Response to editor

Re: Manuscript #hess-2024-384, entitled “A Robust Calibration and Evaluation Framework for Dynamic Catchment Characteristics in Hydrological Modeling.”

Dear Editor,

We sincerely thank both reviewers for recognizing the significance and innovation of this study, as well as for their time and effort. Since the first round of comments was returned to us in early April, we have devoted considerable time to undertaking systematic and substantial revisions. The manuscript has been carefully refined in response to all suggestions to improve its clarity, rigor, and overall contribution.

The manuscript has been reorganized around the central scientific question, and the chain of evidence has been refined to present a clearer and more rigorous narrative. The reviewers expressed encouragement regarding the study's novelty and scientific value, which is sincerely appreciated. In view of the comprehensive revisions undertaken in response to all comments, a careful reconsideration of the manuscript's content and quality is respectfully requested.

The overall structure and logic of the manuscript have been systematically revised. The transition between the Introduction and Methods sections has been rewritten to clarify the respective roles of data preprocessing, sub-period clustering, and calibration experiments within the study framework, forming a coherent chain from problem statement to methodological design, results, and scientific interpretation. To enhance readability, the approach for extracting dynamic catchment characteristics and its function within the calibration framework has been incorporated into the main Methods section in a concise form, with detailed technical procedures retained in the Supplementary Information. Moreover, we have standardized key terminologies, replacing all instances of “validation” with “evaluation”, and redrawn and split figures so that each focuses on a single information point, with clearer legends. This thereby reduces dependence on supplementary materials and strengthens the clarity of the methodological and analytical components.

Transparency of the methodology and consistency across experiments have also been enhanced. In response to the issues raised regarding insufficient detail and inconsistent descriptions of the experimental schemes, we have added a schematic diagram of the HYMOD model structure, a definition table for parameters, states, and fluxes, and a summary of the core logic of the calibration process in the Methods section. Descriptions of all seven experiments have been standardized, including objectives, optimization procedures, objective functions, and parameter treatments, and full correspondence between diagrams and text has been ensured. The SCE-UA algorithm was applied uniformly across all experiments to support comparability, and the Evaluation section now contains mathematical definitions and applicability conditions for each performance metric. These modifications enhance the traceability and reproducibility of the method and respond to the reviewers’ suggestion that the core methodology should appear in the main text.

Moreover, the scope and robustness of the results have been expanded. Diagnostic analyses previously limited to representative catchments have been extended to the full MOPEX dataset,

together with multi-metric evaluations over the entire study region. Sub-period clustering has been performed across all 219 MOPEX catchments, with 217 exhibiting significant seasonality. Summary statistics, including boxplots, have been evaluated at the basin scale. Process-level diagnostic outputs—such as flux mapping and state/flux time series—have been strengthened to substantiate the conclusion that the recommended experiment produces balanced improvements across flow phases and sub-periods based on internal model responses. Parameter transferability and performance consistency during the evaluation period have been further examined, contributing to broader generalizability and clearer hydrological interpretation.

In addition, we have enriched the discussion on academic depth, applicability boundaries, and methodological limitations. Compared to the original manuscript, the revised version adds discussions on key scientific topics such as parameter identifiability and model structural uncertainty. Diagnostic experiments have been used to illustrate how dynamic parameters can partially compensate for structural deficiencies, thereby adding conceptual depth and clarifying the interpretive value of the framework.

In summary, targeted and verifiable improvements have been made to the manuscript's structure, methodological rigor, evidential support, and theoretical interpretation. The revised version presents clearer arguments, a more coherent evidence chain, and a more rigorous assessment of hydrological processes, strengthening the contribution of the proposed framework to calibration and process diagnosis in dynamic catchments.

It is important to note that the current revision involved extensive technical work conducted over a prolonged period of seven months. To complete the required restructuring and recalculation, two additional researchers, Jiajia Zhang and Wenqing Cheng, were invited to contribute. They played central roles in reorganizing the seven calibration experiments, recalculating the full set of results for all experiments across 219 catchments, and designing and implementing the diagnostic analysis used in the Discussion section to examine parameter sensitivity to dynamic catchment characteristics. Their contributions were essential to the successful completion of the revision. With the agreement of all authors, their inclusion as co-authors is respectfully requested to duly acknowledge their significant and substantive contributions to the scientific content of the study.

A revised manuscript and detailed point-by-point responses have been submitted for consideration. Appreciation is expressed to the reviewers for their insightful and constructive comments, which substantially improved the scientific quality of the study and helped refine the direction of future research. Gratitude is also extended to the Handling Editor, Professor Efrat Morin, for providing the opportunity to undertake a comprehensive revision. A careful review of the revised manuscript and its reconsideration is respectfully requested.

Sincerely,

Xiao Wang

xiao_wang@whu.edu.cn

Reply to Reviewer1

Major Comments

Clarity and Structure

Q1: The manuscript's presentation is a major limitation to its scientific communication. While clarity should not override substance in peer review, in this case the lack of structure and clarity severely affects the reader's ability to assess the methods and results.

Response:

We sincerely thank you for your thorough evaluation of the manuscript's structure and scientific communication. In response to the constructive comments, the overall framework and logical flow of the manuscript have been extensively refined over the past seven months, and the necessary recalculations have been completed. Significant contributions are provided by two additional researchers during the revision. Their involvement enabled a comprehensive synthesis across all catchments and resulted in substantial improvements in the manuscript's clarity and structure in accordance with the reviewer's suggestions. The major concerns regarding insufficient structure and clarity have been fully addressed through these revisions.

In the revised manuscript, the entire structure has been reorganized around the main research line: how to effectively incorporate dynamic catchment characteristics into model calibration to enhance the hydrological model's ability to simulate different flow phases. The main adjustments are as follows:

(1) Methods section: The research design has been organized into four conceptual layers: (i) research tools, centered on the hydrological model and its functional modules; (ii) input preparation, featuring sub-period clustering informed by dynamic catchment characteristics; (iii) diagnostic experiment design, comprising seven comparative

calibration experiments; and (iv) the evaluation system, which integrates multi-criteria performance assessment with internal hydrological model diagnostics. The procedure for extracting dynamic catchment characteristics from the original manuscript has been integrated into a dedicated subsection, "3.2 Clustering hydrological processes," with additional methodological detail added to enhance structural clarity and strengthen the methodological progression.

(2) Results section: More technical details about the sub-period clustering are added, which are presented in the newly added section "4.1 Defined sub-periods based on catchment dynamics". The results for all experiments across all catchments are presented more comprehensively in section 4.2, including comparative assessments of model performance, parameter characteristics, and internal flux diagnostics across the full set of calibration experiments.

(3) Discussion section: The section has been refined to emphasize comparative analysis and scientific explanation, such as the impact of different calibration strategies on hydrological model stability and dynamic response capabilities, and the scientific significance of parameter uncertainty and equifinality issues. A focused discussion on parameter sensitivity to dynamic catchment characteristics has also been incorporated as section "5.3 Parameter response to catchment dynamics".

Through the series of adjustments, the revised manuscript forms a clearer logical progression from the formulation of the research question through methodological development, experimental evaluation, and scientific explanation. We hope that our revisions can improve the readability and scientific value of the manuscript. Detailed modifications have been incorporated throughout the revised manuscript, with all adjustments highlighted in red to facilitate identification and review.

Q2: The “Methods” section is disorganized and lacking in detail. For instance: Section 2 on Dynamic catchment characteristics should be integrated into the Methods, as it is a central component of the proposed framework.

Response:

We greatly appreciate your valuable suggestions regarding the manuscript’s structure. In the revised manuscript, comprehensive adjustments have been made to the relevant Method sections to address the concern about disorganization and insufficient detail. The “Dynamic catchment characteristics” part of Section 2 in the original manuscript has been moved into the new Methods section as subsection “3.2 Clustering hydrological processes,” where a complete description of the extraction of dynamic catchment characteristics and the sub-period clustering procedure is provided. The corresponding results of the hydrological process clustering have been relocated to Results 4.1 “Defined sub-periods based on catchment dynamics,” for a clearer and more detailed presentation. As an important preprocessing step within the proposed framework, the clustering of sub-periods forms the foundation for the subsequent seven experiments.

It should be noted that the study investigates calibration challenges in catchments with significant seasonal and interannual dynamic changes and proposes a structured framework to address two major issues: the influence of objective function design on flow-phase-specific performance, and the limitations of sub-period calibration with dynamic parameters. The analysis focuses on evaluating the applicability of different parameter calibration experiments. It also assesses whether sub-period calibration informed by dynamic catchment characteristics can more effectively reduce the mismatch between model structure and actual hydrological processes. While the approach of clustering hydrological processes is not the main innovation of this paper, it serves as a central preprocessing step within the proposed framework. Therefore, a concise but complete description of the index system, feature extraction, and clustering procedure is provided in Section 3.2, with detailed algorithmic settings included in the

Supporting Information to maintain clarity in the main text.

We sincerely appreciate the reviewer's suggestions once again. With the revisions completed, the manuscript now presents a more rigorous structure and clearer logical progression, thereby sharpening the central research theme of comparing the performance and mechanisms of different calibration experiments under dynamic catchment conditions while retaining the necessary background information. The revised sections have been updated accordingly in the main text.

Revised manuscript text:

3.2 Clustering hydrological processes

"Sub-period calibration provides a practical means of linking dynamic catchment characteristics with hydrological models. In sub-period calibration, the simulation period is clustered into multiple sub-periods characterized by relatively homogeneous hydrological conditions, allowing dynamic parameters to better reflect temporal variations in catchment behaviour across different streamflow regimes (Zhang and Liu, 2021). In this study, the clustering of sub-periods is guided by temporal variations in key hydrometeorological and land-surface variables. The methodological framework consists of three key steps: (1) constructing a dynamic catchment characteristic index system to describe catchment states; (2) extracting dynamic catchment characteristics through screening and dimensionality reduction; and (3) applying unsupervised clustering to cluster the time series into sub-periods with similar hydrological processes for subsequent sub-period calibration.

Describing catchment dynamics: *To characterize the temporal dynamics of catchment behaviour, a dynamic catchment characteristic index system comprising a climatic subsystem and a land-surface subsystem is constructed to represent the time-varying states of the catchment. The climatic subsystem includes core hydrometeorological variables such as precipitation (P), temperature (T), and potential evapotranspiration (PE), along with corresponding extreme climatic indicators. The land-surface subsystem reflects evolving*

surface conditions through indicators such as antecedent runoff, runoff coefficient, and the normalized difference vegetation index (NDVI). All indicators are sampled using a moving window approach, with the optimal window length determined through a time-windowed Bayesian inference framework based on predictive log-score (PLS) performance (Hsueh et al., 2024). The framework is designed to preserve long-term trend signals, suppress short-term high-frequency noise, and improve the stability and robustness of dynamic catchment characteristic extraction.

Extracting dynamic catchment characteristics: *Not all indicators exhibit significant dynamic catchment variability; therefore, filtering irrelevant or redundant variables is essential to retain meaningful catchment dynamics. A threshold-based screening is applied to identify variables exhibiting significant seasonality, retaining only relevant subsystems and forming an initial pool of candidate indicators (see Supporting Information S2.1 for detailed criteria). The Maximal Information Coefficient (MIC) is then employed to quantify linear and nonlinear associations between candidate indicators and streamflow, ensuring hydrological relevance. To mitigate multicollinearity and reduce dimensionality, Principal Component Analysis (PCA) is performed, with the first two principal components retained for clustering. This multi-step filtering and reduction procedure ensures robust extraction of dynamic catchment characteristics and provides a solid basis for sub-period clustering according to hydrological similarity.*

Clustering hydrological processes: *Based on the extracted dynamic catchment characteristics, the time series is clustered into distinct sub-periods using the unsupervised Fuzzy C-Means (FCM) clustering algorithm. The optimal number of clusters is determined through a combination of clustering validity indicators, including the Partition Coefficient (SC), Separation Index (S), and Xie–Beni (XB) index, which collectively assess clustering compactness and separation. In addition, the elbow method is employed as a supplementary diagnostic to identify the inflection point beyond which further increases in cluster number yield diminishing returns. Clustering is performed in the principal component space, enabling effective capture of structural patterns in catchment dynamics. The resulting sub-periods*

provide a robust foundation for integrating dynamic parameters into hydrological models.

In addition, the sub-period clustering is developed exclusively using data from the calibration period. To independently evaluate the generalization capability and robustness of the model under unseen conditions, no model training or parameter adjustment is performed during the evaluation period."

4.1 Defined sub-periods based on catchment dynamics

"To support the implementation of sub-period calibration, periods were identified for all 219 catchments based on variations in dynamic catchment characteristics. The results indicate that dynamic catchment patterns are widespread across the study area, with 219 catchments exhibiting significant variation in at least one hydrometeorological variable (precipitation, temperature, potential evapotranspiration, NDVI, or runoff). Spatially, precipitation seasonality is more significant in the central and western regions; potential evapotranspiration seasonality is widespread, especially in northern areas; runoff seasonality is most evident in the central and northeastern regions; and vegetation seasonality is also common, with only a few high-latitude catchments lacking significant dynamic variation.

A data-driven method was applied to extract relevant information and cluster the time series into distinct periods. The optimal sampling window for each catchment was identified using a Bayesian inference approach, with values ranging from 5 to 150 days (mean = 59.45 days). The MIC was then applied to filter out indicators with weak correlation to runoff. PCA was performed for dimensionality reduction, and the first two components explained, on average, 83.5% of the total variance. Based on the reduced feature space, FCM clustering was used to group time steps, with an average of 4.2 periods identified per catchment.

To illustrate the applicability of the method under diverse hydro-climatic conditions, five representative catchments were selected, covering a range of climate zones and dominant hydrological drivers. These catchments were also used in the subsequent modelling experiments. As shown in Fig. 4a and Fig. 4b, their optimal window lengths ranged from 30 to

150 days, with 12 to 31 indicators retained after screening. In all five cases, the number of identified periods ranged from 3 to 5. When compared with hydrographs, the identified periods aligned well with key hydrological processes, such as rising and recession limbs (Fig. 4c). In catchments with strong dynamic signals (e.g., Case A and Case B), the identified periods showed stable interannual patterns, while in catchments with greater variability (e.g., Case D and Case E), the clusterings still captured major dynamic catchment characteristics. These period clusterings provide a physically interpretable structure that supports the dynamic parameterization and modelling experiments introduced in the following sections. Considering the performance of the seven modelling experiments across both calibration and evaluation periods, Experiments 5 and 7 are considered the recommended experiments for capturing dynamic catchment characteristics. Experiment 5, with multi-parameter dynamic calibration, achieves high predictive accuracy across diverse flow regimes, although it may slightly compromise physical consistency in runoff generation. Experiment 7, incorporating smooth parameter transitions, maintains comparable accuracy while promoting more consistent and physically reasonable runoff strategies across sub-periods, thus offering a balanced approach between model performance and hydrological interpretability. Detailed analysis of the results will be presented in the following sections."

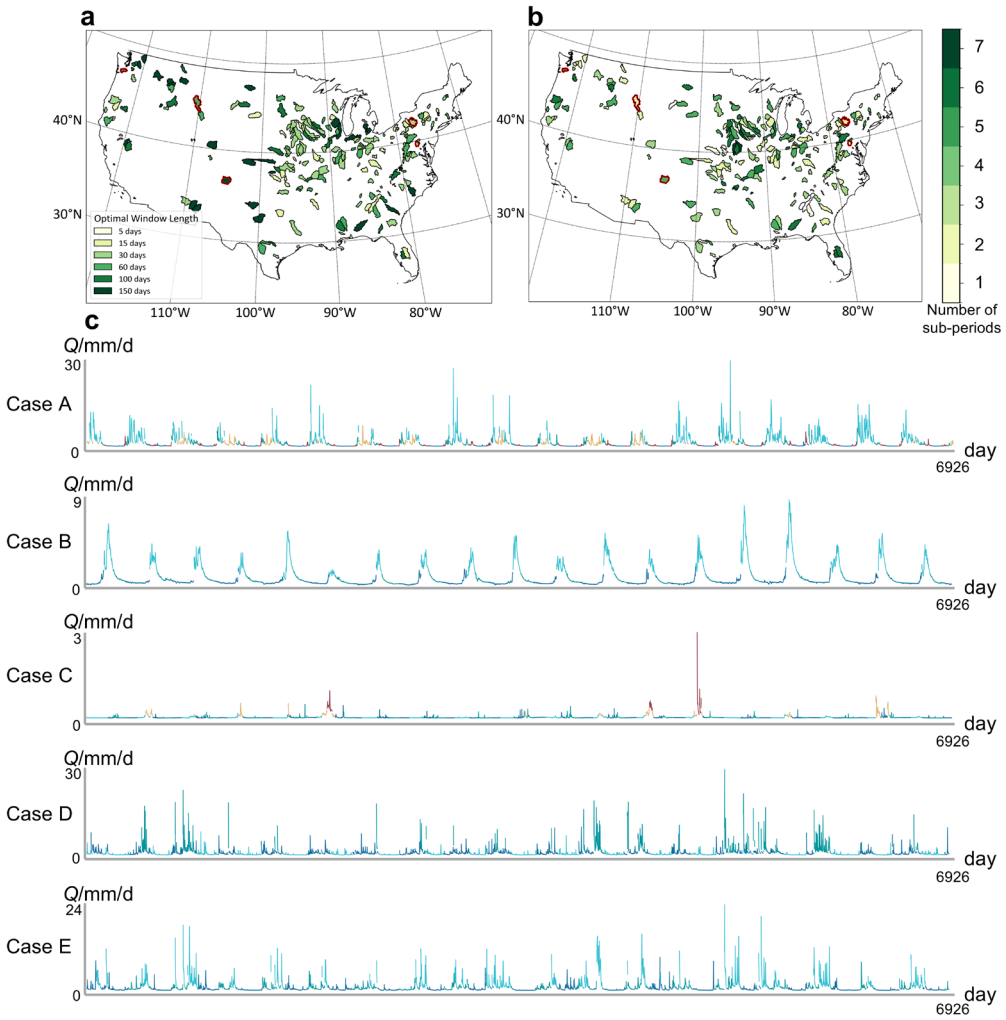

Figure 4. a, *Optimal window lengths of catchment area used in this study for the sub-period clustering.* ***b,*** *Number of subperiods reflecting results from Section 3.2.* ***c,*** *Visualization of clustering results on the hydrograph for the respective study cases.*

Q3: The EDCC approach is only superficially described in the main text, despite its centrality to the framework's novelty. Key implementation and performance results are hidden in the SI, where they are difficult to evaluate.

Response:

We greatly appreciate the reviewer's attention to the description and presentation of the EDCC approach. The concern about the insufficient description of the EDCC method in the main text has been fully addressed in the revised manuscript, where the relevant content has been refined to provide a clearer and more coherent methodological narrative. To enhance the transparency

of the sub-period clustering logic, the key steps—including the index system, feature extraction, and clustering procedures—have been systematically summarized and explained in Section 3.2 of the Methods. The detailed technical workflow was retained in the Supporting Information (SI) to ensure methodological completeness. This structure enables a direct understanding of the operational logic of sub-period clustering within the revised manuscript, without requiring frequent reference to the SI. Additional technical details are presented in Section 4.1 of Results, including the number of sub-periods, the length of the temporal window across the study area, and the specific clustering outcomes for the case-study catchments. Further clustering results are provided in the SI for reference.

As mentioned in our response to Q1 regarding the structure and details of the Methods section, this study focuses on comparing different parameter calibration experiments and examining whether sub-period calibration based on dynamic catchment characteristics can effectively improve the coupling between the model structure and the catchment's hydrological processes. Therefore, the revised manuscript retains the essential description of sub-period clustering in the main text while avoiding excessive detail, so as to emphasize the core scientific questions and innovations. The specific revisions, implemented in Sections 3.2 and 4.1, are detailed in the response to Q2 (see "Revised manuscript text").

Q4: Details on the hydrological model (HYMOD) are insufficient. Given the extensive use of model fluxes and state variables in both analysis and figures, a clearer introduction to the model and its components is essential.

Response:

We greatly appreciate your comments. A complete introduction to the HYMOD model is essential for understanding the experimental design and interpreting the results, especially regarding the dynamic behavior of internal fluxes and state variables. In the revised manuscript, the Methods section (3.1 Hydrological model) has been supplemented to include a schematic diagram illustrating the HYMOD model structure and principles, a detailed explanation of its

operational workflow, and a table defining all model parameters, state variables, and fluxes. These additions enhance the clarity and comprehensibility of the manuscript.

Revised manuscript text:

3.1 Hydrological model

"The two investigated strategies involve only parameter configuration and calibration procedures, without necessitating structural changes to the hydrological model. Such strategies are compatible with lumped, semi-distributed, and fully distributed models, encompassing both conceptual and physically based types. To evaluate and compare the applicability of different calibration strategies under dynamic catchment conditions, the simple conceptual hydrological model, HYMOD (Hydrological MODel) (Moore, 1985), is employed for verification. The HYMOD model is a conceptual rainfall-runoff model with a simple structure (five parameters), low input requirements, and empirical physical interpretations. It has been successfully used in streamflow prediction across America and many other regions (Vrugt et al., 2003; Wagener et al., 2001). In addition, to enhance model performance in snowy areas, the Degree-day model is applied to account for the snow melt (Supporting Information S1.6) (Wang et al., 2022a).

The structure of the HYMOD model is shown in Fig. 2. Precipitation (P) and potential evapotranspiration (PET) drive a probability-distributed soil-moisture store characterized by a maximum capacity (H_{uz}) and a shape parameter (B). Actual evaporation (AE) is limited by potential evapotranspiration and soil water availability. The remaining rainfall infiltrates to recharge the soil-moisture storage (XH_{uz}). When XH_{uz} reaches its maximum capacity (H_{uz}), the surplus is released as excess rainfall (saturation-excess runoff, OV). This excess rainfall is then partitioned by α into inputs to the quick-flow and the slow-flow pathways. Quick flow is routed through a cascade of three linear reservoirs (states X_{q1}–X_{q3}) governed by K_q, producing outflow Q_q, while the slow flow is routed through a single linear reservoir governed by K_s, producing outflow Q_s. The simulated discharge (Q_{sim}) is computed as the sum of Q_q and Q_s (Wang et al., 2022a). Detailed information on the HYMOD model parameters, state variables, and fluxes is

provided in Table 2."

