Responses to the Evaluation of Reviewer #2

General Comments

This manuscript presents a data-driven framework for correcting structural errors in groundwater contaminant transport models. Gaussian process regression is used to represent systematic model discrepancy, while multiple physical constraints are introduced through a combined likelihood function. An adaptive weighting strategy and a stopping criterion are proposed to balance the observational likelihood and the likelihood terms associated with different physical constraints. The framework is tested using a synthetic three-dimensional PCE reactive-transport model with simplified preferential pathways and a Cd–phosphate sand-column experiment in which the cotransport mechanism is omitted from the physical model. The results show that using multiple physical constraints can improve validation-period predictions, reduce mass-balance errors, and narrow posterior parameter and prediction intervals. The topic is relevant to groundwater modeling and uncertainty analysis. Model structural error is an important source of predictive uncertainty but is often treated less explicitly than parameter uncertainty. The main contribution of the study is the extension of a GPR-based structural-error correction method from a single physical constraint to multiple constraints, together with an adaptive weighting procedure. The two case studies represent different types of model inadequacy and provide a useful first demonstration of the method. The manuscript is generally well organized, and the results are encouraging. However, several methodological details and interpretations should be clarified before publication. These revisions do not require changing the main framework, but they would improve reproducibility and strengthen the conclusions.

Response: We thank the reviewer for the positive evaluation and insightful comments to this manuscript, and we have carefully addressed these comments in below responses and revised the manuscript.

Specific Comments

Comment 1: Please clarify how adaptive weighting is implemented within the MCMC procedure.

The dynamic weighting strategy is central to the manuscript, but it is not fully clear whether the weights are updated independently for each Markov chain or shared among all chains. Please also explain how samples generated before and after weight stabilization are used in the posterior analysis. A practical approach would be to treat weight adjustment as an adaptive or burn-in stage, fix the weights after the stopping criterion is reached, and use only the subsequent samples for posterior inference. It would also be helpful to report the iteration at which the weights stabilize, together with standard MCMC diagnostics such as acceptance rate, R^, and effective sample size.

Response: Thanks for pointing out this important problem. We apologize for not clearly describing the implementation of the dynamic weight-updating strategy during MCMC sampling. It should be noted that the dynamic weight-updating strategy is

applied only during the burn-in period. During this period, the weights of the sub-likelihood functions are updated separately for each Markov chain based on its own evolution. Once the stopping criterion for weight updating is satisfied, the weights are fixed, and only the subsequent samples generated using these fixed weights are used for posterior inference. Therefore, the dynamic updating of the sub-likelihood weights does not alter the target posterior distribution used for posterior inference.

Following this suggestion, we made three main revisions. (1) We explicitly clarified in the main text that, during the burn-in period, the sub-likelihood weights are dynamically updated separately for each Markov chain based on its own evolution (Line 275). (2) We added annotations to Figures 10 and 11 indicating the iterations at which the weight updating was terminated. (3) For both case studies, we evaluated the evolution of the Gelman–Rubin statistics for all parameters when multiple physical constraints were coupled (Figures S1 and S2), and used these statistics to assess MCMC convergence.

Line 275: Since multiple Markov chains are employed, the weights of the sub-likelihood functions are updated separately for each chain during the burn-in period based on its own evolution. Once the stopping criterion is satisfied, a common set of fixed sub-likelihood weights is used for subsequent sampling across all chains.

