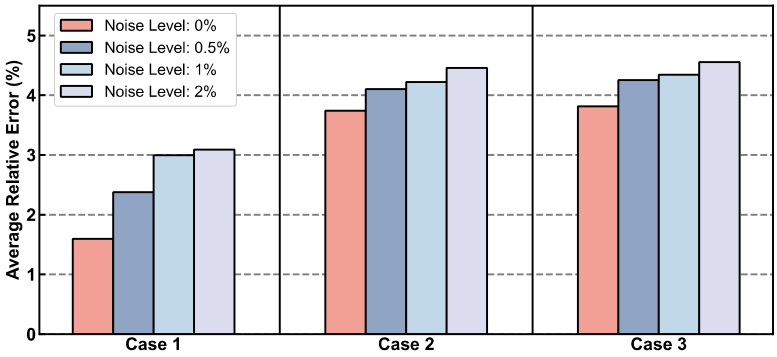


Figure: 23 Comparison of average relative errors for three cases under different noise levels

8 Conclusions

This study proposes a multi-chain surrogate-assisted hybrid optimization algorithm, SA-CSA-TS, to address the challenges of prohibitive computational costs and multi-modal complexity in GCSI. The algorithm incorporates three key innovations. First, surrogate models are embedded to alleviate the computational burden, while continuous iterative updates ensure reliable optimization guidance. Second, a multi-chain synergistic learning framework enables the exchange of evaluated samples among chains, enhancing data diversity and preventing premature convergence caused by limited local information. Third, a two-stage sequential strategy is employed where CSA conducts global exploration and TS performs neighbourhood refinement guided by a shared tabu list, effectively balancing exploration and exploitation.

Through three illustrative case studies, the applicability of different surrogate models and the overall performance of the proposed algorithm were systematically investigated. Results indicate that the Radial Basis Function (RBF) offers the best balance of stability and accuracy, particularly excelling in fitting low-value regions, making it the optimal surrogate for this framework. Comparative experiments with four algorithms (SA-CSA-TS, GA, CSA, and SA-CSA) highlight the superior robustness and accuracy of the proposed framework. While the benchmark algorithms frequently stagnate in local optima due to the parameter-compensation effect, SA-CSA-TS successfully identifies the true contaminant source parameters by leveraging multi-chain cooperation to escape local entrapment. Furthermore, the algorithm achieves a computational cost reduction of approximately 85-88% across the three cases, proving it to be both a precise and efficient tool for GCSI. Future work will focus on extending this framework to highly heterogeneous aquifers and exploring parallel computing techniques to further enhance its applicability in real-time emergency response.

Code and Data availability

The codes and case studies used in this work are available at https://doi.org/10.5281/zenodo.17862863 (Wu, 2025) and maintained at the GitHub repository (https://github.com/smasky/SA-CSA-TS). All numerical experiments are carried out using the UQPyL platform, which is available at http://www.uq-pyl.com (or https://github.com/smasky/UQPyL).

CRediT author statement

Mengtian Wu: Methodology, Software, Writing – original draft, Writing – review & editing, Funding acquisition; Xuan Huang: Methodology, Software; Pengcheng Xu: Methodology, Software; Xu Yang: Software; Han Chen: Methodology, Software; Jin Xu: Methodology; Qingyun Duan: Conceptualization, Methodology, Funding acquisition, Project administration

Acknowledgments

This study was supported by the Jiangsu Provincial Science and Technology Basic Research Program Youth Fund Project [grant number BK20241516], the National Natural Science Foundation of China [grant numbers 42101046 and W2431029], the National Key R&D Program of China [grant number 2021YFC3201102], and the Key Scientific and Technological Project of the Ministry of Water Resources of the P.R.C. [grant number SKS-2022001].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Asher, M.J., Croke, B.F.W., Jakeman, A.J., Peeters, L.J.M., 2015. A review of surrogate models and their application to groundwater modeling: SURROGATES OF GROUNDWATER MODELS. Water Resour. Res. 51, 5957–5973. https://doi.org/10.1002/2015wr016967

Ayvaz, M.T., Elci, A., 2018. Identification of the optimum groundwater quality monitoring network using a genetic algorithm based optimization approach. J. Hydrol. 563, 1078–1091. https://doi.org/10.1016/j.jhydrol.2018.06.006

Bai, T., Tahmasebi, P., 2022. Characterization of groundwater contamination: a transformer-based deep learning model. Adv. Water Resour. 164, 104217. https://doi.org/10.1016/j.advwatres.2022.104217

Bakker, M., Post, V., Langevin, C.D., Hughes, J.D., White, J.T., Starn, J.J., Fienen, M.N., 2016. Scripting MODFLOW Model Development Using Python and FloPy. Groundwater 54, 733–739. https://doi.org/10.1111/gwat.12413

Broomhead, D.S., Lowe, D., 1988. Multivariable functional interpolation and adaptive networks. Complex Systems 2, 321–355.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2, 27:1-27:27. https://doi.org/10.1145/1961189.1961199

