Supporting Information for

Enhancing Inverse Modeling in Groundwater Systems through Machine Learning: A Comprehensive Comparative Study

Junjun Chen[1,2], Zhenxue Dai[2,3], Shangxian Yin[4], Mingkun Zhang[5], Mohamad Reza Soltanian[6]

[1] National and Local Joint Engineering Laboratory of Internet Application Technology on Mine, China University of Mining and Technology, Xuzhou, 221008, China

[2] College of Construction Engineering, Jilin University, Changchun, 130026, China

[3] School of Environmental and Municipal Engineering, Qingdao University of Technology, Qingdao, 273400, China

[4] College of Safety Engineering, North China Institute of Science and Technology, Langfang, 065201, China

[5] Shandong Rui Yi technology development Co., Ltd., Jinan, 250000, China

[6] Departments of Geosciences and Environmental Engineering, University of Cincinnati, OH, 45220, USA

Correspondence to: Zhenxue Dai (dzx@jlu.edu.cn), Shangxian Yin (yinshx03@126.com)

Contents of this file

Table S1 to Table S3
Figures S1 to S44

Introduction

Table S1 is the optimal hyperparameters for MSVR by four metaheuristic algorithms

Table S2 and **S3** are $RMSE_{(All)}$ and R^2_{All} values of FC-DNN with different number of hidden layers

Table S4 is the RMSE values of estimated log-permeability fields for the four metaheuristic algorithms and the TNNA algorithm under Scenario 1-4.

Fig.S1. to **Fig.S2** are detail architectures of LeNet and ResNet.

Fig.S3. to **Fig.S6.** are pair-wise comparisons for four surrogate modeling methods.

Fig.S7. to **Fig.S10.** are spatial distributions of log-permeability field estimation results and absolute errors by four metaheuristic algorithms for Scenarios 1~4

Fig.S11. to **Fig.S14.** are spatial distributions of log-permeability field estimation results and absolute errors by the TNNA method for Scenarios 1~4

Fig.S15. to **Fig.S44.** are spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations using the TNNA algorithm and four metaheuristic algorithms (from day 2 to day 60 recorded every two days).

Table S1 Optimal hyperparameters for MSVR by four metaheuristic algorithms

Training data number	Optimal algorithms	*C*	*ε*	*σ*	MSE
200	GA	18.640	6.0117E-03	0.398	6.2652E-02
	DE	**27.526**	**4.8503E-03**	**0.391**	**6.2498E-02**
	PSO	**27.526**	**4.8503E-03**	**0.391**	**6.2498E-02**
	SA	35.533	8.3451E-06	0.334	6.2499E-02
500	GA	54.278	4.9071E-03	0.509	4.9246E-02
	DE	39.979	3.0950E-03	0.867	4.9729E-02
	PSO	48.596	3.5939E-03	0.706	4.9215E-02
	SA	**32.241**	**5.5964E-03**	**0.615**	**4.8987E-02**
1000	**GA**	**23.296**	**4.2424E-03**	**0.724**	**4.3391E-02**
	DE	40.680	3.9406E-03	0.585	4.3556E-02
	PSO	25.317	6.1069E-03	0.820	4.3510E-02
	SA	71.104	4.0023E-05	0.561	4.3777E-02
2000	GA	61.888	1.1828E-03	0.918	3.5188E-02
	DE	53.579	1.6516E-03	0.964	3.5137E-02
	PSO	**50.431**	**9.4148E-04**	**0.921**	**3.5120E-02**
	SA	50.307	9.0781E-03	1.033	3.5265E-02

Note: The rows in bold represent the optimal hyperparameter configurations corresponding to the smallest MSE values.

Table S2 $\mathrm{RMSE}_{(\mathrm{All})}$ values of FC-DNN with different number of hidden layers

Training data number	Hidden layer number						
	1	2	3	4	5	6	7
200	0.07588	**0.05882**	0.06870	0.17916	0.16125	0.13690	0.13340
500	0.07050	0.04308	0.03788	**0.03786**	0.05824	0.09567	0.10229
1000	0.05118	0.03571	**0.02703**	0.02732	0.02866	0.04213	0.07825
2000	0.03936	0.02944	**0.02090**	0.02168	0.02580	0.03064	0.06887

Note: The bold values represent the smallest MSE values among the considered seven hidden layer numbers.

Table S3 R^2_{All} values of FC-DNN with different number of hidden layers

Training data number	Hidden layer number						
	1	2	3	4	5	6	7
200	0.94140	**0.96479**	0.95197	0.67332	0.73539	0.80926	0.81890
500	0.94942	0.98111	0.98540	**0.98541**	0.96548	0.90685	0.89351
1000	0.97334	0.98703	**0.99256**	0.99240	0.99164	0.98194	0.93768
2000	0.98424	0.99118	**0.99555**	0.99522	0.99323	0.99045	0.95173

Note: The bold values represent the largest R^2_{All} values among the considered seven hidden layer numbers.

Table S4. RMSE values of estimated log-permeability fields for the four metaheuristic algorithms and the TNNA algorithm under Scenario 1-4.

Scenarios	Metaheuristic algorithms					TNNA	
		GA	DE	PSO	SA		
Scenario 1	N_{PC}=100	0.7844	0.5984	0.9423	0.7720	epoch=200	0.4895
	N_{PC}=500	0.8246	0.7639	0.6379	0.8980	epoch=1000	0.4748
	N_{PC}=1000	0.6659	0.6391	0.7127	0.8012		
Scenario 2	N_{PC}=100	0.9554	0.5223	0.8785	0.6987	epoch=200	0.4317
	N_{PC}=500	0.6164	0.4925	1.0293	1.1549	epoch=1000	0.4271
	N_{PC}=1000	0.5389	0.5322	0.9686	0.6288		
Scenario 3	N_{PC}=100	0.5386	0.3892	0.5486	0.5647	epoch=200	0.3161
	N_{PC}=500	0.4339	0.4271	0.5762	0.5714	epoch=1000	0.2970
	N_{PC}=1000	0.4060	0.5042	0.6295	0.5558		
Scenario 4	N_{PC}=100	0.4436	0.3841	0.5723	0.6459	epoch=200	0.2749
	N_{PC}=500	0.4265	0.3971	0.3770	0.5654	epoch=1000	0.2328
	N_{PC}=1000	0.3653	0.3459	0.5367	0.5033		

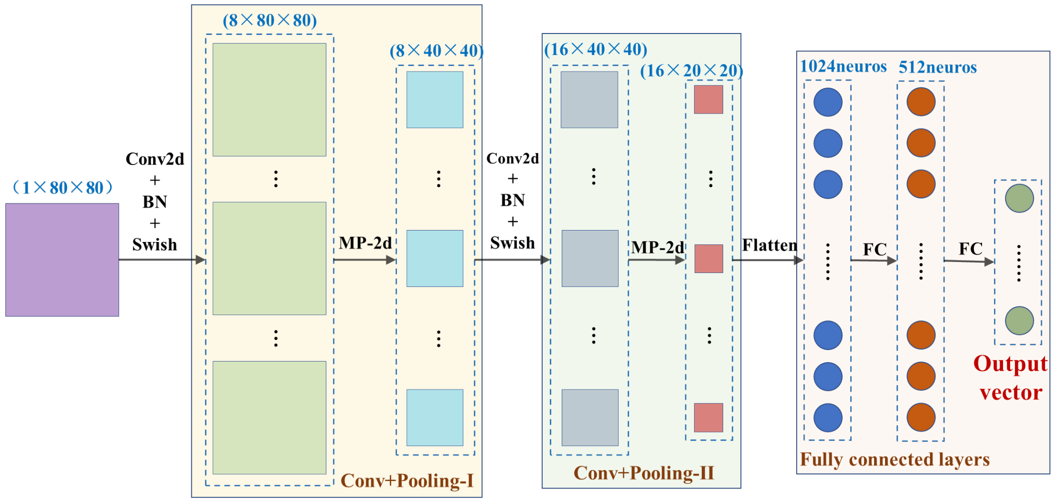

Fig.S1. Detailed architecture of a LeNet based CNN. The input matrix data are obtained according to Figure 2(c) and subjected to feature extraction through a sequence of two convolutional and pooling layers, subsequently connected to the output layer using a flatten layer and two fully connected layers.

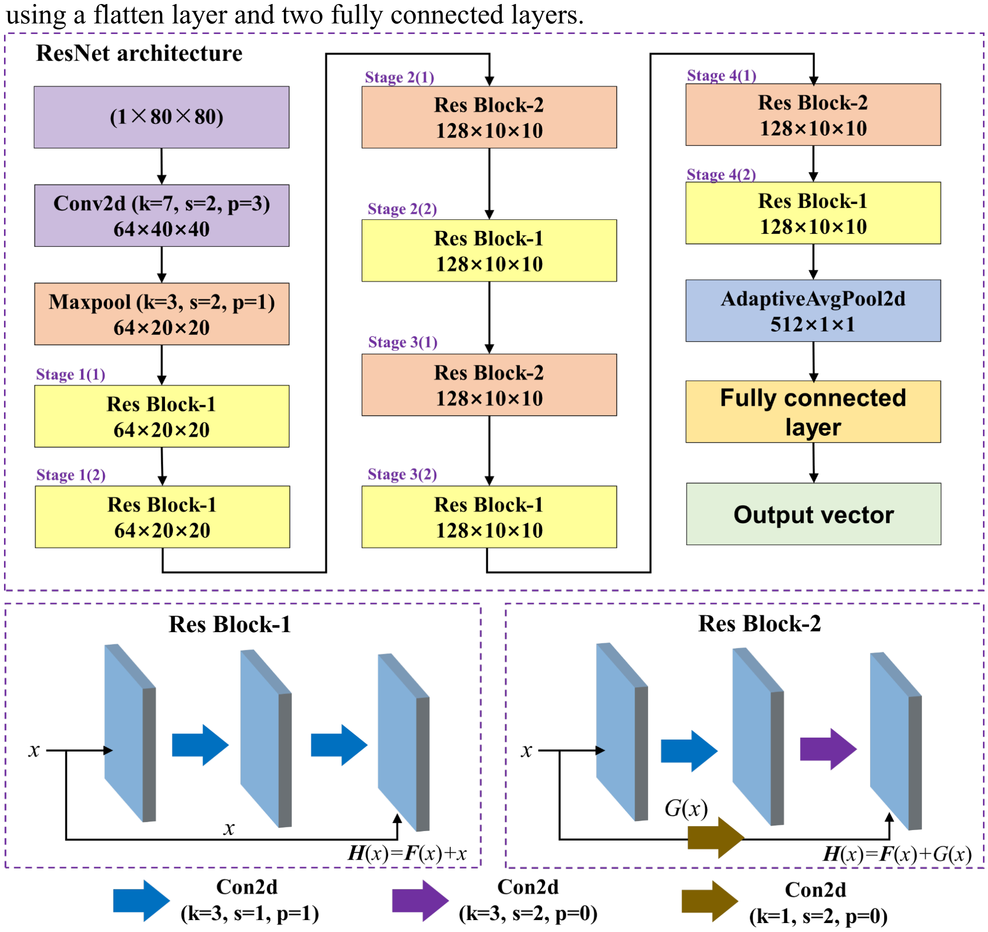

Fig.S2. Detailed architecture of a ResNet based CNN. The input matrix data are obtained according to Figure 2(c). “Res Block-1” and “Res Block-2” are two different types of residual blocks used in this ResNet. Eight residual blocks in four stages are designed in this ResNet. “Stage i (j)” represents the jth residual block used in stage i.

