Response to comments on Manuscript egusphere-2025-6543 (RC1)

Title: What can hydrological modelling gain from spatially explicit parameterization and multi-gauge calibration?

Authors: Xudong Zheng, Dengfeng Liu*, Hao Wang, Chuanhui Ma, Hui Liu, Guanghui Ming, Qiang Li, Mohd Yawar Ali Khan, and Fiaz Hussain

Manuscript ID: egusphere-2025-6543
Citation: https://doi.org/10.5194/egusphere-2025-6543-RC1

Reply on RC1：

Many thanks for taking the time and effort to review our paper. All comments from Referee #1 are addressed below with point-by-point responses.

For better readability, our replies are prefixed with "**R/**". The original reviewer comments are prefixed with "**RC-n/**" and are presented in **bold**. The revisions to be added into the revised manuscript is highlighted in red. The important parts are highlighted in **blue**. The quoted content is displayed in *italics*. All references cited in this response are listed at the end of the reply. Line numbers from the unrevised manuscript are provided where relevant.

Point-to-point response:

RC1-0/ The manuscript presents a timely and highly relevant investigation into the combined effects of spatially explicit parameterization and multi-gauge calibration on hydrological modeling. The paper is well-written, logically structured, and provides a meaningful steppingstone for the advancement of modern Model-Data Infusion frameworks. However, there are several issues needed to be addressed before publication.

R/ Thanks for your positive feedback and recognition of our work. Your comments are valuable for revising and improving our paper. Below, we provide detailed responses to each of your concerns. The corresponding revisions will be incorporated into the subsequent revised manuscript. To maintain consistency with the original numbering, this comment is labeled as 0.

Specific comments:

RC1-1/ 1. In the parameter calibration, the authors did not consider three key VIC parameters that are commonly calibrated, namely Ds, Ws, and Dm. Although Gou et al. (2020) is cited in the manuscript, that study does not provide sensitivity analysis results for the specific basin investigated here. Including these parameters in the calibration process could potentially lead to different results. Therefore, the authors are encouraged to provide sensitivity analysis results for the study basin to justify the exclusion of these parameters. Otherwise, I believe that Ds, Ws, and Dm should be incorporated into the calibration.

R/ Thank you very much for your insightful comment and your expertise regrading the VIC model. As you correctly pointed out, D_s, W_s, D_m, b, and the three soil layer depths are among the most sensitive parameters in VIC and generally require calibration against basin-specific hydrological observations. This has been widely acknowledged in previous studies and in practical applications of the VIC model (Wen et al., 2012; Gou et al., 2020), and we fully agree with this point.

It should be clarified that there may be a **misunderstanding regarding parameter naming and the equivalence of the baseflow formulation**. Specifically, the free parameters adopted in this study include the variable infiltration curve parameter *b*, the **three baseflow parameters (D_1, D_2, D_3), the three soil layer depths (d_1, d_2, d_3)**, the flow velocity *v*, and the diffusion coefficient *D*.

The VIC model employs the Arno baseflow formulation, which is generally expressed as follows:

$$Q_b = \begin{cases} \frac{D_s D_m}{W_s \theta_b^s} \cdot \theta_b, & 0 \le \theta_b < W_s \theta_b^s \\ \frac{D_s D_m}{W_s \theta_b^s} \cdot \theta_b + \left(D_m - \frac{D_s D_m}{W_s}\right)\left(\frac{\theta_b - W_s \theta_b^s}{\theta_b^s - W_s \theta_b^s}\right)^c, & \theta_b \ge W_s \theta_b^s \end{cases} \quad (1)$$

where θ_b is the soil moisture in the bottom layer (mm); θ_b^s denotes the maximum soil moisture in the bottom layer (mm); D_m represents the maximum baseflow (mm step^{-1}); D_s is the fraction of D_m at which nonlinear baseflow begins, with a range of 0-1; W_s denotes the fraction of θ_b^s at which nonlinear baseflow initiates, also ranging from 0 to 1; The parameter *c* is the exponent governing the nonlinear portion of the Arno baseflow curve and is typically set to 2. Clearly, D_s, W_s, and D_m are the key parameters controlling the shape of the Arno baseflow curve and, consequently, the generation of baseflow; therefore, they require calibration.

In 2001, Nijssen et al. (2001) revisited the Arno model from the perspective of linear reservoirs and proposed an equivalent formulation:

$$Q_b = \begin{cases} D_1 \theta_b, & 0 \le \theta_b < D_3 \\ D_1 \theta_b + D_2 (\theta_b - D_3)^{D_4}, & \theta_b \ge D_3 \end{cases} \quad (2)$$

where D_1 denotes the linear reservoir coefficient; D_2 is the nonlinear reservoir coefficient; D_3 represents the soil moisture (mm) at which nonlinear baseflow begins; D_4 is analogous to the parameter *c*. This formulation is also known as the **Nijssen form of the Arno model**. Its advantage lies in effectively **reducing the interactions among parameters, thereby facilitating calibration** (Mizukami et al., 2017). Currently, this formulation has been incorporated into the official VIC model and can be specified through the global parameter settings.

