Response to Reviewer #1

The study quantified the effect of parametric uncertainty on the resulting runoff partitioning using 50 behavioral datasets of the Tsinghua Hydrological model in more than 60 catchments in the Yangtze River Basin. It shows that in catchments with larger runoff ratio, the parametric uncertainty is higher resulting in higher variability of simulated runoff components.

Although I find the analysis of relationship between parametric uncertainty and their possible catchments attributes interesting in general, the investigation performed in this study is rather incomplete and poorly motivated. The model appears to be unvalidated. Moreover, a rather small sample of parameters were considered for the analysis, compromising the reliability of the results. No rationale for selecting possible catchment controls is provided, while the results are based on simulation of a single model and single objective function, making the generalizability rather difficult. Please find my detailed comments below.

Response:

Thank you very much for your constructive comments. The points you raise indeed represent important limitations of our study. Some of these can be addressed by conducting supplementary experiments or clarified in the text, while others may be difficult to address adequately within the current scope of this work. Please find our detailed responses to the general and minor comments below.

General comments

1. Motivation: The motivation of the study as mentioned in Line 68-70 is to provide an a priori guidance on whether to only select single best calibration parameter or instead choose multiple behavioral parameter sets. I rather disagree with such premise. Even if some studies still use a single best parameter set, they are simply ignoring parametric uncertainty altogether, while there is a large body of studies including those referenced by the authors that show that using multiple parameter sets is essential for accounting for and communicating uncertainty. I do find the idea of knowing a priori how much parametric uncertainty one might expect useful for the sake of understanding and improving model limitations and to lower the computational costs of running a large number of simulations.

However, especially given rather inconclusive results of this study, using single parameter set would in any case mean ignoring the uncertainty. See also my detailed comments.

Response:

Thank you for the comment and we agree that our original premise regarding "guidance" for calibration strategy was indeed somehow overstated. This study mainly focuses on understanding the influence factors of the uncertainty in runoff component partitioning simulations, which is also valuable for understanding hydrological model behavior and limitations. Given the current inconclusive results, it is difficult to provide a “guidance” for the choice of calibration strategies. We will rephrase the motivation of the study in the revised manuscript.

2. Lack of Validation: It seems that there is no validation of model performance provided in this study. The method section describes only the calibration approach. Moreover, it seems that the total length of observations used for calibrations is about 2 years, which is too short to result in reliable parameter identification. Without the validation, there is no proof of parameter validity for an uncalibrated period.

Response:

Thank you for the comment. We acknowledge the shortage of data length and the lack of validation. However, we need to clarify that publicly available streamflow records are limited for Chinese basins, and access often depends on collaboration with data-providing agencies (Chen et al., 2025, https://doi.org/10.1038/s41561-025-01833-x; Lin et al., 2023, https://doi.org/10.1038/s44221-023-00039-y). The dataset used in this study covering 63 hydrological stations was obtained through such collaborations and represents a comparatively rare multi-basin observational resource. Even though it covers only two years, obtaining concurrent records from 63 stations was challenging.

Given the data scarcity, we utilized the full 2-year period for calibration to take full advantage of the data. The model performance is expected to be rather stable due to the short period, i.e., the model is unlikely to perform well in one year but extremely poorly in another year because of the similar conditions. However, to address the concern about model stability and validity, we will conduct a supplementary experiment. In specify, we will select representative catchments and split the data into a calibration period (Year 1) and a validation period (Year 2), to analyze the correlation between the KGE performance in the calibration and validation periods. We will include these results in the Supplementary Material and discuss them in the Methods section to justify the robustness of our approach despite the data limitations.

3. Parameter sampling: Given that the focus of this study is to investigate parameter uncertainty, 50 parameter sets seem like a very low number of samples. A much more comprehensive sampling is needed to prove that lower number of behavioral sets is indeed not simply an artefact of small sampling.

Response:

Thank you for the comment. We apologize for the lack of clarity regarding the parameter sampling strategy. The number 50 is the final selected behavioral parameter sets, not the total number of samples evaluated. Our calibration utilizes the pySOT algorithm, which involves a two-layer loop structure: (1) Inner loop: In each pySOT run, the model is evaluated 3,000 times to find a local optimum. (2) Outer loop: We repeated the pySOT run 50 times to avoid being trapped in local optima. Therefore, the total number of parameter evaluations is actually 150,000 (3,000*50). The 50 sets represent the local optimal solutions derived from this extensive sampling. We will clarify the sampling strategy in the Methods section to demonstrate that the parameter space has been adequately explored.

4. Single calibration strategy and single model: While the study aims to provide generalizable results on how to a priori decide calibration strategy, it only examines one single objective function for the calibration (i.e., KGE) and one single conceptual model. In the Limitation section the authors highlight the latter as the limitation themselves. As hydrological models of even rather similar structures are known to behave quite differently in how they partition the fluxes (Merz et al., 2022 https://doi.org/10.1175/BAMS-D-21-0284.1; van Kempen et al., 2020 https://doi.org/10.5194/nhess-21-961-2021), several model structures must be analyzed to confirm reliability of the suggestions presented here. Similarly, given high effect of the objective function on the resulting parameter sets, a comprehensive set of experiments with various objective functions is needed to evaluate the generalizability of

the findings.