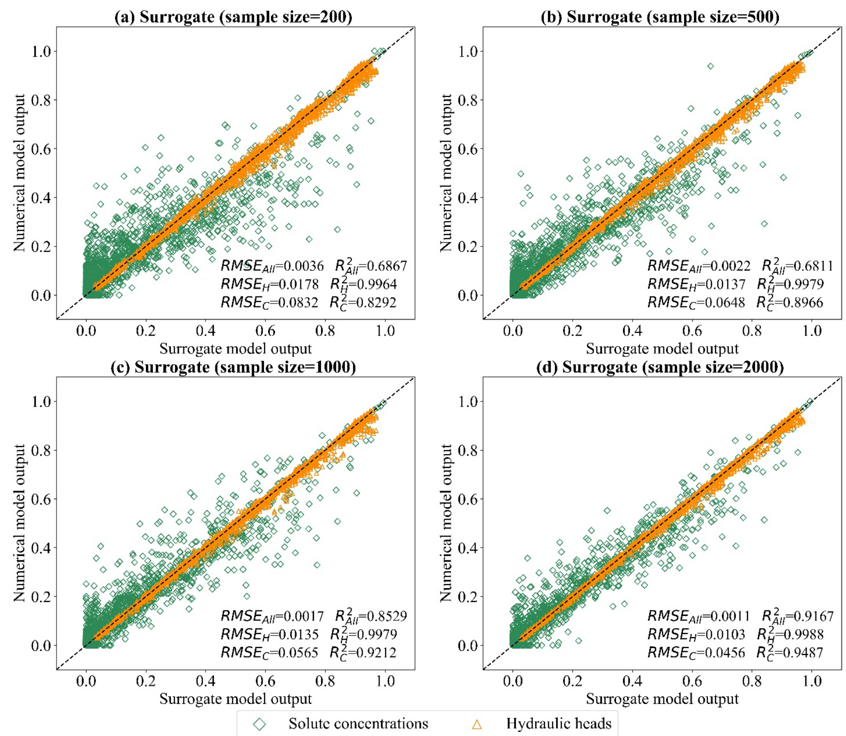

Fig.S3. Performance of MSVR based surrogate models for the solute concentration and hydraulic head prediction. (a~d) are pair-wise comparisons based on surrogate models trained by 200, 500, 1000, and 2000 training samples, respectively.

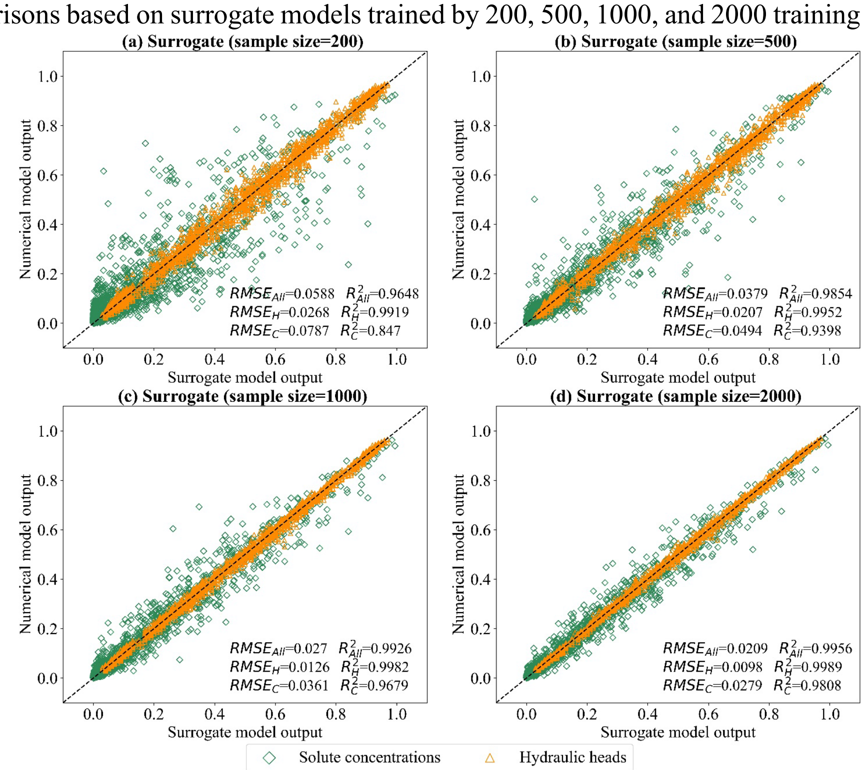

Fig.S4. Performance of FC-DNN based surrogate models for the solute concentration and hydraulic head prediction. (a~d) are pair-wise comparisons based on surrogate models trained by 200, 500, 1000, and 2000 training samples, respectively.

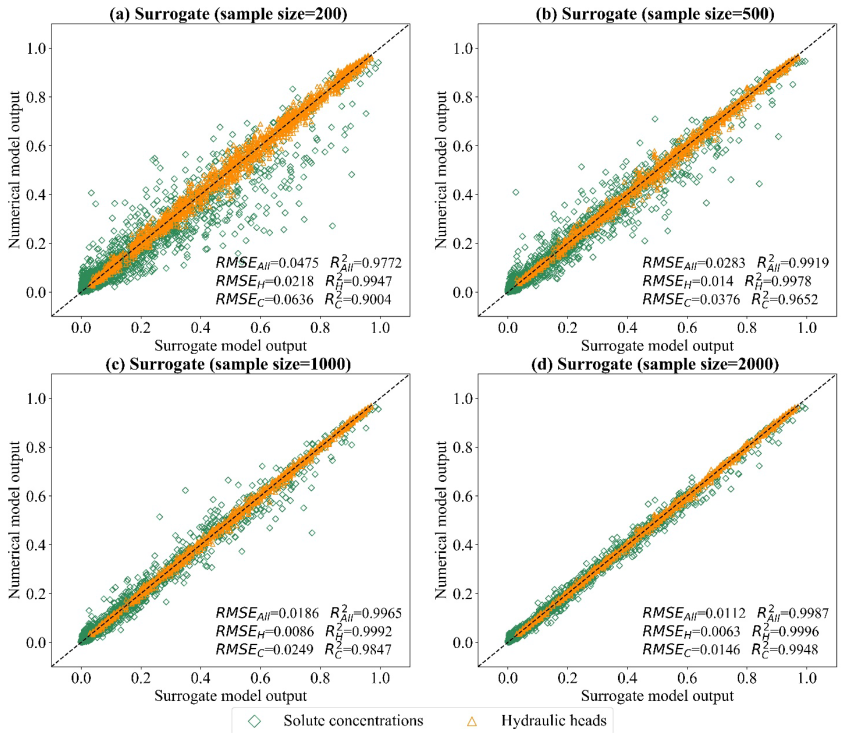

Fig.S5. Performance of LeNet CNN based surrogate models for the solute concentration and hydraulic head prediction. (a~d) are pair-wise comparisons based on surrogate models trained by 200, 500, 1000, and 2000 training samples, respectively.

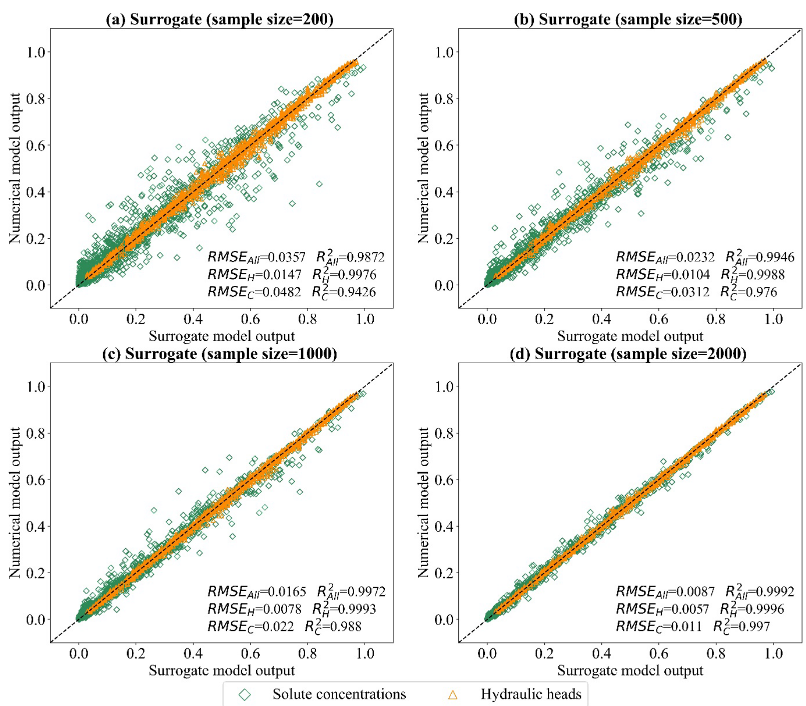

Fig.S6. Performance of ResNet CNN based surrogate models for the solute concentration and hydraulic head prediction. (a~d) are pair-wise comparisons based on surrogate models trained by 200, 500, 1000, and 2000 training samples, respectively.

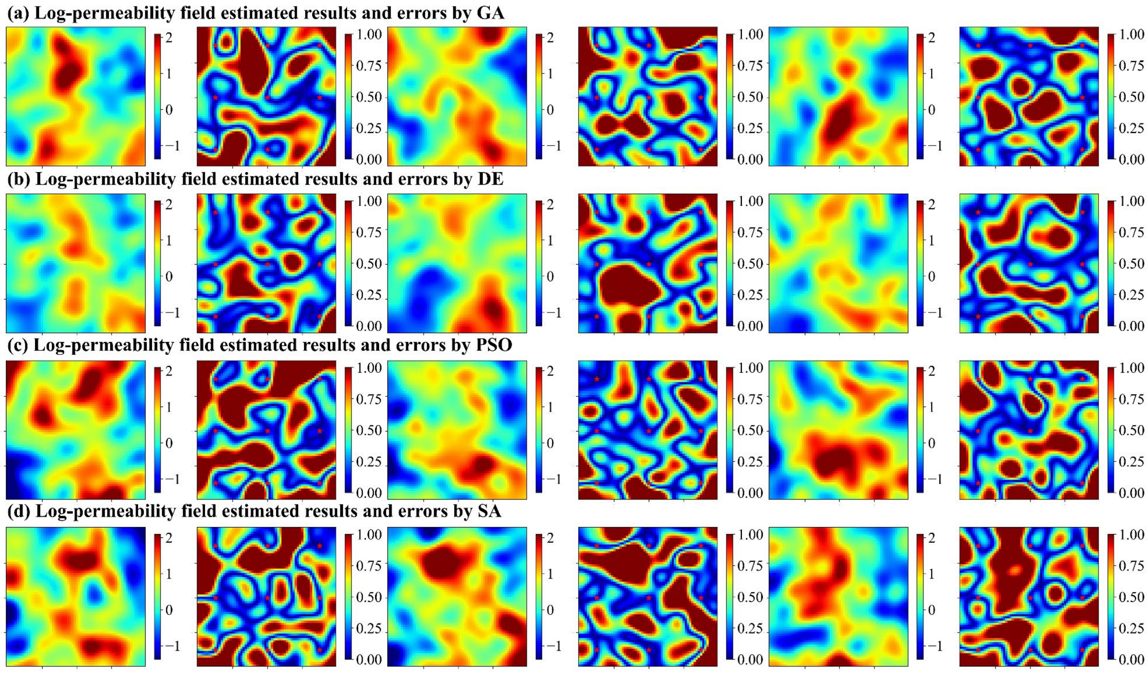

Fig.S7 Spatial distributions of log-permeability field estimation results (row 1, 3, and 5 for N_{PC}=100, 500, and 1000, respectively) and absolute errors (row 2, 4, and 6 for N_{PC}=100, 500, and 1000, respectively) for Scenario 1, achieved by four metaheuristic algorithms.