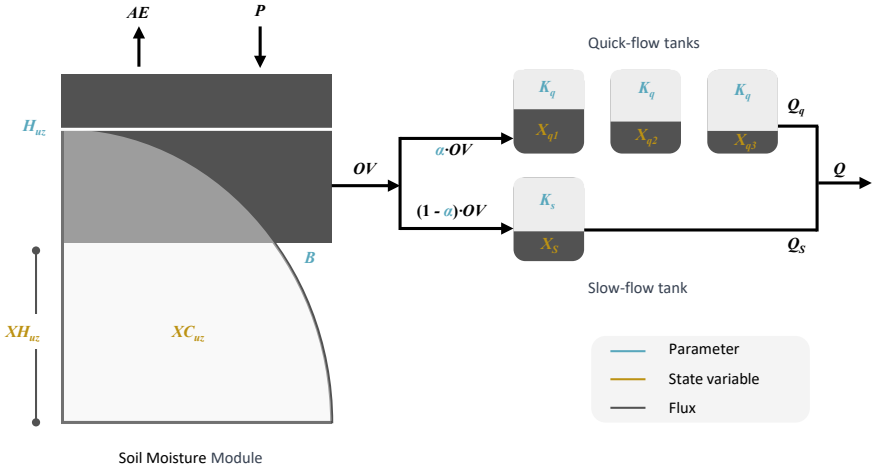

Figure 2. *Schematic diagram of the HYMOD structure and principles (Vrugt et al., 2003; Wagener et al., 2001).*

Table 2. *HYMOD model parameters, state variables, and fluxes (Vrugt et al., 2003; Wagener et al., 2001).*

Label	*Property*	*Range*	*Description*
H_{uz}	*Parameter*	*10–1500 mm*	*Maximum height of the soil moisture accounting tank*
B	*Parameter*	*0–1.99*	*Scaled distribution function shape*
α	*Parameter*	*0–0.99*	*Quick or slow split*
K_q	*Parameter*	*0.5–0.99*	*Quick-flow routing tanks' rate*
K_s	*Parameter*	*0–0.5*	*Slow-flow routing tank's rate*
XH_{uz}	*State variable*	*mm*	*Upper-zone soil moisture tank state height*
XC_{uz}	*State variable*	*mm*	*Upper-zone soil moisture tank state contents*
X_q	*State variable*	*mm*	*Quick-flow tank state contents*
X_s	*State variable*	*mm*	*Slow-flow tank state contents*
AE	*Flux*	*mm* d^{-1}	*Actual evapotranspiration flux*
OV	*Flux*	*mm* d^{-1}	*Excess rainfall flux*
Q_q	*Flux*	*mm* d^{-1}	*Quick-flow flux*
Q_s	*Flux*	*mm* d^{-1}	*Slow-flow flux*
Q_{sim}	*Flux*	*mm* d^{-1}	*Total simulated streamflow flux*

Q5: The calibration experiments, though commendably framed with clear objectives, lack clarity in terms of execution and consistency: It is essential to clarify what is being optimized, when, and how parameters are treated during sub-periods. For example, lines 146-155 reference SCE-UA, while Experiment 2 introduces NSGA-II without clearly stating if other

experiments revert to SCE-UA. The description of Experiment 4, in particular, remains opaque even after repeated readings. Figure 2 is helpful but insufficient.

Response:

We thank you for your concern about the clarity of the experimental design and description. In the revised manuscript, descriptions of Experiments 1–7 have been optimized and standardized. Each experiment is now described using a fixed format, including the research objective, optimization algorithm, objective function settings, and how parameters are treated during sub-periods, allowing readers to quickly grasp the differences between the experiments.

To avoid ambiguity and ensure comparability, Experiment 2, which originally used NSGA-II, has been redesigned. The SCE-UA algorithm is now used as the global optimization tool in all experiments. A weighted objective function, wNSE + (1–w)LNSE (where the weight w ranges from 0 to 1 with a step length of 0.05), is designed to achieve an equivalent multi-objective optimization. This adjustment not only ensures methodological consistency but also enables direct comparison of results among experiments.

Details for Experiment 4 were expanded. Time-varying parameters are introduced, allowing only the most sensitive parameter to vary across sub-periods while other parameters remain constant. State variables and fluxes are passed between sub-periods through an inheritance mechanism.

In addition, Figure 2 has been redesigned and optimized to align precisely with the experiment descriptions. It visually illustrates differences among the seven experiments in terms of algorithms, parameter handling, and sub-period clustering methods.

Revised manuscript text:

3.3 Calibration experiments

"To systematically evaluate how calibration strategies capture catchment dynamics and improve the simulation of diverse flow regimes, a diagnostic framework comprising seven calibration strategies is developed. These experiments sequentially address key challenges in representing time-varying hydrological behaviour, with a focus on objective function design and time-varying parameterization (Fig. 3).

Experiments 1–3 use time-invariant parameters and focus on the design and weighting of objective functions. Experiment 1 establishes a baseline with standard global calibration. Experiment 2 applies a multi-objective approach to explore trade-offs between high and low flows. Experiment 3 designs a composite objective function to enhance simulation performance across a range of flow conditions. Experiments 4–7 incorporate time-varying parameters to better represent temporal catchment variability and examine related calibration challenges. Experiment 4 allows only the most sensitive parameter to vary, assessing partial dynamization and parameter compensation. Experiment 5 makes all parameters dynamic, raising issues of parameter dimensionality. Experiment 6 investigates the effects of abrupt parameter shifts on model continuity. Experiment 7 introduces smooth parameter transitions to reduce instability while preserving responsiveness to catchment dynamics.

Throughout the experiments, the Shuffled Complex Evolution algorithm (SCE-UA) is employed to search for the globally optimal parameter set (Duan et al., 1993). The HYMOD model is configured for catchments over 19 years from 1982 to 2000, with 1982 as the warm-up year, 1983–1995 for calibration, and 1996–2000 for evaluation. All other model parameters are held at their default values. Unless specified otherwise, model calibration is guided by the following objective function:

$$OF = 0.5*NSE+0.5*LNSE \quad (1)$$

Experiment 1 uses time-invariant parameters calibrated over the entire period without sub-period clustering. It serves as a baseline for assessing standard global calibration.

Experiment 2 approximates a multi-objective calibration by combining NSE and LNSE

into a weighted objective: w × NSE + (1–w) × LNSE. The weight w varies from 0 to 1 (step = 0.05), forming a series of single-objective optimizations using SCE-UA with time-invariant parameters. This setup explores trade-offs between flow regimes without changing the optimization algorithm.

Experiment 3 adopts a composite objective function to improve simulation across flow regimes. It integrates RMSE with flow duration curve (FDC)-based metrics (RMSE_Q95, Q70, Qmid, Q20, Q5, as listed in Table 3), representing different flow phases. Weights are derived from Experiment 1 using AHP, PP, and CRITIC methods (refer to Supporting Information S1.7).

Experiment 4 introduces time-varying parameters by allowing only the most sensitive parameter to vary across sub-periods, while all others remain fixed. State variables and fluxes are passed between sub-periods through an inheritance approach.

Experiment 5 extends the dynamic calibration to all parameters, with distinct values assigned to each sub-period. As a result, the number of parameters increases in proportion to the number of sub-periods, generating a high-dimensional calibration space. State and flux continuity between sub-periods is maintained through the same inheritance mechanism used in Experiment 4.

Experiment 6 investigates the impact of abrupt parameter transitions across sub-periods. Parameters are optimized independently for each sub-period. During model runs, parameter sets switch discretely between sub-periods, while state variables and fluxes are inherited to maintain continuity.

Experiment 7 adopts the same calibration structure as Experiment 6 but incorporates smooth parameter transitions during evaluation. The parallel calibration strategy is designed to preserve continuity in parameter evolution while maintaining water balance within each sub-period."

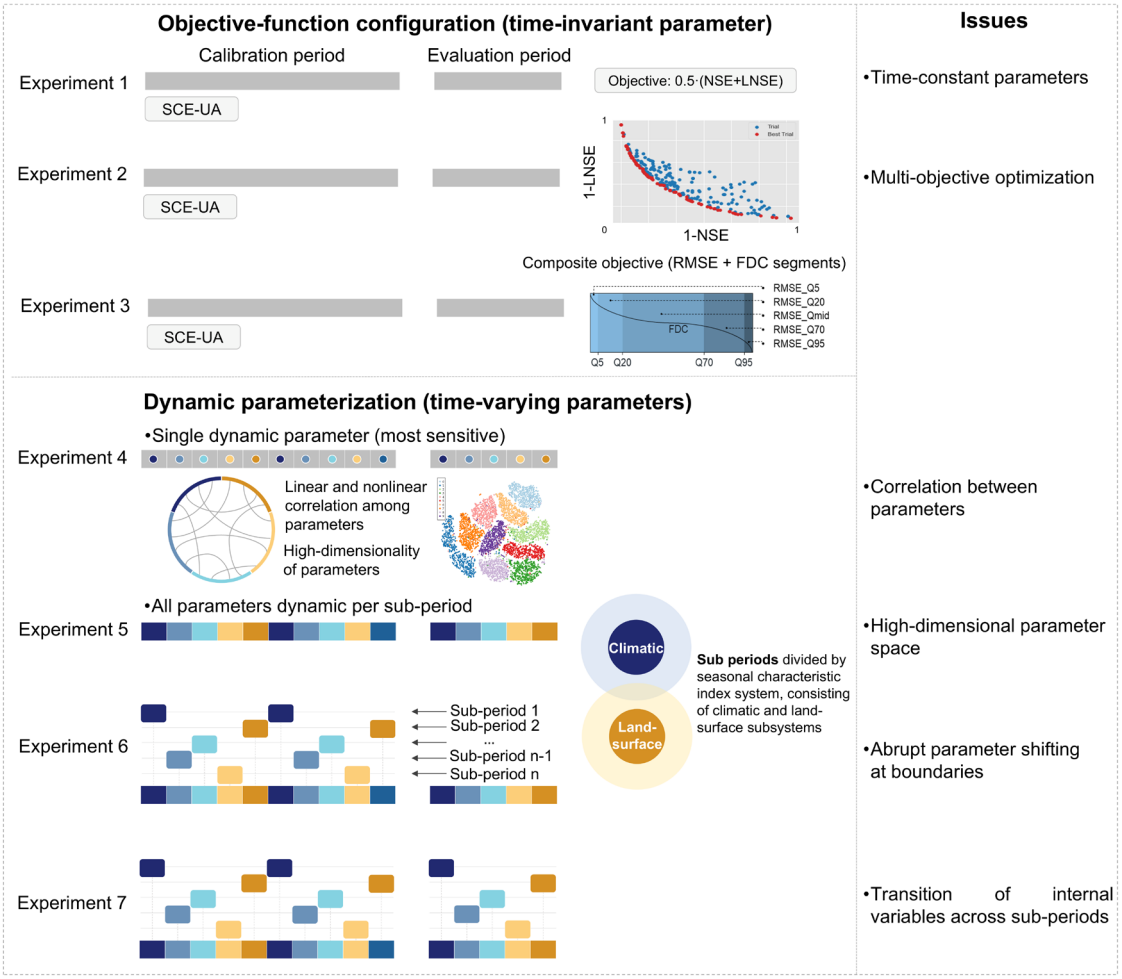

Figure 3. *Schematic illustration of the seven calibration experiments. The colour bands represent state variables and fluxes, which are continuously transferred within the same period. In Experiments 1, 2, and 3, the parameters are time-invariant, but the experiments differ in their objective function configurations. Conversely, experiments 4, 5, and 6 maintain a consistent objective function, but vary the parameters across different experiments. In Experiment 4, the dynamic of only the specific parameter is operated, and the other fixed parameters are optimized simultaneously. In Experiment 5, the parameter set is dynamized. The parameter sets in different sub-periods are optimized simultaneously. In Experiment 6, the data from the individual sub-periods are used for minimizing the objective function, while the model is run for the whole period. In the evaluation period, the parameter set between two consecutive sub-periods is updated accordingly. In Experiment 7, the calibration is the same as in Experiment 6. In the evaluation period, the simulated flow data from each separate sub-period are combined and compared with the observed flow.*

Q6: In general, the reader should not need to reference the SI repeatedly to understand the core methodology.

Response:

Thank you for your constructive suggestion. As mentioned before, the core aim of this study is

to compare and evaluate the suitability of different parameter calibration experiments in catchments with significant dynamic characteristics. Sub-period clustering serves primarily as a central preprocessing step for sub-period calibration, providing the foundation for a systematic comparison and evaluation of these calibration strategies. In the revised manuscript, a summary of the key implementation steps for sub-period clustering—including index system construction, MIC screening, PCA dimensionality reduction, and cluster-based defining—has been integrated into Section 3.2 of the Methods, enabling readers to understand the methodology and results directly from the main text. Meanwhile, without affecting comprehension of the core methodology, a more detailed technical workflow is retained in the SI for readers seeking additional details. Through this adjustment, the revised manuscript can independently present the core methodology and main conclusions in the main text, thereby enhancing the readability and logical coherence. The specific revisions, implemented in Section 3.2, are detailed in the response to Q2 (see "3.2 Clustering hydrological processes" in the "Revised manuscript text").

Q7: With respect to the "Evaluation" section, more detail should be provided on each performance metric, including references and benchmark values. Also note that Table 1 and Table S2 are almost identical and redundant.

Response:

We greatly appreciate your suggestions for improving the "Evaluation" section. In the revised manuscript, a new column has been added to Table 3 that systematically presents the mathematical formulas for all evaluation metrics, along with brief explanations of their meaning and applicability. Also, Table S2 has been removed to avoid redundancy and confusion, resulting in a concise and coherent structure between the main text and the SI. These revisions can significantly improve the completeness and readability of the "Evaluation" section, enabling a full understanding of the paper's performance evaluation system without consulting the SI.

Revised manuscript text:

Table 3. *Description of performance metrics.*

Metric	*Formula*	*Description*
NSE	$NSE = 1 - \frac{\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^{n}(Q_{obs,i} - \overline{Q}_{obs})^2}$	*Sensitive to peaks and discharge dynamics*
LNSE	$LNSE = 1 - \frac{\sum_{i=1}^{n}(log\ Q_{obs,i} - log\ Q_{sim,i})^2}{\sum_{i=1}^{n}(log\ Q_{obs,i} - \overline{log\ Q}_{obs})^2}$	*Emphasizing low flows with the log of discharge*
RMSE_Q5	$RMSE_{Q5} = \sqrt{\frac{1}{n_{Q5}} \sum_{i \in I_{Q>Q5}} (Q_{obs,i} - Q_{sim,i})^2}$	*RMSE in FDC Q5 very-high-segment volume*
RMSE_Q20	$RMSE_{Q20} = \sqrt{\frac{1}{n_{Q20}} \sum_{i \in I_{Q5<Q<Q20}} (Q_{obs,i} - Q_{sim,i})^2}$	*RMSE in FDC between Q5 and Q20 high-segment volume*
RMSE_Qmid	$RMSE_{Qmid} = \sqrt{\frac{1}{n_{Qmid}} \sum_{i \in I_{Q20<Q<Q70}} (Q_{obs,i} - Q_{sim,i})^2}$	*RMSE in FDC between Q20 and Q70 mid-segment volume*
RMSE_Q70	$RMSE_{Q70} = \sqrt{\frac{1}{n_{Q70}} \sum_{i \in I_{Q70<Q<Q95}} (Q_{obs,i} - Q_{sim,i})^2}$	*RMSE in FDC between Q70 and Q95 low-segment volume*
RMSE_Q95	$RMSE_{Q95} = \sqrt{\frac{1}{n_{Q95}} \sum_{i \in I_{Q<Q95}} (Q_{obs,i} - Q_{sim,i})^2}$	*RMSE in FDC Q95 very-low-segment volume*
RMSE	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i})^2}$	*RMSE sensitive to flood peaks*
MSE	$MSE = \frac{1}{n}\sum_{i=1}^{n}(Q_{obs,i} - Q_{sim,i})^2$	*MSE is sensitive to high flow*
MSEL	$MSEL = \frac{1}{n}\sum_{i=1}^{n}(log\ Q_{obs,i} - log\ Q_{sim,i})^2$	*MSEL is sensitive to low flow*
MAE	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert Q_{obs,i} - Q_{sim,i}\rvert$	*MAE is measuring the overall discharge*

Q8: The discussion of flux mapping is vague; although described as a ternary plot method, such plots do not appear in the main text.

Response:

We thank you for your valuable feedback on the flux mapping section. The role of flux mapping in this paper is to assist in diagnosing the model's internal behavior and uncertainty, particularly

in analyzing the equifinality issues that may arise from dynamic parameter calibration.

In the revised manuscript, the relevant sections have been improved. In Section 4.3 of Results, we have added ternary plot results to visually demonstrate the differences in flow composition under different experiments. In the Discussion section, we further clearly explain these plots, illustrating how the distribution of point clouds reflects uncertainty, parameter sensitivity, and the structural limitations of the model. Also, consistent use of the term "flux mapping" throughout the manuscript ensures alignment between figures and text, avoiding potential misunderstandings arising from inconsistent terminology.

Revised manuscript text:

Section 4.3 State variables and fluxes (lines 395-430)

"The comparative analysis of Experiments 1, 5, and 7 further illustrates the performance improvements introduced by Experiments 5 and 7. Fig. 9 illustrates the flux mapping of various sub-periods in the study case A, comparing Experiments 1, 5, and 7. Flux-mapping figures for the other study cases are detailed in the Supporting Information (Fig. S13-S16). Each scatter point in the figures represents a parameter set generated during the SCE-UA algorithm optimization process. The colour and relative position of each scatter point on the axes illustrate the variation in runoff components for sub-periods under specific parameter sets, as well as the corresponding objective function value. To facilitate comparison, the results of Experiment 1 are also presented by the same sub-periods as Experiments 5 and 7. Notably, the differences in optimization performance between Experiments 1 and 7 reveal key insights into model behaviour. Across all study cases, both Experiments 1 and 7 show the poorest results in sub-periods 1 and 2, with the largest (worst) objective function values. In the remaining three sub-periods, the objective function values are significantly better. Compared to Experiment 1, Experiment 7 consistently identified more optimal parameter sets with smaller objective function values within the same period. For example, in Fig. 9b (Experiment 7), most of the dark blue scatter points for sub-period 5 cluster around a vertical axis value of approximately

0.25, whereas in Experiment 1, scatter points for the same sub-period are more widely distributed near 0.5. Shifting the focus to flux components, the spatial distribution of scatter points in the flux maps reveals varied runoff components and internal model behaviour for each sub-period. In Experiment 7, clusters of scatter points of the same colour appear more compact, while in the traditional scheme, they are more dispersed along both vertical and horizontal axes. This pattern indicates that, despite similar objective function values, Experiment 7 possesses a narrower range of optimal equifinality parameters during the parameter evolution process, reducing the model's internal fluxes equifinality and uncertainty. Furthermore, Fig. 9b shows that in Experiment 7, the colour bars along the vertical axis are shorter and more evenly distributed, demonstrating that from sub-periods 1 to 5, the SCE-UA algorithm more rapidly converges to near-optimal solutions, showing a narrower range of variability in the optimization process.

Further comparison with Experiment 5 (Fig. 9c) shows that parameter sets within each sub-period were tightly clustered in the vertical direction, indicating consistently high performance within individual sub-periods. However, these clusters were widely dispersed along the horizontal axis. For instance, the cluster for Sub-period 2 (dark red) is concentrated at higher Q_s values (approximately 0.9), whereas the cluster for Sub-period 4 (orange) is concentrated at much lower values (approximately 0.5). Such horizontal separation suggests that different runoff generation mechanisms (fluxes) are adopted across sub-periods to achieve high performance, which may compromise the physical consistency of the overall simulated discharge (Q_{sim}). This inconsistency is particularly evident in case E, where runoff generation mechanisms across sub-periods appeared nearly independent, while the separation is less significant in case B. In contrast, scatter clusters in Experiment 7 (Fig. 9b) are more tightly aligned along the horizontal axis, indicating the adoption of more consistent and physically reasonable runoff strategies across sub-periods. Nevertheless, Experiment 7 poses a potential risk of discontinuities in internal state variables at sub-period boundaries, a phenomenon that was particularly evident in case D (Fig. S15). In summary, the improvements observed in Experiments 5 and 7 underscore both the importance of refining dynamic parameters and the

model's ability to simulate complex hydrological processes across sub-periods. However, Experiment 5 may compromise physical consistency in runoff generation processes, while Experiment 7 faces the challenge of ensuring smooth transitions of state variables across boundaries."

