General author comments

This document contains the executive editor and reviewer comments, and our responses to these comments, regarding the discussion of: https://www.geosci-model-dev-discuss.net/gmd-2018-313/

The editor and reviewer comments are shown here as indented text. We have added numbering to certain comments, so they may be more directly referenced, if needed. Our responses are shown after each comment as unindented text (such as this).

Based on the comments, it seems that the APIS algorithm description was not clear enough and needs both clarification and revising. We will add a comparison to a basic multichain MCMC, to highlight the differences. We will add information about the convergence of the algorithm, as discussed in more detail when answering the specific comments. We will also separate the “Results and discussion” into two different sections.

The main differences between APIS and a basic multichain MCMC sampler will be outlined in the methods section along the following lines:

1. In our simulations APIS is set up as 40 simultaneous, independent IS samplers that have their own prior distributions and locations. This is similar to a (simultaneous, independent) 40-chain MCMC sampler, where each chain corresponds to one IS sampler. The priors in our simulations are truncated Gaussian distributions, with initial locations randomly sampled from a uniform distribution that is defined by the ranges of each parameter. The deviations were also randomised.
2. At each iteration (this is called an epoch, but we have tried to avoid the use of the name), we take 50 draws with each of the 40 IS samplers (altogether 2000 draws). In basic MCMC we would take one draw for each chain (40 altogether).
3. The location and shape parameters for each individual IS sampler is then updated (based on their “own” draws) – in APIS the new values are automatically accepted (this is called blind adaptation). MCMC chains, in contrast, would deal with acceptance probabilities etc. and accept/reject the draw accordingly (also possibly adapt the prior distribution at certain intervals).
4. Additionally, the APIS global estimates of the parameter expected values are updated (using deterministic mixture and all samples).

Each individual IS sampler (at 2) generates a sample of 50 draws which, reweighted according to the cost function, form an estimate of the posterior distribution. The parameter distributions, that we have presented, are formed from the location parameters of these individual IS samplers. They do not represent the _true_ posterior distributions, rather they are a collection of estimates of the mean. These values are expected to be around all the modes of the target. The deterministic mixture ensures the stability of the estimation of the parameter (global) expected values (4 above).

We originally finished the APIS simulations about 9 months ago. At that time we had ran the algorithm for all conductance formulations for the duration of 100 epochs. Afterwards we discovered a coding error in the Ball-Berry conductance formulations (affecting all of these models). As a result of this error, these simulations were wrong. They were, however, very indicative of the parameter distributions. After the error was corrected, we ran the APIS sampler for the BB model and verified that the resulting distributions were similar to the ones before. Because of this, we did not run the APIS for the other models, but ran the optimiser from the end state. For the sake of comparability, the optimiser was also run for Baseline and Bethy simulations.

In the APIS descriptions we followed the same notation as in the paper originally describing the algorithm. This was done, so it would be easier for readers to refer to this paper. Since this seems to be only confusing, we will revert to the more common notation (when applicable).

Executive editor comment

- So in your case, JSBACH must be included into the title of the manuscript. Additionally an identifier / version number indicating the exact version of the code must be added. Note, that the code modifications you are discussing in your manuscript need to be made available. Especially note, that the exact version of JSBACH used in your manuscript needs to be permanently archived. The information how this is achieved need to be added to the code availability section.

The JSBACH model is version controlled (svn) and the information about the model branch and version number will be added to the code availability section. However, we cannot grant access to the model, as it is under the Max Planck Institutes License agreement. Our modifications have been made mostly within the existing model structure, so the separation of original model and our modifications is not meaningful. We will archive our modifications e.g. in GitHub. Accessing these modules will require acceptance of the MPI-M License agreement (as the modifications are made within the model code), after which access to the modified modules can be granted on request.

Comments by Reviewer 1

Overview:

The authors apply adaptive population importance sampler and a simple stochastic optimization algorithm to optimize the parameter sets of six different stomatal conductance models using measurements from 10 FLUXNET sites. For the validation, the experiment period is split into the optimization period and validation period at the six study sites. The remaining four study sites were used only for the validation. The reproducibility of GPP and ET was investigated with the optimized parameters. For the drought event at one flux site, the effectiveness of additional parameter optimization for water use efficiency was also investigated.

The results indicate that the optimization scheme presented in this paper successfully improve the estimation of GPP and ET, even for the drought event. The model was also modified to use a delayed effect of temperature for photosynthesis activity.

