“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.

We would like to thank the first anonymous reviewer for their thorough and insightful feedback on our manuscript. We respond (normal text) to the comments (*italics*) inline below with a plan for producing an improved version for the reviewer and the wider community’s consideration. We are also more than happy to discuss specific points again before proceeding with these changes.

Reviewer 1

General comments

The paper introduces a complex numerical method for robust model comparison using Bayes Factors. More precisely, the authors propose a pipeline for estimating the marginal likelihood (and consequently, the Bayes Factor) by combining thermodynamic integration with Replica exchange Monte Carlo for power posterior ensemble simulation, and Preconditioned Hamiltonian Monte Carlo (pHMC) for efficient gradient-based sampling. This appears to be one of the first approaches that integrate all these sub-algorithms into one pipeline, in conjunction with an innovative implementation of the methodology within a probabilistic programming framework paired with a differentiable programming language. The paper provides a comprehensive overview of the contributions and related work, as well as an extensive explanation of the methodology and all relevant numerical methods. The authors then discuss the implementation aspects and model before presenting the results. While the results section is well-organized, it left me, as a reader, wanting more. Specifically, the results section does not convincingly demonstrate that the presented methodology effectively addresses the problems mentioned in the introduction. Additionally, certain details in the results section appear to be rushed over and sporadically mentioned without proper references or prior introduction. In conclusion, I would recommend enhancing the results section with more convincing evidence and a clearer exposition of the details before considering the paper for publication.

Specific comments

1. In the abstract, the sentence detailing the prior calibrated posterior predictive p-value may be too intricate for readers unfamiliar with the basic concept of p-value and 'posterior predictive p- value.'

We will remove this point which is too technical for the abstract and introduction - we will also discuss this again in our response to Question 13.

2. The introduction's layout, which typically presents background and related works before concluding with the paper's contributions, is more familiar to me. Perhaps swapping sections 1.2 Background and 1.1 Contribution could be considered for a more traditional structure.

We will switch the subsections around and make some additional changes if needed to maintain the flow of the overall section.

3. Bayes Factors (BF) are a crucial component of the paper, yet the formula for calculating them within a multimodal context, beyond just two models, is absent. I expected to find this expression, potentially as an extension of Equation 6

We will remove the specific case and add the general case for two models indexed with e.g. i and j.

4. The sentence after line 280 stating, 'the samples of the replica with β = 1 are used to estimate the posterior parameters,' highlights a significant procedure that is not adequately explained.

We agree, this is an important point for readers that was only briefly touched on; although all of the replicas are used to improve chain mixing and in the subsequent marginal likelihood calculation, the parameter estimates are derived only from the statistics of the beta = 1 chain. We will improve this section with comments to this effect.

5. In Algorithm 3, index 'j' iterates from 1 to L, representing the number of leapfrog steps. However, the index 'j' does not appear clearly within the algorithm, leading to potential confusion.

We will correct this issue.

6. The No-U-Turn sampler (NUTS) and Metropolis-adjusted Langevin algorithm (MALA) are suddenly introduced in Section 3.1, without prior mention or any references, and are then used for comparing the results obtained with pHMC. Given their relevance to the results section, introducing these MCMC variants earlier in the Background would enhance the paper's cohesiveness.

We added MALA and NUTS (which are not used in the subsequent result section, rather (p)HMC) in response to a previous question about our work (would NUTS or other samplers do a better job than (p)HMC on the Gaussian shell problem?).

The key point is that NUTS, MALA or (p)HMC are very unlikely to transition across the gap between the two Gaussian shells, so the answer to the question is no. The important addition here is the use of the Replica Exchange algorithm (it could be used with either NUTS, HMC or MALA). We could also run this test with HMC and get the same isolated shell. We propose mentioning in the text and caption that HMC produces similar plots to NUTS and MALA and explain this more clearly. We could also produce the figure with (p)HMC and simply mention that MALA and NUTS produce the same single shell, not exploring the other half.

