https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

A Skill-Guided LLM Agent for Process-Aware Calibration of Large-Scale Hydrological Models: Demonstration with VIC v5.1.0

ChengYang Qian[1,2], Guoqiang Peng[1], Yujing Sun[1], Cheng Zeng[2], Junzhi Liu[3]

[1]School of Geography and Tourism, Shaanxi Normal University, Xi'an, 710062, China
[2]SIPSG Infomation Technology Co., Ltd., Suzhou, 215100, China
[3]Center for the Pan-Third Pole Environment, Lanzhou University, Lanzhou, 730000, China

Correspondence to: Guoqiang Peng (pgq@snnu.edu.cn)

Abstract. Calibrating large-scale hydrological models is challenging because heterogeneous land-surface processes and parameter interactions complicate error diagnosis and physically meaningful parameter adjustment. Conventional optimization can identify high-performing parameter sets but provides limited process-level explanations. Although large language models (LLMs) offer opportunities for knowledge-guided calibration, explicit hydrological reasoning linked to testable adjustment hypotheses and simulation-based verification remains underdeveloped. We develop a Skill-guided LLM agent for process-aware calibration. The framework organizes hydrological knowledge into reusable procedural guidance through four domain-specific Skills: Diagnosis, Adjustment, Evaluation, and Verification. In a closed-loop workflow, the agent interprets streamflow error signatures, formulates process-level hypotheses, proposes physically constrained parameter adjustments, and evaluates their effects through deterministic simulations. The framework was implemented using VIC v5.1.0 coupled with RVIC in the Jialing River Basin. Under the same model-evaluation budget, it was compared with an LLM-based approach without Skill guidance and the SCE-UA algorithm. Relative to the baseline parameter configuration, the Skill-guided approach increased the Nash–Sutcliffe efficiency (NSE) from 0.690 to 0.716 and reduced mean absolute peak-timing error across the same stations and flood events from 23.22 to 9.89 days. Improvements were not consistent across all hydrological signatures. Calibration trajectories showed how process diagnoses informed parameter interventions and how simulation feedback supported their verification. Local sensitivity analysis showed that parameter sensitivities differed between the initial and optimized configurations. These results demonstrate the feasibility of integrating structured hydrological knowledge, LLM reasoning, and deterministic simulations into a traceable, hypothesis-driven calibration framework for process-based hydrological models.

1 Introduction

Process-based hydrological models are widely used to investigate terrestrial water-cycle processes, predict streamflow, and assess hydrological responses to environmental and climatic changes (Clark et al., 2015; Fatichi et al., 2016). For large-scale applications, these models represent spatially heterogeneous land-surface processes through parameterized formulations of infiltration, soil-water storage, evapotranspiration, runoff generation, and baseflow. Interactions among these processes and

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

their uncertain parameters complicate the diagnosis of simulation errors and the identification of hydrologically meaningful adjustments (Renard et al., 2010; Guse et al., 2024). Manual trial-and-error calibration is especially demanding because modifying one parameter can affect several aspects of the simulated hydrograph. The Variable Infiltration Capacity (VIC) model is a representative large-scale, process-based hydrological model that describes spatially variable land-surface processes while remaining computationally suitable for regional applications (Liang et al., 1994; Hamman et al., 2018). Effective calibration should therefore consider not only the agreement between simulated and observed discharge but also the hydrological interpretation of parameter adjustments (Clark et al., 2017).

Conventional calibration methods generally formulate parameter estimation as an optimization problem in which uncertain parameters are iteratively adjusted to improve agreement between simulated and observed streamflow (Gupta et al., 1998; Hrachowitz et al., 2014). Global optimization algorithms, such as the Shuffled Complex Evolution method developed at the University of Arizona (SCE-UA), have been widely adopted because of their ability to explore nonlinear and interacting parameter spaces more effectively than local search procedures (Duan et al., 1992). Although these algorithms provide systematic search strategies, distributed hydrological calibration remains computationally demanding because each candidate parameter set may require a complete land-surface and routing simulation (Ercolani and Castelli, 2017). These methods are designed to efficiently explore parameter spaces, but they generally do not provide explicit mechanisms for interpreting calibration errors or linking parameter changes to specific hydrological processes (Yilmaz et al., 2008; Zhang et al., 2009). Process-based diagnostic approaches can partly address these limitations by evaluating multiple hydrograph signatures rather than relying on a single aggregate metric (McMillan, 2021). However, translating an error signature into a diagnosis of the underlying process deficiency and then into an appropriate parameter-adjustment strategy still relies substantially on expert knowledge and accumulated calibration experience. This knowledge is often tacit and dispersed across scientific literature, model documentation, and individual practice, making it difficult to reproduce and systematically transfer across users, basins, and hydrological models (Knoben et al., 2022).

Recent advances in artificial intelligence have introduced new approaches for hydrological model calibration beyond conventional simulation-based parameter optimization (Wang et al., 2026). Machine learning methods reduce computational costs by approximating process-model responses and enabling efficient parameter searches across large catchment samples (Farahani et al., 2025). Differentiable parameter-learning frameworks infer model parameters from catchment characteristics, while knowledge-informed deep-learning approaches integrate sensitivity analysis and information-theoretic measures to identify influential parameters (Tsai et al., 2021; Jiang et al., 2023). However, although these approaches improve calibration efficiency and parameter estimation, the knowledge guiding their decisions is generally encoded implicitly in training data, surrogate models, objective functions, or learned policies (Herrera et al., 2022). Consequently, existing AI-based hydrological calibration approaches still lack an explicit and reusable framework for formalizing the iterative loop of discrepancy diagnosis, process interpretation, constrained parameter adjustment, re-simulation, and evaluation.

Large language models (LLMs), when integrated into agent-based systems, can combine domain information, task planning, external tool execution, and iterative feedback to support scientific workflows (Wang et al., 2024; Wei et al., 2022). Their

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

application to hydrological modeling nevertheless requires safeguards against unsupported reasoning, limited reproducibility, and insufficient physical grounding (Foroumandi et al., 2023). Recent work has explored LLM-assisted VIC calibration through interactions with a numerical simulator (Zhu et al., 2026), demonstrating the feasibility of this emerging approach. A further methodological challenge is to make the relationship among error signatures, process hypotheses, parameter interventions, and simulation outcomes explicit and reusable. In the present study, domain-specific Skills are used to structure this knowledge (Nearing et al., 2021). Each Skill provides procedural guidance and constraints for a particular task, enabling the agent to distinguish process diagnosis from action generation and to evaluate proposed interventions against subsequent model responses.

In this study, we develop and evaluate a Skill-guided LLM framework for VIC model calibration. Following the reasoning–acting paradigm of LLM agents , the framework integrates hydrological Skills with LLM reasoning and deterministic VIC–RVIC tools, enabling process-aware diagnosis, constrained parameter adjustment, and feedback-based iterative verification. Specifically, we (1) develop a Skill-based calibration framework that formalizes hydrological calibration expertise by organizing the links among hydrograph-error signatures, process-level diagnoses, and physically constrained parameter adjustment hypotheses; (2) establish an auditable closed-loop calibration workflow that separates LLM reasoning from deterministic VIC–RVIC simulation, evaluation, and verification; and (3) systematically compare the proposed framework with a matched NoSkill-LLM baseline and the SCE-UA algorithm under controlled experimental conditions. The evaluation focuses on calibration and validation performance, optimization behavior under limited simulation budgets, and the hydrological consistency of calibration decisions.

2 Methodology

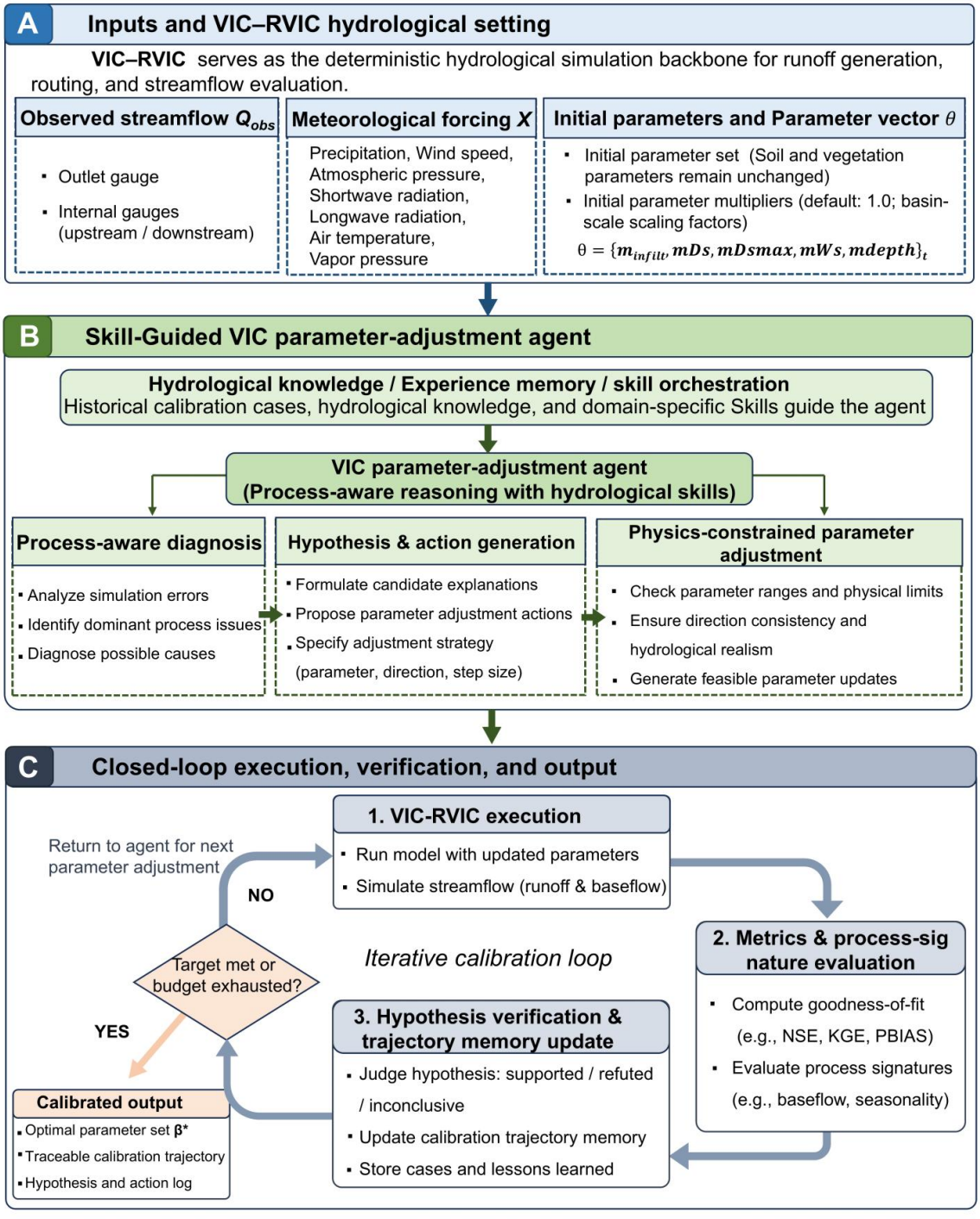
2.1 Overall framework of Skill-guided LLM calibration

