Dear Editor:

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.

1. The answer to comment "b. There is no discussion about parameter sensitivity and hyperparameter optimization of the TNNA algorithm" by referee #1 is not satisfactory. I warmly ask the authors to carefully reconsider this issue.

Response:

Thanks for this comment. We have added a paragraph at the end of Section 2.1.2 clarifying the critical hyperparameters affecting DNN training when using the Adam optimizer within a PyTorch implementation. Specifically, the paragraph discusses how these hyperparameters influence the training outcomes and highlights that their optimal values for different scenarios were determined through a trial-and-error approach. The added paragraph reads as follow:

"*When conducting DNN training, hyperparameter selection primarily influences the update process of trainable parameters. Besides the weight decay mentioned above, learning rate and the number of epochs are two other crucial hyperparameters directly affecting training stability and convergence speed. A larger learning rate accelerates initial convergence but may lead to oscillations near the optimal solution, whereas a smaller learning rate tends to improve final accuracy but requires more epochs to achieve convergence. In this study, we first set a relatively large number of epochs to ensure that the trainable parameters are adequately updated. Subsequently, appropriate learning rates and weight decay values for different scenarios are determined through a trail-and-error approach.*"

2. The answer to comment "d. There is insufficient detail about how the surrogate models are trained and on which parameters they are trained on" by Referee #1 is not fully satisfactory. Some information has been added, but I am afraid that the information is not sufficient for a reader who would like to apply the same procedure.

Response:

Thanks for this comment. We have added the hyperparameter settings used for training the neural network models as follows:

"When training the FC-DNN, LeNet, and ResNet for Case 1, the hyperparameters for batch size and learning rate were consistently set to 50 and 1×10^{-4}, respectively. The weight decay values for LeNet and ResNet were both set to 1×10^{-5}, while FC-DNN used a weight decay of 0. The number of training epochs was uniformly set to 500 for all three models."(Line 507-509)

"For ResNet training in Case 2 (Gaussian random field), the hyperparameters were set as follows: batch size =100, learning rate =1×10^{-4}, and weight decay =1×10^{-6}. For Case 3 (non-Gaussian random field), the corresponding values were batch size = 50, learning rate = 1×10^{-3}, and weight decay= 1×10^{-8}. In both cases, the number of training epochs was also set to 500." (Line 555-557)

3. Lines 370ff. the description of the noise added to the data is now very clear. Unfortunately, my comment remains still valid and basically not answered. In fact, if is the "true" physical quantity, then the normalized value is $X=(x-x_{min})(x_{max}-x_{min})^{-1}$ and $x = x_{min} + X(x_{max} - x_{min})$. The noisy normalized value is $X' = (1 + \delta)X$, where δ follows a normal distribution, with zero mean and unit standard deviation. The corresponding noisy physical quantity is given by $x'= x_{min}+(1+\delta)X(x_{max}-x_{min})$.

The absolute error on the physical quantity is then $x' - x = \delta X(x_{max} - x_{min})$. It is clear from this formula, that the absolute error is proportional to X, so that it is negligible for values of x close to $xmin$. On the other hand it is maximum, when x is close to x_{max}. I do not see any physical reason for such a choice. Moreover, the relative error is given by

$$\frac{x' - x}{x} = \frac{\delta X(x_{max} - x_{min})}{x_{min} + X(x_{max} - x_{min})}$$

If $x_{min} = 0$, then the relative error is equal to δ, otherwise it depends also on the value of X.
I think that a proper discussion of this issue is very important.

Response:

We appreciate this detailed and insightful comment regarding the formulation of the multiplicative noise model and its implications for the distribution of errors in the physical domain. We agree that this type of noise leads to absolute errors that are proportional to the normalized values, resulting in negligible errors near the lower bound ($x_{\min}$) and more significant errors near the upper bound ($x_{\max}$).

The multiplicative noise model adopted in this study was based on its usage in some previous studies, and the physical motivation behind this choice was not thoroughly examined at the time. Based on the literature reviewed during this revision, we recognize that the multiplicative noise model has the practical advantage of ensuring non-negativity of perturbed observations, which is important near plume boundaries with low concentrations. Additionally, whether observation noise depends on the measured values in practice is often determined by the specific measurement technique employed. In the revised manuscript, we have supplemented the discussion of this issue, including two measurement scenarios where multiplicative noise may naturally arise in real-world environmental applications. The complete revised paragraph added to the manuscript is as follows:

