Response letter of 2025GL120393R

Dear Editor and Referees,

Please find the revised version of our manuscript “Technical Note: Benefits of Bayesian estimation of model parameters in a large hydrological model ensemble”, which we would like to resubmit for publication in *HESS*

Comments made by the reviewer were highly insightful. They allowed us to greatly improve the quality of the manuscript. Our responses to the comments are provided below.

Each comment made by the reviewers is written in *italic* font. We numbered each comment as (n.m) in which n is the reviewer number and m is the comment number. In the revised manuscript, changes are highlighted in yellow.

Sincerely,
Yohei Sawada, Shinichi Okugawa

Response to the comments from Referee #1.

(1.1) Lines 116-117: How was this model selection made? I understand the motivation to "preserve diversity in model complexity while minimizing the total computation cost" but what process was used to select, for example, model ID 07 instead of model ID 09 etc?

→ First, we decided not to use models which took more than 2-day for MCMC computation in our workstation. Many higher ID models were excluded in this process. Second, we excluded models which achieved lower performance in our previous work (Sawada et al. 2022). The models with IDs of 01, 14, 20, 22, and 23 were excluded in this process. Lastly, we chose the models shown in the manuscript to reduce the total number of models while maintaining the diversity of the models. For example, ID 09 is the family of ID 10, so that we used only ID 10. These points were unclear in the original version of the paper, and we have clarified it in the revised version of the paper:

> Lines 119-123: First, we decided not to use models whose MCMC computation takes longer than 2 days in our workstation. Many higher ID models were excluded in this decision. Second, we excluded models which achieved lower performance in our previous work (Sawada et al. 2022). The models with IDs 01, 14, 20, 22, and 23 were excluded. Lastly, we did not use multiple similar structure models.

(1.2) Line 192: How were the 51 basins selected out of the 135 in MERV-Jp?

→ Thank you very much for this comment. Although MERV-Jp originally includes 135 river discharge gauge observations corresponding to 135 river basins, some of the basins are the subset of the other larger basins. We did not choose overlapping basins. Also, we excluded river basins whose discharge record was shorter than 25 years. Consequently, we used 51 basins in MERV-Jp. This point was not explained in the original paper. We clarified it in the revised version of the paper:

> Lines 208-212: While MERV-Jp originally includes 135 gauges' observations corresponding to 135 river basins, some of them are the subset of the other larger basins. In this paper, we use one of those overlapping river basins. Also, we excluded river basins whose discharge records were shorter than 25 years. Consequently, the number of river basins in this study is smaller than that in the original MERV-Jp work (Sawada et al. 2022).

(1.3) Lines 211-213: Does this mean that for the 1-year calibration scenario and a typical station, the spin-up was performed from 1986 to 1990, calibration was then done for 1986, and then the timeseries was for evaluation was started from 1991? If so, from what set of initial conditions was the evaluation timeseries generated as it would have been the uncalibrated parameters that were used for the spin-up? Please clarify.

→ Given a parameter set, we run a model from 1986 to 1990 as a spin-up. Then, using state variables at the last timestep in the spin-up period, we run a model again in 1986 (we recognize that we potentially have a gap at the beginning of the calibration), which is evaluated against observation as a calibration. In an evaluation period, we simply ran a model with the estimated parameter set from 1986 to 2015 and discarded 1986-1990 results as a spin-up. Whenever we evaluate parameters, we run it from 1986 to 1990 as a spin up. In the revised version of the paper, we added some sentences to improve our description of spinning up:

> Lines 240-224: Given a parameter set, we run a model from 1986 to 1990 as a spin-up. Then, using state variables at the last timestep of the spin-up period, we run a model again in 1986 and evaluate the estimated river discharge. In the validation process, we simply run a model with the estimated parameter set from 1986 to 2015 and discard

results in 1986-1990 as a spin-up.

(1.4) Figures: Some figures are too small or cluttered to see the point the authors are making.

Figure 2: The subplots are quite small and it is difficult to see the authors main point (that the BPA surpasses the deterministically optimized models particularly for data-sparse scenarios) due to the clutter within the axes and the broad scales used on the y-axes.

Figure 3: It is not easy to see that more complex models benefit more from BPA partly because many model-basin pairs have near-white squares due to the colormaps. It would also be beneficial to add some indication of whether a model/basin benefits from BPA or not as currently the reader has to count the squares to find this information.