The proposed framework is a Skill-guided large language model (LLM) framework for hydrological model calibration that integrates domain-specific reasoning with deterministic VIC–RVIC simulations. The central design principle is to separate procedural knowledge, reasoning capability, and numerical computation. The Skill layer defines the calibration procedures and hydrological knowledge required for each calibration stage. The LLM layer performs contextual reasoning within the constraints defined by active Skills, including interpreting hydrological errors, formulating process-level hypotheses, and generating structured parameter adjustment actions. The Tool layer provides deterministic computational functions, including parameter updating, VIC–RVIC execution, metric calculation, and calibration trajectory recording.

The complete calibration process follows a closed-loop workflow consisting of six stages (Fig. 1): (1) initialization of calibration states and parameter configurations; (2) process-aware diagnosis of simulated and observed streamflow discrepancies; (3) hydrological hypothesis formation and structured action generation; (4) physics-constrained filtering of proposed parameter modifications; (5) deterministic VIC–RVIC simulation and performance evaluation; and (6) hypothesis

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

verification, trajectory-memory update, and stopping assessment. The updated calibration state, simulation evidence, and hypothesis verification results are then fed back to the agent for the next iteration until the stopping criteria are met.



https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Figure 1: Technical workflow of the study.

2.2 Skill design and hydrological knowledge representation

A Skill is defined as an executable representation of hydrological calibration knowledge that provides procedural and domain-specific guidance for an LLM-based agent. Unlike a fixed prompt or deterministic rule set, a Skill explicitly specifies the required information, relevant domain knowledge, applicable constraints, and expected outputs for a specific calibration task. Four task-specific Skills are developed in the present agent: Diagnosis, Adjustment, Evaluation, and Verification. These Skills provide an explicit structure for interpreting model errors and formulating parameter interventions, while deterministic software components execute parameter modifications, hydrological simulations, and numerical metric calculations.

(1) Diagnosis Skill. The Diagnosis Skill guides the LLM agent to interpret discrepancies between simulated and observed streamflow and identify possible hydrological process deficiencies based on hydrograph error signatures, including runoff generation imbalance, insufficient baseflow response, recession mismatch, and timing errors.

(2) Adjustment Skill. The Adjustment Skill provides parameter – process relationships, candidate calibration parameters, adjustment constraints, and physically admissible modification ranges. It supports the generation and validation of feasible parameter adjustment actions while maintaining physical consistency.

(3) Evaluation Skill. The Evaluation Skill defines the numerical evaluation procedures and performance indicators used to assess model responses, including NSE, KGE, PBIAS, and process-level signatures.

(4) Verification Skill. The Verification Skill enables closed-loop hypothesis testing by evaluating whether the simulated response after parameter modification supports, refutes, or provides inconclusive evidence for the proposed hydrological hypothesis. The verification outcome guides subsequent calibration iterations.

Together, the four Skills organize existing hydrological expertise into a documented decision workflow. They do not independently compute streamflow or prescribe universal numerical parameter-update rules; instead, they guide the agent's interpretation of model responses and its selection of admissible interventions (Fig. 2).

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

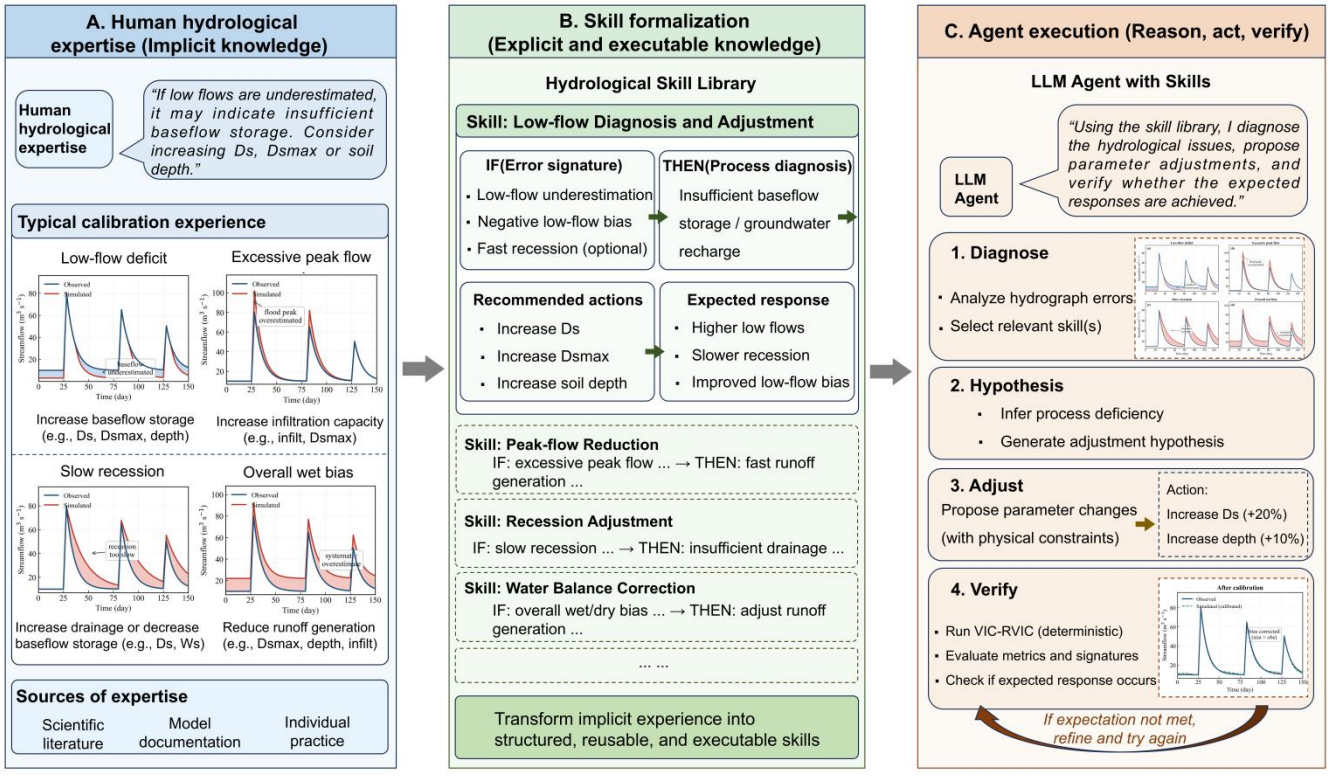


Figure 2: From implicit hydrological expertise to executable calibration reasoning skills.

2.3 Process-aware diagnosis and hypothesis-driven calibration

2.3.1 Calibration state representation

At calibration iteration t, the current model condition is represented as a structured calibration state rather than a single objective-function value:

$$S_t = \{\theta_t, M_t, D_t, H_t\} \quad (1)$$

where θ_t denotes the current parameter vector, M_t represents global performance metrics, D_t represents hydrological process signatures, and H_t denotes the accumulated calibration trajectory memory. The parameter state is expressed as:

$$\theta = \{m_{\inf ilt}, m_{Ds}, m_{Ds\max}, m_{Ws}, m_{depth}\} \quad (2)$$

where each term denotes the current multiplier applied to the corresponding VIC parameter field (Sepúlveda et al., 2022). Nash–Sutcliffe efficiency (NSE) is the primary calibration objective, while Kling–Gupta efficiency (KGE) and percentage bias (PBIAS) provide complementary information on streamflow agreement and water-volume bias (Gou et al., 2020; Gupta et al., 2009). The process signatures include volume bias (VB), peak-flow error (PE), low-flow error (LE), recession-response mismatch (RR), and timing error (TE), which characterize hydrological deficiencies not fully captured by aggregate

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

metrics (Mizukami et al., 2019). The trajectory memory records previous parameter states, diagnoses, actions, simulated responses, and verification outcomes so that subsequent decisions can use both current and historical information.

2.3.2 Multi-label process diagnosis

The Diagnosis Skill interprets discrepancies between simulated and observed discharge using the process signatures. Because multiple hydrological deficiencies can occur simultaneously, diagnosis is treated as a multi-label task with six categories: S1, excessive runoff generation; S2, insufficient runoff generation; S3, peak-flow mismatch; S4, recession-response mismatch; S5, timing mismatch; and S6, persistent or structural mismatch. More than one category may be activated at a given iteration. For example, a simulation may reproduce total runoff volume while underestimating flood peaks or misrepresenting low-flow behavior. The diagnostic labels do not directly prescribe parameter values. Instead, they identify plausible process deficiencies to be examined during hypothesis formation and parameter adjustment.

2.3.3 Hypothesis formation and structured action generation

Following process diagnosis, the agent formulates an explicit hydrological hypothesis describing a plausible cause of the observed model deficiency (Shinn et al., 2023; Wang et al., 2023). The hypothesis represents a plausible hydrological mechanism rather than an immediate parameter prescription. Guided by the Adjustment Skill, the agent then proposes a structured action specifying the target parameter, adjustment direction and magnitude, hydrological rationale, and expected model response. Separating hypothesis formation from action generation makes the link between diagnosed processes and proposed interventions explicit, rather than directly mapping aggregate performance metrics or error signatures to parameter values. The resulting hypothesis remains a testable model-level explanation rather than a definitive attribution of catchment physical behavior.

