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

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 comments have been invaluable in enhancing the quality of our paper, and for that, we are deeply grateful. In response, we have undertaken a comprehensive and thorough major revision based on your valuable recommendations.

We have reorganized the entire structure of the manuscript, particularly redesigning the “Methods” section into four logical, progressive levels to ensure a clear and coherent narrative flow from the model introduction, sub-period division, and experimental design to the evaluation system. To address your concern about the imbalanced presentation of results, we have incorporated a large-sample statistical analysis for all dynamic catchments in the revised manuscript. This crucial addition not only robustly demonstrates the generalizability and stability of our recommended calibration schemes but also showcases the broad applicability and value of our framework across diverse hydro-climatic conditions. Furthermore, we have carefully reviewed and revised the entire manuscript to ensure key terms are clearly defined, the language is concise, and that core methodological details from the supplementary materials are integrated into the main text to enhance its self-containment and readability.

We hope these systematic revisions could significantly improve both the scientific rigor and the clarity of the manuscript. Your expert guidance has not only helped us rectify the deficiencies of the initial draft but has also pushed us to think more deeply about and better articulate the core contributions of our research. Thank you once again for your time and expertise.

## **Specific Comments:**

### 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, we have carefully reviewed and unified the use of terms throughout the manuscript to ensure conceptual consistency and terminological precision.

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 substantial intra-annual and/or inter-annual variability, making their simulation particularly challenging for models. 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. Finally, dynamic parameter is a concept refers to model parameters that are allowed to vary across sub-periods, rather than remaining fixed over the entire simulation period.

We are grateful for the reviewer’s suggestion, which has led to a clearer presentation of our conceptual framework and improved the overall quality of the manuscript.

**Revised manuscript text:**

*Introduction*

*“A dynamic catchment is defined as one in which hydrological processes exhibit pronounced intra-annual or inter-annual variability, making their simulation particularly challenging.”*

*“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.”*

*“Sub-periods are segments of the simulation period characterized by relatively homogeneous hydrological conditions, which are typically identified through clustering of the time series.”*

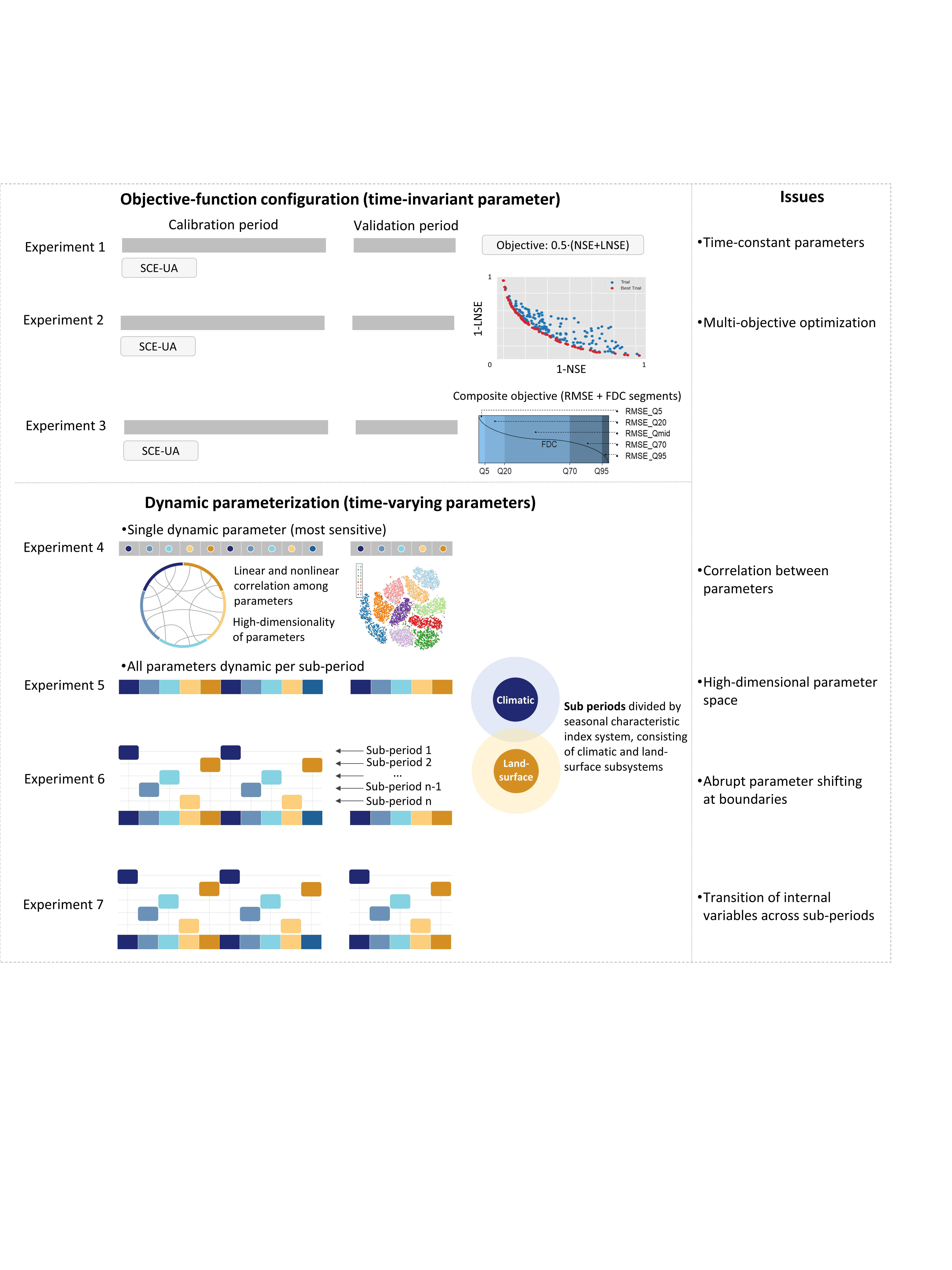
*“A dynamic parameter is defined as a model parameter that varies across sub-periods rather than remaining fixed over the entire simulation period.”*

