https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

Technical Note: Benefits of Bayesian estimation of model parameters in a large hydrological model ensemble

Yohei Sawada[1], and Shinichi Okugawa[1]

[1] Department of Civil Engineering, Graduate School of Engineering, the University of Tokyo, Tokyo, Japan

Corresponding author: Y. Sawada, Department of Civil Engineering, the University of Tokyo, Tokyo, Japan, 7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan, yoheisawada@g.ecc.u-tokyo.ac.jp

Abstract

Quantifying and mitigating parametric and structural uncertainties in hydrological models are crucial to accurately understand and predict the rainfall-runoff process. Despite recent advances in Bayesian approaches for quantifying structural uncertainty using very large hydrological model ensembles, the simultaneous quantification of both parametric and structural uncertainties has yet to be implemented since previous works on large model ensembles have relied on deterministic optimization of parameters. Here we present the potential benefits of Bayesian estimation of parametric uncertainty within a large hydrological model ensemble. We find that Bayesian estimation of model parameters (more generally, change in calibration methods) potentially influences the interpretation

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

of model comparisons. Specifically, Bayesian parametric uncertainty quantification greatly benefits complex models with many parameters, thereby affecting discussions of the appropriate level of model complexity. We also find that Bayesian parametric uncertainty quantification does not substantially improve multi-model hydrological predictions. The adverse effects of parameter misspecification in individual models are effectively mitigated by combining models with diverse structures. Thus, the high computational cost of Bayesian parameter estimation is not paid for to improve rainfall-runoff analysis in a large hydrological model ensemble.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

1. Introduction

Hydrological models are essential tools to simulate the relationship between meteorological conditions and runoff in river basins. Quantifying and mitigating uncertainties in these models can greatly contribute to the accurate flood and drought prediction, water resources management, and climate change assessment. Assuming minimal error in input data, uncertainties in hydrological models can be broadly classified into two categories: parametric uncertainty and structural uncertainty (e.g., Beven 2005; Gupta et al. 2012). While parametric uncertainty arises from the misspecification of model parameters, which are coefficients of equations represented in the model, structural uncertainty originates from the specification of the equations themselves. The quantification and mitigation of both parametric and structural uncertainties based on hydrological observations, particularly river discharge data, remain grand challenges in hydrology.

Mitigating parametric uncertainty has been extensively investigated in hydrology. Early research largely focused on estimating a single set of parameters that minimized the cost function measuring the fit between simulation and observation. In this paper, we refer these approaches as deterministic optimization. A wide variety of gradient-based and evolutionary algorithms have been applied to this task (e.g., Duan et al. 1993; Tolson et al. 2007; Fowler et al. 2014; Qin et al. 2017, among others). Deterministic optimization methods suffer from equifinality (Beven 2006), in which multiple combinations of parameters reproduce observations equally well, and degrade the accuracy of simulating unseen data. Bayesian estimation, by contrast, explicitly produces posterior probability distributions based on observations and prior knowledge of parameters. This approach

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

offers several advantages: it can address equifinality (Vrugt et al. 2008a), enables probabilistic forecasting (e.g., Hung et al. 2025), and consider errors in data such as rainfall and river discharge measurements (e.g., Vrugt et al. 2008b). The golden standard of Bayesian estimation is Markov Chain Monte Carlo (MCMC; Hastings 1970; Geman and Geman 1984). In hydrology, the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al. 2008c; Laloy and Vrugt 2012) has been widely used as a variant of MCMC algorithms. While applying MCMC to hydrological models is computationally expensive, sequential data assimilation offers a computationally cheap alternative to estimate posterior distributions of model parameters and state variables (e.g., Moradkhani et al. 2005; Vrugt et al. 2013; Sawada 2022).

A common approach to mitigating structural uncertainty in process-based hydrological models is the applications of machine-learning, which can provide fully data-driven modeling (e.g., Kratzert et al. 2018, 2019) and correct the bias of process-based models (e.g., Funato and Sawada 2025). These methods are intrinsically deterministic and remain subject to equifinality. Large hydrological model ensembles, such as Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT; Knoben et al. 2019) and Raven (Craig et al. 2020) opened the door to easily perform Bayesian uncertainty quantification of model structure. Previous works employed Generalized Likelihood Uncertainty Estimation (GLUE)-like methods (Beven and Binley 1992), in which “reasonably good” models are selected based on predefined performance thresholds to identify equally plausible models within a large multi-model ensemble. For instance, Knoben et al. (2020) evaluated 36 hydrological models across 559 river basins using MARRMoT and identified a high degree of structural equifinality. Knoben et al. (2025) further

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

demonstrated that this high equifinality can be mitigated by considering sampling uncertainty in evaluation data. Chlumsky et al. (2021) performed simultaneous calibration of model structures and parameters (see also Mai et al. 2020). They revealed a high degree of equifinality within hydrological models implemented in the Raven framework. Although these works on large samples of hydrological models have advanced the quantification of structural uncertainty, they relied on deterministic parameter optimization. Therefore, they did not explicitly consider parametric uncertainty in the Bayesian way. It has yet to be clarified how Bayesian quantification of parametric uncertainty benefits large hydrological model ensemble-based assessment of structural uncertainty.