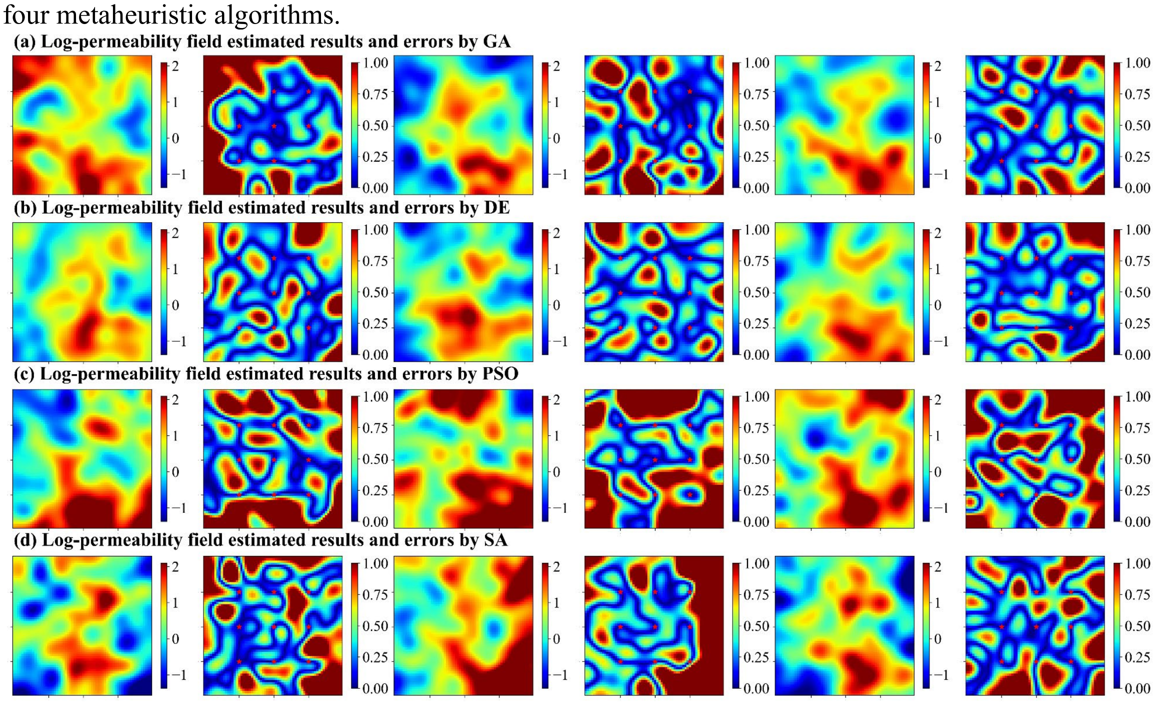

Fig.S8 Spatial distributions of log-permeability field estimation results (row 1, 3, and 5 for N_{PC}=100, 500, and 1000, respectively) and absolute errors (row 2, 4, and 6 for N_{PC}=100, 500, and 1000, respectively) for Scenario 2, achieved by four metaheuristic algorithms.

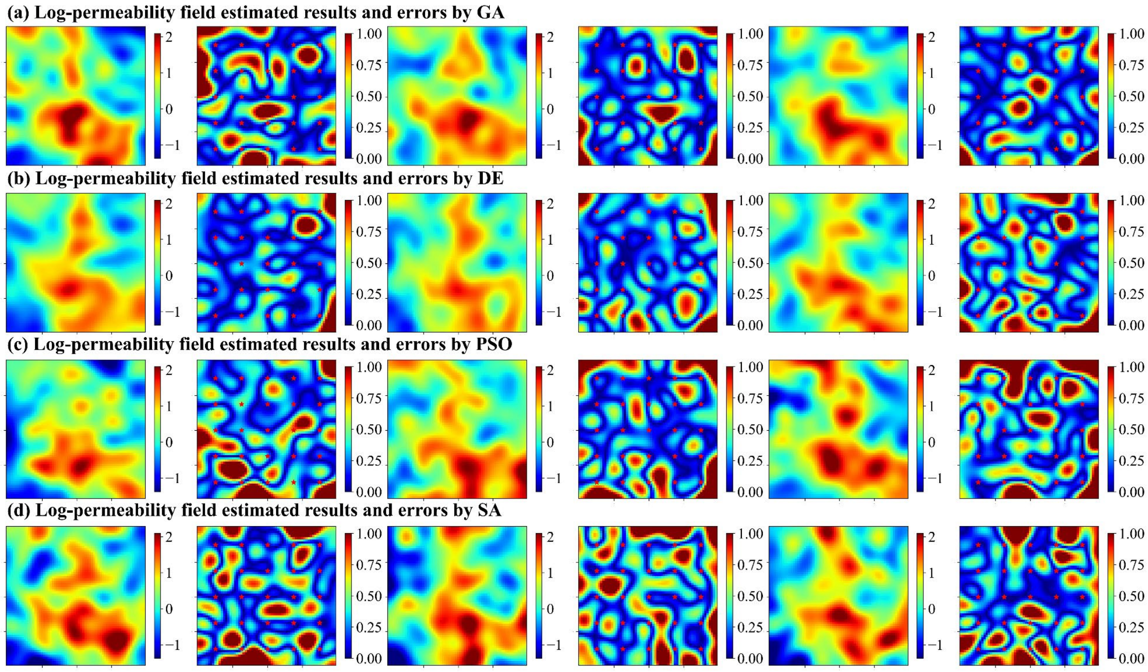

Fig.S9 Spatial distributions of log-permeability field estimation results (row 1, 3, and 5 for N_{PC}=100, 500, and 1000, respectively) and absolute errors (row 2, 4, and 6 for N_{PC}=100, 500, and 1000, respectively) for Scenario 3, achieved by four metaheuristic algorithms.

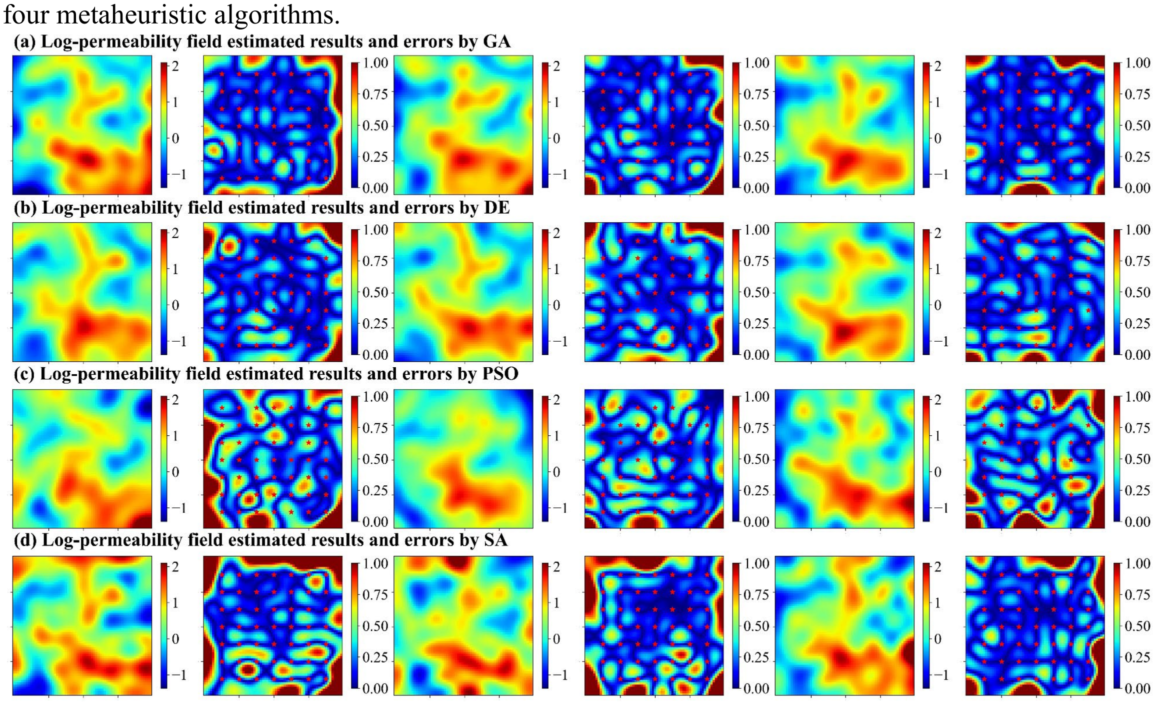

Fig.S10 Spatial distributions of log-permeability field estimation results (row 1, 3, and 5 for N_{PC}=100, 500, and 1000, respectively) and absolute errors (row 2, 4, and 6 for N_{PC}=100, 500, and 1000, respectively) for Scenario 4, achieved by four metaheuristic algorithms.

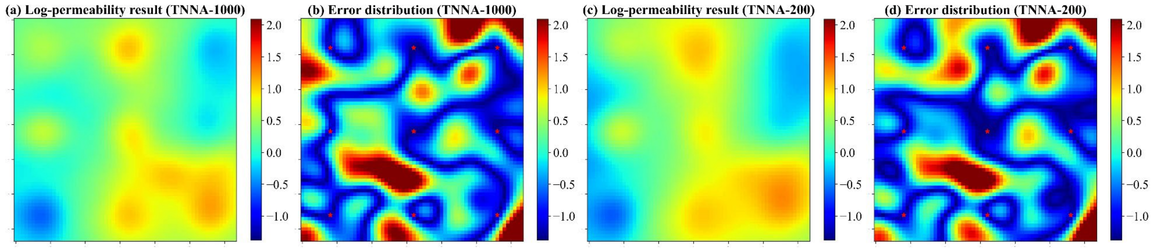

Fig.S11. Spatial distributions log-permeability field estimation results and absolute errors for Scenario 1, achieved by the TNNA inversion algorithm.