7. In Figure 6, the produced prior predictive 95% pointwise confidence interval seems quite narrow, which is unexpected given the variability one would anticipate when sampling from a 13-dimensional (prior) uncertainty space. Additionally, the observed discharge should have been plotted for comparison, to evaluate how well it is bracketed by the prior uncertainty interval.

We plan to remove Figure 6 and revise Figure 5 to include both the prior predictive data (not just one sample “synthetic discharge”) and the observed discharge (already shown).

8. The Deviance Information Criterion (DIC) and Widely Applicable Information Criterion (WAIC) are introduced at the end of Section 3.2 without any explanation or references. In Paragraph 430, the IAT number and Geweke diagnostics are also mentioned without reference. It is unclear if these are assumed to be general knowledge. The placement of these terms is somewhat non-intuitive as they are subsequently used throughout the results section.

We will add these definitions, possibly in an Appendix - we already feel the paper is quite long!

*9. The report lacks a clear statement regarding the number of forward model runs that were evaluated or needed. Is the correct interpretation that 10*4000 runs were conducted, multiplied by 15 for each model?*

It’s a good point, thanks for mentioning this. We haven’t expanded on this point properly at all in the paper - we will add a general estimate of the number of forward model runs (and additionally, adjoint/backwards runs to get the gradient, that are of a similar complexity to the forward model run) and explain where they take place in the algorithm.

10. The results of the synthetic experiments from Sections 3.2.1 and 3.2.2, depicted in Figures 10 and 13, are confusing. The Model 4 with four buckets (M_4) seems to be well calibration with the data originating from a much simpler model. This raises the question of why the hydrographs of Model 4 aligns so closely with those generated by the 'true' model, which would not be expected.

Model 4 contains a superset of the components in Model 2 - consequently, it can reproduce the dynamics of the data produced from one of the simpler models (Model 2, in this case). We set this up intentionally to demonstrate that the Bayes factor will penalize the more parametrically complex model. However, we did not expand on this point properly - we will revise the text to make this clearer.

11. What does the conclusion from Figure 10 mean? “Hence, BF penalizes models with more parameters.” How does one conclude this?

This is connected to our response to question 10. The model fit is the same, so the BF penalizes the model with more parameters (the expected result). We will revise this caption to be more precise.

12. In Section 3.3, the authors compare the uncertainty bounds in Figure 16 with a prior-predicted hydrograph from Figure 5. However, the hydrograph in Figure 5 represents only a single random realization from the prior, which seems like an inappropriate comparison. It would be more informative to compare the Monte Carlo mean derived from the prior with the mean hydrograph obtained from the learned posterior. As it stands, Figures 5 and 16 do not seem to be compared on an equitable basis.

We agree, so we propose to change Figure 5 as proposed in our answer to Question 7.

13. The results and discussion in the results section have not convincingly demonstrated the ability of the prior calibrated posterior predictive p-value to detect prior data conflicts, a capability that was highlighted in the abstract and introduction.

We will drop this point from the abstract and introduction as we agree it could almost justify a paper on its own (we will leave it in as a diagnostic, but not as a core part of the paper).

14. Section 3.4.2, titled “Convergence of marginal likelihood”, feels brief and incomplete, as if the discussion in unfinished.

We will add extra details of the marginal likelihood stabilizing for increasingly fine discretisation of the thermodynamic integral (TI) and some extra discussion on the convergence.

15. The Nash Sutcliffe efficiency (NSE) obtained for the selected model is 0.397, which is low for a model deemed to be calibrated. Typically, NSE values below 0.6 are considered ‘low’.

We mentioned around line 503 that the NSE shows that the model is better than the mean, but we agree that the value is still ‘low’ - we will mention this in this text.

On the broader point of model (in-)adequacy, we agree that if these results were shown in the context of a paper proposing new hydrological models (defined by the operator G), there would be room for improvement. We picked this ‘HBV-like’ model with an extendable number of buckets largely for simplicity. However, the key contribution of the paper is on an approach for the model selection problem via Bayes Factors and in that context we think the results make sense.