### 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, we have rigorously reviewed and revised all figures and their captions to ensure that each caption contains sufficient information, clearly explains the specific meaning of each element in the figure, conveys the core scientific message, and provides explanations for all symbols and abbreviations. For example, we have redrawn Figure 2 from the original manuscript (now Figure 3) and used a more detailed caption to explain the core differences among the seven experiments in terms of objective functions, parameter settings, and sub-period handling. We believe these changes will significantly improve the readability of the figures, allowing readers to understand our research process and results more intuitively and accurately. Thank you again for your valuable suggestion.

**Revised manuscript text:**



*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 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 validation 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 validation period, the simulated flow data from each separate sub-period are combined and compared with the observed flow.*

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

**Response:**

Thank you very much for your question. Clarifying the basis for selecting the case study catchments is indeed crucial for highlighting the generalisability of the framework. In the revised manuscript, we have expanded the number of case studies from four to five and have clearly stated the selection criteria and rationale in Section 2, “Study area.” Our main objective was to ensure that the case studies are sufficiently diverse and representative to comprehensively test the robustness and applicability of the proposed framework under different natural conditions. Specifically, these five catchments cover different climate zones, topographic conditions, and hydrological characteristics, ranging from humid to semi-arid and from plains to mountainous regions. Their detailed information is listed in Table 1. Through this arrangement, we hope to more strongly support the applicability and reliability of the framework in different scenarios. We sincerely thank you for your detailed review, which has helped us to further improve the design and description of our case studies.

**Revised manuscript text:**

*Section 2*

*“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 analysed 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.”*

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ID | 12027500 | 6192500 | 7211500 | 1643000 | 1531000 |
| Location | 122.99°W  46.63°N | 110.40°W  44.84°N | 104.76°W  36.65°N | 77.25°W  39.63°N | 77.24°W  42.10°N |
| 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 (m) | 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,  Pasture/Hay | Evergreen Forest,  Shrub/Scrub | Evergreen Forest,  Grassland/Herbaceous | Deciduous Forest,  Cultivated Crops | Deciduous Forest,  Pasture/Hay |

### 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. The EDCC method is not the core innovation of this paper, but is a result of our previous research. In this study, it serves mainly as a preprocessing step for dividing the “sub-periods,” while the core work lies in comparing and conducting different parameter calibration experiments. Nevertheless, we completely agree with your suggestion to improve readability. In the revised manuscript, we have integrated and summarized the key implementation steps for sub-period division—which were previously in the supplementary materials (including the construction of the indicator system, feature extraction, dimensionality reduction, and clustering methods)—into Section 3.2, “Partitioning hydrological processes,” of the Methods, so that readers can directly understand its operational logic within the main text. A more detailed technical workflow is retained in the Supplementary Information (SI) for interested readers. At the same time, in the Results section, we further conduct the effectiveness of the sub-period calibration experiments based on the sub-period division across all catchments. We sincerely thank you for your suggestion; this modification has significantly improved the structural integrity of the paper and has also made the line of argument clearer and more rigorous.

