Dear Editor and Reviewers:

Thank you very much for your valuable comments and insightful suggestions on our manuscript. We have carefully considered each comment and revised the manuscript accordingly. Our detailed responses to each comment are listed below.

Responses to the issues raised by Referee #1 are as follows:

a. The manuscript fails to identify which is the clear innovation brought forward.

Response:

In the revised manuscript, we have reorganized the second-to-last paragraph of the introduction (lines 86-117). Specifically, the innovation of this study is summarized in line 111-117 as follows:

"*In summary, the primary contributions of this study are as follows:*

(1) Proposed a novel inversion framework that integrates the TNNA algorithm with dimensionality reduction techniques, including KLE and generative machine learning methods, thereby extending its applicability to high-dimensional heterogeneous fields characterized by Gaussian and non-Gaussian stochastic processes, respectively.

(2) Conducted a comprehensive comparative analysis between the TNNA algorithm and four conventional metaheuristic algorithms across three case scenarios, highlighting the advantages of machine learning in inverse estimation under different heterogeneous conditions."

b. There is no discussion about parameter sensitivity and hyperparameter optimization of the TNNA algorithm.

Response:

Thank you for your comment. Regarding the hyperparameter optimization for the TNNA algorithm, we did not conduct a formal sensitivity analysis, as the training process of the reverse neural network is guided by the constraints of the inverse objective function, which requires only a set of observation data as input. Therefore, in a GPU hardware environment, we are able to quickly determine suitable hyperparameters based on prior research through an empirical trial-and-error approach.

Recognizing the significant influence of hyperparameters on neural network performance in some certain scenarios, we have emphasized the importance of hyperparameter optimization in future research (see the last sentence of Section 5):

"*Furthermore, hyperparameters can significantly influence neural network performance in certain scenarios. It is necessary for future research to explore hyperparameter optimization and sensitivity analysis to identify the optimal neural network structures and training strategies, ultimately enhancing model performance across diverse hydrological conditions.*"

c. There is no sufficient detail about the computational advantage of TNNA with respect to the other techniques (other than saying that you have to run less number of times the forward model for TNNA).

Response:

Thank you for raising this issue. We acknowledge that methodological details supporting these advantages were insufficiently clarified in our original manuscript. In the revised manuscript, we have emphasized the computational characteristics of the four metaheuristic algorithms in the last paragraph of Section 2.3.1, and clarified in Section 2.3.2 why the TNNA algorithm requires only one forward simulation per epoch when training the reverse network:

"*A common characteristic of all the methods described above is that each iterative update of model parameters requires multiple evaluations of the objective function, and sufficient iterations are necessary to balance local exploitation and global exploration.*" (see the last paragraph of Section 2.3.1)

……

……

"*In the above process, each backpropagation step involves only a single forward calculation of the loss function. After establishing the computational graph, gradients of the trainable parameters* $\theta_{Reverse}$ *are computed through backpropagation combined with automatic*

differentiation. These gradients are then used to update the trainable parameters $\theta_{Reverse}$. Thus, only one forward simulation is executed during each epoch of the reverse network $F_{Reverse}$ training procedure. This presents a marked computational advantage of TNNA compared to the four selected metaheuristic algorithms, which require numerous forward simulations for parameter updates at each iteration." (see Section 2.3.2)

The differences in implementation between these two categories of methods, combined with their comparative results presented, clearly illustrate the computational advantage of the TNNA algorithm.

d. There is insufficient detail about how the surrogate models are trained and on which parameters they are trained on.

Response:

We appreciate this comment. In Section 4.1, we have provided additional details on the specific data structures of the model parameters and outputs used in the surrogate models for the three case scenarios.

"*Surrogate models were first compared using the Case 1 with low-dimensional parameter. For this scenario, the input parameters for the surrogate models consist of a 9-dimensional vector, including 8 permeability parameters and the contaminant source release concentration. The output consists of the simulated hydraulic heads and solute concentrations at 25 observation points.*" (see the first paragraph in Section 4.1)

"*In the two high-dimensional scenarios, the input parameters for the surrogate models are single-channel matrix data representing the heterogeneous parameter field, while the output consists of vector formed by flattening the multi-channel matrix data, representing the simulated hydraulic heads and solute concentrations at predefined time steps within the simulation domain. The training and testing datasets for these two case scenarios consist of 2000 and 500 samples, respectively.*" (see the last paragraph in Section 4.1, line 503-506)

e. The manuscript is too long, and it has too many details on the different methodologies that could be moved to an appendix to make the reader more comfortable while reading the main part of the text.