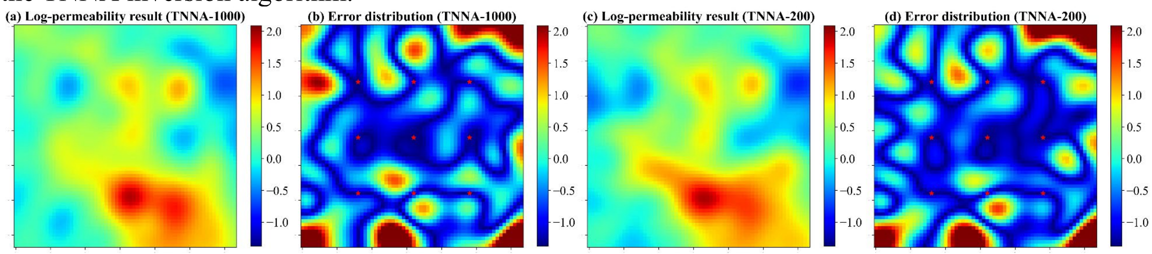

Fig.S12. Spatial distributions log-permeability field estimation results and absolute errors for Scenario 2, achieved by the TNNA inversion algorithm.

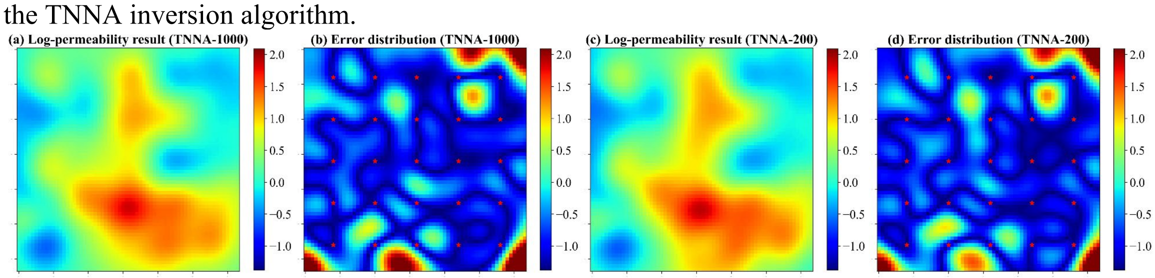

Fig.S13. Spatial distributions log-permeability field estimation results and absolute errors for Scenario 3, achieved by the TNNA inversion algorithm.

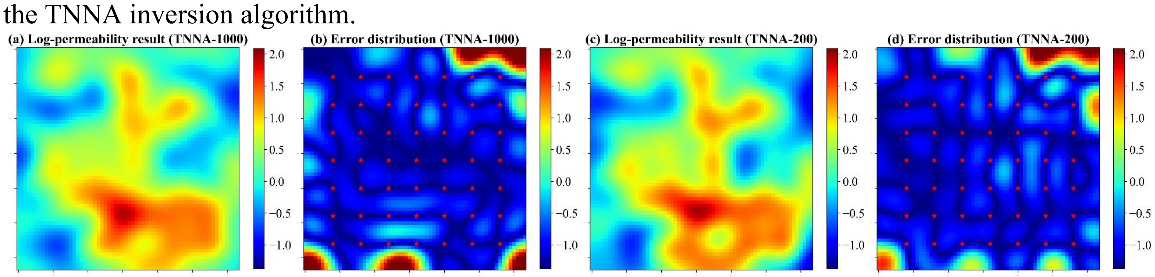

Fig.S14. Spatial distributions log-permeability field estimation results and absolute errors for Scenario 4, achieved by the TNNA inversion algorithm.

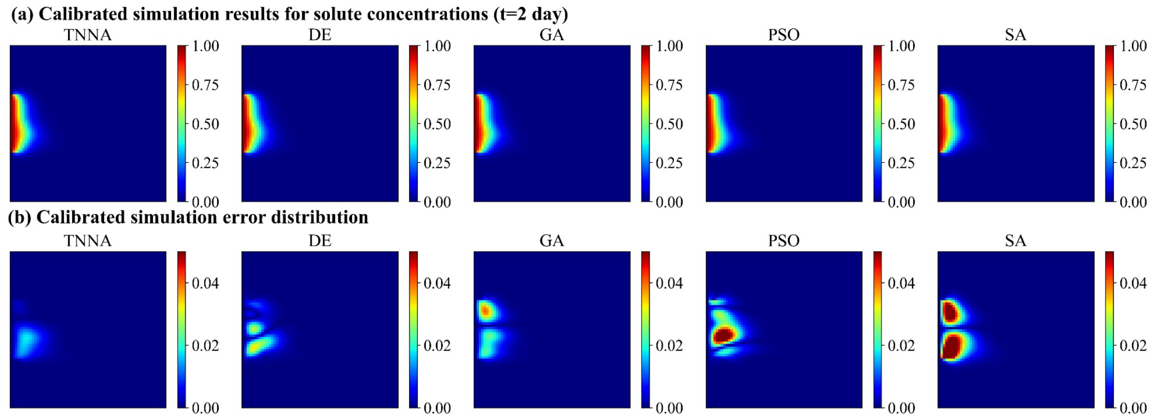

Fig.S15. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=2 day) using the TNNA algorithm and four metaheuristic algorithms.

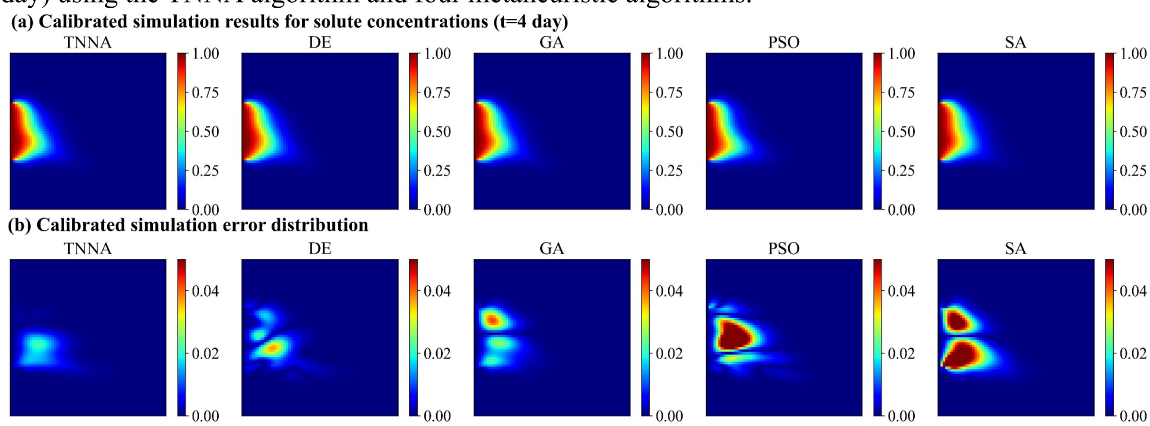

Fig.S16. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=4 day) using the TNNA algorithm and four metaheuristic algorithms.

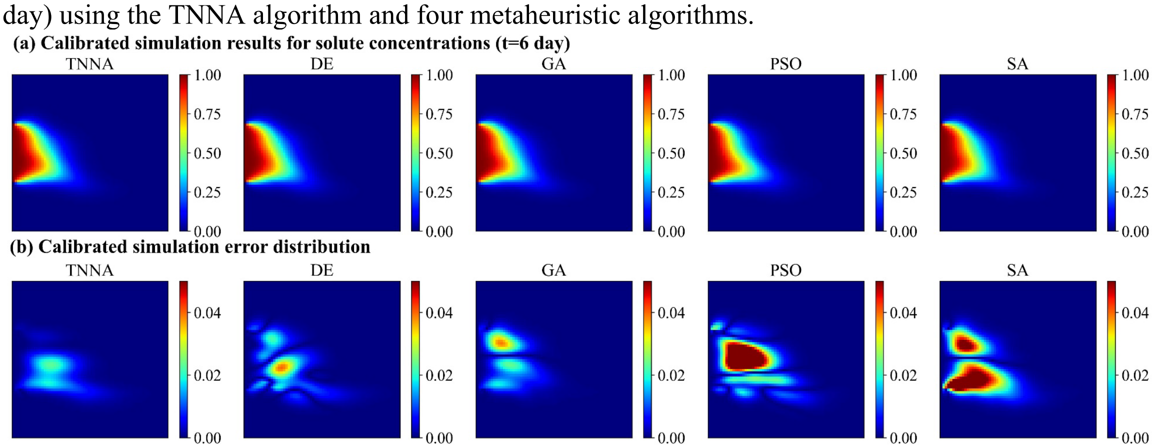

Fig.S17. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=6 day) using the TNNA algorithm and four metaheuristic algorithms.

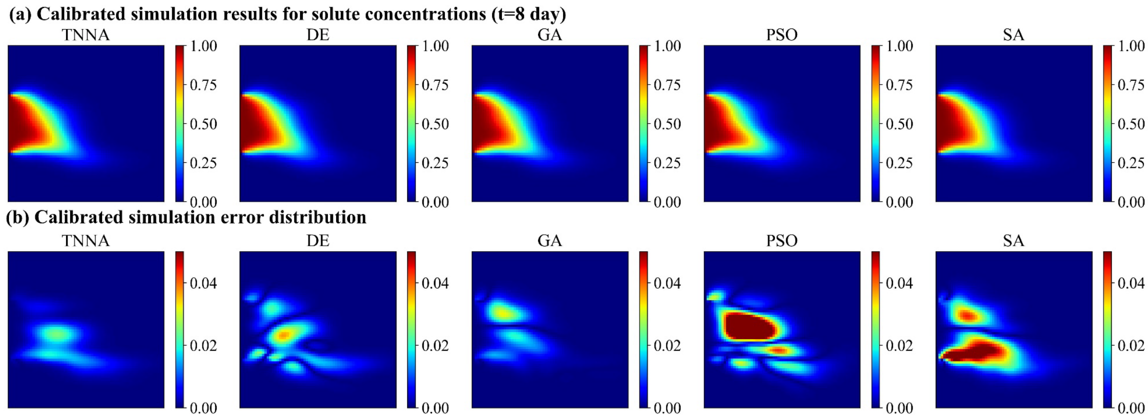

Fig.S18. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=8 day) using the TNNA algorithm and four metaheuristic algorithms.