Chang, Z., Lu, W., Wang, Z., 2021. A differential evolutionary markov chain algorithm with ensemble smoother initial point selection for the identification of groundwater contaminant sources. J. Hydrol. 603, 126918. https://doi.org/10.1016/j.jhydrol.2021.126918

Delshad, M., Pope, G.A., Sepehrnoori, K., 1996. A compositional simulator for modeling surfactant enhanced aquifer remediation, 1 formulation. J. Contam. Hydrol. 23, 303–327. https://doi.org/10.1016/0169-7722(95)00106-9

Feng, Z., Niu, W., Liu, S., 2021. Cooperation search algorithm: A novel metaheuristic evolutionary intelligence algorithm for numerical optimization and engineering optimization problems. Appl. Soft Comput. 98, 106734. https://doi.org/10.1016/j.asoc.2020.106734

Gorelick, S.M., Zheng, C., 2015. Global change and the groundwater management challenge. Water Resour. Res. 51, 3031–3051. https://doi.org/10.1002/2014WR016825

Guneshwor, L., Eldho, T.I., Kumar, A.V., 2018. Identification of groundwater contamination sources using meshfree RPCM simulation and particle swarm optimization. Water Resour. Manage. 32, 1517–1538. https://doi.org/10.1007/s11269-017-1885-1

Harbaugh, A.W., 2005. MODFLOW-2005, the US Geological Survey modular groundwater model-the groundwater flow process. Cent. Integr. Data Anal. Wis. Sci. Cent.

Hou, Z., Lu, W., 2018. Comparative study of surrogate models for groundwater contamination source identification at DNAPL-contaminated sites. Hydrogeol. J. 26, 923–932. https://doi.org/10.1007/s10040-017-1690-1

Hughes, J.D., Langevin, C.D., Banta, E.R., 2017. Documentation for the MODFLOW 6 framework (USGS Numbered Series No. 6-A57), Documentation for the MODFLOW 6 framework, Techniques and Methods. U.S. Geological Survey, Reston, VA. https://doi.org/10.3133/tm6A57

Jha, M., Datta, B., 2013. Three-dimensional groundwater contamination source identification using adaptive simulated annealing. J. Hydrol. Eng. 18, 307–317. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000624

Li, J., Lu, W., Fan, Y., 2021. Groundwater pollution sources identification based on hybrid homotopy-genetic algorithm and simulation optimization. Environ. Eng. Sci. 38, 777–788. https://doi.org/10.1089/ees.2020.0117

Li, P., Karunanidhi, D., Subramani, T., Srinivasamoorthy, K., 2021. Sources and consequences of groundwater contamination. Arch. Environ. Contam. Toxicol. 80, 1–10. https://doi.org/10.1007/s00244-020-00805-z

Li, Y., Lu, W., Pan, Z., Wang, Z., Dong, G., 2023. Simultaneous identification of groundwater contaminant source and hydraulic parameters based on multilayer perceptron and flying foxes optimization. Environ. Sci. Pollut. Res. 30, 78933–78947. https://doi.org/10.1007/s11356-023-27574-1

Lophaven, S.N., Nielsen, H.B., Søndergaard, J., 2002. DACE: a Matlab kriging toolbox. Citeseer.

Luo, C., Wang, X., Xu, Y.J., Jia, S., Liu, Z., Mao, B., Lv, Q., Ji, X., Rong, Y., Dai, Y., 2025. Synergistic identification of hydrogeological parameters and pollution source information for groundwater point and areal source contamination based on machine learning surrogate–artificial hummingbird algorithm. Hydrol. Earth Syst. Sci. 29, 5719–5736. https://doi.org/10.5194/hess-29-5719-2025

Mahar, P.S., Datta, B., 2001. Optimal identification of ground-water pollution sources and parameter estimation. J. Water Resour. Plan. Manag.-asce 127, 20–29. https://doi.org/10.1061/(ASCE)0733-9496(2001)127:1(20)

Meenal, M., Eldho, T.I., 2012. Simulation-optimization model for groundwater contamination remediation using meshfree point collocation method and particle swarm optimization. Sadhana-acad. Proc. Eng. Sci. 37, 351–369. https://doi.org/10.1007/s12046-012-0086-0

Mirghani, B.Y., Mahinthakumar, K.G., Tryby, M.E., Ranjithan, R.S., Zechman, E.M., 2009. A parallel evolutionary strategy based simulation-optimization approach for solving groundwater source identification problems. Adv. Water Resour. 32, 1373–1385. https://doi.org/10.1016/j.advwatres.2009.06.001

Ouyang, Q., Lu, W., Miao, T., Deng, W., Jiang, C., Luo, J., 2017. Application of ensemble surrogates and adaptive sequential sampling to optimal groundwater remediation design at DNAPLs-contaminated sites. J. Contam. Hydrol. 207, 31–38. https://doi.org/10.1016/j.jconhyd.2017.10.007

Pan, Z., Lu, W., Wang, H., Bai, Y., 2023. Groundwater contaminant source identification based on an ensemble learning search framework associated with an auto xgboost surrogate. Environ. Modell. Software 159, 105588. https://doi.org/10.1016/j.envsoft.2022.105588

Rasmussen, C.E., Williams, C.K.I., 2006a. Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA, USA.