Response:

Thank you for your suggestion. We have moved the detailed implementation procedures of the metaheuristic algorithms to the supplementary materials and added a summarized paragraph about the four methods in the revised manuscript.

Responses to the issues raised in the Editor's attachment are as follows:

1. Lines 13 & 14. Rephrase the sentences, possibly as follows: "Tandem neural network architecture (TNNA) is a machine learning algorithm which has been recently proposed for estimating uncertain parameters with inverse mappings".

Response:

Suggestion followed.

2. Lines 22 & 25. Here a percentage of noise is mentioned (1% or 10%), but it is not clearly stated which are the measured quantities and what is used as reference value.

Response:

This sentence has been revised as follows:

"*Additionally, we evaluate algorithm performance under two different noise level conditions (multiplicative Gaussian noise with standard deviations of 1% and 10%) for normalized hydraulic head and solute concentration data in the non-Gaussian random field scenario, which exhibits the most complex parameter characteristics.*"

3. Lines 31 to 36. Recent publications only have been considered. However, these concepts are well-established since a long time and can be considered text-book material. Moreover, I wonder whether

the papers referenced for inverse modeling are the most relevant ones. Many other review papers on inverse problems in hydrology are available and should be considered (e.g., 10.1016/0022-1694(87)90207-1, 10.2136/sssabookser5.4.c40, 10.1016/S0167-5648(04)80146-1, 10.2166/nh.2007.024, 10.1029/96WR00160, 10.1007/BF01547729, 10.1016/0309-1708(91)90039-Q, 10.3390/hydrology11110189, 10.1029/WR022i002p00095, 10.1002/hyp.3360060305, and many others).

Response:

Suggestion followed.

4. Line 39. I would not use "deterministic" to characterize Bayesian methods, which are based on the theory of stochastic processes.

Response:

This sentence has been revised as:

"*Among available algorithms, methods based on objective functions established from maximum a posteriori estimation and solved by optimization techniques represent a significant category*"

5. Line 62. Substitute "CNN" with "convolutional neural network (CNN)".

Response:

Suggestion followed.

6. Line 70. Substitute "DNN" with "deep neural network (DNN)".

Response:

Suggestion followed.

7. Line 80. Methods based on the minimization of an objective function can be improved, from the point of view of the computational effort, through the use of the adjoint equation for the computation of the gradient of the objective function. This should be considered by the authors and possibly mentioned or discussed in the manuscript.

Response:

We appreciate this suggestion and have added the following statements at the end of the surrogate modeling section (lines 65–70):

"*Specifically, inversion approaches based on objective function minimization can also benefit from adjoint methods (Plessix, 2006). Integrating adjoint methods with machine learning-based surrogate models enables efficient gradient computation in high-dimensional and complex scenarios, making their practical implementation tractable (Xiao et al., 2021).*"

8. Lines 83 & 360. Is "designs" the best word? May be, "is based on", "considers" or "proposes"? Similarly for "designed" at line 360.

Response:

Thanks for this comment and this word is replaced by "considers".

9. Line 85. Substitute "was" with "is", because the present tense is used in the following sentences.

Response:

Suggestion followed.

10. Line 91. Expression "parameter values transition smoothly across space" could be rephrased, possibly as "the spatial variation of parameter values is quite smooth".

Response:

We have revised the expression as suggested.

11. Line 95. Is "curse" the best word?

Response:

The phrase "curse of dimensionality" is a widely used term in the field of machine learning. However, to avoid potential misunderstanding for readers unfamiliar with this domain, we have rephrased the sentence as follows (line 100):

"*Additionally, dimensionality reduction techniques are necessary for the two high-dimensional cases to reduce computational complexity associated with high-dimensional parameter spaces.*"

12. Lines 95 to 102. These sentences could be improved to explain why different methods have been used for the different scenarios and to motivate the specific choice of each method. This should help to improve the description of what is novel in this work, otherwise the comment by one of the reviewers remains crucial ("The manuscript presents a thorough comparison, but it fails to identify which is the clear innovation brought forward.")