→ Due to a restriction in the submission process of EGUSphere, the resolution and size of the figure is not good in the document which referees received. We believe that the quality of the figure is not problematic when published. We put Figure 2 below for a reference, since we can control what to show in this pdf.

Following the instructions of the reviewer, we have changed the color scale in Figure 3. The new Figure 3 can be found below.

Also, we have counted the number of river basins in which BPA significantly outperforms the deterministically optimized model and shown it in Figure 4, which is attached below.

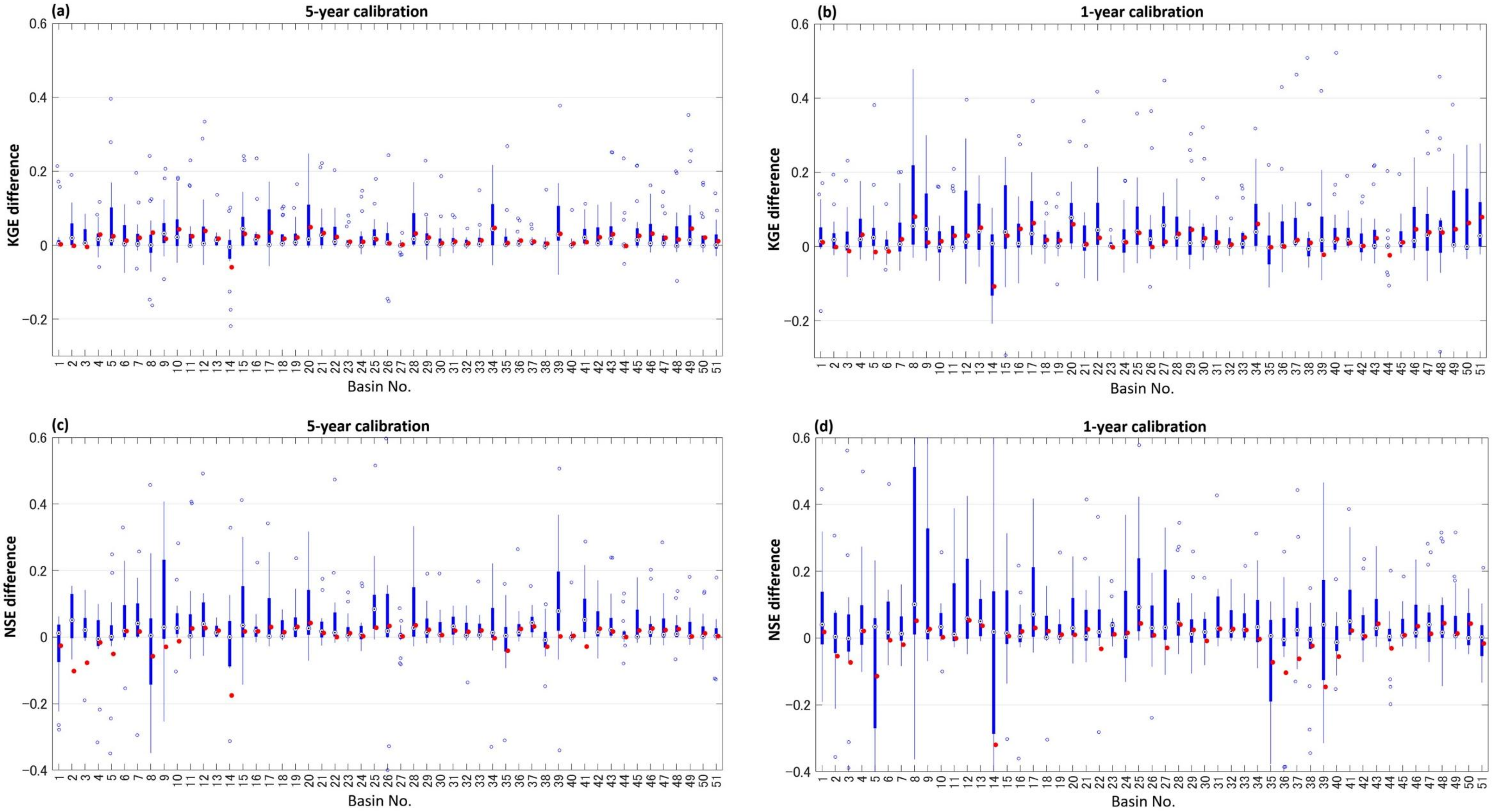

Figure 2. Differences in KGE (a, b) and NSE (c, d) between BPA and deterministically optimized models (boxplots) in 5-year (a, c) and 1-year (b, d) calibration scenarios. Red dots show the performance differences between BMPA and BMA (see Section 3).