To address this research gap, we performed MCMC-based parametric uncertainty quantification for 17 hydrological models with different structures across 51 river basins, thereby enabling Bayesian estimation of both parametric and structural uncertainty. We then examined scientific and practical benefits of applying Bayesian estimation of model parameters in a large model ensemble. Specifically, we posed the following research questions:

- Scientific benefits: Does Bayesian estimation of model parameters (or, more broadly, changes in calibration methods) potentially affect the interpretation of model comparisons?
- Practical benefits: Does Bayesian estimation of model parameters improve the accuracy of rainfall-runoff simulations in a large hydrological model ensemble? Is this improvement sufficient to justify the large computational cost of MCMC?

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

2. Method

2.1. Hydrological model

We used MARRMoT v1.3 (Knoben, 2019). MARRMoT includes 46 lumped conceptual hydrological models with a wide range of complexities. The models are driven by daily total precipitation, daily mean temperature, and daily mean potential evapotranspiration, and they estimate daily basin-averaged runoff. From the 46 available models, we selected models with IDs 02, 03, 04, 06, 07, 10, 11, 12, 13, 17, 18, 21, 24, 27, 30, and 31 (see the supplement material of Knoben et al. 2019 for model details). These 17 models were chosen to preserve diversity in model complexity while minimizing the total computational cost, since Bayesian estimation of model parameters is computationally expensive.

2.2. Parameter estimation

2.2.1. Deterministic optimization

As a deterministic optimization method, we used the Nelder-Mead algorithm (Nelder and Mead, 1965). We performed the deterministic parameter optimization for each model in each basin. We used the *fminsearchbnd* function in MATLAB. Parameter ranges were specified according to the original setting of MARRMoT (see Knoben et al. 2019). Although our optimization method is a classical method and may be less capable of avoiding local minima than modern evolutionary algorithms (e.g., Duan et al. 1993; Tolson et al. 2007), we have found that it achieved high performance in our testbed (Sawada et al. 2022; see Section 3).

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

2.2.2. Bayesian optimization

For Bayesian optimization, we adopted the method proposed by Liu et al. (2022). As the MCMC method, we used the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al. 2008c; Laloy and Vrugt 2012). 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. 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. We used the MATLAB implementation of DREAM (https://github.com/Zaijab/DREAM). Besides the use of the KGE-based informal likelihood function, we used the default hyperparameter setting in this implementation. From the resulting Markov chains, we sampled 200 parameter sets to represent the posterior distribution of model parameters.

2.3. Bayesian model averaging

To evaluate the practical benefits of Bayesian parametric uncertainty quantification in a large hydrological model ensemble, we combined models with different structures and parameters into a single prediction using Bayesian Model Averaging (BMA).

First, we combined 17 models calibrated by deterministic optimization. Let the quantity of interest generated by the k-th model with deterministically optimized parameters, $M_k(\boldsymbol{\theta}_{DO,k})$, be denoted as Δ_k, where $\boldsymbol{\theta}_{DO,k}$ represents the optimized parameters of the k-th model. In this study, Δ_k is runoff in the validation period (see Section 3). Given observation $\boldsymbol{y}$, the posterior mean of the quantity of interest is:

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

$$E[\Delta|\boldsymbol{y}] = \sum_{k}^{N} w_k \, \Delta_k \tag{1}$$

with model weights defined as:

$$w_k = P\left(M_k(\boldsymbol{\theta}_{DO,k})|y\right) = \frac{P\left(\boldsymbol{y}|M_k(\boldsymbol{\theta}_{DO,k})\right) P(M_k(\boldsymbol{\theta}_{DO,k})}{\sum_{i=1}^{N} P\left(\boldsymbol{y}|M_i(\boldsymbol{\theta}_{DO,i})\right) P(M_i(\boldsymbol{\theta}_{DO,i})} \tag{2}$$

where N is the total number of models (=17 in this case). $\boldsymbol{y}$ should be recognized as river discharge observation in a calibration period. Equation (1) indicates that the posterior mean is a weighted average of all model outputs, with weights proportional to their posterior probabilities. We parameterized the weights w_k using KGE:

$$w_k = \frac{\exp\left(-\left(KGE_{max} - KGE\left(M_k(\boldsymbol{\theta}_{DO,k})\right)\right)\right)}{\sum_{i=1}^{N} \exp\left(-\left(KGE_{max} - KGE\left(M_i(\boldsymbol{\theta}_{DO,i})\right)\right)\right)} \tag{3}$$

where KGE_{max} is the maximum KGE among the *N* models, and $KGE(M_k(\boldsymbol{\theta}_{DO,k})$ is the KGE of the k-th model. In this paper, this averaging of hydrological models calibrated by deterministic optimization is specifically referred to as Bayesian Model Averaging (BMA).

Next, we compared the individual models calibrated by deterministic optimization with those calibrated by MCMC. To do so, we sampled 200 parameter sets from the posterior distributions, ran models with those parameter sets, and then combined these 200 hydrological simulations by Bayesian model averaging. Define the quantity of interest generated by the k-th model with the *l*-th parameter set, $M_k(\boldsymbol{\theta}_l)$, as $\Delta_{k,l}$. Similar to model averaging, the posterior mean of the quantity of interest estimated by the k-th model, Δ_k follows:

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

$$E[\Delta_k|\boldsymbol{y}] = \sum_{l}^{M} w_l \, \Delta_{k,l} \tag{4}$$

with weights:

$$w_l = \frac{\exp\left(-\left(KGE_{max,k} - KGE\left(M_k(\boldsymbol{\theta}_l)\right)\right)\right)}{\sum_{i=1}^{M} \exp\left(-\left(KGE_{max,k} - KGE\left(M_k(\boldsymbol{\theta}_i)\right)\right)\right)} \tag{5}$$

where M is the total number of parameter sets (i.e. 200), $KGE_{max,k}$ is the maximum KGE among the M parameter sets of the k-th model. This averaging within a single model with different parameters is referred to as Bayesian Parameter Averaging (BPA).

Finally, we averaged all models and parameter sets. In this case, the posterior mean of the quantity of interest is:

$$E[\Delta|\boldsymbol{y}] = \sum_{k}^{N} \sum_{l}^{M} w_{k,l} \, \Delta_{k,l} \tag{6}$$

with weights:

$$w_{k,l} = \frac{\exp\left(-\left(KGE_{max,k,l} - KGE\left(M_k(\boldsymbol{\theta}_l)\right)\right)\right)}{\sum_{i=1}^{N} \sum_{j=1}^{M} \exp\left(-\left(KGE_{max,k,l} - KGE\left(M_i(\boldsymbol{\theta}_j)\right)\right)\right)} \tag{7}$$

where $KGE_{max,k,l}$ is the maximum KGE among all models and parameter combinations. This joint averaging is called Bayesian Model and Parameter Averaging (BMPA).

3. Experiment design

We applied the aforementioned methods to 51 river basins in Japan (Figure 1). We used meteorological forcings from the Multi-model Ensemble for Robust Verification of hydrological modeling in Japan (MERV-Jp) dataset (Sawada et al. 2022; Sawada and

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

Okugawa 2022). The river basins shown in Figure 1 cover a wide range of climatic, soil, land-use, anthropogenic, and topographic conditions. Sawada et al. (2022) reported that 44 deterministically calibrated models in MARRMoT achieved high accuracy to reproduce observed runoff. The best KGEs in 44 models exceeded 0.7 in nearly all basins, which are comparable to those reported in other large model ensemble studies across different regions (e.g., Knoben et al. (2020)). Therefore, our findings are expected to be transferable to similar studies in the context of large hydrological model ensembles.

The study period spans 1986-2015. Two calibration periods were considered: a 5-year calibration period representing data-rich conditions, and a 1-year calibration period representing data-poor conditions. In the 5-year calibration scenario, the initial 5-year (1986-1990) data were used for calibration with both deterministic optimization and MCMC, and the remaining 25-year (1991-2015) data were used for evaluation. In basins where complete discharge records were unavailable for 1986-1990, the calibration period was shifted to ensure a continuous 5-year record, which slightly reduced the validation period. We used the same 5-year data for model spin-up, resulting a 10-year model integration for each parameter evaluation step. In the 1-year calibration scenario, we applied the same 5-year data spin-up, followed by evaluation of parameters in the subsequent 1-year simulation (1986 in most basins). The validation data for the 1-year calibration scenario were identical to those for the 5-year calibration scenario. Model performance was evaluated using KGE and Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe 1970; see also Duc and Sawada 2023 for a modern interpretation of NSE) during the validation period.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