Response:

Thank you for this suggestion. We have added the reason for choosing KLE and generative machine learning methods for dimensionality reduction for Gaussian random fields and non-Gaussian random fields, respectively:

"*Specifically, the Karhunen-Loève Expansion (KLE) method is feasible for Gaussian random fields. It reconstructs the Gaussian random field through a linear combination of orthogonal basis functions, …… These methods can establish relationships between low-dimensional standard distributions (e.g., uniform distribution) and high-dimensional distributions, effectively representing non-Gaussian random fields as low-dimensional latent vectors (i.e., parameters after dimensionality reduction).*"

For the innovation of this study, we have revised the description as follows:

"*In summary, the primary contributions of this study are as follows:*

(1) Proposed a novel inversion framework that integrates the TNNA algorithm with dimensionality reduction techniques, including KLE and generative machine learning methods, thereby extending its applicability to high-dimensional heterogeneous fields characterized by Gaussian and non-Gaussian stochastic processes, respectively.

(2) Conducted a comprehensive comparative analysis between the TNNA algorithm and four conventional metaheuristic algorithms across three case scenarios, highlighting the advantages of machine learning in inverse estimation under different heterogeneous conditions."

13. Line 136 to 138. These sentences could be rephrased, possibly as "These four methods were proposed at different stages of the development of machine learning, but the application for constructing surrogate models in most groundwater modeling scenarios is still relevant." Did I interpret correctly your thoughts? If so, this sentence remain rather nevertheless rather apodictic and I wonder whether it can be supported in a better way from physical arguments or is it necessary.

Response:

Thank you for this suggestion. We agree that the original statement was somewhat general and could be better clarified. To address this, we have revised the description to directly introduce the four machine learning models along with their respective architectural characteristics. The revised text reads as follows:

"Specifically, four popular machine learning models with distinct architectural differences are evaluated for surrogate modeling. These are: multi-output support vector regression (MSVR), a kernel-based architecture for data mapping; fully connected deep neural network (FC-DNN), composed of stacked fully connected layers; LeNet, a classical convolutional neural network (CNN) architecture; and deep residual convolutional neural network (ResNet), which incorporates residual connections into the CNN structure."

14. Line 140. Sentence "The surrogate model for inversion will be constructed using the most accurate among them" remains vague.

Response:

This sentence has been rephrased as: "*The predictive accuracy of four surrogate modeling approaches will be compared in this study, and the best-performing approach among them will subsequently be selected for inversion computations.*" (line 158)

15. Line 141. Expression "the values for different simulation components" is not fully clear to me. All the data sets used for the training are normalized with the formula $X_i = (x_i - x_{min})/(x_{max} - x_{min})$, where x_i is the i-th value of the data set, x_{min} and x_{max} are respectively the minimum and maximum

value of the data set, and X_i is the normalized value. Is this right?

Response:

Yes, your understanding is correct. We have revised the original sentence as follows and added a reference describing the specific normalization method:

"*Before constructing surrogate models, the training datasets are normalized separately for each simulation component using Min-Max Normalization, in which each component is scaled independently based on its minimum and maximum values, ensuring that all normalized values fall within the range [0, 1] (Chen et al., 2021).*" (line 159)

16. Equation (4). how is this equation related to the parameters of equations (1) to (3)? Are x and y scalar or vector quantities?

Response:

Thank you for this helpful comment. We have revised the notation in Equation (4): the original x and y are replaced with $\boldsymbol{m}$ and $\hat{\mathbf{y}}$ to maintain consistency with Equations (1) to (3).

17. Line 150. Substitute "Eq.(5)~(6)", possibly with "equations (5) and (6)".

Response:

Suggestion followed.

18. Lines 151 to 155. The notation has to be modified. What is w^j ? In the second line of equation (6) it could be better to use $(u-\varepsilon)^2$. Remark "ε… insensitive tube" can be erased.

Response:

Thank you for this suggestion. The notation w^j represents the regression vector within matrix W corresponding to the j^{th} observed dataset. We have revised the definitions of W and B in the manuscript as follows: $W=[w^1,\ldots,w^{N_{obs}}]^T \in \mathbb{R}^{N_{obs}\times N_{samples}}$ and $B=[b^1,\ldots,b^{N_{obs}}]^T \in \mathbb{R}^{N_{obs}\times 1}$. Equation (6) has been updated to use the term $(u-\varepsilon)^2$, and the remark "ε… insensitive tube" has been removed accordingly.

19. Lines 164 & 165. Erase "the penalty parameter" and "the kernel function parameter", the name of the variable is sufficient. However, σ is not defined, is it?

Response:

Thank you for this comment. We have erased the phrases "the penalty parameter" and "the kernel function parameter" as suggested. In addition, σ is a bandwidth parameter of the kernel function. So, we supplemented the definition of the kernel function explicitly in the revised manuscript.