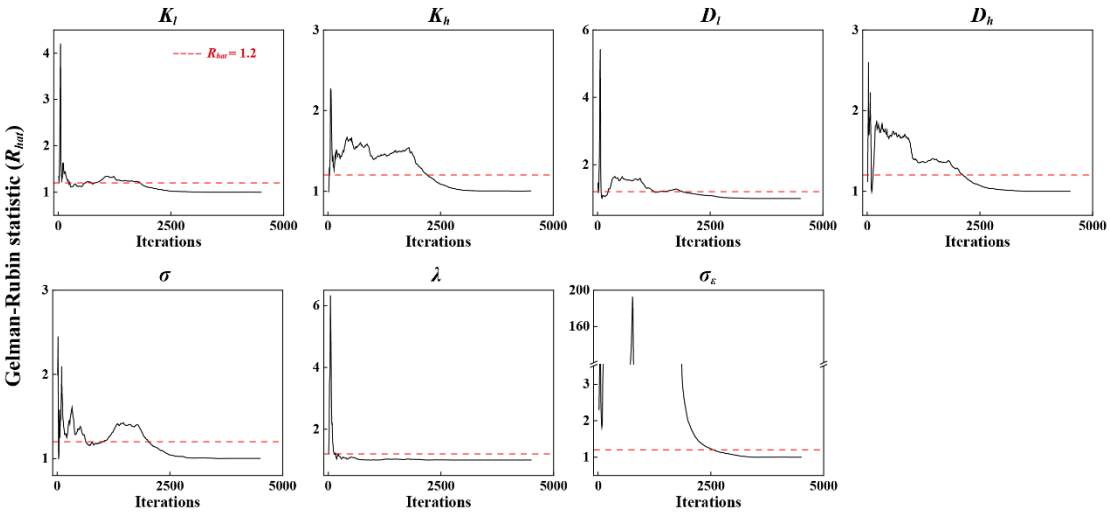

Figure S1. Gelman-Rubin diagnostics of Markov chains for the synthetic groundwater PCE reactive transport simulation under the multiple-constraint scenario.

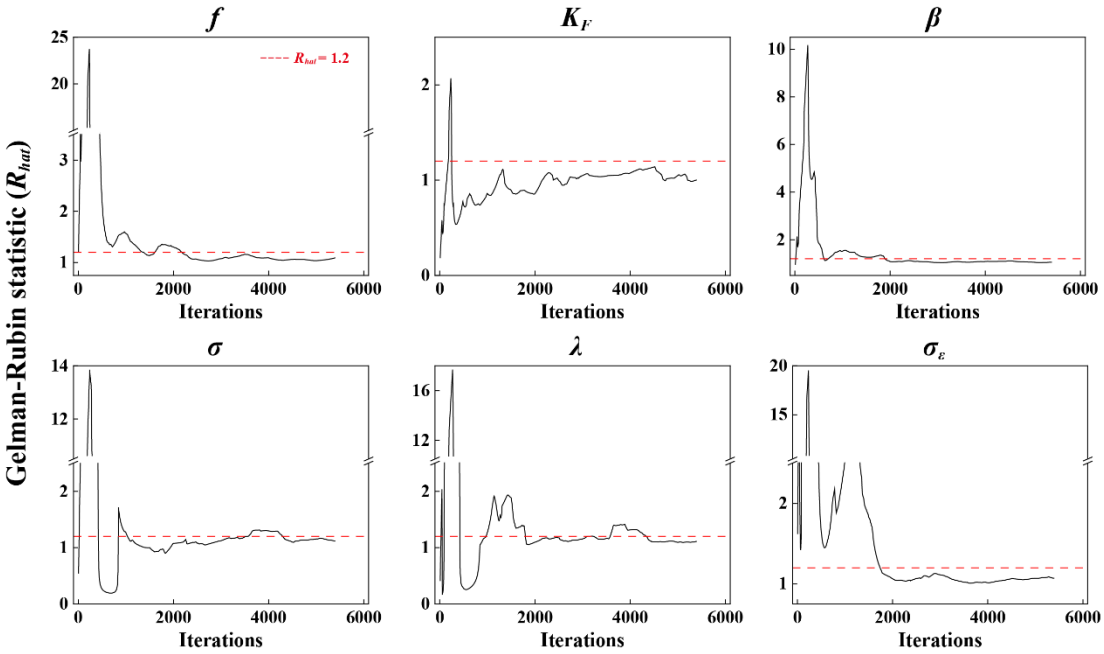

Figure S2. Gelman-Rubin diagnostics of Markov chains for the sand column Cd-phosphate cotransport simulation under the multiple-constraint scenario.

Comment 2: The formulation and scaling of the physical constraints need further explanation.

In Eqs. (9)-(10), the constrained variable includes the physical-model prediction, GPR structural error, and measurement error. Since measurement error is not expected to satisfy mass conservation, the authors may consider applying the physical constraints to the latent corrected state, f(θ)+b(x,ϕ), rather than to the observation noise. Alternatively, the current implementation should be justified more clearly. Please also explain why the same constraint variance of 0.2 is used for constraints with different units and numerical scales. A short sensitivity test using one or two alternative values, or a description of how the constraint residuals were normalized, would be sufficient.

Response: We thank the reviewer for this important suggestion. In the framework proposed in this study, the corrected groundwater model predictions consist of three components: the physical model predictions, structural error, and measurement error. Since measurement error remains an important component of the corrected predictions, we consider it appropriate to impose the physical constraints on the sum of these three components.

