“Selecting a conceptual hydrological model using Bayes' factors computed with Replica Exchange Hamiltonian Monte Carlo and Thermodynamic Integration” by Damian N. Mingo, Remko Nijzink, Christophe Ley, and Jack S. Hale.

Reviewer #2

https://doi.org/10.5194/egusphere-2023-2865-RC2

We would like to thank the second reviewer for their thoughtful comments. We will address their specific comments in this response and move towards a final response in the coming weeks. We are also more than happy to discuss specific points with the reviewer again.

I have now finished reviewing the work by Mingo et al. The authors have combined Replica-Exchange Hamiltonian Monte Carlo (HMC) with Thermodynamic Integration in order to do Bayesian inference for the parameters of a conceptual hydrologic model, while simultaneously they compute the marginal likelihood of the model; the latter, facilitates model inter-comparison via the Bayes Factor (BF). In general, the manuscript is well written and has novelty in the sense that the proposed algorithm has never been applied before to hydrological modeling. As a result, I am overall positive! However, I think the manuscript would benefit from a more in-depth discussion (possibly toward the end of the article) about the scientific problem that the authors address, the limitations, and what are some possible alternatives.

Indeed, there is an implicit assumption in our paper that computing the BF is something one might want to do in the first place! We agree we should have been more expansive on this point, so we make some specific answers to your points below. We will then paraphrase this into some new paragraph(s) in the discussion.

In light of the extensive comments (major and editorial) of Reviewer #1 with which I completely agree, I would like to raise some concerns about the usefulness of BF as a hydrologic model inter-comparison metric. Please see my comments below:

1. For the synthetic experiments, Tables 4 and 5 show that both DIC and WAIC could correctly indicate the data-generating model, i.e., M2 and M3, respectively. For the average reader, this might practically mean that we do not need BF as an additional metric to "tell" us which model to choose. Please provide an explanation to show why employing BF matters. If you cannot demonstrate that the BF can capture the true underlying model while the other, simpler metrics, cannot, then it is hard to justify your analysis.

There is an example of the BF succeeding to identify the underlying data generating model in our paper, whereas the DIC does not, and the WAIC only provides at best weak evidence.

In Experiment 2 (data generated from the three-bucket model M3), the DIC values for M3 and M4 differ by ~1, while the WAIC values for M3 and M4 differ by ~3 (Table 5 and Figure 12b).

(Burnham & Anderson 2002) on page 71, discusses ~4-7 (less evidence) and >~10 (substantial evidence) in favour of one model over another when using IC-type measures for model selection. There are other similar values in the literature but this seems to be a commonly used interpretation, akin to the table of Kass and Raftery for the BF (1995) which we show in Table 1.

Information and Likelihood Theory: A Basis for Model Selection and Inference. (2002). In K. P. Burnham & D. R. Anderson (Eds.), Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (pp. 49–97). Springer New York. doi: 10.1007/978-0-387-22456-5 2

With this interpretation in mind, the DIC does not provide evidence to prefer M3 over M4. The WAIC possibly provides weak evidence in favour of M3 over M4, but we would be wary of making that conclusion by noting the substantial size of the error bar for M4 WAIC in Fig 12b. The BF (Table 5) decisively selects M3, the data generating model, over both M2 and M4.

This is clearly not evidence for the superiority of the BF as a model selection tool in all circumstances, and so we would not feel comfortable framing the BF as being superior in the paper. However, it is indicative that there may be cases where the BF succeeds where other approaches do not. The BF comes at a substantial computational cost over IC-type measures, necessitating improved algorithms such as the one we proposed in this paper if the BF is ever to be used at all in practice.

We will tweak the discussion of Experiment 2 to highlight this point better.