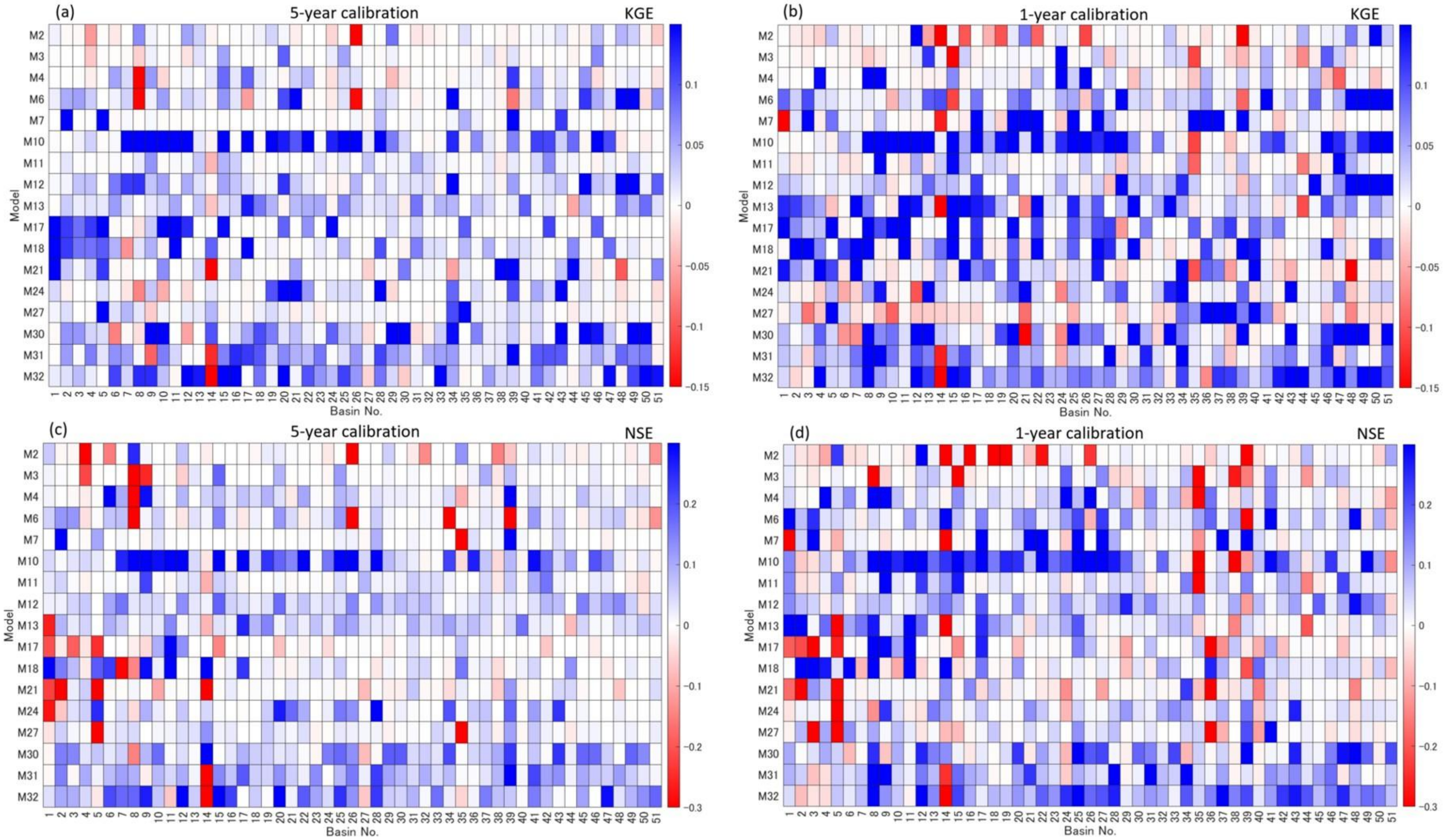

Figure 3. Differences in KGE (a, b) and NSE (c, d) between BPA and deterministically optimized models for each basin and model in 5-year (a, c) and 1-year (b, d) calibration scenarios.