In addition, following this suggestion, we conducted a sensitivity analysis of the constraint variance, σ_c^2. For both case studies, we evaluated prediction performance and mass conservation errors under different constraint variances, i.e., σ_c^2 = 0.01, 0.1, 0.2, and 0.5 (Figures S3 and S4).

The results show that when σ_c^2 is relatively small or large, such as σ_c^2 = 0.01 or 0.5, the agreement between the predictions and observations deteriorates, resulting in relatively large RMSE values and mass conservation errors. When σ_c^2 = 0.2, both RMSE and mass conservation error are minimized. Therefore, σ_c^2 = 0.2 was adopted in

this study, consistent with the findings of our previous study (Tian et al., 2026).

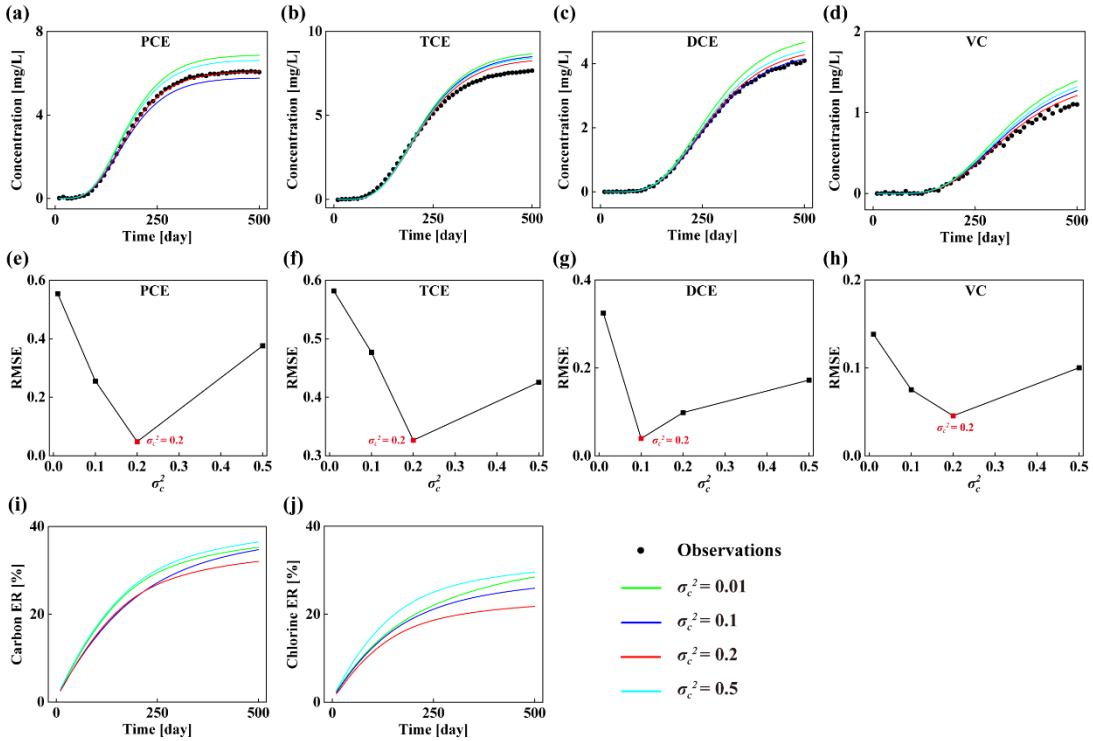

Figure S3. Predictions of synthetic groundwater PCE reactive transport simulation under different constraint variances (σ_c^2).

