Review of 'Fundamentals of Data Assimilation,' by P. Rayner, A.M. Michalak and F. Chevallier

Summary of review

The article aims to provide an outline of the fundamentals of data assimilation to biogeochemists. While the intention is good, I find that the authors have missed the opportunity of reconciling the approaches taken by different researchers in a 'review fashion', and it does not focus on biogeochemistry. Instead, the focus of the article is on the description of data assimilation from a statistical viewpoint. From this perspective, I find that the authors have some fundamental misconceptions, and that others have made presentations (see below) that are much clearer and achieve the necessary rigor. Further, while I appreciate the authors' wish to avoid a lot of mathematical notation, I find their technical presentation and the logic hard to follow on many occasions.

I hope the authors find the following comments helpful.

General comments

- The authors may not be aware of some articles that already discuss data assimilation from the same viewpoint. The most relevant is, to the best of my knowledge, that by Wikle and Berliner (2007), titled a 'Bayesian tutorial for data assimilation', which adopts the following working definition of data assimilation: 'An approach for fusing data (observations) with prior knowledge (e.g., mathematical representations of physical laws; model output) to obtain an estimate of the distribution of the true state of a process'. This working definition is the same as that of the authors of the present manuscript. Wikle and Berliner (2007) discuss Bayesian inference, the choice of prior distributions, Kalman filtering, particle filters, MCMC, Bayesian hierarchical models, and they give a lot of intuitive examples. What is the contribution of this article over that of Wikle and Berliner (2007)? Furthermore, there are other relevant works related to data assimilation from the meteorological sciences (e.g., Bocquet et al., 2010), that are not cited. Additionally, quantitative network design based on posterior uncertainties has been done elsewhere (e.g., Krause et al., 2008); see Pg.13 l.13.

- I found the omission of 'mathematically precise language' confusing, and despite the good intentions I am not convinced this is good practice. There are several problematic parts in the article. I just give a few examples:

  - Equation (2) is problematic. First, the left-hand side has to be a function of y and y^t unless one is presuming that these are known constants. Second, one usually seeks either the joint posterior distribution $p(y^t, x \mid y)$ or, depending on the scope, the posterior distributions $p(x \mid y)$ and $p(y^t \mid y)$. Both y^t and x are not directly observed. If (2) is referring to the posterior distribution $p(x \mid y)$, then one must take into account the fact that y^t is unknown. This will inevitably involve an integral over

y^t on the right-hand-side. Also, what role is $p(y^t \mid y)$ playing in (2) if it is not a function of x? (Such terms are usually absorbed into the constant of proportionality).

- Pg.5: The sentence 'x^a can correspond to a very small probability' cannot be interpreted when x is endowed with a probability *density* function. The follow-up clause 'possibly even smaller than for x^b' needs to be qualified – under what distribution are you comparing probabilities?
- Pg.8: The sentence 'Note that neither the measurement nor the true value are random variables, it is only our state of knowledge that introduces uncertainty' comes across as a hybrid Bayesian/frequentist argument. I agree that the distributions may be constructed based on state of knowledge (this is the subjective argument for eliciting distributions), but in the Bayesian framework both x and y are random variables. One may condition on an observed y to carry out inference on x, but this is different from saying that y is not a random variable.
- Equation (3) and its subsequent description are incorrect. First, I suspect that the authors wanted the left-hand-side to be a posterior (or conditional) distribution. Second, the right-hand side is the factorized joint distribution. Third, one cannot obtain an MLE from (3), only posterior inferences (except in the special case that the prior distribution is uniform).

- As outlined in the previous point, equations (2) and (3) are incorrect; further, Figure 1 is very hard to interpret (see below). I believe these misconceptions arise because the authors have not placed data assimilation into a hierarchical modelling framework as Wikle and Berliner (2007) did (although the manuscript mentions the hierarchical model once on Pg.10). The authors need to condition on a set of data for inference on the state, and I was not able to find where they do this (see also (4)). One can also view data assimilation as a state-space estimation problem which is another connection that is not made.

