Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2018-5-RC2, 2018
© Author(s) 2018. This work is distributed under
the Creative Commons Attribution 4.0 License.

Nonlinear Processes
in Geophysics
Discussions

Interactive comment on "Ensemble Variational Assimilation as a Probabilistic Estimator. Part I: The linear and weak non-linear case" *by* Mohamed Jardak and Olivier Talagrand

M. Bocquet (Referee)

marc.bocquet@enpc.fr

Received and published: 12 March 2018

I have been following the progress of M. Jardak and O. Talagrand on this work for about five years and I am very happy to finally see its fruits.

With a very few exceptions, it is a very well written and very enjoyable paper to read. The numerical experiments are carefully designed and defined (and averaged over sufficient runs for statistical significance), a quality often missing in similar papers. The discussion, theoretical and numerical tests are sophisticated and refined, and yet, all useful. To me, this clearly indicates that a lot of energy and intelligence have been put

into this study and paper.

There are a few flaws that need addressing, some of them requiring some care. However, they should be quick to address. Hence, I recommend a minor revision of the manuscript.

The weak points of the paper are:

1. An insufficient discussion in the introduction of the fact that we already know that naive RTO / EDA / EnsVAR is not (perfectly) Bayesian and proper references to Oliver et al. (1996); Bardsley et al. (2014); Liu et al. (2017) should be added or better referred to;
2. An abstract that – very surprisingly given how well the paper is written – does not faithfully reflect the findings of the paper;
3. Some technical issues such as:
    - The use of capital letters in the titles of the sections in an inconsistent manner; please check all titles.
    - Improper use of `\citet` and `\citep`. Please check this throughout the manuscript.
    - Many inconsistencies in how you refer to equations. This should be made consistent throughout the whole manuscript. Note that there is a recommended notation for Nonlinear Processes in Geophysics (see the guidelines).
    - One speaks of and writes "the ensemble Kalman filter", not "Ensemble Kalman Filter" (which could also let the reader think that you are not very familiar with the EnKF literature).
    - Spaces in between text and ":" need to be removed (especially in the figure captions).

Specific remarks, in connection, or not, to the previous remarks are:

1. Abstract: lines 5-7: even though the authors have worked on this idea for many years, the ensemble of data assimilation (EDA) principle implemented at Météo-France and ECMWF predates the EnsVAR and conceptually coincides with it. Moreover, it could be confused with the general terminology of EnVar (Ensemble Variational) methods. That is why I do not believe that this terminology should be put forward, especially in the abstract. That said, this obviously does not lessen the findings of the paper in any way. So, it is up to the authors.

2. Abstract, line 9: "the standard variational procedure" seems too vague. I am not sure of its meaning.

3. Abstract, line 10-14: the emphasis is on the performance (accuracy) of the method compared to, e.g., the EnKF. I do not believe that this is wise in the absence of a proper cycling with which the EnKF could shine. I do not understand why the emphasis is not on the discussion of the Bayesian (or not Bayesian) trait of the method and the quality of the updated ensemble, which is the strong point of this study.

4. Abstract, line 13: It is customary to write "the ensemble Kalman filter" instead of "Ensemble Kalman Filter".

5. lines 10-20: "of the observations proper": what do you mean?

6. line 41: The EnKF definitely uses a definite article and should read: "The ensemble Kalman filter...". This should be systematically checked throughout the manuscript.

7. line 42: "by of" $\longrightarrow$ "by".

8. line 80: "The present work is devoted to the study of that algorithm, and of its properties as a Bayesian estimator..." Precisely! That is why the abstract should reflect this point instead of focusing on the accuracy.

9. line 103-104: The connection between RTO and EDA as used in geophysical data assimilation has first been made, put forward and discussed in Liu et al. (2017) (and much more[1]). This must be mentioned here. (Incidentally this is how the authors of the present manuscript became aware of Bardsley et al. (2014).) Moreover, Oliver et al. and were the first to discuss this problem in 1996 (Oliver et al., 1996), which is something that Liu et al. (2017) recalled. You must cite this reference as well.

10. line 116: "succinctly": you could avoid the adverb, as it lets the reader think that it will not be enough information to get a clear idea on the numerical results.

11. line 124: "multi-dimensional/one-dimensional Gaussian" more precisely called "multivariate/univariate Gaussian".

12. line 125, before section 2: As already discussed (and illustrated) in Oliver et al. (1996); Bardsley et al. (2014); Liu et al. (2017), RTO/EDA (hence called naive RTO in Liu et al. (2017)) produces a biased nonlinear sampling. This should be briefly mentioned in the introduction as this is an established and published fact, important to your paper.

13. line 133: "data operator": why invent a new name when there exist "observation / forward / Jacobian / source-Receptor operator" in geophysical data assimilation?