Our wider point (line 160) is that the community should consider differentiable models and PPL as a standard methodology when developing new modelling toolboxes. This will open up the practical range of questions about models from parametric calibration (currently common) to model selection (still rare using BFs) enabled by algorithms like the ones we propose in this paper.

16. The results section does not sufficiently demonstrate the efficacy of the model. The findings presented in Figures 14 and 15 lack in-depth discussion. Although convergence diagnostics for real-world data suggest 'good' outcomes, the presentation falls short of being persuasive. Furthermore, the methodology appears to struggle with definitively identifying the most likely model in real-data scenarios, as indicated by the results in Table 8.

We will add some extra explanations on Figure 14 and 15.

The second point we have already addressed in response to Question 15.

The process identifies model 3 as the most likely model in the real-data scenario under the assumptions we choose to make (the true data-generating model is unknown and by definition, not in the predefined set of models). This is why we show many results for when the data generating model is in the set to persuade the reader that the algorithm works.

17. It would be beneficial to include a visual comparison, such as hydrographs, of the calibrated models M2 and M4 against the real data to better illustrate their performance.

We will include the hydrographs against real data.

18. Based on the results of Figure 16, the uncertainty band drawn from the posterior seems wide, even for small streamflow values, which does not give a good hint of adequate parameter estimation / model calibration.

Indeed, the bands are wide, suggesting model improvements, hybrid models or modern approaches such as neural ODEs may be necessary from a hydrological modelling perspective.

Technical corrections

1. Review the formatting and positioning of Equations 1 and 2e.

We aligned the equals but we can equally align on the left.

2. Clarify the paragraph containing lines 105, 110, and 115 to eliminate repetitive information and streamline the content for better readability.

We will rewrite this part removing the duplicate sentences on DREAM, and merge the sentences on HMC, and generally make it easier to read.

3. Check Equation 2b for a possible typographical error: it should state (V_2)t = and n=2, rather than (V_1)t = n>=2.

We will correct this.

4. In Equation 3, there appears to be a typo with k_{2,1}, it should likely read k_{1,2}.

We will correct this.

5. The use of 'p' to denote both the number of uncertain parameters (dimensionality of \Theta) and the number of discrete time steps in sections 2.1 and 2.2.1 is confusing. Consider using distinct notation for these two different concepts.

We will make the notation distinct.

6. The clarity of the paragraph on lines 210-215 could be enhanced. Simplifying the text and focusing on the key points would help to make the paragraph more comprehensible.

We will simplify the text.

7. Confused by the structure – not sure why section 2.4, titled "Preconditioned Hamiltonian Monte Carlo" is a standalone subchapter. It may be more logical to include it in Section 2.3 "Numerical Methods" along with other algorithm steps.

Indeed, this should be a subsubsection 2.3.x as it is an extension on HMC algorithm with a special inner product structure.

8. Address the overall layout beginning from page 25 to ensure that the content is well-organized and visually accessible to readers.

We agree it's not very smooth at the moment. We are currently letting the text and figures flow according to LaTeX's default rules - after the final text is set we will tweak the layout and pin the ordering to get a better flow through this section.

9. The caption for Figure 8. "Posterior distributions for model M2" may be missing content. Should it be "Prior and posterior distributions for model M2" to accurately reflect the content of the figure?

We will make this change.

10. Revise the description of Figure 10 for precision; it should likely specify that the "mean discharge data was generated from the posterior predictive distribution of each model and plotted," assuming that is the intended meaning.

We will adjust this.

11. Reformulate the awkward phrasing in line 465 to correct the sentence. It should read "which implies the model can generate the data", removing the extra "is".

We will make this change.

12. Correct the reference error in line 475: “The mean log marginal likelihood is presented in Table 3” should be updated to “The mean log marginal likelihood is presented in Table 5” to direct readers to the correct table.

We will make this change.