Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2020-24-RC2, 2020

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

Interactive comment on "Application of ensemble transform data assimilation methods for parameter estimation in nonlinear problems" *by* Sangeetika Ruchi et al.

Marc Bocquet (Referee)

marc.bocquet@enpc.fr

Received and published: 29 August 2020

1 General opinion

This is a nicely written paper. It is very technical and the manuscript requires quite a few mathematical pieces of knowledge to follow the methodology! But, in my opinion, this technicality is justified. This, however, requires the notation and naming to be very clear and homogeneous. Although not too numerous, there are quite a few typos that need correcting. The numerical application is very appealing and enlightening, although not

Printer-friendly version

Discussion paper

entirely convincing because of the gap in performance between the ensemble sizes 100 and 500 (for the permeability field, not the parameters of the vein).

Overall, I believe the manuscript only requires a minor revision but that it should be very carefully addressed.

2 Remarks and suggestions:

1. Page 1: I believe that the title of the paper is too generic, not specific enough. It could suit dozens of papers already published. I strongly suggest that you revise it. I understand that this is not easy since you use a large collection of methods. Although quick to amend, I believe this point is problematic for the visibility/identification of the paper.

2. Page 1, line 2: "Kalman inversion" is not a widespread terminology, "randomized maximum likelihood" is better known, even beyond the reservoir community. See Oliver et al. (1996) and many references since then.

3. Page 1, line 4: "of the associated statistics.": I am not sure to get what you mean.

4. page 1, line 18, "a just approximation": do you mean a "correct approximation"?

5. page 2, line 28, "The main drawback of MCMC is that this approach is not parallelizable.": You know that there are parallel (multiple tries) versions of MCMCs. It actually seems that you are yourself using multiple parallel MCMCs. So I believe you should mitigate that statement.

6. Page 2, line 41-43: "However for nonlinear problems, Ernst et al. (2015); Evensen (2018) showed that in the large ensemble size limit an EKI approximation is not consistent with the Bayesian approximation.": To the best of my knowledge this

is has been pointed out first by Oliver et al. (1996). The mathematical problem has also been clearly defined by Bardsley et al. (2014), and nicely named 'Randomize-Then-Optimize". There is also a recent discussion on the issue in Liu et al. (2017), p. 2894.

7. Page 2, line 57-58: Yes, but you should at this point mention here that the idea originates from the optimal transport community, and that it is by now widespread.

8. Page 3, line 73: even though obvious, it would be better to mention explicitly that $\mathcal{N}$ is the Gaussian distribution.

9. Page 4, line 114: "Mutation" is applied mathematics Pierre Del Moral's terminology. You could briefly explain what it corresponds to in the geophysics particle filter community (rejuvenation?)

10. Page 4, line 122: "we use random walk" $\longrightarrow$ "we use the following random walk"

11. Page 5, line 135: "where C is computational cost of a forward model F" $\longrightarrow$ "where C is the computational cost of the forward model F"

12. Page 6, line 136: 'is not effected" $\longrightarrow$ "is not affected"

13. Page 6, line 150: "we seak" $\longrightarrow$ "we seek"

14. Page 6, line 160, Eq.(10): What is the definition of the norm of the random variables used in this equation?

15. Page 7, line 181: "One the other hand" $\longrightarrow$ "On the other hand"

16. Page 7, line 193: "where Z is matrix with entires" $\longrightarrow$ "where Z is the matrix with entries"

17. Page 7: It would be worth referring to the monograph by Peyré and Cuturi (Peyré and Cuturi, 2019) on optimal transport (in particular section 4), since it is very well done and freely available.

18. Page 8, line 8, "ensemble Kalman inversion (EKI) is one of the widely used algorithm." it has other (better known) names such as Randomized Maximum Likelihood (RML) and Randomize-Then-Optimize. Its sequential variant is known as the very well known EDA (Ensemble of Data Assimilation) in the numerical weather prediction/data assimilation community.

19. Page 8, line 218: "By implementing a sequential observation update of Whitaker et al. (2008),": what do you mean by this statement?

20. Page 8, line 224: "is remarkable robust" $\longrightarrow$ "is remarkably robust"

21. Page 9, Eqs.(14,15): I don't understand the intermediate member of both equations. The β or $1-\beta$ should be powers of g, and not multiply g. Or is this a notation? What did I miss?