2.4 Physics-constrained execution and hypothesis verification

2.4.1 Physics-constrained action filtering

Before any VIC parameters are modified, the controller evaluates each LLM-generated action using a deterministic feasibility filter. The checks cover parameter bounds, permitted single-step adjustment magnitudes, direction consistency, parameter – process consistency, and output-schema validity. An invalid action is rejected and returned for bounded regeneration; if no admissible action is obtained, the parameter state is retained.

The accepted parameter multipliers are applied to the fixed spatial parameter fields. For parameter i in grid cell g at iteration t, the corresponding VIC parameter value is calculated as:

$$p_{i,g}^{(t)} = m_{i,t} p_{i,g}^{(base)} \tag{3}$$

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

where $p_{i,g}^{(base)}$ denotes the base parameter value and $m_{i,t}$ is its current calibration multiplier. For a controller-validated adjustment, the parameter state is updated as:

$$m_{i,t+1} = m_{i,t} + \Delta m_{i,t}^{acc} \tag{4}$$

Here, $\Delta m_{i,t}^{acc}$ is the effective change to multiplier i accepted by the controller, as determined by the implemented adjustment rule and prescribed constraints. It is zero for parameters not updated, including cases in which no valid action is obtained. Thus, the LLM proposes an intervention while the controller determines and executes the admissible state change.

After validation, the controller writes the updated parameter configuration and executes the VIC–RVIC simulation chain. VIC calculates runoff generation, soil-water dynamics, and baseflow; RVIC routes the resulting runoff components to simulated discharge at the target stations. Streamflow metrics and process signatures are then computed deterministically from simulated and observed discharge. NSE is the primary calibration objective, while KGE, PBIAS, and process signatures remain available for diagnosis and verification. The LLM does not execute the hydrological simulations or replace the deterministic calculation of evaluation metrics.

2.4.2 Hypothesis verification and trajectory memory update

After each VIC–RVIC simulation, the Verification Skill compares the expected response e_t generated before simulation with the realized response r_{t+1}.

The realized response is defined as: $r_{t+1} = \{\Delta m_{t+1}, \Delta d_{t+1}\}$ (5)

where: $\Delta m_{t+1} = m_{t+1} - m_t$ (6)

and: $\Delta d_{t+1} = d_{t+1} - d_t$ (7)

The verification outcome is classified as Supported when the simulated response agrees with the expected hydrological mechanism, Refuted when it contradicts the hypothesis or degrades the targeted behavior, and Inconclusive when the evidence is insufficient or conflicting. The classification draws on the expected response and the observed metric and signature changes rather than on a single objective-function value.

Each iteration is recorded in the trajectory memory, including the parameter state, diagnosed symptoms, hypothesis, proposed and executed action, expected and realized responses, verification verdict, and resulting model performance. This record informs subsequent diagnosis and parameter adjustments. Calibration terminates when the predefined stopping criteria are met, including attainment of the target performance criterion or exhaustion of the model-evaluation budget. The final outputs are the best-performing parameter set, its evaluation metrics, and the calibration trajectory.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

2.5 VIC model and calibration parameterization

The Variable Infiltration Capacity (VIC) model was used as the deterministic hydrological simulator within the proposed Skill-guided LLM calibration framework. VIC represents the interactions among meteorological forcing, vegetation characteristics, soil properties, and water storage processes through physically based land-surface formulations (Mizukami et al., 2017). In this study, vegetation parameters and non-calibrated soil parameters were kept unchanged throughout the optimization process to maintain a consistent model structure.

Surface runoff was generated using the variable infiltration formulation, while baseflow was simulated using the nonlinear groundwater response formulation. Runoff and baseflow generated from individual VIC grid cells were subsequently routed through RVIC to obtain simulated streamflow at hydrological stations.

Five hydrologically sensitive VIC parameters were selected as the calibration parameter space: infilt, D_s, W_s, D_{smax}, and depth. These parameters represent infiltration capacity, nonlinear baseflow initiation, maximum baseflow generation, soil moisture threshold controlling nonlinear baseflow response, and soil storage capacity, respectively (Hamman et al., 2018). The baseflow recession exponent was fixed at 2.0 and excluded from calibration.

The initial values of the five calibrated parameters were used to construct the baseline VIC parameterization. Specifically, the initial values of infilt, D_s, W_s, and depthwere assigned as 0.3, 0.001, 0.8, and 1.9, respectively, based on representative values reported in previous VIC applications. Unlike these globally assigned parameters, D_{smax} was initialized using its spatially distributed baseline field, where values vary among VIC grid cells according to the corresponding parameter dataset. A multiplier of 1.0 preserves the baseline parameter field, whereas deviations from 1.0 modify the parameter magnitude while maintaining the original spatial heterogeneity of the VIC parameter dataset.

3 Experimental design

3.1 Study area and datasets

The Jialing River Basin, located in the upper Yangtze River Basin, was selected as the study area. The basin covers approximately 160,000 km^2 and spans parts of Shaanxi Province, Gansu Province, Sichuan Province, and Chongqing Municipality (Fig. 3).

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

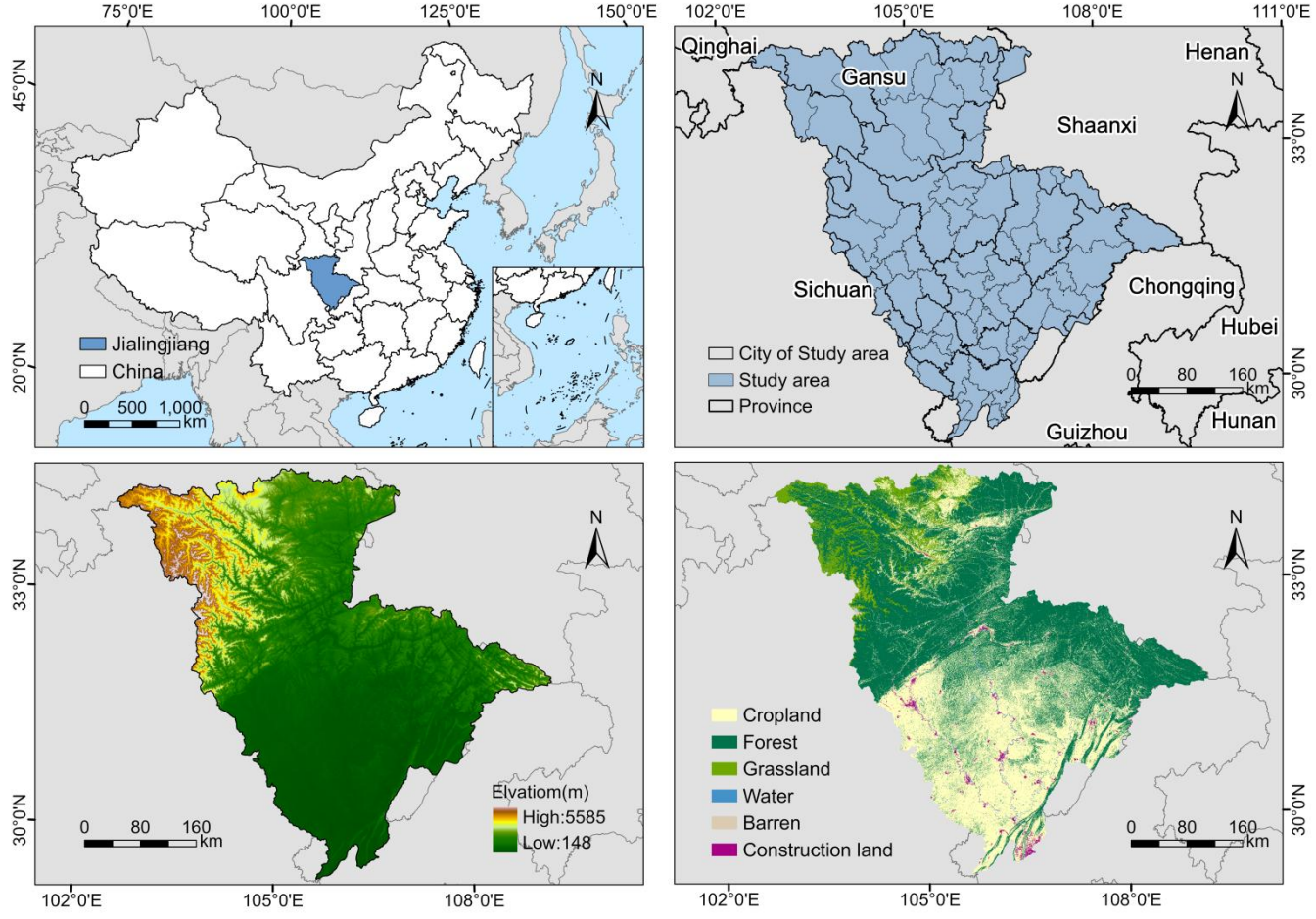


Figure 3: Location of the study area.

A 90 m digital elevation model (DEM) from the Global Hydrodynamics Project was used to derive terrain information for hydrological routing (https://global-hydrodynamics.github.io/; last access: 30 October 2026). Meteorological forcing data were obtained from ERA5 hourly data on single levels from 1940 to present (https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels?tab=download; last access: 30 October 2026), including air temperature, precipitation, surface pressure, wind speed, downward shortwave radiation, downward longwave radiation, and vapor pressure. Soil parameters were derived from the Harmonized World Soil Database (HWSD v1.2), with soil hydraulic properties estimated using pedotransfer functions. Vegetation parameters were generated from the MODIS MCD12Q1 land-cover product and the VIC vegetation parameter library (Yeste et al., 2020). The base soil and vegetation parameter datasets were held fixed across all experiments.

