Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-350-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "Efficient Bayesian inference for large chaotic dynamical systems" *by* Sebastian Springer et al.

Anonymous Referee #2

Received and published: 31 December 2020

General comments

This paper aims to solve the high computational challenge in the statistical inference of chaotic dynamic models. This is a practical and challenging problem. It has great potential in practice. This paper combines two methods to solve the intractability of inference of chaotic dynamic models. CIL was adopted to incorporate more information of the observations into the summary statistics. LA-MCMC was utilized to reduce computational time and thus make the proposed method more practical.

The paper was written in a very good manner. First of all, the problem was introduced in Section 1. Then CIL and LA-MCMC were described in detail in Section 2. Three examples were demonstrated in Section 3. The authors have put a lot of effort to CIL and LA-MCMC, especially CIL. This makes the paper self-contained. To improve the

readability, the authors should state clearly in Section 2 which parts are novel. This helps readers to understand the contribution of this work.

It seems this paper does not match well with the scope of Geoscientific Model Development (Methods for assessment of models). It should be submitted to an applied statistical journal. This paper described the state-of-art statistical methods and implemented several experiments to evaluate the performance of such methods in the applications of geoscience. The paper provided codes for practitioners to run their own data. The paper itself is rich in information that is very useful for practitioners, but it may not be appropriate to publish in GMD journal. According to the scope of GMD journal, the publication should develop new metrics for assessing model performance and novel ways of comparing model results with observational data. I cannot see such work in this paper. It also worries me that the authors made many decisions without enough support either from theory or empirical evidence. For example, in line 201, the authors claim that their likelihood function does not depend on the initial conditions of the forward model. Is this statement true only to your forward model? In general, the paper merits publication given the following comments can be responded properly.

Specific comments

1. The statistical properties of CIL are crucial to determine the overall performance of the proposed method. Does CIL converge to the true likelihood function and in what conditions it converges to the true likelihood? In approximate Bayesian computation, the approximate likelihood will converge to the true likelihood function given that the summary statistics are sufficient and the threshold approaches to zero. What conditions are essential for CIL to converge?

2. It is not clear what novelty of this paper is. Both CIL and LA-MCMC are well developed methods. What is the contribution of this paper? In section 2, the authors basically reviewed two methods: CIL and LA-MCMC. This paper does not propose new geoscience models either. It is not clear which parts are proposed by the au-

thors and what novelty is. Please state in the paper clearly which parts are new to the literature, either in methodological or domain area.

3. Line 201: the authors claim that their likelihood function does not depend on the initial conditions of the forward model. This is a very ambitious declaration. From my understanding, the initial conditions of the forward model will impact the synthetic data and thus impact the summary statistics significantly. Please explain why your statement is true and show some evidence.

4. Line 206: the authors stated their approach can save computational time by reducing the length of simulated forward model. They only require one single epoch to compute the CIL for the later inference. The idea is beneficial to save computational resources. However, this may lead to skewed posterior distributions. The reason is that the mean vector and covariance matrix in Equation (4) are computed based on all the combinations of s and l. Normally, a synthetic data should be of the same length as the observation and Equation (5) can be computed correspondingly. The current version of Equation (5) is likely to lead to a skewed posterior distribution, because only partial comparison between the synthetic data and the observations has been incorporated into the likelihood function. Intuitively, the Figure 2 and Figure 4 have shown some skewness in the posterior distributions. Can you explain and justify your reasonings behind the line 206?

5. As the data set in the simulation of Lorenz 63 model, we should expect LA-MCMC to matches standard MCMC very well. Figure 2 demonstrates the pairwise marginal distribution of Lorenz 63 model. Does Lorenz 63 model pay more attention to the accuracy of pairwise marginal distributions? Otherwise, the authors should show the results of each single parameters, so the readers can evaluate the performance more easily. For the other models, can you show the marginal distributions of each single parameter as well?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-350,

2020.