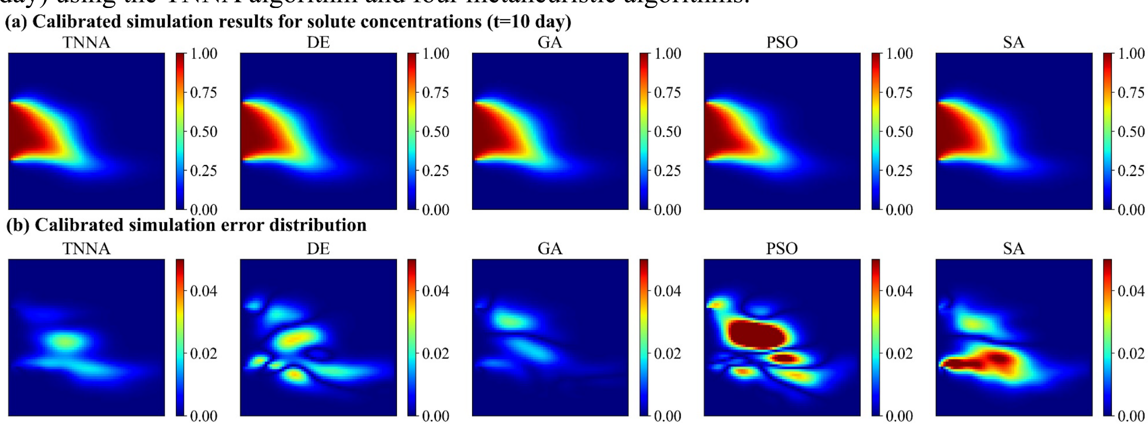

Fig.S19. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=10 day) using the TNNA algorithm and four metaheuristic algorithms.

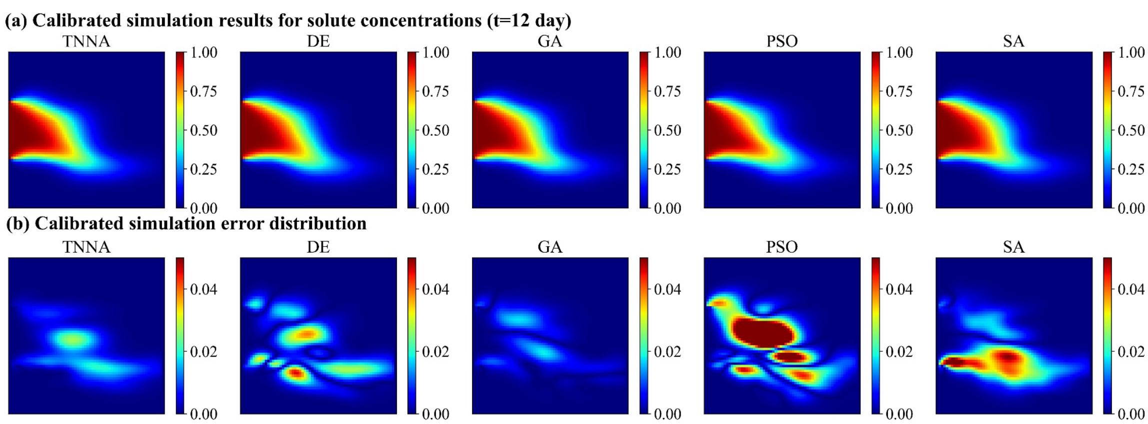

Fig.S20. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=12 day) using the TNNA algorithm and four metaheuristic algorithms.

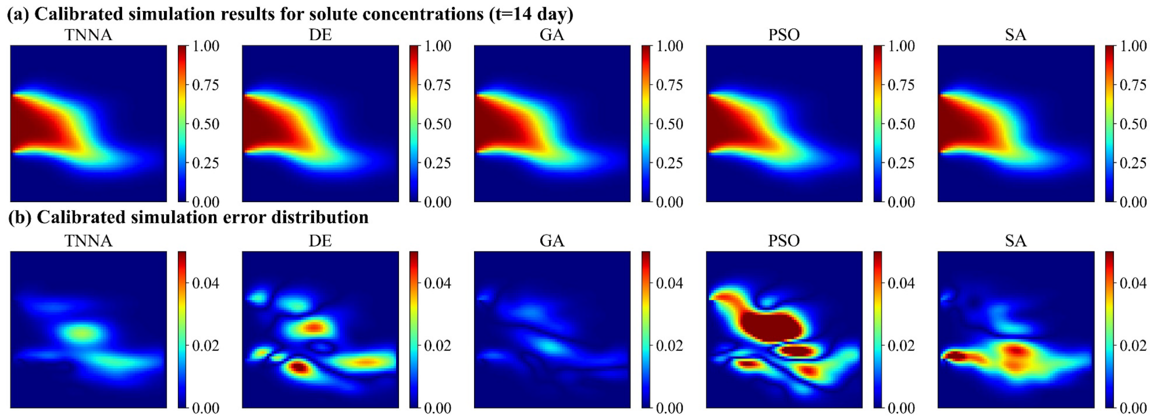

Fig.S21. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=14 day) using the TNNA algorithm and four metaheuristic algorithms.

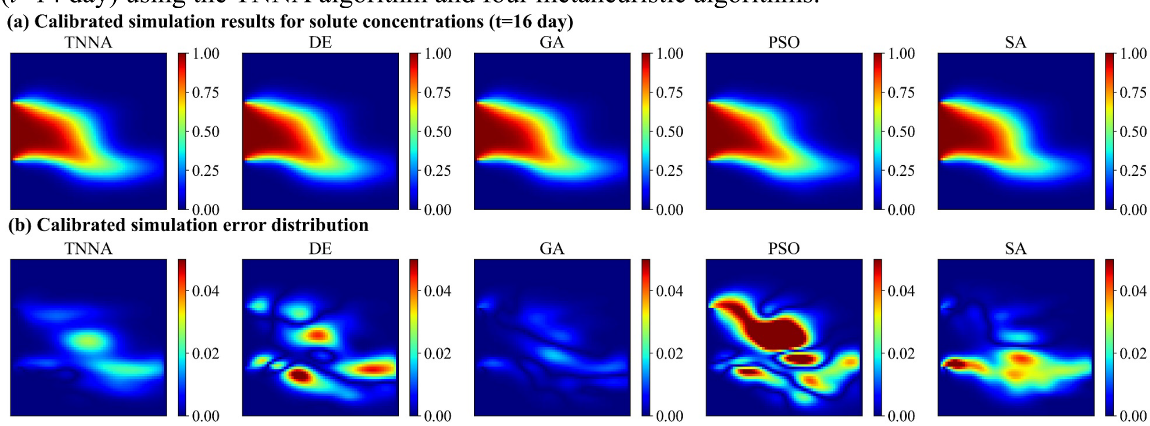

Fig.S22 Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=16 day) using the TNNA algorithm and four metaheuristic algorithms.

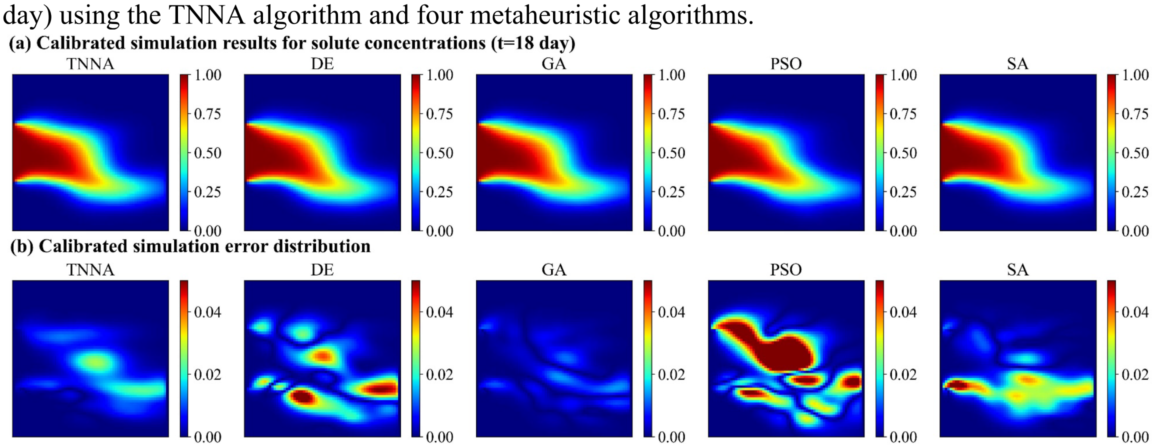

Fig.S23 Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=18 day) using the TNNA algorithm and four metaheuristic algorithms.

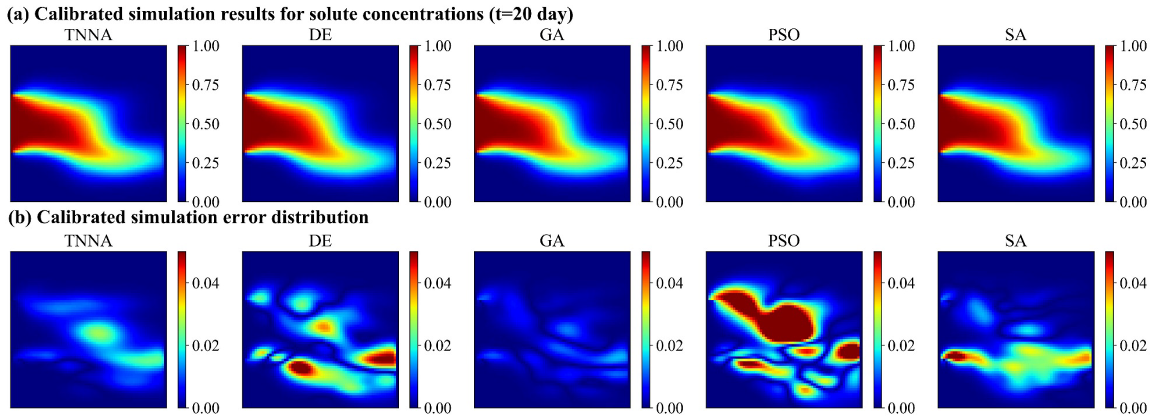

Fig.S24 Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=20 day) using the TNNA algorithm and four metaheuristic algorithms.

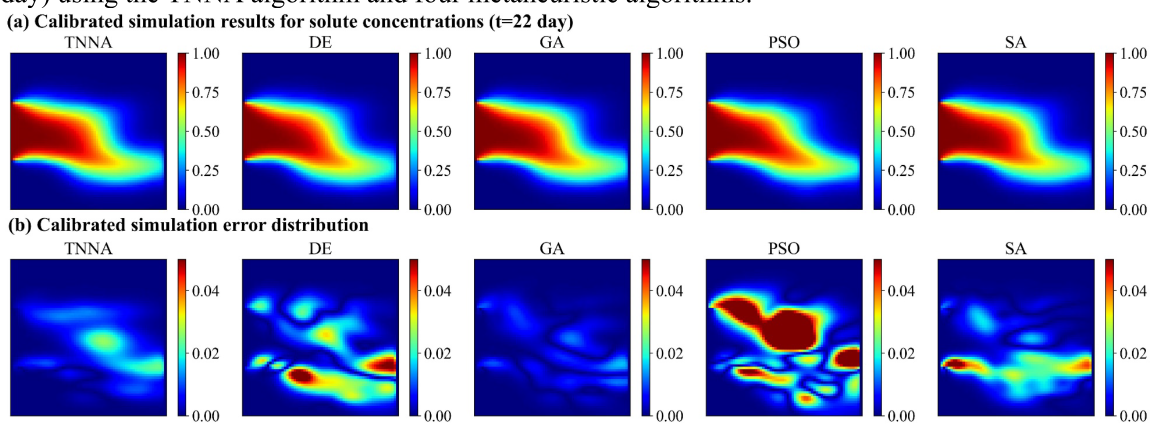

Fig.S25. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=22 day) using the TNNA algorithm and four metaheuristic algorithms.