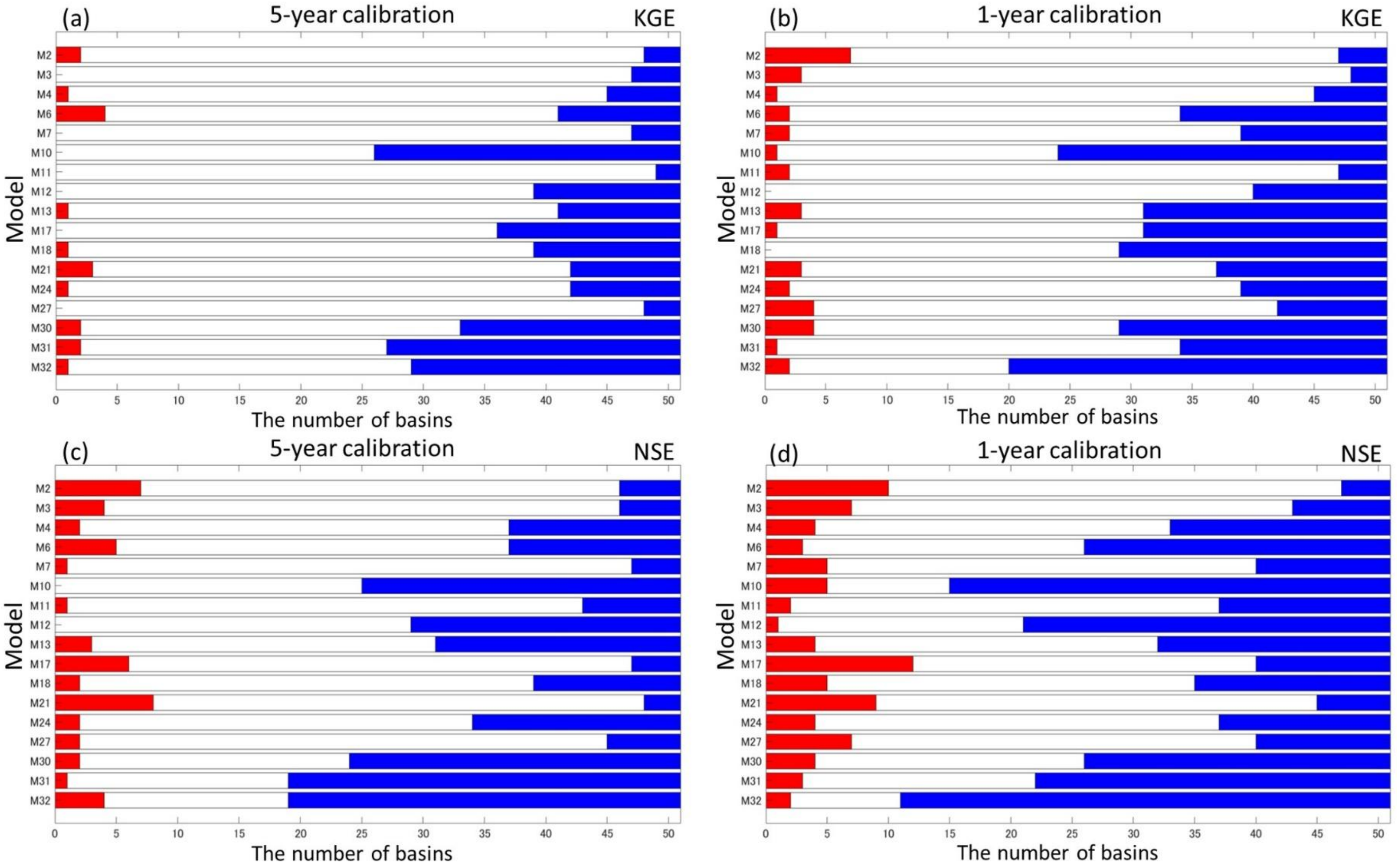

Figure 4. (a, b) The number of river basins in which (red) KGE of BPA is at least 0.05 lower than that of the deterministically optimized model, (blue) KGE of BPA is at least 0.05 higher than that of the deterministically optimized model, and (white) the difference in KGE between BPA and the deterministically optimized model is less than 0.05, in (a) 5-year and (b) 1-year calibration scenarios. (c, d) Same as (a, b) but for NSE.

(1.5) Lines 286-287: "BPA achieves 0.8 KGE for nearly all models, while some deterministically optimized models (i.e., ID 7, 24, 30, and 31) perform poorly." ID 31 deterministic performs as well as ID 21 BPA so this sentence may be misleading.

→ Thank you for the comment. We added ID 21 as a poorly performed model.