22. Page 9, line 238-239: "This ansatz can also be understood as using the EKI as an more elaborate proposal density for the importance sampling step within SMC.": Using RML as a proposal density was already proposed and tested by Oliver et al. (1996).

23. Page 9, line 244-245: Are (x,y) horizontal dimensions or is y the depth? I believe it is worth explaining.

24. Page 10, 258: "δ Dirac function" $\longrightarrow$ "δ the Dirac function"

25. Page 10, line 262: "We assume log permeability for" $\longrightarrow$ "We assume that the log permeability for"

26. Page 10, line 267: Ok, but which type of solver did you use? (multigrid, linear algebra solver, etc.)

27. Page 11, line 272: "The grid dimension is 70" $\longrightarrow$ "The grid dimension is N=70"

28. Page 11, line 277: "The grid dimension is 50" $\longrightarrow$ "The grid dimension is N=50"

29. Page 11, lines 293-295: "Such a small noise makes the data assimilation problem hard to solve, since the likelihood is very peaked and a non-iterative data assimilation approach fails.": the explanation is very unclear to me. Please clarify.

30. Page 12, line 301: "An MCMC solution was obtained by combining 50 independent chains each of length 10^6": this contradicts to some extent the statement made about its serial nature in the introduction.

31. Page 12, line 223: 9 observation seem too few, are they? Your experiments might rely too much on the prior. I guess for reservoir or hydrological applications, there are indeed just a a few points, but they are many measurements over time at the same well.

32. Page 12, line 323: "distributed observations. which are displayed" $\longrightarrow$ "distributed observations, which are displayed"

33. Page 13, Figure 2: please add a label (β) to the x-axis.

34. Page 13, Figure 2: At $\beta = 0$ there is quite a discrepancy between the N=100 and the N=500 experiments. This could show that EKI (alone) is not working very well here. Moreover, quite often, the whiskers for N=100 and N=500 have no overlap. We would expect some overlap, would we? Do you have an interpretation?

35. page 13, Figure 3: By “Optimal” in the labelling of the panels, do you mean optimal transport, or something else?

36. page 14, Figure 4: Now, there is some consistent overlap between the N=100 and N=500 experiments, because, I guess, of the limited number of parameters (the curse of dimensionality is avoided in this case).

37. Page 14, line 333: "is lowest though" $\longrightarrow$ "is lowest although".

38. Page 15, lines 362-363: "This makes the proposed method a promising option for the high dimensional nonlinear problems one is typically faced with in geophysical applications.". Your problem do not have time dependence (does it?) which often makes many geophysical applications (like meteorology and ocean forecasting) very difficult. So you could mitigate that statement.

39. Page 16, Figure 6: What about the prior? How does it compare to the posterior?

40. Page 17: "approach provides all the desirable properties required to obtain robust and highly accurate approximate solutions of nonlinear high dimensional Bayesian inference problems.": You cannot really make such a bold statement from one (however nice) example. Please mitigate your statement.

41. General question which is worth discussing a bit: In practice, how fast is the Sinkhorm numerical solution compared to the exact optimal transport?

References

Bardsley, J.M., Solonen, A., Haario, H., Laine, M., 2014. Randomize-then-optimize: A method for sampling from posterior distributions in nonlinear inverse problems. SIAM J. Sci. Comput. 36, A1895–A1910.

Liu, Y., Haussaire, J.M., Bocquet, M., Roustan, Y., Saunier, O., Mathieu, A., 2017. Uncertainty quantification of pollutant source retrieval: comparison of Bayesian methods with application to the Chernobyl and Fukushima-Daiichi accidental releases of radionuclides. Q. J. R. Meteorol. Soc. 143, 2886–2901. doi:10.1002/qj.3138.

Oliver, D.S., He, N., Reynolds, A.C., 1996. Conditioning permeability fields to pressure data, in: ECMOR V-5th European Conference on the Mathematics of Oil Recovery, pp. 259–269.
Peyré, G., Cuturi, M., 2019. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning 11, 355–607. doi:10.1561/2200000073.

Printer-friendly version

Discussion paper