"*where $F_{MSVR}(\boldsymbol{m})$ denotes the dataset regression model operator constructed based on MSVR; $\varphi(\boldsymbol{m})$ is a nonlinear regression function that implicitly maps the input vector $\boldsymbol{m}$ into a high-dimensional feature space. Its inner product defines the kernel function $K(\boldsymbol{m}, \boldsymbol{m}_i)$ (here we use the Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{m}, \boldsymbol{m}_i) = \varphi(\boldsymbol{m})^T\varphi(\boldsymbol{m}_i) = exp(-0.5 \parallel \boldsymbol{m}-\boldsymbol{m}_i \parallel^2/\sigma^2)$).*" (line 171)

20. Equation (7). Do W and B have the same meaning as the same quantities in (4)? σ was defined to be a parameter at line 165, here is a function: this is confusing for the Readers who are not familiar with the applied methods. Erase × from the formula.

Response:

We have revised Equation (7) by changing the notations W and B to W_{DNN} and B_{DNN}, respectively, to explicitly indicate that they represent the weight parameter matrix and bias vector of the fully-connected layer in DNNs. Additionally, we have updated the notation of the lth activation function to $f_{\sigma\text{-}l}(\cdot)$.

21. Lines 177 to 204. The notation is unclear, it does not correspond with the notation introduced in the previous part of the manuscript. For instance, symbols F and G have already been used for different quantities. H is not defined is it? The loss function has the same symbol as an hyperparameter of MSVR. ωi in equation (12) is not defined, is it?

Response:

Thank you for your suggestion. We have revised the notation accordingly in the revised

manuscript.

22. Lines 205 to 215. Is the information about the number of neurons in each hidden layer relevant here, namely, in the description of the methodology? It should be stated later and the motivation for the choice of this value should be given. The same comment applies for the type of activation functions. The whole paragraph could be moved to another point, i.e., after the description of the data sets and where the method is applied.

Response:

We appreciate your valuable comment. The determination of the optimal number of hidden layers for the FC-DNN has been moved to Section 4.1. The description of the two CNNs (LeNet and ResNet) architectures remain in the methodology section, as these architectures are fixed throughout this study. The description regarding the hidden-layer design of the FC-DNN has been revised as follows:

"*The performance of the FC-DNN is sensitive to the number of hidden layers, whose optimal value is determined based on specific case studies presented in the application section.*"

The motivations behind the selection of activation functions are supplemented as follows:

"*The activation function for the output layer is Sigmoid to constrain outputs within the range of 0 to 1. For hidden layers, the Swish activation function is adopted due to its smooth form with non-monotonic and continuously differentiable properties, which helps improve the DNN training procedures (Elfwing et al., 2018).*"

Additionally, the explanation of the hidden-layer selection in Section 4.1 has been updated accordingly:

"*For the FC-DNN, the optimal number of hidden layers was separately determined for each of the four datasets. The candidate range for the number was set from 1 to 7. According to the* $RMSE_{All}$ *and* R^2_{All} *values in Table S2 and Table S3, optimal number of hidden layers for in the FC-DNN for* $\boldsymbol{D}_{train\text{-}200}$, $\boldsymbol{D}_{train\text{-}500}$, $\boldsymbol{D}_{train\text{-}1000}$ *and* $\boldsymbol{D}_{train\text{-}2000}$ *are 2, 4, 3, and 3, respectively.*"

23. Lines 217 & 218. I partially disagree with statement "the purpose of a surrogate model is to minimize the difference between the predicted outputs and the numerical modeling outputs": the purpose of a surrogate model is to substitute a high-dimensional model with a low-dimensional model. So the surrogate model must provide outputs which closely resemble those of a high-dimensional model.

Response:

Thank you for highlighting this issue. This statement may lead to misunderstandings. Our intention here was to introduce the formulation of the loss function for the surrogate model. Therefore, we have revised the original text as follows:

"*The surrogate models are trained by minimizing the difference between the predicted outputs* $\hat{y}_i = \boldsymbol{F}_{DNN}(\boldsymbol{m}_i, \theta_{DNN})$ *and the numerical modeling outputs* y_i*. Following prior researches (Mo et al., 2019, 2020; Chen et al., 2021), the loss function is formulated with L1 norm constraints:*" (line 252-254)

24. Line 217. Why an L1 norm? L2 norms have been used so far in the work!

Response:

We are thankful for this comment. The difference between L1 norm and L2 norm is that L1 norm is based on Laplace distribution hypothesis and L2 norm is based on Gaussian distribution hypothesis. For the observation noises are considered as Gaussian distribution in this paper, the objective function for inversion is based on L2 norm. While training surrogate models with L1-norm is primarily based on recommendations from previous studies and insights gained from our own research experience. In fact, L2-norm is also widely used and may be applicable for this study. In the revised manuscript, the relevant reference citations are added.