Daily observed streamflow data for 2004–2018 were obtained from the National Earth System Science Data Center, National Science & Technology Infrastructure of China (http://www.geodata.cn; last access: 30 October 2026). The Beibei station was used for calibration, while the remaining stations were used for spatial transfer evaluation.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

3.2 VIC–RVIC simulation configuration

VIC (version 5.1.0) was configured at a spatial resolution of 0.25° × 0.25°, covering 305 grid cells. Hourly meteorological data were aggregated to six-hourly forcing inputs for VIC simulations. VIC performed water and energy balance calculations at a six-hour time step, with runoff and baseflow from four consecutive time steps summed to daily totals. These daily outputs were routed using RVIC, and the resulting daily streamflow was evaluated against daily observations. RVIC used the routing configuration described below, the simulation covered 2004 – 2018, with 2004 – 2006 as spin-up, 2007 – 2015 for calibration, and 2016–2018 for independent temporal validation.

A preliminary routing experiment was conducted using five candidate routing velocities (0.5, 1.0, 1.5, 2.0, and 3.0 m s^{-1}). Among these candidates, 2.0 m s^{-1} produced the best streamflow simulation performance and was therefore selected as the fixed routing velocity for the subsequent calibration and validation experiments. The routing output interval was set to 86,400 s to match the temporal resolution of VIC runoff aggregation and streamflow evaluation.

3.3 Calibration experiments and evaluation strategy

Three calibration approaches were compared using the same VIC – RVIC model setup, parameter space, calibration and validation periods, and evaluation metrics. The same model-evaluation budget was applied across all approaches.

(1) Skill-guided LLM

The proposed Skill-guided LLM approach integrates hydrological Skills with LLM reasoning to perform process-aware diagnosis, hypothesis generation, constrained parameter adjustment, and hypothesis verification. The LLM temperature was set to 0 to reduce sampling variability and improve the consistency of calibration decisions. The calibration followed a closed-loop workflow in which simulation results and hypothesis-verification outcomes were fed back to guide subsequent iterations.

(2) NoSkill-LLM

The NoSkill-LLM approach was designed as an LLM-assisted calibration method without hydrological Skill guidance. The LLM received a predefined prompt containing the current calibration information and task description and was asked to recommend parameter adjustments directly. These recommendations were manually applied in VIC – RVIC simulations without the automated trajectory-based feedback mechanism used in the Skill-guided LLM approach.

(3) SCE-UA

The Shuffled Complex Evolution algorithm (SCE-UA) was used as a conventional global optimization method for comparison.

The comparison therefore evaluates the performance of the implemented calibration workflows rather than isolating the contribution of hydrological Skills alone.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

4 Results

4.1 Performance comparison of parameter optimization approaches

All optimization approaches were evaluated using the same VIC–RVIC configuration, parameter space, calibration period, and evaluation budget to ensure a fair comparison. The baseline VIC–RVIC simulation and three optimization strategies were compared, including the proposed Skill-guided LLM framework, an LLM without hydrological Skill guidance, and the conventional SCE-UA optimization method.

As summarized in Table 1, the Skill-guided LLM achieved the highest NSE value (0.716) among all evaluated approaches, followed by the LLM without Skill guidance (0.706). Both LLM-based approaches improved the baseline simulation, while the conventional SCE-UA method did not identify an improved parameter configuration under the given evaluation budget, with its best solution remaining identical to the initial parameter set (NSE = 0.690). These results indicate that incorporating LLM-based reasoning into the parameter adjustment process provided additional capability for exploring the constrained VIC parameter space compared with conventional optimization under limited evaluation budgets.

Beyond the overall efficiency metric, the Skill-guided LLM showed a distinct advantage in representing flood response timing. The mean absolute peak timing error was reduced from 23.22 days for the baseline simulation to 9.89 days after Skill-guided optimization. In contrast, the LLM without Skill guidance improved NSE but did not show a comparable improvement, suggesting that numerical optimization alone was insufficient to improve the temporal characteristics of flood responses. The improved peak timing representation indicates that incorporating hydrological error signatures and constrained parameter-action reasoning helped guide the LLM toward parameter adjustments associated with catchment response dynamics.

However, the improvement in overall hydrograph agreement was not uniformly reflected across all hydrological signatures. Compared with the baseline simulation, the Skill-guided LLM resulted in larger deviations in peak-flow magnitude and low-flow reproduction. Specifically, the peak-flow error changed from −7.29% to −15.92%, while the low-flow error changed from −11.38% to −19.01%. These results suggest that the optimization process primarily reduced errors associated with the overall hydrograph evolution rather than simultaneously minimizing all individual hydrological characteristics.

Table 1. Performance comparison of optimization approaches.

Method	NSE	KGE	PBIAS (%)	Peak error (%)	Low-flow error (%)	Peak timing error (d)
Baseline	0.690	0.787	+0.32	-7.29	-11.38	23.22
SCE-UA	0.690	0.787	+0.32	-7.29	-11.38	23.22
LLM without Skill (best GPT)	0.706	0.770	-0.69	-14.51	-10.67	23.22
Skill-guided LLM (best DS)	**0.716**	0.774	-4.81	-15.92	-19.01	**9.89**

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

285 The Skill-guided LLM improved overall streamflow agreement and substantially enhanced the representation of flood peak timing, while residual discrepancies in peak magnitude and low-flow simulation revealed trade-offs associated with single-objective optimization (Fig. 4). Such trade-offs are expected when optimizing a single objective function such as NSE (detailed results for the calibration period are presented in Table A1 of the Supplementary Material.), which emphasizes overall temporal agreement between simulated and observed streamflow but does not explicitly optimize individual
290 signatures such as peak magnitude or low-flow conditions.

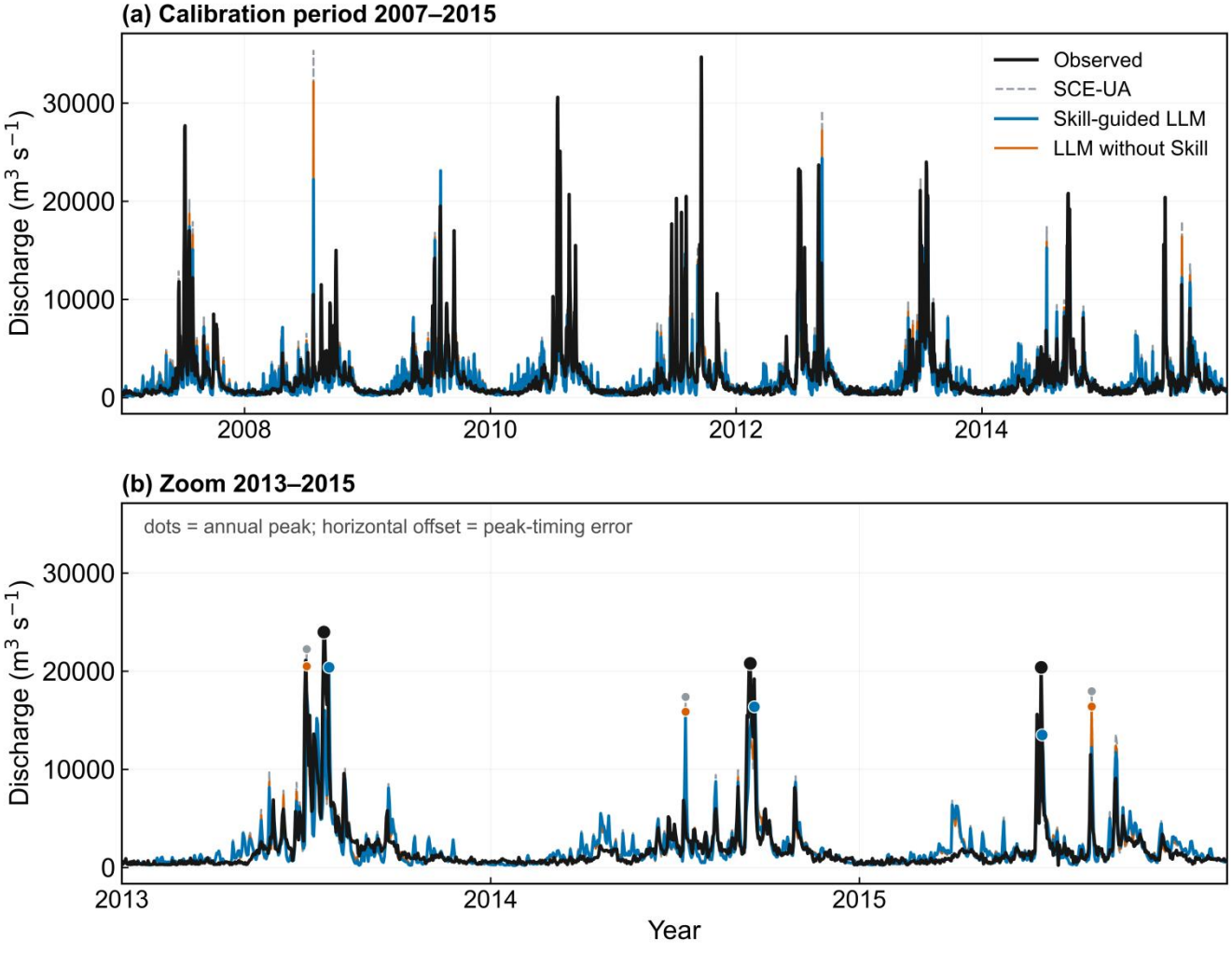


Figure 4: Comparison of observed and simulated streamflow hydrographs obtained using different calibration approaches during the calibration period, with an enlarged view of representative events (2013-2015).

Validation period results (Table A2), which assess the transferability of the optimized parameter configurations, show that
295 NSE improvements were not consistently maintained across all configurations; however, the improved peak timing

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

representation observed during calibration was largely preserved, suggesting that the framework primarily enhanced the temporal consistency of simulated hydrological responses.