**Revised manuscript text:**

*Section 3.2*

*“****Describing catchment dynamics****: To characterize the temporal dynamics of catchment behavior, 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 several 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, where the optimal window length is determined via a time-windowed Bayesian inference framework based on predictive log-score (PLS) performance (Hsueh et al., 2024). This method is designed to preserve long-term trend signals, suppress short-term high-frequency noise, and enhance 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. First, a threshold-based screening is applied to identify variables with significant seasonality, thereby retaining only the relevant subsystems and forming an initial pool of candidate indicators (see Supporting Information S2.1 for detailed criteria). Subsequently, the Maximal Information Coefficient (MIC) is used to quantify both linear and nonlinear associations between candidate indicators and streamflow, ensuring that selected indicators are hydrologically relevant. To further address multicollinearity and reduce dimensionality, Principal Component Analysis (PCA) is conducted, and the first two principal components are retained for clustering. This multi-step filtering and reduction process ensures robust extraction of dynamic catchment characteristics and establishes a solid foundation for sub-period clustering based on 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 conducted in the principal component space, allowing structural patterns in the catchment dynamics to be captured effectively. This data-driven clustering approach reveals the temporal heterogeneity of hydrological processes and provides a robust basis for integrating dynamic parameters into hydrological models.”*

### Is the EDCC applied separately for each basin?

**Response:**

Thank you very much for raising this key methodological question. Your understanding is completely correct; 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 dividing sub-periods for each catchment individually can we more accurately capture its unique dynamic processes. This “tailored” approach is a prerequisite for ensuring that the subsequent dynamic parameter calibration can effectively reflect the true hydrological processes of a specific catchment. In the revised manuscript, we have clarified this point in Section 3.2 of the Methods to avoid any potential misunderstanding. We sincerely thank you for your rigorous review; this modification has helped us to further improve the description of our methods.

**Revised manuscript text:**

*Section 3.2*

*“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 via screening and dimensionality reduction; and (3) applying unsupervised clustering to cluster the time series into hydrologically coherent sub-periods for further model integration.”*

*“All indicators are sampled using a moving window approach, where the optimal window length is determined via a time-windowed Bayesian inference framework based on predictive log-score (PLS) performance (Hsueh et al., 2024). This method is designed to preserve long-term trend signals, suppress short-term high-frequency noise, and enhance the stability and robustness of dynamic catchment characteristic extraction.”*

### 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. In designing the sub-period division process, we adopted an unsupervised, data-driven approach to minimize manual intervention as much as possible. In the revised manuscript, we have provided a clearer explanation of this in Section 3.2 of the Methods. Specifically, the determination of the number of clusters does not rely on manual specification but is based on a comprehensive evaluation of multiple objective clustering validity indices (including SC, S, and XB indices), supplemented by the elbow method as a diagnostic tool to provide a quantitative reference for selecting the optimal number of clusters. In short, the core steps of the sub-period division are all based on data, and the decision-making process has been openly described in the paper, thereby ensuring the transparency and reproducibility of the method. We sincerely thank you for your suggestion; this modification has helped us to further enhance the rigor of our method description.

**Revised manuscript text:**

*Section 3.2*

*“****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 conducted in the principal component space, allowing structural patterns in the catchment dynamics to be captured effectively. This data-driven clustering approach reveals the temporal heterogeneity of hydrological processes and provides a robust basis for integrating dynamic parameters into hydrological models.”*

### 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; this indeed touches upon the core of testing the framework’s generalization capability 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 division model 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. Instead, we input the data from the evaluation period into the partitioning model established during the calibration period. This classifies each day of the evaluation period into its corresponding sub-period, and the parameter set corresponding to that sub-period is then used. 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. We sincerely thank you for your insightful suggestion; this modification has helped us to articulate the design logic of the framework more clearly.

**Revised manuscript text:**

*Section 3.2*

*“In this study, 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 validation period.”*