Specific comments

- Pg.1 l.14 and Pg.7 l.9: The authors talk about a model 'choice', but in a Bayesian setting care is needed, and uncertainty arising from the consideration of multiple models has to be taken into account.
- Pg.1 l.17: I agree that 'data assimilation', 'parameter estimation', 'inverse modelling', and 'model-data fusion', are often used interchangeably, but I thought this article should not ignore this source of confusion, rather it should take the opportunity to resolve it.
- Pg.1 l.20: It should be made clear that the model's predictive performance should be assessed on out-of-sample data and not just any data.
- Pg.2 l.30: What are x and ξ?
- Pg.3 l.13: The likelihood function, which is important to both frequentists and Bayesians, needs to be considered in such a discussion.

- Pg.6: Figure 1 (top) is misleading. The axes have arrows in both directions (what does this mean?), words are used for axes labels (I assume the 'unknown' is x, the data y, but then where is y^t?). This figure aims to instruct but it is very difficult to relate to the mathematics.

- Pg.7 l.18: Does 'Generate PDFs' mean 'Elicit prior PDFs'? Does 'Calculate the PDF for the quantity of interest' mean 'compute the posterior PDF?'. Since the authors are advocating a Bayesian approach they need to be more precise in their terminology.

- Pg.7 l.29: Jeffreys (1957) is given as a reference but it is not in the reference list.

- Pg.8 l.2: This is incorrect. The objective Jeffrey's prior is highly dependent on the parameter being inferred and the model.

- Pg.8 l.11: It is not clear what the sentence 'upscaling or downscaling properties of these statistics, for instance through correlations' is implying.

- Pg.9 l.2: The use of 'simulated quantity' in this context is misleading – I believe the authors mean $H(x)$ as the 'simulation' but it could also mean 'simulation of a random quantity'.

- Pg.9 l.9: The statement on adding the errors quadratically is both mathematically and statistically incorrect. First, one must be operating on a log scale, and second, this statement should be considering covariances and not errors.

- Pg.12: The discussion on MCMC is misleading. First, the sentence 'An advantage of the Gibbs sampler is that it can lead to a higher rate of acceptance' is inaccurate. The Gibbs sampler ensures that the acceptance rate is exactly 1 (guaranteed acceptance). Second, increased computational cost of the Gibbs sampler is not only due to the required multiple sampling, but also due to high intra-chain correlation. Finally, MCMC is *not* exceedingly robust. In fact it is quite a messy approximate inference approach, as applied Bayesians will attest to.

- Pg.13 l.10: It is incorrect that one can calculate the posterior mean without knowledge of the posterior covariance.

- Pg.13 l.14: Given the previous discussion it should have been mentioned that in a Bayesian framework one may (and should) also invoke prior distributions on the forward models, since this is usually a highly uncertain component.

- Pg.14 l.16: It is not a maximum likelihood estimate but a maximum-a-posteriori estimate. This difference is crucial in this context.

- Pg.15 l.27: It should be 'EnKF' not 'NKF'.

Concluding remark

All in all, after accounting for the works of Tarantola (2005) and Wikle and Berliner (2007) and for the authors' misconceptions, I do not see the added value of this article. It would be much more valuable to the community if it were transformed into a 'review article' of methods used in biogeochemistry (illustrating how those methods fit into a common biogeochemical framework). The title would need to be changed to reflect this and the contents would need to reflect the considerable work already done in Bayesian connections to data assimilation.

References

Bocquet, M., Pires, C. A., and Wu, L. (2010). Beyond Gaussian statistical modeling in geophysical data assimilation. *Monthly Weather Review*, 138:2997–3023.

Krause, A., Singh, A., and Guestrin, C. (2008). Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9:235–284.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia, PA.

Wikle, C. K. and Berliner, L. M. (2007). A Bayesian tutorial for data assimilation. *Physica D: Nonlinear Phenomena*, 230:1–16.