Therefore, Eq. (1) and Eq. (2) are equivalent, and **calibrating D_1, D_2, D_3 is effectively the same as calibrating D_s, D_m, W_s**. These two sets of parameters can be converted into each other (Mizukami et al., 2017).

We acknowledge that this misunderstanding may have been caused by the similar naming of D_1, D_2, D_3 and d_1, d_2, d_3, as well as the unclear description in Nijssen formulation. To enhance clarity, the three soil layer depths have been renamed h_1, h_2, h_3, and the following explanation has been added into the revised manuscript:

Line 223-224:

It is worth noting that VIC supports two equivalent baseflow parameterizations. Following Mizukami et al. (2017), we adopt the Nijssen formulation (i.e., D_1, D_2 and D_3) to avoid the parameter interactions present in the original Arno formulation. The derivation can be found in Nijssen et al. (2001). Namely, calibration of D_1, D_2 and D_3 is equivalent to calibrating D_s, D_m and W_s, which have been identified as sensitive parameters in previous studies (Wen et al., 2012; Gou et al., 2021).

Renaming d_1, d_2, and d_3 as h_1, h_2, and h_3:

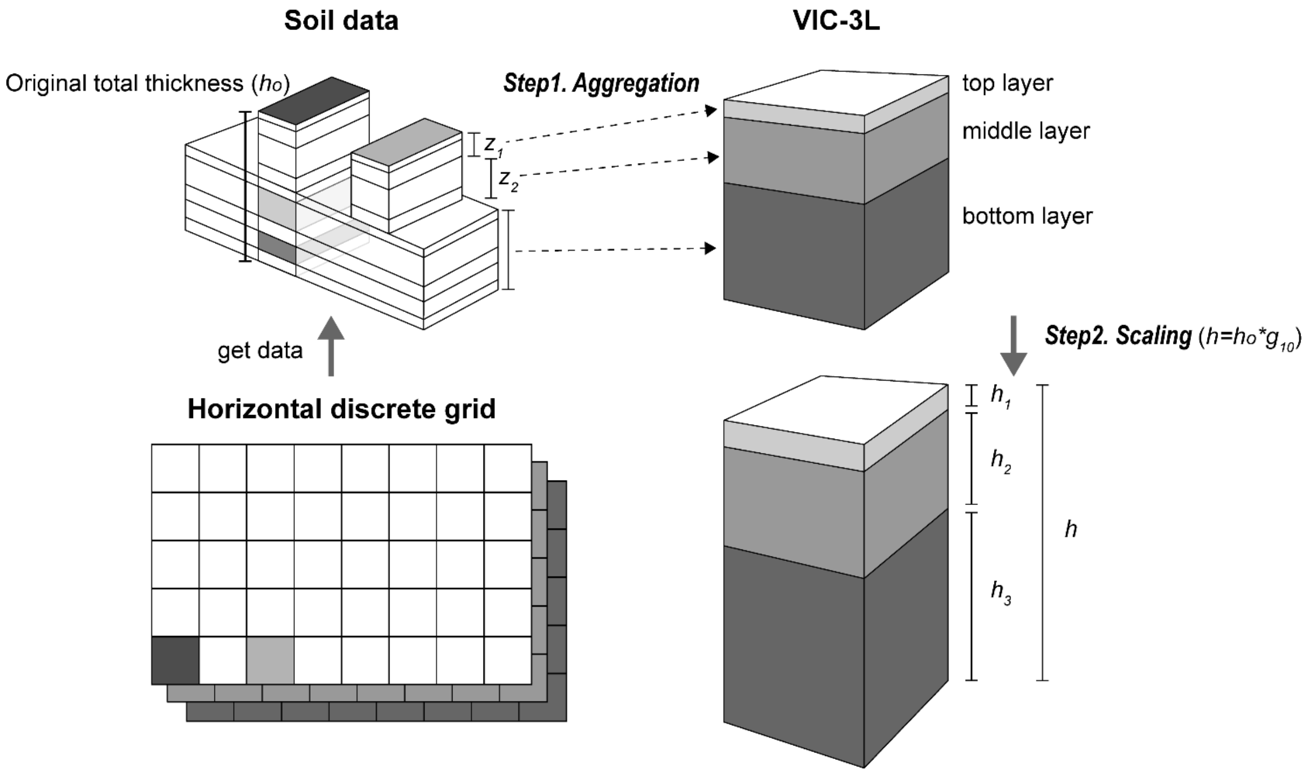

Figure 2. Schematic of the two-step mechanism for reconciling soil data with VIC three-layer (VIC-3L) vertical structure. In the aggregation step, the original soil layers are merged into three VIC soil layers based on the layering numbers z_1 and z_2. The scaling step then applies a scaling factor (g_{10}) to the original total thickness (h_o) to obtain the model-prepared layer thicknesses h_1, h_2, and h_3.