For deterministic optimization, we have 17 hydrological predictions by 17 models in 51 river basins, yielding 17×51 = 867 hydrographs. For Bayesian estimation of parametric uncertainty, each of the 17 models was run with 200 posterior parameter sets, producing 17×200 = 3400 simulations per basin. Across 51 basins, this amounted to 3400×51 = 173,400 hydrographs. First, we compared the performance of individual models calibrated by deterministic optimization with those calibrated by MCMC and combined by BPA (see Section 2.3) across all 17 models in 51 river basins, to assess if the evaluation of the model structures is affected by parameter estimation method. Second, we compared the performance of BMA and BPMA (see Section 2.3), to discuss the potential benefits of considering parametric uncertainty through Bayesian estimation in improving hydrological predictions.

4. Results and discussions

Bayesian estimation of model parameters (i.e., BPA) systematically outperforms deterministic optimization. Boxplots of Figure 2 show the differences in KGE and NSE between BPA and deterministically optimized models. The superiority of MCMC-based optimization is consistent with earlier findings on the DREAM (e.g., Laloy and Vrugt 2012) algorithm. To our best knowledge, we verified this superiority within a large model ensemble for the first time. Two main factors explain this result. First, MCMC explores the parameter space more extensively to maximize KGE than deterministic optimization. Second, MCMC accounts for equifinality by sampling multiple parameter sets that reproduce observation equally well, which leads the higher robustness to unseen data. Even in the rich-data scenario (i.e., the 5-year calibration), improvements in KGE exceeds 0.2 in some models. This trend is more pronounced in the data-scarce scenario (i.e., 1-

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

year calibration scenario). When calibration data are limited, parameter estimates are inherently uncertain, since many different parameter combinations may be able to equally explain the limited data. In such cases, Bayesian estimation is more appropriate than deterministic methods.

Figure 3 reveals that the improvements achieved by BPA over deterministically optimized models systematically appear. For instance, model ID 10 (Susannah Brook model v2; see Son and Sivapalan 2007, Knoben et al. 2019) shows substantial gains from Bayesian parameter estimation in more than 20 river basins. Except for this model, more complex models with the larger numbers of parameters (e.g., ID 30, 31, and 32) tend to benefit more from Bayesian estimation than simpler models (e.g., ID 2, 3, and 4) (note that models with higher IDs generally correspond to greater structural complexity; see Knoben et al. 2019) especially in the data-rich scenario (i.e., 5-year calibration scenario). This indicates that models with many parameters are particularly affected by equifinality and therefore gain substantially from Bayesian parameter estimation.

Previous studies have evaluated model structures by comparing the performance of deterministically optimized models. Our results imply that such evaluations may be affected by the choice of parameter optimization methods. For instance, although results in Knoben et al. (2020) and Knoben et al. (2025) indicated that complex models with many parameters do not necessarily outperform simpler models, this conclusion may partly reflect an underestimation of the maximum potential performance of complex models due to reliance on deterministic optimization. While these complex models have been shown not to suffer from overfitting (Knoben et al. 2020), there might be room for

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

improvement through Bayesian estimation, which explicitly addresses parameter equifinality.

When a large number of calibrated models are available, a practical way to improve prediction accuracy is to use the (weighted) average of their outputs (e.g., Kimizuka and Sawada 2022; Zhang and Yang 2018). The red dots in Figure 2 show the performance differences between BPMA and BMA. Although Bayesian parameter estimation substantially improves the performance of individual models, the overall improvement from BMA to BPMA is marginal. Even in the data-scarce scenario (i.e., the 1-year calibration scenario), the improvement in KGE (NSE) by Bayesian estimation is less than 0.1 (0.05). Surprisingly, even the classic optimization method remains competitive with DREAM-based Bayesian optimization when models are combined in a large ensemble. Considering the substantial computational costs of the MCMC-based Bayesian parameter estimation, we conclude that Bayesian parametric uncertainty quantification provides limited practical benefits for improving predictions in large hydrological model ensembles.

This occurs because poorly performing models produced by deterministic optimization are assigned lower weights in the BMA framework. Figure 4 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, 24, 30, and 31) perform poorly. Nevertheless, BMA remains competitive with BPMA, since the poorly performing models receive smaller weights during averaging. Therefore, the adverse effects of suboptimal calibration are effectively mitigated.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

5. Conclusions

Here we performed MCMC-based parameter optimization for 17 hydrological models across 51 river basins to clarify the potential benefits of Bayesian parametric uncertainty quantifications in a large hydrological model ensemble. Scientifically, Bayesian parametric uncertainty quantification is important because it can influence the interpretation of structural uncertainty assessment. The benefits of the Bayesian parameter estimation appear systematically across models rather than randomly. Certain models gain substantial improvements, and more complex models tend to benefit more strongly than simpler models. Considering Bayesian parameter uncertainty potentially affects the discussion of the appropriate complexity of hydrological models.