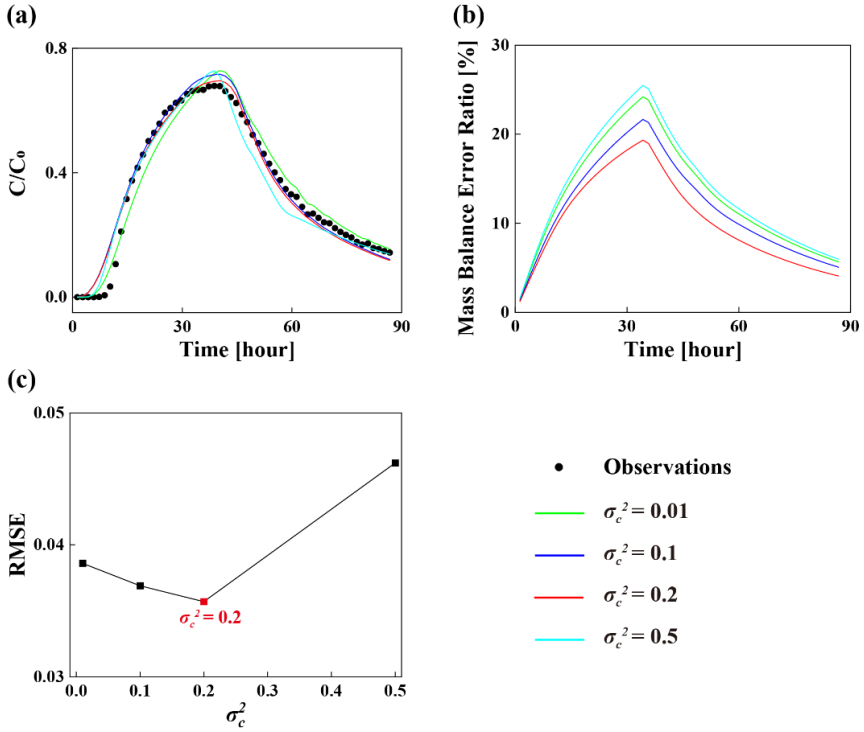

Figure S4. Predictions of sand column Cd-phosphate cotransport simulation under different constraint variances (σ_c^2).

Comment 3: Please provide a more complete assessment of predictive uncertainty. The

manuscript mainly interprets narrower 95% prediction intervals as reduced uncertainty. Narrower intervals are useful only when they retain adequate coverage. Please report, at least for the validation period, the coverage rate and average width of the 95% prediction intervals. An interval score could also be included if readily available. For the synthetic case, it would be useful to indicate whether the known true parameter values fall within the corresponding posterior credible intervals. These additions would distinguish improved uncertainty quantification from simple interval contraction.

Response: We thank the reviewer for this important suggestion. Following this suggestion, we have added the coverage rates and average widths of the 95% confidence intervals for the simulated results of both case studies (Tables S3 and S4) and included the corresponding discussion in the revised manuscript (Lines 574 and 615).

Line 574: In contrast, simultaneously coupling multiple physical constraints markedly reduces the prediction intervals for all chlorinated hydrocarbons (Figs. 8i-8l), and the corresponding average widths of the 95% prediction intervals are substantially smaller (Table S3 in Supporting Information), indicating reduced predictive uncertainty.

Line 615: In contrast, when the Cd concentration bound constraint is additionally coupled, all posterior predictions remain within the physically admissible range, and the corresponding 95% prediction intervals exhibit substantially smaller average widths (Table S4 in Supporting Information), indicating a marked reduction in predictive uncertainty.

In addition, we consider the reviewer's suggestion to verify whether the true parameter values fall within the corresponding posterior intervals to be highly valuable. However, the present study adopts an external correction approach for model structural error rather than an internal correction approach. Therefore, the identified posterior distributions of the parameters should be compared with the true parameter values of the simplified model.

In the first case study, the preferential flow regions differ in shape between the true and simplified models. Consequently, the parameter values of the true model (hydraulic conductivity and dispersivity) cannot be regarded as the true values of the corresponding parameters in the simplified model, and the true parameter values of the simplified model are therefore unavailable. The second case study is a real case based on a laboratory sand column experiment, in which the physical parameters to be identified (the proportion coefficient of equilibrium sorption f, the Freundlich isotherm coefficient k_d and the isotherm shape parameter β) cannot be directly measured. In addition, the GPR hyperparameters in both case studies are statistical parameters used to characterize the structural error and lack direct physical meaning. Therefore, their posterior distributions cannot be validated against true values.

It is worth noting that, in the second case study, if hydraulic conductivity were selected as an unknown parameter, its experimentally measured value could be used as the true value to validate the posterior distribution of the identified parameter.

Table S3. Evaluation of predictive uncertainty for the sand column Cd-phosphate cotransport model under the three scenarios.