RC1-2/ 2. The use of MPR is indeed an effective approach for deriving distributed parameters; however, one of its key limitations lies in the uncertainty associated with the transfer functions. Previous studies have shown that different transfer functions can lead to substantially different calibration results. For example, Gou et al. (2021) adopted transfer functions for D1–D3 that differ from those used in this study. It remains unclear whether such differences could lead to different conclusions. The manuscript currently lacks analysis and discussion on this issue, which should be addressed to strengthen the robustness of the study.

Gou, Jiaojiao, et al. "CNRD v1. 0: a high-quality natural runoff dataset for hydrological and climate studies in China." Bulletin of the American Meteorological Society 102.5 (2021): E929-E947.

R/ This is a similar issue caused by the confusion noted in RC1-1, and we thank you for bringing it to our attention. In our parameter estimation using the MPR approach, we consulted a wide range of literature to find suitable transfer functions. Among these, **Gou et al. (2020), Gou et al. (2021), and Mizukami et al. (2017)** were particularly important, as they are pioneering works that provide essential transfer functions for applying the MPR approach within the VIC model.

However, we note that the **naming of parameters** and the **formulation of baseflow** module vary across different studies, which we found to be a challenge during our practical implementation. For example, Gou et al. (2020, 2021) tend to use the traditional Arno baseflow formulation, representing the baseflow parameters with D_s, D_m, and W_s, and the three soil layer depths with $D_{1/2/3}$. In contrast, Mizukami et al. (2017) prefer the Nijssen version of the Arno baseflow formulation, using $D_{1/2/3}$ to denote the baseflow parameters equivalent to D_s, D_m, and W_s, while representing the three soil layer depths with $d_{1/2/3}$. **Integrating the formulations from these studies can be quite confusing and prone to errors**.

Since we adopted the Nijssen version of the Arno baseflow formulation, we followed the naming convention used by Mizukami et al. (2017) to maintain consistency with their work. Consequently, our notation appears different from that of Gou et al. (2021). Nevertheless, fundamentally, **calibration of $D_{1/2/3}$ is equivalent to that of D_s, D_m, and W_s**, as we discussed in our reply to RC1-1.

RC1-3/ 3. The manuscript currently lacks sufficient statistical validation of the calibration experiments. Without statistical evaluation, it is difficult to determine whether the reported improvements reflect meaningful advancements or merely small fluctuations. For instance, an KGE difference between 0.715 and 0.716 in Table 6 is not necessarily meaningful without significance testing. The authors are encouraged to incorporate appropriate statistical tests, such as the Wilcoxon signed-rank test or paired t-tests for comparisons. In addition, reporting the standard deviation (you only showed ensemble mean) across ensemble runs would substantially strengthen the credibility of the results.

R/ We sincerely appreciate your constructive suggestion. We fully agree that incorporating more explicit statistical tests is essential for strengthening the robustness and credibility of our conclusions. Following your recommendation, we have performed **pairwise paired t-tests across all cases**. The results are presented in Table 8, and a corresponding analysis has been added to Section 4.1 to further interpret these findings. For your convenience, the relevant revisions, including the updated text and table, are provided below.

Line 369:

To further examine the statistical significance of performance differences across various cases, we conducted pairwise t-tests on the KGE at the Shiquan station (basin outlet), as shown in Table 8. It is evident that, at a significance level of 0.05, Case 4 performs significantly better than all other cases, while Case 8 performs significantly worse than all other cases. Moreover, the patterns identified in the previous analysis are further confirmed by this significance test, with the more complex configurations generally outperforming the simpler ones. It is worth noting that this comparison is conducted under a single-objective context, as Cases 6–8 only consider streamflow at the basin outlet as objective functions.

Table 8. Pairwise t-test results of the ensemble performance metrics (KGE) for the top 40 ranked optimization results across all calibration cases. The table shows the t-statistics for the comparison between each pair of cases, with significant differences indicated by asterisks ($p < 0.05$).

