Comment on gmd-2021-273

Florian Hartig (Referee)

Referee comment on "Calibrating the soil organic carbon model Yasso20 with multiple datasets" by Toni Viskari et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-273-RC2, 2021

Viskari et al. present a (re-)calibration of a slightly modified Yasso SOC model. The authors highlight in particular that calibrating to multiple data streams lead to better global model performance, and that the calibration was more successful with advanced population-based MCMC algorithms, compared to more traditional Metropolis-Hasting samplers.

While I agree with the remarks of my colleague (Rev 1) that the results regarding the superiority of multiple constraints and population MCMCs by themselves are rather unsurprising, I come to a different assessment regarding the overall novelty of this study. I see the issue highlighted by Rev 1 mainly in the presentation of the results, which concentrates in my option too strongly on generic technical aspects of the calibration, instead of stressing the improvements to the Yasso model that are facilitated by this calibration.

More specifically, my understanding is that the model used in this study is a previously unpublished improvement of the popular Yasso model that is calibrated and to some extent also validated in this study. To me, this seems valuable / novel, but this value would be more easily seen if the authors could better highlight the resulting benefits for SOC modelling. Moreover, if model improvements take a more central role of this paper, I would re-consider the decision to not compare the performance the new (calibrated) Yasso20 model with the older Yasso07 model – it seems to me that an improvement in the performance of the model would be a great argument to counter the novelty concern of Rev 1.

Regarding the comments of Rev 1 that a global SA is necessary before calibration, and that the likelihood should be weighted: I think both are good points that should be considered, but I also think that the approach taken by the authors is not necessarily

wrong. Performing an SA prior to calibration has the main purpose of reducing the number of parameters in the MCMC, which speeds up calculations. If the authors manage to calibrate their model despite not performing an SA, I don't see a problem. The topic of weighting is a bit more tricky: statistically, arbitrarily-reweighting data is difficult to defend (although this is widely applied). As we show in Oberpriller et al. (2021) Ecology Letters, if the model is 100% correct, weighting has no benefits to the calibration. However, as we show in the same paper, if there are systematic model errors, weighting can be beneficial for obtaining reasonable fits when data are strongly unbalanced and weighting is done appropriately. As these conditions may be met here, I would agree that the authors could experiment with whether re-weighting the data improves model performance, however, I wouldn't say that re-weighting is a categorically better or absolutely needed.

In summary, I think that this is solid model calibration study. There are a few minor technical issues that can be found below, the most critical is probably that I would recommend also calibrating the error terms in the likelihood. Also, the issue of re-weighting could be considered. To clarify the novelty of the study, I would recommend to re-structure the presentation around the overall goal of model improvement and useability (e.g. that you show how to perform quick / efficient calibrations for this particular model). The question of multiple data-streams and sampler choice is interesting, but with the focus on model improvement, it would take a more supportive role in the overall story. Also, while being perfectly intelligible, I believe that the general conciseness / flow of the text could still be improved.

Best wishes

Florian Hartig

=== Detailed comments: ===

Title: consider erasing „the impact of" (conciseness)?

10: "tools in determining" -> to determine ?

16 erase "the"

21 In terms of the logical flow, I would recommend starting with the topic of soils here, as this would allow you to move more naturally to the models at the end of the paragraph (as opposed to the current structure, which is going models -> soils -> models)

21 FOR estimating?

42 The third challenge feels a bit like an add-on. More generally, I wasn't convinced about the sense of classifying these three distinct challenges, because they are (as you note) connected. Maybe it would be easier to state something along the lines that there is evidence that we should add more complexity to the models (e.g. nonlinearities), but that empirical (data availability, spatial variation) and methodological challenges (data assimilation) have so far hindered successful expansions of model complexity.

46 The logical flow / connection of this new to the last paragraph is hard to grasp. That you need multiple data streams was already stated in the last paragraph (challenge 2). Possibly, you could remove this from the previous paragraph and say here that calibration challenges could be addressed by combining multiple data streams. In this context, the comments in the intro of Oberpriller et al., 2021, https://onlinelibrary.wiley.com/doi/full/10.1111/ele.13728 may be of interest.

86 It sounds here as if you refer to the model + calibration protocol as Yasso20, but below (114) you refer to the model alone as Yasso20. I think you mean the latter, right?

115 I realize this information is provided later, but I think it would help the reader at this point to have one sentence that clarifies how Yasso20 differs from Yasso07. Also, clarify in the description that follows whether descriptions refer to the Yasso07 or Yasso20.

164 On the github repo that you link, there also seems a Yasso15?

Table 1,3,4: These are quite long, maybe some of these could be combined, presented visually or moved to the supplementary? I think the main text should concentrate on the central messages of the paper.

221 It seems that you assume in this section (I also looked at the code to make sure) that data uncertainties (i.e. sd in your likelihood) have to be fixed a priori from the data. This, however, is rarely a good idea. Even if you know the observation error perfectly, there can be other reasons for your model to deviate from the observed data (e.g. model error, or some variability in the environment that has nothing to do with the observation process). Consequently, if you fix the sd in the likelihood based on your observation uncertainty, you will get wrong (typically too narrow) posterior distributions. I would highly

recommend calibrating with variable sds. If you want, you can set priors to reflect your data uncertainties, but you should give the calibration a chance to correct those if necessary.

290 In general, the calibration part provides very little info on the most crucial part of the calibration, which is the likelihood that you calibrate. The details on the algorithms are useful, but this could also go in the appendix.

333 Different or not converged? It seems not converged, but then you shouldn't interpret it and just run it longer.

354 Via the arguments in Oberpriller et al., 2021 (cited above) these observations could be interpreted as a hint of systematic model / data error. Would you agree? Possibly to be added to the discussion.

375 Here, but also other sections: I think it would be helpful for the reader if you would re-state at the beginning of each result what the purpose / motivation for the respective result was. This section, for example, doesn't trivially connect to a research question of yours, nor is it mentioned in the methods that you would look at this, so you should give the reader a bit of context. I also wonder if the correlations wouldn't better be mixed with a discussion of the mean parameter estimates, which seems to be missing. Such a discussion would imo logically be better placed BEFORE the discussion of predictive performance, but after the discussion of the calibration performance (convergence)

385 As I said in my general comments, I believe you should put the model improvements (i.e. updated parameters, improved performance) in the center, and present the results about MCMC algorithms and multiple data streams rather as a byproduct.

394 Maybe I missed it, but did you show that multiple maxima were the problem? In my experience, trade-offs between parameters are a far more common in the context that you consider here

396 OK, but isn't that to be expected?

404 I also don't know, usually one would presume these two methods to be very similar.

420 One could also read your results as showing that you should directly include prior information if you have it, or else you might get a worse / non-sensible result.

435 You could of course also think about re-weighting the different data streams. It is not necessary to cite it here, but I think the discussion in Oberpriller et al., 21 could be useful in this section.