Constraints	Calibration/ Validation	Species	Coverage rate of 95% PI	Average width of 95% PI
Without physical constraint	Calibration	PCE	100.00%	4.8549
		TCE	100.00%	3.3433
		DCE	100.00%	1.7583
		VC	100.00%	0.8372
	Validation	PCE	100.00%	4.6832
		TCE	100.00%	4.8017
		DCE	100.00%	3.3493
		VC	100.00%	1.0976
With single physical constraint	Calibration	PCE	100.00%	4.0288
		TCE	100.00%	2.2447
		DCE	100.00%	0.682
		VC	100.00%	0.2225
	Validation	PCE	100.00%	4.1261
		TCE	100.00%	3.7011
		DCE	100.00%	3.0829
		VC	100.00%	1.0726
With multiple physical constraints	Calibration	PCE	100.00%	1.1897
		TCE	100.00%	1.1904
		DCE	100.00%	1.1828
		VC	100.00%	1.0828
	Validation	PCE	100.00%	1.2039
		TCE	100.00%	1.2668
		DCE	100.00%	1.2304
		VC	100.00%	1.0867

Table S4. Evaluation of predictive uncertainty for the synthetic groundwater PCE reactive transport model under the three scenarios.

Constraints	Calibration/ Validation	Coverage rate of 95% PI	Average width of 95% PI
Without physical constraint	Calibration	100.00%	0.4517
	Validation	90.00%	0.2395
With single physical constraint	Calibration	100.00%	0.3839
	Validation	100.00%	0.2332
With multiple physical constraints	Calibration	86.67%	0.1260
	Validation	100.00%	0.1195

Comment 4: Several case-study details should be checked and clarified.
The measured bulk density of the sand column is reported as 1.69 g cm^{-3}, whereas Table 2 gives 1.24 g cm^{-3}. Please confirm which value was used in the model. The upper Cd concentration bound of 0.75 mg cm−3 also appears inconsistent with the stated influent concentration of 0.1 mM and may represent a normalized concentration; the unit and definition should therefore be verified. In addition, the exponential and gamma priors in Tables 3 and 4 should be defined by their distribution parameters, not only by ranges.

Response: We thank the reviewer for pointing out this important issue. We apologize for the confusion caused by the inconsistent parameter descriptions. After carefully rechecking the model settings, we confirmed that the bulk density of the sand column used in the model is 1.24g/cm. We have corrected the corresponding parameter description in the revised manuscript (Line 352).

Regarding the upper-bound constraint on Cd concentration, we carefully rechecked the model settings and computational codes and confirmed that both the definition and implementation of this constraint are correct. The only error was the y-axis label in Figure 9. Specifically, Figure 9 presents the normalized Cd concentration, C/C_0, rather than the absolute Cd concentration, where C/C_0 represents the ratio of the Cd effluent concentration to the initial injected concentration. We have therefore corrected the y-axis label in Figure 9 and the corresponding description of the physical constraint (Line 453). In addition, it should be clarified that the upper-bound constraint on Cd concentration was not directly determined from the initial injected concentration. Instead, it was specified based on the maximum Cd concentration observed in the experiment to constrain the model predictions within a physically reasonable range.

Following the reviewer's suggestion, we have also added the parameter values of the exponential and Gamma prior distributions in Tables 3 and 4 to explicitly define these priors.

Line 352: The column was uniformly packed with dry soil, and the average porosity (0.4) and bulk density (1.24 g/cm^3) were determined using the gravimetric method.

Line 453: The experimental breakthrough curve indicates that the upper bound of Cd concentration does not exceed 75% of the initial influent concentration C_0.

Comment 5: A simple comparison would help isolate the value of the adaptive weighting strategy.
The current scenarios compare no constraint, one constraint, and multiple constraints. It is therefore difficult to determine how much of the improvement comes from adding physical constraints and how much comes from dynamically updating their weights. Please consider adding one simple baseline using multiple constraints with fixed equal weights or fixed weights after residual normalization. This comparison could be limited to one representative case if computational cost is a concern. The discussion should also slightly moderate statements such as "general framework" and "physical constraints function analogously to additional observations." More precise wording would be that

the framework has been demonstrated in two cases and that physical constraints provide additional prior or regularization information. Because the remaining mass-balance errors are not negligible, “reduces violations of mass conservation” would also be more accurate than implying that conservation is fully satisfied.

Response: We thank the reviewer for this important suggestion. We agree that it is necessary to compare the simulation results obtained using the dynamic and fixed weighting strategies. Following this suggestion, we have added the simulation results using the fixed weighting strategy when multiple physical constraints are simultaneously coupled (Figures S5–S8), together with the corresponding performance metrics (Tables S5 and S6), and discussed these results in the revised manuscript (Line 655). In addition, we have revised several statements in the manuscript (Lines 692 and 763) to more accurately describe the applicability of the proposed method and avoid overstatement.