4.2 Analysis of optimization trajectories and parameter adjustments

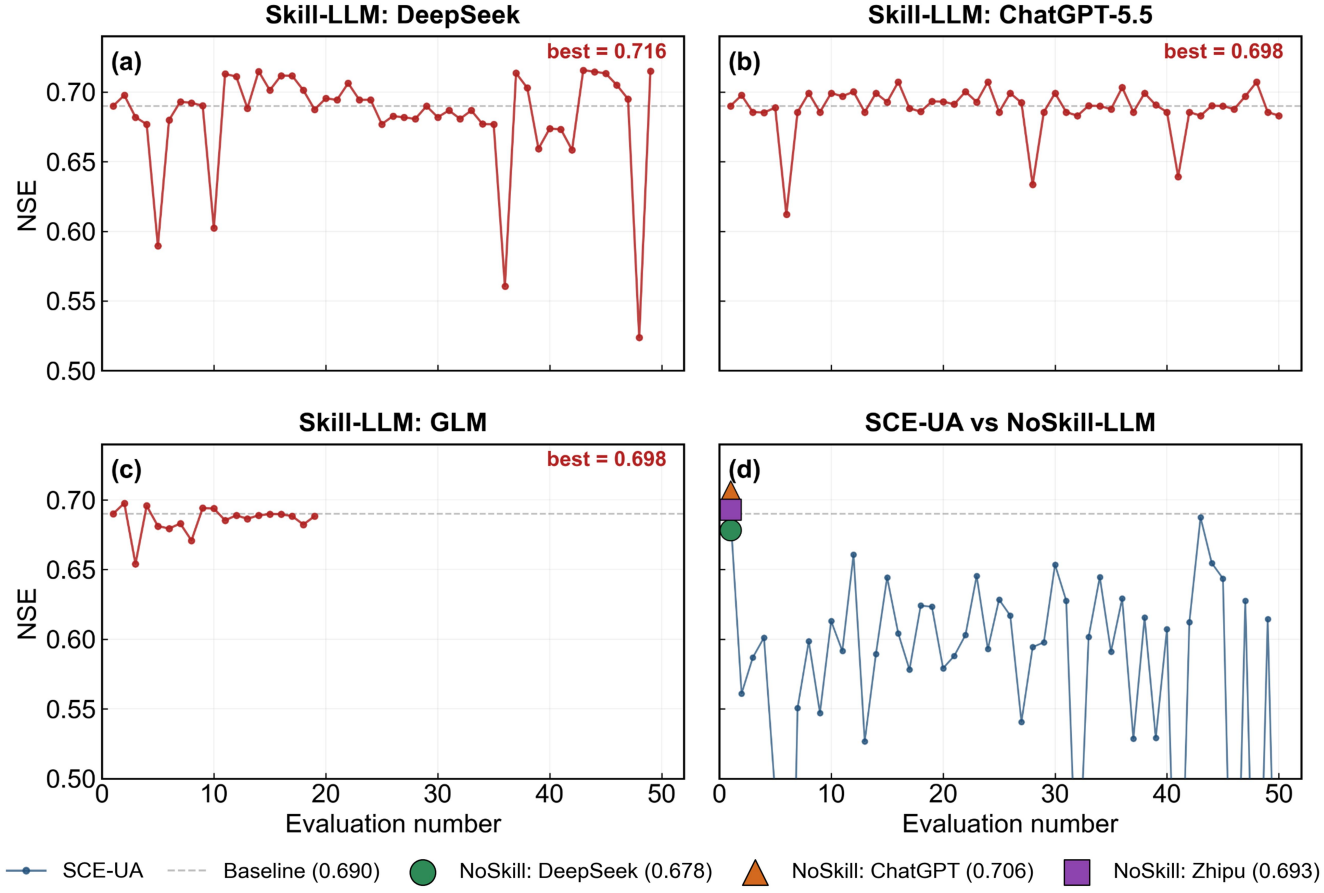
Beyond the final calibration performance, the optimization trajectories provide insights into how different approaches explored the VIC parameter space and how the Skill-guided LLM translated hydrological diagnostics into parameter adjustments. Unlike conventional calibration approaches that only provide the final parameter configuration, the proposed framework records intermediate parameter proposals, model responses, and hypothesis verification outcomes, allowing an explicit analysis of the calibration process.

4.2.1 Optimization trajectory of different calibration approaches

Among the tested Skill-guided LLM configurations, DeepSeek-v4-flash achieved the highest calibration performance, reaching a maximum NSE of 0.716. The optimization trajectory shows that high-performing solutions were identified during the later exploration stage rather than obtained from the initial parameter configuration.

Other Skill-guided LLM (ChatGPT-5.5 and GLM-5.3-flash) implementations achieved moderate improvements, with maximum NSE values of 0.698, indicating that the effectiveness of Skill-guided LLM depends not only on the reasoning framework but also on the capability of the underlying LLM. In comparison, SCE-UA remained close to the initial solution within the available evaluation budget, while NoSkill-LLM exhibited a wider but less constrained exploration pattern (Fig. 5). These results suggest that the hydrological Skill does not simply increase the number of parameter trials but guides the search toward more hydrologically plausible regions of the parameter space.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.



315 **Figure 5: Compares the optimization trajectories of SCE-UA, LLM without Skill, and Skill-guided LLM under their respective evaluation budgets.**

4.2.2 Parameter exploration trajectories

To understand which parameters contributed to the improved calibration performance, the evolution of the five VIC parameter multipliers during Skill-guided LLM was examined.

320 The largest adjustment occurred in depth, which was repeatedly modified during the exploration process and finally retained at 1.8 times the initial value. This behavior indicates that soil storage capacity was identified as an important control for correcting the diagnosed recession and low-flow-related errors. Ws was gradually increased during the search and approached its prescribed upper boundary. This suggests that the optimization relied on modifying soil moisture conditions controlling nonlinear baseflow generation. Dsmax showed moderate exploration and remained close to the upper range, 325 suggesting a potential role in regulating recession characteristics. In contrast, infilt and Ds experienced exploratory changes but showed limited deviations in the final retained solution, suggesting that these parameters were not the dominant controls under the calibrated configuration (Fig. 6).

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Overall, the parameter trajectories demonstrate that the Skill-guided LLM did not uniformly perturb all calibration parameters but concentrated the search on hydrologically interpretable parameters related to storage and baseflow processes.

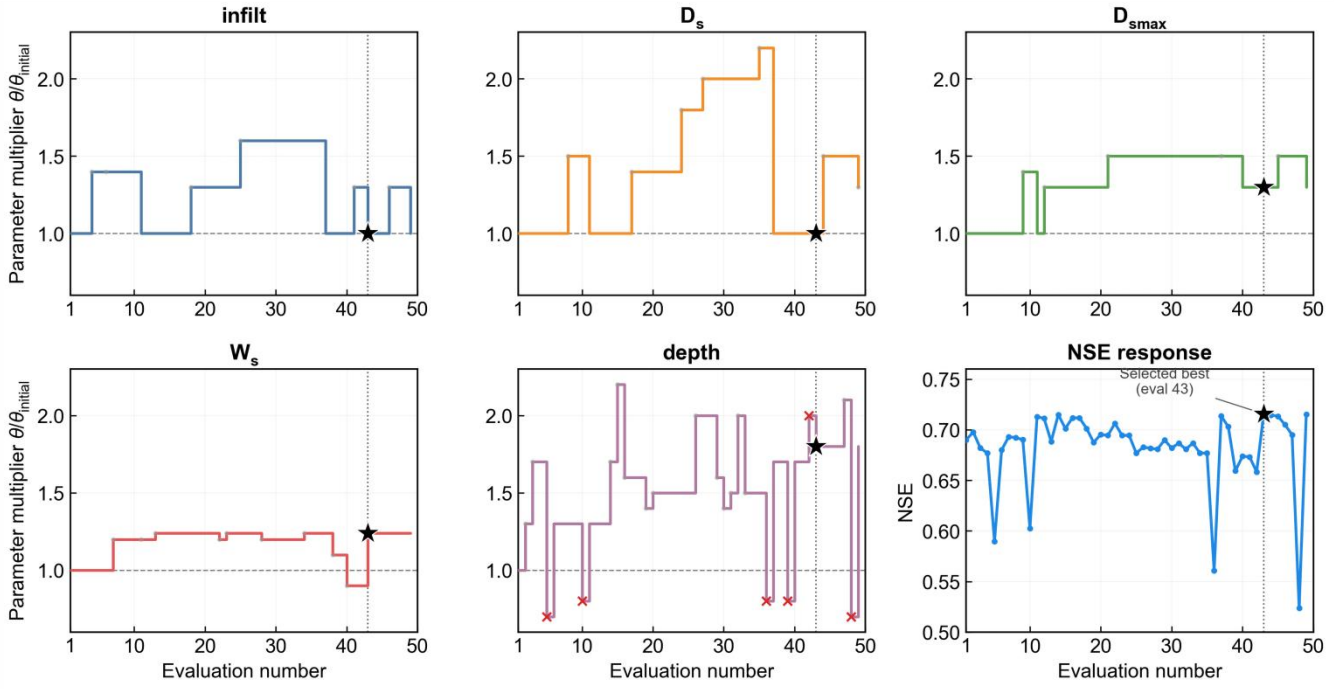


Figure 6: Parameter exploration trajectories and corresponding NSE responses during Skill-guided LLM calibration. The trajectories show the actual parameter values evaluated at each trial rather than the best-so-far solutions. Crosses denote rejected proposals, and the star marks the selected best solution.

4.2.3 Diagnosis-driven parameter adjustment

While Figure 6 illustrates what parameters were adjusted, it does not explain why these adjustments were proposed. Therefore, representative optimization episodes were further analyzed to reveal the diagnosis–action–verification mechanism of the Skill-guided LLM (Fig. 7).