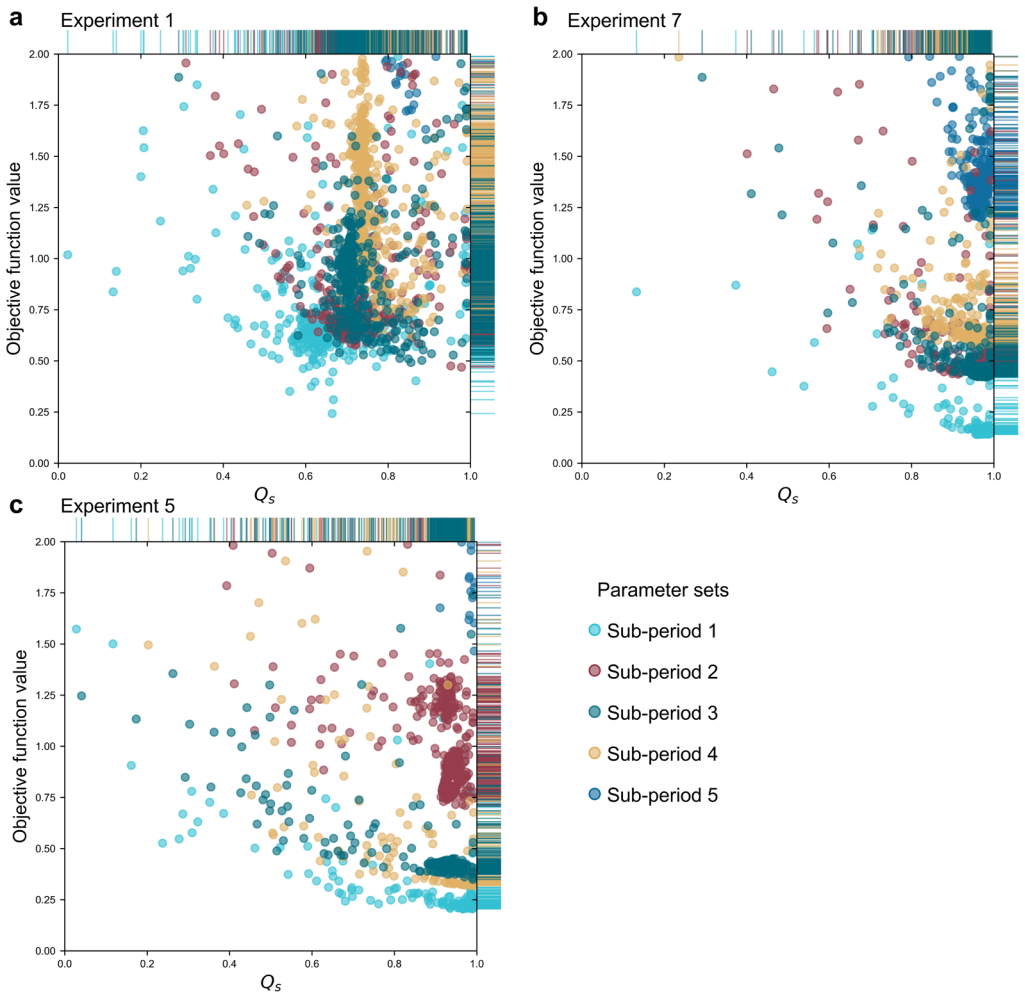

***Figure 9. a**, Flux mapping for case A in the conventional scheme, **b**, Experiment 7, and **c**, Experiment 5, where the horizontal axis represents the proportion of Qs in the runoff.*

Q9: Although four case studies are introduced, results are shown primarily for Case A. In most instances, other cases are summarised with a single sentence asserting similarity. If the intention is to generalise the proposed framework, this is inadequate. Either more contrast

between cases should be shown or a more compelling rationale for their selection should be provided.

Response:

Thank you for your detailed review of the case study setup and presentation. In the revised manuscript, five representative catchments are selected based on differences in climatic zones, topographic conditions, and hydrological characteristics. In Section 2, "Study area," now explicitly explains the rationale for this selection, highlighting coverage from humid to semi-arid climates and from plains to mountainous regions, ensuring that the case studies are representative and provide a solid basis for evaluating the framework's applicability. Basic information for each catchment, including geographical location, catchment area, climate type, and main hydrological features, has been added to enhance transparency of the selection criteria. This addition also addresses the reviewer's concern regarding insufficient justification for the case study selection. In Section 4.2 of Results, the analysis of the core case study has been retained, and key results for all catchments have been added. Box plots of performance metrics are presented to visually demonstrate the framework's performance across different scenarios, ensuring that contrasts among cases are explicitly shown rather than described only in summary.

Revised manuscript text:

Section 2 Study area

"The Model Parameter Estimation Experiment (MOPEX) is an international project aimed at developing enhanced techniques for a priori estimation of parameters in hydrologic models and land surface parameterization schemes of weather and climate models (Duan et al., 2006). A comprehensive MOPEX database has been developed that contains historical hydrometeorological data and land-surface characteristics data for numerous hydrological catchments in the United States (US) and other countries. This study utilizes the dataset from 219 catchments spatially distributed across the contiguous US (Fig. 1a). Rigorous screening

criteria were applied to ensure the acquisition of high-quality data. The screening process involved three key considerations: (1) no missing or non-physical data throughout the study period; (2) minimal interference from anthropogenic influences in both temporal and spatial dimensions; and (3) a large spatial distribution scale of the selected catchments, including diverse meteorological and underlying surface conditions. The dataset for selected catchments includes the hydrometeorological forcing data, land-surface data, and streamflow data, covering the period from 1983 to 2000. Hydrometeorological data includes daily precipitation data (P), temperature data (T), and streamflow (Q) provided by the MOPEX dataset, as well as potential evaporation data (PE) calculated by the Hamon model (Mccabe et al., 2015). The Normalized Difference Vegetation Index (NDVI) was used as one of the land-surface indicators to represent the vegetation coverage of the catchments, which had a spatial resolution of 8 km and a temporal resolution of half-monthly intervals (Tucker et al., 2010). Based on these criteria, a total of 219 catchments were selected (Fig. 1a), spanning a wide range of hydrological and meteorological characteristics, making them ideal for testing various model structures under diverse conditions (Duan et al., 2006).

In addition to the large-sample analysis of the MOPEX dataset, five representative catchments, Case A (12027500), Case B (6192500), Case C (7211500), Case D (1643000), Case E (1531000), are analyzed in more detail as case studies. These catchments encompass a variety of Köppen climate classifications and different dominant dynamic catchment characteristics, facilitating comparison of calibration strategies and evaluation of their robustness under diverse hydroclimatic conditions. Their locations and characteristics are listed in Table 1 and will be analyzed in depth in the subsequent sections."

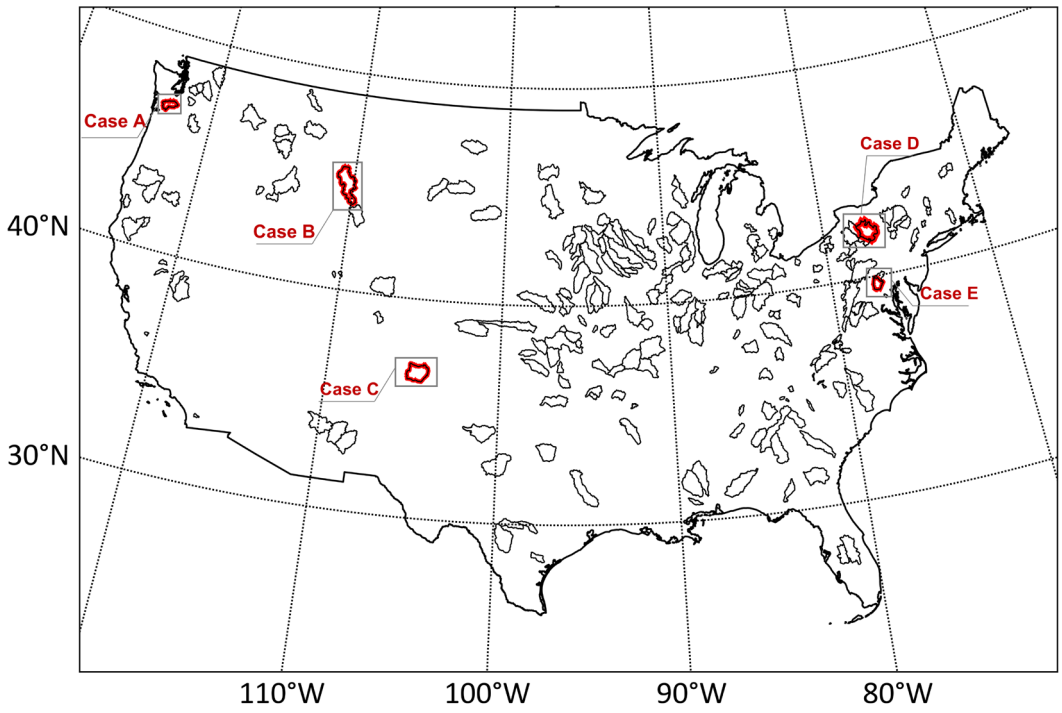

Figure 1. *Location map of the catchment area used in this study, where cases A, B, C, D, and E correspond to catchments 12027500, 6192500, 7211500, 1643000, and 1531000 (from west to east) are highlighted with red outlines for reference.*

Table 1. *Summary of catchment characteristics for study cases.*

ID	*12027500*	*6192500*	*7211500*	*1643000*	*1531000*
Location	*122.99°W*	*110.40°W*	*104.76°W*	*77.25°W*	*77.24°W*
Area (km²)	*895*	*3551*	*2850*	*817*	*2056*
Climate	*Csb*	*Dfc*	*Bsk*	*Cfa*	*Dfb*
Mean P (mm)	*1548.78*	*735.71*	*491.70*	*1068.49*	*870.53*
Mean PE (mm)	*596.53*	*731.59*	*1279.88*	*897.63*	*711.06*
Mean Q (mm)	*1110.19*	*369.79*	*10.08*	*430.15*	*366.76*
Mean elevation	*253.06*	*2441.28*	*2262.91*	*191.80*	*492.25*
Mean slope (°)	*12.16*	*15.26*	*9.44*	*4.99*	*8.25*
Runoff ratio	*0.72*	*0.50*	*0.02*	*0.40*	*0.42*
Aridity index	*2.60*	*1.01*	*0.38*	*1.19*	*1.23*
Forest cover (%)	*71.96*	*36.95*	*16.76*	*31.31*	*57.36*
Land use	*Evergreen Forest,*	*Evergreen Forest,*	*Evergreen Forest, Grassland/Herbace*	*Deciduous Forest,*	*Deciduous Forest,*

4.2 Model performance

"To compare seven experiments in dynamic catchments and to identify potential limitations in model calibration, the evaluation is conducted across 219 catchments characterized by hydrological variability. As shown in Fig. 5, the NSE and LNSE values during

both calibration and evaluation periods reveal differences in the ability of diverse calibration schemes to capture high- and low-flow conditions. The median NSE reached only 0.4–0.5 in Experiments 1 and 2, and although the LNSE approached 0.7, negative values are frequently observed. It is suggested that global optimization or simple weighted objective functions often lead to an averaging of catchment responses, thereby limiting accuracy for both high- and low-flow conditions. Experiment 3 employed an objective function defined as: $OF = 0.27 \cdot RMSE\_Q5 + 0.16 \cdot RMSE\_Q20 + 0.08 \cdot RMSE\_Qmid + 0.24 \cdot RMSE\_Q70 + 0.25 \cdot RMSE\_Q95$*, the weighting scheme explicitly accounted for extremely high (Q95), high (Q70), medium (Qmid), low (Q20), and extremely low (Q5) flows. Despite this design, both NSE and LNSE declined relative to Experiment 1. The decrease may be attributed to excessive parameter adjustments aimed at fitting a limited number of extreme events, which reduced the predictive accuracy of the overall streamflow process. When single dynamic parameters are introduced in Experiment 4, median NSE and LNSE increased to approximately 0.55 and 0.8, respectively, with narrower interquartile ranges. These outcomes indicate that dynamic parameters enhanced the ability of the hydrological model to capture temporal variability, although structural errors persisted, as reflected in local outliers. More significant improvements emerged with multiple dynamic parameters. Experiment 5 achieved median NSE and LNSE values of approximately 0.7–0.8 in both calibration and evaluation periods. Although high-dimensional optimization increased computational demand and LNSE variability in some basins, overall performance represented a balanced trade-off between dynamic adaptability and physical consistency. Experiment 6 also performed well during the calibration period; however, its abrupt parameter switching led to a particular decline of LNSE and increased dispersion in the evaluation period. Experiment 7 addressed these shortcomings by applying a gradual parameter-switching strategy during the evaluation period. As shown in Fig. 5, the boxplots are more compact and shifted toward higher values, indicating that stable and consistent performance was achieved across most basins. However, compared with Experiment 5, Experiment 7 displayed a greater number of outliers, particularly in LNSE, where they tended to cluster at lower values, suggesting higher variability in model performance across catchments. The overall accuracy remained comparable to that of Experiment 5. In summary,*

compared with static calibration schemes (Experiments 1–3), single dynamic parameter calibration (Experiment 4) improved simulative accuracy, while multi dynamic parameter calibration produced further gains. Among all experiments, Experiments 5 and 7 demonstrated the most robust and accurate performance.

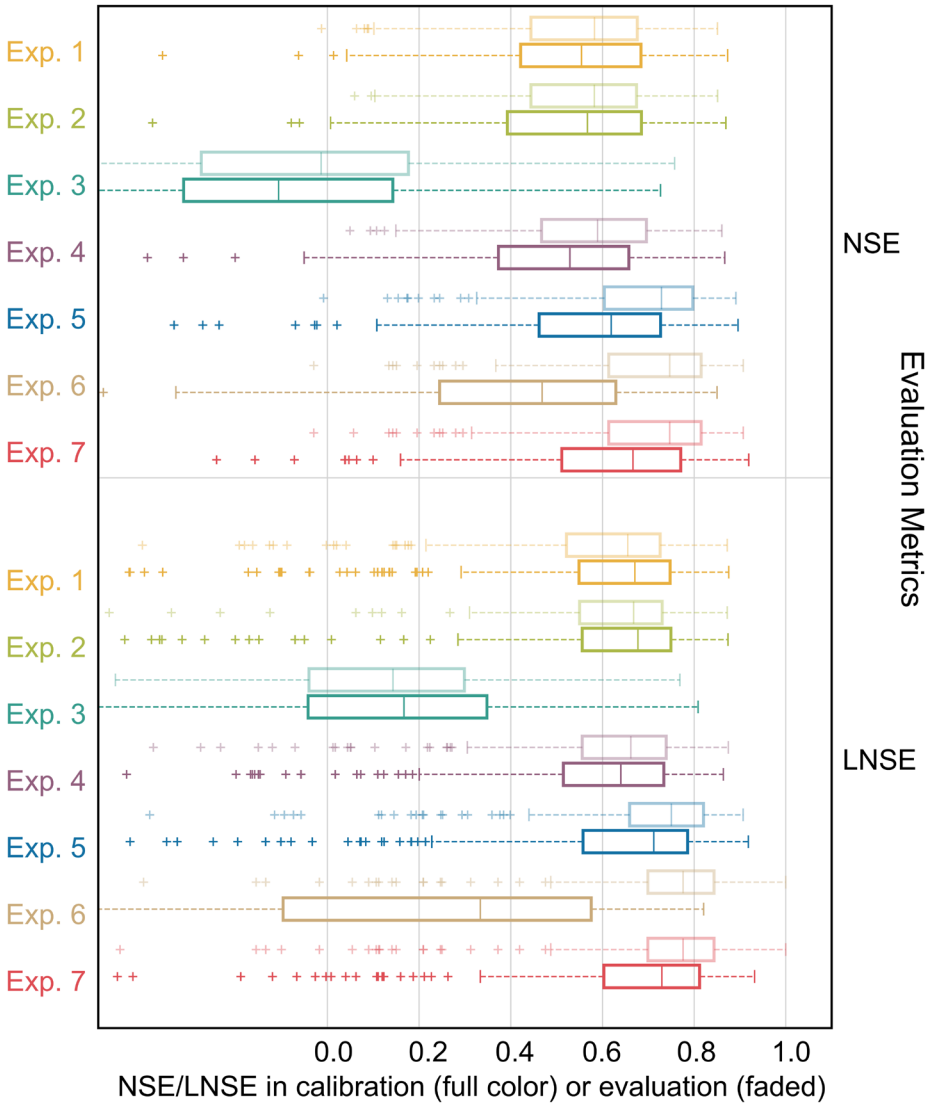

Figure 5. *Performance of seven calibration experiments on the MOPEX dataset across 219 catchments. Boxplot color denotes different experiments. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and are denoted as +.*

To further examine how the different experiments under various hydrological conditions, a detailed assessment of five representative catchments is conducted with diverse dynamic patterns and baseline model performance. Fig. 6 presents the model performance of the seven

experiments in five study cases. In all study cases, Experiment 1 demonstrated low simulation accuracy and limited parameter transferability across different flow phases, particularly under extremely low-flow and high-flow conditions. The results in Experiments 2 and 3 show limitations on both objective functions compared to Experiments 5 and 7. Adjusting the weights between NSE and LNSE improved accuracy for mid-phase flows but failed to account for other flow phases. For instance, in case A, considering NSE, the metric increased from 0.48 (Experiment 1) to 0.62 (Experiment 2) during the calibration period, and from 0.50 to 0.64 during the evaluation period. However, both RMSE_Q5 and RMSE_Q95 increased. Relative to Experiment 1, the RMSE_Q95 exhibited a diminished performance during both the calibration and evaluation periods in Experiment 2. Despite prioritizing high and low flows through a weighted objective function, Experiment 3 underperforms compared to Experiment 1. While the objective function emphasizes these targeted phases, adjusting its weights unexpectedly failed to improve performance in the target flow phase and even worsened the model's performance in other evaluation metrics, indicating that this scheme exhibits instability in its performance across different flow phases. For instance, in case A, the NSE decreased from 0.48 to -0.74 in the calibration period, and from 0.50 to -0.27 in the evaluation period, compared with Experiment 1. In case C, the performance decline was more significant, with NSE values during both the calibration and evaluation periods approaching zero. Experiment 4 exhibited only marginal improvements over Experiment 1 across most metrics. In contrast, Experiments 6 and 7, which employed the same calibration procedures, achieved strong overall performance during the calibration period, particularly in reproducing high flows and flood peaks. However, during the evaluation period, Experiment 6 showed inconsistent performance—while excelling in certain aspects such as high-flow simulation, it experienced significant degradation in others (e.g., NSE, MAE, and RMSE_Q95). The extent of performance decline in Experiment 6 varied among catchments: in case D, RMSE_Q95 increased by only 0.61 mm/d compared to the calibration period, whereas in case C, the deterioration was most severe, with RMSE_Q95 increasing by 17.64 mm/d. The decline can be attributed to extremely dry conditions, where runoff volumes approached zero (less than 0.01 mm), making small deviations translate into large relative errors. Notably, across all study

cases, Experiments 5 and 7 consistently maintained excellent performance during the evaluation period, closely mirroring their calibration results and outperforming other experiments in nearly all metrics. Moreover, analysis of parameter transferability revealed minimal differences between calibration and evaluation periods for Experiments 5 and 7. Hence, Experiments 5 and 7 demonstrate the superior performance across all evaluation metrics, exhibiting improvements in simulations across various flow phases."

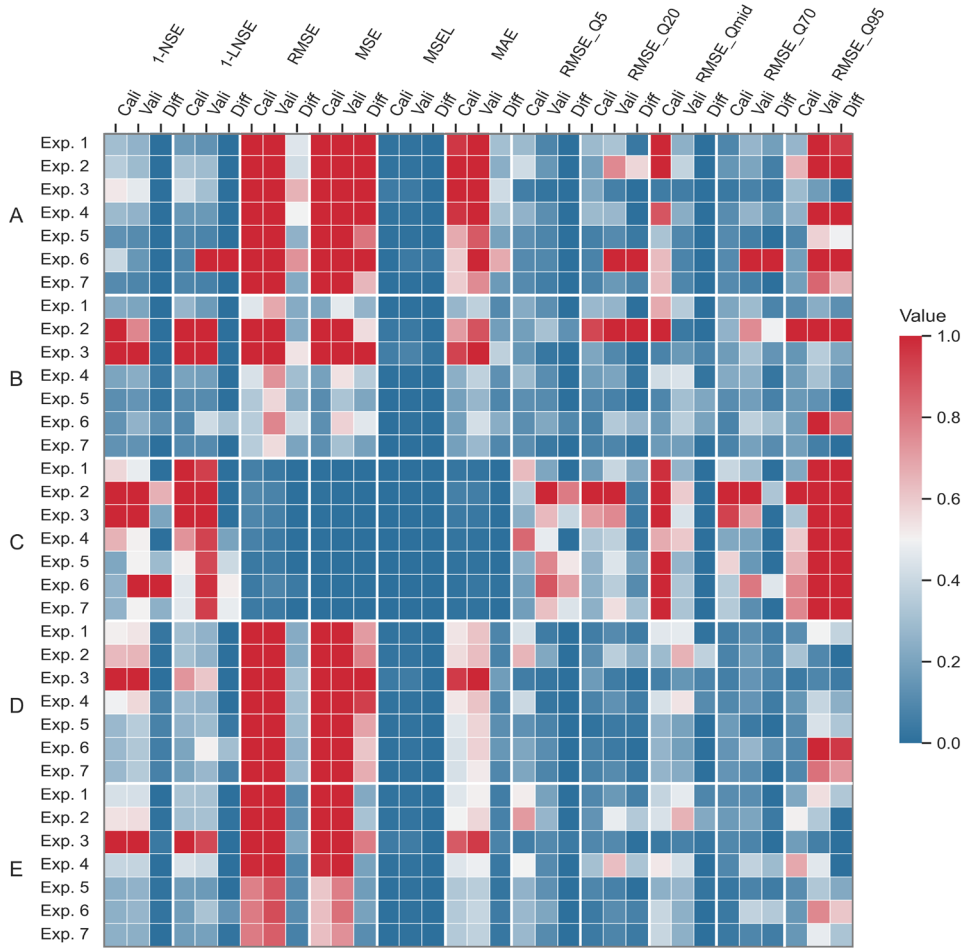

Figure 6. *The model performance of seven experiments in five study cases was assessed using multiple evaluation metrics. Lower values reflect superior performance.*

Q10: A comprehensive synthesis across the full set of catchments (e.g., the 130 dynamic basins in MOPEX) is conspicuously missing. Only a few lines (291–297) address this aggregation, and no discussion is offered on spatial or climatic variability in model performance.

Response:

Thank you very much for your valuable suggestion. In the revised manuscript, statistical results for all catchments have been systematically added in Section 4.1 of the Results. Box plots are used to visually display performance differences among the calibration experiments across all catchments. These results not only cover comparison of metrics such as NSE and LNSE but also reveal the performance distribution and variability of the different experiments. By retaining the in-depth analysis of a typical case while providing statistical support from the full sample, the robustness and strength of the conclusions are enhanced in the revised manuscript. The specific revisions in Section 4.2 (lines 300–330) of the revised manuscript are detailed in the response to Q9 (see "4.2 Model performance" in the "Revised manuscript text").

Q11: EDCC results should be presented within the Results section, not just described or relegated to Supporting Information.

Response:

We greatly appreciate your valuable feedback. In the revised manuscript, the role of sub-period clustering has been further clarified: it serves as the foundational clustering method for all subsequent comparative experiments. Its effectiveness is not demonstrated by individually displaying clustering charts, but is reflected in the superior performance of the seven experiments across all catchments (see Section 4.1 of the Results). These results indicate that the sub-period calibration based on this clustering consistently improves hydrological model performance and process consistency in the majority of catchments, supporting the applicability of the method under diverse catchment conditions.