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8
Case 1	–	7.357*	6.820*	-11.652*	0.512	11.331*	2.965*	7.056*

Case 2	-7.357*	–	5.431*	-12.681*	0.189	11.040*	-0.471	6.920*
Case 3	-6.820*	-5.431*	–	-13.425*	-1.082	9.668*	-5.316*	6.343*
Case 4	11.652*	12.681*	13.425*	–	3.842*	13.970*	12.084*	8.459*
Case 5	-0.512	-0.189	1.082	-3.842*	–	8.108*	-0.234	6.308*
Case 6	-11.331*	-11.040*	-9.668*	-13.970*	-8.108*	–	-11.050*	1.568
Case 7	-2.965*	0.471	5.316*	-12.084*	0.234	11.050*	–	6.935*
Case 8	-7.056*	-6.920*	-6.343*	-8.459*	-6.308*	-1.568	-6.935*	–

On the other hand, reporting the standard deviation is generally advisable. However, in the context of this study, Table 7 only presents the ensemble mean of **top five optimal solutions**, resulting in a **very limited sample size**. Under such conditions, the calculation of standard deviation may introduce considerable uncertainty and potentially be misleading. For this reason, we did not include this metric. Nevertheless, our additional analysis indicates that, owing to the characteristics of the optimization algorithm and the relatively large number of calibration iterations performed, **the top five optimal solutions remain stable**. This point is important for clarifying the robustness of the results and for addressing your concern.

RC1-4/ 4. The description of the two-step mechanism for reconciling soil data with the VIC three-layer (VIC-3L) vertical structure is confusing. Since Table 2 already presents the transfer functions for D1–D3, it is unclear what additional role this two-step procedure plays. Because this component underpins the subsequent analysis, a more detailed and transparent explanation is essential.

R/ This issue arises from the naming inconsistency, as discussed in our responses to RC1-1 and RC1-2. D_1–D_3 refer to the parameters governing the Arno baseflow curve (Eq. (2)), not the soil layer depths d_1–d_3 associated with the two-step mechanism discussed here. The misunderstanding likely stems from the fact that **our manuscript follows a different convention than Gao et al. (2021), in which D1–D3 were used to represent soil layer thicknesses**.

On the other hand, unlike some previous VIC applications, the two-step mechanism adopted here accounts for the **dynamic, layer-specific assignment of soil attributes**. As a result, the soil layer delineation not only determines the layer thicknesses but also influences the corresponding soil attributes, such as soil texture.

Specifically, the original SoilGrids dataset provides six soil layers (0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm, and 100–200 cm, hereafter SoilGrids-SL), each with distinct soil texture and bulk density attributes. However, the VIC model typically adopts a three-layer soil structure (hereafter, VIC-3L-SL). Therefore, the multiple soil attributes from SoilGrids-SL need to be assigned to the three layers of VIC-3L-SL. **To enable this process to be dynamically calibrated**, we implemented a two-step mechanism (Fig. 2). First, two layering number parameters (z_1 and z_2) determine the correspondence between SoilGrids-SL and VIC-3L-SL. Based on this mapping, the soil attributes from the relevant SoilGrids layers are aggregated to derive the soil attributes of the three VIC layers (i.e., averaged). Second, a scaling factor (g_{10}) is applied to uniformly adjust the total soil column thickness. This mechanism introduces three calibratable parameters (z_1, z_2, and g_{10}). **The first two parameters control the allocation of soil attributes and the relative thickness proportions of the three VIC soil layers, while the third parameter regulates the overall soil column thickness in VIC-3L**. This mechanism has been applied in Mizukami et al. (2017), and has been shown to be important for overall model performance. Our results also demonstrate this effect, as discussed in Section 4.4.1 with respect to soil layer depth discretization.

We believe that this issue arose from insufficient clarity in the original manuscript. Therefore, we have addressed it in two ways: first, by revising the parameter nomenclature (see our response to RC1-1); and second, by providing additional clarification of the two-step mechanism, which has been incorporated into the revised manuscript as follows.

Line 232-235:

Different soil layer thicknesses produce significantly different runoff responses; consequently, these thicknesses (h_1, h_2, and h_3) are typically treated as free parameters during calibration to fine-tune the VIC behavior (Gou et al., 2020). In addition, the original soil dataset (i.e., SoilGrids1km) provides six soil layers, each characterized by distinct properties (e.g., soil texture), whereas VIC-3L adopts a three-layer soil structure. In order to feed original soil data into the transfer function, it is necessary to reconcile its soil-layer structure with the VIC vertical structure.

RC1-5/ 5. The authors designed eight calibration experiments; however, the current numerical labeling makes it difficult for readers to remember the specific configurations during subsequent discussion. It is recommended that the authors adopt clearer and more descriptive naming conventions to distinguish the different experiments, which would improve readability and interpretability.