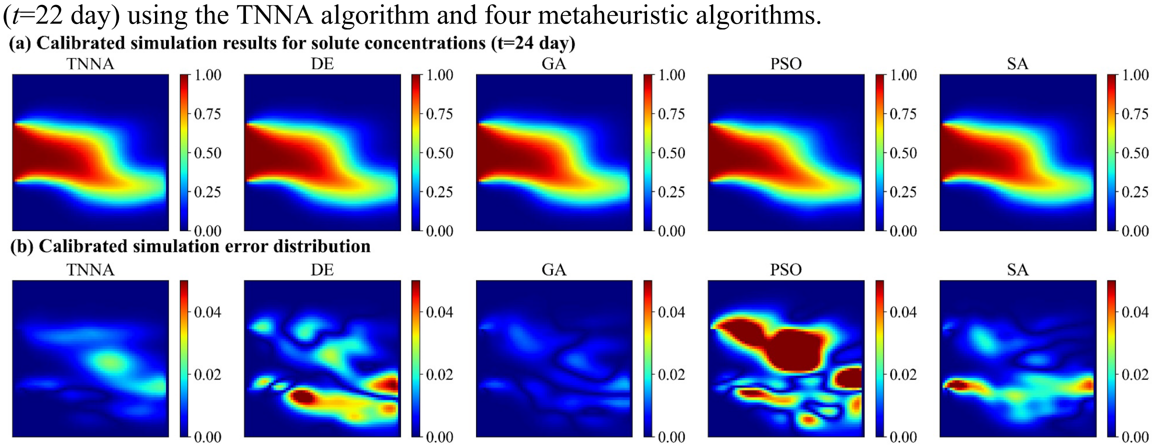

Fig.S26. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=24 day) using the TNNA algorithm and four metaheuristic algorithms.

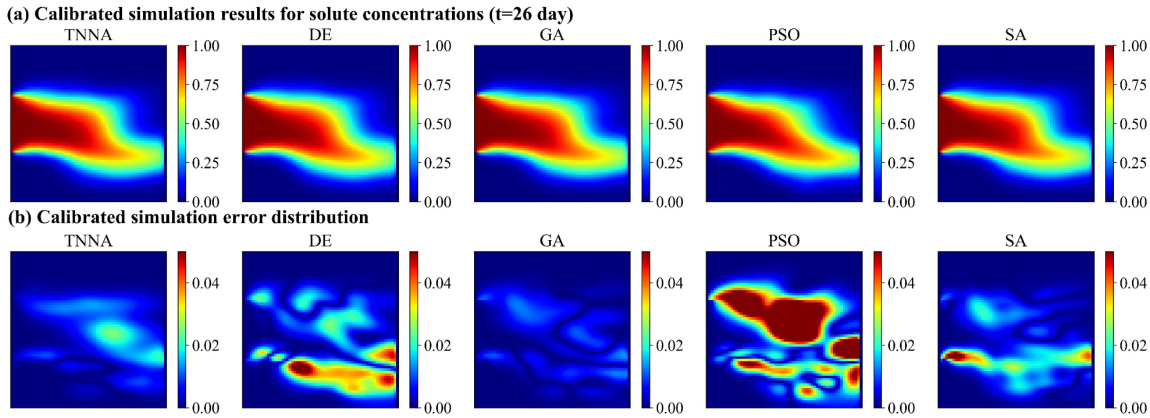

Fig.S27. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=26 day) using the TNNA algorithm and four metaheuristic algorithms.

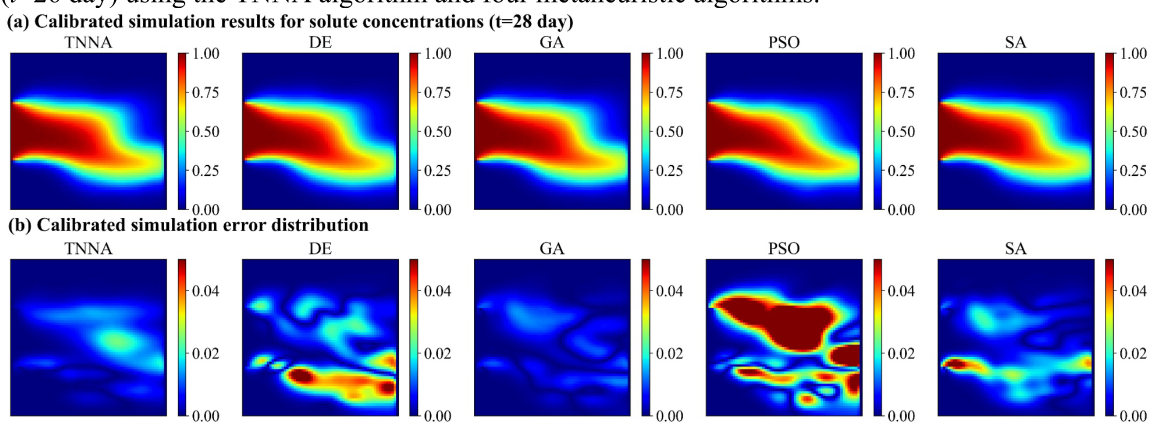

Fig.S28. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=28 day) using the TNNA algorithm and four metaheuristic algorithms.

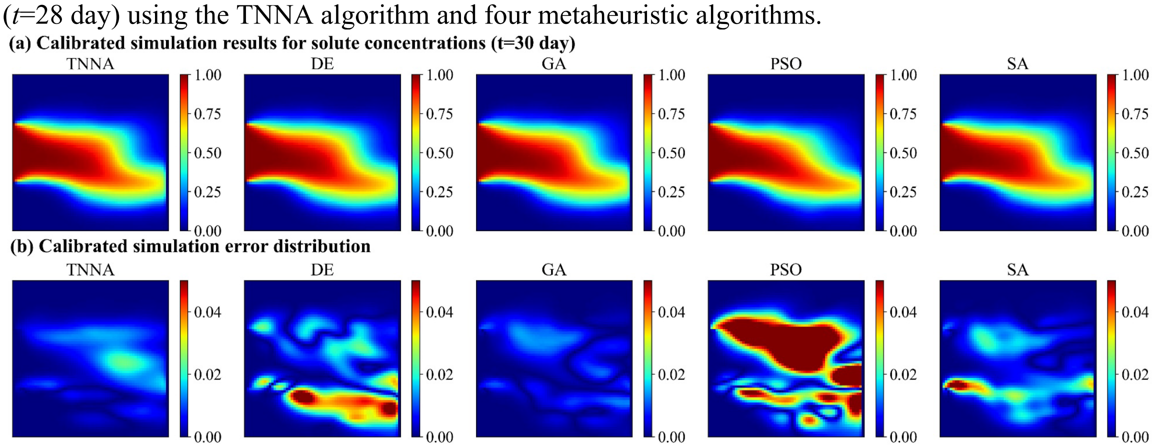

Fig.S29. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=30 day) using the TNNA algorithm and four metaheuristic algorithms.

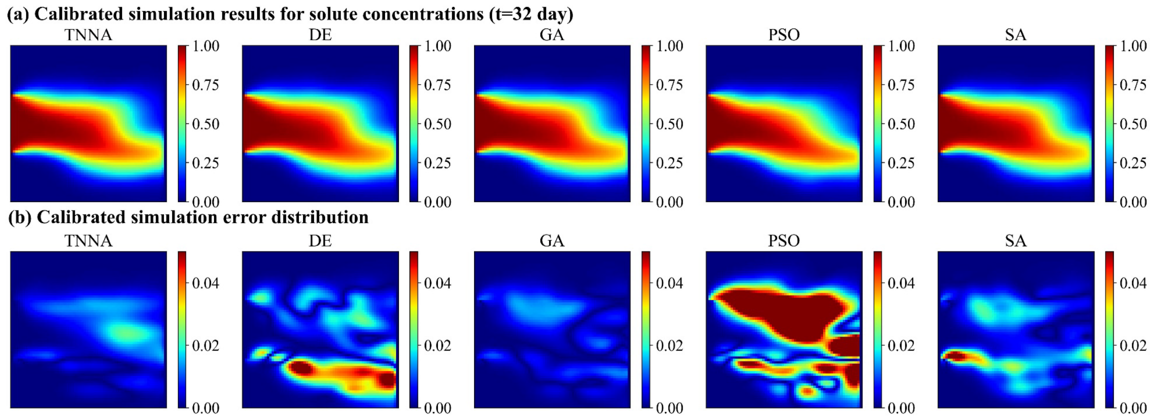

Fig.S30. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=32 day) using the TNNA algorithm and four metaheuristic algorithms.

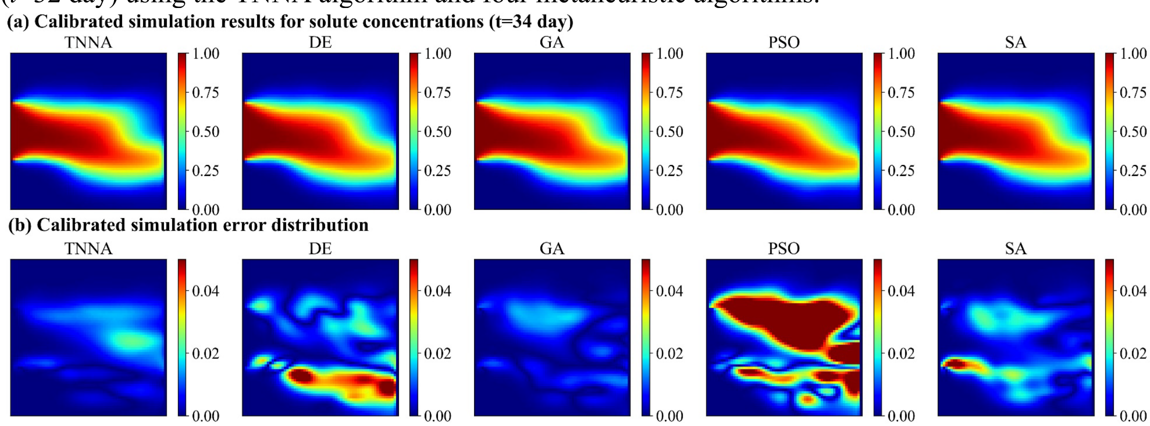

Fig.S31. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=34 day) using the TNNA algorithm and four metaheuristic algorithms.