In addition, in the revisions, we have tried to balance the completeness and logical coherence of the main research line, ensuring that the explanation of the sub-period clustering provides the necessary support, without diverting attention from the main focus on comparing parameter calibration methods. These adjustments enhanced the clarity and coherence of the Results

section. The specific revisions in Section 4.1 of the revised manuscript are detailed in the response to Q2 (see "4.1 Defined sub-periods based on catchment dynamics" in the "Revised manuscript text").

Q12: The analysis of parameter correlation and flux mapping—currently discussed in the Discussion—should be integrated as part of the core results. These are not interpretive reflections, but rather diagnostic outputs central to evaluating the model framework.

Response:

We greatly appreciate your insightful suggestions regarding the structure of the paper and the presentation of results. Parameter correlation analysis and flux mapping are recognized as important diagnostic tools for validating the effectiveness of the framework. In the revised manuscript, flux mapping results have been moved into the Results section (Section 4.3) and presented as core results, including ternary plots and textual explanations, allowing readers to directly observe their contribution to diagnosing the hydrological model's mechanisms. Regarding the parameter correlation analysis, the complete correlation matrices have been provided in Section S4 of the Supporting Information (Figures S17–S20), and their conclusions have been referenced in the Discussion to support interpretation of the framework. Through this arrangement, we highlight the intuitive diagnostic value of flux mapping while maintaining the integrity of the parameter correlation analysis, dividing the results and discussion sections more rationally. We sincerely thank you for your suggestion again; this modification has significantly improved the scientific rigor and readability of the paper.

Revised manuscript text:

4.3 State variables and fluxes

"The state variables and fluxes reflect the internal operation of the hydrological model. The assessment results of state variables and fluxes through seven calibration experiments for

case study A are illustrated in Fig. 7 and Fig. 8 (results of cases B, C, D, and E are shown in S3 of Supporting Information). Experiments 1, 2, and 3 exhibited only minimal differences in both state variables and flux time series, with only the results of Experiments 1 and 3 shown for clarity. A slight improvement is shown in Experiment 4 compared with the time-invariant parameter schemes; however, small mismatches remain during flow recessions and peak timings. This indicates that the dynamic adjustment of a single parameter is insufficient to represent the full range of catchment dynamics. Notably, the state variable X_q *and flux* Q_q *in Experiment 4, the display is abnormally flat compared to Experiment 1, showing a wrong response of the rapid runoff module to input variations.*

In Experiment 6, abrupt parameter switching is applied across sub-periods. The state variable X_q *and flux* Q_q *in Experiment 6, exhibit step changes or even discontinuities at the switching boundaries, with large deviations during low-flow sub-periods. The phenomenon is particularly evident in cases B and D. These results indicate that abrupt switching disrupts water balance continuity, thereby reducing performance in low-flow simulations. Despite these setbacks, Experiments 5 and 7 introduced significant improvements across all study cases. In Experiment 5, multi-parameter dynamic calibration is applied while continuity of state variables and fluxes is maintained. As shown in Fig. 7 and Fig. 8, in case A, the flux variables* Q_q *and* Q_S *transition smoothly across sub-periods without visible discontinuities, the state variables* XH_{uz} *and* XC_{uz} *also connect consistently across sub-periods, indicating that multi-parameter dynamic calibration captures the catchment dynamics of soil moisture and storage processes. However, Experiment 5 shows limitations in maintaining the consistency of simulated discharge (*Q_{sim}*). For example, in case B, the baseline extent of* Q_{sim} *exhibited slight drift, reflected in systematic differences in response intensity to similar rainfall events across adjacent sub-periods. The fluxes and state variables in Experiment 7 exhibit results similar to those in Experiment 5. However, when sub-period simulations are concatenated, slight inconsistencies occasionally emerge at the sub-period boundaries, with flood peaks being slightly overestimated or baseflows being underestimated. Overall, Experiments 1, 2, and 3 exhibit negligible differences in state variables and flux series, although Experiment 3 produces*

a decline in low-flow accuracy. Experiment 4 shows marginal improvements compared with time-invariant parameterization; however, it indicates that a single dynamic parameter is insufficient to capture overall dynamic catchment characteristics. Experiment 6 applies abrupt parameter switching across sub-periods, which disrupts water continuity. In contrast, Experiments 5 and 7 display significant improvements in simulation performance, particularly by mitigating the underestimation of high flows and the overestimation of low flows, as evidenced by the behaviour of internal model variables.

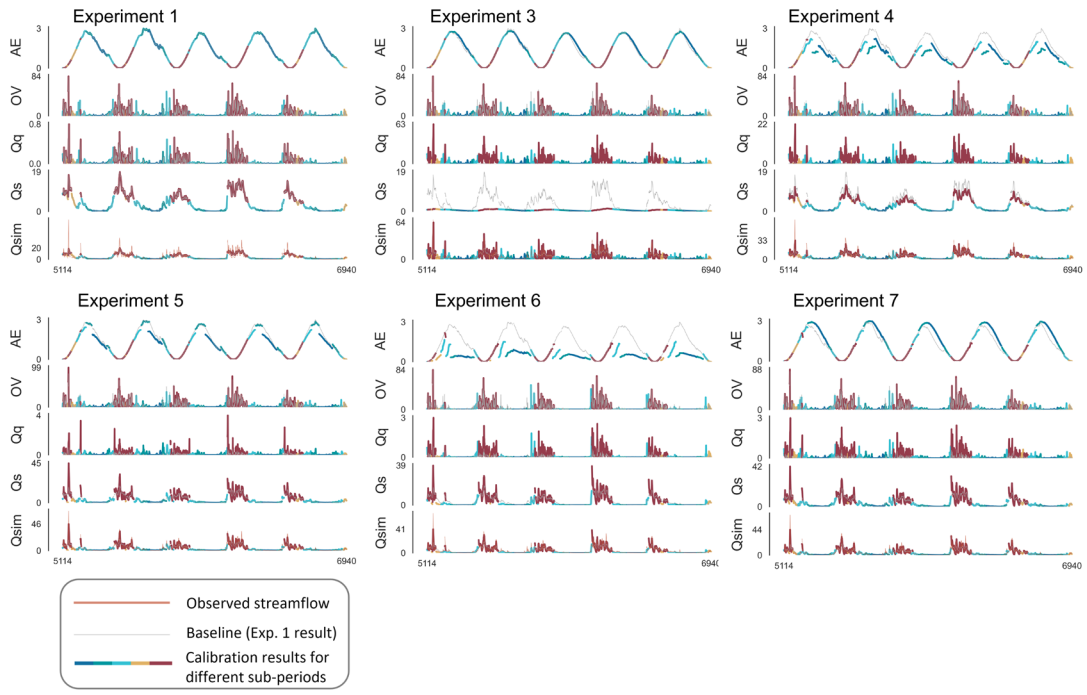

Figure 7. *Fluxes simulation results of experiments during the representative evaluation period for case A. The figure shows the flux simulation results from Experiments 1 to 7, with different colours representing different sub-periods.*

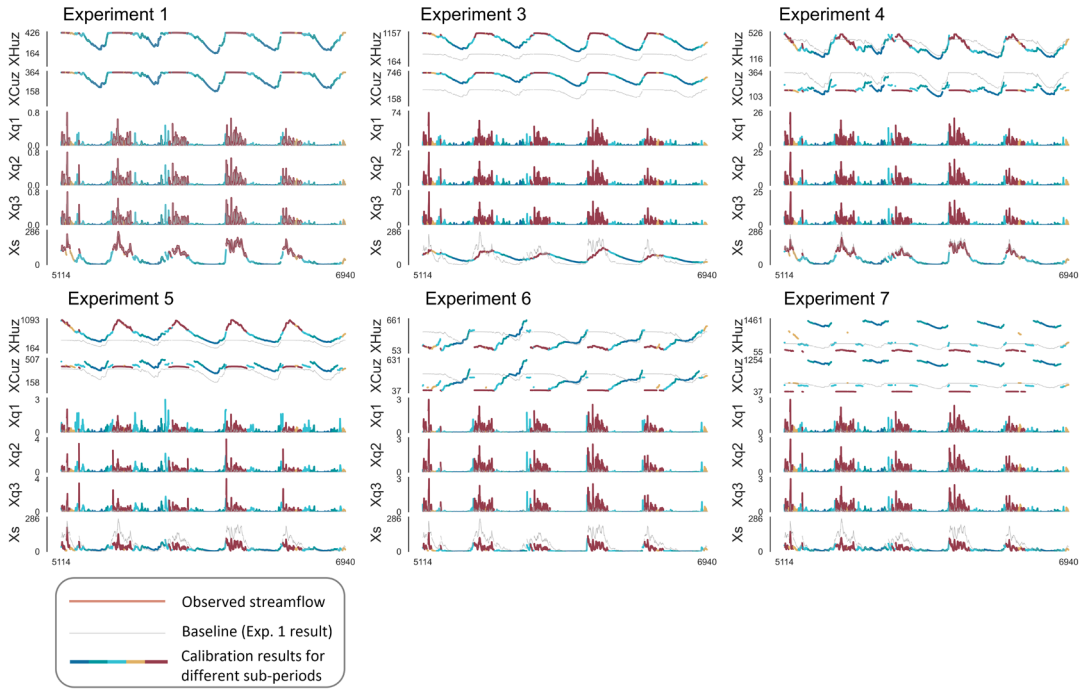

Figure 8. *State variables simulation results of experiments during the representative evaluation period for case A. The figure shows the state variable simulation results from Experiments 1 to 7, with different colours representing different sub-periods.*

The comparative analysis of Experiments 1, 5, and 7 further illustrates the performance improvements introduced by Experiments 5 and 7. Fig. 9 illustrates the flux mapping of various sub-periods in the study case A, comparing Experiments 1, 5, and 7. Flux-mapping figures for the other study cases are detailed in the Supporting Information (Fig. S13-S16). Each scatter point in the figures represents a parameter set generated during the SCE-UA algorithm optimization process. The colour and relative position of each scatter point on the axes illustrate the variation in runoff components for sub-periods under specific parameter sets, as well as the corresponding objective function value. To facilitate comparison, the results of Experiment 1 are also presented by the same sub-periods as Experiments 5 and 7. Notably, the differences in optimization performance between Experiments 1 and 7 reveal key insights into model behaviour. Across all study cases, both Experiments 1 and 7 show the poorest results in sub-periods 1 and 2, with the largest (worst) objective function values. In the remaining three sub-periods, the objective function values are significantly better. Compared to Experiment 1, Experiment 7 consistently identified more optimal parameter sets with smaller objective function values within the same period. For example, in Fig. 9b (Experiment 7), most of the

dark blue scatter points for sub-period 5 cluster around a vertical axis value of approximately 0.25, whereas in Experiment 1, scatter points for the same sub-period are more widely distributed near 0.5. Shifting the focus to flux components, the spatial distribution of scatter points in the flux maps reveals varied runoff components and internal model behaviour for each sub-period. In Experiment 7, clusters of scatter points of the same colour appear more compact, while in the traditional scheme, they are more dispersed along both vertical and horizontal axes. This pattern indicates that, despite similar objective function values, Experiment 7 possesses a narrower range of optimal equifinality parameters during the parameter evolution process, reducing the model's internal fluxes equifinality and uncertainty. Furthermore, Fig. 9b shows that in Experiment 7, the colour bars along the vertical axis are shorter and more evenly distributed, demonstrating that from sub-periods 1 to 5, the SCE-UA algorithm more rapidly converges to near-optimal solutions, showing a narrower range of variability in the optimization process.

Further comparison with Experiment 5 (Fig. 9c) shows that parameter sets within each sub-period were tightly clustered in the vertical direction, indicating consistently high performance within individual sub-periods. However, these clusters were widely dispersed along the horizontal axis. For instance, the cluster for Sub-period 2 (dark red) is concentrated at higher Q_s values (approximately 0.9), whereas the cluster for Sub-period 4 (orange) is concentrated at much lower values (approximately 0.5). Such horizontal separation suggests that different runoff generation mechanisms (fluxes) are adopted across sub-periods to achieve high performance, which may compromise the physical consistency of the overall simulated discharge (Q_{sim}). This inconsistency is particularly evident in case E, where runoff generation mechanisms across sub-periods appeared nearly independent, while the separation is less significant in case B. In contrast, scatter clusters in Experiment 7 (Fig. 9b) are more tightly aligned along the horizontal axis, indicating the adoption of more consistent and physically reasonable runoff strategies across sub-periods. Nevertheless, Experiment 7 poses a potential risk of discontinuities in internal state variables at sub-period boundaries, a phenomenon that was particularly evident in case D (Fig. S15). In summary, the improvements observed in Experiments 5 and 7 underscore both the importance of

refining dynamic parameters and the model's ability to simulate complex hydrological processes across sub-periods. However, Experiment 5 may compromise physical consistency in runoff generation processes, while Experiment 7 faces the challenge of ensuring smooth transitions of state variables across boundaries."

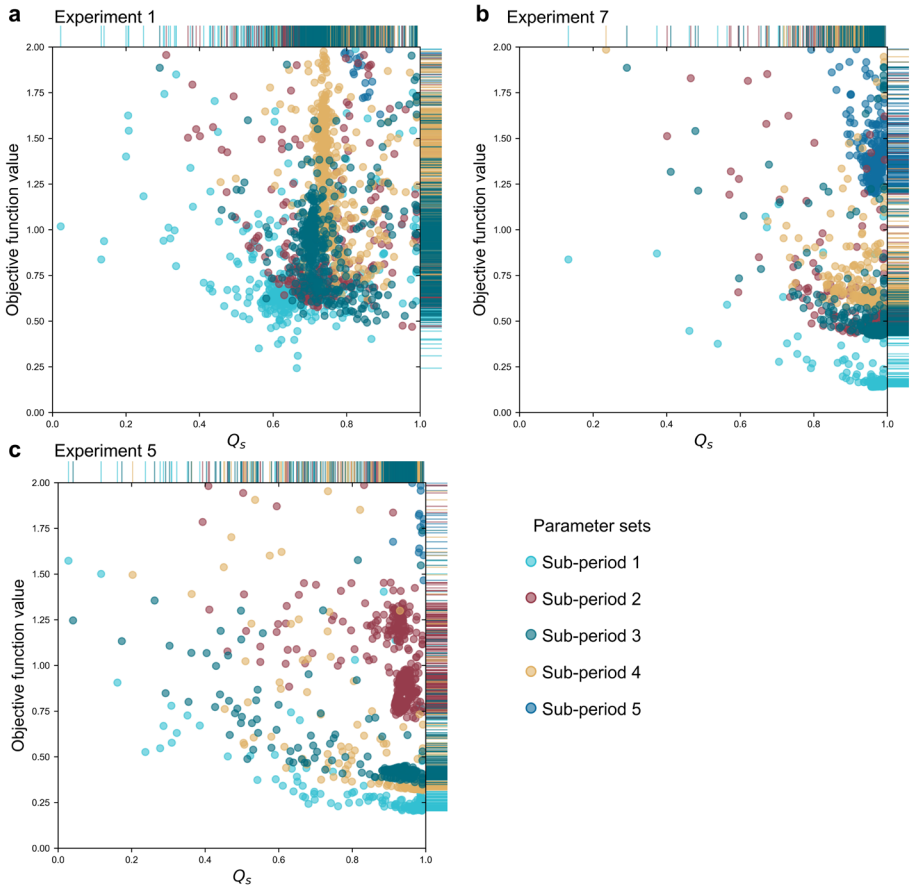

***Figure 9. a**, Flux mapping for case A in the conventional scheme, **b**, Experiment 7, and **c,** Experiment 5, where the horizontal axis represents the proportion of Q_s in the runoff.*

Q13: These limitations and in the structure of the manuscript are compounded by the choice of the figures, which, while informative in parts, suffer from poor organisation and mixed messaging: Figures such as Figure 1 and Figure 6 combine multiple purposes (contextual information, results, conceptual illustrations) in a way that muddles their message. Each figure should ideally present a single, focused point.

Response:

We thank you for your valuable feedback on the presentation of the figures. The initial draft intended to clarify, hoping that readers could see both the framework structure and some results in a single figure. In the revised manuscript, the figures have been reconstructed and separated to improve logical presentation. Figure 1 now displays only the locations of the typical catchments, while the previously included results have been moved to Figure 4 in the Results section. Figure 6 has been split into three separate figures, one focusing on the flux mapping results (Figure 9) and the others on the comparison of parameters or performance metrics (Figures 6 and 10). Captions have been clarified to specify the purpose and content of each figure, allowing readers to interpret the information at a glance. These revisions ensure that each figure addresses a single, focused scientific question, enhancing clarity, readability, and the overall logical flow of the manuscript.

Revised manuscript text:

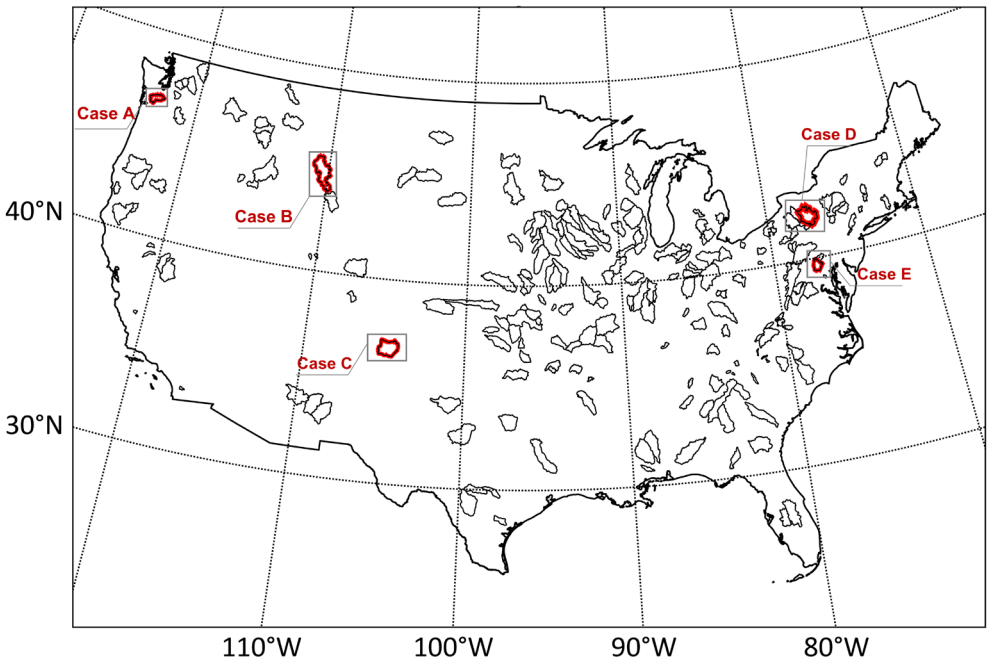

Figure 1. *Location map of the catchment area used in this study, where cases A, B, C, D, and E correspond to catchments 12027500, 6192500, 7211500, 1643000, and 1531000 (from west to east) are highlighted with red outlines for reference.*

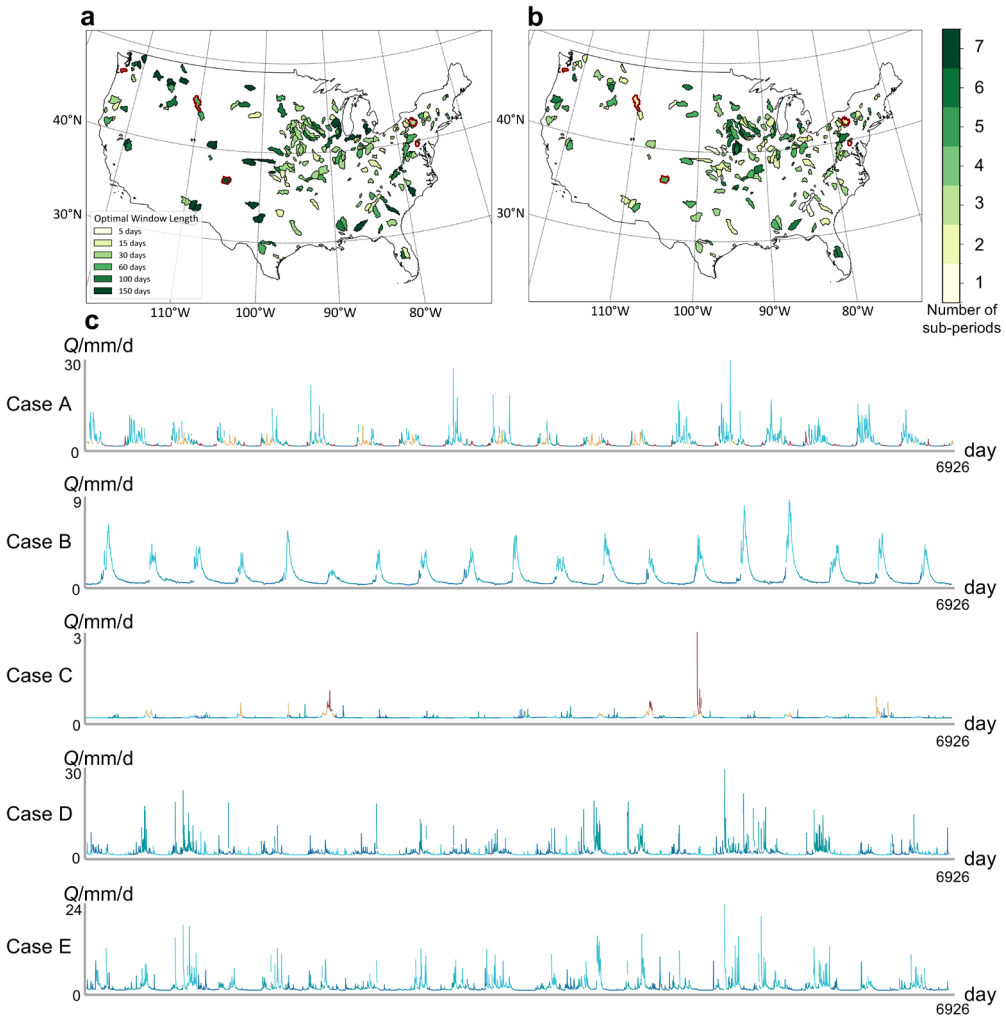

***Figure 4. a**, Optimal window lengths of catchment area used in this study for the sub-period clustering. **b**, Number of subperiods reflecting results from Section 3.2. **c,** Visualization of clustering results on the hydrograph for the respective study cases.*