"Following prior researches (Mo et al., 2019, 2020; Chen et al., 2021), the loss function is formulated with L1 norm constraints:" (line 252-254)

25. Line 221. Statement "a widely used machine learning framework" can be erased.

Response:

Suggestion followed.

26. Line 226. Symbol G has already been used to denote other quantities, functions, etc.

Response:

The symbol $\boldsymbol{G}(\boldsymbol{s})$ is replaced by $\boldsymbol{Y_G}(\boldsymbol{s})$.

27. Line 238 & 239. Sentence "For example,… the reduced-dimensional parameters" can be erased, the citation could be sufficient. However, I wonder whether it is the optimal one.

Response:

This sentence has been removed, and we have added references Loève (1955) and Mariethoz and Caers (2014) for the Karhunen–Loève expansion. (line 273)

28. Section 2.2.2. Once again the notation is confusing: symbols that have been used previously for some quantities are used here to denote different quantities. Formula $\boldsymbol{z}\sim\boldsymbol{q}(\boldsymbol{z})$ is given without an explanation.

Response:

We have revised the symbols used in the manuscript, unifying the representation of the low-dimensional vectors in Sections 2.2.1 and 2.2.2 as $\boldsymbol{z}$. The explanation for the $\boldsymbol{z}\sim\boldsymbol{q}(\boldsymbol{z})$ is supplemented as:

"*The distribution of the latent vectors* $\{\boldsymbol{z}_1,\ldots,\ \boldsymbol{z}_N\}$*, obtained by mapping the N prior model parameter samples* $\{\boldsymbol{m}_1,\ldots,\ \boldsymbol{m}_N\}$*, is denoted as* $\boldsymbol{q(z)}$*.*" (line 278-279)

29. Section 2.3.1. Once again the notation is confusing and sometimes not rigorous. These parts could be moved to the appendix, or, even better in the supplementary material.

Response:

We have revised the notation in this section, and have moved the detailed steps of metaheuristic algorithms to the supplementary material. A summarized paragraph is retained in the main text as follows:

"*The four metaheuristic algorithms used in this paper essentially update model parameters through distinct heuristic stochastic search strategies. ……Simulated Annealing (SA) starts from a random initial solution and iteratively explores neighbouring solutions, accepting them probabilistically based on the Metropolis criterion, while gradually decreasing temperature parameter until convergence (Metropolis et al., 1953; Kirkpatrick et al., 1983).*

A common characteristic of all the methods described above is that each iterative update of model parameters requires multiple evaluations of the objective function, and sufficient iterations are necessary to balance local exploitation and global exploration. Detailed implementation procedures and theoretical foundations of these methods are provided in the supplementary materials. The metaheuristic algorithms used in this study were implemented using the open-source Python package scikit-opt (https://scikit-opt.github.io/)."

30. Section 2.3.2. This section requires a thorough revision, with a clear definition of individual quantities.

Response:

Suggestion followed. The definitions of individual quantities in Section 2.3.2 have been revised.

31. Line 362 & 366. "Measurement unit" is a different concept from"relevant temporal scale".

Response:

To avoid potential confusion with the term "scale," we have revised the manuscript to explicitly describe the differences in geometric sizes among the cases.

"Both Case 1 and Case 2 are approximately tens of meters in size, with simulation time measured in days.……Case 3 simulates contaminant plume migration, has a size of approximately one kilometre, and simulation time measured in years." (line 357-358)

32. Line 366. Could "plain" be substituted with "alluvial"?

Response:

"Plain" has been replaced with "alluvial" as it is indeed more appropriate.

33. Line 374. Which observation data have been simulated? Hydraulic head? Solute concentrations? This is stated much later only.

Response:

Thank you for the comment. We have clarified this in the manuscript as follows (line 369):

"The observation data (hydraulic heads and solute concentrations) for model parameter inversion are generated by adding Gaussian noise perturbations to the numerical model simulation results."