For example, at iteration 2, recession bias was diagnosed and depth was increased (×1.3). The subsequent model response supported the hypothesis, and the proposed adjustment was accepted, resulting in an improved NSE. In contrast, the increase of Ws at iteration 11 (×1.2) resulted in a supported hypothesis, indicating that the adjustment improved the targeted hydrological behavior. At iteration 43, the framework identified combined recession and low-flow issues and proposed increasing depth to 1.8. This configuration produced the best NSE and was retained as the final solution (Fig. 7). Importantly, not all parameter proposals were accepted. Later adjustments, such as the depth reduction at iteration 48 (×0.7), were rejected after evaluation (due to diagnosed peak bias), demonstrating that the proposed framework incorporates an explicit hypothesis verification mechanism rather than blindly accepting LLM-generated parameter changes.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

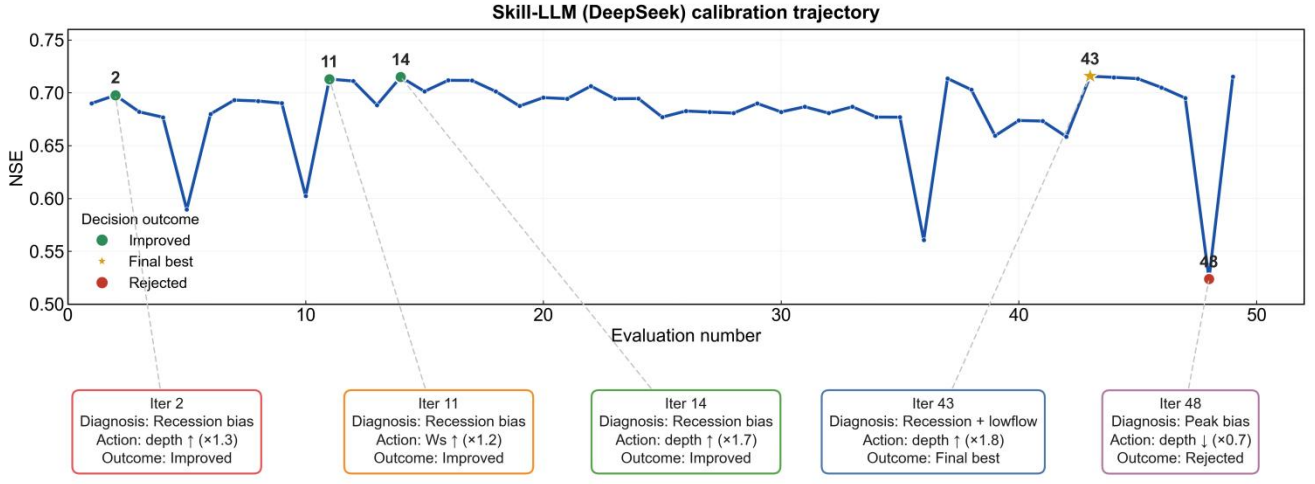


Figure 7: Representative diagnosis–action–verification records from the Skill-guided calibration run.

4.2.4 Final parameter states and hydrological interpretatio

The representative final parameter configurations of the major calibration approaches are summarized in Table 2. The complete parameter configurations for all tested LLM variants are provided in Table 3.

The Skill-guided LLM solution showed substantial adjustments in storage-related parameters compared with the initial configuration. In particular, depth was increased to enhance catchment storage representation, while Ws and Dsmax were adjusted to regulate soil moisture and baseflow responses. In contrast, SCE-UA retained the initial parameter configuration under the given evaluation budget, whereas the LLM without Skill approach produced a different adjustment pattern without explicit hydrological diagnosis.

Table 2. Representative selected best configuration obtained by different calibration approaches.

Parameter	Initial	Skill-guided LLM (best DS)	LLM without Skill (best GPT)	SCE-UA
depth	1.0	1.8	1.1	1.0
infilt	1.0	1.0	0.85	1.0
W_s	1.0	1.24	1.05	1.0
D_{smax}	1.0	1.3	1.2	1.0
D_s	1.0	1.0	1.1	1.0

The different parameter states obtained by each calibration strategy also provide a basis for investigating parameter sensitivity and associated hydrological responses.

Table 3. Final parameter configurations of all optimization experiments.

Parameter	Skill-guided LLM (GLM)	Skill-guided LLM (GPT)	LLM without Skill (DS)	LLM without Skill (GLM)
depth	1.3	1.3	1.2	1.2

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

infilt	1.0	1.0	0.9	1.3
W_s	1.0	1.0	0.7	1.1
D_{smax}	1.0	1.0	1.5	1.2
D_s	1.0	1.0	1.0	0.8

4.3 Local Sensitivity Analysis and Hydrological Interpretation

To evaluate whether the parameter adjustment space used by the Skill-guided LLM corresponds to physically meaningful hydrological controls, a local one-at-a-time (OAT) sensitivity analysis was conducted around both the initial VIC parameter configuration and the final optimized parameter configuration. The analysis focuses on the relationship between parameter perturbations and hydrological responses, rather than providing a global ranking of VIC parameter importance.

In this experiment, five adjustable VIC parameters were evaluated. For the initial configuration, each parameter was perturbed by ±20% while keeping all other parameters unchanged. The same procedure was repeated around the final Skill-guided LLM optimized parameter set (DeepSeek). For parameters located at prescribed upper bounds (Dsmax and Ws), only feasible downward perturbations were evaluated (Table 4).

Table 4. Comparison of local sensitivity before and after optimization.

Parameter	Baseline OAT ΔNSE	Optimized OAT ΔNSE	Hydrological interpretation
depth	+0.0331	+0.0019	Dominant storage control in the initial state; sensitivity largely reduced after optimization.
infilt	-0.0206	-0.0025	Controls runoff partitioning; initial correction potential was reduced after optimization.
W_s	+0.004	-0.0372 (downward only)	Strong boundary-sensitive parameter; optimized solution relies on a high Ws value.
D_{smax}	+0.003	+0.0009 (downward only)	Limited local influence around the optimized state.
D_s	-0.001	-0.0006	Minor effect on NSE under the evaluated configuration.

Among the five parameters, Depth exhibited the strongest sensitivity under ±20% perturbations, with NSE values of 0.6647 and 0.6978 for the − 20% and +20% cases, respectively, compared with 0.690 for the baseline configuration. This improvement was mainly associated with enhanced low-flow simulation, indicating that soil water storage was an important source of initial model error. In addition, decreasing infilt improved NSE, suggesting that runoff generation and water partitioning were also influential during the initial stage (Fig. 8). The corresponding process-level responses further show that these perturbations affected peak-flow error, low-flow error, and recession behavior in different ways (Fig. B1).

Around the final Skill-guided LLM parameter set (NSE = 0.716), the sensitivity of most parameters became substantially weaker. Perturbations of infilt, depth, and Ds resulted in NSE changes smaller than 0.003, indicating that optimization moved the model toward a locally stable parameter region.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

An exception was Ws. A downward perturbation of Ws reduced NSE by 0.0372, while further upward perturbation was infeasible because the parameter had reached its prescribed upper bound. This boundary-constrained sensitivity suggests that the optimized solution relies strongly on the soil moisture threshold represented by Ws.

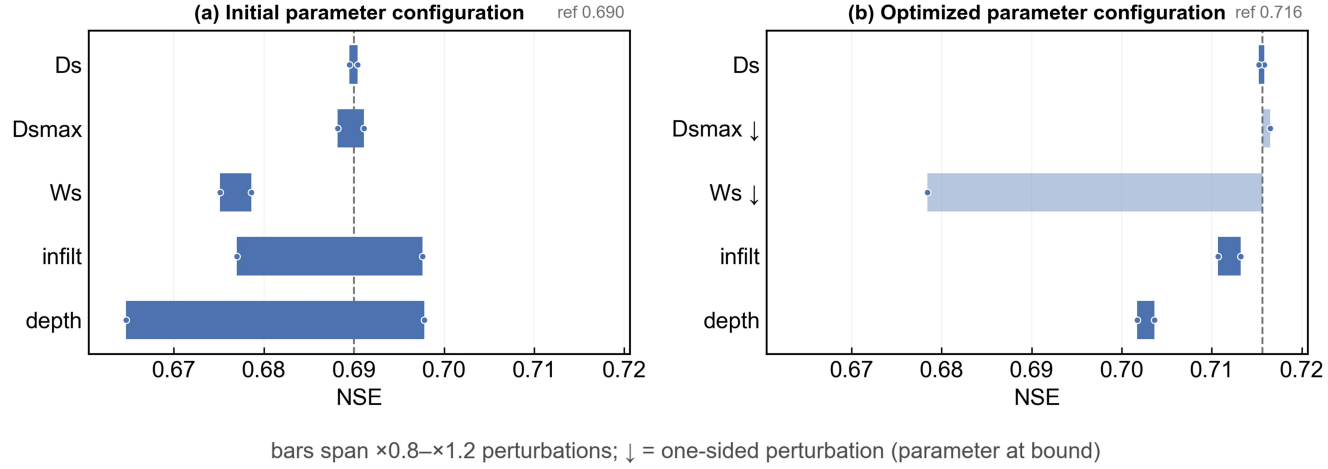


Figure 8: Comparison of local VIC parameter sensitivity around the initial and Skill-guided LLM-optimized configurations based on NSE responses to one-at-a-time parameter perturbations.

The sensitivity analysis demonstrates that parameter importance is state dependent. The initial configuration was mainly controlled by storage and runoff partition parameters, whereas the optimized configuration showed reduced local sensitivity except for the boundary-constrained Ws parameter. These findings support the design of Skill-guided LLM adjustment, where parameter actions are associated with specific hydrological error signatures rather than treated as unconstrained parameter searches.

5 Dsicussion

5.1 From expert-dependent calibration to knowledge-guided agents

Although conventional optimization algorithms automate parameter searching, hydrological model calibration still relies heavily on expert knowledge for diagnosing simulation errors, interpreting model behavior, and identifying physically meaningful parameter adjustments (Efstratiadis and Koutsoyiannis, 2010; Beven and Freer, 2001). The proposed Skill-guided LLM framework provides a means of incorporating such expertise into an automated calibration agent. In this approach, expertise is represented not as fixed optimal parameter values but as procedural guidance that links error signatures, plausible process explanations, admissible actions, and model-based tests. Rather than focusing solely on numerical optimization, it integrates hydrological reasoning, constrained parameter adjustment, and iterative verification into a knowledge-guided calibration process.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

More importantly, agent-based calibration offers a potential pathway toward making hydrological models more accessible to non-specialist users. By embedding domain knowledge into reusable Skills, such agents could reduce the technical expertise required for routine calibration and facilitate a transition from expert-dependent workflows toward broader adoption of hydrological modeling. Therefore, the broader value of knowledge-guided calibration agents lies not only in improving calibration efficiency and interpretability, but also in lowering barriers to the practical application of hydrological models.

5.2 Knowledge-guided calibration behavior and performance trade-offs

The results demonstrate that Skill guidance improved calibration performance compared with both conventional optimization and LLM calibration without explicit hydrological Skills. The improvement in NSE indicates that the framework was able to identify parameter configurations with better overall hydrograph consistency under the same computational budget.