> Lines 316-318: Figure 5 illustrates a typical case: in basin no. 43, BPA achieves 0.8 KGE for nearly all models, while some deterministically optimized models (i.e., ID 7, 21, 24, 30, and 31) perform poorly.

Response to the comments from Referee #2.

(2.1) 1. Justify and test the posterior sample size (200 draws). You draw 200 parameter sets from the MCMC output to represent the posterior. Please justify this choice (e.g., computational budget vs diminishing returns) and add a brief sensitivity check (e.g., 100 vs 200 vs 500

draws for a subset of basins/models) to demonstrate stability.

→ Figure R1 shows the performance of BPA sampled from the different numbers (100, 200, and 500) of parameters in the calibration period. We could not find any significant differences and concluded that 200 parameter sets are already enough to represent the posterior. We have briefly mentioned this result in the revised version of the paper:

> Lines 184-185: The performance is insensitive to the number of the sampled parameters if it is larger than 100 (not shown).

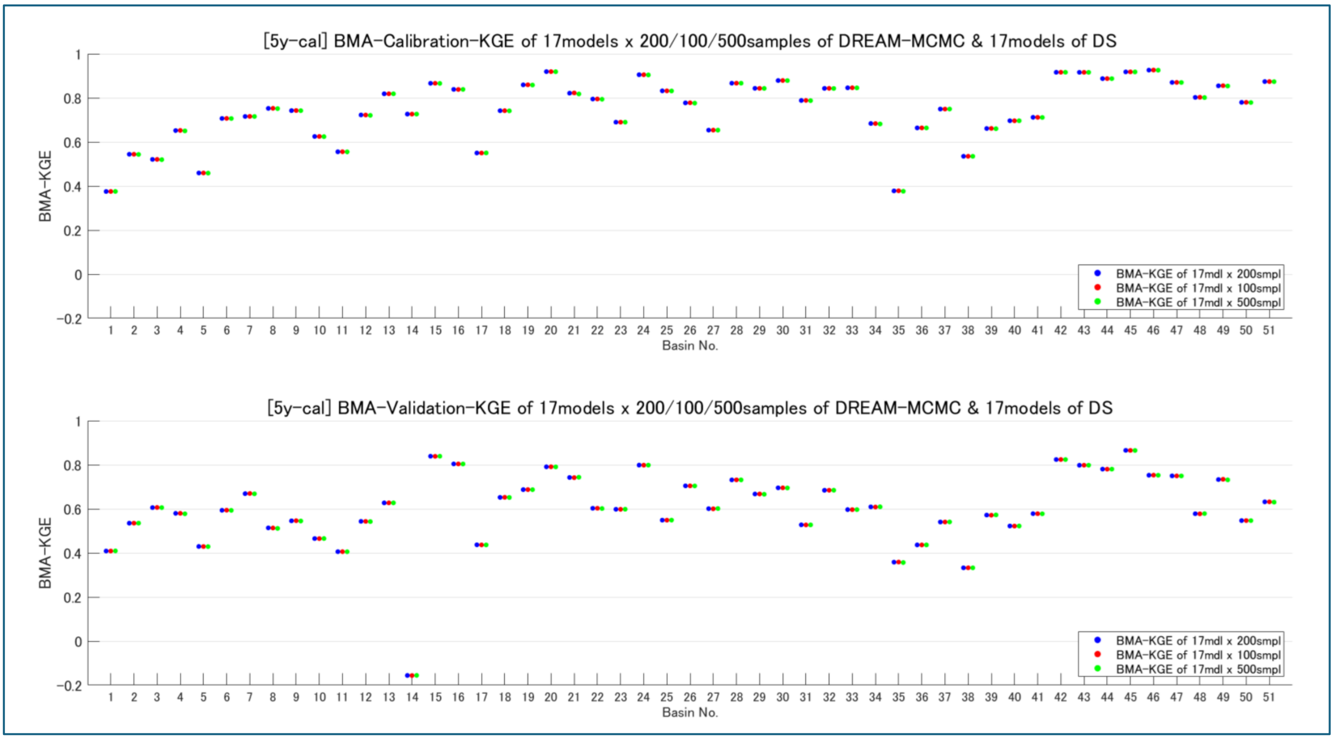

Figure R1. Performance of BPA with the different numbers of samples.

(2.2) 2. Clarify/justify the KGE-based "likelihood" used for DREAM. Since KGE is not a conventional likelihood, please briefly explain why this informal likelihood is appropriate for your goals and discuss known modes of failure.

