Reply to Reviewer's Comments on egusphere-2025-5694

We are grateful to the Editor, Dr. Yonggen Zhang, for his time and effort in handling our manuscript and for providing us with the opportunity to revise and improve it. We also sincerely appreciate the two anonymous reviewers for their thoughtful and constructive comments. Below, we provide a detailed, point-by-point response to the reviewers' comments. The reviewers' comments are presented in black, our responses in blue, and the changes made in the revised manuscript in red. All modifications in the revised manuscript are highlighted in red for ease of cross-reference.

Reply to RC1's Comment

General Assessment

Space-time varying sensitivity analysis is an ongoing topic for research, given its potential value for deeper analysis of distributed models on the one side, and it's very high computational demand on the other side. The study presented is thus relevant for the community.

My main points for revisions are the Discussion section – which does not compare the current results with previous findings, and a lack of robustness analysis and testing of the influence of choices and assumptions. These can both be rectified. Please see may detailed comments below.

Response: We sincerely thank the reviewer for the positive assessment of the relevance and significance of this study, as well as for the insightful and constructive suggestions. In response, we have thoroughly revised the manuscript to improve its clarity, rigor, and scholarly depth. The main revisions are summarized as follows:

(1) To enhance focus and clarity, the time-varying sensitivity analysis was removed, and the manuscript now concentrates exclusively on spatial sensitivity patterns. The title has been revised accordingly. The new title now reads "Identifying Dominant Parameters in SWAT Across Subbasin and HRU Scales Using a Two-Step Deep Learning-Assisted Spatial Sensitivity Analysis".

(2) The Discussion section has been comprehensively reorganized. Four new subsections were added to evaluate the effects of performance metrics, parameter distributions, upper stream gauge inclusion, and screened parameters on sensitivity patterns.

(3) Additional tests were conducted to ensure the reliability of the sensitivity results, including Morris screening stability analysis, convergence diagnostics of Sobol' indices with bootstrap-based confidence intervals, and rank stability assessments.

These revisions directly address the reviewer's concerns regarding the Discussion section, robustness analyses, and the influence of methodological choices. We believe that the revised manuscript has been substantially strengthened and is now clearer, more rigorous, and better aligned with the expectations of the hydrological modeling community.

Below, we provide a detailed, point-by-point response to the reviewer's comments. The reviewer's comments are presented in black, our responses in blue, and the corresponding changes made in the revised manuscript in red.

Main Comments:

Comment 1: Any sequential strategy, in which different methods are used in series, depends on early steps not overly constraining the outcome of later steps. For example, in this case, ensuring that parameters are not eliminated that might later become relevant when assessed in a distributed manner. How can it be ensured that this problem does not occur in this sequential application proposed here?

Response: We thank the reviewer for this important comment. We fully agree that any sequential sensitivity analysis strategy must demonstrate that the initial screening step does not unduly constrain the results of the subsequent distributed analysis. This concern is central to the credibility of the proposed two-step sensitivity analysis framework, and in the revised manuscript we have addressed it from three complementary perspectives: (1) the methodological rationale for using Morris screening prior to Sobol' -based distributed sensitivity analysis, (2) an explicit robustness test of the Morris screening results with different numbers of trajectories, and (3) an additional sensitivity analysis in which the second-stage Sobol' analysis was repeated with an expanded set of screened parameters.

First, we clarify more explicitly the theoretical justification for the sequential design. The Morris screening step was not intended as a strict or aggressive elimination procedure, but rather as a conservative filter used to reduce the dimensionality of an otherwise computationally prohibitive distributed Sobol' analysis. This choice is supported by previous studies showing that the Morris μ^* index and the Sobol' total-effect index S_{Ti} are conceptually related, in the sense that both reflect the overall influence of a parameter, including nonlinear and interaction effects (e.g., Campolongo et al., 2007; Saltelli et al., 2008; Wainwright et al., 2014; Dai et al., 2024). In the revised manuscript, we therefore emphasize that the Morris method is used as a qualitative but deliberately conservative pre-screening step prior to the Sobol' method, rather than as a definitive identification of all relevant controls.

Second, we have added a new robustness analysis to evaluate the stability of the Morris screening itself. Specifically, the Morris analysis was repeated using 100, 200, 300, 400, and 500 trajectories, and the results are now presented in Figure S3, which is copied below. These additional experiments show that the five parameters, i.e., CN2, CANMX, ESCO, CH_K1, and RCHRG_DP, remain consistently dominant across all trajectory settings. While minor variations occur among weakly sensitive parameters, the identity of the leading screened parameters does not change. This demonstrates that the screening outcome is not an artifact of a particular Morris design or of an insufficient number of trajectories. We have thus noted in the revised manuscript that the adopted setting of 500 trajectories is substantially larger than that used in many typical Morris applications, precisely to increase the robustness of the screening step.

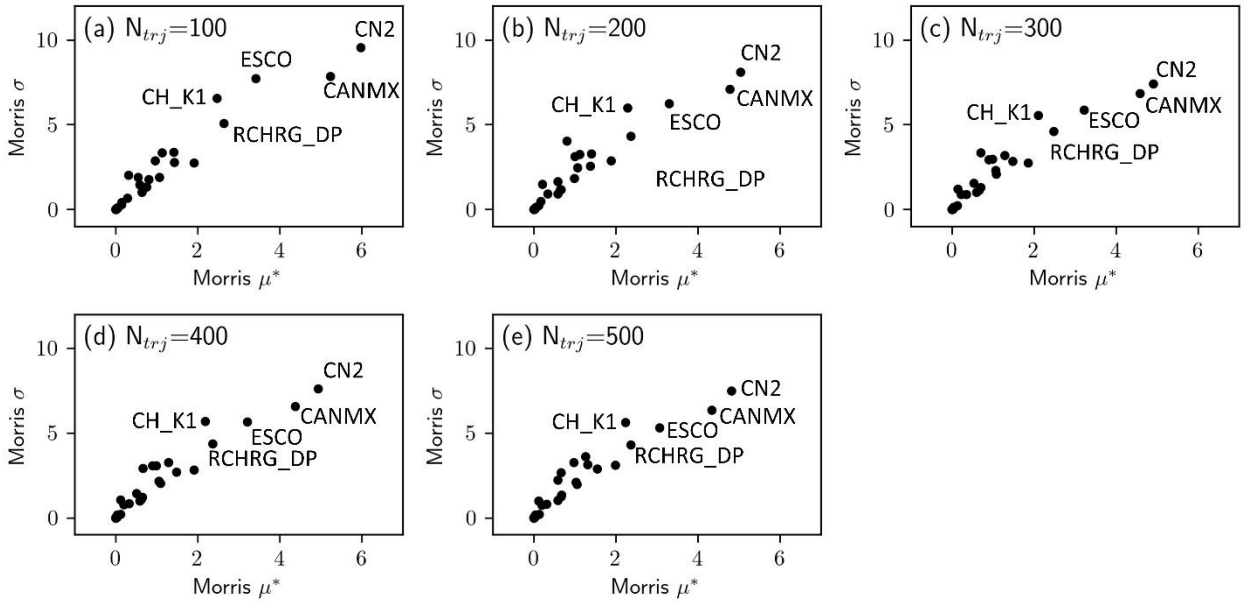

Figure S3: Morris sensitivity results obtained using different numbers of trajectories, illustrating the stability of five screened parameters across (a) 100, (b) 200, (c) 300, (d) 400, and (e) 500 trajectories.

Third, and most importantly, we now directly test the reviewer's concern by evaluating whether the second-stage distributed Sobol' results change when more parameters are retained after screening. In the revised manuscript, we added a new discussion section, Section 4.4 ("4.4 Effects of screened parameters on sensitivity patterns"), in which the Sobol'-based spatial sensitivity analysis was extended from the original five screened parameters to nine parameters. The four additionally included parameters, i.e., SOL_AWC, SOL_BD, GWQMN, and GW_DELAY, have lower Morris μ^* values than the original five, but higher values than the remaining candidates. Under this expanded setup, the number of distributed parameters increased from 195 to 351 at the subbasin scale and from 2,559 to 5,079 at the HRU scale. New MLP surrogate models were then trained, and the corresponding distributed Sobol' analysis was repeated. The results, now shown in Figure 9, indicate that including these additional moderately influential parameters redistributes variance contributions to some extent, but does not fundamentally alter the dominant spatial sensitivity patterns. In particular, the same major hotspot regions and the same primary controlling parameters identified in the original 5-parameter framework remain dominant. This provides direct evidence that the proposed sequential framework is robust, and that the main conclusions are not an artifact of overly restrictive screening in the first step.

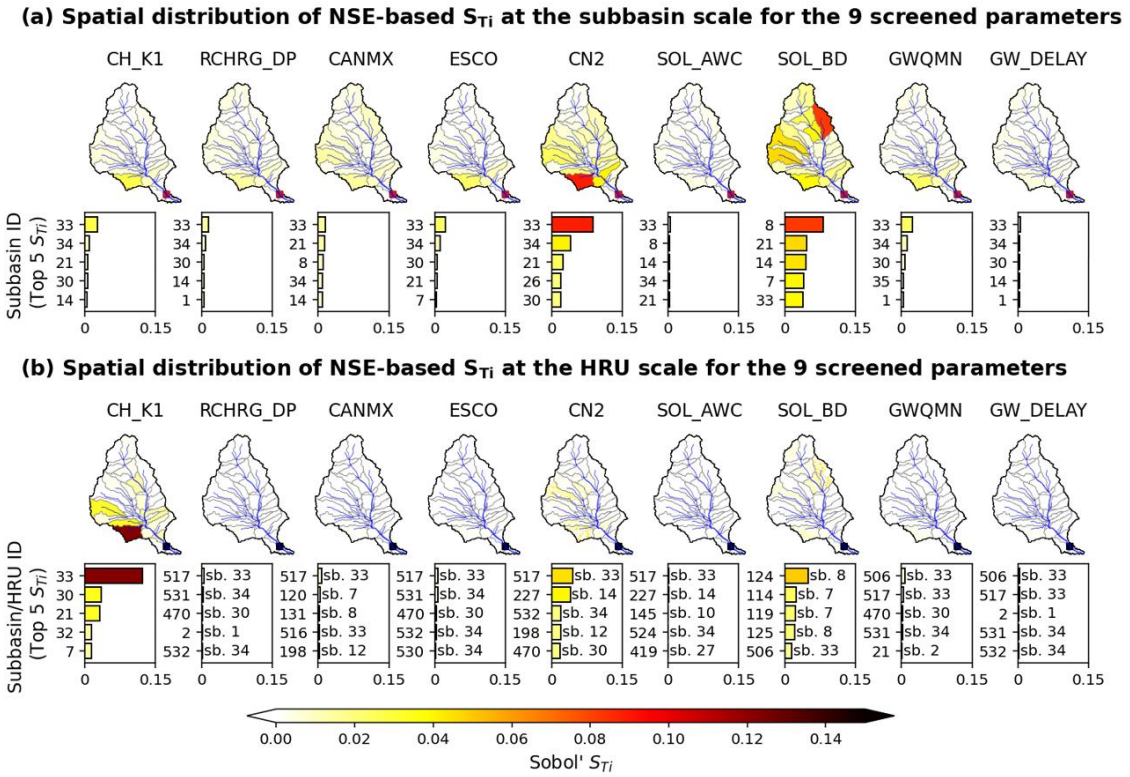

Figure 9: Spatial distribution of the total-effect sensitivity index S_{Ti} for the nine screened parameters at (a) the subbasin and (b) HRU scales.