Rasmussen, C.E., Williams, C.K.I., 2006b. Gaussian processes for machine learning, Adaptive computation and machine learning. MIT Press, Cambridge, Mass.

Razavi, S., Tolson, B.A., Burn, D.H., 2012a. Numerical assessment of metamodelling strategies in computationally intensive optimization. Environ. Modell. Software 34, 67–86. https://doi.org/10.1016/j.envsoft.2011.09.010

Razavi, S., Tolson, B.A., Burn, D.H., 2012b. Review of surrogate modeling in water resources. Water Resour. Res. 48, W07401. https://doi.org/10.1029/2011wr011527

Singh, A., 2015. Review: computer-based models for managing the water-resource problems of irrigated agriculture. Hydrogeol. J. 23, 1217–1227. https://doi.org/10.1007/s10040-015-1270-1

Singh, R.M., Datta, B., 2006. Identification of groundwater pollution sources using GA-based linked simulation optimization model. J. Hydrol. Eng. 11, 101–109. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:2(101)

Song, J., Yang, Y., Chen, G., Sun, X., Lin, J., Wu, Jianfeng, Wu, Jichun, 2019. Surrogate assisted multi-objective robust optimization for groundwater monitoring network design. J. Hydrol. 577, 123994. https://doi.org/10.1016/j.jhydrol.2019.123994

Song, J., Yang, Y., Wu, Jianfeng, Wu, Jichun, Sun, X., Lin, J., 2018. Adaptive surrogate model based multiobjective optimization for coastal aquifer management. J. Hydrol. 561, 98–111. https://doi.org/10.1016/j.jhydrol.2018.03.063

Swetha, K., Eldho, T.I., Singh, L.G., Kumar, A.V., 2025. Groundwater contaminant source identification using swarm intelligence-based simulation optimization models. Environ. Sci. Pollut. Res. Int. 32, 1626–1639. https://doi.org/10.1007/s11356-024-35850-x

Wang, J.L., Lin, Y.H., Lin, M.D., 2015. Application of heuristic algorithms on groundwater pumping source identification problems, in: 2015 IEEE International Conference on Industrial Engineering and Engineering Management (Ieem).

Presented at the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE, New York, pp. 858–862.

Wang, Z., Lu, W., Chang, Z., Zhang, T., 2024. Joint identification of groundwater pollution source information, model parameters, and boundary conditions based on a novel ES-MDA with a wheel battle strategy. J. Hydrol. 636, 131320. https://doi.org/10.1016/j.jhydrol.2024.131320

Wu, M., Wang, L., Xu, J., Hu, P., Xu, P., 2022a. Adaptive surrogate-assisted multi-objective evolutionary algorithm using an efficient infill technique. Swarm Evol. Comput. 75, 101170. https://doi.org/10.1016/j.swevo.2022.101170

Wu, M., Wang, L., Xu, J., Wang, Z., Hu, P., Tang, H., 2022b. Multiobjective ensemble surrogate-based optimization algorithm for groundwater optimization designs. J. Hydrol. 612, 128159. https://doi.org/10.1016/j.jhydrol.2022.128159

Wu, M. 2025. SA-CSA-TS: A multi-chain surrogate-assisted hybrid optimization algorithm combining CSA and TS. Zenodo. https://doi.org/10.5281/zenodo.17862863.

Xing, Z., Qu, R., Zhao, Y., Fu, Q., Ji, Y., Lu, W., 2019. Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model. J. Hydrol. 572, 501–516. https://doi.org/10.1016/j.jhydrol.2019.03.020

Yin, J., Tsai, F.T.-C., 2020. Bayesian set pair analysis and machine learning based ensemble surrogates for optimal multi-aquifer system remediation design. J. Hydrol. 580, 124280. https://doi.org/10.1016/j.jhydrol.2019.124280

Zhao, Y., Lu, W., Xiao, C., 2016. A kriging surrogate model coupled in simulation-optimization approach for identifying release history of groundwater sources. J. Contam. Hydrol. 185, 51–60. https://doi.org/10.1016/j.jconhyd.2016.01.004

Zheng, C., Wang, P.P., 1999. MT3DMS: a modular three-dimensional multispecies transport model for simulation of advection, dispersion, and chemical reactions of contaminants in groundwater systems; documentation and user’s guide.

Zhu, L., Lu, W., Luo, C., Xu, Y., Wang, Z., 2024. An ensemble optimizer with a stacking ensemble surrogate model for identification of groundwater contamination source. J. Contam. Hydrol. 267, 104437. https://doi.org/10.1016/j.jconhyd.2024.104437