Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-313-RC2, 2019

Geoscientific Model Development Discussions

Interactive comment on "Land surface model photosynthesis and parameter calibration for boreal sites with adaptive population importance sampler" *by* Jarmo Mäkelä et al.

Anonymous Referee #2

Received and published: 3 April 2019

Dear editors, dear reviewers,

I enjoyed reading this study by Mäkelä on the calibration of a new version of the JSBACH model. I find that topic and general approach fit well to Geoscientific Model Development, and that the paper has the potential to make an informative and useful contribution in this field.

That being said, I currently see two major problems with the study (detailed below), as well as a number of smaller issues that need to be cleared up before publication.

MAJOR / GENERAL ISSUES

1) The current abstract, and much of the method section, are concerned with the calibration of the model. At the same time, however, the authors make several modifications of the model (which are mostly described in the appendix), apparently in response to shortcomings that were identified in earlier studies, and present an additional case study (summer droughts in Hyytiälä) to demonstrate the improved properties of the model. As a reader, one gets the feeling that at least two studies were combined in one: i) a study about the calibration of JSBACH, with side notes about the effectiveness of the APIS algorithm ii) a study about model improvements. The case study on Hyytiälä seems to me, with due respect, a bit of a Finnish obsession – many models have problems with properly reproducing flux characteristics of Hyytiälä, I guess due the somewhat unusual soil / climate combination of this site, and although it's nice that the improved model fares better than the previous version, I'm not sure if there is a scientific reason for giving so much space to the performance of this site for a general vegetation model. To address this entire point, I would urge the authors to consider my comments, and think about whether this paper could / should be restructured, possibly by giving more space to model improvements in intro / methods.

2) The calibration procedure has several severe technical shortcomings that should be resolved. The most important is the formulation of the likelihood, which, at the moment, is not a real likelihood, but just an arbitrary cost function. You could also keep it like that, but then you shouldn't call the result posterior (as it is not based on a reasonable estimate of p(D|M)). Moreover, you should provide convergence diagnostics. If you want to make strong statements about the quality of APIS, I would recommend a benchmark against a suitable algorithm for you problem (I make a recommendation later). Moreover, I did not understand why an optimization is necessary on top of the posterior estimation. The MAP could also be estimated from the posterior sample (unless high precision about the exact location of the MAP is required)

3) Again, a bit of a broad comment, but I found that many of the conclusions are only weakly supported by the results, and also in the results and discussions, there are of

points of interpretation that seem only weakly linked to the results. Could you please make sure, throughout the manuscript, that the discussion concentrates on tangible, numerical results, and that it is clear to the reader on what results you base your interpretation (e.g. by appropriate references to tables, figures, SI)

DETAILED COMMENTS

Title: says nothing about model improvements

1.3 Name algorithm

1.7 this sounds very vague – how was performance compared, and why do you say on the one hand that there was no clear best model, and then that some models were better.

1.8 why would the improvement in the Finnish site be important?

1.10 This seems a completely unconnected question that is suddenly introduced here at the end of the abstract. See main comment 1

2.12 logical gap here – not clear why the problems named before call for species / zone specific parameterizations.

2.19 Why can this be hypothesized? To me, the only logical reason is that all models have (different) model errors, which is thus always compensated (differently) by other parts of the model. Is that your logic? If so, please make this explicit

2.25 It does not become clear why it would be necessary to study inter-site variability or the specific drought even for the questions you have posed before

2.29 This seems to contradict the abstract, where you state you use an optimizer for the optimum. But you could of course have estimated the mode via a suitable density estimator, or just take the parameter value with highest posterior within the sampled points.

2.30 I'm not sure why you provide this information at this point – an overview about the methods would have been more logical

3.3 by aggregates you mean means?

3.8 The temporal split is of course a less independent validation than a new site, but OK, why not . . .

4.3 What do you mean by "observational metereological dataset" – where the weather stations on the flux sites? If not, how far away, and is that a problem?

4.5 I'm not really sure why you would want to consider a feedback in this context, i.e. if you have climate measurements on site. Probably relates to previous question

4.8 Why do you need two citations for that fact that you don't work on a grid, but plot-based, i.e. for what are those references cited?