At the same time, we also acknowledge in the revised Discussion that no sequential strategy can completely rule out the possibility that a weakly ranked parameter in a lumped screening step may become locally relevant under finer spatial discretization. We therefore clarify that the purpose of the proposed framework is to identify the dominant spatial controls on runoff response in a computationally feasible way, rather than to exhaustively enumerate every parameter that could become conditionally important in isolated locations. This limitation is now explicitly discussed in the manuscript.

References:

Campolongo, F., Cariboni, J., and Saltelli, A.: An effective screening design for sensitivity analysis of large models, Environmental Modelling & Software, 22, 1509-1518, https://doi.org/10.1016/j.envsoft.2006.10.004, 2007.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: Global Sensitivity Analysis. The Primer, https://doi.org/10.1002/9780470725184, 2007.

Wainwright, H. M., Finsterle, S., Jung, Y. J., Zhou, Q. L., and Birkholzer, J. T.: Making sense of global sensitivity analyses, Computers & Geosciences, 65, 84-94, 10.1016/j.cageo.2013.06.006, 2014.

Dai, H., Ju, J., Gui, D., Zhu, Y., Ye, M., liu, Y., Cui, J., and Hu, B. X.: A two-step Bayesian network-based process sensitivity analysis for complex nitrogen reactive transport modeling, Journal of Hydrology, 632, 130903, https://doi.org/10.1016/j.jhydrol.2024.130903, 2024.

Changes made in the revised manuscript:

- Revised the methodological description to clarify that the Morris step is used as a conservative qualitative filter before the quantitative Sobol' analysis.
- Added a robustness test of Morris screening with 100, 200, 300, 400, and 500 trajectories, now shown in Figure S3.
- Added a new discussion section, Section 4.4, and a new figure (Figure 9) presenting the Sobol' sensitivity analysis with nine screened parameters instead of five.
- Expanded the Discussion to explicitly acknowledge the limitation of sequential screening and to clarify the intended scope of the framework.

Comment 2: In this case study, have the authors tested whether the distributed sensitivity analysis is changing when different parameters are kept after the first stage? E.g. by trying to pursue step 2 with all parameters for smaller test cases?

Response: We thank the reviewer for this insightful and constructive comment. To address this concern, we evaluated whether the results of the distributed sensitivity analysis depend on the set of parameters retained after the first-stage Morris screening.

Following the reviewer's suggestion, we extended the spatial sensitivity analysis by including four additional parameters, SOL_AWC, SOL_BD, GWQMN, and GW_DELAY. This expansion increased the number of distributed parameters from 195 to 351 at the subbasin scale and from 2,559 to 5,079 at the HRU scale. New multilayer perceptron (MLP) surrogate models were trained, and the Sobol'-based spatial sensitivity analysis was repeated under these expanded parameter sets.

The results demonstrate that incorporating additional parameters leads to a redistribution of variance contributions but does not fundamentally alter the spatial sensitivity patterns. In particular, the dominant parameters and key sensitivity hotspots, especially those located near the basin outlet, remain consistent with those identified in the original five-parameter analysis. This confirms that the proposed two-step framework reliably captures the principal controls on model performance without being overly constrained by the initial screening step.

These new analyses have been incorporated into the revised manuscript as Section 4.4, "*Effects of screened parameters on sensitivity patterns*" and are illustrated in Figure 9. Together with the robustness tests presented in response to Comment 1, these results provide strong evidence for the reliability and stability of the proposed sequential sensitivity analysis framework.

Changes made in the revised manuscript:

- Added Section 4.4 and Figure 9 presenting the Sobol' sensitivity analysis using nine screened parameters instead of five.

Comment 3: In how far have the authors tested the "performance" of the MLP in terms of consistency of sensitive parameters?

Response: We thank the reviewer for this insightful and important comment. In the original submission, the performance of the MLP surrogate was primarily evaluated based on its ability to reproduce SWAT-simulated monthly runoff. It was implicitly assumed that accurate emulation of model outputs would also ensure consistency in the resulting sensitivity indices. Following the reviewer's suggestion, we have now explicitly verified this assumption by directly assessing the ability of the MLP surrogates to preserve the sensitivity structure of the original SWAT model.

Beyond predictive accuracy, the consistency of sensitivity patterns between the surrogate and the original model was evaluated using a reduced set of parameter realizations (N = 64). At both the subbasin and HRU scales, spatial sensitivity analysis based on the Sobol' method was performed using both the SWAT model and the corresponding MLP surrogates under identical parameter samples. This comparative analysis enabled a direct assessment of the agreement between surrogate-based and model-based sensitivity results.

The results are presented in Figures 3a5–3a6 and 3b5–3b6, which compare the Sobol' total-effect sensitivity indices S_{Ti} derived from SWAT outputs with those obtained from the MLP surrogates at the subbasin and HRU scales, respectively. The comparison shows strong agreement in both the magnitude and spatial distribution of dominant sensitivity signals, although minor differences in absolute values are observed. Importantly, the ranking and identification of the top 20% most influential parameters remain highly consistent (highlighted by white circles). These findings demonstrate that the MLP surrogates successfully preserve the relative importance and ranking of parameters, thereby providing confidence that the surrogate-based Sobol' analyses accurately capture the key controls of the original SWAT model.

The revised figure is copied below:

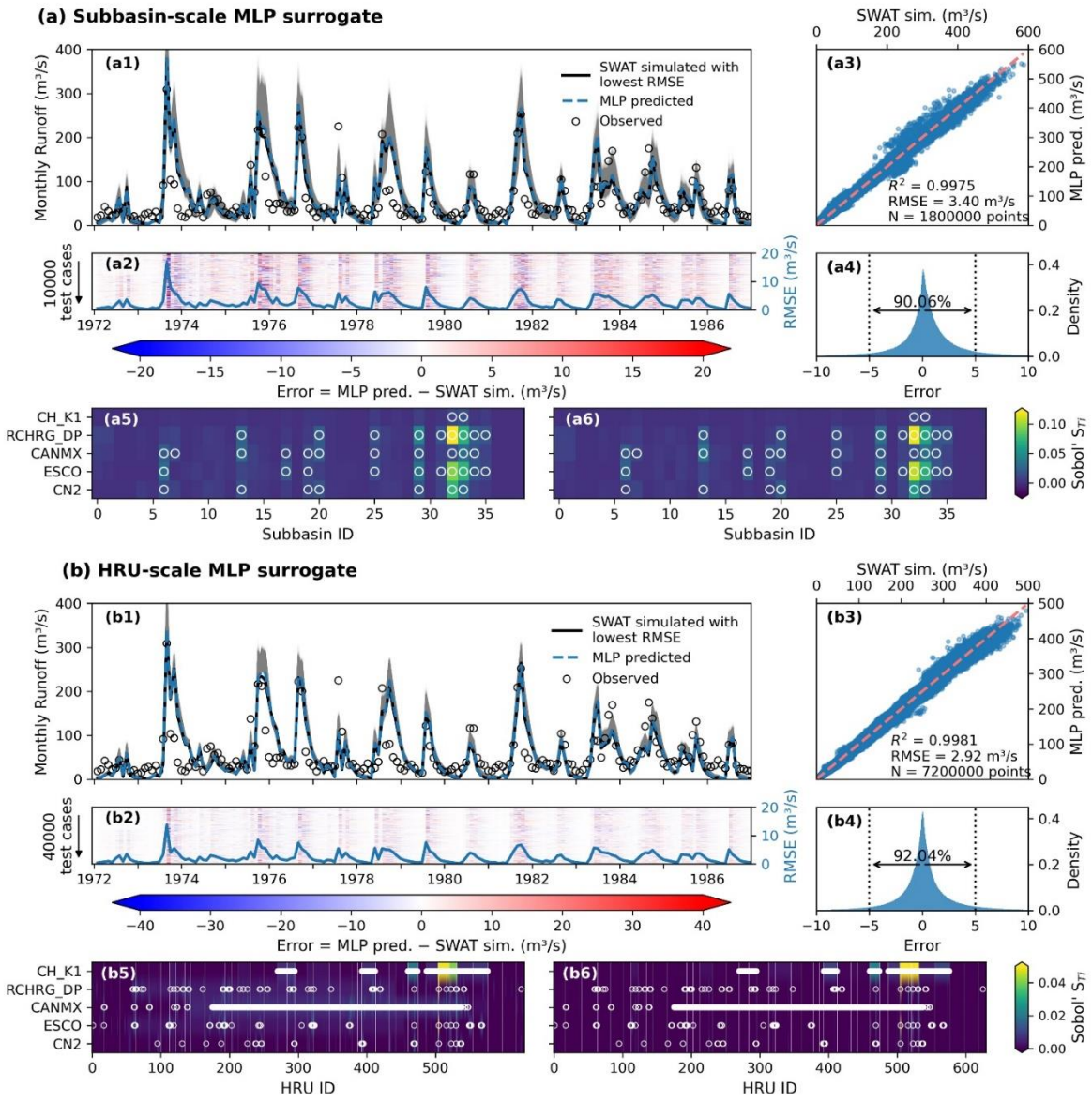

Figure 3: Performance of the trained MLP surrogates at the (a) subbasin and (b) HRU scales. Panels (a1) and (b1) compare ensemble SWAT simulations for the test cases (grey lines; 10,000 cases at the subbasin scale and 40,000 cases at the HRU scale), the specific SWAT-simulated runoff with lowest RMSE relative to observed runoff (black lines) and the corresponding MLP-predicted runoff (blue dashed lines), as well as observed runoff (black circles) at the ZJS gauge station. Panels (a2) and (b2) show heatmaps of prediction errors across 180 months from 1972 to 1986 for all test cases; blue solid lines indicate the RMSE values calculated for each month across realizations. Panels (a3) and (b3) present scatter plots of MLP-predicted versus SWAT-simulated runoff, showing strong agreement with global R2 values of 0.9975 and 0.9981, and global RMSE values of 3.40 $m^3\ s^{-1}$ and 2.92 $m^3\ s^{-1}$, respectively. Panels (a4) and (b4) show the error density, which further indicate that 90.06% and 92.04% of monthly prediction errors fell within ±5 $m^3\ s^{-1}$ for the subbasin- and HRU-scale MLP surrogates, respectively, confirming high predictive fidelity. Panels (a5) and (b5) show the S_{Ti} using SWAT model outputs with small parameter realizations N

= 64 for subbasin-scale SSA and HRU-scale SSA, respectively, and Panels (a6) and (b6) show the S_{Ti} using outputs from the corresponding MLP surrogates with the same parameter realizations. White circles denote the top 20% S_{Ti} values.