[1](i) a shorter and heuristic derivation of Bardsley et al.' derivation, (ii) a higher dimensional illustration of RTO versus naive RTO, (iii) the suggestion and illustration that RTO is not as efficient as hoped for for higher-dimensional models, (iv) the claim that EDA as used in operational meteorology is (probably very moderately) impacted by this.

14. line 149: this statement only makes sense in an infinite numerical precision context. In practice, it of course depends on the condition number of operators built with Γ and Σ.
15. line 175: "there is of course no reason to think...": again this has been settled in Oliver et al. (1996); Bardsley et al. (2014); Liu et al. (2017). So why not be more straightforward and factual here? Such as: “In general, this procedure does not lead to an unbiased Bayesian estimation, but can nonetheless provide a very useful approximation (Oliver et al., 1996; Bardsley et al., 2014; Liu et al., 2017)”.
16. line 195: since it is here linear, you should use a bold upright character for H. See also line 291.
17. line 240-251: nice remark!
18. line 266: "bayesianity" $\longrightarrow$ "Bayesianity".
19. line 291: It is worth mentioning that this time-step is 0.05 time unit.
20. line 293: Since the results do not depend on σ, why not resort to the very commonly used (in data assimilation) value of 1? Moreover, it is not the one used for the weakly nonlinear case. That may seem odd to the reader.
21. line 294: “(however, ...impact)”: no need to place the statement within parentheses. In my opinion, your remark is legitimate and important.
22. lien 364: "in non-linear and non-Gaussian situations where Bayesianity does not hold" is not a consistent statement with your current introduction (you never mentioned this explicitly even when referring to Bardsley et al. (2014)). You should have referred to Oliver et al. (1996), Bardsley et al. (2014) (with the main result clearly mentioned) and Liu et al. (2017) for this statement to fully make sense.

23. line 369: “where one day is equal to 0.44 time unit...”: I would rather say 0.20 time unit(?).
24. line 375: "succesive" $\longrightarrow$ "successive".
25. line 392: "effect ." $\longrightarrow$ "effect.".
26. line 292: It is certainly a nonlinear effect, but do you have an explanation for it?
27. line 402: “(top, middle and bottom p and accuracy” should be removed.
28. lines 411-412: “The degradation of reliability in the lower two panels may therefore be due here to non-linearity.“: this implies that you believe that naive RTO samples are not perfect draws from the conditional pdf. But, again, your introduction should have discussed this point more clearly.
29. lines 413-414: “is much larger for decrease of spatial density than for decrease of temporal density (middle and top panels respectively).“ $\longrightarrow$ “is much larger for the decrease of spatial density than for the decrease of temporal density (middle and top panels respectively).“
30. line 416: “panelsrespectively” $\longrightarrow$ “panels respectively”.
31. lines 415-416: “consistent with the top two panels of Figure 4, which suggest that the model fields are more correlated in time than in space...”: yes, a very well known fact about the L96 model (with this configuration), which is actually what E. Lorenz wanted to achieve with this low-order model.
32. lines 440-442: You might want to slightly improve this discussion since comparing 10^{-3} to 10^{-6} could let the reader think the ensemble is actually very non-Gaussian (the appendix clarifies this point, but the text should not require the appendix to be clear).

33. line 449: “We present in this Section comparison...” $\longrightarrow$ “We present in this Section a comparison...”.

34. lines 450-451: “Fair comparison is therefore possible only at the end of the assimilation window.”: yes, but not only. A fair comparison of DA methods would also imply cycling, which is not the case of EnsVAR here. I am very fine about your using the EnKF and particle filter to compare the ensemble qualities; but not really when it comes to comparing RMSE at the end of the window. At the very least this should be briefly discussed.

35. line 459: “...is the one described by Evensen (2003)”: which one? G. Evensen’s book describes both stochastic and deterministic EnKFs. (Of course I know the answer, you just need to improve the statement.) By the way, you should, from time to time, insists on the fact that you picked up the stochastic EnKF since the deterministic is now more popular. Moreover, choosing the stochastic EnKF makes sense in this study as the EnsVAR is also stochastic. You could mention this as this would strengthen your choice for the stochastic EnKF.

36. line 461: With a fully observed system and an ensemble of size $N = 30$, you do not need to use localisation. In the present context it could actually be detrimental to the quality of the EnKF ensemble! (e.g., Bocquet and Carrassi, 2017). I would recommend that you do not use localisation here.

37. lines 479-481: It is fine to report these numbers in here, but not allude to them in the abstract, where, out of context, they do not make much sense.

38. line 503: “Kuramuto” $\longrightarrow$ “Kuramoto”, as well as in both references by Yoshiki Kuramoto et al.

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.

Bocquet, M., Carrassi, A., 2017. Four-dimensional ensemble variational data assimilation and the unstable subspace. Tellus A 69, 1304504. doi:10.1080/16000870.2017.1304504.

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.

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2018-5, 2018.