"Here we applied the multiplicative noise is intended to ensure that all perturbed observation values remain non-negative, which is particularly important in regions near plume boundaries where concentrations are close to zero. Generally, observation errors are assumed to be independent of the measured values, whereas the multiplicative noise model introduces value-proportional perturbations, resulting in a positive correlation between the standard deviation of observation noise and the true values. This type of error dependence may also exist in real-world studies when certain measurement techniques are used. For example, in hydraulic head monitoring, pressure transducers may exhibit drift (i.e., a persistent deviation in output not caused by actual pressure changes) due to aging and fatigue of components such as the diaphragm or strain gauge, leading to reduced measurement accuracy (Sorensen and Butcher, 2011). Variation in hydraulic pressure can lead to different levels of drift among transducers, with those installed at higher pressure (i.e., high hydraulic heads) environments tending to experience more significant drift and thus being more prone to elevated observation noise. For the analysis of solute concentrations in laboratory settings, when the concentrations of water samples exceed the detection range of the instrument, a common approach is to dilute these samples prior to measurement. While analytical instruments may introduce additive errors at a relatively fixed level, the rescaling process following dilution (i.e., multiplying the measured value by the dilution factor) amplifies these errors. As a result, the final measurement error becomes approximately proportional to the original solute concentration (Kabala and Skaggs, 1998). Given that the goal of this study is to evaluate the robustness of five inversion algorithms under different noise levels, both additive and multiplicative noise models are suitable for representing observational uncertainty. Prior work by Neupauer et al. (2000) demonstrated that the choice between these two noise types has minimal influence on the comparative performance of inversion methods."

4. The answer to the comment on Figure 14, namely on the "wavy" behavior shown in the second row of plot (a), opens a relevant question. If I understood properly, the hydraulic head has been simulated with a threshold of 0.01 m on the iterative method applied to solve the flow equation: if this is the case, the simulation error is greater than 0.01 m. However, let us assume that this is the order of magnitude of the simulation error on "noise-free" heads. Then, this is the same order of magnitude as the error added in the low-noise tests, when a 1% standard deviation is considered. In other words, the noise-free data share an error with the same order of magnitude as the noise in the low-level tests. However, the basic difference is that the simulation error could have the same absolute error everywhere, whereas the added noise is proportional to X, as demonstrated above. Am I wrong?

Response:

We appreciate this insightful comment. The numerical accuracy in this study is controlled based on the relative error in the conservation equations. Specifically, we set the convergence criterion to a relative error threshold of 10^{-5}, meaning that the maximum local mass imbalance in any grid cell does not exceed one part in 100,000 of the total mass in that cell. After completing the numerical simulation that meets this convergence criterion, the hydraulic head outputs were

recorded with a precision of one centimeter. We have added a clarification regarding the solver's numerical precision in lines 380–385 of the revised manuscript, as follows:

"*In all the three cases, the relative error tolerance for the conservation equations was uniformly set to 10^{-5}, ensuring that the maximum imbalance of conserved quantities within each discrete grid cell remains below one part in 100,000 of the total quantity in that cell.*"

The results compared in Figure 4 represent the absolute differences between numerical model outputs generated using the true parameter field and those obtained using the parameter fields estimated through inversion. These error distributions do not reflect numerical approximation errors, nor do they include the added multiplicative noise. In Lines 700–702, we have explicitly clarified the meaning of symbols used in Figure 4 as follows:

"*Note that in Figure 14, $\hat{\boldsymbol{y}}_{\boldsymbol{H}}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{C}}$ represent the simulated spatial distributions of hydraulic heads and solute concentrations based on the estimated permeability fields through inverse modeling, while $\boldsymbol{y}_{\boldsymbol{H}}$ and $\boldsymbol{y}_{\boldsymbol{C}}$ represent those simulated using the true permeability field.*"

5. Line 43 & 44. Sentence "Among available algorithms, methods based on objective functions established from maximum a posteriori estimation and solved by optimization techniques represent a significant category" remains quite ambiguous. May be, it could be substituted with "Methods based on the minimization of objective functions or the maximization of posterior distributions require the application of optimization techniques".

Response:

Suggestion followed.

6. Lines 67 to 70. Indeed, my comment intended to stress that the use of the adjoint equation limits the number of runs of the forward problem for the application of gradient-based optimization algorithms.

