Review of the manuscript: “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. Hale1

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

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.
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?
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.
11. What does the conclusion from Figure 10 mean? “Hence, BF penalizes models with more parameters.” How does one conclude this?
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.
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.
14. Section 3.4.2, titled “Convergence of marginal likelihood”, feels brief and incomplete, as if the discussion in unfinished.
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’.
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.
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.
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.

Technical corrections

1. Review the formatting and positioning of Equations 1 and 2e.
2. Clarify the paragraph containing lines 105, 110, and 115 to eliminate repetitive information and streamline the content for better readability.
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
4. In Equation 3, there appears to be a typo with k_{2,1}, it should likely read k_{1,2}.
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.

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