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

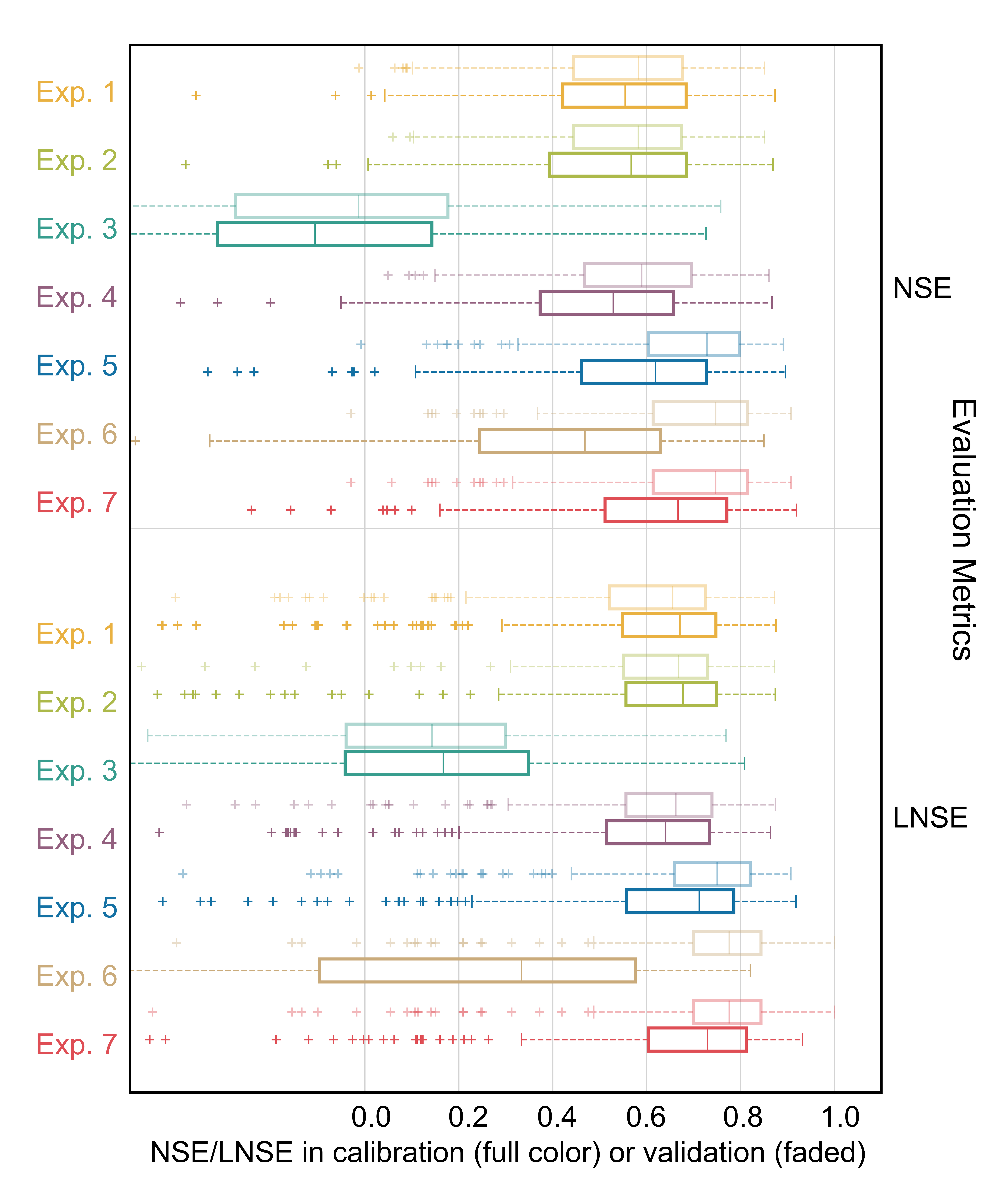
**Response:**

Thank you very much for your suggestion. Due to limitations in length and presentation in the initial draft, we were not able to systematically present the large-sample statistical results in the main text. In this revision, we have addressed the relevant data processing and visualization issues and have made a significant addition to the Results section. In Section 4.2, we use box plots to visually display the distribution of performance metrics (NSE and LNSE) for 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, allowing the study to both provide mechanistic explanations for the framework in individual catchments and demonstrate its robustness at the large-sample level. We sincerely thank you for your suggestion; this modification has significantly improved the completeness and persuasiveness of the results presentation.

**Revised manuscript text:**

*Section 4.2*

*“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 validation periods reveal differences in the ability of different calibration schemes to capture high- and low-flow conditions.”*



*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 +.*

### 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 this highly 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 division and its impact on calibration complexity. At the beginning of the Results section in the revised manuscript, we have added statistical information about the clustering at the scale of all catchments. Specifically, we present 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 used for feature extraction, and the average variance explanation rate of the principal component analysis (PCA), among other results. These statistics intuitively reflect the typical characteristics of the sub-period division method when applied to different catchments and quantify the degree of complexity it introduces. We sincerely thank you for your valuable suggestion; this modification has significantly improved the completeness and depth of the results presentation.

**Revised manuscript text:**

*Section 4.1*

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