Response:

Thank you for your comment. We agree that selection of models and objective functions influence the results. However, given the vast diversity of hydrological models and the substantial workload involved in establishing and calibrating models for 63 catchments, it is challenging to implement a comprehensive multi-model comparison within the scope of the current study. Similarly, regarding the objective function, we selected KGE as a representative metric because it is a comprehensive metric that balances bias, variability, and correlation. In the revised manuscript, we will explicitly acknowledge this as a limitation, and that future work should extend this analysis to other model structures and objective functions to test generalizability.

5. Potential controls of uncertainty: The manuscript does not provide the rationale for selecting a few examined catchment attributes as potential controls of parameter uncertainty. This is necessary to understand why the examined attributes were selected in a first place and if there any other potential properties that can be used to explain observed parameter uncertainty.

Response:

Thank you for the comment. At this point, there was no formal selection rationale applied in this exploratory study. We selected a small set of basic, commonly used, physically meaningful catchment attributes that either influence the rainfall-runoff processes (e.g., climate and topography conditions) or reflect the fundamental characteristics of these processes (e.g., runoff volume and runoff ratio). We will add a short note in the Methods to make this explicit and include a brief limitation in the Discussion clarifying that other properties may also affect parameter uncertainty and remain for future work.

Specific comments

Line 18, 111 and elsewhere: It is not quite clear what is meant by “high-quality rainfall” here? It is also was not clarified later in the Methods part. It would be much more instructive to specify temporal resolution, length of observations or density of the observations to highlight a particular aspect of data quality. Please revise.

Response:

The term “high-quality” mainly referred to the high temporal resolution (< 1h) and integrity (low missing rate) of the data. We will remove the vague term and instead explicitly describe the data attributes in the Methods section.

Line 23: Given the theme of this paper, I do not think that the model performance is relevant for the abstract. Consider omitting this.

Response:

We agree that the model performance is not so relevant for the abstract, but we believe that good model performance is a necessary basic for the subsequent uncertainty analysis. Consequently, we will simplify this part in the abstract to only brief state that the model performs adequately well.

Line 25-26: At this point, this statement is not clear, because it is not yet clarified how the uncertainty is computed. Please clarify and revise.

Response:

We will replace the vague phrasing with more precise terms such as “range” to clearly define how the uncertainty was quantified.

Line 34-36: This recommendation is rather general. It would be helpful to specifically highlight what sort of guidance this study can provide for the choice of the calibration strategy.

Response:

The specific recommendation is actually the third point of result about the relation between model uncertainty and catchment attributes. However, as pointed out by the reviewer and acknowledge in the response to the general comments, the current results are not conclusive enough for providing an operational guidance. Consequently, we will remove the statement regarding guidance from the Abstract and focus on the results.

Line 49-51: This recent review of Wagner et al., 2025 (https://doi.org/10.1002/wat2.70018) might be worth mentioning here.

Response:

We will include this reference in the introduction as suggested.

Line 53: I disagree that these are two different calibration approaches. Considering multiple behavioral parameter sets vs taking one best single parameter set stands for "accounting for parametric uncertainty" vs "ignoring it". While most of the current modeling studies account for parametric uncertainty using various approaches, using a single best parameter set remains a poor practice that unfortunately still persists in some studies. Yet, this does not make it a distinct calibration strategy. Please revise.

Response:

We agree that adopting multiple behavioral parameter sets and one best single parameter set are actually not two different calibration approaches, but two different treatments to model parameter uncertainty. We will revise the statement in the introduction. However, we think that the necessity of a detailed uncertainty analysis often depends on the specific research objectives and questions. While parameter uncertainty is undoubtedly important, it may not always be the primary focus of every hydrological study, and we cannot expect every single study to address all aspects of modeling comprehensively. Even in our current work, which prioritizes parameter uncertainty, we specifically focus on the uncertainty of runoff partitioning. Therefore, we will modify the text to reflect that using a single parameter set is a simplification often driven by specific research goals, rather than presenting it as a competing strategy to uncertainty analysis.

Line 57-59: Given the relevance of the behavioral parameter sampling, this part seems rather short to me and incomplete, especially in terms of the references mentioned. It seems that most of seminal works of Keith Beven on this topic are overlooked here. Please add.

Response:

We will add the relevant citations by Keith Beven and others.

Line 63-64: As I mentioned in my comment above, even if some studies still use a single parameter set, there is a consensus that in order to represent parametric uncertainty multiple parameter sets should be considered as the study referenced here also emphasizes.

Response:

We will adjust the statement.

Line 99 and elsewhere: The term “average” is ambiguous. Please specify if this is mean or median.

Response:

It refers to the arithmetic mean. We will specify "mean" throughout the manuscript.

Line 101: Please clarify here what is meant here by runoff ratio and how it was computed in this study.

Response:

The runoff ratio is calculated as the ratio of total discharge to total precipitation. We will clarify this definition in the revised Methods section.