Changes made in the revised manuscript:

- Added a surrogate validation analysis assessing the consistency of sensitivity indices between SWAT and the MLP surrogates.
- Included new comparison plots in Figures 3a5–3a6 and 3b5–3b6.
- Revised Section 3.1 to explicitly describe the methodology and results of this validation.

Comment 4: Have the authors estimated confidence limits on the resulting sensitivity indices to ensure that their analyses have converged? It is difficult to assess the robustness of the results without such convergence tests. (e.g. Sarrazin et al. 2016, https://doi.org/10.1016/j.envsoft.2016.02.005)

Response: We thank the reviewer for this important comment and for highlighting the valuable work of Sarrazin et al. (2016). In the original submission, convergence of the Morris screening results was assessed by increasing the number of trajectories, and convergence of the Sobol' sensitivity indices was evaluated through stability analyses; however, these diagnostics were not explicitly reported. In the revised manuscript, we have incorporated a comprehensive convergence assessment and quantified the uncertainty associated with the sensitivity indices, following the recommendations of Sarrazin et al. (2016).

Specifically, the convergence of the Sobol' sensitivity results was evaluated by progressively increasing the sample size N. A bootstrap procedure with 100 replicates was employed to estimate the 95% confidence intervals of the first-order S_i and total-effect S_{Ti} indices. In addition, rank stability was assessed using Spearman's rank correlation coefficient by comparing the rankings of all parameters at each intermediate sample size with those derived from the final reference sample size. This analysis was conducted for all distributed parameters, including 195 parameters at the subbasin scale and 2,559 parameters at the HRU scale.

The results demonstrate clear convergence. The 95% bootstrap confidence intervals progressively narrow as the sample size increases, and parameter rankings stabilize with increasing N. In particular, rankings based on S_{Ti} become effectively identical beyond N = 512 for the subbasin-scale SSA and N = 2,048 for the HRU-scale SSA. These findings confirm that the identification of dominant parameters is robust with respect to the final reference sample sizes of N = 4,096 and N = 32,768, respectively. The convergence diagnostics thus provide strong evidence for the reliability and numerical stability of the sensitivity analysis.

These results are presented in Figure S4, which is copied below, and described in the revised manuscript, ensuring transparent verification of convergence and uncertainty in the estimated sensitivity indices.

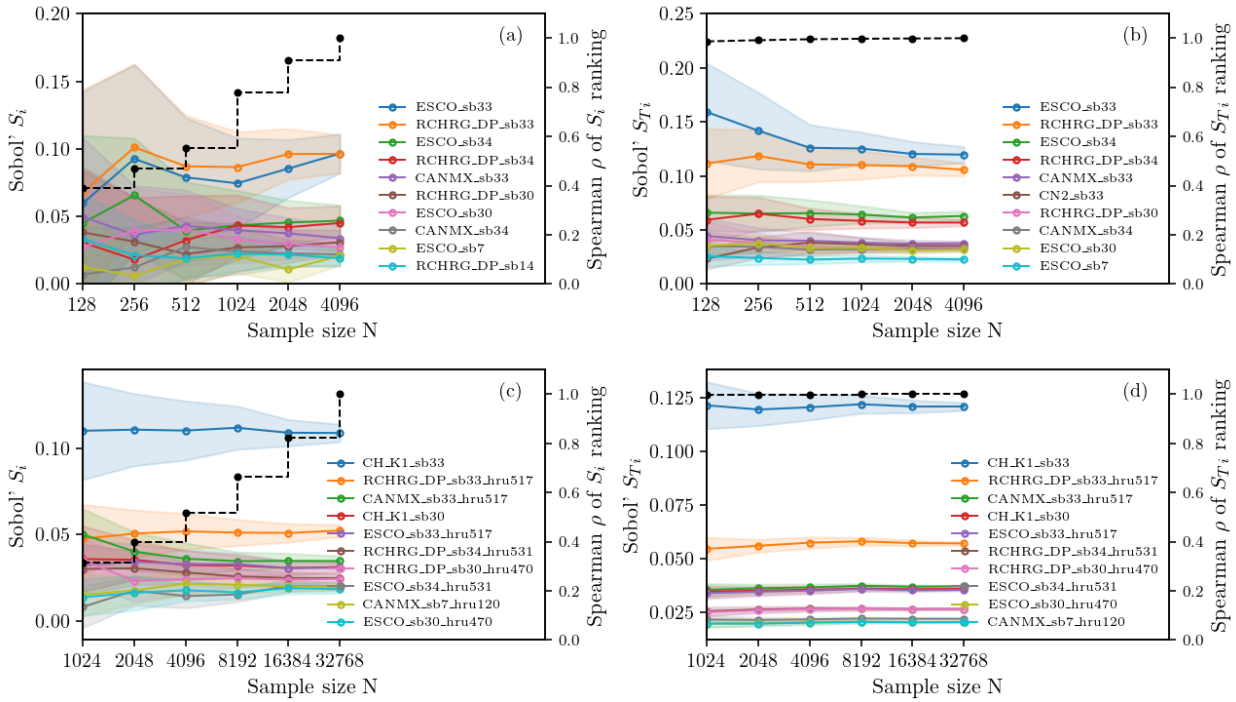

Figure S5: Convergence assessment of Sobol' sensitivity indices at subbasin and HRU scales. Panels (a) and (b) show the convergence of first-order S_i and total-effect S_{Ti} indices at the subbasin scale, respectively, while panels (c) and (d) present the corresponding results at the HRU scale. Colored lines represent the evolution of sensitivity indices for the ten most influential parameters as a function of sample size N and shaded areas denote the associated 95% bootstrap confidence intervals. Black dashed step lines indicate the stability of parameter rankings, quantified using the Spearman rank correlation coefficient relative to the largest sample size (N = 4,096 for the subbasin scale and N = 32,768 for the HRU scale), computed across all parameters (195 at the subbasin scale and 2,559 at the HRU scale). For clarity, only the top ten parameters ranked by S_i and S_{Ti} are shown.

Changes made in the revised manuscript:

- Added convergence diagnostics for Sobol' sensitivity indices using progressive sample sizes.
- Estimated 95% confidence intervals through bootstrap resampling with 100 replicates, following Sarrazin et al. (2016).
- Evaluated rank stability using Spearman's rank correlation coefficient,
- Included new results in Figure S4 and described them in Section 3.

Comment 5. Why do you see such strong differences in results for sub-basins and HRUs? If it is potentially due to the different number of parameters, then can this not be tested and confirmed?

Response: We thank the reviewer for this insightful comment. We agree that the pronounced differences between the subbasin- and HRU-scale sensitivity results warrant further clarification. In the original manuscript, we hypothesized that these discrepancies might be primarily

attributable to the increased dimensionality of the parameter space at the HRU scale. However, in the revised manuscript, we conducted additional analyses to explicitly test this assumption and to provide a more comprehensive explanation.

First, we evaluated whether the observed differences arise solely from the increased number of distributed parameters by performing spatial sensitivity analyses using alternative performance metrics. In addition to the NSE, we conducted SSA using PBIAS and KGE. The results, presented in Figure 6, which is copied below, reveal that the differences between subbasin- and HRU-scale sensitivity patterns become less pronounced under PBIAS and KGE, indicating that parameter dimensionality alone does not fully explain the observed discrepancies.

A notable exception is observed for the channel routing parameter CH_K1. Under NSE, CH_K1 exhibits substantially higher sensitivity at the HRU scale, whereas its influence becomes considerably weaker when evaluated using PBIAS and KGE. This discrepancy cannot be attributed to spatial aggregation, as CH_K1 is consistently defined at the subbasin level in all analyses. Instead, it reflects the distinct aspects of model behavior emphasized by each performance metric. CH_K1 primarily governs flow routing processes, influencing the timing and attenuation of streamflow. NSE is particularly sensitive to such dynamics because it strongly penalizes errors in peak flow and temporal variability. In contrast, PBIAS measures cumulative bias and is largely insensitive to timing errors, while KGE, although incorporating a correlation component, balances multiple performance aspects, thereby reducing the relative influence of timing-related discrepancies. Consequently, the sensitivity of CH_K1 is significantly diminished under PBIAS and KGE.

These findings demonstrate that the differences between subbasin- and HRU-scale sensitivity results arise from a combination of factors, including parameter dimensionality, spatial discretization, and the specific aspects of hydrological behavior emphasized by different performance metrics. Accordingly, we have revised the manuscript to provide a more nuanced interpretation of scale-dependent sensitivity patterns. Nevertheless, we acknowledge that fully disentangling these effects requires further investigation and remains an important direction for future research.

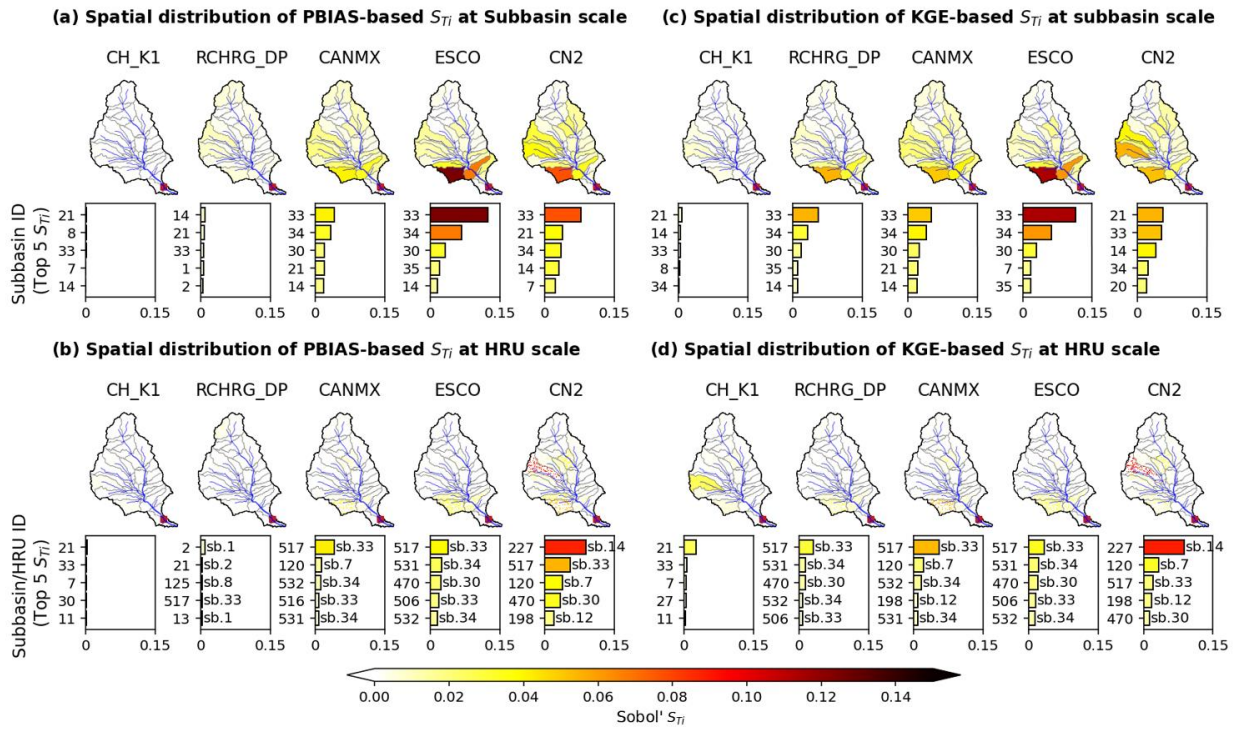

Figure 6: Spatial distribution of the total-effect sensitivity index S_{Ti} at the subbasin and HRU scales using different performance metrics. Panels (a) and (b) show results based on PBIAS, and panels (c) and (d) show results based on KGE.