Practically, Bayesian parametric uncertainty quantification does not greatly contribute to improving multi-model ensemble hydrological prediction. It implies that structural errors are larger than parametric errors in hydrological prediction. Given the high computational cost of Bayesian estimation, multi-model ensembles calibrated by deterministic optimization are sufficient in many cases.

Our analysis was limited to 17 models, fewer than in previous large ensemble studies (Knoben et al. 2020, 2025; Chlumsky et al. (2021)). We had to exclude computationally expensive models in MARRMoT to make the MCMC applications feasible in this initial attempt. Future work should expand to all MARRMoT models and pursue GLUE-like assessments of structural uncertainty using Bayesian parameter uncertainty quantification by unleashing the power of high-performance computers.

Acknowledgement

This study was supported by the Foundation of River & basin Integrated CommunicationS (FRICS) grant, the JST Moonshot R&D Program (JPMJMS2281), the KAKENHI grant (25H00760, 23K20967), and Joint Usage and Research of JHPCN (jh230003, jh240064, jh2500035). The computation was performed on the Wisteria/BDEC-01 supercomputer in the University of Tokyo.

Code availability

MARRMoT v1.3 is available at https://doi.org/10.5281/zenodo.3235664 (Knoben, 2019). DREAM is available at https://github.com/Zaijab/DREAM.

Data availability

Results of hydrological models in this work can be found at https://doi.org/10.5281/zenodo.17282833 (Sawada and Okugawa 2025).

Author contribution

YS designed the study, interpreted results, and wrote the initial version of the paper. SO performed numerical experiments, analyzed the results, and contributed to editing the paper.

References

Beven, K.: On the concept of model structural error, Water Sci. Technol., 52, 167–175, https://doi.org/10.2166/wst.2005.0165, 2005.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Chlumsky, R., Mai, J., Craig, J. R., and Tolson, B. A.: Simultaneous calibration of hydrologic model structure and parameters using a blended model, Water Resour. Res., 57, https://doi.org/10.1029/2020WR029229, 2021.

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A. P., and Tolson, B. A.: Flexible watershed simulation with the Raven hydrological modelling framework, Environ. Model. Softw., 129, 104728, https://doi.org/10.1016/j.envsoft.2020.104728, 2020.

Duan, Q. Y., Gupta, V. K., and Sorooshian, S.: Shuffled complex evolution approach for effective and efficient global minimization, J. Optim. Theory Appl., 76, 501–521, https://doi.org/10.1007/BF00939380, 1993.

Duc, L. and Sawada, Y.: A signal-processing-based interpretation of the Nash–Sutcliffe efficiency, Hydrol. Earth Syst. Sci., 27, 1827–1839, https://doi.org/10.5194/hess-27-1827-2023, 2023.

Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved rainfall-runoff calibration for drying climate: Choice of objective function, Water Resour. Res., 54, 3392–3408, https://doi.org/10.1029/2017WR022466, 2018.

Funato, M. and Sawada, Y.: Multi-model ensemble and reservoir computing for river discharge prediction in ungauged basins, arXiv [preprint], https://arxiv.org/abs/2507.18423, 2025.

Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell., 6, 721–741, 1984.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resour. Res., 48, https://doi.org/10.1029/2011WR011044, 2012.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hastings, W. K.: Monte Carlo sampling methods using Markov chains and their applications, Biometrika, 57, 97–109, https://doi.org/10.1093/biomet/57.1.97, 1970.

Kimizuka, T. and Sawada, Y.: How robust is a multi-model ensemble mean of conceptual hydrological models to climate change?, Water, 14, 2852, https://doi.org/10.3390/w14182852, 2022.

Knoben, W. J. M.: MARRMoT_v1.2, Zenodo [code], https://doi.org/10.5281/zenodo.3235664, 2019.

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., and Woods, R. A.: Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations, Geosci. Model Dev., 12, 2463–2480, https://doi.org/10.5194/gmd-12-2463-2019, 2019.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments, Water Resour. Res., 56, e2019WR025975, https://doi.org/10.1029/2019WR025975, 2020.

Knoben, W. J. M., Raman, A., Gründemann, G. J., Kumar, M., Pietroniro, A., Shen, C., Song, Y., Thébault, C., van Werkhoven, K., Wood, A. W., and Clark, M. P.: Technical note: How many models do we need to simulate hydrologic processes across large

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

geographical domains?, Hydrol. Earth Syst. Sci., 29, 2361–2375, https://doi.org/10.5194/hess-29-2361-2025, 2025.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

Laloy, E. and Vrugt, J. A.: High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing, Water Resour. Res., 48, W01526, https://doi.org/10.1029/2011WR010608, 2012.