However, the improvement was not uniform across all hydrological signatures. The Skill-guided LLM achieved better flood response timing representation but showed trade-offs in peak magnitude and low-flow simulation. This behavior reflects a fundamental challenge in hydrological calibration: optimizing a single objective function such as NSE may improve overall agreement while affecting specific components of the hydrograph (Ved and Vimal, 2023; Naz et al., 2018).

The advantage of the proposed framework therefore does not come from simply searching more parameter combinations. Instead, the trajectory analysis shows that the LLM performed diagnosis-driven exploration. Parameter adjustments were associated with hydrological hypotheses, and subsequent simulations were used to evaluate whether these hypotheses were supported or rejected. This closed-loop process provides an interpretable mechanism for calibration decisions.

5.3 Hydrological interpretation of parameter adjustments

The parameter trajectories indicate that the LLM did not modify all parameters equally. Instead, adjustments concentrated on parameters associated with catchment storage and runoff response, particularly soil depth and baseflow-related parameters. This behavior is consistent with the hydrological interpretation that soil storage controls water retention, recession behavior, and delayed runoff responses.

The sensitivity analysis further reveals that parameter importance is state-dependent (Sepúlveda et al., 2022). Around the initial parameter configuration, soil depth and infiltration parameters showed stronger influences on NSE, whereas the optimized parameter state exhibited different sensitivity patterns. This finding suggests that sensitivity analysis based only on initial parameter states may not fully represent the behavior of optimized hydrological models.

The state-dependent sensitivity also explains why parameter exploration and final parameter selection should not be interpreted independently. A parameter that is highly sensitive does not necessarily require a large adjustment, while a parameter with moderate sensitivity may become critical near an optimized state (Mai et al., 2022). The combination of trajectory analysis and sensitivity analysis therefore provides complementary information for understanding model calibration behavior.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

5.4 Limitations and future perspectives

Several limitations remain. First, the present experiments primarily optimize NSE, whereas other hydrological signatures are used diagnostically; this can lead to trade-offs in peak-flow magnitude, low-flow behavior, and water-balance bias. Second, the current framework relies on short-term optimization history within individual calibration tasks and does not yet support long-term knowledge accumulation across different basins and calibration cases. Future agent architectures may incorporate reusable calibration experiences across watersheds while maintaining physical constraints. Third, the present study focuses on one distributed hydrological model and one basin. Further evaluation across different models, and catchments is required to assess the transferability of Skill-guided calibration.

Overall, this study demonstrates that LLMs can contribute to hydrological model calibration when embedded within a physically constrained and knowledge-guided workflow. The key contribution is not replacing traditional calibration algorithms, but enabling a new form of interpretable calibration agent that combines hydrological expertise, language-based reasoning, numerical simulation, and iterative verification.

6 Conslusion

This study developed an agent-based framework for process-aware hydrological model calibration, centered on a Skill-guided LLM agent. The main methodological contribution lies in transforming hydrological calibration expertise into explicit procedural knowledge that can guide the agent in formulating and evaluating model-based parameter interventions. Within this framework, domain-specific Skills support the agent in diagnosing simulation errors, generating hydrological hypotheses, and selecting process-relevant parameter adjustments, while deterministic modeling tools execute the proposed interventions and evaluate their hydrological responses. This integration establishes an explicit and traceable connection between hydrological diagnosis, parameter decisions, and simulation feedback.

The proposed approach was demonstrated using VIC v5.1.0 coupled with RVIC in the Jialing River Basin. The results showed that the agent could generate hydrologically interpretable parameter adjustments, evaluate their consequences through numerical simulations, and use the resulting feedback to inform subsequent calibration decisions. Analysis of the calibration trajectories illustrated the diagnosis–action–verification process, while local sensitivity analysis revealed state-dependent parameter influences. Although improvements were not consistent across all hydrological signatures, the experiments demonstrated the feasibility of incorporating explicit hydrological reasoning into an iterative model calibration process.

More broadly, this study provides a methodological basis for integrating domain knowledge, LLM-based reasoning, and numerical simulation into knowledge-guided hydrological modeling. By making expert calibration knowledge actionable and model-based decisions traceable, the proposed approach offers a potential pathway toward reducing dependence on manual trial-and-error and specialized calibration experience. Although demonstrated using a single hydrological model and

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

basin, the agent-based design provides a basis for future adaptation to other process-based hydrological models. Further evaluation across different models and catchments is needed to establish its generality and transferability.

Appendix A: Additional results

Table A1. Performance comparison of optimization approaches (calibration period).

Method	NSE	KGE	PBIAS (%)	Peak-flow err (%)	Low-flow err (%)	Peak timing err (d)
SCE-UA (baseline)	0.690	0.787	+0.32	-7.29	-11.38	23.22
Skill-guided LLM (DeepSeek)	**0.716**	0.774	-4.81	-15.92	-19.01	**9.89**
Skill-guided LLM (GPT-4o/GLM)	0.698	0.736	-0.44	-17.49	+3.44	23.22
LLM without Skill (DeepSeek)	0.678	0.704	+0.27	-18.33	+25.43	23.22
LLM without Skill (ChatGPT)	0.706	0.770	-0.69	-14.51	-10.67	23.22
LLM without Skill (GLM)	0.693	0.802	+0.73	-8.14	-18.64	23.22

Table A2. Performance comparison of optimization approaches (validation period).

Method	NSE	KGE	PBIAS (%)	Peak-flow err (%)	Low-flow err (%)	Peak timing err (d)
SCE-UA (baseline)	0.644	0.788	-5.36	-1.90	-18.42	59.00
Skill-guided LLM (DeepSeek)	0.644	0.793	-5.68	+4.01	-15.99	27.00
Skill-guided LLM (GPT-4o/GLM)	0.659	0.759	-5.67	-11.47	-6.70	27.00
LLM without Skill (DeepSeek)	0.659	0.724	-5.89	-14.94	+9.54	27.00
LLM without Skill (ChatGPT)	0.668	0.781	-6.74	-2.27	-15.09	27.00
LLM without Skill (GLM)	0.614	0.794	-3.31	-1.68	-22.94	59.00

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

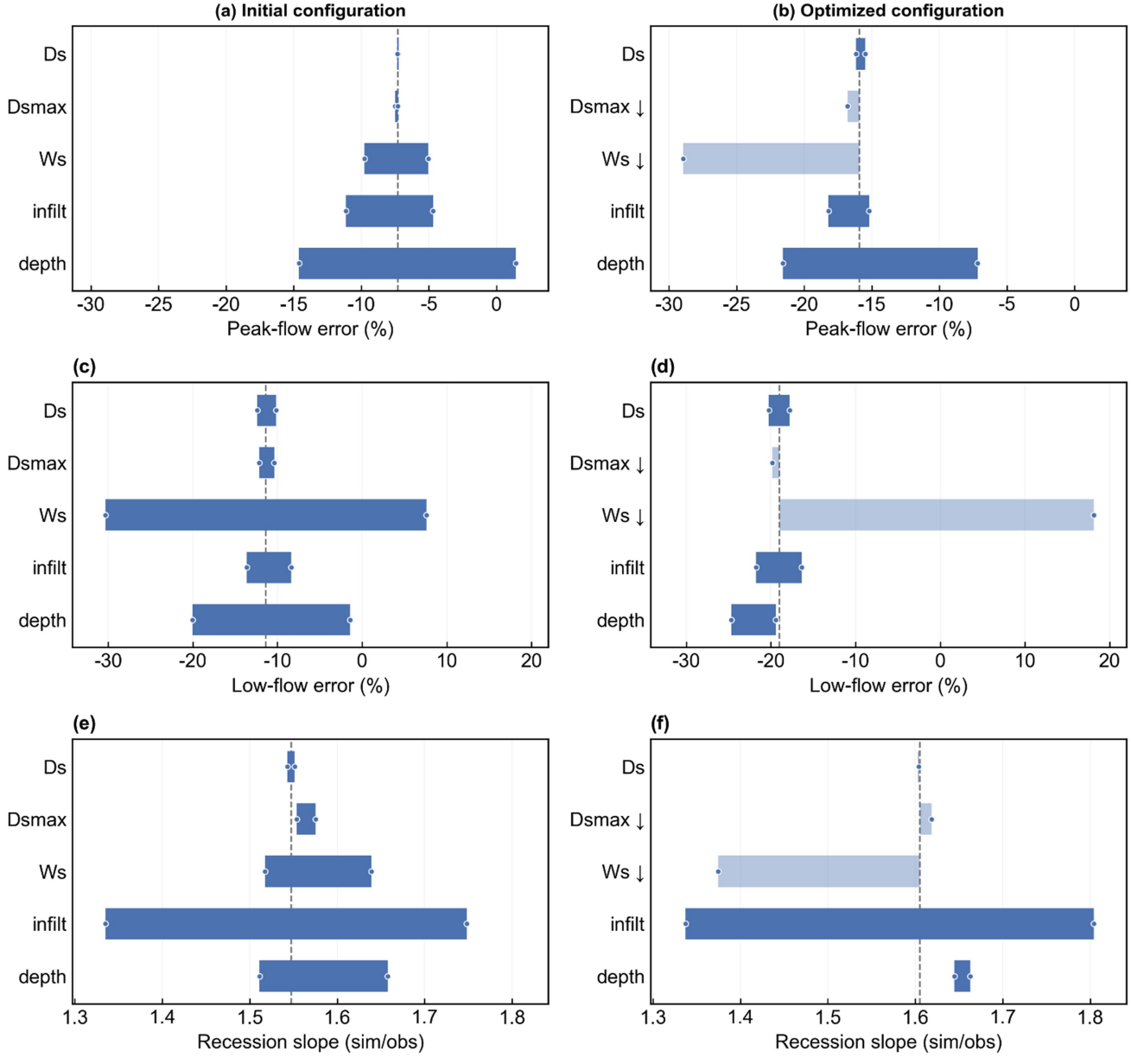


Figure B1: Process-level hydrological responses to local VIC parameter perturbations around the initial and Skill-guided LLM optimized configurations. The left and right columns show responses around the initial and optimized parameter configurations, respectively. Rows show changes in peak-flow error, low-flow error, and recession slope ratio (simulated/observed). Horizontal bars indicate the response ranges induced by local one-at-a-time parameter perturbations, while downward arrows denote one-sided perturbations for parameters constrained by the prescribed bounds.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Code and data availability

The source code of the HydroSkill-Agent framework (version v1.0.0) is archived on Zenodo (Qian et al., 2026), with a permanent DOI: https://doi.org/10.5281/zenodo.23919903. The development repository is available at https://github.com/lianyi732-hue/HydroSkill-Agent (last access: 31 October 2026). Other data generated or analyzed during this study are available from the corresponding author upon reasonable request.