Response:

Thank you for this comment. We have added a discussion at the end of the second paragraph of the Introduction regarding the application of adjoint equations in optimization algorithms, and highlighted potential implementation challenges including the overwhelming programming effort and complexity involved in deriving adjoint equations:

"*The efficiency of optimization algorithms can be enhanced by integrating them with adjoint methods, particularly when extended to high-dimensional parameter spaces. Adjoint methods are capable of efficiently computing gradients for all parameters simultaneously through solving adjoint equations derived from the original forward model (Plessix, 2006). This gradient information can directly accelerate local optimization algorithms (Epp et al., 2023) and facilitate gradient-enhanced global optimization methods (Kapsoulis et al., 2018), significantly improving efficiency in complex inverse problems. However, practical implementation of adjoint methods remains challenging due to the overwhelming programming effort and the complexity associated with deriving adjoint equations, especially for strong nonlinear system models (Xiao et al., 2021; Ghelichkhan et al., 2024).*"

In the third paragraph of the Introduction, the advantages of integrating adjoint methods with DNN-based surrogate models was revised as follows:

"*Additionally, due to their inherent differentiability and continuity, DNN-based surrogate models can be integrated with adjoint equations, enabling efficient gradient computations, and significantly facilitating their practical implementation in high-dimensional and complex scenarios (Xiao et al., 2021).*"

7. Line 115. The sentence should be rephrased. It might be ambiguous to what word "respectively" refers.

Response:

Thank you for this comment. This sentence has been revised as follows:

"Proposed a novel inversion framework that integrates the TNNA algorithm with dimensionality reduction techniques, including KLE for Gaussian stochastic processes and

generative machine learning methods for non-Gaussian stochastic processes, thereby extending its applicability to high-dimensional heterogeneous fields."

8. Line 117. Specify which is the benchmark with respect to which ML methods give an advantage.
Response:

We appreciate this comment. The benchmark here refers to "metaheuristic stochastic search strategies", and the machine learning methods mentioned here correspond to "DNN-based reverse mapping". Accordingly, this sentence is revised as:

"*Conducted a comprehensive comparative analysis between the TNNA algorithm and four conventional metaheuristic algorithms across three case scenarios, highlighting the advantages of DNN-based reverse mapping over metaheuristic stochastic search strategies for inverse estimation under different heterogeneous conditions.*"

9. Lines 119 & 120. Sentence "With advancements… for future studies" could be erased.
Response:

Suggestion followed.

10. Line 141. $\boldsymbol{z}$ is not defined, is it?
Response:

Thank you for this comment. The definition of z has been supplemented as follows:

"*……, where $\boldsymbol{z}$ is a low-dimensional vector whose parameter space is commonly defined as an easily sampled probability distribution (e.g., standard Gaussian or uniform distribution).*"

11. Line 145, equation (3). If I understand correctly the notation, the operator for parameter dimensionality reduction $\boldsymbol{G}$ computes the high-dimensional parameter vector $\boldsymbol{m}$ starting from a low-dimensional vector $\boldsymbol{z}$. Then the computed vector $\boldsymbol{m}$ is used in the forward (high-fidelity?) model. I think this is not the proper description of surrogate models.
Response:

Thank you for this comment. We understand your concern and recognize that our original wording might have caused confusion. To clarify, in high-dimensional parameter scenarios, it is necessary not only to construct a surrogate model to enhance the computational efficiency of forward simulations, but also to address the dimensionality reduction of high-dimensional model parameters. We have enriched the original paragraph with more details, explicitly explaining how the dimensionality-reduced parameter representation allows indirect inversion based on Equation (3). The revised paragraph is as follows:

"*In high-dimensional parameter scenarios, directly optimizing the model parameter $\boldsymbol{m}$ can lead to computational difficulties due to its high dimensionality. To mitigate this issue, in addition to constructing a surrogate model $\boldsymbol{F}_{Forward}(\cdot)$ to improve the computational efficiency of forward simulations,……*".

12. Line 150. Expression "calculating their responses" should be rephrased.
Response:

Thanks for this comment. This sentence is revised as follows.

"The process begins by sampling model parameters from prior distributions. The corresponding system responses for these parameter samples are simulated using a high-fidelity numerical model."

13. Line 155. Expression "convolutional neural network" can be erased, because the acronym CNN has already been defined.
Response:

Suggestion followed.

14. Line 162. Is the reference Chen et al. (2021) significant? The min-max normalization was introduced long time before that paper.