R/ You are right that the original experiment names could be confusing. We have improved their readability by **adding aliases**, as reflected in the updated Table 4 and Fig. 3. Corresponding references throughout the manuscript have also been updated to include these aliases for enhanced clarity.

Table 4. Summary of experiments within the EF-SPM framework, showing the cases considered, the configurations compared, and their purposes.

Experiment	Cases	Defined purpose	Alias
Exp. 1-1	Case 1 and Case 5	Uniform vs. fully distributed parameterization (under multi-gauge calibration)	UD-MG
Exp. 1-2	Case 7 and Case 8	Uniform vs. fully distributed parameterization (under single-gauge calibration)	UD-SG
Exp. 2-1	Case 5 and Case 8	Multi-gauge vs. single-gauge calibration (under distributed parameterization)	MS-DP
Exp. 2-2	Case 1 and Case 7	Multi-gauge vs. single-gauge calibration (under uniform parameterization)	MS-UP
Exp. 3-1	Case 4 and Case 5	Uniform vs. spatially explicit soil layer depths parameterization (under multi-gauge calibration)	UD-MG: Depth
Exp. 3-2	Case 3, Case 2 and Case 4	Fixed vs. uniform vs. distributed RVIC parameterization (under multi-gauge calibration)	FUD-MG: RVIC
Exp. 4	Case 6 and Case 5	Simplest baseline setup vs. Full-Complexity setup	Baseline-Complex

*Alias abbreviation: UD = uniform vs. distributed parameterization; MS = multi-gauge vs. single-gauge calibration; FUD = fixed vs. uniform vs. distributed parameterization; MG = multi-gauge calibration; SG = single-gauge calibration; DP = distributed parameterization; Depth = soil layer depths; RVIC = RVIC-specific parameterization; Baseline-Complex = Simplest baseline setup vs. Full-Complexity setup.