Author contributions

CQ contributed to the investigation, visualization, and validation. GP provided supervision, acquired funding, and contributed to reviewing and editing the manuscript. YS made substantial contributions to the design and development of the proposed model framework, including methodology development and software implementation, as well as data curation, formal analysis, and writing of the original manuscript. CZ contributed to the conceptualization of the study and software development. JL was responsible for project administration and resources.

Competing interests

The contact author has declared that none of the authors has any competing interests.

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

Acknowledgement for the data support from "National Earth System Science Data Center, National Science & Technology Infrastructure of China. (https://www.geodata.cn; last access: 30 October 2026)".

Financial support

This research was funded by the National Natural Science Foundation of China (grant number 42201517).

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

References

Beven, K., Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE metho dology, J. Hydrol., 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Clark, M. P., Bierkens, M. F. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N., Cai, X., Wood, A. W., Peters-Lidard, C. D.: The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism, Hydrol. Earth Syst. Sci., 21, 3427–3440, https://doi.org/10.5194/hess-21-3427-2017, 2017.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., Rasmussen, R. M.: A unified approach for process-based hydrologic modeling: 1. Modeling concept, Water Resour. Res., 51(4), 2498–2514, https://doi.org/10.1002/2015WR017198, 2015.

Duan, Q., Sorooshian, S., Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28(4), 1015–1031, https://doi.org/10.1029/91WR02985, 1992.

Efstratiadis, A., Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review, Hydrol. Sci. J., 55(1), 58-78, https://doi.org/10.1080/02626660903526292, 2010.

Ercolani, G., Castelli, F.: Variational assimilation of streamflow data in distributed flood forecasting, Water Resour. Res., 53(1), 158-183, https://doi.org/10.1002/2016WR019208, 2017.

Farahani, M. A., Wood, A. W., Tang, G., Mizukami, N.: Calibrating a large-domain land/hydrology process model in the age of AI: The SUMMA CAMELS emulator experiments, Hydrol. Earth Syst. Sci., 29, 4515–4537, https://doi.org/10.5194/hess-29-4515-2025, 2025.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., Downer, C. W., Camporese, M., Davison, J. H., Ebel, B., Jones, N., Kim, J., Mascaro, G., Niswonger, R., Restrepo, P., Rigon, R., Shen, C., Sulis, M., Tarboton, D.: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, J. Hydrol., 537, 45–60, https://doi.org/10.1016/j.jhydrol.2016.03.026, 2016.

Foroumandi, E., Moradkhani, H., Sanchez-Vila, X., Singha, K., Castelletti, A., Destouni, G.: ChatGPT in hydrology and Earth sciences: Opportunities, prospects, and concerns, Water Resour. Res., 59, e2023WR036288, https://doi.org/10.1029/2023WR036288, 2023.

Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., Wu, J., Zhou, R.: Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China, Water Resour. Res., 56, e2019WR025968, https://doi.org/10.1029/2019WR025968, 2020.

Gupta, H. V., Kling, H., Yilmaz, K. K., Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377(1–2), 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.

Gupta, H. V., Sorooshian, S., Yapo, P. O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 34(4), 751–763, https://doi.org/10.1029/97WR03495, 1998.

Guse, B., Han, L., Kumar, R., Rakovec, O., Luedtke, S., Herzog, A., Thober, S., Samaniego, L., Wagener, T.: Spatio-temporal consistency and variability in parameter dominance on simulated hydrological fluxes and state variables, Water Resour. Res., 60(12), e2023WR036822, https://doi.org/10.1029/2023WR036822, 2024.

Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., Mao, Y.: The Variable Infiltration Capacity model version 5 (VIC-5): Infrastructure improvements for new applications and reproducibility, Geosci. Model Dev., 11, 3481–3496, https://doi.org/10.5194/gmd-11-3481-2018, 2018.

Herrera, P. A., Marazuela, M. A., Hofmann, T.: Parameter estimation and uncertainty analysis in hydrological modeling, WIREs Water, 9, e1569, https://doi.org/10.1002/wat2.1569, 2022.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resour. Res., 50(9), 7445–7469, https://doi.org/10.1002/2014WR015484, 2014.

Jiang, P., Shuai, P., Sun, A., Mudunuru, M. K., Chen, X.: Knowledge-informed deep learning for hydrological model calibration: An application to Coal Creek Watershed in Colorado, Hydrol. Earth Syst. Sci., 27, 2621–2644, https://doi.org/10.5194/hess-27-2621-2023, 2023.

Knoben, W. J. M., Clark, M. P., Bales, J., Bennett, A., Gharari, S., Marsh, C. B., Nijssen, B., Pietroniro, A., Spiteri, R. J., Tang, G., Tarboton, D. G., Wood, A. W.: Community workflows to advance reproducibility in hydrologic modeling: Separating model-agnostic and model-specific configuration steps in applications of large-domain hydrologic models, Water Resour. Res., 58, e2021WR031753, https://doi.org/10.1029/2021WR031753, 2022.

Liang, X., Lettenmaier, D. P., Wood, E. F., Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res. Atmos., 99(D7), 14415–14428, https://doi.org/10.1029/94JD00483, 1994.

Mai, J., Craig, J. R., Tolson, B. A., Arsenault, R.: The sensitivity of simulated streamflow to individual hydrologic processes across North America, Nat. Commun., 13(1), https://doi.org/10.1038/s41467-022-28010-7, 2022.

McMillan, H.,K.: A review of hydrologic signatures and their applications, WIREs Water, 8, e1499, https://doi.org/10.1002/wat2.1499, 2022.

Mizukami, N., Clark, M.P., Newman, A.J., Wood, A.W., Gutmann, E.D., Nijssen, B., Rakovec, O., Samaniego, L.: Towards seamless large-domain parameter estimation for hydrologic models, Water Resour. Res., 53(9), 8020-8040, https://doi.org/10.1002/2017WR020401, 2017.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., Kumar, R.: On the choice of calibration metrics for “high-flow” estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601-2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.



Naz, B. S., Kao, S. C., Ashfaq, M., Gao, H., Rastogi, D., Gangrade, S.: Effects of climate change on streamflow extremes and implications for reservoir inflow in the United States, J. Hydrol., 556, 359-370, https://doi.org/10.1016/j.jhydrol.2017.11.027, 2018.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., Gupta, H. V.: What role does hydrological science play in the age of machine learning? Water Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

Qian, C., Peng, G., Sun, Y., Zeng, C., and Liu, J.: lianyi732-hue/HydroSkill-Agent: HydroSkill-Agent v1.0.0, Zenodo [code and data set], https://doi.org/10.5281/zenodo.23919903, 2026.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resour. Res., 46, W05521, https://doi.org/10.1029/2009WR008328, 2010.

Sepúlveda, U.M., Mendoza, P.A., Mizukami, N., Newman, A.J.: Revisiting parameter sensitivities in the variable infiltration capacity model across a hydroclimatic gradient, Hydrol. Earth Syst. Sci., 26, 3419–3445, https://doi.org/10.5194/hess-26-3419-2022, 2022.

Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., Yao, S.: Reflexion: Language agents with verbal reinforcement learning, Adv. Neural Inf. Process. Syst, 36, 8634-8652, https://doi.org/10.52202/075280-0377, 2023.

Tsai, W.-P., Feng, D., Pan, M., Beck, H., Lawson, K., Yang, Y., Liu, J., Shen, C.: From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling, Nat. Commun., 12, 5988, https://doi.org/10.1038/s41467-021-26107-z, 2021.

Ved, P., Vimal, M.: Soil Moisture and Streamflow Data Assimilation for Streamflow Prediction in the Narmada River Basin, J. Hydrometeorol., 24(8), 1377-1392, https://doi.org/10.1175/JHM-D-22-0124.1, 2023.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., Anandkumar, A.: Voyager: An Open-Ended Embodied Agent with Large Language Models, arXiv [preprint], arXiv:2305.16291. https://doi.org/10.48550/arXiv.2305.16291, 2023.

Wang, J., Fu, G., Savic, D.: Leveraging large language models for automating water distribution network optimization, Water Res., 288, 124536, https://doi.org/10.1016/j.watres.2025.124536, 2026.

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents, Front. Comput. Sci., 18, 186345, https://doi.org/10.1007/s11704-024-40231-1, 2024.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models, Adv. Neural Inf. Process. Syst., 35, 24824-24837, https://doi.org/10.52202/068431-1800, 2022.

https://doi.org/10.5194/egusphere-2026-6762
Preprint. Discussion started: 6 November 2026
© Author(s) 2026. CC BY 4.0 License.



Yeste, P., Ojeda, M. G., Gámiz-Fortis, S. R., Castro-Díez, Y., Esteban-Parra, M. J.: Integrated sensitivity analysis of a macroscale hydrologic model in the north of the Iberian Peninsula, J. Hydrol., 590, 125230, https://doi.org/10.1016/j.jhydrol.2020.125230, 2020.

Yilmaz, K. K., Gupta, H. V., Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, https://doi.org/10.1029/2007WR006716., 2008.

Zhang, X., Srinivasan, R., Zhao, K., Liew, M. W.: Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model, Hydrol. Process., 23(3), 430–441, https://doi.org/10.1002/hyp.7152, 2009.

Zhu, Z., Tang, Y., Tang, X., Zhang, J., Gao, C., Zhang, S., Xu, H., Guan, T.: Large language models as calibration agents in hydrological modeling: Feasibility and limitations, Geophys. Res. Lett., 53, e2025GL120043, https://doi.org/10.1029/2025GL120043, 2026.