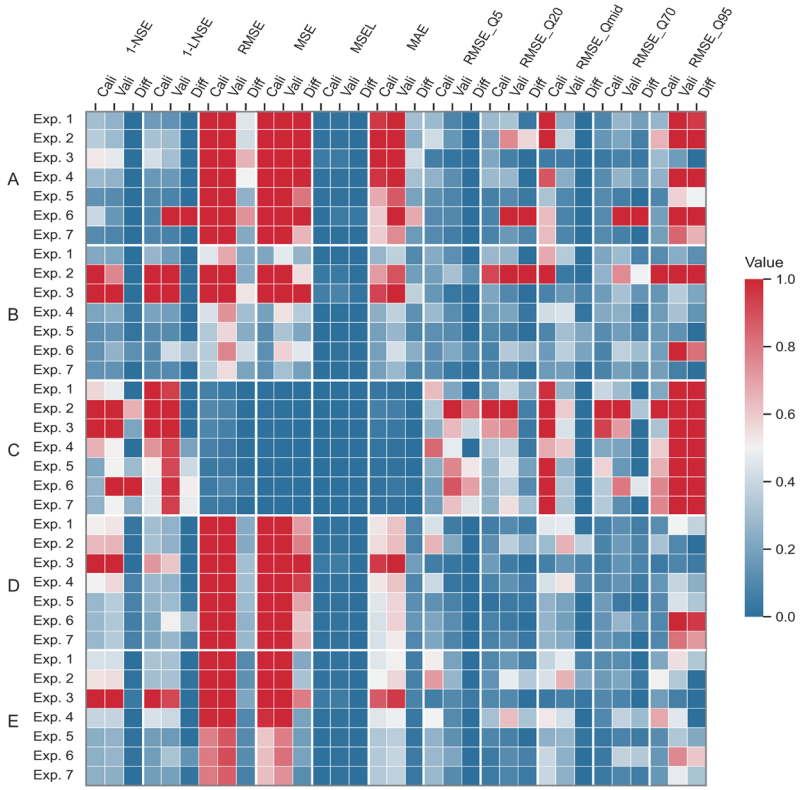

Figure 6. *Model performance of seven experiments in five study cases was assessed using multiple evaluation metrics. Lower values reflect superior performance.*

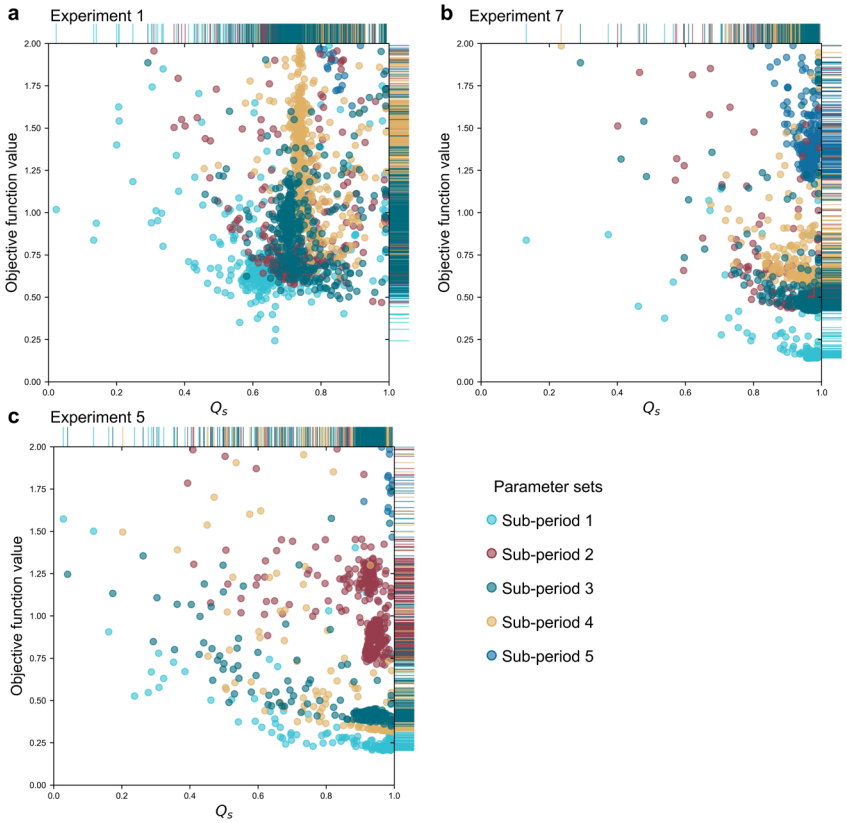

Figure 9. a, *Flux mapping for case A in the conventional scheme,* ***b,*** *Experiment 7, and* ***c,*** *Experiment 5, where the horizontal axis represents the proportion of Q_s in the runoff.*

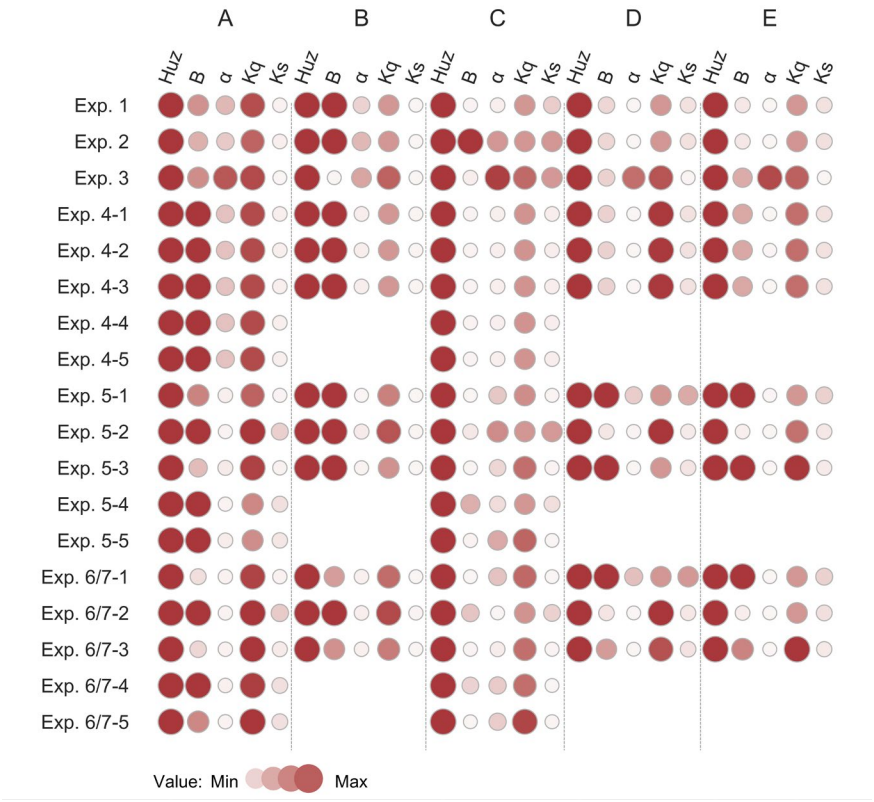

Figure 10. *Assessment of dynamic parameter sets across various calibration experiments. The parameter boundaries shown in the figure are H_{uz} (0-1500), B (0-2), α (0-1), K_q (0.5-1), and K_s (0-0.5).*

Q14: Visuals that directly illustrate performance improvements are lacking. Figure S4 (which compares calibration outcomes across all basins) should be elevated to the main text.

Response:

We greatly appreciate your constructive suggestion. Figure S4 was initially placed in the SI to limit the length of the main text. In the revised manuscript, Section 4.1 of the Results presents a new, comprehensive performance comparison box plot (Figure 5), systematically comparing the performance of all seven calibration experiments across all catchments. The figure clearly illustrates the magnitude of performance improvements across all catchments and provides richer comparative information, thus more strongly supporting the superiority of our recommended experiment. By elevating Figure S4 to the main text in this manner, the completeness and clarity of the results presentation are enhanced. The added Figure 5 is detailed in the response to Q9 (see “4.2 Model performance” in the “Revised manuscript text”).

Q15: Conversely, time series plots (Figures 3 and 4) do not meaningfully add to the manuscript and could be moved to the SI if needed.

Response:

Thank you very much for your valuable suggestion. In the revised manuscript, the presentation of results has been adjusted to enhance focus in the main text. Section 4.3 now retains streamlined versions of Figure 3 (fluxes) and Figure 4 (state variables) with a more compact layout, emphasizing typical sub-periods and key anomalies. Diagnostic annotations, such as arrows and text boxes, have been added to guide readers to the core scientific questions. Meanwhile, complete time series plots have been moved to the Supporting Information (SI) to ensure the data and analysis remain fully accessible. Through this arrangement, the main text can more concertedly highlight the core scientific finding of "the improvement in the response mechanism of hydrological processes to dynamic parameters," while the SI retains the complete chain of evidence for reference. We sincerely thank you for your suggestion again; this adjustment has significantly improved the focus and logical clarity of the paper. The specific revisions in Section 4.3 (lines 361-394) of the revised manuscript are detailed in the response to Q12 (see "4.3 State variables and fluxes" in the "Revised manuscript text").

Scientific Significance and Depth

Q16: Despite its potential, the scientific contribution of this work is undermined by superficial analysis and poor framing of the results: The conclusions (lines 444–451) include trivial points (e.g., trade-offs in multi-objective calibration) that do not substantively add to the field. The third point—regarding sub-period calibration as a remedy to structural model deficiencies—is more interesting, but is not adequately supported by clear, main-text results.

Response:

We greatly appreciate your recognition of the significance and potential of this study. Your recognition provided important motivation to enhance the manuscript and to more effectively demonstrate the scientific value of the proposed framework. In the revised manuscript, the Conclusion has been rewritten to emphasize the core finding that sub-period calibration can, under certain conditions, alleviate model structural deficiencies, with this conclusion directly linked to specific results. In the Results section (Sections 4.1–4.3), key diagnostic results have been incorporated into the main text, including box plots of the performance statistics across all catchments and plots of state and flux changes that reveal the model's internal behavior. In Section 5.3 of the Discussion, by introducing parameter constraints, we found that when the compensatory ability of the overall dynamic parameters is limited, the model performance declines. This demonstrates that the dynamic variation of parameters is largely to compensate for the model's own structural deficiencies. Through these revisions, the conclusions of the revised manuscript more clearly reflect the scientific value of the framework and are closely aligned with the results. Thank you again for recognizing the potential of our research and for your precise guidance on improvement, which has enabled us to more fully present the contributions and significance of the study.

Revised manuscript text:

5.3 Parameter response to catchment dynamics

"In this study, the sub-period clustering method (Section 3.3) is employed to extract the dynamic catchment characteristics of hydrological processes, enabling model parameters to adjust across hydrological periods. This approach improved simulation accuracy and robustness in dynamic catchments, demonstrating the necessity and effectiveness of incorporating dynamic parameters into conceptual hydrological models (Refsgaard et al., 2021). However, a critical question arises: To what extent do dynamic parameter variations represent the true dynamic variability of catchment properties, and to what extent do they

compensate for structural deficiencies of the model itself (Thornton et al., 2022)? To address this problem, a diagnostic experiment is designed. Building on the sub-period calibration framework (Experiment 7), a soft constraint based on globally optimal parameters is introduced, integrating prior information on overall catchment behaviour into sub-period parameter estimation. This design balances the flexibility of dynamic parameter adjustment with the need to preserve physical consistency. The diagnostic objective function is defined as:

$$OF=1-(0.5*NSE+0.5*LNSE) + Penalty \quad (2)$$

where the penalty term quantifies the deviation of the sub-period parameter set $\hat{\theta}_i$ *from the globally optimal parameter set* θ_i*. The penalty is formulated as the mean of the absolute relative errors:* $Penalty = \frac{1}{N} \times \sum \left|\frac{\hat{\theta}_i - \theta_i}{\theta_i}\right|$*, where* i *denotes the parameter index, and* N *is the total number of parameters (five in the HYMOD model). This setting allows assessment of how model responses change when parameter variability is constrained within a more stable and physically consistent range.*

As shown in Fig. 12a, imposing the constraint leads to posterior distributions that are more concentrated within each sub-period, with reduced dispersion, reflecting greater stability. Parameter transferability between calibration and evaluation periods also improved, as illustrated in Fig. 12b, with smaller declines in model performance across periods. However, these gains in parameter stability are accompanied by significant reductions in NSE and LNSE, rendering performance inferior to unconstrained sub-period calibration. This trade-off highlights the compensatory role of dynamic parameters in addressing structural limitations of fixed model formulations. When the capacity of parameters to compensate is constrained, the observed performance decline reflects underlying structural inadequacies in representing key hydrological processes.

The demand for dynamic parameters is often symptomatic of structural insufficiency. A structurally adequate model should maintain stable parameters that represent physical catchment properties. When the model formulation fails to capture essential processes, the

"optimal" parameters must vary dynamically to compensate for these omissions (Beven, 2019). Evidence from the GLUE framework has shown that posterior parameter distributions can diverge almost completely between wet and dry seasons, implying that sub-period calibration with distinct parameter sets effectively corrects structural errors and improves accuracy (Blasone et al., 2008). Moreover, concepts from Data-Based Mechanistic (DBM) modelling and state-dependent parameter (SDP) approaches suggest that time- or state-dependent gains—such as nonlinear filters linked to soil moisture or runoff—can be identified from data. These gains compensate for missing nonlinearities in effective rainfall, often exhibiting dynamic catchment patterns over longer periods. On this basis, the proposed sub-period calibration framework is positioned as a practical means of using parameters as proxy variables to alleviate structural deficiencies, thereby enhancing streamflow simulation accuracy in dynamic catchments (Bouaziz et al., 2022; Terrier et al., 2021)."

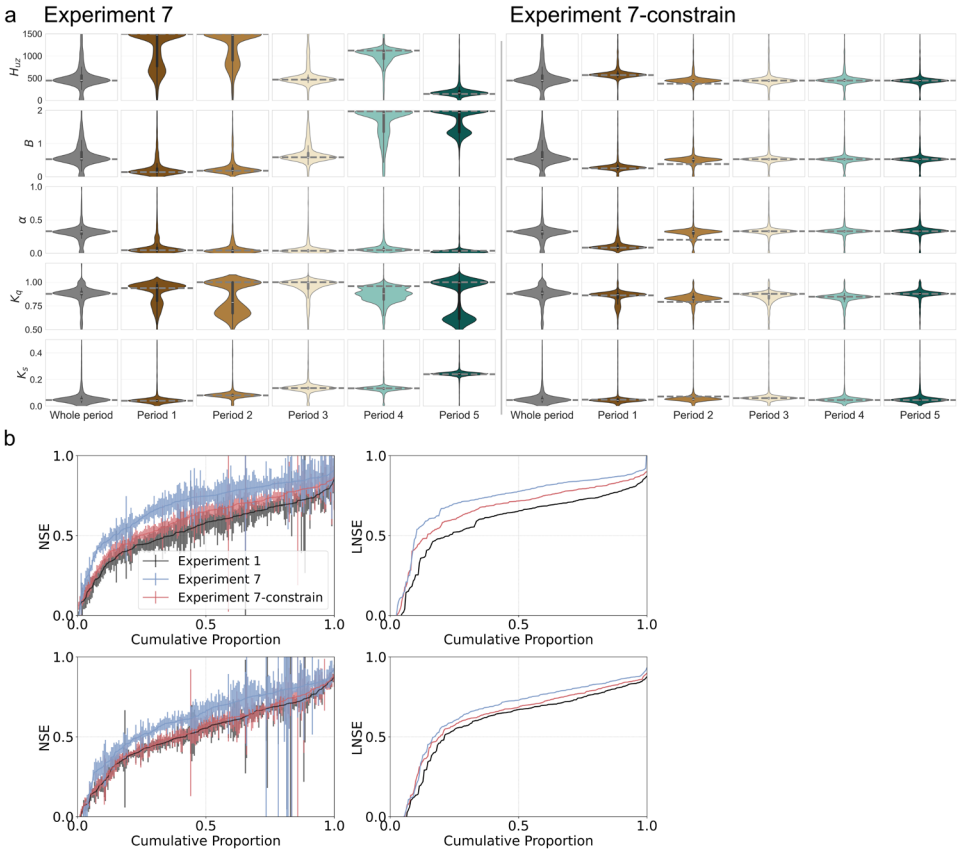

Figure 12. a, *Distributions of the optimal parameter spaces across sub-periods under different climatic and land-surface conditions for case A. Each violin depicts one parameter space, with parameter values on the y-axis; the violin width reflects the probability density of the parameter values. Parameter bounds are: H_{uz} (0-1500), B (0-2), α (0-1), K_q (0.5-1), and K_s (0-0.5). Results for all study cases are provided in S5 of the Supporting Information.* ***b,*** *Cumulative distribution functions (CDFs)*

of NSE and LNSE comparing the experiment built on Experiment 7 with added parameter constraints (blue), Experiment 7 (red), and Experiment 1 (black); higher values indicate better performance. Upper panels show calibration; lower panels show verification. Shaded bands denote 90% bootstrap confidence limits to indicate sampling uncertainty.

Q17: The claim that this is a "novel framework" requires broader evidence of generalisability and applicability across a wide range of catchments. How does performance vary with catchment type, climate regime, or data quality?

Response:

Thank you for your valuable suggestion regarding the generalizability of the framework. In the initial draft, a few typical case studies were presented to explain the mechanistic differences between different calibration experiments through in-depth analysis. In the revised manuscript, in addition to retaining the case catchments, summary results for all catchments have been added (see Section 4.1 of the Results) to show the overall performance of the different calibration experiments. And box plots in Figure 5 present this information, enabling readers to assess performance variability and general trends across all catchments. These results show that the recommended sub-period calibration experiment performs better than the time-invariant parameter experiment in most catchments, especially in mid-to-low flow periods and in terms of process consistency. Moreover, the title is adjusted to "A Robust Calibration and Evaluation Framework for Dynamic Catchment Characteristics in Hydrological Modeling", which will better reflect the contribution of our work. We sincerely thank you for your suggestion. By combining the case study catchments with the full sample results, the robustness of the framework in different scenarios can be more adequately demonstrated, thereby enhancing the reliability of the research conclusions. The specific revisions in Section 4.2 (lines 300-330) of the revised manuscript are detailed in the response to Q9 (see "4.2 model performance" in the "Revised manuscript text").

Q18: The "Discussion" section falls short of its purpose. It lacks depth, avoids key limitations, and does not engage with broader literature on parameter identifiability or structural uncertainty.

In particular, there is a missed opportunity to discuss important limitations and implications. For example: What are the limitations or assumptions of the EDCC clustering approach?

Response:

We appreciate the reviewer's comments on the EDCC approach. The EDCC method primarily serves as a tool for identifying and segmenting dynamic catchment characteristics based on hydrological, meteorological, and land surface information.

To illustrate this limitation, diagnostic experiments were conducted in Section 5.3 in which dynamic calibration was constrained using whole-period optimal parameters. The performance dropped significantly, showing that much of the temporal variability in parameter functions as compensation for structural deficiencies rather than a faithful reflection of environmental signals. This indicates that while EDCC is effective for identifying temporal shifts, bridging the gap to model parameterization requires additional methodological development, such as model structural enhancement. The specific revisions in Section 5.3 of the revised manuscript are detailed in the response to Q16 (see the "Revised manuscript text").

Q19: How does the framework handle equifinality and model realism, beyond scalar performance metrics?

Response:

Thank you very much for pointing out that relying solely on scalar performance metrics is insufficient for assessing model realism. We fully agree with this view and have strengthened the process-level diagnostics for equifinality and model realism in the revised manuscript. In Sections 4.2 and 4.3 of the Results, we have added ternary plots and diagnostic plots of state variables, fluxes, and parameters. These results visually reveal that in some experiments, abrupt parameter changes lead to discontinuities or non-physical anomalies in key state variables like soil moisture content. In contrast, the recommended experiment, combining multi-parameter

dynamization with sub-period calibration, exhibits smoother and more reasonable dynamic responses. Such a comparison not only reflects the impact of different calibration strategies on the internal mechanisms but also provides direct evidence for evaluating model realism beyond scalar metrics, thus supporting the superiority of the recommended experiment at the process level. The specific revisions in Sections 4.2 and 4.3 of the revised manuscript are detailed in the response to Q9 and Q12 (see the "Revised manuscript text").

Q20: Why do dynamic parameters fail to reflect environmental signals in some experiments (e.g., Experiment 4)? Are the algorithms or model structures to blame?

Response:

We appreciate the reviewer's insightful comments. Upon closer examination, we found that the observed inability of dynamic parameters to consistently capture environmental signals is not attributable to a single factor, but to the combined influence of parameter interdependence, equifinality, and the complexity of the optimization landscape. Strong correlations among parameters lead to compensatory effects, whereby temporal variation in one parameter can be offset by static or slowly varying others, effectively masking the signal of interest. In addition, the model exhibits equifinality, meaning that different parameter combinations may yield comparable objective function values while reflecting distinct internal processes, which reduces the identifiability of specific dynamic responses.

The response surface of the model is highly nonlinear and multi-modal, which imposes path dependence and increases the likelihood of convergence toward local optima, even when the optimization algorithm and initial conditions are varied. As a result, introducing time variability in a single parameter, as in Experiment 4, only marginally improved model performance. By contrast, when groups of parameters were allowed to vary jointly or smooth transitions were imposed, as in Experiments 5 and 7, the model achieved more consistent improvements across both high- and low-flow phases, suggesting that collective parameter adaptation is necessary to overcome the identifiability limits imposed by correlation and equifinality. To further clarify

this mechanism, we designed diagnostic experiments with penalty constraints that anchored dynamic parameters toward whole-period optima. The significant performance decline under these conditions indicates that dynamic parameters serve, at least in part, as compensatory mechanisms for structural limitations of the model, which explains why their environmental signals may not always be directly observable.

Taken together, these results lead us to conclude that the phenomenon should not be ascribed simply to either algorithmic or structural deficiencies, but rather to the interplay of parameter dependence, process equifinality, and optimization constraints. This interpretation, along with the supporting diagnostic evidence, is elaborated in Section 5.3 of the Discussion, which can be seen in the response to Q16 ("Revised manuscript text").

Q21: How are the results influenced by the specific structure of the HYMOD model? How generaliseble are they?