34. Lines 374 to 379. The added noise is proportional to the value of the "measured" value. Therefore, this means that the error on hydraulic head is assumed to be very small close to the boundaries where the prescribed head is 0 m and to be the highest at the opposite border of the domain, where the prescribed head attains high values. Unfortunately, hydraulic head represent a potential and as such it could be changed by adding a constant value, without changing the hydraulic gradient, which is the "engine" of groundwater flow. Therefore, if one used a different reference height for hydraulic head, the absolute value of errors on hydraulic head and the errors on hydraulic gradients would change a lot.

Response:

Thanks for this valuable comment. I fully agree with your point of view, and your insights will provide significant inspiration for our future research. We admitted that our original description led to your misunderstanding. The observation noise in this study was added after data normalization. Thus, no matter how large the measured hydraulic head values are, their normalized values always range from 0 to 1, and these normalized values directly reflect the relative differences in hydraulic head. Additionally, the primarily purpose of this study is to examine how varying noise levels affect the inversion results. Applying multiplicative noise with different standard deviations provides a feasible method to design two distinct observational noise levels. To avoid misunderstandings, we have explicitly clarified in the revised manuscript that noise was introduced based on the normalized numerical simulation results:

"*Specifically, observational noise is introduced by multiplying the* ***min-max normalized*** *simulated data by a random noise factor* $\epsilon \sim N(1, \sigma^2)$," (line 370)

35. Line 376. Once again, ε is used to denote a different quantity.

Response:

The "ε" is replaced by "ϵ" here.

36. Lines 383, 515, 649. Symbol "~" should be substituted with "to".

Response:

Suggestion followed.

37. Sections 3.1 to 3.3. Which method is used for the simulation of flow and transport? Finite differences, finite elements, finite volumes,…? Eulerian or Lagrangian methods for solute transport? Which time spacing is used? Is the transport model purely convective?

Response:

Thank you for this comment. We have supplemented additional information about the forward modeling solution in the last sentence of the first paragraph of Section 3, as follows:

"The numerical models of the three cases are established using TOUGHREACT, which employs an integral finite difference method with sequential iteration procedures and adaptive time stepping to solve the flow and transport equations. Dispersion effects are inherently incorporated through molecular diffusion and numerical dispersion induced by upstream weighting and grid discretization (Xu et al., 2011)." (line 362-365)

38. Lines 390, 402, 423. Words "meshes" or "grids" should be substituted, possibly with "cells" or "elements".

Response:

Thank you for this comment. We have uniformly replaced "meshes" and "grids" with "cells" throughout the manuscript.

39. Figure 2, Lines 565ff. Here upper case K is used for permeability, whereas lower case k is used in the text. I prefer the latter choice, but a uniform symbol should be used throughout the whole manuscript.

Response:

We have uniformly revised the notation to use the lowercase *k* throughout the manuscript.

40. Line 394. Word "uncertain" can be erased.

Response:

Suggestion followed.

41. Line 406. Expression "are as:" should be corrected.

Response:

"are as:" has been modified as "are"

42. Line 409. Word "stable" should be substituted with "stationary" or "steady-state".

Response:

The word "stable" has been replaced with "stationary" as recommended.

43. Line 410. Add a reference for "equifinality". Indeed, in this way an important prior information and regularization is introduced, without proper discussion.

Response:

Suggestion followed. We have added relevant discussions and references:

"*It should be noted that in high-dimensional parameter scenarios, the increased degrees of freedom typically result in greater parameter uncertainty. Insufficient observational information may fail to effectively constrain parameter estimation, resulting in potential uncertainty and equifinality (Mclaughlin and Townley, 1996; Zhang et al., 2015; Cao et al., 2025). ……introducing these constraints ensures the stability and robustness of the inversion outcomes without affecting the inherent performance characteristics of the five optimization algorithms compared in this study.*" (line 406-413)

44. Line 428. Expression "t=2~24 years" could be substituted as "from 2 years to 24 year".

Response:

Suggestion followed.

45. Line 450. Expression "Figure S3~Figure S6" should be substituted, possibly with "Figures S3 to S6 in the supplementary material".

Response:

Suggestion followed.

46. Lines 458ff, Figures 5, 6 & 13, Tables 1 & 3. Measurement units for RMSE are missing. How is RMSE computed for all the model outputs? Head and concentration errors cannot be simply summed up, as they bear different measurement units.