Liu, Y., Fernández-Ortega, J., Mudarra, M., and Hartmann, A.: Pitfalls and a feasible solution for using KGE as an informal likelihood function in MCMC methods: DREAM(ZS) as an example, Hydrol. Earth Syst. Sci., 26, 5341–5355, https://doi.org/10.5194/hess-26-5341-2022, 2022.

Mai, J., Craig, J. R., and Tolson, B. A.: Simultaneously determining global sensitivities of model parameters and model structure, Hydrol. Earth Syst. Sci., 24, 5835–5858, https://doi.org/10.5194/hess-24-5835-2020, 2020.

Moradkhani, H., Hsu, K.-L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, Water Resour. Res., 41, https://doi.org/10.1029/2004WR003604, 2005.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models. Part 1: A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nelder, J. A. and Mead, R.: A simplex method for function minimization, Comput. J., 7, 308–313, https://doi.org/10.1093/comjnl/7.4.308, 1965.

Qin, Y., Kuczera, G., and Kavetski, D.: Comparison of Newton-type and SCE optimisation algorithms for the calibration of conceptual hydrological models, Australas. J. Water Resour., 20, 169–176, https://doi.org/10.1080/13241583.2017.1298180, 2016.

Sawada, Y.: An efficient estimation of time-varying parameters of dynamic models by combining offline batch optimization and online data assimilation, J. Adv. Model. Earth Syst., 14, e2021MS002882, https://doi.org/10.1029/2021MS002882, 2022.

Sawada, Y., Okugawa, S., and Kimizuka, T.: Multi-model ensemble benchmark data for hydrological modeling in Japanese river basins, Hydrol. Res. Lett., 16, 73–79, https://doi.org/10.3178/hrl.16.73, 2022.

Sawada, Y., and Okugawa, S.: Dataset of "Benefits of Bayesian estimation of model parameters in a large hydrological model ensemble" [Dataset], https://doi.org/10.5281/zenodo.17282833, 2025

Son, K. and Sivapalan, M.: Improving model structure and reducing parameter uncertainty in conceptual water balance models through the use of auxiliary data, Water Resour. Res., 43, W01415, https://doi.org/10.1029/2006WR005032, 2007.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resour. Res., 43, https://doi.org/10.1029/2005WR004723, 2007.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

Vrugt, J. A., Ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson, B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resour. Res., 44, https://doi.org/10.1029/2007WR006720, 2008b.

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., and Hyman, J. M.: Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, Int. J. Nonlinear Sci. Numer. Simul., 10, 273–290, 2008c.

Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V., and Robinson, B. A.: Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, Stoch. Environ. Res. Risk Assess., 23, 1011–1026, https://doi.org/10.1007/s00477-008-0274-y, 2008a.

Vrugt, J. A., Ter Braak, C. J. F., Diks, C. G. H., and Schoups, G.: Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications, Adv. Water Resour., 51, 457–478, https://doi.org/10.1016/j.advwatres.2012.04.002, 2013.

Zhang, L. and Yang, X.: Applying a multi-model ensemble method for long-term runoff prediction under climate change scenarios for the Yellow River Basin, China, Water, 10, 301, https://doi.org/10.3390/w10030301, 2018.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