Figure 1: This figure would be more instructive if it would also display precipitation or runoff ratios that are used as explanatory controls of the uncertainty in this study.

Response:

We will display these basin attributes in Figure 1 in the revised manuscript.

Line 112-113: This is a rather misleading way report missing values. Please specify the study period and the portion of missing data across all catchments.

Response:

The study period for all the 63 basins is consistently 2014.1.1-2015.12.31. We will report the portion of missing data in the revised manuscript.

Line 115: Please quantify this by specifying how much water level data vs discharge data is available.

Response:

We will add the specific number of water level and discharge data in the revised manuscript.

Line 116-119: Please specify how these relationships were built. Was it a linear relationship?

Response:

Yes. The linear relationship was used. We will clarify this.

Line 128-132: It is not clear how this information was used. Is it needed for model input? This has to be clarified.

Response:

Yes, these datasets were used as model inputs. We will add few sentences to clarify this in the manuscript.

Line 149: Even if the model was used in the previous studies key equation and especially parameters have to be specified. Parameters can easily be added near the corresponding compartment in Figure 2. Please add.

Response:

We will update Figure 2 to illustrate each parameter controls which process. However, as the THREW model is a complex distributed model containing several subregions, it is difficult to say which equations are the most important. Consequently, we will provide brief description in the manuscript and cite the papers about the model development for the detailed mathematical formulations.

Section 2.3: This section must provide a complete information on the length of calibration period and validation period.

Response:

As clarified in the response to general comment, we utilize all the observation streamflow data to calibrate the model considering the short length of data series.

Table 2: Are these all the parameters of the model? Please clarify.

Response:

These are the parameter determined by calibration. There are also some parameters determined by related dataset (e.g. soil property parameters) or fixed due to small influence on model performance. We will clarify this in the table caption and main text.

Line 172: It is not clear what is meant by the "optimal KGE", the highest one?

Response:

Yes, it means the highest KGE. We will rephrase for clarity.

Line 179: Even if the river bed might be impervious, this is not a primary reason for surface runoff. For perennial rivers as here, rainfall falling on the channel surface is much more likely to flow horizontally along the channel than vertically towards river bed. Please revise.

Response:

Yes, the original expression was inaccurate. We intended to refer to the water surface of the river channel. We will delete "impermeable areas" and directly specify "river channel water surface" avoid confusion.

Section 3.1: It is not clear of the reported performance corresponds to the calibration or to the validation method. Please clarify.

Response:

As clarified in the response to general comment, we utilize all the observation streamflow data to calibrate the model considering the short length of data series. We will clarify it here.

Figure 8: Please add explanation of clusters in the caption. Please also clarify what is meant here by maximum rainfall. Event? Rate? Volume?

Response:

It is mean annual rainfall as illustrated in Table 1. We will also add explanation of these terms in the caption.

Line 206: Please clarify what makes these two catchments typical and why for another example later another set of two catchments were selected. Please avoid subjective choice of catchments for examples and present the results for all study catchments.

Response:

Visualizing detailed time-series for all 63 catchments in the main text is not feasible, so we only presented the results for some catchments. Different typical catchments were selected for Figure 4 and Figure 5 to represent different aspects of behaviors. Figure 4 shows the model performance, so the catchments with highest and lowest KGE were selected as typical catchments. Figure 5 illustrates the relation between KGE and the contribution of subsurface runoff (C_{sub}), so the catchments with high and low sensitivity of KGE to C_{sub} were selected. We will clarify this selection logic in the text. Additionally, although presenting results for all catchment in main text is not feasible, we will provide the data producing Figure 4 and Figure 5 for all catchments in the Supplementary Information to ensure completeness.

Figure 3: This figure could be more efficiently presented as a boxplot or a violin plot.

Response:

We will redraw the Figure 3 accordingly.

Line 216 and elsewhere: Please avoid term “significantly” if no statistical test was used.

Response:

We will check through the manuscript and remove all such terms.

Line 260: Please be more specific on which catchments are these.

Response:

We will provide the information of the specific catchment.

Figure 6 and all other figures: Please explain all the terms and acronyms from the figure in the caption.

Response:

We will explain the terms and acronyms in the caption of all figures.

Table 3 and elsewhere: Please specify type of correlation used here.

Response:

The “correlation coefficient” reported here and elsewhere is the Pearson correlation coefficient (r), which measures the linear correlation between the two variables. We will clarify it in the revised manuscript.

Line 306: I am not sure if the term “model sensitivity” is suitable here. Given that “parameter sensitivity” is a rather established term, I would avoid using it in a different context.

Response:

We will replace “model sensitivity” with a more accurate term to avoid confusion with the established term.

Line 338: It is not clear what is meant here by the assimilation of the datasets. Please clarify.

Response:

It means use multiple datasets to constrain uncertainty and improve model performance. As we are not referring to a specific practice, so a rather general term is used here.

Line 367-371: Please specify if these studies were in the same study region/ catchments.

Response:

Those studies were not in the same region with this manuscript. We will clarify it in the revise manuscript.