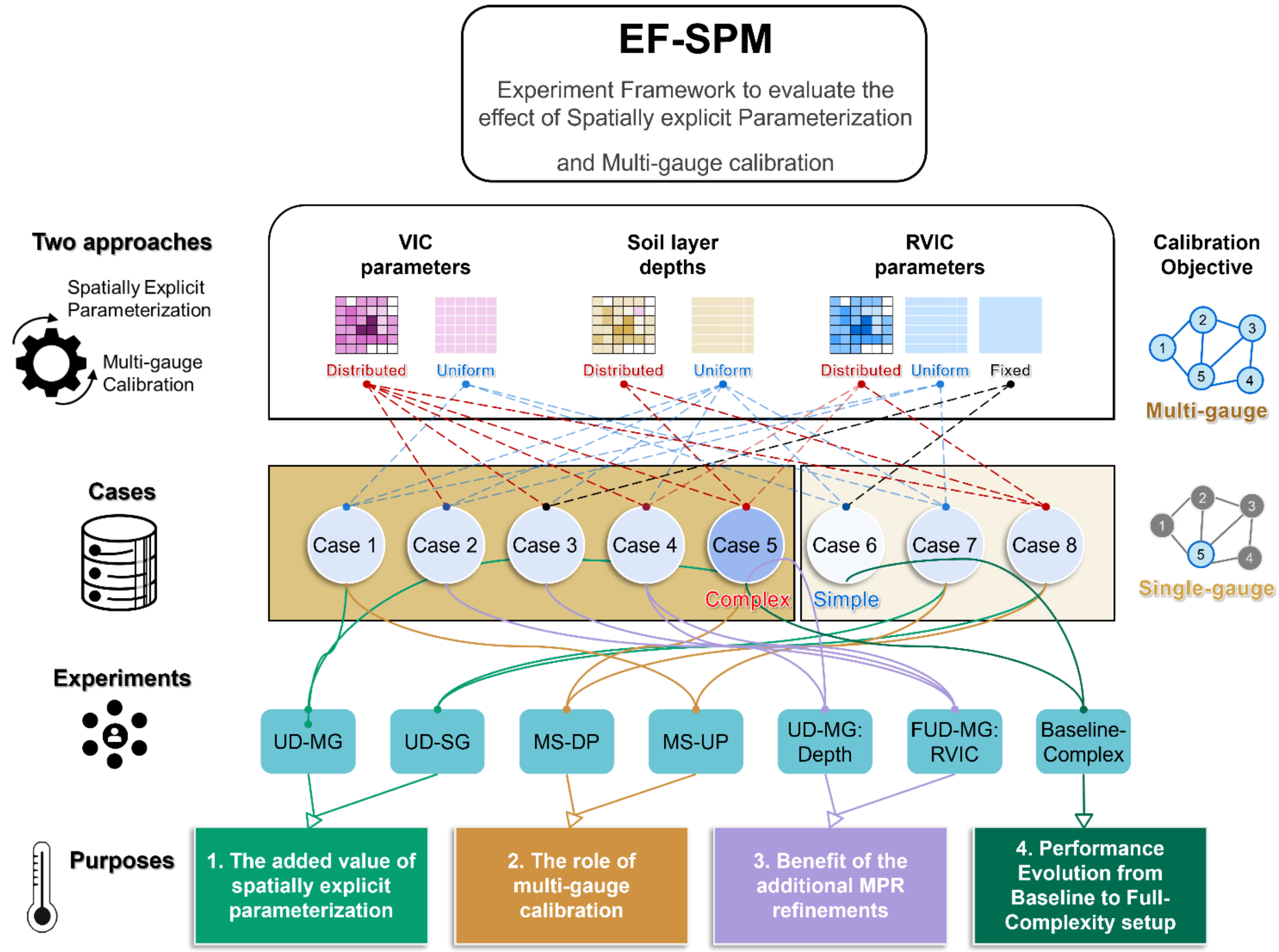

Figure 3. Schematic illustration of the Experiment Framework to evaluate the effect of Spatially explicit Parameterization and Multi-gauge calibration (EF-SPM).

RC1-6/ 6. When applying the NSGA-II algorithm for multi-objective optimization, the implementation details are not sufficiently described. Given that the study basin includes five gauging stations, does this imply that five separate objectives were defined for calibration? The authors should clearly specify how the objective functions were formulated and aggregated, and how the multi-objective framework was structured in practice.

R/ We thank you for pointing out the lack of clarity in our methodology. In response, we have added the following clarification in Section 3.4.

Line 327:

Under the two optimization algorithms, CMA-ES optimizes the KGE at the basin outlet (Shiquan) as the single-objective function (KGE_{Q5}), whereas NSGA-II performs multi-objective optimization in the multi-dimensional Pareto space, considering the KGE of streamflow at the five subbasin gauges as separate objectives (KGE_{Q1}, KGE_{Q2}, KGE_{Q3}, KGE_{Q4}, KGE_{Q5}).

It should be noted that, under the NSGA-II framework, the five objectives **are treated as five independent dimensions in the Pareto space and are not aggregated**. Further clarification of Fig. 10 will indeed help to better address this point and provide a more comprehensive understanding of the optimization process.

In Fig. 10 (below), each subplot represents one of the optimization objectives, corresponding to KGE_{Q1}, KGE_{Q2}, KGE_{Q3}, KGE_{Q4}, and KGE_{Q5} (i.e., the streamflow at five gauges). Within this five-dimensional Pareto space, NSGA-II explores the trade-offs between the conflicting objectives. The Pareto front, highlighted in red in the figure, represents a set of solutions where no other solution is strictly better across all objectives. In other words, each point on the Pareto front is non-dominated, meaning that it cannot be improved in any objective without worsening at least one other objective. However, it should be emphasized that there is no single solution that is superior across all objectives; instead, the Pareto front provides a range of equally optimal solutions, each reflecting a different trade-off between the objectives.

Therefore, **the multi-objective optimization algorithm of NSGA-II ultimately produces a set of solutions on the Pareto front through this process, without the need to aggregate the five objectives**. The final optimal solution is then selected from this set of solutions on the Pareto front by manually choosing the most balanced solution, considering the trade-offs between the objectives.

We hope that our response addresses your concerns and provides the clarity needed.