Changes made in the revised manuscript:

- Added additional spatial sensitivity analyses using PBIAS and KGE as alternative performance metrics.
- Included new results and comparisons in Figure 6.
- Expanded the Discussion to explain the scale-dependent sensitivity of CH_K1 and the influence of different objective functions.
- Clarified those differences between subbasin- and HRU-scale results are not solely attributable to parameter dimensionality.

Comment 6. Please check again for spelling issues. E.g. in the caption of Fig. 5 “show lagged Spearman’s rank correlations (r) between and runoff”, there is a word missing after between.

Reply: We sincerely thank the reviewer for carefully identifying this error and for highlighting the need for a thorough language check. We apologize for the oversight.

To improve the focus and clarity of the study, we have removed the time-varying sensitivity analysis and now concentrate exclusively on spatial sensitivity analysis. Consequently, the figure containing the reported error, as well as all related descriptions, has been removed from the revised manuscript. In addition, the entire paper has undergone a comprehensive proofreading to correct spelling, grammatical, and typographical errors.

Changes made in the revised manuscript:

- Removed the time-varying sensitivity analysis and the associated figure and captions.
- Thoroughly proofread the manuscript to ensure linguistic accuracy and clarity.

Comment 7. The Discussion section is a good start, but it is currently not fulfilling its actual role. It is meant to discuss the results of this specific study in the context of previous studies. However, section 4.1 just reviews the results, while section 4.2 makes some references to potential future explorations. So, the current 4.1 should be part of the results section. In section 4, the authors need to discuss how their findings different (or not) from previous findings regarding the sensitivity of the SWAT parameters. Did they find new influences of processes compared to other studies? Did the different approach yield different results? Etc. Also, what did the authors find in their methodology compared to previous space-time varying analyses? The authors made some different choices and assumptions, how did this influence the results and findings?

Response: We thank the reviewer for this insightful and constructive comment. We fully agree that the Discussion section should interpret the results within the context of existing literature rather than reiterate findings. Accordingly, the manuscript has been substantially revised to improve its structure, clarity, and scholarly depth.

First, the original Section 4.1, which primarily summarized sensitivity hotspots and hot moments, has been relocated to the Results section as Section 3.3. In addition, as noted in our response to Comment 6, all time-varying sensitivity analyses have been removed to enhance the focus of the study on spatial sensitivity patterns.

Second, the Discussion section has been comprehensively restructured to emphasize interpretation, comparison with prior studies, and methodological insights. In the revised manuscript, Section 4 now comprises the following four subsections:

- **Section 4.1**: Effects of performance metrics on sensitivity patterns
- **Section 4.2**: Effects of parameter distributions on sensitivity patterns
- **Section 4.3**: Effects of including an upstream gauge on sensitivity patterns
- **Section 4.4**: Effects of screened parameters on sensitivity patterns

These additions enable a systematic evaluation of the robustness of the proposed framework and provide a deeper interpretation of the results.

Third, the revised Discussion explicitly compares our findings with previous SWAT sensitivity studies. The dominant parameters identified, such as CN2, ESCO, and RCHRG_DP, are consistent with established literature, confirming the reliability of the approach. At the same time, the cross-scale analyses at the subbasin and HRU levels provide new insights into spatial heterogeneity and sensitivity hotspots that are seldom explored in earlier studies. Furthermore, the proposed two-step, deep learning-assisted framework offers methodological advances in

handling high-dimensional distributed parameters, distinguishing this study from traditional lumped or basin-scale analyses.

These revisions ensure that the Discussion fulfills its intended role by situating the findings within the broader scientific context, highlighting consistencies and novelties relative to previous research, and clarifying the implications of methodological choices.

Changes made in the revised manuscript:

1. The original Section 4.1 has been moved to the Results section as Section 3.3.
2. All time-varying sensitivity analyses have been removed to focus on spatial sensitivity analysis.
3. The Discussion has been reorganized into four new subsections (Sections 4.1–4.4).
4. Comparisons with previous SWAT sensitivity studies have been added to contextualize the findings and highlight methodological contributions.

Reply to RC2's Comment

General Assessment

The paper tackles an important problem, but several method choices and interpretations need improvement, and some conclusions appear over-claimed. I would recommend major revision.

Response: We sincerely thank the reviewer for the careful evaluation and constructive recommendation. We appreciate the recognition of the importance of this research and fully acknowledge the need to strengthen methodological justification, improve interpretation, and avoid overstatement of conclusions. In response, we have thoroughly revised the manuscript to enhance its clarity, rigor, and scientific credibility. The major revisions are summarized as follows:

(1) To enhance focus, time-varying sensitivity analyses have been removed, and the manuscript now concentrates on spatial sensitivity patterns. The new title now reads "Identifying Dominant Parameters in SWAT Across Subbasin and HRU Scales Using a Two-Step Deep Learning-Assisted Spatial Sensitivity Analysis". The manuscript has been thoroughly proofread to correct typographical and grammatical errors and improve overall readability.

(2) We conducted additional robustness and convergence analyses, including bootstrap-based confidence intervals, rank stability assessments, and posterior-constrained sensitivity analyses, to ensure the reliability of the Sobol' sensitivity results.

(3) Detailed descriptions of spatial parameterization strategies, parameter perturbation methods, and SWAT input file modifications have been added. The distinctions between the "replace" and "factor" approaches are now clearly explained, improving methodological transparency and reproducibility.

(4) Additional experiments were conducted to evaluate the effects of parameter distributions, performance metrics (NSE, PBIAS, and KGE), and observation locations, including the incorporation of an upstream gauge and multi-site constraints.

These revisions directly address the reviewer's concerns regarding methodological rigor, interpretation, and validity of conclusions. We believe that the revised manuscript has been substantially strengthened and now meets the standards required for publication. A detailed, point-by-point response to all comments is provided below. The reviewer's comments are presented in black, our responses in blue, and the corresponding changes made in the revised manuscript in red.

Major Comments

Comment 1: The authors state that no model calibration was performed to maintain "diagnostic integrity". While this isolates parameter sensitivity from calibration bias, it risks performing SA on a model that does not represent the physical reality of the Jinghe River Basin. Will sensitivity patterns change significantly once the model is constrained into a realistic posterior parameter space? You might want to report baseline SWAT performance vs observations and discuss how poor/mediocre fit would distort SA.

Response: We sincerely thank the reviewer for this insightful and important comment. We agree that assessing the influence of parameter distributions and model performance on sensitivity analysis is essential to ensure the physical relevance and robustness of the results.

In the original submission, sensitivity analysis was conducted using prior parameter ranges derived from SWAT documentation and previous studies, with the aim of diagnosing dominant process controls within a broadly defined feasible parameter space. However, we acknowledge that the absence of calibration may limit the interpretation of the results as basin-specific physical controls. To address this concern, we have undertaken additional analyses and revised the manuscript accordingly.

First, we evaluated baseline SWAT performance. The default parameterization produced poor agreement with observed runoff, with a full-period NSE of −6.0. We further examined the ensemble of 17,000 simulations generated during the Morris screening stage. As reported in our previous response, the explored prior parameter space contains behavioral parameter sets capable of reproducing observed runoff reasonably well, with the best-performing simulation achieving an NSE of 0.70. This demonstrates that the sensitivity analysis was conducted within a feasible parameter space capable of representing realistic basin behavior.

To further evaluate the influence of parameter realism on sensitivity patterns, we conducted an additional spatial sensitivity analysis using NSE-constrained posterior parameter distributions. This new analysis has been incorporated into the revised manuscript as Section 4.2, "Effects of parameter distributions on sensitivity patterns." In this approach, the original prior distributions were conditioned by retaining only parameter realizations associated with acceptable model performance. Specifically, NSE values were computed for 50,000 SWAT simulations at the subbasin scale and 200,000 simulations at the HRU scale. The results are shown in Figure S7 and also copied below. Two performance thresholds (NSE > 0.1 and NSE > 0.3) were applied to derive behavioral ensembles.

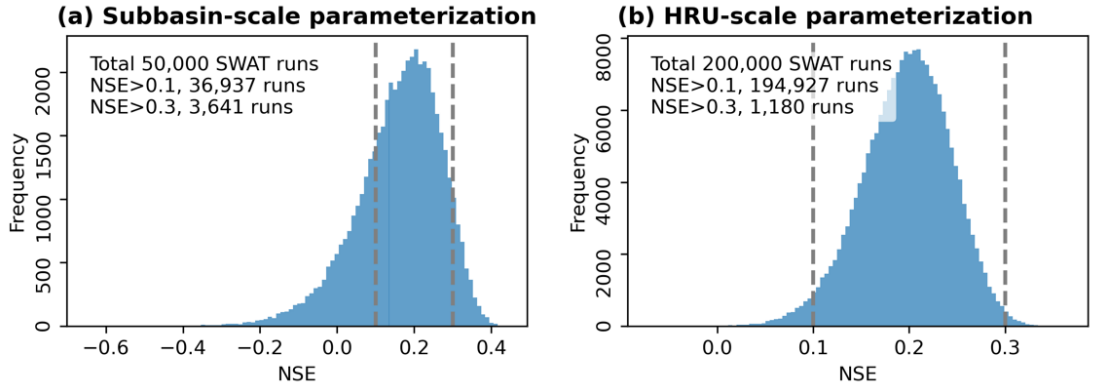

Figure S8: Histograms of NSE calculated from (a) 50,000 SWAT simulations used to construct the subbasin-scale MLP surrogate and (b) 200,000 SWAT simulations used to construct the HRU-scale MLP surrogate. The two vertical grey dashed lines indicate the NSE thresholds of 0.1 and 0.3, respectively.