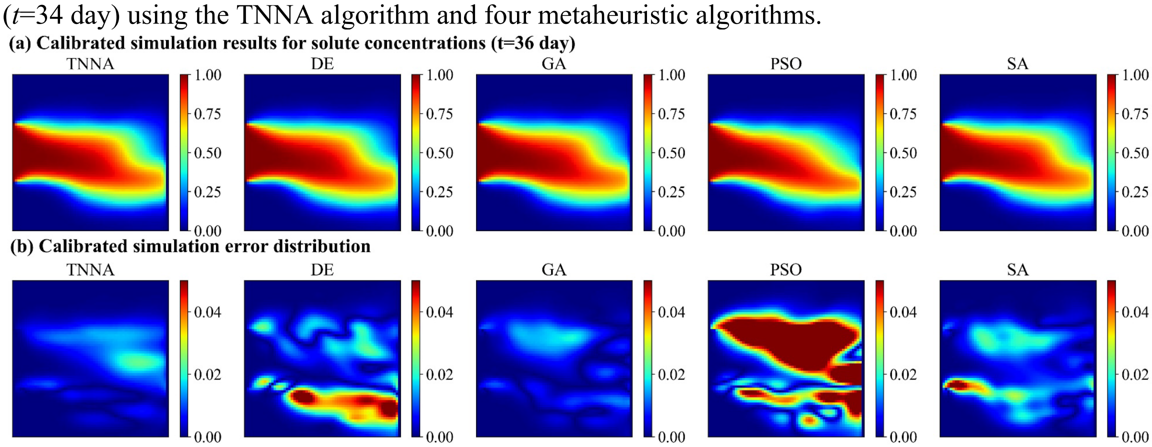

Fig.S32. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=36 day) using the TNNA algorithm and four metaheuristic algorithms.

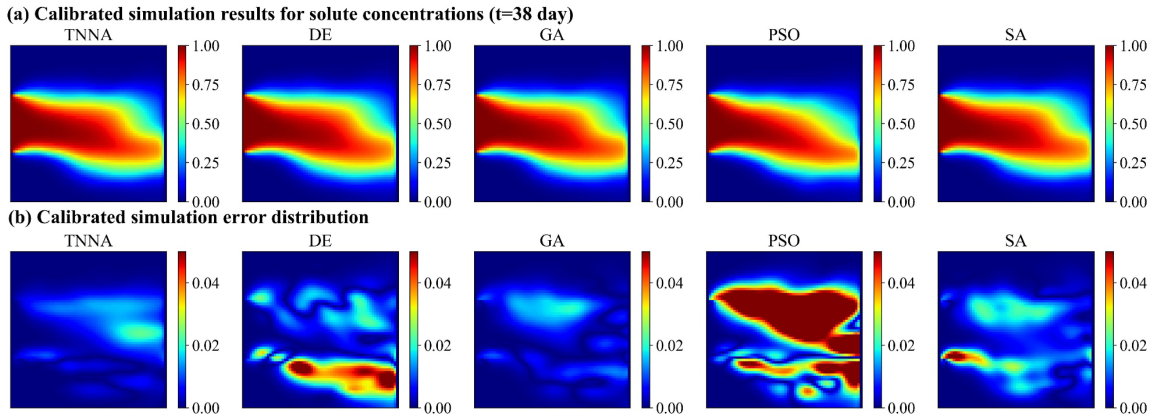

Fig.S33. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=38 day) using the TNNA algorithm and four metaheuristic algorithms.

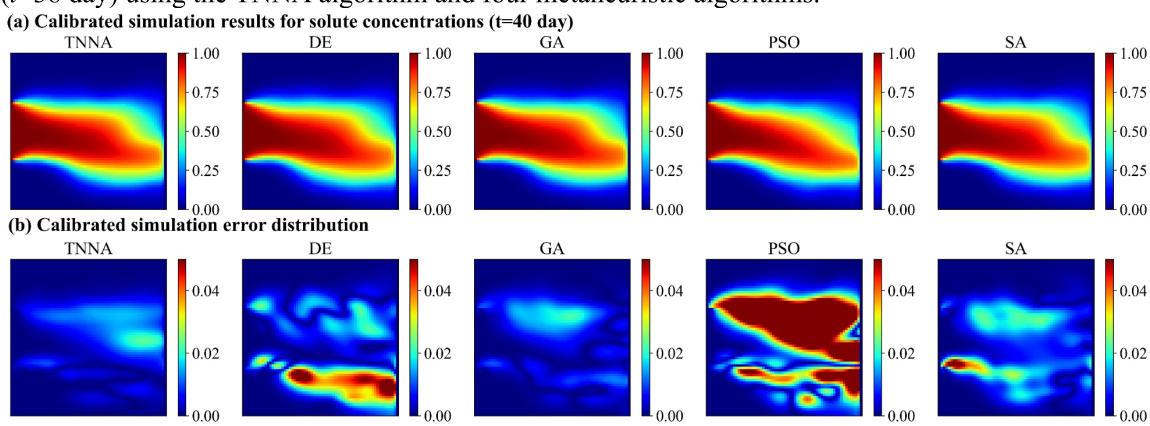

Fig.S34. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=40 day) using the TNNA algorithm and four metaheuristic algorithms.

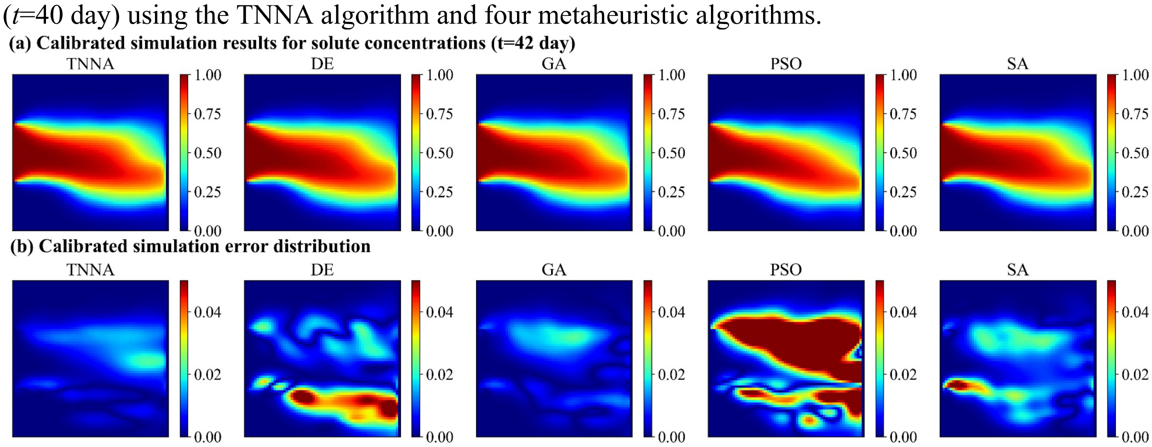

Fig.S35. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=42 day) using the TNNA algorithm and four metaheuristic algorithms.

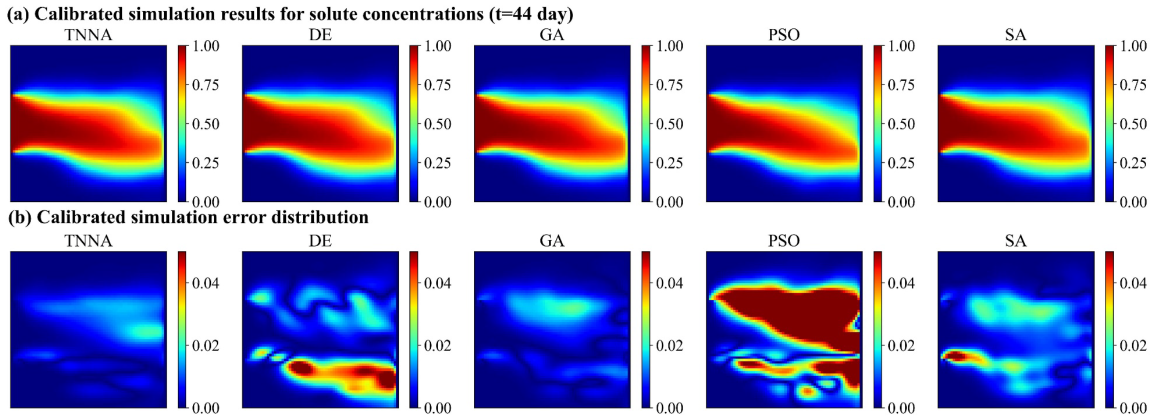

Fig.S36. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=44 day) using the TNNA algorithm and four metaheuristic algorithms.

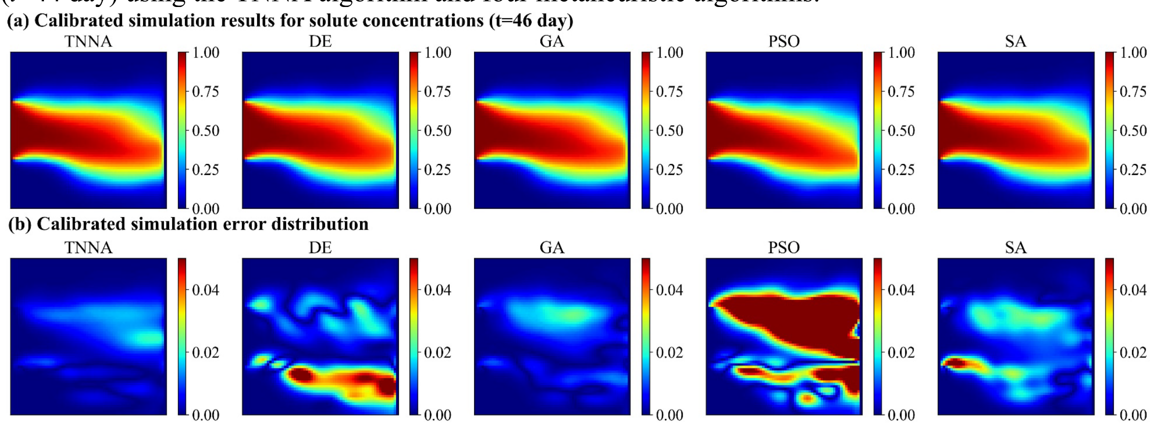

Fig.S37. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=46 day) using the TNNA algorithm and four metaheuristic algorithms.

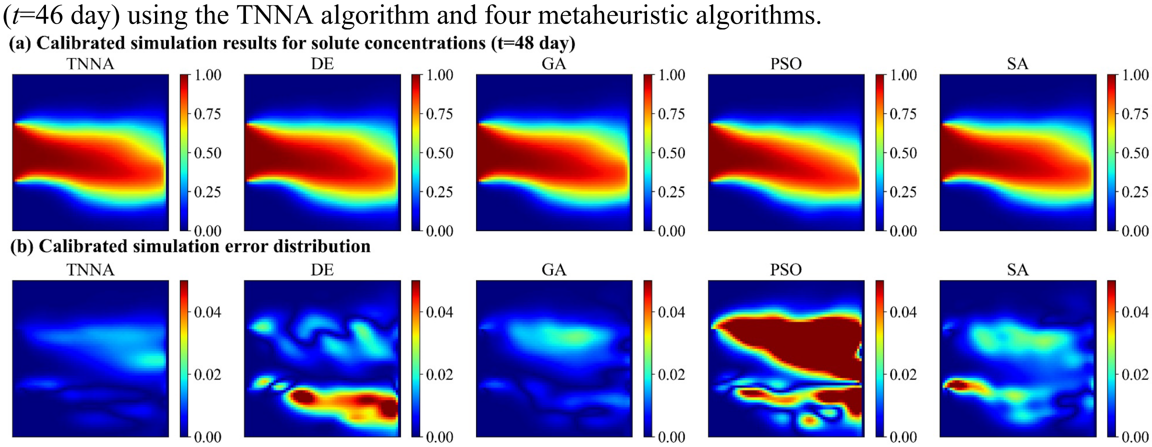

Fig.S38. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=48 day) using the TNNA algorithm and four metaheuristic algorithms.