Compared with the dynamic weight-updating strategy, the fixed weighting strategy results in relatively large carbon and chlorine mass conservation errors in Case Study 1 (Figure S5), and substantially poorer predictive performance for all four chlorinated hydrocarbons (Figure S6). Similarly, in Case Study 2, the fixed weighting strategy results in a substantially larger Cd mass conservation error (Figure S7) and greater bias in the predicted Cd concentrations (Figure S8). These results demonstrate that although multiple physical constraints can provide additional information, appropriately balancing the contributions of different physical constraints is critical for improving predictive performance and physical consistency, further supporting the effectiveness and necessity of the dynamic weight-updating strategy.

Line 655: In contrast, the proposed dynamic weight-updating strategy is well suited to combined likelihood functions that integrate multiple sub-likelihoods associated with distinct physical mechanisms, resulting in substantially improved predictive accuracy and physical consistency compared with the fixed weighting strategy (Figure S5-S8 in Supporting Information).

Line 692: Within the Bayesian uncertainty analysis, these physical constraints provide additional information for parameter identification, thereby further reducing parameter uncertainty。

Line 763: This study proposes a structural error correction framework capable of coupling multiple physical constraints, which has been demonstrated through two case studies.

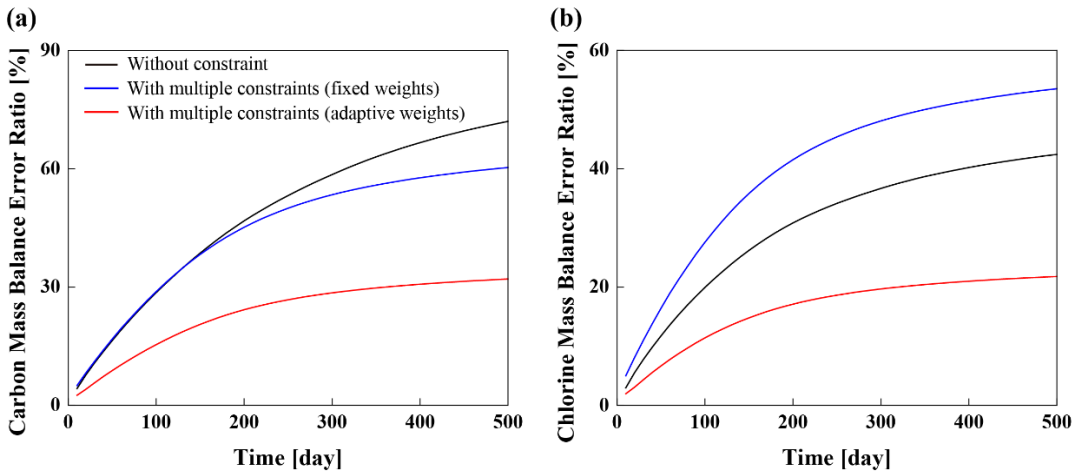

Figure S5. Variation in the carbon/chlorine mass balance error rate for the synthetic groundwater PCE reactive transport model under the multiple-constraint scenario with fixed weights.