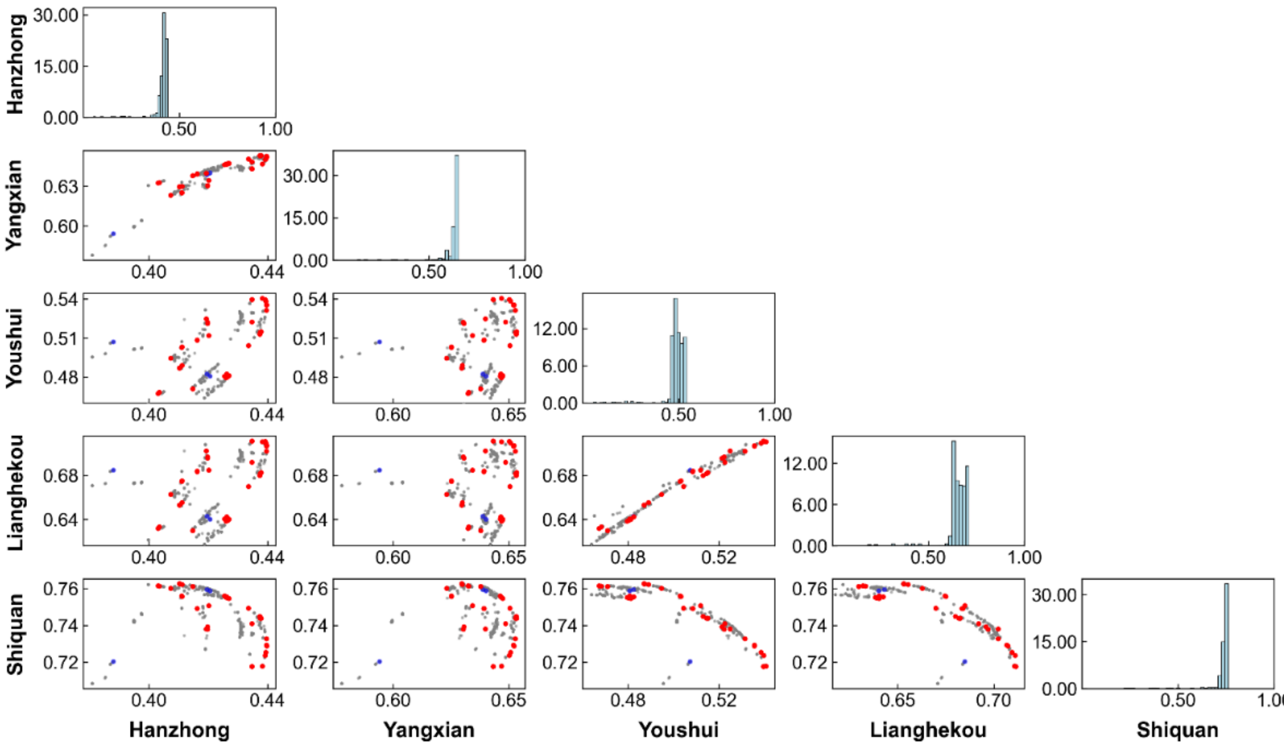

Figure 10. Objective space scatter plot matrix for Case 1. The matrix depicts the trade-offs between calibration objectives, where each axis corresponds to the Kling-Gupta Efficiency (KGE) of simulated streamflow for a sub-basin (gauge). Grey dots represent all candidate solutions from the calibration process. The initial and final Pareto fronts are highlighted in blue and red, respectively. Histograms along the diagonal show the distribution of the KGE for each individual objective.

RC1-7/ 7. The authors should include, in Table 5, the ranges for all parameters to be optimized. Providing the parameter bounds would improve transparency and allow readers to better assess the robustness and reproducibility of the calibration procedure.

R/ Thank you for pointing out the insufficient methodological description in our manuscript. Following your suggestion, we have added Table 6 to provide a detailed description of the parameters subject to calibration and their feasible ranges, in order to improve the transparency of the methodology. The table and the added text are provided below for your convenience.

Line 304:

The parameters to be calibrated in the EF-SPM experimental framework, along with their descriptions and feasible ranges, are summarized in Table 6.

Table 6. Description and feasible ranges of the free parameters subject to calibration.

Category	Parameters	Unit	Related parameters and description	Feasible range	
				Lower	Upper
VIC parameters	g_1, g_2	–	g-parameters for Variable infiltration curve parameter (Distributed)	−2.0, 0.8	1.0, 1.2
	b	–	Variable infiltration curve parameter (Uniform)	0.001	0.5
	g_3	–	g-parameter for D_1 (Distributed)	1.75	3.5
	g_4	–	g-parameter for D_2 (Distributed)	1.75	3.5
	g_5	–	g-parameter for D_3 (Distributed)	0.001	2.0
	D_1	–	Linear reservoir coefficient (Uniform)	0.001	1.0
	D_2	–	Nonlinear reservoir coefficient (Uniform)	0.001	1.0
	D_3	–	Percentile of the bottom soil layer thickness (Uniform)	0.001	1.0
Soil layer depths	g_{10}	–	Scaling factor of total depth	0.1	4.0
	z_1, z_2	–	The layering number of the top two soil layers in the SoilGrids 6-layer soil profile	1, 3	2, 5
RVIC parameters	t_p	[h]	Time of peak occurrence	1.0	24.0
	μ	[h^{-1}]	Rising coefficient	2.0	10.0
	m	–	Recessing coefficient	0.5	6.0
	g_6, g_7, g_8	–	g-parameters for flow velocity (Distributed)	0.01, 0.1, 0.2	0.5, 0.3, 0.4
	g_9	–	g-parameter for flow diffusivity (Distributed)	0.01	0.5
	v	[$m\ s^{-1}$]	Flow velocity (Uniform)	0.01	3.0
	D	[$m^2\ s^{-1}$]	Flow diffusivity (Uniform)	10.0	4000.0

RC1-8/ 8. Although the authors provided a zoomed-in view in Figure 5, the differences remain unclear. From my perspective, the four schemes perform almost identically in the high-flow segment.