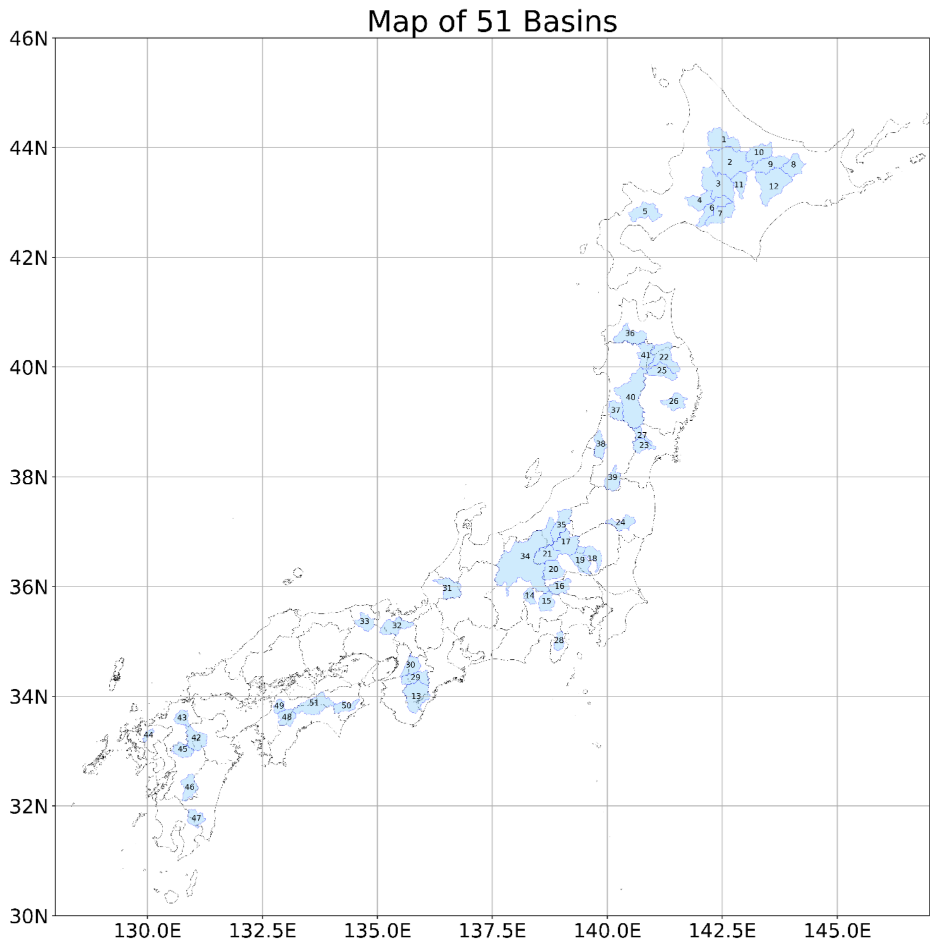


Figure 1. Study area of 51 river basins.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

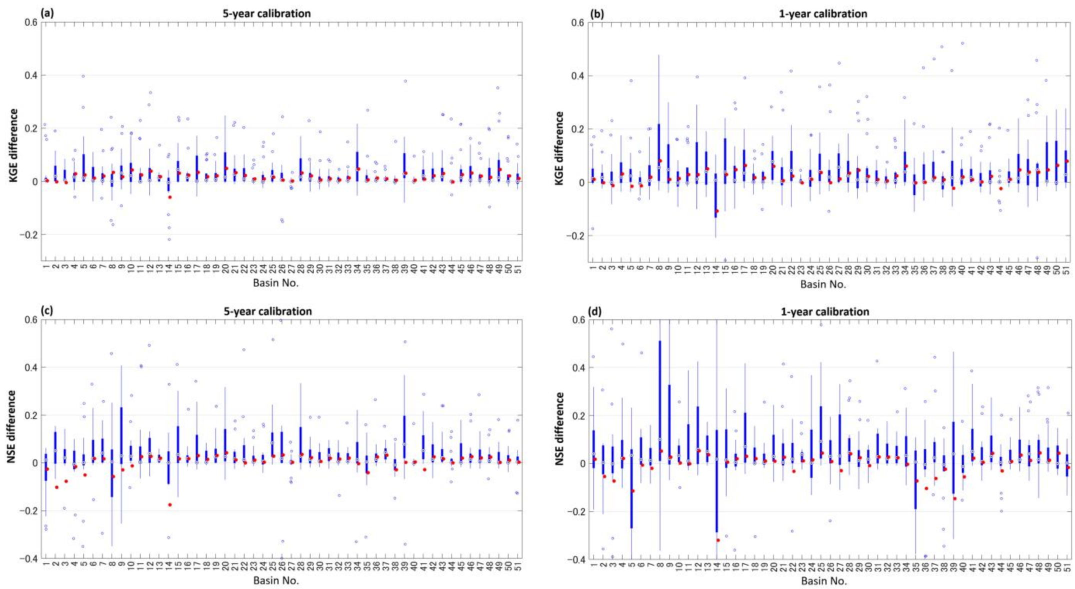


Figure 2. Differences in KGE (a, b) and NSE (c, d) between BPA and deterministically optimized models (boxplots) in 5-year (a, c) and 1-year (b, d) calibration scenarios. Red dots show the performance differences between BMPA and BMA (see Section 3).