4.15 "fractional structure"? I think the first part of the sentence is just clutter, just say: In JSBACH, the land surface is divided into grid cells, and the grid cells are divided into tiles . . .

4.17 site-level

4.21 Although this seems logical on the first glance, it's not always clear if the "right" LAI setting for a model is the measured LAI, because models often assumes homogenous leaf distributions, but real leaf distribution is inhomogeneous, a lower-than-measured LAI will sometimes produce more appropriate photosynthesis values (cf Medlyn, Belinda E. "Physiological basis of the light use efficiency model." Tree physiology 18.3 (1998): 167-176.). It depends on the model structure. I wonder if this would better be calibrated as well, or at least I'd like to hear your comments about the assumptions about leaf distribution in JSBACH and if setting observed LAI is clearly appropriate.

5.2 You give no reasons, but I assume the modifications were done to facilitate the calibration?

5.7 Why give numbers for the groups and not a name?

5.5. The sentence is unintelligible. Moreover, the explanation of how parameter ranges (i.e. priors – why don't you call them priors) are derived is not sufficient. Provide a clear rationale for prior elicitation.

6.1 "The" lacking. In general, you are very economic regarding the use of articles.

6.3 heatsum sum

6.6 ready

7.11 You didn't define Chi, but I assume this is your prior space? Also, there is no need that this space is a subset of Rˆn (you can have discrete parameters)

7.11 Likewise, observations don't have to be continuous, thus not element R

6.15 What do you mean by directly assessed?

eq 3: the sense of the three different formulations of the right side of the formula evades me. The middle one is Bayes formula, the other two seem nonsense. If you want to define p(x|y) as l(x|y), why not define this directly. Moreover, usual notation for Likelihood is curly capital L. Same for g(x) – why first introduce the prior as p(x) and then rename it to g(x)? I also see no need for Z – if we keep on writing p(y), eq. 4 is much easier to understand

7.25 It is a VERY unorthodox notation to define pi(x) as the posterior pi(x) is often used for the prior, to distinguish it from p(x|D). I found this highly confusing

eq. 4 This seems to me a crazy reformulation of the formula, as it is so much harder to see why this holds as if you would just write the standard p(D) = int p(D |x) p(x) dx, which shows that if you marginalize the posterior over the space X, you are left with p(D).

8.30 This entire procedure remained nebulous to me. First of all, if you favor the pro-

posal mean, you should correct this in the acceptance probabilities, right?. How was this done? Secondly, when I understand correctly, you use the same spin-up (from the mean) for all parameters? I don't see how this can be justified, and how this could be corrected. What does "slightly scale" mean, do you increase or decrease weights?

9.9 If I understand correctly, you developed a new optimizer here? Why not use a well-known, tested optimization algorithm? In general, what you do here looks like a pretty standard gradient descent method. I would suggest to re-run this with an established optimizer (apart from the fact that I don't understand why you need an optimizer)

9.20 I have many doubts whether this algorithm makes sense / performs better than alternatives, and would recommend to test optima against a reliable algorithm (DEoptim package in R is very reliable for complicated target in my experience), but OK, it's probably not the main point about this paper. I just don't understand why you wouldn't fall back on standard solutions wherever possible.

10.4 The spin-up procedure seems to favor the proposal mean and could thus distort the posterior. Please discuss

10.9 I'm not sure if I understand correctly – you are applying a KDE on the sampled posterior, and then create samples from that for the posterior predictive distribution? Why would you do that? Would that (potentially) distort the posterior? Please discuss and if you do what I think you do, prove that this does not distort the posterior.

10.13 What's the sense of this effectiveness? It seems this is something like sensitivity, but this could be calculated directly from the difference prior – posterior. Moreover, when I understand correctly, this is a kind of conditional calculation, where you keep the other parameters at the optimum? Makes kind of sense, but is also loosing info about the parameter correlations in the posterior, so in theory, a parameter could be very "effective", despite being globally poorly constrained (due to a trade-off with another parameter). Please discuss.

10.25 Based on your exposition, you should define a likelihood, and not a cost function. This word has no meaning in a Bayesian context.