Response:

When calculating RMSE, both hydraulic head and solute concentration data are normalized to a unified scale between 0 and 1. Consequently, the RMSE values are dimensionless and have no specific measurement units. To clarify this point, we added the following statement at the end of the second paragraph in Section 4.1:

"*Additionally, it should be noted that the above RMSE and R^2 metrics are computed based on the normalized hydraulic head and solute concentration data.*" (line 454-455)

47. Figures 5 and 6. Expression "(a) ~(c) are" should be substituted, possibly with "Plots (a) to (c) show".

Response:

Suggestion followed.

48. Section 4.2.1. What is the "logarithmic average convergence" represented in Figure 7? Is it the RMSE?

Response:

"logarithmic average convergence" represents logarithmic objective function values during inversion iterations. We have added a clarification in line 513:

"*Figure 7 presents the logarithmic average convergence curves (i.e., log10 of the average objective value during inversion iterations) of four metaheuristic algorithms and the TNNA algorithm throughout 100 parameter scenarios.*" (line 523-524)

49. Figure 7. Why the initial value is different among different algorithms? The caption does not specify what is the difference between the four plots. It would be important to recall that the TNNA curve is the same for all the plots. Why is the curve of DE so "noisy"? I have not recognized such an irregular behavior in my experience with that algorithm. The TNNA curve is quite smooth, but it shows very small bumps, in particular slightly after 150 iterations. Is there any explanation for that behavior?

Response:

Thank you for your insightful comments and questions. Below, we provide clarifications regarding your concerns:

The differences in initial values arise from the distinct initialization strategies of the algorithms. For the four metaheuristic algorithms (DE, GA, PSO, SA), the initial objective value corresponds to the best among N_{PC} candidates randomly sampled from the prior distribution. In this study, each algorithm was run independently without a fixed seed, resulting in slight variations due to randomness, though the values remain within a similar range. In contrast, the initial model parameters of TNNA method are not directly sampled, but are determined by the randomly initialized weights of the reverse network. Therefore, its initial objective value typically differs significantly from those of the metaheuristic algorithms. For the purpose of this study, the inconsistency in initial points does not affect the comparison of results.

The caption for Figure 7 has been revised as:

"***Figure 7.*** *Comparative convergence trends (log_{10} of the average objective value) of five optimization algorithms on 100 parameter scenarios. Plot (a)~(d) compare the four metaheuristic algorithms and TNNA under N_{PC}=20, 40, 60, and 80, respectively; TNNA was executed only once on the same 100 parameter scenarios, and its curve is identical across all plots; Markers indicate convergence values every 10 iterations.*"

Regarding the issue of the noisy curve of DE, model parameter optimization by metaheuristic algorithms is a stochastic process, and it is normal for fluctuations in objective function values to occur during convergence processes. For the DE method, when N_{PC}=80 for instance, the objective function values after 150 iterations range between 9.05×10^{-5}~1.32×10^{-4} (corresponding to logarithmic values of -4.04~-3.88 in Figure 7(d)). Fluctuations between consecutive iterations typically remain within 1×10^{-5} (mostly around 3×10^{-6}), which is a reasonable magnitude for optimization algorithms. The DE curve appears more noticeably noisy in Figure 7 due to its relatively larger fluctuation amplitude compared to other methods.

50. Line 505. Is the noise additive or multiplicative? It seems to be additive, now. So there is a difference with respect to what has been described at lines 374 to 379.
Why?

Response:

Thank you very much for pointing out this issue. We used multiplicative noise in all numerical examples presented in this study. We have also rechecked our computational procedures and confirmed that this inconsistency was a typo. We have corrected it in the revised manuscript accordingly.

51. Lines 516ff. Validation should refer to the use of data sets corresponding to different physical situations from those considered during calibration. So this is not a standard and thorough "model validation".

Response:

Thank you very much for this clarification. We acknowledge that our understanding and usage of the term "validation" were indeed not precise enough. To avoid any misunderstanding, we have revised the description related to this aspect accordingly in the manuscript.

52. Line 572. Numbers in “K4 and K6” should be subscripts.
Response:
Suggestion followed.

53. Figures 8 & 9. The captions do not provide full descriptions of the figure content.
Response:
Suggestion followed.

54. Lines 584ff. Once again, “deterministic” is used in a context where the Bayesian, stochastic approach is mentioned.
Response:
Thank you for your comment. The “deterministic” is replaced by “well-defined”.

55. Figure 10. The figure caption should be rewritten. Six rows are mentioned, but the figure has 4 rows and 6 columns. No explanation is given for (a) to (d).
Response:
The figure caption has been revised as:
“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 5, achieved by four metaheuristic algorithms (plots (a) to (d) correspond to GA, DE, PSO and SA, respectively).”