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025

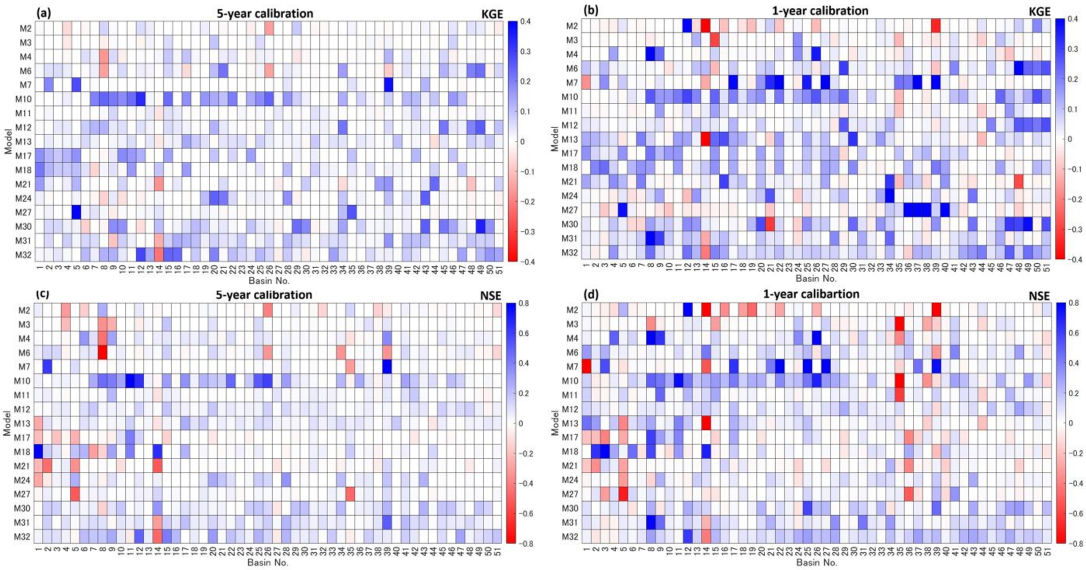
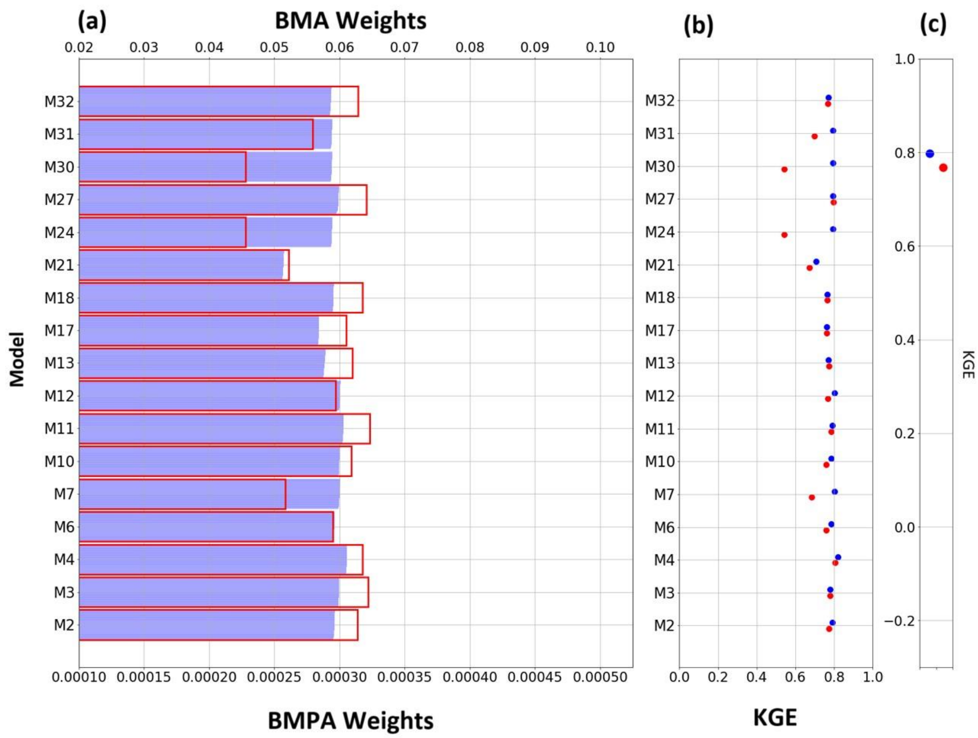


Figure 3. Differences in KGE (a, b) and NSE (c, d) between BPA and deterministically optimized models for each basin and model in 5-year (a, c) and 1-year (b, d) calibration scenarios.

https://doi.org/10.5194/egusphere-2025-4984
Preprint. Discussion started: 5 December 2025



466

467 **Figure 4.** (a) Weights in BMA (red bars) and BMPA (blue bars). (b) KGE of BPA (blue dots) and
468 deterministically optimized models (red dots). (c) KGE of BMPA (blue dots) and BMA (red dots). Results of
469 basin No. 43 (see Figure 1) are shown.