10.25 If this is a likelihood, correct interpretation would be that likelihood is normal

10.26 a) what do you mean by “successfully done” – that it went through the review? please give a reason for this b) so, the cost function is NOT the MSE

10.29 Not clear to me what you mean by “covariance vector”, and “combining model and observation error”. You don’t know the model error as such, but of course, as in a linear regression, you can fit the sd of the normal distribution to the effective spread around the predicted value, which is the standard approach.

eq 9 – OK, if you want to define this as your likelihood, then you should simply state the correct assumptions. What you assume here is that the relative error is normally distributed, with a standard deviation EQUAL the mean, and additionally you divide the likelihood by the number of data points (dividing by N_ET), which makes no sense if you truly want to create a likelihood. Why does this not make sense?

a) you don’t know a priori how your residual scatter around the model predictions. The scale of the normal distribution (essentially the denominator in the likelihood) affects the shape of the posterior, i.e. makes it more or less wide. As you can’t know the correct scale, you have to fit a parameter here

b) also, it doesn’t make sense to have residual go to zero for small observations. A sensible expression for the scale of the normal would be to fit

log likelihood = (predicted – observed / (a0 + a1 * predicted))ˆ2

where parameters a0, a1 have to be optimized.

c) there is no good statistical reason to divide by N_ET, i.e. by using a mean squared error as the likelihood. Essentially, by doing this, you scale the likelihood to have the evidence of one data point, making the posterior much wider than it would naturally be

In general, it seems to me that the likelihood you use here creates a posterior that is far wider than any sensible statistical assumptions would allow.

11.8 I found the structure with results and discussion together not very helpful. It seems to me that this is adding to the fragmentation of this paper, which seems to address several questions (model improvement, calibration, drought case study) at the same time. A discussion which summarizes the results and puts them in a common perspective would have seemed preferable to me

11.13ff it seems you suggest in this paragraph that identifiability equals or is related to convergence, and it’s not clear to me why (in general, these are two different issues). Moreover, I can see no visual difference in convergence speed between the three examples. If you claim the first one converges faster, please back this up by numeric estimates of convergence, e.g. Gelman-Rubin.

11.13 Moreover, you should provide convergence diagnostics / proof of convergence (typically Gelman-Rubin) for all your results! Not having checked convergence is not acceptable.

11.16 It’s not the algorithm that is unable to constrain the parameter, it’s the likelihood. Also, you seem to suggest that this is a problem, but that’s perfectly normal for a Bayesian analysis.

11.20 What does reasonably stable mean? See comment about convergence diagnostics above

11.24 Again, not really clear to me why you do the optimization in the first place, instead of estimating the MAP from the posterior sample

11.25 Near or at. Why is near the limits a problem? If you have flat priors, you state that all these values are equally likely, so near limit is no problem. I suspect though that you have MAPs at the limit, posterior medians are of course never at the limit.

2.8 You can parallelize the chains in DREAM, which means that, assuming you run 3

MCMCs as usual, you can use at least 9 cores. I'm not sure how many cores you were using. On the plus side, I'd bet that DREAM or DEzs algorithms converge faster than APIS. I think it would be useful to benchmark against one of these algorithms. Both are implemented in the R package BayesianTools.

12.20 So, why not calibrate them right away?

12.20 / 12.26 Most of the info in these paragraphs are not results

16.22 The entire section reads like an independent case study with its own methods, results and discussion.

19.8 Comments about APIS: I could not see a serious evaluation of the convergence and quality of this algorithm in the paper. At least, you should provide convergence checks. If you want to say anything about the quality of APIS, I think you should compare to a reasonable referece. For example, DEzs or DREAMzs in the R package BayesianTools would be suitable reference algorithms that have proven to work well for these kinds of problems.

19.10 Define successful.

19.10 General comment: for any claim you make in this section about your findings, please refer to a specific result in section 3 that is the basis of your claim. Specifically, I can see no results that provide hard support for your first two claims.

19.28 Code and data availability is insufficient. Unless there is a good reason against this, please provide all code and empirical data (drivers and calibration / validation) with the paper, or in an appropriate repository. FLUXNET could be updated or changed, and in any case, it would be more convenient for the reader to have your entire data set at hand. Moreover, you should ensure computational reproducibility, but storing random seeds etc. for the algorithms. Ideally, also results, in particular MCMC chains should be saved, if space permits this.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-313, 2019.