Comparisons between prior and posterior parameter distributions indicate that most parameters exhibit only minor deviations, as evidenced by low Kolmogorov–Smirnov (KS) statistics (Figure S8, which is copied below). Even under the stricter constraint (NSE > 0.3), over 92% of parameters at the subbasin scale and 99% at the HRU scale show KS values below 0.1, indicating minimal distributional changes.

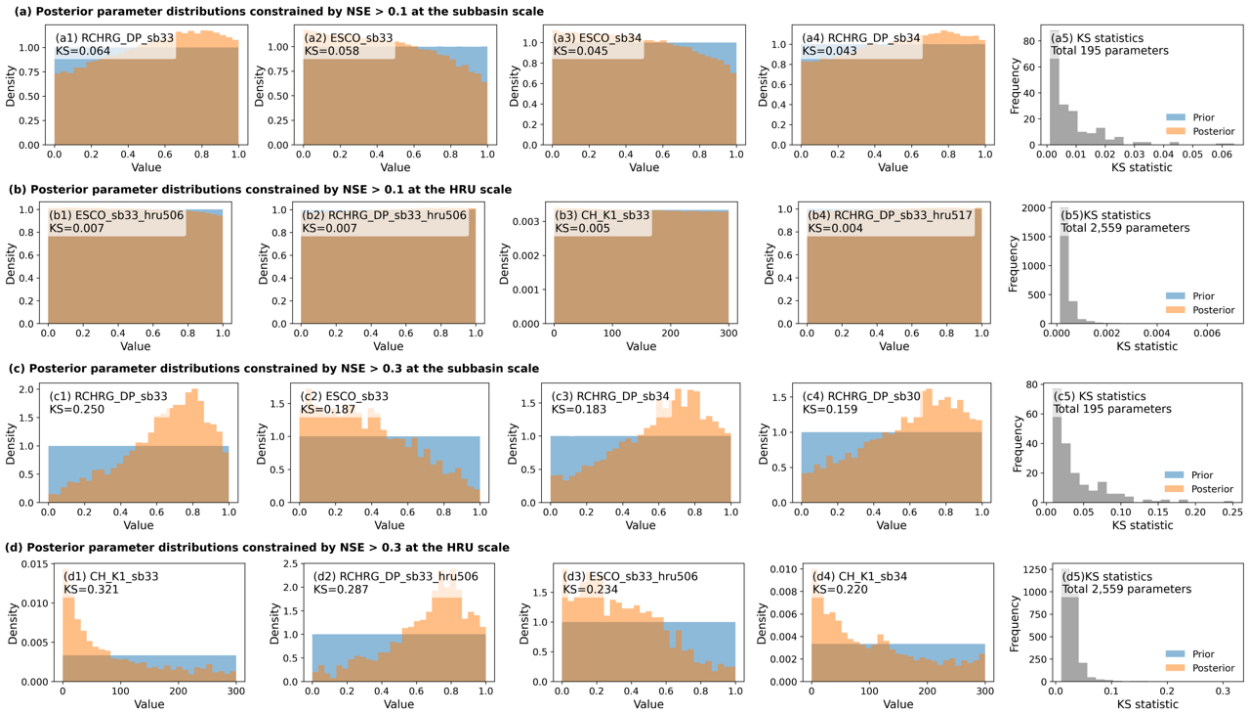

Figure S9: Comparison of the prior and posterior distributions of the four parameters exhibiting the largest Kolmogorov–Smirnov (KS) statistics, together with histograms of KS values across all distributed parameters constrained by (a) NSE > 0.1 at the subbasin scale, (b) NSE > 0.1 at the HRU scale, (c) NSE > 0.3 at the subbasin scale and (d) NSE > 0.3 at the HRU scale. KS values close to 0 indicate minimal deviation between the prior and posterior distributions, whereas values approaching 1 signify strong posterior constraints.

Based on these posterior distributions, the MLP-assisted the Sobol' spatial sensitivity analysis was repeated. The resulting sensitivity maps (Figure 7, which is copied below) demonstrate that the principal spatial hotspots and dominant parameters remain largely unchanged. Differences are mainly observed in the magnitude of the Sobol' indices rather than in their spatial organization.

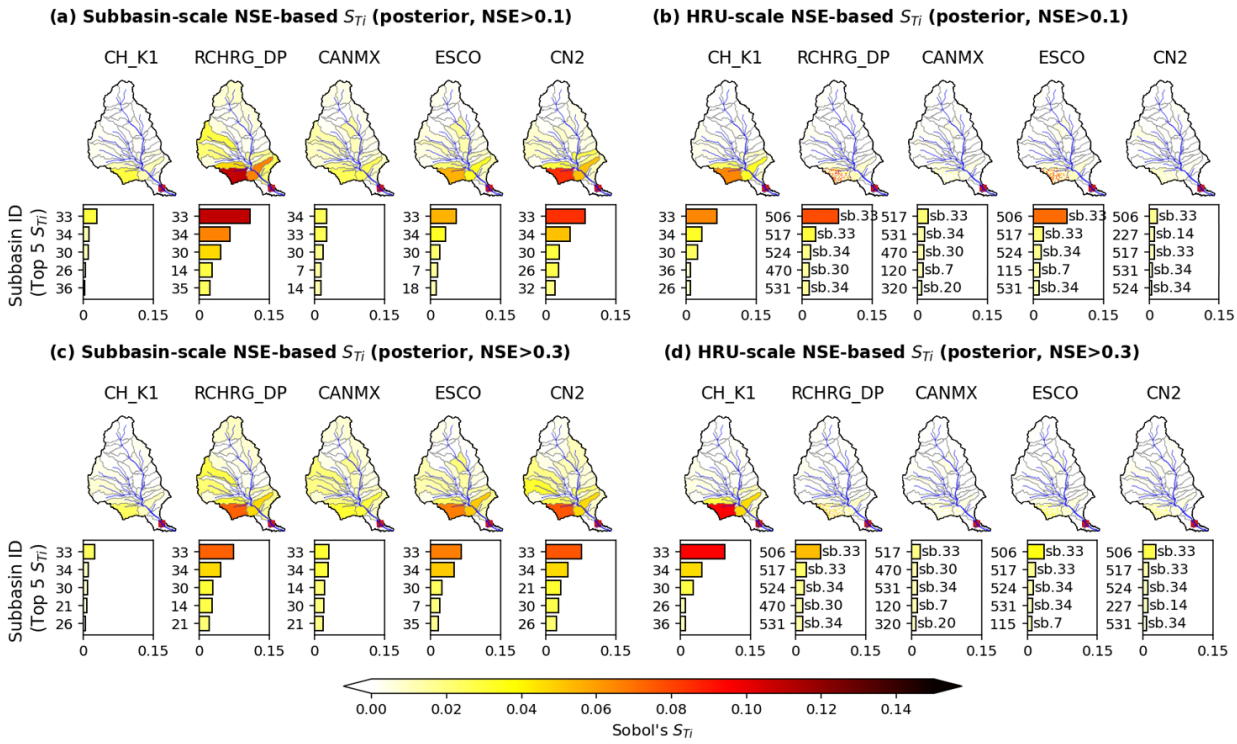

Figure 7: Spatial distribution of the total-effect sensitivity index S_{Ti} at the subbasin and HRU scales using posterior parameter distributions constrained by different NSE threshold. Panels (a) and (b) show results constrained by NSE > 0.1, and panels (c) and (d) show results constrained by NSE > 0.3.

These results indicate that constraining the parameter space to behaviorally acceptable regions affects the absolute sensitivity magnitudes more than the spatial patterns themselves. Consequently, the identified sensitivity hotspots are not artifacts of broad prior parameter ranges but reflect persistent structural characteristics of the SWAT model and the hydrological behavior of the Jinghe River Basin. Overall, this additional analysis confirms that the inferred spatial sensitivity patterns are robust to moderate posterior conditioning. At the same time, it underscores the importance of incorporating parameter realism when interpreting quantitative sensitivity estimates.

Changes made in the revised manuscript:

- Introduced a new subsection, **Section 4.2: "Effects of parameter distributions on sensitivity patterns."**
- Conducted additional SSA using NSE-constrained posterior parameter distributions.
- Included new supporting figures illustrating NSE distributions, KS statistics, and posterior sensitivity patterns (Figures S7–S8 and Figure 7).

Comment 2: The author chose fully connected MLPs over sequence models like LSTMs arguing that the temporal structure is "encoded in the output vector". However, hydrologic processes are inherently autoregressive. Would not a 180-neuron output layer treat each month runoff as an

independent regression target and potentially ignoring the temporal dependencies and mass balance continuity inherent in the SWAT model?

Response: We thank the reviewer for this insightful comment. We agree that hydrologic processes are inherently autoregressive and that sequence models such as LSTMs are well suited to represent temporal dependencies. However, in this study the surrogate modeling task is formulated as a deterministic vector-to-vector regression problem, where a static parameter vector is mapped to the complete 180-month runoff series simulated by SWAT.

Because all model parameters are time-invariant and the meteorological forcing remains identical across parameter samples, the objective is not to predict runoff sequentially in time but to approximate the functional relationship between parameter configurations and the resulting system response. Consequently, the 180-neuron output layer represents a single structured response rather than independent monthly targets. The temporal dependencies and mass balance constraints inherent in SWAT simulations are implicitly embedded in the training data and learned jointly by the MLP.

Furthermore, the MLP surrogates demonstrated high predictive accuracy and successfully preserved the sensitivity structure of the SWAT model, confirming their suitability for the the Sobol'-based sensitivity analysis. Nevertheless, we acknowledge that sequence models such as LSTMs may provide a more explicit representation of temporal dynamics and could be explored in future studies.

Changes made in the revised manuscript:

- Clarified the rationale for adopting fully connected MLPs in Section 2.5.2.
- Explicitly stated that the surrogate approximates a vector-to-vector mapping from static parameters to the full runoff time series.
- Added a discussion on the applicability of sequence models such as LSTMs.

Comment 3: You find hotspot subbasins near the gauge (e.g., 33–34) dominate sensitivities. This is unsurprising when the response is a single-gauge performance metric and proximity/connectivity to the gauge will mechanically increase leverage as briefly acknowledged by the authors. However, the authors interpret patterns as reflecting spatial heterogeneity (land use/soil/topography). Would hotspot locations remain under alternative responses and/or multi-site constraints? You might want to discuss this including the effect of additional gauges, internal variables (ET, soil water, baseflow index), or alternative spatially distributed responses.

Response: We sincerely thank the reviewer for this insightful comment. We agree that the use of a single-gauge performance metric may introduce a spatial leverage effect, whereby subbasins closer to the outlet exert a disproportionate influence on sensitivity patterns. To address this concern, we have conducted additional analyses by incorporating an upper stream gauge and evaluating multi-site constraints.

First, we considered only the upper stream YJP station, located in the middle reach of the basin. A Morris screening analysis was performed using runoff observations at YJP, which identified the same five dominant parameters as those derived from the outlet-based ZJS analysis (Figure S9).