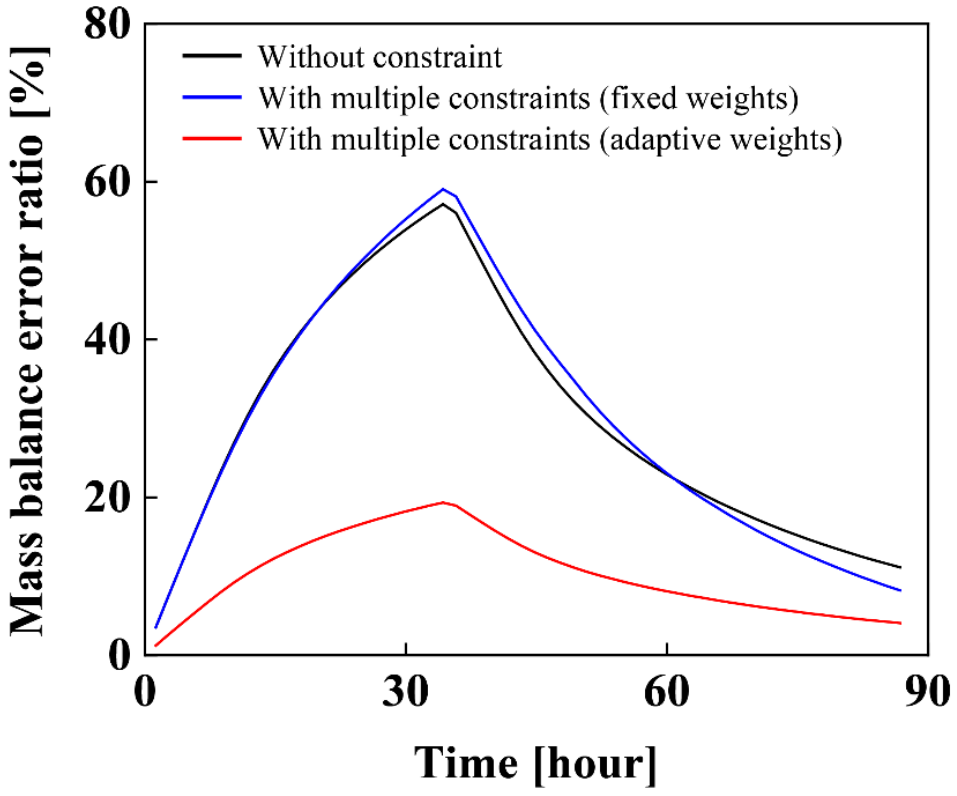

Figure S6. Variation in the cadmium mass balance error rate for the sand column Cd-phosphate cotransport model under the multiple-constraint scenario with fixed weights.

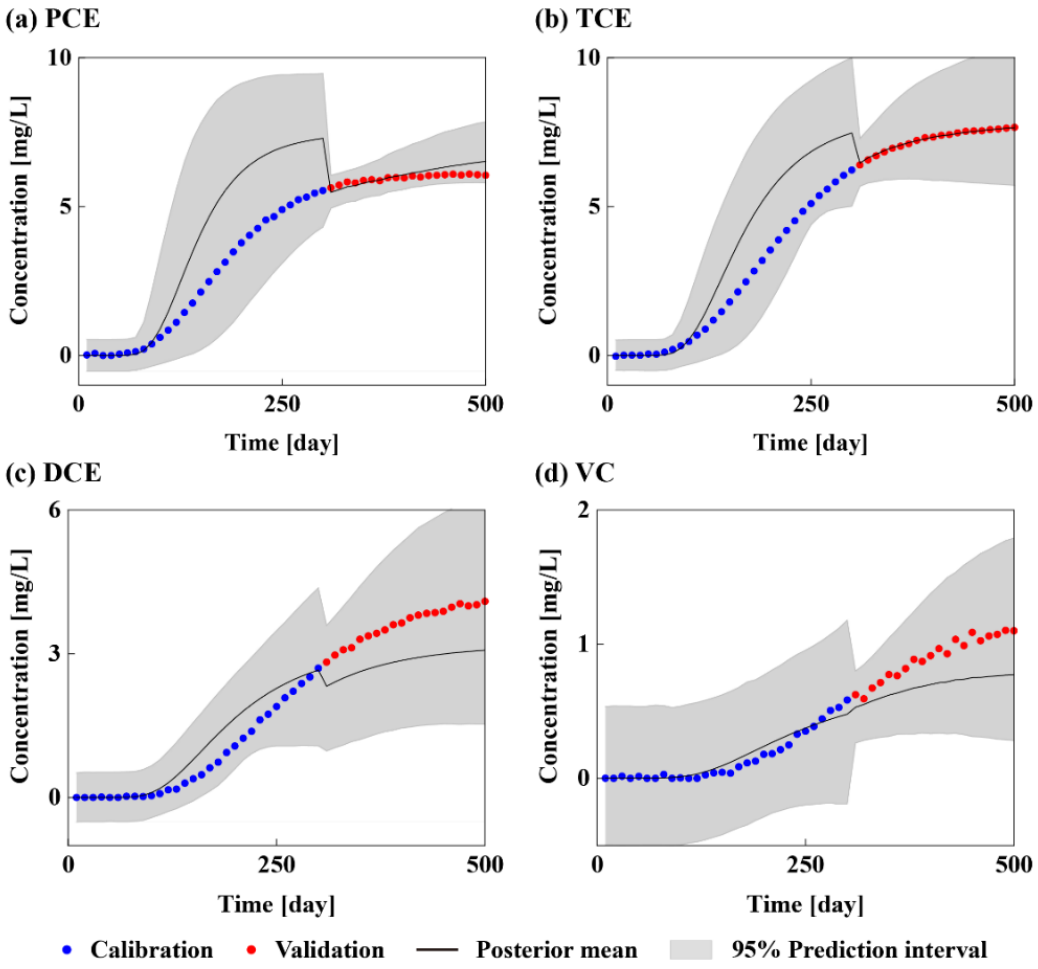

Figure S7 DDM predictions of chlorinated ethenes concentrations under the multiple-constraint scenario with fixed weights.