R/ We appreciate your careful examination of Fig. 5. To better highlight the high-flow differences among the four schemes, we have updated Fig. 5 to include a **box-and-scatter representation**. This visualization makes the distribution of simulated high flows more apparent, showing that the spatially explicit parameterization schemes (Cases 5 and 8) tend to produce **higher peaks** compared to the uniform schemes (Cases 1 and 7). These differences are further supported by the signature metrics presented in Table 9 and 10. We have added a clarifying statement in the manuscript to highlight these distinctions.

Line 396-398:

In general, spatially explicit parameterization (Cases 5 and 8) offer improved simulation of the high-flow segment, yielding higher peak values (as shown in the box-and-scatter representation Fig. 5); however, this comes at the expense of accurately capturing extreme low flows, namely, an overestimation of baseflow.

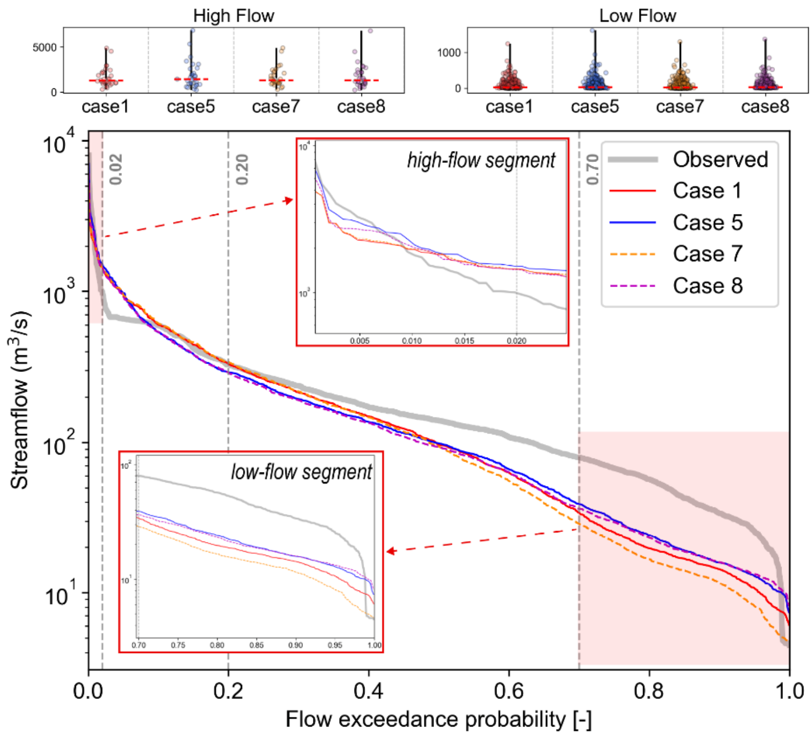

Figure 5. Comparison of observed and simulated flow duration curves for different case configurations at the Shiquan station during the validation period.

Reference

Gou, J., Miao, C., Duan, Q., Tang, Q., di, Z., Liao, W., Wu, J., and Zhou, R.: Sensitivity analysis - based automatic parameter calibration of the variable infiltration capacity (VIC) model for streamflow simulations over China, Water Resources Research, 56, 10.1029/2019WR025968, 2020.

Gou, J., Miao, C., Samaniego, L., Xiao, M., Wu, J., and Guo, X.: CNRD v1.0: A High-Quality Natural Runoff Dataset for Hydrological and Climate Studies in China, Bulletin of the American Meteorological Society, 102, 1-57, 10.1175/BAMS-D-20-0094.1, 2021.

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., Rakovec, O., and Samaniego, L.: Towards seamless large-domain parameter estimation for hydrologic models, Water Resources Research, 53, 8020-8040, 10.1002/2017wr020401, 2017.

Nijssen, B., Lettenmaier, D., Lohmann, D., and Wood, E.: Predicting the Discharge of Global Rivers, Journal of Climate - J CLIMATE, 14, 3307-3323, 10.1175/1520-0442(2001)014<3307:PTDOGR>2.0.CO;2, 2001.

Wen, Z., Liang, X., and Yang, S.: A new multiscale routing framework and its evaluation for land surface modeling applications, Water Resources Research, 48, 8528, 10.1029/2011WR011337, 2012.

Finally, we would like to once again thank the Editor and all the Reviewers for your thorough review and support of our paper. If you have any questions, suggestions, or discussions, please feel free to contact us.