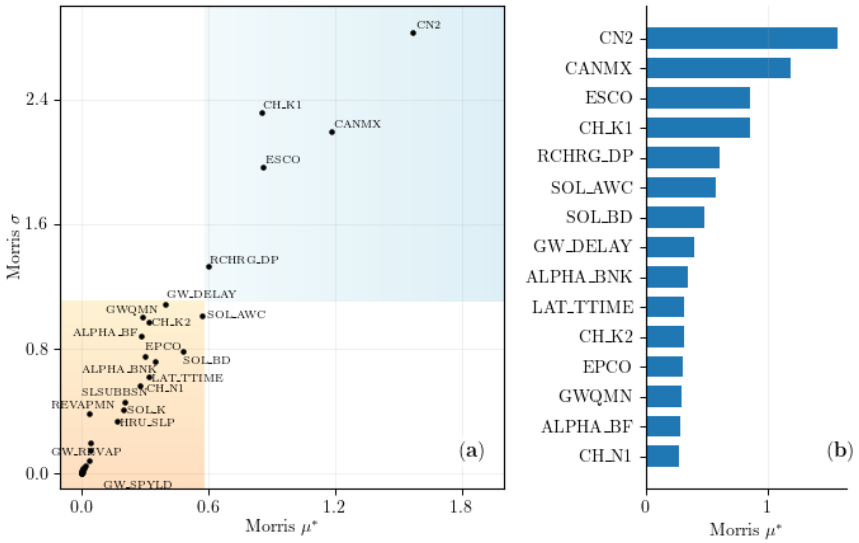

Figure S10: Sensitivity analysis results by Morris screening method for NSE-based runoff simulations at YJP gauge station: (a) scatter plot of the mean absolute elementary effect (μ*) versus its standard deviation (σ) for 33 selected parameters. (b) bar chart ranking the top 15 parameters by μ*.

This result indicates that parameter influence is largely consistent across observation locations. Based on these screened parameters, a new MLP surrogate was constructed and trained for the YJP station, and Sobol'-based SSA was subsequently conducted. The results show that sensitivity hotspots shift toward the vicinity of the YJP station, while parameters located downstream exhibit near-zero sensitivity, confirming that sensitivity patterns are constrained by hydrological connectivity and flow direction (Figure 8).

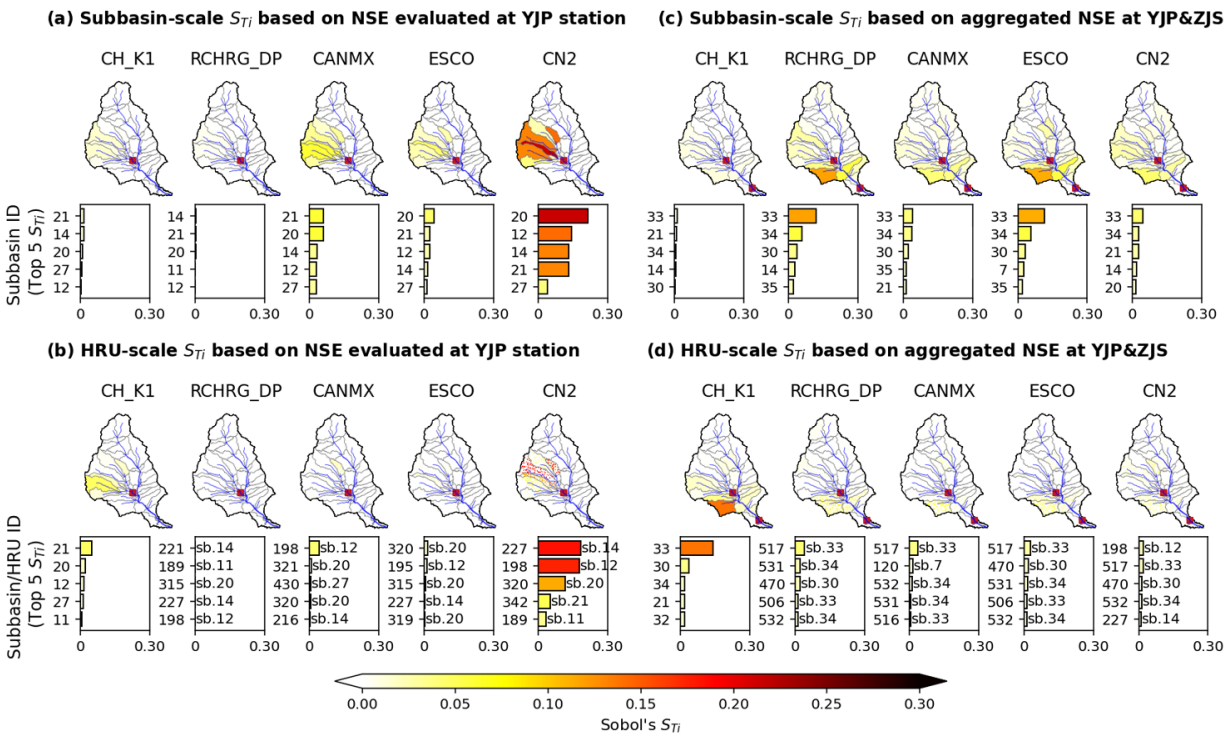

Figure 8: Spatial distribution of Sobol' total-effect sensitivity index S_{Ti} at the subbasin and HRU scales using. Panels (a) and (b) show results based on NSE evaluated at the YJP gauge station, and panels (c) and (d) show results based on aggregated NSE evaluated across YJP and ZJS stations.

Next, we evaluated a multi-site constraint by employing an aggregated NSE that combines observations from both the upstream YJP and downstream ZJS stations. Morris screening under this combined objective again identified the same five dominant parameters (Figure S11). A corresponding MLP surrogate was then reconstructed and used to perform Sobol'-based SSA. The resulting sensitivity patterns remain broadly consistent with those obtained using the outlet-based objective (Figure 8). However, due to its integrative nature, the downstream ZJS station exerts a stronger influence, leading to intermediate sensitivity magnitudes under the combined objective.

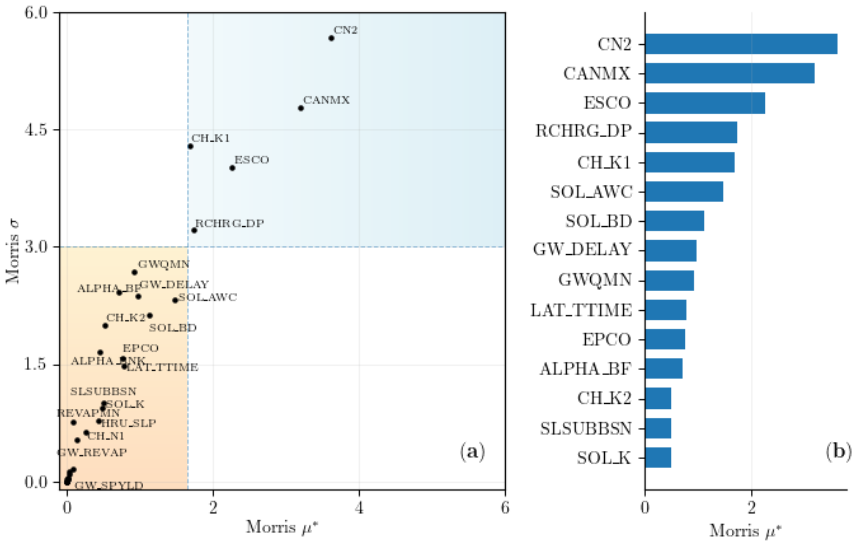

Figure S12: Sensitivity analysis results by Morris screening method for aggregated NSE across the YJP and ZJS stations: (a) scatter plot of the mean absolute elementary effect (μ*) versus its standard deviation (σ) for 33 selected parameters. (b) bar chart ranking the top 15 parameters by μ*.

Collectively, these analyses demonstrate that while hotspot locations are influenced by observation placement, they are fundamentally governed by basin structure and hydrological connectivity. They also confirm that the identified sensitivity hotspots are not merely artifacts of a single-gauge objective function but reflect persistent spatial controls within the watershed. At the same time, the results highlight that multi-site constraints do not necessarily balance sensitivity contributions, as downstream gauges tend to dominate owing to their larger contributing areas.

These additional analyses have been incorporated into the revised manuscript as Section 4.3, "Effects of including an upper stream gauge on sensitivity patterns" where their implications are discussed in detail.

Changes made in the revised manuscript:

- Added a new subsection, **Section 4.3: "Effects of including an upper stream gauge on sensitivity patterns" .**
- Incorporated an upstream gauge (YJP) to evaluate the influence of observation location on sensitivity patterns.
- Introduced a multi-site aggregated NSE objective function and conducted additional Morris and Sobol' analyses.
- Added new results and figures **(Figures 8, S9–S11)** illustrating sensitivity patterns under single-site and multi-site constraints.

Comment 4: For the HRU-scale Sobol' analysis the authors used k = 2559 and N = 32768. The authors noted that sensitivity magnitudes (S_T) are smaller at this scale due to "variance dilution" Is the identified sensitivity a physical signal or a mathematical artifact of the Sobol' method when the input space is massively expanded? With thousands of parameters, many S_T estimates can be

noisy/biased; negative indices and non-closure can occur and should be diagnosed. You might want to include convergence checks including reporting fraction of negative indices and demonstrating stability of rankings/hotspots. Otherwise, soften your claims and frame findings as exploratory.

Response: We sincerely thank the reviewer for this important and insightful comment. We agree that Sobol' sensitivity analysis in high-dimensional parameter spaces requires rigorous diagnostics to ensure that the results represent meaningful physical signals rather than mathematical artifacts. To address this concern, we have incorporated a comprehensive convergence and robustness assessment in the revised manuscript, following the recommendations of Sarrazin et al. (2016).

Specifically, the convergence of the Sobol' sensitivity indices was evaluated by progressively increasing the sample size N. A bootstrap procedure with 100 replicates was employed to estimate the 95% confidence intervals of the first-order S_i and total-effect S_{Ti} indices. In addition, rank stability was assessed using Spearman's rank correlation coefficient by comparing the rankings of all parameters at each intermediate sample size with those derived from the final reference sample size (Figure S4, which is copied below). This analysis was conducted for all distributed parameters, including 195 parameters at the subbasin scale and 2,559 parameters at the HRU scale.