Table S5 Evaluation of predictions for the synthetic groundwater PCE reactive transport model under the multiple-constraint scenario with fixed weights.

Constraints	Calibration/ Validation	Species	NSE	MAE	RMSE
With multiple physical constraints (fixed weights)	Calibration	PCE	0.3562	1.3212	1.6362
		TCE	0.6518	1.0199	1.2937
		DCE	0.8649	0.2452	0.3297
		VC	0.9274	0.0381	0.0492
	Validation	PCE	-2.1390	0.1848	0.2252
		TCE	0.9921	0.0253	0.0328
		DCE	-3.5710	0.7924	0.8078
		VC	-1.0849	0.2150	0.2324

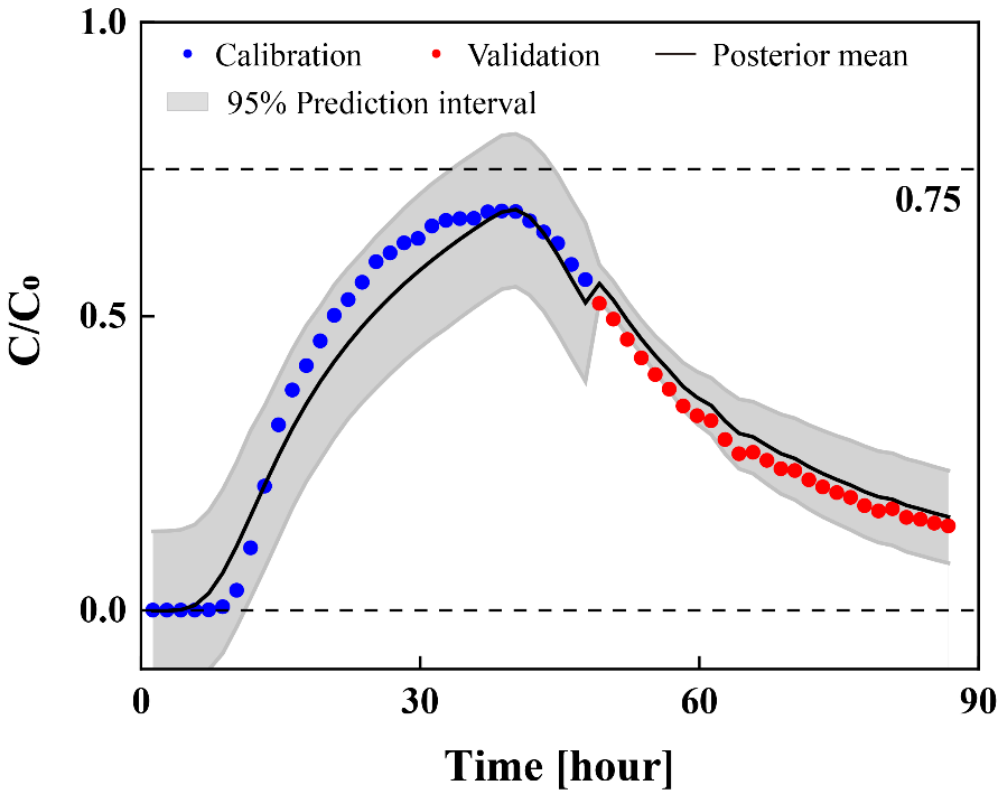

Figure S8. DDM predictions of cadmium concentration under the multiple-constraint scenario with fixed weights.

Table S6 Evaluation of predictions for the sand column Cd-phosphate cotransport model under the multiple-constraint scenario with fixed weights.

Constraints	Calibration/ Validation	NSE	MAE	RMSE
With multiple physical constraints (fixed weights)	Calibration	0.9112	0.0408	0.0499

References

Tian, J., Zeng, X., Wang, D., & Wu, J. (2026). A data-driven approach coupled with physical constraints to improve groundwater models with structural error. Water Resources Research, 62, e2025WR040247. https://doi.org/10.1029/2025WR040247