Response:

Thanks for this comment. We agree that min-max normalization (0–1 normalization) has indeed been introduced much earlier than Chen et al. (2021). The purpose of citing this paper here was primarily to provide a recent reference that clearly describes the complete formulation and implementation details, particularly regarding data normalization methods specifically tailored for groundwater model parameter inversion involving multiple simulated components. This is beneficial for readers to directly understand the detailed implementation processes. We have accordingly clarified this in the revised manuscript by adding the following note:

"*Details regarding the specific formulation and implementation of normalization in groundwater models involving multiple simulated components can be found in Chen et al. (2021)*"

15. Lines 173ff. $\boldsymbol{m}_i$ is not defined, is it?

Response:

Thank you for this comment. The definition of $\boldsymbol{m}_i$ has been supplemented as follows:

"*……$\boldsymbol{m}_i$ denotes the ith model parameter vector from the surrogate model training dataset).*"

16. Line 237. Other choices, e.g., hyperbolic tangent, could provide output in the range from 0 to 1. So, why the Sigmoid was chosen?

Response:

Thank you for this insightful comment. Indeed, the primary consideration in choosing the activation function for the output layer is ensuring that the output values align with the target value range. We agree that other activation functions, such as the hyperbolic tangent (-1 to 1) and ReLU (0 to $+\infty$), could also theoretically serve as suitable output-layer activations, and these functions have also proven effective in practice to guarantee surrogate model accuracy. We specifically chose the Sigmoid function to strictly constrain initial model outputs within the target range (0–1), thereby reducing the risk of occasional extreme or anomalous predictions, particularly in the early stages of training. We have included the following explanation in the revised manuscript:

"*Note that other activation functions whose outputs cover this range can also be adopted, ……, thereby reducing the risk of occasional extreme or anomalous predictions, particularly in the early stages of training*"

17. Lines 253 & 254. The motivation for the use of L1 norm for the loss function is not very informative. Moreover, is "constraints" the right word here?

Response:

We appreciate this comment. This sentence related to the word of "constraints" has been modified as:

"*……, the L1 norm-based loss function is adopted and formulated as:*"

We have added the following description to clearly illustrate the implications of selecting the L1 or L2 norm for surrogate modeling:

"*It should be note that the L2 norm can also be employed as a loss function in constructing surrogate model tasks. Due to its squared-error formulation, the L2 norm provides smoother gradients and more stable parameter updates near convergence compared to the L1 norm; however, this formulation also makes it more sensitive to extreme outliers. When the sampled parameters sparsely cover the parameter space, adopting the L1 norm loss function can improve the robustness of surrogate model predictions.*"

18. Line 286. G has already been used at line 141, even if here it is a scalar and there it was a vector.

Response:

The "G" has been replaced by "$\mathcal{G}$"

19. Line 333. “After obtain” should be corrected as “After obtaining”.
Response:
Thanks for this comment. We have corrected it.

20. Line 338. Word “vetcor” should be corrected as “vector”.
Response:
Done.

21. Line 355. Expression “measured in days” should be substituted, possibly with “of 60 days”.
Response:
Suggestion followed.

22. Line 358. Expression “measured in year” should be substitute, possibly with “of several years (up to 30 years)”.
Response:
Suggestion followed.

23. Line 409. The concept of equifinality has been introduced in hydrology at least in 1992 by Beven & Binley (DOI: 10.1002/hyp.3360060305).
Response:
This reference has been added accordingly.

24. Line 429. Word “focus” should be corrected as “focuses”.
Response:
Thanks for this comment. The “focus” has been corrected as “focuses”.

25. Line 524. Word “during” should be substituted, possibly with “as a function of the”.
Response:
Suggestion followed.

26. Lines 555ff, Figure 7. The remark about the “noisy” curve of DE that the authors included in their “Response to comments” should be added to the text of the manuscript.
Response:
Suggestion followed. The added sentences are as:
“It is worth noting that the five optimization algorithms rely on stochastic processes for parameter updates. Therefore, the objective function values are not guaranteed to decrease monotonically with each iteration. According to Figure 7, the DE algorithm exhibits more noticeable fluctuations compared to other algorithms. Nevertheless, these fluctuations remain within a reasonable range. For example, at N_{PC}=80, 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 considered reasonable for optimization algorithms.”

27. Lines 740ff. Punctuation should be checked.
Response:
Done.

28. Line 743. Words “also” and “more” could be erased.
Response:
Suggestion followed.

29. Lines 743 & 744. Expression “representative parameter field datasets” should be clarified.

Response:

This sentence is revised as follows:

“*However, obtaining representative parameter field datasets that accurately capture the spatial variability and heterogeneous geostatistical characteristics of the target aquifer remains challenging in practical research.*”

30. Line 761. Meteorological factors can affect head measurements, but other factors could be much more relevant.

Response:

We appreciate this comment acknowledge that the original description was overly absolute. This sentence was revised as follows:

“*Similarly, hydraulic head measurements may be influenced by other factors, including meteorological conditions, human groundwater extraction, and engineering disturbances, among others*.”