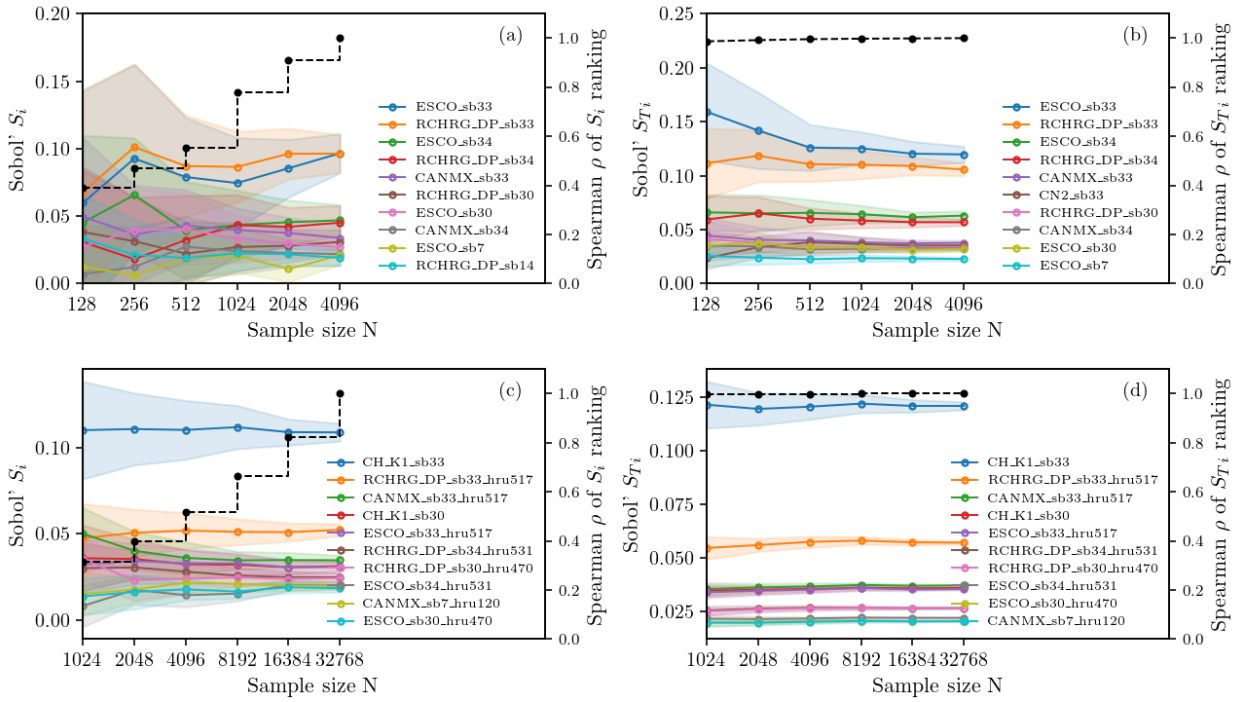

Figure S5: Convergence assessment of Sobol' sensitivity indices at subbasin and HRU scales. Panels (a) and (b) show the convergence of first-order S_i and total-effect S_{Ti} indices at the subbasin scale, respectively, while panels (c) and (d) present the corresponding results at the HRU scale. Colored lines represent the evolution of sensitivity indices for the ten most influential parameters as a function of sample size N and shaded areas denote the associated 95% bootstrap confidence intervals. Black dashed step lines indicate the stability of parameter rankings, quantified using the Spearman rank correlation coefficient relative to the largest sample size (N = 4,096 for the subbasin scale and N = 32,768 for the HRU scale), computed across all parameters (195 at the subbasin

scale and 2,559 at the HRU scale). For clarity, only the top ten parameters ranked by S_i and S_{Ti} are shown.

Changes Made in the Revised Manuscript:

- Added convergence diagnostics for Sobol' sensitivity indices using progressively increasing sample sizes.
- Estimated 95% confidence intervals through bootstrap resampling with 100 replicates, following Sarrazin et al. (2016).
- Evaluated rank stability using Spearman's rank correlation coefficient.
- Included new results in Figure S4 and described them in Section 3 and the Supporting Information.

Comment 5: The use of NSE as the sole sensitivity target can be problematic, especially for rolling windows. NSE is highly sensitive to variance and peaks, can behave poorly for low flows, and is nonlinearly bounded; windowed NSE can become unstable for low-variance windows. You mention this as a limitation, but the paper's main results rely on it. You might want to include at least one complementary metric and a compare whether dominant parameters/hot moments persist across metrics. Otherwise, the "hot moments" identified (e.g., CN2 peaking during wet months) may reflect the flaws or the mathematical structure of NSE rather than the shift in the physical process.

Response: We sincerely thank the reviewer for this insightful comment. We agree that relying solely on NSE as the sensitivity target may introduce metric-dependent biases, particularly due to its sensitivity to peak flows and variance. To address this concern, we conducted additional analyses using two complementary performance metrics, Percent Bias (PBIAS) and Kling–Gupta Efficiency (KGE), and incorporated the results into the revised manuscript.

Specifically, we introduced a new subsection, Section 4.1: "Effects of performance metrics on sensitivity patterns" These metrics provide distinct yet complementary perspectives on hydrological model performance: PBIAS quantifies systematic bias in water balance, while KGE offers a balanced assessment by integrating correlation, bias, and variability. Morris screening results based on PBIAS and KGE exhibit some differences in parameter rankings compared with NSE. However, all parameters identified as sensitive under NSE remain within the top-ranked subset identified by the alternative metrics, indicating that the core set of influential parameters is robust to the choice of performance metric.

To ensure direct comparability, the five parameters screened using NSE were retained for spatial sensitivity analysis, with PBIAS and KGE replacing NSE as response metrics. The trained multilayer perceptron (MLP) surrogates were reused to compute these metrics, eliminating the need for retraining. The resulting Sobol' sensitivity maps (Figure 6, which is copied below) demonstrate that the overall spatial patterns are highly consistent across metrics at both the subbasin and HRU scales. Sensitivity hotspots remain concentrated in key subbasins near the watershed outlet, confirming that these patterns are governed primarily by hydrological connectivity rather than by the mathematical structure of NSE.

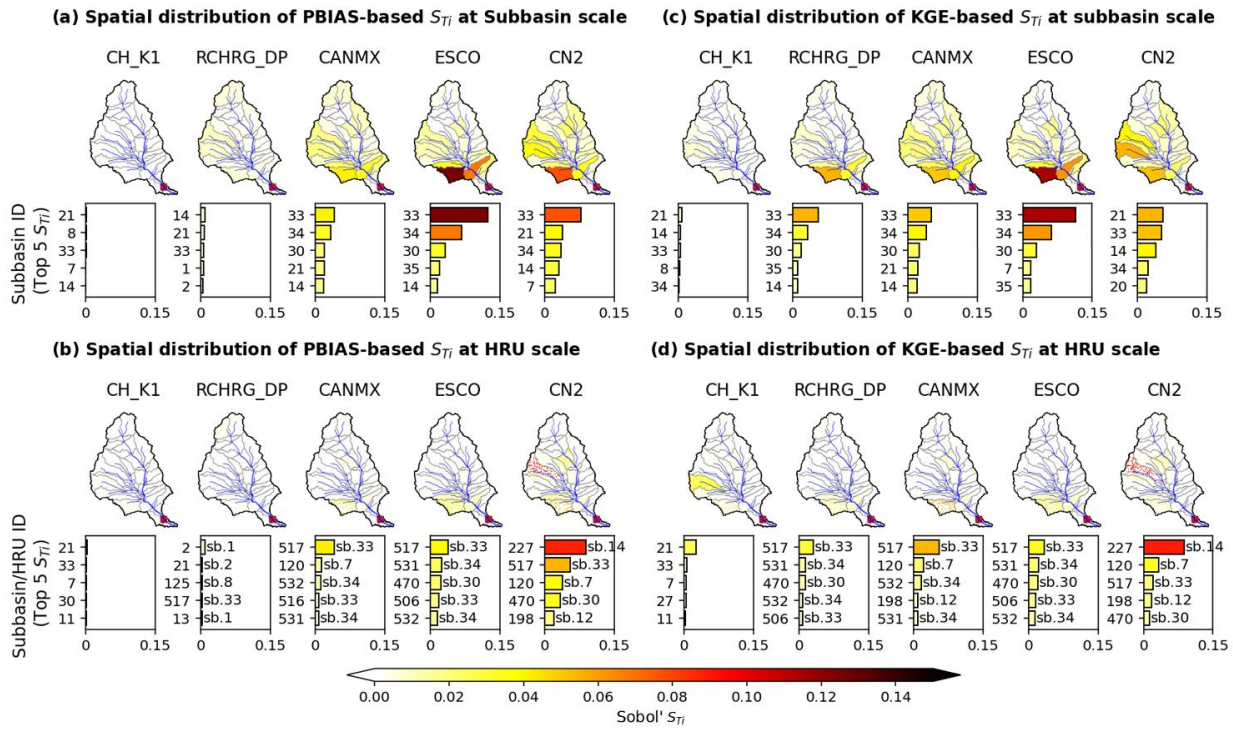

Figure 6: Spatial distribution of the total-effect sensitivity index S_{Ti} at the subbasin and HRU scales using different performance metrics. Panels (a) and (b) show results based on PBIAS, and panels (c) and (d) show results based on KGE.

A notable exception is observed for the routing parameter CH_K1, whose sensitivity decreases under PBIAS and KGE. This difference arises because NSE is particularly sensitive to peak flow timing and variability, whereas PBIAS emphasizes cumulative bias and KGE balances multiple performance components. This finding highlights the influence of metric selection on sensitivity magnitudes while confirming the robustness of the identified spatial patterns.

Overall, these additional analyses demonstrate that the dominant parameters and sensitivity hotspots persist across performance metrics. Therefore, the conclusions of the study are not artifacts of the NSE metric but reflect consistent and physically meaningful controls within the watershed.

Changes made in the revised manuscript:

- Added a new subsection, **Section 4.1: "Effects of performance metrics on sensitivity patterns"**
- Introduced complementary performance metrics (PBIAS and KGE) to assess robustness.
- Conducted additional Morris and Sobol' sensitivity analyses using these metrics.
- Included new results in **Figure 6** and Supporting Information (Figures S5–S6).
- Expanded the Discussion to clarify the influence of performance metric selection.

Comment 6: The manuscript states the model was "simulated at a monthly time step … consistent with available meteorological data," yet earlier you describe daily meteorological records and daily runoff data availability. Clarify if SWAT time step was daily or monthly or whether outputs were aggregated to monthly for NSE/SSA.

Response: We sincerely thank the reviewer for highlighting this inconsistency. We apologize for the confusion caused by the previous wording. In the revised manuscript, we have clarified that the SWAT model was executed at a daily time step, consistent with the temporal resolution of the meteorological inputs. Since observed streamflow data were available only at the monthly scale, the simulated daily runoff was aggregated to monthly values for model evaluation and subsequent sensitivity analysis.

The revised description in the manuscript reads as follows:

"*The model was run at a daily time step over 1962–1986, consistent with the daily meteorological inputs. Since observed runoff was available only at the monthly scale, simulated daily runoff was aggregated to monthly values for model evaluation and sensitivity analysis*."

Comment 7: You highlighted the use of lagged Spearman's rank correlations between sensitivity and runoff is a highlight. However, the 1-month lag in 3-month windows needs clearer physical attribution. Is this a signature of soil moisture memory or a delay in the surrogate response?