→ We fully agree that KGE should be recognized as an informal likelihood in the context of MCMC. This point was admitted in the original version of the paper:

> While mean squared error is usually used as the formal likelihood function in the DREAM algorithm and many other MCMC applications, Liu et al. (2022) proposed using Kling-Gupta Efficiency (KGE: Gupta et al. 2009) as an informal likelihood function.

The issue of using KGE as an informal likelihood function has already been thoroughly discussed in Liu et al. (2022). KGE can have a negative value, so that putting it to exp(-x) does not work. Liu et al. (2022) suggested using a gamma function to solve this issue. We've already mentioned this point in the original paper:

> Lines 146-148: Because KGE ranges from -infinity to one, Liu et al. (2022) applied a gamma density function to handle negative KGE values and developed a proper informal likelihood for the DREAM algorithm.

In the original version of the paper, the motivation to use KGE as an informal likelihood was indeed unclear. While MCMC often uses formal likelihood (i.e. mean squared error), KGE (or NSE) is often used in deterministic optimization. When we used different likelihood functions for MCMC and deterministic optimization, it become complicated to interpret the results between BMA, BPA, and BMPA. To avoid this complication, we have decided to use the method of Liu et al. (2022). This point has been mentioned in the revised version of the paper. We found that we forgot to mention that deterministic optimization maximizes KGE, which has been fixed in this revision:

> Lines 129-130: We used the *fminsearchbnd* function in MATLAB to search parameter sets which maximize Kling-Gupta Efficiency (KGE: Gupta et al. 2009).
>
> Lines 140-149: While mean squared error is usually used as the formal likelihood function (e.g., mean squared error) in the DREAM algorithm and many other MCMC applications, the informal likelihood function such as KGE is used for deterministic optimization. In our work, we will compare MCMC and deterministic optimization, so that it is necessary to use the same likelihood measure for both methods. Liu et al. (2022) proposed using KGE as an informal likelihood function. Because KGE ranges from -infinity to one, Liu et al. (2022) applied a gamma density function to handle negative KGE values and developed a proper informal likelihood for the DREAM algorithm. To use KGE in our MCMC algorithm, we used the method of Liu et al. (2022).

(2.3) 3. Writers can report computational cost and MCMC configuration. Given the emphasis on computational expense, please report the key DREAM settings and approximate compute:
number of chains, iterations, burn-in/thinning (if any), and typical runtime per basin/model (or total core-hours).

→ We have clarified the DREAM settings in the revised paper:

> Lines 153-156: The numbers of chains and maximum iterations were set to 3 and 10000, respectively. We sampled backwards from the end of the chains to avoid sampling in the burn-in period. The number of thinning intervals was set to 10.

We have clarified the typical runtime of MCMC computation.

> Lines 155-158: The typical computation time of MCMC to estimate posterior parameter distributions for 17 models in a single river basin using a single Intel Xeon Gold 6348 CPU processor is 85 hours.

(2.4) 4. Define calibration/validation windows and the "shifted" calibration rule. Please state explicitly how calibration/validation periods are chosen in the two data regimes, and how the calibration window is shifted when discharge records are incomplete (rule used, and how many basins are affected).

→ We have given an example of the "shifted" calibration rule in the revised version of the paper. We believe that the description greatly helps understand our calibration/validation window settings.

> Lines 233-237: For instance, if a complete discharge record is unavailable for 1987, we extend the calibration period to 1986-1991 to evaluate models by 5-year long observation (i.e., observation in 1986, 1988, 1989, 1990, and 1991). In this case, the remaining 24-year (1992-2015) data were used for evaluation, so that the validation period is slightly shortened.

We also explained the number of basins in which the periods are shifted.

> Lines 237: We performed the shift of the calibration period in 8 river basins.

(2.5) 5. Missing data handling. Please clarify how missing discharge values are treated in calibration and validation (e.g., removed days, gap-filling, objective computed only on overlapping periods), and whether forcing inputs ever contain gaps and how those are handled

→ We compute objectives only in overlapping periods. This point has been clarified in the revised paper:

> Lines 215-216: Daily discharge data in MERV-Jp have missing values. The likelihood function (i.e., KGE) is computed only on days when river discharge data is available.

Forcing inputs do not contain gaps. They are reanalysis-satellite-in-situ merged products. We have clarified that we do not have any missing value in forcing input in the revised version of the paper:

> Lines 216-217: There are no missing values in our forcing inputs.