Dear Editor,

Thank you for considering our publication after major revision and for providing additional feedback. We appreciate your thorough review and the opportunity to improve our work further. Below, we provide a point-by-point response to your comments.

Figure 1 caption: consider to rephrase ", in green the model experiment inputs are shown, in red the models, and in grey the analysis components." to " The model experiment inputs are shown in green, the models in red, and the analysis components in grey."
Change capital A, B, and C letters in the flowchart to (a), (b), and (c) to be consistent.

We appreciate your suggestion to rephrase the caption. We have revised the caption to: "The model experiment inputs are shown in green, the models in red, and the analysis components in grey." We have also updated the labels in the flowchart, changing the capital A, B, and C letters to A, b, and c to maintain consistency with the labeling convention used elsewhere in the publication. In addition, we have adjusted the font to improve text clarity.

Lines 251-256 and Figure 2: Lines 251-256 need more clarity to what exactly you are presenting in Figure 2 and how you derived the so-called uncertainties. For example, you state "The discharge observation uncertainty estimates of the CAMELS-GB dataset are processed by averaging the upper and lower bounds of uncertainty estimates per flow percentile (5, 25, 75, 95)." Did you average all the four flow percentiles? How did you get +20% and -15%?

Thank you for pointing out the need for more clarity in explaining Figure 2 and the derivation of the uncertainty estimates. We have revised the text to provide a more detailed and precise explanation of the process. Specifically, we clarified that:

- *The absolute error, model difference, between calibrated model simulations and observed discharges for each flow category and catchment is calculated first, which is now explicitly visualized in Figure 2a (blue line).*
- *The upper and lower uncertainty bounds were taken from the CAMELS-GB dataset. We revised the text to state how these bounds were derived clearly: for example, the upper uncertainty percentages at flow category boundaries correspond to 25% and 15%, respectively, which we averaged to obtain an upper uncertainty bound of 20% for the low flow category. This is now clearly shown by the orange and red lines in Figure 2b.*
- *We also clarified the derivation of the overall observation uncertainty percentage, which is now explained as the average of the absolute values of the upper and lower bounds, resulting in a 17.5% uncertainty.*

The legend used in (b) and (c) can be confusing as they use the same green line and the same name to refer to two different things. Please consider updating one of them so one can differentiate the two lines that are currently both shown in green as 'Observation Uncertainty (17.5%). The way the name is given to a time series is already unclear. I guess the green line in (b) refers to 'the mean observation error' and in (c), 'the observed discharge with added uncertainties'.

We acknowledge your concern regarding the potential confusion caused by using the same green line and label in Figures 2b and c. To address this, we have made several revisions to improve the example. These include changing the legend colors. Updating the time series data to another catchment ensures that the presentation is clearer and more consistent. These changes better *illustrate the method. process. In addition, We have clearly stated in the text that observation uncertainty refers to " the portion of the observed discharge attributed to uncertainty"*

Figure 3b Here, the optimized wflow_sbm model performs better in 75% of the catchments than the PCR-GLOBWB model. Both models demonstrate poor results for approximately 25% of the evaluated catchments (<0.40 KGE-NP).
Can you please double check the number 25%? It does not look right to me.

We have adjusted the text to: "between approximately 18 % and 24 %".

Table 2
Why do you need the last column to show total instances of 299? It is just repeated three times in the table. You already stated 299 catchments in the table caption.

We removed the mention of the total instances in Table 2.

Line 356 The HBV-96 and XINANJIANG models that most closely resemble each other based on the number of stores, process descriptions, and parameters
This is a wrong statement. The HBV-96 and XINANJIANG models are very different in process descriptions (especially in runoff generation) and parameters, although they share some common features and they both show robust model performances.

We agree with your comment and added nuance to this statement by only referring to the number of model stores and the number of parameters. The text is revised: "The HBV-96 and XINANJIANG models most resemble each other regarding the number of stores and parameters; the parameters themselves and process descriptions differ. Both models contain a low number of instances, allowing the identification of the better-performing model.".