Response:

Thank you for reminding us to reflect on the influence of the model structure on the results. HYMOD was chosen due to its relatively simple structure and the clear physical interpretation of its parameters, which facilitates an intuitive exploration of the relationship between parameter dynamization and process responses. The core contribution of this study is to propose a general diagnostic calibration framework to explore the common problems faced in dynamic calibration in dynamic catchments. Moreover, the Results and Discussion sections provide targeted analyses of the variations in HYMOD's parameters, state variables, and fluxes, demonstrating how the framework captures both internal process responses and structural limitations. The simple structure of the model actually magnifies these challenges, making phenomena such as the parameter compensation effect in Experiment 4 and the physical discontinuities caused by abrupt parameter changes in Experiment 6 more apparent. This, in fact, strengthens our conclusion that these are fundamental issues inherent in dynamic calibration, not just specific noise that appears only in complex models. Therefore, as

emphasized in the discussion section, the contribution of this study does not depend on the specificity of a particular model. Instead, it proposes a diagnostic and calibration framework that can be applied in different modeling environments, providing a general approach for evaluating parameter dynamization and diagnosing parameter identifiability and structural uncertainty, which has broader applicability. The rationale for selecting HYMOD in Section 3.1 is detailly in the response to Q4, and the application results of HYMOD in Sections 4.2 and 4.3 of the revised manuscript are detailed in the response to Q9 and Q12, see the "Revised manuscript text".

Revised manuscript text:

Section 6 (lines 578-581)

"The calibration and evaluation framework proposed in this study not only addresses defects caused by the simplification of model structure for hydrological models but also enhances model simulation accuracy across different flow phases and effectively reduces model uncertainty. The evaluation framework comprehensively assesses the performance of hydrological models through multi-criteria evaluation and reveals sources of uncertainty in model internal operation from the perspectives of state variables and fluxes."

Minor Comments

- Line 1 (title): I'm not sure if the novel approach relates to evaluation, also I have never heard of "dynamic" catchments, maybe seasonal is a better term here.

Response:

Thank you for your valuable suggestion. The term "dynamic catchment characteristics" was selected instead of "seasonal catchments" to more precisely capture the focus of this study. Our work centers on proposing a calibration and evaluation framework that addresses the time-varying nature of a catchment's hydro-meteorological and land-surface conditions. While

seasonality is a significant component of these variations, the term “dynamic characteristics” encompasses a broader range of temporal scales, including inter-annual and non-periodic changes, which also present challenges for hydrological model simulation (Reusser et al., 2009). Our framework is designed to capture these time-varying features through dynamic parameterization to compensate for model structural deficiencies. Therefore, we feel that “dynamic catchment characteristics” better reflect the universality and central scientific question of our research (Chagas et al., 2024). The term “Evaluation” has been retained in the title, as it accurately describes the comprehensive diagnostic analysis of the hydrological model’s internal states and fluxes—not just the outlet streamflow—which is a key component of the framework.

Revised manuscript text:

Title

“A Robust Calibration and Evaluation Framework for Dynamic Catchment Characteristics in Hydrological Modeling”

- Line 35: projecting, rather than predicting is a better term in this contextual.

Response:

We greatly appreciate your detailed review and valuable comment. In the context of this paper, using “projecting” is more accurate and appropriate than “predicting.” The former is typically used for long-term trend extrapolation and warning, which better fits the context of our study discussing the model’s future applicability based on dynamic catchment characteristics. In contrast, “predicting” is more common for short-term forecasting and does not adequately convey the research's scope. Following your suggestion, the “predicting” has been changed to “projecting” in the revised manuscript to ensure scientific rigor and precision. Thank you again for helping us improve the professionalism and accuracy of our paper.

Revised manuscript text:

Section 1 (lines 36-37)

"Hydrological models serve as essential tools in water management, supporting tasks such as runoff projection, disaster warning, and water-resource planning."

- Line 85 (and throughout): the term "dimensionality disaster" sounds very grandiose and pompous, I would avoid it, but needs to at best be better defined.

Response:

Thank you for your suggestion. We have accepted this comment and revised the term "dimensionality disaster." Considering that the context of our study primarily concerns the uncertainty and identifiability issues arising from an increased number of parameters, rather than the "curse of dimensionality" in its strict sense, we have uniformly replaced it with "issues of parameter dimensionality" in the revised manuscript. This phrasing more accurately reflects the actual problems discussed in the study while avoiding potentially misleading terminology. We sincerely thank you for your reminder, which has made our academic expression more rigorous.

Revised manuscript text:

Section 1 (lines 94-95)

"Experiments 4–7 explore issues in dynamic parameter calibration, such as parameter correlation, dimensionality, and state transitions."

Section 3.3 (lines 201-202)

"Experiment 5 makes all parameters dynamic, raising issues of parameter dimensionality."

- Line 90: experiments are conducted, not verified.

Response:

Thank you very much for the reminder. We have carefully checked the entire text and have standardized the verb used with "experiments" to "conducted" to maintain accuracy. We sincerely thank you for your comment, which has made our academic expression more standard.

Revised manuscript text:

Abstract (lines 23-24)

"Seven calibration experiments were conducted to explore issues related to time-invariant parameters, objective function configurations, parameter correlations, dimensionality in global optimization, and abrupt parameter shifts."

- Lines 99-101: the criteria used here should be better explained (even in the SI). In particular criteria 2 and 3 feel very subjective.

Response:

Thank you very much for the reminder. In response to your concerns about subjectivity, we have elaborated on these criteria in the revised manuscript under Section 2, "Study area." In the revised manuscript, the three objective principles have been clarified to guide the selection process to ensure its transparency and reproducibility. (1) Data integrity: Only catchments with complete, continuous, and physically reasonable records for the entire study period (1983–2000) were included. This served as a strict data quality control requirement. (2) Minimal anthropogenic influence: We prioritized catchments with the least interference from human activities (e.g., reservoir regulation, large-scale land-use changes) in both temporal and spatial dimensions. This is not a subjective judgment but a standard criterion in hydrological research aimed at isolating and studying natural hydrological processes as much as possible. (3) Coverage of diverse hydro-climatic conditions: We intentionally selected catchments with a broad spatial distribution to encompass diverse meteorological and underlying surface

conditions. The principle is crucial for evaluating the generalizability and robustness of our proposed framework and is an objective requirement to ensure our conclusions are broadly representative. By explicitly articulating these principle-based criteria, the selection process is rendered clearer, more rigorous, and less prone to subjective interpretation. Thank you again for your constructive feedback.

Revised manuscript text:

Section 2 (lines 102-106)

"Rigorous screening criteria were applied to ensure the acquisition of high-quality data. The screening process involved three key considerations: (1) no missing or non-physical data throughout the study period; (2) minimal interference from anthropogenic influences in both temporal and spatial dimensions; and (3) a large spatial distribution scale of the selected catchments, including diverse meteorological and underlying surface conditions."

- Lines 111-123: this section requires better referencing of the methods described.

Response:

Thank you for pointing out the issue of insufficient referencing. In the revised manuscript, this section has been integrated into the new Methods section (3.2 Clustering hydrological processes), and corresponding classic or representative references have been added to enhance the credibility of the methods used. The specific revisions, implemented in Section 3.2, are detailed in the response to Q2 (see "Revised manuscript text").

Revised manuscript text:

Section 3.2 (lines 155-158)

"Sub-period calibration provides a practical means of linking dynamic catchment characteristics with hydrological models. In sub-period calibration, the simulation period is

clustered into multiple sub-periods characterized by relatively homogeneous hydrological conditions, allowing dynamic parameters to better reflect temporal variations in catchment behaviour across different streamflow regimes (Zhang and Liu, 2021).”

- Lines 114, 117, 119: check the format of the bullet points here.

Response:

Thank you very much for your detailed review. We have standardized and formatted the corresponding list items to ensure the layout complies with academic standards.

Revised manuscript text:

Section 3.2 (lines 159-162)

“The methodological framework consists of three key steps: (1) constructing a dynamic catchment characteristic index system to describe catchment states; (2) extracting dynamic catchment characteristics through screening and dimensionality reduction; and (3) applying unsupervised clustering to cluster the time series into sub-periods with similar hydrological processes for subsequent sub-period calibration.”

- Lines 132-133: “calibrating” and see point on line 1 with respect to “dynamic catchments”.

Response:

Thank you for your correction. We have replaced it in the revised manuscript to better introduce the calibration experiments.

Revised manuscript text:

Section 3.2 (line 192-193)

“To systematically evaluate how calibration strategies capture catchment dynamics and

improve the simulation of diverse flow regimes, a diagnostic framework comprising seven calibration strategies is developed."

- Line 133: the Pareto-based method need better introduction and references.

Response:

Thank you very much for the reminder. In the revised manuscript, we have modified the optimization method for Experiment 2, uniformly adopting the SCE-UA algorithm and achieving equivalent multi-objective optimization through a weighted single-objective function. Therefore, the Pareto-based method (NSGA-II) mentioned in the original manuscript has been removed. We have provided a detailed explanation of the new optimization method in Section 3.3 of the Methods and have accordingly deleted the Pareto-related descriptions and citations to maintain consistency between the methods and the main text.

Revised manuscript text:

Section 3.3 (lines 210-212)

"Experiment 2 approximates a multi-objective calibration by combining NSE and LNSE into a weighted objective: $w \times NSE + (1–w) \times LNSE$. The weight w varies from 0 to 1 (step = 0.05), forming a series of single-objective optimizations using SCE-UA with time-invariant parameters. This setup explores trade-offs between flow regimes without changing the optimization algorithm."

- Line 141: despite their name, in a simple conceptual model, parameters don't really have "clear physical meaning".

Response:

Thank you for your suggestion. We have changed "clear physical meaning" to the more rigorous and conceptually appropriate "empirical physical interpretations" to indicate that these

parameters are generalized representations of complex physical processes, rather than direct physical quantities.

Revised manuscript text:

Section 3.1 (lines 135-138)

"To evaluate and compare the applicability of different calibration strategies under dynamic catchment conditions, the simple conceptual hydrological model, HYMOD (Hydrological MODel) (Moore, 1985), is employed for verification. The HYMOD model is a conceptual rainfall-runoff model with a simple structure (five parameters), low input requirements, and empirical physical interpretations."

- • Line 142: the reference to Fig 2 here seems out of place, please check.

Response:

Thank you for your careful correction. We have checked and adjusted the reference to Figure 2 at this location, making its position more logical and consistent with the surrounding text.

Revised manuscript text:

Section 3.3 (lines 193-195)

"These experiments sequentially address key challenges in representing time-varying hydrological behaviour, with a focus on objective function design and time-varying parameterization (Fig. 3)."

- Line 152 (and throughout): "validation" is a rather controversial term when it comes to modelling. Please use "evaluation" instead.

Response:

We greatly appreciate this important comment. We have carefully reviewed the entire manuscript and have uniformly changed all instances of "validation" to "evaluation" to ensure the scientific and standard use of terminology. We sincerely thank you for your detailed review; this change has significantly improved the professionalism and accuracy of the manuscript.

- Lines 152-154: the sentence "It should be noted…", while true feels deceptive. The authors did not actually test this with any other model.

Response:

Thank you very much for pointing out that this sentence could be misleading. The relevant wording has been revised to ensure the rigor of our conclusions. The core contribution of this study is to propose and test a general diagnostic calibration framework to systematically investigate the challenges in seasonal catchment modeling. We explicitly state that this framework is designed to apply to all conceptual models, but the scope of the current study is indeed limited to the systematic testing of one model. The framework would be extended to more models for systematic evaluation in our future work. With this revision, the related discussion is more focused and clearer. The specific revisions, implemented in Section 3.1 Hydrological model, are detailed in the response to Q4 (see in the "Revised manuscript text").

- • Line 181: be careful with the use of acronyms, and do not introduce acronyms that haven't been explained previously.

Response:

Thank you for the reminder. We have carefully checked the entire text to ensure that all acronyms are defined with their full names upon their first appearance.

Revised manuscript text:

Section 3.2 (lines 175-177)

"The Maximal Information Coefficient (MIC) is then employed to quantify linear and nonlinear associations between candidate indicators and streamflow, ensuring hydrological relevance. To mitigate multicollinearity and reduce dimensionality, Principal Component Analysis (PCA) is performed, with the first two principal components retained for clustering."

- Line 186: "Parameters" … "are", or "The parameter" … "is".

Response:

Thank you for your careful review. The relevant grammatical issue has been corrected in the revised manuscript.

Revised manuscript text:

Section 4.4 (line 433)

"The dynamic parameter sets, optimized by various calibration experiments across five case studies, are depicted in Fig. 10."

- Line 205: The parenthesis "(n is the number of sub-periods)" is redundant.

Response:

Thank you for the reminder. The redundant parentheses and their content have been removed.

Revised manuscript text:

Section 3.3 (lines 218-219)

"As a result, the number of parameters increases in proportion to the number of sub-periods, generating a high-dimensional calibration space."

- Line 212: In this example, n is five, this sentence needs rewriting.

Response:

Thank you for your suggestion. We have rewritten this section to make its logic clearer.

Revised manuscript text:

Section 4.1 (line 285)

"In all five cases, the number of identified periods ranged from 3 to 5."

- • Line 255: "Flux mapping".

Response:

Thank you for the reminder. To ensure terminological consistency, we have uniformly used the key term "flux mapping" throughout the manuscript.

- Line 291: What "evaluation metrics" are being referred to here? Is this an average of all of them? There needs to be additional clarity on how this is being evaluated.

Response:

Thank you very much for pointing out the lack of clarity. In the revised manuscript, the ambiguous collective term "evaluation metrics" has been avoided. In Section 4.1 of the Results, we directly present the performance distribution of two specific and representative metrics, NSE and LNSE, across all catchments (see Figure 5), thereby clearly comparing the performance of different experiments in high and low flows. Furthermore, in the in-depth analysis of the case study catchments (see Figure 6), a series of specific metrics is presented, including RMSE, MAE, and RMSE for different flow segments, making the performance evaluation more comprehensive and transparent. The related revisions in Section 4.1 is detailly detailed in the response to Q2, and the revised Section 4.2 is detailed in the response to Q9, see the "Revised manuscript text".

- Line 343 and 348 (Figures 3 and 4): Are these calibration or evaluation results? The caption says one thing, the legend a different one.

Response:

Thank you for your detailed review. The captions of all figures have been strictly checked and corrected in the revised manuscript to resolve the ambiguity you pointed out. For the time series plots you mentioned (now Figure 7 and Figure 8), their captions now clearly state that the results shown are from their respective evaluation periods. With these changes, we have ensured the consistency and clarity of the figure information to avoid any misunderstanding by the readers. The related revisions of figures in Section 4.3 is detailly detailed in the response to Q12, see the "Revised manuscript text".

- Line 410: "abnormal" seems like a charged word for this.

Response:

Thank you for the reminder. We have replaced "abnormal" with "non-physical" to use a more neutral and academic expression to describe the discontinuity in parameters or state variables.

Revised manuscript text:

Section 5.1.3 (line 492-493)

"While the causes of non-physical dynamic parameter values are complex, they might be partially attributed to the failure of global optimization algorithms to converge and find approximated global optimal solutions during the evolutionary process."

References

Beven, K.: How to make advances in hydrological modelling. Hydrology Research, 50(6), 1481-1494, https://doi.org/10.2166/nh.2019.134, 2019.

Bouaziz, L. J., Aalbers, E. E., Weerts, A. H., Hegnauer, M., Buiteveld, H., Lammersen, R., ... & Hrachowitz, M.: Ecosystem adaptation to climate change: the sensitivity of hydrological predictions to time-dynamic model parameters. Hydrology and Earth System Sciences, 26(5), 1295-1318, https://doi.org/10.5194/hess-26-1295-2022, 2022.

Chagas, V. B., Chaffe, P. L., & Bloeschl, G.: Regional low flow hydrology: Model development and evaluation. Water Resources Research, 60(2), e2023WR035063, 2024.

Duan, Q. Y., Gupta, V. K., and Sorooshian, S.: Shuffled Complex Evolution Approach for Effective and Efficient Global Minimization, Journal of Optimization Theory and Applications, 76, 501-521, https://doi.org/10.1007/BF00939380, 1993.

Hsueh, H. F., Guthke, A., Wöhling, T., & Nowak, W.: Optimized predictive coverage by averaging time-windowed Bayesian distributions. Water Resources Research, 60(5), e2022WR033280, https://doi.org/10.1029/2022WR033280, 2024.

Moore, R. J.: The probability-distributed principle and runoff production at point and basin scales, Hydrological Sciences Journal, 30, 273-297, https://doi.org/10.1080/02626668509490989, 2009.

Refsgaard, J. C., Stisen, S., & Koch, J.: Hydrological process knowledge in catchment modelling–Lessons and perspectives from 60 years development. Hydrological Processes, 36(1), e14463, https://doi.org/10.1002/hyp.14463, 2021.

Reusser, D. E., Blume, T., Schaefli, B., & Zehe, E.: Analysing the temporal dynamics of model performance for hydrological models. Hydrology and earth system sciences, 13(7), 999-1018, https://doi.org/10.5194/hess-13-999-2009, 2009.

Terrier, M., Perrin, C., De Lavenne, A., Andréassian, V., Lerat, J., & Vaze, J.: Streamflow naturalization methods: a review. Hydrological Sciences Journal, 66(1), 12-36, https://doi.org/10.1080/02626667.2020.1839080, 2021.

Thornton, J. M., Therrien, R., Mariéthoz, G., Linde, N., & Brunner, P.: Simulating fully-integrated hydrological dynamics in complex alpine headwaters: potential and challenges. Water Resources Research, 58(4),

e2020WR029390, https://doi.org/10.1029/2020WR029390, 2022.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., & Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water resources research, 39(8), https://doi.org/10.1029/2002WR001746, 2003.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., & Sorooshian, S.: A framework for development and application of hydrological models. Hydrology and Earth System Sciences, 5(1), 13-26, https://doi.org/10.5194/hess-5-13-2001, 2001.

Wang, Y., Wang, J., Xie, J., and Lu, H.: Improvements in the degree-day model, incorporating forest influence, and taking China's Tianshan Mountains as an example, Journal of Hydrology: Regional Studies, 44, https://doi.org/10.1016/j.ejrh.2022.101215, 2022a.

Zhang, X. and Liu, P.: A time-varying parameter estimation approach using split-sample calibration based on dynamic programming, Hydrology and Earth System Sciences, 25, 711-733, https://doi.org/10.5194/hess-25-711-2021, 2021.

Reply to Reviewer2

Title: A Robust Calibration and Evaluation Framework for Dynamic Catchment Characteristics in Hydrological Modelling

The paper “A Novel Framework for Calibration and Evaluation of Hydrological Models in Dynamic Catchments” by Lan et al. addresses the important issue of model calibration and proposes a novel framework for calibrating models in so-called "dynamic catchments."

However, in my opinion, the paper suffers from a substantial lack of clarity, an imbalanced presentation of results (with a disproportionate focus on the case study), and omits essential information from the main text.

I recommend that the authors undertake major revisions, reorganize the paper, and include results for all catchments, while shortening the case study analysis. The manuscript should be made clearer and more concise.

Response:

Thank you very much for your meticulous review and for providing such insightful and constructive feedback on our manuscript. Your thoughtful evaluation and your recognition of the scientific value and potential contribution of this work are sincerely appreciated by us. Your comments have been instrumental in improving the overall quality and clarity of the manuscript. In response, we have undertaken a comprehensive major revision based on your valuable recommendations.

The entire structure of the manuscript has been reorganized to improve conceptual clarity and logical progression. The “Methods” section has been redesigned into four integrated components, forming a coherent sequence from hydrological model introduction and sub-period clustering to experimental design and the evaluation system. To address concerns regarding the balance of the results, a large-sample statistical analysis across all dynamic catchments has been incorporated. This addition

enhances the demonstration of the generalizability and stability of the recommended calibration schemes and highlights the broader applicability of the framework under diverse hydro-climatic conditions. Furthermore, the Discussion section has been refined to strengthen comparative analyses and to provide mechanistic explanations. These revisions include a clearer examination of how different calibration strategies influence model stability and dynamic response behaviour, an expanded analysis of parameter uncertainty, equifinality, and sensitivity to dynamic catchment characteristics. The manuscript has also been reviewed throughout to ensure consistency in key terms, simplify the expression, and integrate essential methodological details previously presented in the supplementary materials, thereby improving completeness and readability. In addition to adding summary statistics and boxplots for all 219 MOPEX catchments (Section 4.2), we have also streamlined the descriptions of the five case studies. The case sections now focus on the most representative differences between the calibration experiments, while some detailed time series plots have been moved to the Supporting Information (SI).

Please find our detailed point-by-point responses below. For clarity, all comments are given in black, and responses are shown in blue text. We hope the revised manuscript and our responses adequately address all concerns raised. We look forward to hearing from you regarding the suitability of our revised manuscript for acceptance.

Specific Comments:

Q1: The authors do not clearly define several key terms foundational to the study, including “dynamic catchments”, “dynamic parameters”, “dynamic features”, “seasonal catchments”, “sub-period” and others. While some meanings can be inferred, proper definitions should be provided in the main paper.

Response:

We sincerely thank the reviewer for highlighting the importance of clarity in key

terminology. In response, a careful, thorough review of the manuscript has been conducted to ensure consistency and precision in the use of key terms, and explicit definitions have been provided at their first appearance in the Introduction.

Specifically, we now consistently use four core terms to describe the key components of our framework. Dynamic catchment refers to catchments in which hydrological processes exhibit significant intra-annual and/or inter-annual variability, posing challenges for model simulation. Dynamic catchment characteristic describes the time-varying states of a catchment, characterizing the temporal dynamics of hydrological processes (e.g., the seasonality of precipitation, changes in vegetation cover) within the catchment. Sub-periods are segments of the simulation period characterized by relatively homogeneous hydrological conditions, identified through clustering of the time series. Dynamic parameter refers to model parameters allowed to vary across sub-periods, rather than remaining fixed over the entire simulation period.

Revised manuscript text:

Introduction

"A dynamic catchment is defined as one in which hydrological processes exhibit significant intra-annual or inter-annual variability, making their simulation particularly challenging." (lines 42-44)

"Dynamic catchment characteristics denote the time-varying states of a catchment that describe the temporal evolution of hydrological processes, such as precipitation seasonality and changes in vegetation cover under significant human disturbances." (lines 44-46)

"Sub-periods are segments of the simulation period characterized by relatively homogeneous hydrological conditions, which are typically identified through clustering of the time series." (lines 73-75)

"A dynamic parameter is defined as a model parameter that varies across sub-

periods rather than remaining fixed over the entire simulation period." (lines 72-73)

Q2: In general, the figures are difficult to interpret. The captions lack sufficient explanation, requiring readers to infer too much on their own.