56. Figure 11. The figure caption must be completed with the description of what is represented in the four images.
Response:
The figure caption has been revised as:
“*Figure 11. Spatial distributions log-permeability field estimation results and absolute errors for Scenario 5, achieved by the TNNA. Plots (a) and (c) show the log-permeability fields estimated using 1000 (TNNA-1000) and 200 (TNNA-200) training samples, respectively; plots (b) and (d) present the corresponding absolute error distributions.*”

57. Figure 13. Upper case letter should not be used for measurements units: “days”, not“Days”.
Response:
Suggestion followed.

58. Figure 14. The second row of plot (a) shows a “wavy” behavior. Can it be explained?
Response:
This "wavy" behavior primarily results from the numerical precision of the simulated hydraulic head data. In this study, the hydraulic head simulation precision is 0.01 m (i.e., 1 cm), which means that the minimum scale of simulated error is also 0.01 m. Given that the color bar for displaying hydraulic head errors ranges from 0 to 0.1m, this discretization at intervals of 0.01 m creates a visual "wavy" pattern.

59. Section 4.2.3. It is not clear if the values of permeability of the two hydrofacies have been estimated or have been fixed. In other words, which are the parameters to be identified in this tests?
Response:
We appreciate this comment and apologize for the lack of clarity. In this case, the permeability values of the two hydrofacies were fixed. The parameters to be identified are the hydrofacies types assigned to each discrete grid cell, essentially formulates a high-dimensional binary inverse problem. We have supplemented this clarification at the end of the first paragraph of Section 3.3 as follows:
“This case focus on a high-dimensional binary inverse problem aimed at identifying the lithofacies type of each discrete grid cell within the domain. Note that the permeability values of the two lithofacies are fixed in this case.” (line 429-431)

60. Line 669. Expression “Figure 15-16” should be substituted with “Figures 15 and 16”.
Response:

Suggestion followed.

61. Lines 716 & 717. Sentence “three key aspects should be considered to extended for real-world applications” should be rephrased.

Response:

Thank you for this suggestion. The sentence has been revised to:

“*Given the complexities of subsurface systems, three key aspects should be considered to extend the TNNA method to real-world applications.*” (line 736-737)

62. Lines 724ff. The statement “heterogeneity exhibits ambiguous statistical features” is not clear to me and this makes it unclear also the following remarks.

Response:

This part aims to emphasize that the primary challenge faced in practical research is obtaining representative parameter field samples. The original description has now been revised accordingly:

“*Generative machine learning methods (including state-of-the-art variants) also have the potential to characterize more complex non-Gaussian fields. However, obtaining representative parameter field datasets remains challenging in practical research. For instance, spatial variations in non-stationary stochastic aquifer systems may result in significant discrepancies in geostatistical parameters across sampling windows (Mariethoz and Caers, 2014). Therefore, developing appropriate generator training strategies is essential for these practical scenarios.*” (line 742-747)

63. Line 732 & 733. Expression “such designs are also to eliminate” should be rephrased.

Response:

Thank you for this comment. This sentence has been revised to

“Such monitoring strategies for comparing inversion methods also aim to minimize external interferences, ensuring that performance differences are primarily determined by inversion algorithms themselves.” (line 751-753)

64. Line 763 to 774. These sentences discuss potential future developments, which are not based on the results of this work: therefore, they can be erased.

Response:

Suggestion followed.

65. Lines 787, 795, 804, 814, 843, 855, 859, 874, 886, 915, 918, 921, 926, 954, 973, 981. The page numbers or the paper numbers of these scientific articles are missing.

Response:

This missing page numbers have been completed. Note that the references originally listed in lines 804 and 843 have been deleted along with the corresponding paragraphs, as suggested in comment 64.

66. Line 835. Volume number and page or paper numbers are missing.

Response:

This reference has been deleted along with the corresponding paragraphs, as suggested in comment 64.

67. Lines 839, 848, 913. Several details are missing for these references.

Response:

Missed details have been supplemented, primarily including the conference locations and DOI information.

68. Line 847. Details of this reference should be checked.

Response:

Done.

69. Lines 850, 904. DOI is missing for these references.

Response:

The DOI for these references have been supplemented.

70. Line 908. Details of the reference should be corrected.

Response:

Suggestion followed.

71. Line 911. “npj Digital Medicine” should be checked.

Response:

This reference has been deleted along with the corresponding paragraphs, as suggested in comment 64.