2. Although I am not a Hydrologist myself, I have a hard time understanding the usefulness of BF within the context of hydrologic model comparison. Traditional hydrologists calibrate models using algorithms like Shuffled Complex Evolution (SCE) based on optimization of a deterministic metric, e.g., NSE. I do understand that Bayesian inference of hydrologic model parameters, on the other hand, is appealing because it naturally provides a measure of uncertainty, which is always important. But the BF provides no pragmatic information to the modeler as per which model is performing better. For example, one would still have to compute NSE or KGE for all models M2, M3, and M4 for the real-world data (Table 8) to get an idea of what's happening. On the contrary, I would argue that for conceptual hydrologic models, which are not computationally demanding and time-intensive, likelihood-free methods like Approximate Bayesian Computing (ABC) might be more suitable for model comparison, as the posterior distributions of parameters for different models are obtained on the basis of an actually useful (to the modeler) distance metric, e.g., NSE, KGE, or even a metric tailored only to river discharge peaks!!!

Again, I am positive about your article and I believe it should be considered for publication, but please provide a better discussion about the practical use of BF as a hydrologic model comparison metric…

We will first discuss the issue of the metric or goodness-of-fit measure.

In the paper we use a iid Gaussian likelihood function which in a deterministic setting can be seen as being equivalent to weighted minimisation in an l^2 norm, with the Bayesian prior being equivalent to some regularization term added to the l^2 norm.

According to (Cheng et al. 2014) the NSE is “equivalent to a log-likelihood function with iid Gaussian residuals”. Consequently, if the modeler wishes to use NSE as a metric for parameter calibration (Cheng et al. 2014) proposes that they could simply use a iid Gaussian as a likelihood in a formal Bayesian analysis.

Qin-Bo Cheng, Xi Chen, Chong-Yu Xu, Christian Reinhardt-Imjela, Achim Schulte, Improvement and comparison of likelihood functions for model calibration and parameter uncertainty analysis within a Markov chain Monte Carlo scheme, Journal of Hydrology, Volume 519, Part B, 2014, Pages 2202-2214. https://doi.org/10.1016/j.jhydrol.2014.10.008

We are unable to find results formally linking the KGE with a likelihood function, which means that if the modeler wants to use KGE, they cannot use a formal Bayesian analysis. We found the paper (Liu et al. 2022) which derives an object called an ‘informal pseudo probability density based on the KGE” which is then used in a formal Bayesian analysis, along with some discussion in the introduction of similar adaptations to the NSE. Perhaps this could then be used to evaluate the BF but this is a conjecture at this stage.

Yan Liu, Jaime Fernández-Ortega, Matías Mudarra, and Andreas Hartmann, Pitfalls and a feasible solution for using KGE as an informal likelihood function in MCMC methods: DREAM(ZS) as an example, Hydrology and Earth System Sciences,, 26, 5341–5355, https://doi.org/10.5194/hess-26-5341-2022

Likelihood-free methods such as Approximate Bayesian Computing (ABC) that bypass the evaluation of a likelihood function are a potentially good alternative if an explicit link with between a metric and the likelihood function is unavailable, as in the case of the KGE or perhaps with the example you mention with calibrating to capture peak discharge.

In summary, it seems that the adaptation of commonly used metrics such as NSE and KGE to a full Bayesian setting is still an active area of research. We will add a remark discussing metric choice and ABC to the section “Likelihood construction 2.2.1”. This more detailed discussion will remain here if the reader is interested.

We now discuss the point on metrics vs model selection criteria.

The Gaussian likelihood (implying the weighted l^2 norm), NSE and KGE can be used as measures of goodness-of-fit for both training/fitting/calibration and nested model comparison via e.g. likelihood ratios. However, this is distinct from the model selection criteria (the BF or IC-type measures) which attempt to give a measure of fit balanced by an explicit or implicit penalisation for model complexity. Fit alone cannot choose between two models with free parameters which reproduce the data ‘similarly’ well. This is shown in our paper where we have deliberately constructed M4 with a strict superset of the model elements of M3 (and similarly M3 with M2).

So we agree that computing fit measures and performing graphical and formal posterior predictive checks are still an essential part of the modeling process. What the IC-type and BF offer is an additional measure for comparing models, potentially allowing the choice between models with similar fits.