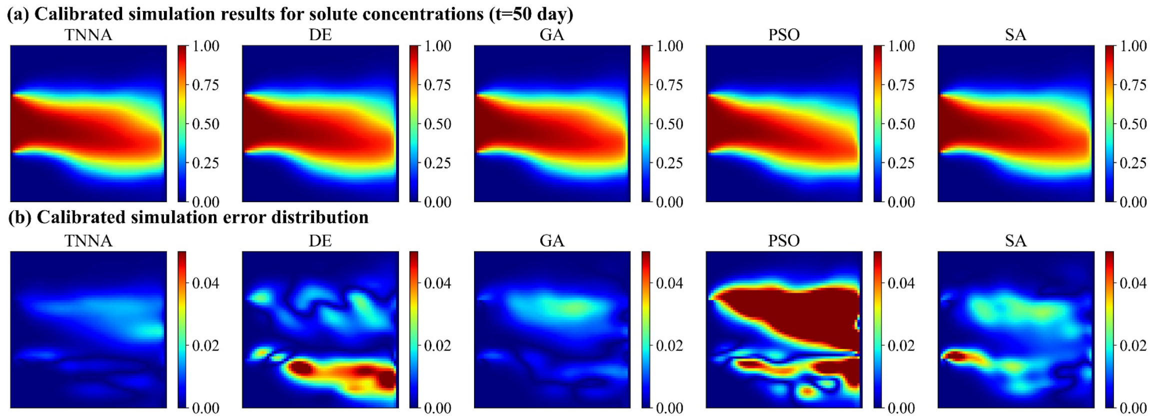

Fig.S39. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=50 day) using the TNNA algorithm and four metaheuristic algorithms.

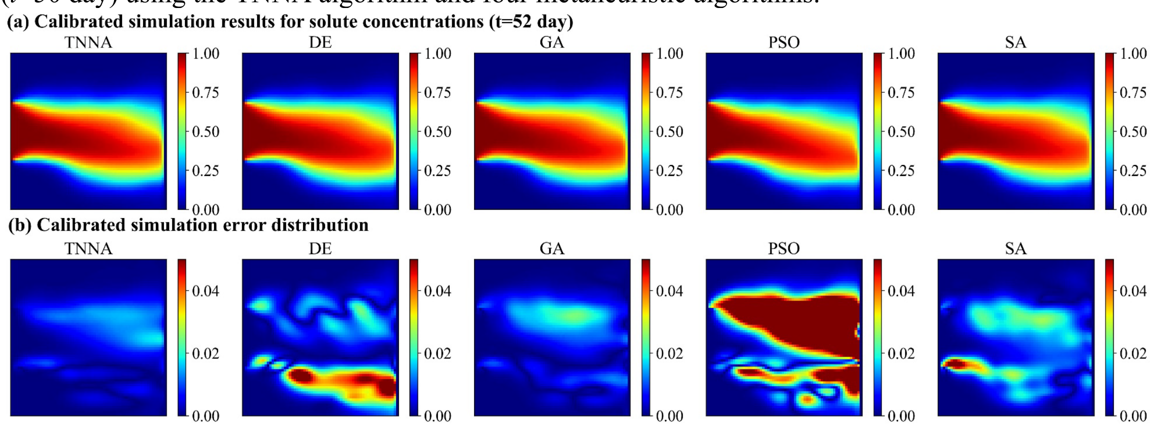

Fig.S40. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=52 day) using the TNNA algorithm and four metaheuristic algorithms.

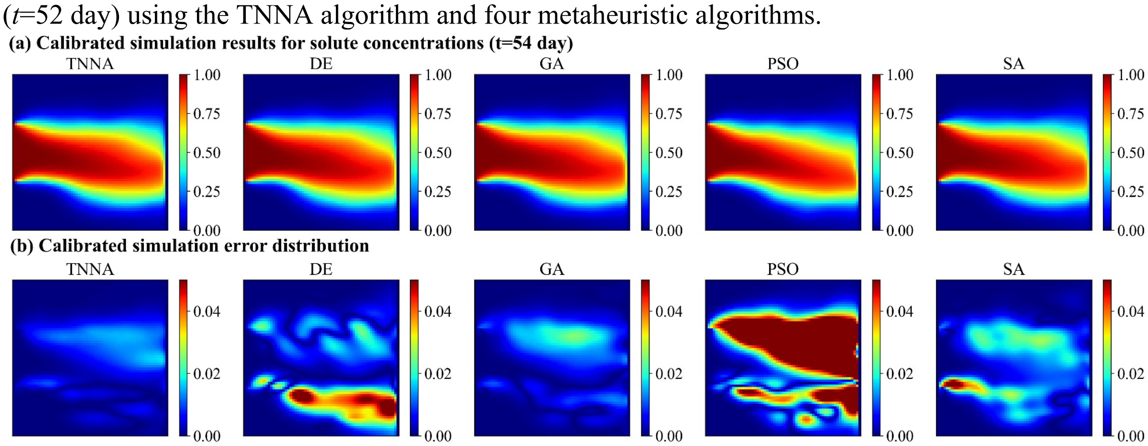

Fig.S41. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (*t*=54 day) using the TNNA algorithm and four metaheuristic algorithms.

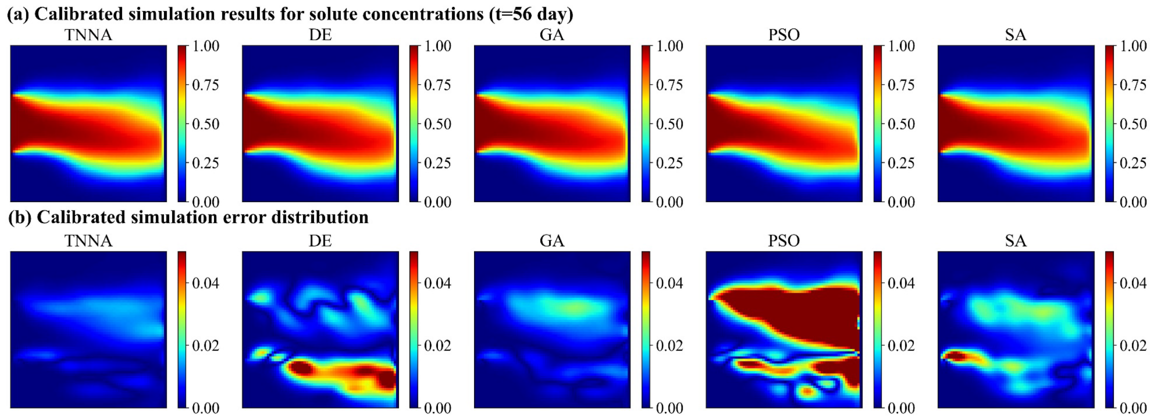

Fig.S42. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=56 day) using the TNNA algorithm and four metaheuristic algorithms.

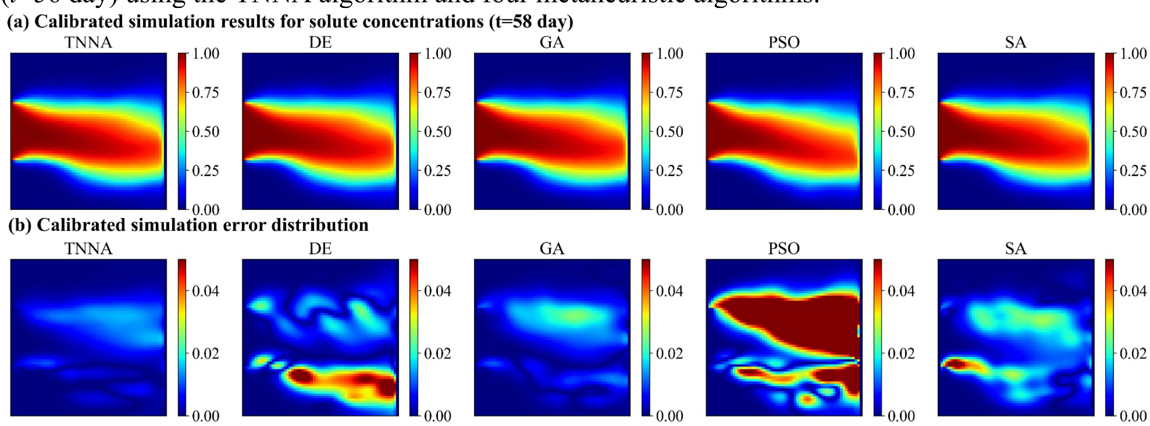

Fig.S43. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=58 day) using the TNNA algorithm and four metaheuristic algorithms.

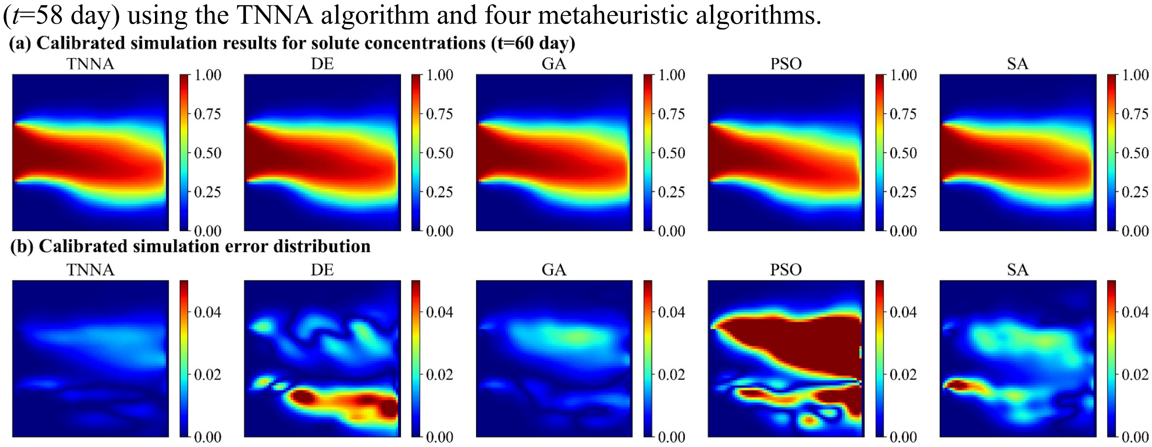

Fig.S44. Spatial distributions of calibrated numerical simulation results and absolute errors for solute concentrations (t=60 day) using the TNNA algorithm and four metaheuristic algorithms.