Response:

Thank you very much for your valuable feedback on the presentation of the figures. We recognize that clear and intuitive figures are crucial for effectively communicating research information. In the revised manuscript, all figures and captions have been thoroughly revised, and each caption has been rewritten to provide explicit explanations of the figure elements, the scientific information conveyed, and all included symbols and abbreviations.

For example, Figure 2 from the original submission (now Figure 3) has been redrawn with a caption that outlines the distinctions among the seven experiments, including core differences in objective functions, parameter configurations, and sub-period treatments. The overall figure structure has also been reorganized to strengthen the logical flow of the manuscript. The revised Figure 1 presents the locations of representative catchments; the former result-related components have been transferred to the Results section and are now included in Figure 4. Figure 6 has been divided into three standalone figures to improve clarity. One figure now presents the flux mapping outcomes (now Figure 9), while the others present parameter comparisons and performance metrics (now Figures 6 and 10). The captions have been rewritten to articulate the specific analytical purpose and content of each figure, enabling straightforward interpretation by readers.

These adjustments ensure that each figure addresses a distinct scientific objective and that the visual narrative supports the rigor, clarity, and coherence of the manuscript. Thank you again for the insightful comment.

Revised manuscript text:

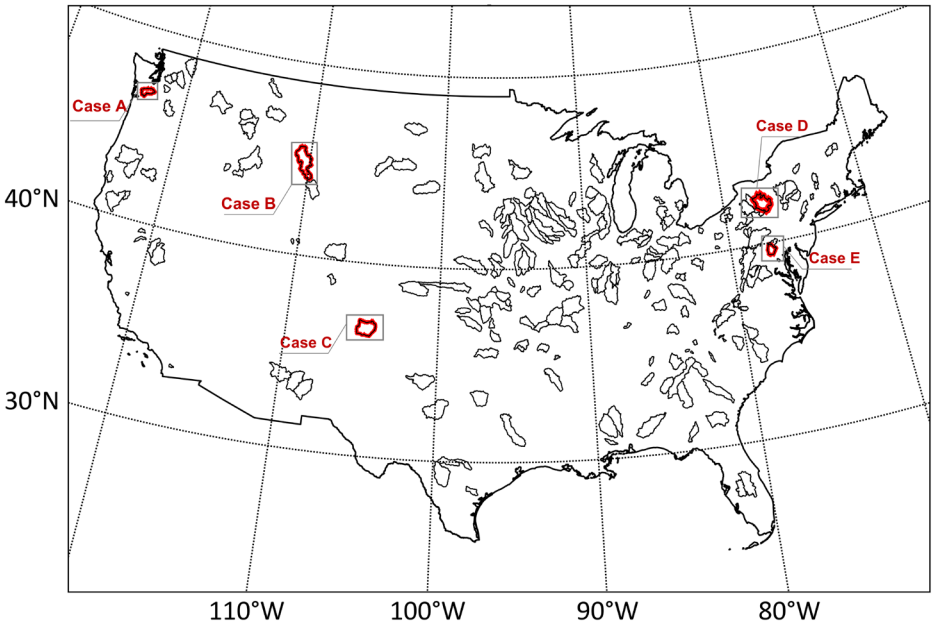

Figure 1. *Location map of the catchment area used in this study, where cases A, B, C, D, and E correspond to catchments 12027500, 6192500, 7211500, 1643000, and 1531000 (from west to east) are highlighted with red outlines for reference.*

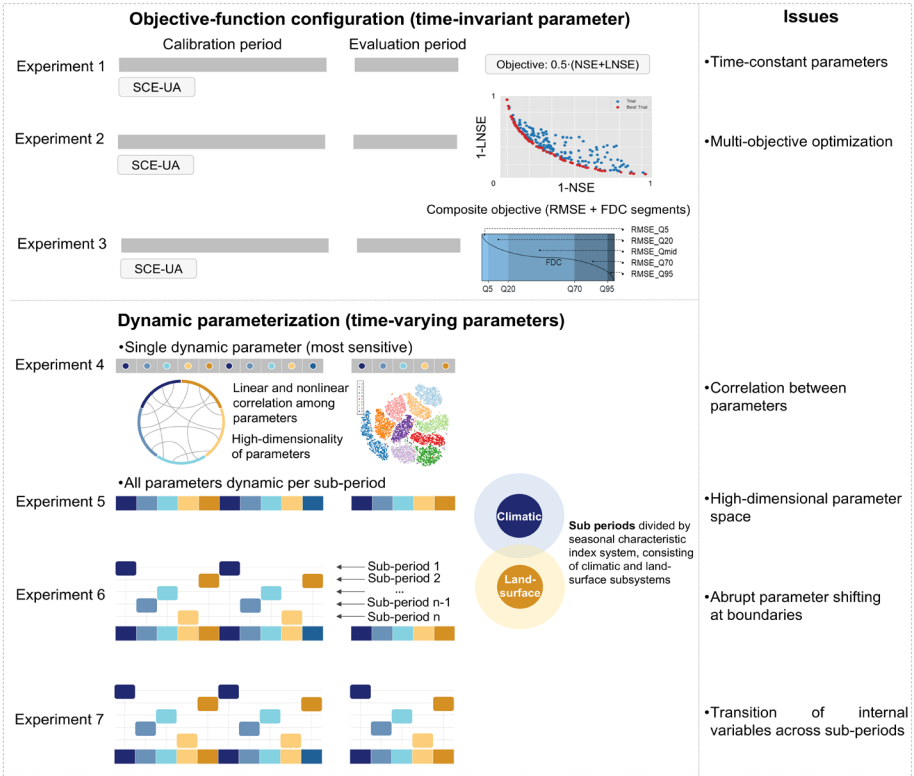

Figure 3. *Schematic illustration of the seven calibration experiments. The colour bands represent state variables and fluxes, which are continuously transferred within the same period. In Experiments 1, 2, and 3, the parameters*

are time-invariant, but the experiments differ in their objective function configurations. Conversely, experiments 4, 5, and 6 maintain a consistent objective function, but vary the parameters across different experiments. In Experiment 4, the dynamic of only the specific parameter is operated, and the other fixed parameters are optimised simultaneously. In Experiment 5, the parameter set is dynamized. The parameter sets in different sub-periods are optimized simultaneously. In Experiment 6, the data from the individual sub-periods are used for minimizing the objective function, while the model is run for the whole period. In the evaluation period, the parameter set between two consecutive sub-periods is updated accordingly. In Experiment 7, the calibration is the same as in Experiment 6. In the evaluation period, the simulated flow data from each separate sub-period are combined and compared with the observed flow.

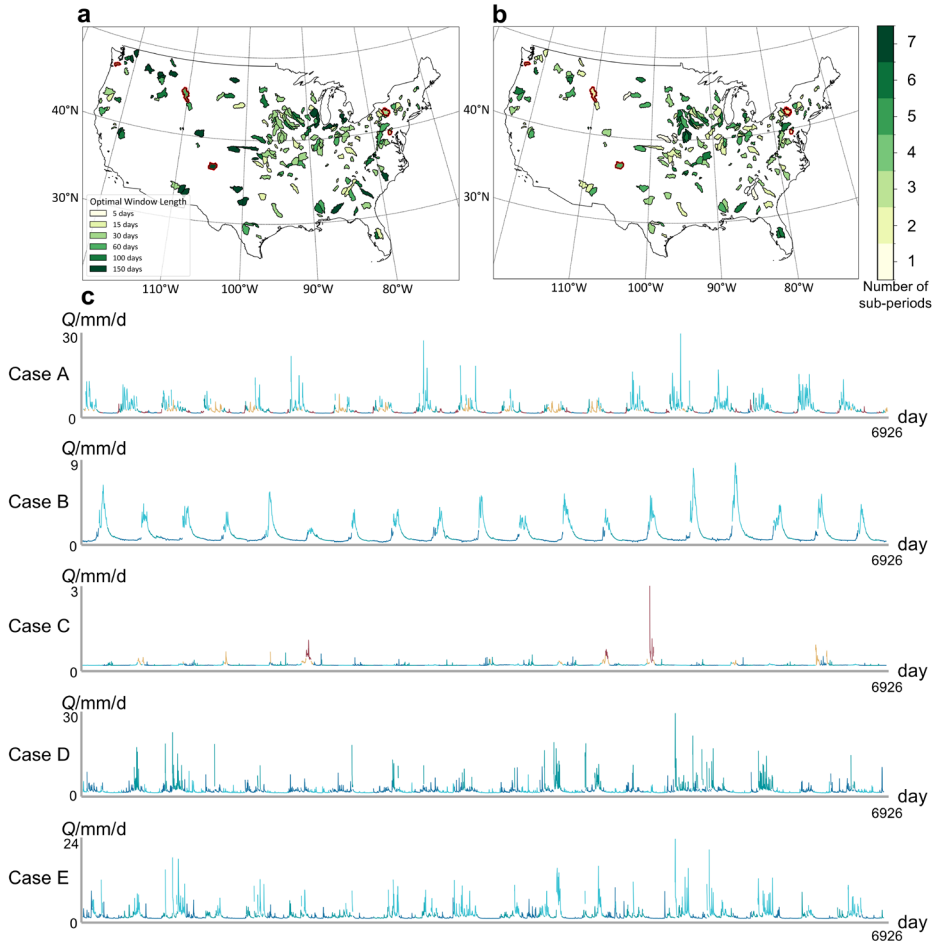

***Figure 4. a**, Optimal window lengths of catchment area used in this study for the sub-period clustering. **b**, Number of subperiods reflecting results from Section 3.2. **c,** Visualization of clustering results on the hydrograph for the respective study cases.*

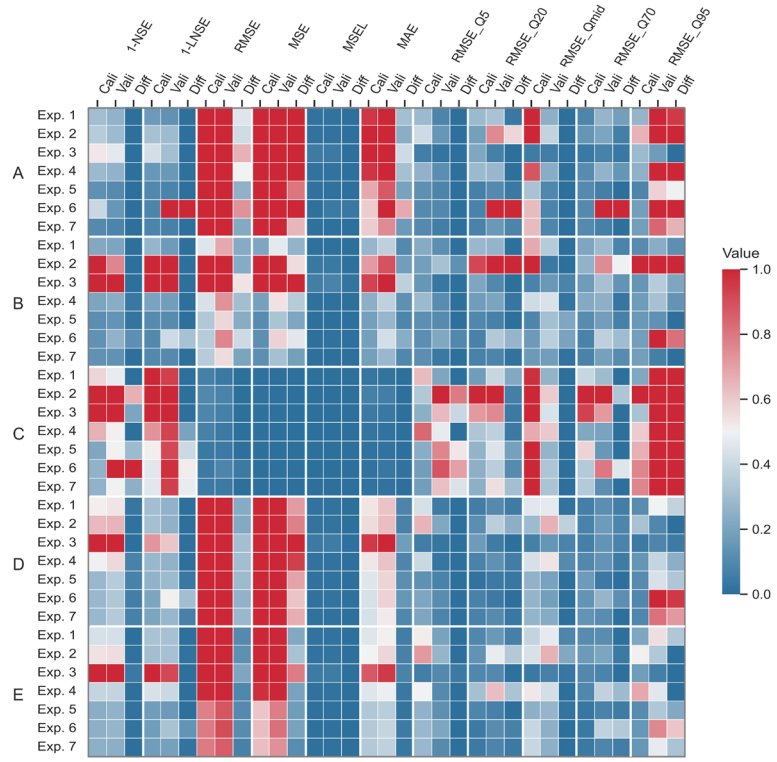

Figure 6. *Model performance of seven experiments in five study cases was assessed using multiple evaluation metrics. Lower values reflect superior performance.*

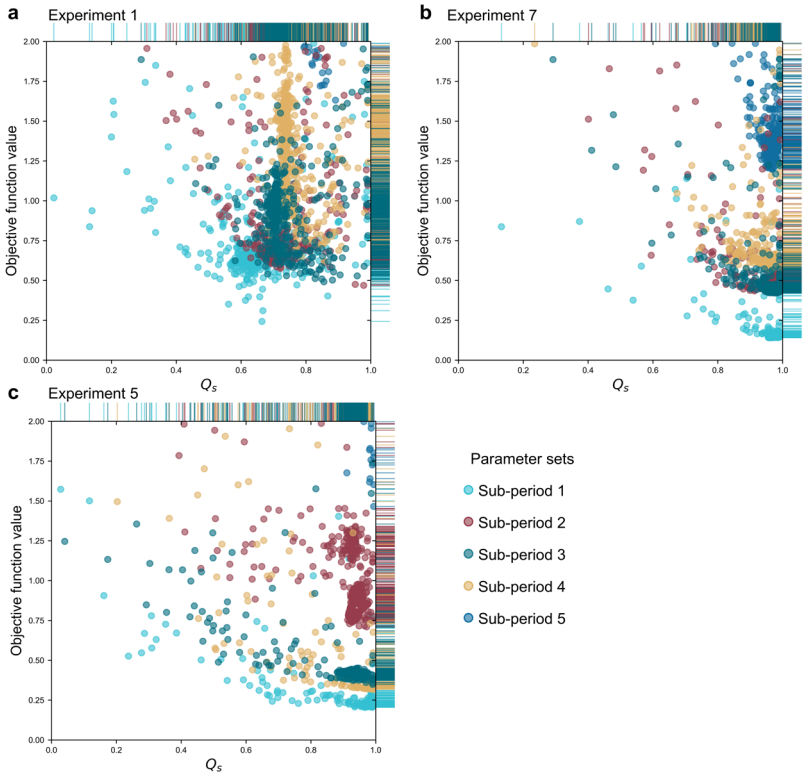

Figure 9. a, *Flux mapping for case A in the conventional scheme,* ***b,*** *Experiment 7, and* ***c,*** *Experiment 5, where the horizontal axis represents the proportion of Q_s in the runoff.*

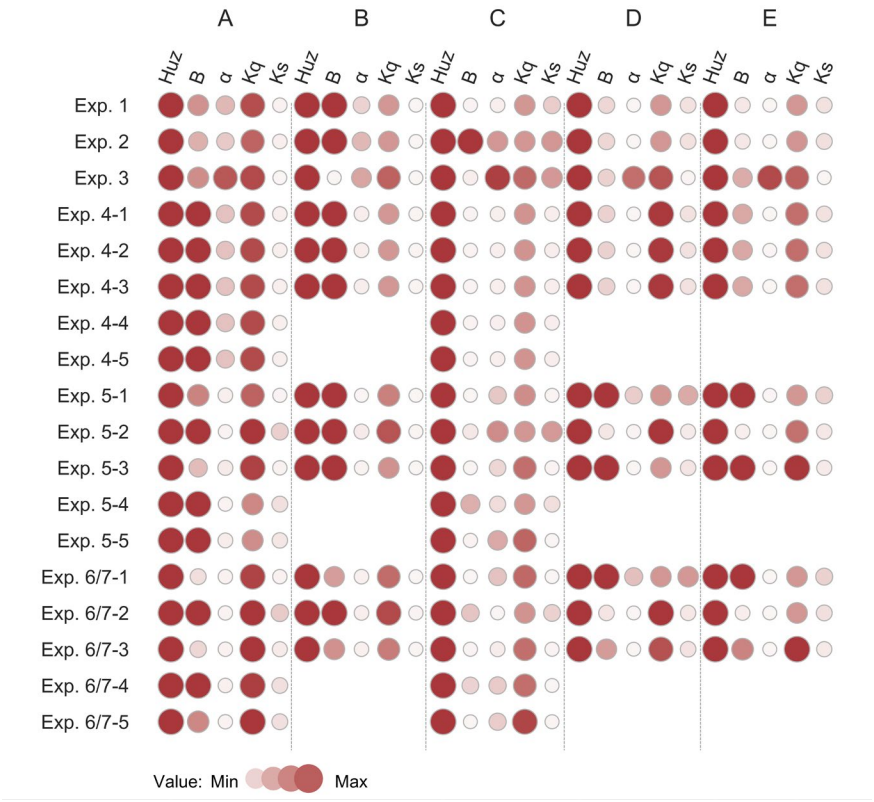

Figure 10. *Assessment of dynamic parameter sets across various calibration experiments. The parameter boundaries shown in the figure are H_{uz} (0-1500), B (0-2), α (0-1), K_q (0.5-1), and K_s (0-0.5).*

Q3: What are the characteristics of the four selected basins? Why were these basins specifically chosen?

Response:

Thank you for your comments. Clarifying the basis for case study selection is essential for demonstrating the framework's generalizability. In the revised manuscript, the number of case studies has been expanded from four to five, and the selection criteria and rationale are detailed in Section 2 ("Study area"). The chosen catchments were selected to provide sufficient diversity and representativeness for testing the robustness and applicability of the proposed framework across varying natural conditions. These five catchments span different climate zones, topographies, and hydrological characteristics, ranging from humid to semi-arid regions and from plains to mountainous areas. Detailed information is provided in Table 1 of the revised

manuscript. This arrangement ensures that the framework's performance can be evaluated under a range of scenarios, supporting its applicability and reliability.

Revised manuscript text:

Section 2 Study area

"The Model Parameter Estimation Experiment (MOPEX) is an international project aimed at developing enhanced techniques for a priori estimation of parameters in hydrologic models and land surface parameterization schemes of weather and climate models (Duan et al., 2006). A comprehensive MOPEX database has been developed that contains historical hydrometeorological data and land-surface characteristics data for numerous hydrological catchments in the United States (US) and other countries. This study utilises the dataset from 219 catchments spatially distributed across the contiguous US (Fig. 1a). Rigorous screening criteria were applied to ensure the acquisition of high-quality data. The screening process involved three key considerations: (1) no missing or non-physical data throughout the study period; (2) minimal interference from anthropogenic influences in both temporal and spatial dimensions; and (3) a large spatial distribution scale of the selected catchments, including diverse meteorological and underlying surface conditions. The dataset for selected catchments includes the hydrometeorological forcing data, land-surface data, and streamflow data, covering the period from 1983 to 2000. Hydrometeorological data includes daily precipitation data (P), temperature data (T), and streamflow (Q) provided by the MOPEX dataset, as well as potential evaporation data (PE) calculated by the Hamon model (Mccabe et al., 2015). The Normalized Difference Vegetation Index (NDVI) was used as one of the land-surface indicators to represent the vegetation coverage of the catchments, which had a spatial resolution of 8 km and a temporal resolution of half-monthly intervals (Tucker et al., 2010). Based on these criteria, a total of 219 catchments were selected (Fig. 1a), spanning a wide range of hydrological and meteorological characteristics, making them ideal for testing various model

structures under diverse conditions (Duan et al., 2006).

In addition to the large-sample analysis of the MOPEX dataset, five representative catchments, Case A (12027500), Case B (6192500), Case C (7211500), Case D (1643000), Case E (1531000), are analyzed in more detail as case studies. These catchments encompass a variety of Köppen climate classifications and different dominant dynamic catchment characteristics, facilitating comparison of calibration strategies and evaluation of their robustness under diverse hydroclimatic conditions. Their locations and characteristics are listed in Table 1 and will be analyzed in depth in the subsequent sections."

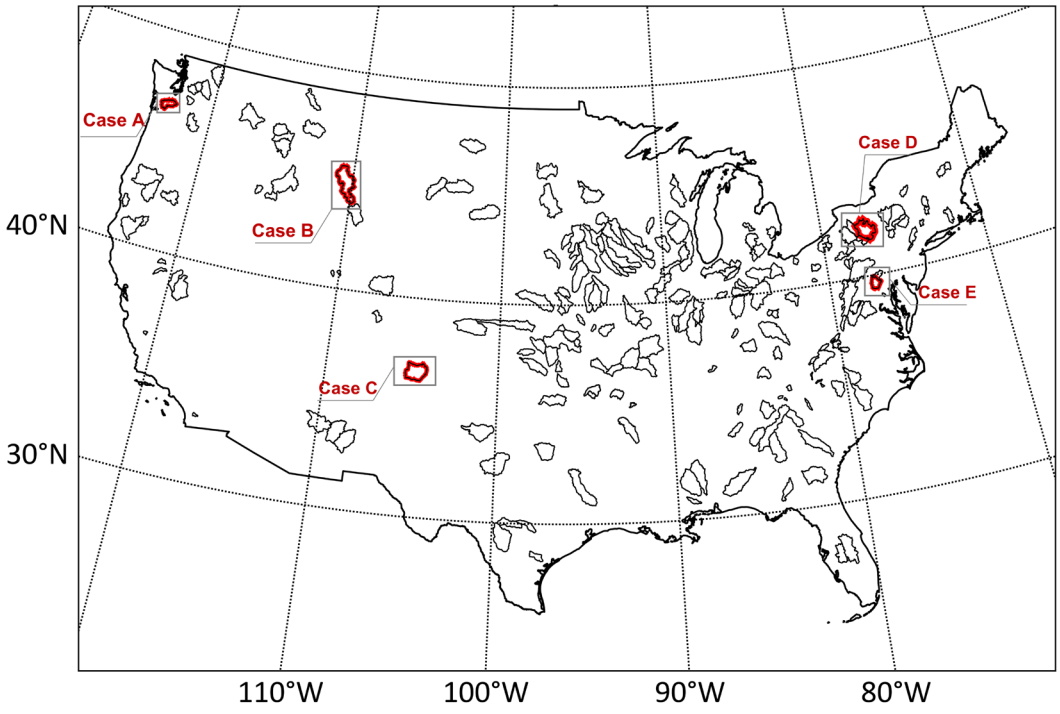

Figure 1. *Location map of the catchment area used in this study, where cases A, B, C, D, and E correspond to catchments 12027500, 6192500, 7211500, 1643000, and 1531000 (from west to east) are highlighted with red outlines for reference.*

Table 1. *Summary of catchment characteristics for study cases.*

ID	*12027500*	*6192500*	*7211500*	*1643000*	*1531000*
Location	*122.99°W*	*110.40°W*	*104.76°W*	*77.25°W*	*77.24°W*
Area (km²)	*895*	*3551*	*2850*	*817*	*2056*
Climate	*Csb*	*Dfc*	*Bsk*	*Cfa*	*Dfb*
Mean P (mm)	*1548.78*	*735.71*	*491.70*	*1068.49*	*870.53*
Mean PE (mm)	*596.53*	*731.59*	*1279.88*	*897.63*	*711.06*
Mean Q (mm)	*1110.19*	*369.79*	*10.08*	*430.15*	*366.76*
Mean elevation	*253.06*	*2441.28*	*2262.91*	*191.80*	*492.25*
Mean slope (°)	*12.16*	*15.26*	*9.44*	*4.99*	*8.25*
Runoff ratio	*0.72*	*0.50*	*0.02*	*0.40*	*0.42*
Aridity index	*2.60*	*1.01*	*0.38*	*1.19*	*1.23*
Forest cover	*71.96*	*36.95*	*16.76*	*31.31*	*57.36*
Land use	*Evergreen Forest,*	*Evergreen Forest,*	*Evergreen Forest, Grassland/Herbac*	*Deciduous Forest,*	*Deciduous Forest,*

Q4: The EDCC (presumably a core procedure in the study) is inadequately explained in the main text. While additional details are provided in the supplementary materials, key components should be moved into the main manuscript for better accessibility.

