Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2017-52-RC3, 2017

Nonlinear Processes in Geophysics
Discussions
Open Access

Interactive comment on "Feature-based data assimilation in geophysics" *by* Matthias Morzfeld et al.

Anonymous Referee #3

Received and published: 22 December 2017

The manuscript presents a novel approach to data assimilation in which the assimilation algorithm is applied not to the raw data but to 'features' extracted from the data. A key insight is that one can readily construct these algorithms by devising a noise model for the features, allowing the use of standard DA algorithms.

The article makes a worthy contribution to the literature and in general I am happy with the presentation and content. I have one 'big picture' criticism which is that the problem of estimating the likelihood for features, and the solution presented in the manuscript, are not discussed except in quite mundane terms. I would like to see a few more sentences of interpretation, intuition or speculation introduced on the feature-based noise model, for instance discussing the Gaussian assumption made in the perturbed observations approach, or the nonGaussian precursor to the feature likelihood (5).

I think for the former case that the Gaussian assumption is mostly a modelling construct – but some intuition suggests that if the feature is appropriate to the parameters and the parameters are not redundant in their effect, then the likelihood function is probably unimodal, and in this case the Gaussian assumption might not be so bad. Is this intuition correct?

For evaluating (5), I wonder if a particle-based approach could represent the likelihood in more fidelity. In this case, also using a particle-based approach for the prior, would one require an ensemble of particles to represent the likelihood for each particle in the prior? Can this possibly be constructed as a Rao-Blackwellized particle filter?

Everything I am asking for is speculative – because I feel some speculation is called for. Let me emphasise that I do not necessarily want each of the above questions to be answered definitively – that would use a great deal of space. I would, however, like to hear more of the authors' opinions on the viability of alternate approaches, and realistic appraisals of the current approach.

P2, line 8-11: From P1 "The likelihood connects the model and its parameters to the data, and is often based on the mismatch of model output and data. A typical example is the squared two-norm of the difference of model output and data."

I would then expect that feature-based data assimilation would entail extracting features from both, model output and data. I think it is clear that this is done in section 4.1. But lines 8-11 on P2 make it sound like one only extracts features from the data.

P7: the noise model for the features is constructed by studying perturbations of the observations in a manner similar to the EnKF. It might be useful to note that the number of perturbed data N_z will be limited by computational power to depend on the number of observations and the complexity of the feature extraction algorithm, and that as in the EnKF the rank of the extracted covariance matrix will depend on N_z.

P7, line 16: a comment on the (non)Gaussianity of the distribution of singular vectors would be welcome, if anything is known, given that SVD is a common method for extracting features or summary statistics.

I think some extra care needs to be taken when defining the cases and effective dimension on p8. These are referred to extensively on p9 and I think the discussion is unnecessarily hard to follow.

P8: Case (ii) is simply “several aspects of the data are neglected” but what is really meant (I think) is that several important or information carrying aspects of the data are neglected, in order to contrast with case (i). I will proceed assuming this is the case; even if I am incorrect, I think a later paragraph (P9, line 11, see below) will need careful rephrasing.

P8: line 39 defines an ‘effective’ dimension, but line 33 appears to me to refer to the same object as ‘intrinsic’ dimension, which I see is the term used in the reference Agapiou et al. . .

P8, final line: “Intuitively, the more information the data contains about the parameters, the harder is the problem.”

This sentence is very confusing on the first several readings. Suppose the data consist of perfect recordings of the model parameters – then the DA problem is easy. I suppose this is a way to describe situations in which e.g. the likelihood is narrow and/or has little intersection with the prior, among many other pathological scenarios for DA schemes – but it is worth remembering that for someone unfamiliar with DA, the idea that problems arise from having very informative data is not intuitive at all but instead extremely unintuitive.

P9, line 11: “In case (ii) however, the feature changes the posterior distribution and, hence, the effective dimension. Since the feature neglects several aspects of the data, assimilating the feature will introduce a more gradual change from prior to posterior distribution than if all data are used. Thus, the feature-based approach reduces the

effective dimension of the problem. For chaotic systems, this reduction in effective dimension can be so dramatic that the original problem is infeasible, while a feature-based approach becomes feasible, see Hakkarainen et al. (2012); Haario et al. 10 (2015); Maclean et al. (2017) and section 4.4.”

I think the middle sentence must be incorrect, and I am sceptical about the third. If assimilating the feature always leads to a more gradual progression from prior to posterior (from the second sentence), and if this is the main source of the reduction in effective dimension, then how could feature DA be feasible in scenarios when the original problem is infeasible (from the third)? I think what is missing is that the computational cost of these schemes falls if the feature is lower dimensional than the data.

The loose definition of case(ii) trips us up again here – it should be made clear that some useful data is being thrown out. Second, I find that case (ii) is not established clearly enough later on in Sec 4.4. Some more care should be taken to discuss whether information, and what information, is being lost or not in the numerical example considered. I bring this point up again later but it is crucial for the discussion here that it be addressed.

P13: While examples 1,3,4 are well justified, there is not really a need to use features to resolve the LV DA problem in 4.2. This is mentioned in for instance the conclusion, but I would appreciate a motivating note at the opening of 4.2, informing the reader that this is a demonstration of the feature method in case (i).

P20-21: I suggest figure 8 come before figure 7, and be referred to on line 7 on p19 as ‘our results’. Figure 8 clearly shows the improvement in the models over the poor initial representations in figure 4, while figure 7 is an evaluation in feature space. If at all possible, the horizontal axis of figure 4 should be extended out to -150 Myr to match the scale of figure 8.

P22: The feature is well chosen to capture the influence of the parameter. What information is lost by choosing to use the feature? Can one show that the feature chosen is not a sufficient statistic, even by a heuristic argument? Again this comment harks back to my old complaints about case (i) vs case (ii). . .

P25: “The feature-based approach reduces computational requirements only if we truly reduce the dimension of the data by focussing only on some of the features of the data.”

I think this is a useful comment and in some ways the focus on effective dimension is unfortunate because it obscures this component of the discussion. I wish a comment like this appeared more clearly back on P8.

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2017-52, 2017.