Response: We sincerely thank the reviewer for this insightful comment. We agree that the physical interpretation of the observed 1-month lag requires careful consideration. A plausible explanation is soil moisture memory, which can introduce delayed hydrological responses by influencing infiltration, evapotranspiration, and runoff generation. Such memory effects have been widely reported in hydrological systems and could lead to temporally lagged sensitivity patterns. Alternatively, the lag may arise from temporal aggregation effects associated with monthly data or from methodological artifacts related to surrogate modeling.

However, to improve the focus and clarity of the manuscript, we have removed all time-varying sensitivity analyses from the revised version and now concentrate exclusively on spatial sensitivity patterns. Consequently, the lagged Spearman's rank correlation analysis and the associated discussions have been omitted. These topics will be pursued in future work using higher-resolution data and more advanced diagnostic methods.

Although the analysis has been removed, we appreciate the reviewer's suggestion and briefly note potential interpretations for future research. Distinguishing between soil moisture memory and surrogate-induced delays could be achieved through such as higher temporal resolution simulations, particularly at daily or sub-daily scales and cross-correlation and impulse-response analyses to quantify hydrological memory effects. These approaches will be considered in subsequent studies.

Changes made in the revised manuscript:

- Removed the lagged Spearman's rank correlation analysis and all time-varying sensitivity results.

- Revised the title, methods, results, and discussion to focus exclusively on spatial sensitivity analysis.

Comment 8: It seems that CN2 is described as adjusted by a multiplicative factor in screening, and then later treated as independently distributed at subbasin and HRU scales. The replacement vs factor approach is not consistently explained. The reader might want to know how "distributed parameters" are represented and thus what the Sobol indices refer to. You might want to describe how key parameters are perturbed at each scale (subbasin vs HRU), including how SWAT input files are edited and whether spatial structure is preserved or broken for the SA purpose. This would help reproducibility if other researchers are interested in this approach.

Response: We sincerely thank the reviewer for highlighting this important issue. We agree that the representation of distributed parameters and the distinction between the "replace" and "factor" approaches were not sufficiently clear in the original manuscript. The apparent inconsistency arises from incomplete explanations rather than differences in parameter treatment. In all analyses, CN2 and similar parameters are perturbed using multiplicative scaling factors; however, the level at which these perturbations are applied varies depending on the stage of analysis and the spatial parameterization scheme.

To clarify this issue, we have revised the manuscript to provide a detailed and consistent description of parameter perturbation strategies, SWAT input file modifications, and the preservation of spatial structure.

During the Morris screening stage, a spatially lumped parameterization strategy was adopted to reduce dimensionality while preserving dominant hydrological controls. Instead of perturbing each spatial instance independently, each parameter was controlled by a single basin-wide perturbation applied uniformly across all spatial units. Two perturbation approaches were implemented depending on parameter characteristics:

- The "replace" approach, in which parameters representing basin-wide or process-based thresholds (e.g., SFTMP) were assigned a uniform value across all spatial units; and
- The "factor" approach, in which parameters with inherent spatial heterogeneity (e.g., CN2 and SOL_BD) were adjusted using multiplicative scaling factors applied to their original spatially distributed values.
  This strategy preserves the relative spatial variability of parameters while reducing the number of independent variables. Such approaches are widely adopted in SWAT calibration and sensitivity studies (Li et al., 2021; Mao et al., 2024).

For the Sobol'-based spatial sensitivity analysis, two distributed parameterization schemes were implemented:

- Subbasin-scale parameterization. All five parameters were independently specified for each of the 39 subbasins, resulting in 195 distributed parameters. CH_K1 was assigned directly to subbasin files (.sub). The remaining parameters (CN2, CANMX, ESCO, and RCHRG_DP), although stored in HRU-related files (.mgt, *.hru, and *.gw), were

parameterized at the subbasin scale by assigning a single value—implemented through either the "replace" or "factor" approach—to all HRUs within each subbasin. This configuration preserves variability among subbasins while suppressing variability within them.

- HRU-scale parameterization. In this scheme, CH_K1 remained defined at the subbasin level, while CN2, CANMX, ESCO, and RCHRG_DP were allowed to vary independently among the 630 HRUs. This resulted in 2,559 distributed parameters (39 subbasin-specific CH_K1 parameters plus 630 HRU-specific parameters for each of the four remaining variables). Each HRU received an independently sampled value, thereby explicitly representing within-subbasin heterogeneity associated with land use, soil properties, and topography.

These clarifications demonstrate that the spatial structure of parameters is preserved during the screening stage and progressively refined during distributed Sobol' analyses. Consequently, the computed Sobol' indices directly quantify parameter sensitivities at the corresponding spatial scales. To enhance transparency and reproducibility, we have also summarized the perturbation strategies in Table S2 and expanded the methodological descriptions accordingly.

Changes made in the revised manuscript:

- Clarified the distinction between the "replace" and "factor" perturbation approaches.
- Provided detailed descriptions of parameter perturbation strategies at the screening, subbasin, and HRU scales.
- Expanded Sections 2.4 and 2.5 to improve methodological transparency and reproducibility.

Minor Comments

Comment 1: The largest errors occur during high-flow months, attributed to the "limited representation of extreme events in the training dataset". For a model intended to support "flood warnings," this is a significant deficiency.

Response: We thank the reviewer for this valuable comment. We acknowledge that larger errors during high-flow periods represent a limitation of the surrogate model. However, the primary objective of this study is to facilitate efficient sensitivity analysis rather than to support operational flood forecasting. To clarify this distinction, we have revised the manuscript to state that the surrogate model is designed to emulate SWAT for sensitivity analysis, not to replace process-based simulations of extreme events.

"*Nevertheless, such limitations do not compromise the primary objective of this study, as the surrogate models are intended to emulate SWAT for sensitivity analysis rather than to replace process-based simulations of extreme hydrological events.*".

Comment 2: There is a duplicated/incorrect subsection header ("2.5.2 Spatial Parameterization…" and 2.5.1 Spatial Parameterization for Distributed Parameters)

Response: We apologize for this mistake. The subsection title of Section 2.5.1 should be "Spatial parameterization for distributed parameters" and subsection title of Section 2.5.2 should be "Construction of deep learning surrogates". This has been corrected in the revised manuscript.

Comment 3: Numerous typos/grammar issues (e.g., "trransform", "unfiorm", "predication error", etc.)

Response: We apologize for these errors. The manuscript has been thoroughly proofread and carefully revised to correct all typographical, grammatical, and formatting issues. In addition, terminology and language usage have been standardized throughout the text to improve clarity and readability.

Comment 4: Resampling land use to 3000 m and DEM to 150 m is a major preprocessing decision; provide quantitative justification that hydrologic response and HRU composition are not materially altered and implications.

Response: We thank the reviewer for highlighting this important preprocessing choice. The resampling resolutions for the DEM and land use data were selected based on a series of quantitative sensitivity tests to ensure that the basin geometry, drainage structure, and HRU composition were not materially altered.

Specifically, DEM resolutions ranging from 30 to 3000 m (30, 60, 90, 150, 300, 500, 750, 1000, 2000, and 3000 m) were tested. The results indicate that watershed area, drainage characteristics, and the number of subbasins remain essentially unchanged when the DEM resolution varies between 30 and 150 m, whereas coarser resolutions lead to noticeable reductions in basin area and subbasin number. Accordingly, a DEM resolution of 150 m was selected as it preserves key hydrological features while improving computational efficiency.

In addition, the influence of land-use resolution on HRU composition was evaluated in combination with different DEM resolutions. The results show that HRU numbers are primarily controlled by DEM resolution, decreasing significantly as the DEM becomes coarser, while land-use resolution exerts a comparatively minor influence under a fixed DEM. Therefore, the selected combination, i.e., DEM resampled to 150 m and land-use data to 3000 m, represents an optimal balance between preserving essential hydrological characteristics and ensuring computational feasibility for the large number of model evaluations required in this study.

These results have been incorporated into the revised manuscript and are supported by Figure S2 (which is copied below) in the Supporting Information.

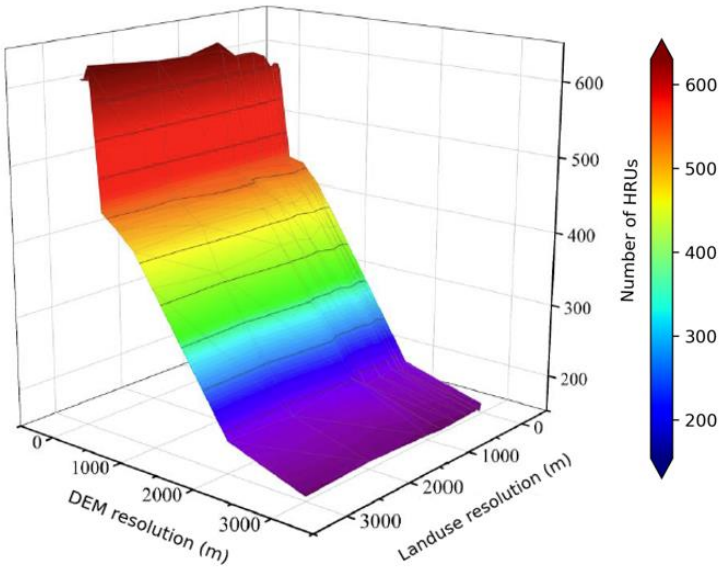

Figure S2： Number of HRUs derived from DEM and land-use data at different spatial resolutions. Both DEM and land-use resolutions were resampled from 30 to 3000 m (i.e., 30, 60, 90, 150, 300, 500, 750, 1000, 2000, and 3000 m), resulting in 100 different HRU delineations based on their combined configurations.

Comment 5: The "hierarchical calibration strategy" is discussed, but the study does not actually demonstrate calibration improvement; tone this down or add an illustrative experiment.

Response: We thank the reviewer for pointing this out. Accordingly, we have toned down the relevant statements in the revised manuscript to avoid implying an empirically validated improvement in calibration performance. We now clarify that the findings suggest the importance of considering the spatial distribution of parameters and their varying sensitivities across spatial scales during model calibration. Rather than advocating a hierarchical calibration strategy, the revised manuscript highlights that spatial sensitivity information can serve as a diagnostic guide for more informed and physically meaningful parameter adjustment.

The revised text reads as follows:

"This study demonstrates that parameter sensitivity in distributed hydrologic models is inherently spatially dependent, highlighting the limitations of traditional lumped calibration strategies. The results suggest that calibration practices may benefit from explicitly considering the spatial distribution of parameters and their varying sensitivities across spatial scales."

Recommendation

The workflow is promising, but the paper needs stronger validation, clearer reproducibility, and more cautious interpretation.

Response: We thank the reviewer again for the overall assessment and recommendation. We hope that these revisions have significantly improved the quality, clarity, and scientific rigor of the manuscript.