Response:

We greatly appreciate your attention to the description and presentation of the EDCC method. Although EDCC is not the primary innovation of this study, it provides a critical preprocessing step, that lays the foundation for the systematic comparison and evaluation of different parameter calibration experiments, which form the main focus of the research. To improve readability, the revised manuscript integrates the key implementation steps for sub-period clustering, previously included in the SI, into Section 3.2 (“Clustering hydrological processes”) of the Methods. These steps include the construction of the indicator system, feature extraction, dimensionality reduction, and clustering procedures, allowing for understanding the operational logic directly within the main text. A detailed algorithmic setting is retained in the Supplementary Information for reference. Meanwhile, the corresponding results of the hydrological

process clustering have been relocated to Results 4.1 “Defined sub-periods based on catchment dynamics,” for a clearer and more detailed presentation.

Revised manuscript text:

3.2 Clustering hydrological processes

“Sub-period calibration provides a practical means of linking dynamic catchment characteristics with hydrological models. In sub-period calibration, the simulation period is clustered into multiple sub-periods characterized by relatively homogeneous hydrological conditions, allowing dynamic parameters to better reflect temporal variations in catchment behaviour across different streamflow regimes (Zhang and Liu, 2021). In this study, the clustering of sub-periods is guided by temporal variations in key hydrometeorological and land-surface variables. The methodological framework consists of three key steps: (1) constructing a dynamic catchment characteristic index system to describe catchment states; (2) extracting dynamic catchment characteristics through screening and dimensionality reduction; and (3) applying unsupervised clustering to cluster the time series into sub-periods with similar hydrological processes for subsequent sub-period calibration.

Describing catchment dynamics: *To characterize the temporal dynamics of catchment behaviour, a dynamic catchment characteristic index system comprising a climatic subsystem and a land-surface subsystem is constructed to represent the time-varying states of the catchment. The climatic subsystem includes core hydrometeorological variables such as precipitation (P), temperature (T), and potential evapotranspiration (PE), along with corresponding extreme climatic indicators. The land-surface subsystem reflects evolving surface conditions through indicators such as antecedent runoff, runoff coefficient, and the normalized difference vegetation index (NDVI). All indicators are sampled using a moving window approach, with the optimal window length determined through a time-windowed Bayesian inference framework*

based on predictive log-score (PLS) performance (Hsueh et al., 2024). The framework is designed to preserve long-term trend signals, suppress short-term high-frequency noise, and improve the stability and robustness of dynamic catchment characteristic extraction.

Extracting dynamic catchment characteristics: *Not all indicators exhibit significant dynamic catchment variability; therefore, filtering irrelevant or redundant variables is essential to retain meaningful catchment dynamics. A threshold-based screening is applied to identify variables exhibiting significant seasonality, retaining only relevant subsystems and forming an initial pool of candidate indicators (see Supporting Information S2.1 for detailed criteria). The Maximal Information Coefficient (MIC) is then employed to quantify linear and nonlinear associations between candidate indicators and streamflow, ensuring hydrological relevance. To mitigate multicollinearity and reduce dimensionality, Principal Component Analysis (PCA) is performed, with the first two principal components retained for clustering. This multi-step filtering and reduction procedure ensures robust extraction of dynamic catchment characteristics and provides a solid basis for sub-period clustering according to hydrological similarity.*

Clustering hydrological processes: *Based on the extracted dynamic catchment characteristics, the time series is clustered into distinct sub-periods using the unsupervised Fuzzy C-Means (FCM) clustering algorithm. The optimal number of clusters is determined through a combination of clustering validity indicators, including the Partition Coefficient (SC), Separation Index (S), and Xie–Beni (XB) index, which collectively assess clustering compactness and separation. In addition, the elbow method is employed as a supplementary diagnostic to identify the inflexion point beyond which further increases in cluster number yield diminishing returns. Clustering is performed in the principal component space, enabling effective capture of structural patterns in catchment dynamics. The resulting sub-periods provide a robust foundation*

for integrating dynamic parameters into hydrological models.

In addition, the sub-period clustering is developed exclusively using data from the calibration period. To independently evaluate the generalization capability and robustness of the model under unseen conditions, no model training or parameter adjustment is performed during the evaluation period."

4.1 Defined sub-periods based on catchment dynamics

"To support the implementation of sub-period calibration, periods were identified for all 219 catchments based on variations in dynamic catchment characteristics. The results indicate that dynamic catchment patterns are widespread across the study area, with 219 catchments exhibiting significant variation in at least one hydrometeorological variable (precipitation, temperature, potential evapotranspiration, NDVI, or runoff). Spatially, precipitation seasonality is more significant in the central and western regions; potential evapotranspiration seasonality is widespread, especially in northern areas; runoff seasonality is most evident in the central and northeastern regions; and vegetation seasonality is also common, with only a few high-latitude catchments lacking significant dynamic variation.

A data-driven method was applied to extract relevant information and cluster the time series into distinct periods. The optimal sampling window for each catchment was identified using a Bayesian inference approach, with values ranging from 5 to 150 days (mean = 59.45 days). The MIC was then applied to filter out indicators with weak correlation to runoff. PCA was performed for dimensionality reduction, and the first two components explained, on average, 83.5% of the total variance. Based on the reduced feature space, FCM clustering was used to group time steps, with an average of 4.2 periods identified per catchment.

To illustrate the applicability of the method under diverse hydro-climatic

conditions, five representative catchments were selected, covering a range of climate zones and dominant hydrological drivers. These catchments were also used in the subsequent modelling experiments. As shown in Fig. 4a and Fig. 4b, their optimal window lengths ranged from 30 to 150 days, with 12 to 31 indicators retained after screening. In all five cases, the number of identified periods ranged from 3 to 5. When compared with hydrographs, the identified periods aligned well with key hydrological processes, such as rising and recession limbs (Fig. 4c). In catchments with strong dynamic signals (e.g., Case A and Case B), the identified periods showed stable interannual patterns, while in catchments with greater variability (e.g., Case D and Case E), the clusterings still captured major dynamic catchment characteristics. These period clusterings provide a physically interpretable structure that supports the dynamic parameterization and modelling experiments introduced in the following sections. Considering the performance of the seven modelling experiments across both calibration and evaluation periods, Experiments 5 and 7 are considered the recommended experiments for capturing dynamic catchment characteristics. Experiment 5, with multi-parameter dynamic calibration, achieves high predictive accuracy across diverse flow regimes, although it may slightly compromise physical consistency in runoff generation. Experiment 7, incorporating smooth parameter transitions, maintains comparable accuracy while promoting more consistent and physically reasonable runoff strategies across sub-periods, thus offering a balanced approach between model performance and hydrological interpretability. Detailed analysis of the results will be presented in the following sections."

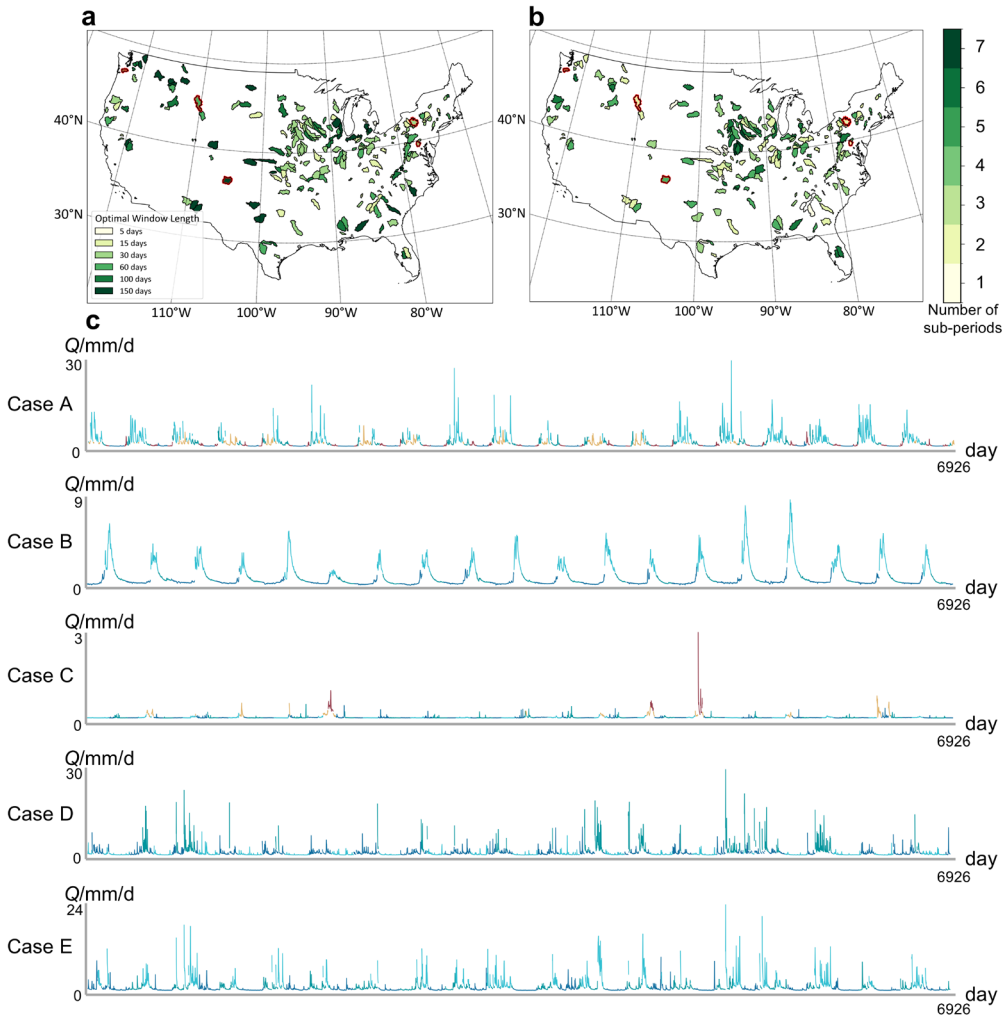

Figure 4. a, *Optimal window lengths of catchment area used in this study for the sub-period clustering.* ***b,*** *Number of subperiods reflecting results from Section 3.2.* ***c,*** *Visualization of clustering results on the hydrograph for the respective study cases.*

Q5: Is the EDCC applied separately for each basin?

Response:

Thank you very much for raising this key methodological question. The EDCC method is applied independently to each catchment. The reason for this approach is that each catchment has significant differences in its climate drivers and underlying surface conditions. Only by constructing an index system and clustering sub-periods for each catchment individually can we more accurately capture its unique dynamic processes. The "tailored" approach is an essential requirement for ensuring that the subsequent

dynamic parameter calibration can effectively reflect the true hydrological processes of a specific catchment. In the revised manuscript, this point has been clarified in Section 3.2 of the Methods to prevent potential misunderstanding. The specific revisions are detailed in the response to Q4 (see "3.2 Clustering hydrological processes" in the "Revised manuscript text").

Q6: Does the EDCC involve any manual decisions, such as determining the number of clusters? Please clarify.

Response:

Thank you very much for your concern about the objectivity of the method. The sub-period clustering process was designed as an unsupervised, data-driven approach to minimize manual intervention. In the revised manuscript, we have provided a clearer explanation of this in Section 3.2 of the Methods (lines 182-189). Specifically, the determination of the number of clusters does not rely on manual specification; it is based on a comprehensive evaluation using multiple objective clustering validity indices, including SC, S, and XB, with the elbow method applied as a diagnostic tool to provide quantitative guidance for selecting the optimal cluster number. All core steps of the sub-period clustering are therefore data-driven, and the decision-making process is fully described in the manuscript, ensuring transparency and reproducibility. The specific revisions in Section 3.2 of the revised manuscript are detailed in the response to Q4 (see "Clustering hydrological processes" in the "3.2 Clustering hydrological processes").

Q7: Is the EDCC applied only during the calibration period? If so, how are its results used in the validation period? Could the method be evaluated using the validation data, for instance by comparing sub-periods defined in the validation period against those from the calibration?

Response:

Thank you very much for your insightful question regarding the connection between the calibration and evaluation periods, which addresses the framework's generalization and robustness. In the revised manuscript, at the end of Section 3.2 of the Methods, we have clearly explained the specific process: the sub-period clustering process is built entirely based on data from the calibration period, without using any information from the evaluation period. During the evaluation phase, we do not perform any new clustering or additional training. Data from the evaluation period are input into the partitioning model established during calibration, assigning each day to its corresponding sub-period, with the associated parameter set applied. This design ensures that the evaluation period remains independent of the calibration process, thus allowing for an objective assessment of its generalization capability and robustness. The specific revisions in Section 3.2 of the revised manuscript are detailed in the response to Q4 (see "3.2 Clustering hydrological processes" in the "Revised manuscript text").

Q8: EDCC results are presented only for the four case studies. Statistical summaries for all catchments should be included.

Response:

Thank you very much for your suggestion. Limitations in length and presentation in the initial draft prevented systematic presentation of large-sample statistical results in the main text. In this revision, these issues have been addressed, and the Results section

has been substantially expanded. In Section 4.2, box plots are presented to visually display the distribution of performance metrics (NSE and LNSE) across all catchments under the seven calibration experiments, comparing the performance of the different experiments at the overall sample level. These statistical results complement the in-depth analysis of the typical case studies, providing mechanistic insights for individual catchments while demonstrating the robustness of the framework across a large sample. We sincerely thank you for your suggestion again; the addition enhances the completeness and clarity of the results presentation.

Revised manuscript text:

4.2 Model performance (lines 300-330)

"To compare seven experiments in dynamic catchments and to identify potential limitations in model calibration, the evaluation is conducted across 219 catchments characterized by hydrological variability. As shown in Fig. 5, the NSE and LNSE values during both calibration and evaluation periods reveal differences in the ability of diverse calibration schemes to capture high- and low-flow conditions. The median NSE reached only 0.4–0.5 in Experiments 1 and 2, and although the LNSE approached 0.7, negative values are frequently observed. It is suggested that global optimization or simple weighted objective functions often lead to an averaging of catchment responses, thereby limiting accuracy for both high- and low-flow conditions. Experiment 3 employed an objective function defined as: $OF = 0.27 \cdot RMSE\_Q5 + 0.16 \cdot RMSE\_Q20 + 0.08 \cdot RMSE\_Qmid + 0.24 \cdot RMSE\_Q70 + 0.25 \cdot RMSE\_Q95$*, the weighting scheme explicitly accounted for extremely high (Q95), high (Q70), medium (Qmid), low (Q20), and extremely low (Q5) flows. Despite this design, both NSE and LNSE declined relative to Experiment 1. The decrease may be attributed to excessive parameter adjustments aimed at fitting a limited number of extreme events, which reduced the predictive accuracy of the overall streamflow process. When single dynamic parameters are introduced in Experiment 4, median NSE and LNSE increased*

to approximately 0.55 and 0.8, respectively, with narrower interquartile ranges. These outcomes indicate that dynamic parameters enhanced the ability of the hydrological model to capture temporal variability, although structural errors persisted, as reflected in local outliers. More significant improvements emerged with multiple dynamic parameters. Experiment 5 achieved median NSE and LNSE values of approximately 0.7–0.8 in both calibration and evaluation periods. Although high-dimensional optimization increased computational demand and LNSE variability in some basins, overall performance represented a balanced trade-off between dynamic adaptability and physical consistency. Experiment 6 also performed well during the calibration period; however, its abrupt parameter switching led to a particular decline of LNSE and increased dispersion in the evaluation period. Experiment 7 addressed these shortcomings by applying a gradual parameter-switching strategy during the evaluation period. As shown in Fig. 5, the boxplots are more compact and shifted toward higher values, indicating that stable and consistent performance was achieved across most basins. However, compared with Experiment 5, Experiment 7 displayed a greater number of outliers, particularly in LNSE, where they tended to cluster at lower values, suggesting higher variability in model performance across catchments. The overall accuracy remained comparable to that of Experiment 5. In summary, compared with static calibration schemes (Experiments 1–3), single dynamic parameter calibration (Experiment 4) improved simulative accuracy, while multi dynamic parameter calibration produced further gains. Among all experiments, Experiments 5 and 7 demonstrated the most robust and accurate performance."

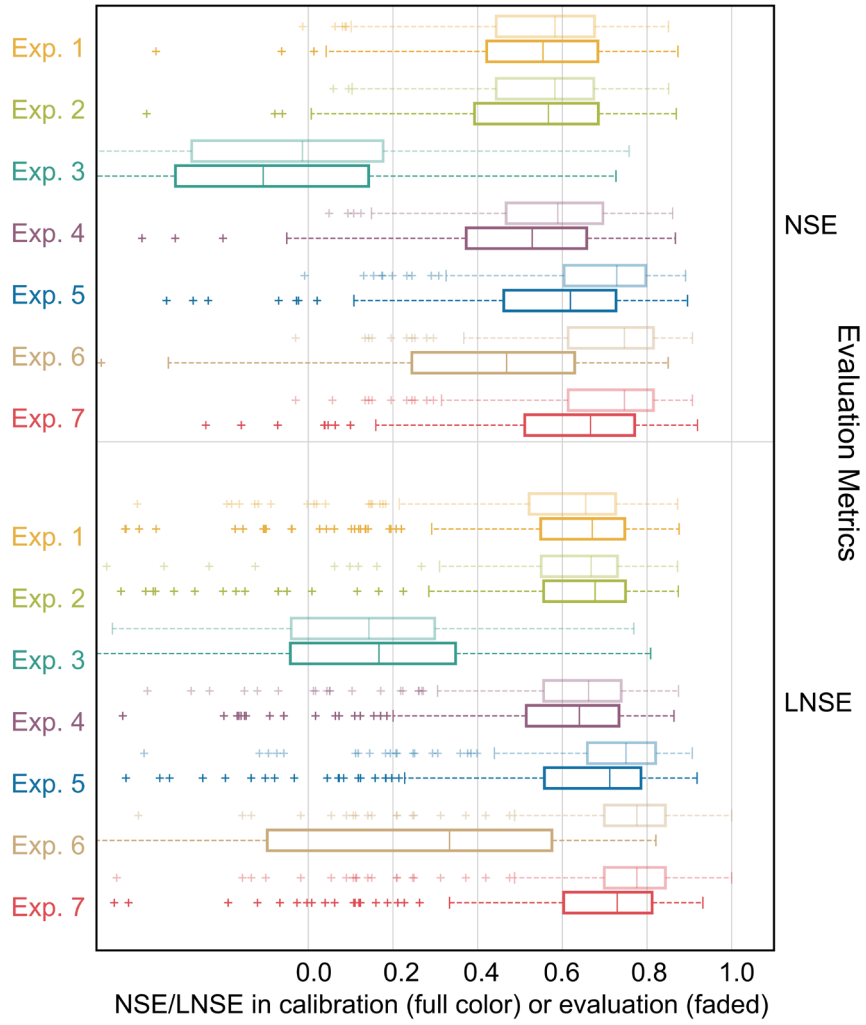

Figure 5. *Performance of seven calibration experiments on the MOPEX dataset across 219 catchments. Boxplot color denotes different experiments. The whiskers extend a maximum of 1.5 times the interquartile range. Values beyond the whiskers are marked as outliers and are denoted as +.*

Q9: The clustering results should capture more about temporal sequencing. I suggest presenting catchment-wide statistics such as distributions of the number of clusters, sub-period lengths, and other relevant metrics. These would clarify the added complexity introduced by sub-period calibration and should appear in the main paper.

Response:

Thank you very much for the constructive suggestion. We completely agree that providing a statistical description of the clustering results is crucial for helping readers

understand the overall characteristics of the sub-period clustering and its impact on calibration complexity. In the revised manuscript, the beginning of the Results section (Section 4.1) presents statistical information across all catchments. The distribution of the number of clusters for all catchments (averaging about 4.2 sub-periods identified), the distribution of the optimal sampling window lengths for feature extraction, and the average variance explained by principal component analysis (PCA) are reported. These statistics illustrate the typical characteristics of the sub-period clustering method across diverse catchments and quantify the complexity introduced by the procedure. The specific revisions in Section 4.1 of the revised manuscript are detailed in the response to Q4 (see "4.1 Defined sub-periods based on catchment dynamics" in the "Revised manuscript text").

References

Bouaziz, L. J., Aalbers, E. E., Weerts, A. H., Hegnauer, M., Buiteveld, H., Lammersen, R., ... & Hrachowitz, M.: Ecosystem adaptation to climate change: the sensitivity of hydrological predictions to time-dynamic model parameters. Hydrology and Earth System Sciences, 26(5), 1295-1318, https://doi.org/10.5194/hess-26-1295-2022, 2022.

Duan, Q. Y., Gupta, V. K., and Sorooshian, S.: Shuffled Complex Evolution Approach for Effective and Efficient Global Minimization, Journal of Optimization Theory and Applications, 76, 501-521, https://doi.org/10.1007/BF00939380, 1993.

Hsueh, H. F., Guthke, A., Wöhling, T., & Nowak, W.: Optimized predictive coverage by averaging time-windowed Bayesian distributions. Water Resources Research, 60(5), e2022WR033280, https://doi.org/10.1029/2022WR033280, 2024.

Zhang, X. and Liu, P.: A time-varying parameter estimation approach using split-sample calibration based on dynamic programming, Hydrology and Earth System Sciences, 25, 711-733, https://doi.org/10.5194/hess-25-711-2021, 2021.