General comments:

Parameter optimization is essential for the model development. The methods proposed in this paper successfully optimize the parameter sets which control carbon flux or water flux. The procedure of this paper seems generally adequate, and I think the paper is relevant for GMD. However, the manuscript is needed to be improved from the two aspects:

- The readers of this paper may not understand and reproduce the experiment because some procedures in the paper are not clear. n addition, the descriptions are sometimes too much redundant or too much simple, and the argument becomes unclear. The authors need to improve he manuscript carefully according to the specific comments.
- The authors indicate that some of the settings are not appropriate for the USO model to simulate the drought event. Nevertheless, the authors concluded that the estimation with USO is one of the “best” results. The authors should run the experiment again with the appropriate setting if they would like to use the USO result for the discussion.

We agree with the first general comment and will improve the manuscript. The second comment we disagree with as our message of how appropriate the situation is, may have come across too negatively. In the simulations, one of the parameter bounds for the USO model does not properly reflect the values in the literature reference (Medlyn et al. 2011). This is true for both optimisations, not only for the drought. This

parameter is not optimised to the boundary (in either optimisations). Considering also, that the “same” parameter (g1) for the other conductance formulations has been unimodal, we feel that rerunning these simulations with a smaller lower bound is not worth the effort.

Specific Comments:

(1 is related to general comment 1, and 2 is related to general comment 2)

1-1. P2 Lines 3-8: The explanation for soil drought is confused. I think it is better to explain “general” soil drought first, and then emphasis the importance of soil draught at the boreal forests. This section is important to explain why the authors chose boreal sites for the experiment. I also do not understand the sentence “reversing the development”.

We will re-order the sentences describing the global and Boreal relevance of soil moisture drought. “Reversing the development” connects to the start of the sentence and to the “recovery of photosynthesis”. We will reformulate these sentences.

1-2. P2 Lines 19-20: I do not understand the sentence; “However, it can be … conductance formulations.”

We hypothesise, that the choice of conductance model will affect the optimal values of the parameters that are not conductance model specific (so it is not enough to only optimise conductance model parameters). For example the maximum carboxylation rate (Vc,max) has different optimised values, when using the different conductance formulations. We will amend the sentence.

1-3. P2 Lines 25-27: “We will assess the inter-site variability ...one site.” and “We will provide an assessment of ... descriptions.” are about the validation. I think it is better to explain the validation process explicitly starting with e.g. “The validity of the optimized parameters is assessed ...”. The explanation should include the two points:

1) At the six flux sites, the experiment period is split into the optimization period and validation period, and the reproducibility of GPP and ET with the optimized parameters were investigated. The remaining four sites were used only for validation.

2) The drought event at one flux site is also investigated with some of the optimized (fixed) model parameters and with additional parameter optimization for water use efficiency.

These comments are more related to the methodology of the paper, not the introduction, and we will take them into account in the proper section of the paper.

1-4. P2 Lines 27-28: I think the sentence about the optimization method, “We utilize the adaptive population importance sampler ... (peak of high probability).” is too much short. One paragraph may be needed to explain this part. Many optimization schemes have been used for land ecosystem models. Therefore, it is better to review some of them, and the authors should explain the difference between APIS and well-known methods (e.g. MCMC and some other optimization methods). There are also several studies which estimated model parameters using flux site measurement. Therefore, advantages of this study also needed to be explained comparing to those previous studies.

We have moved here the comparison to other algorithms, from the start of “Results and discussion”. These paragraphs will be also modified to better suit the new context.

1-5. P3 Lines 2-5: I do not understand the procedure clearly. - What is the time resolution of the model? - What is the difference between “half-hourly” and “daily” timeseries? How these different time series are used in the experiments?

The difference between the time series is that the other consists of measurements at every 30 minutes and the other consists of daily values. The model does not have a fixed time resolution, and it can be run with both datasets. We used the half-hourly dataset to quality check the daily values (as described in the paper). The model is run with the half-hourly dataset, but we examine the output (GPP and ET) at the daily level. We will modify the text accordingly.

1-6. P3 Lines 7-10: I think the explanation is confused. - Please explain the experimental setting more intelligible and clearly (refer to 1-3 of my comments), and the detailed explanations written in these sentences should be added. - "(with measurements separated into successive time period)": How many years are used for optimization and validation respectively?

Detailed information will be added.

1-7. P5 Table2: Definition and range of g1: "Table3" instead of "-" may be better to understand. I could not find the explanation about initial distributions for these parameters. How do the initial ranges in Table 2 relate to the black lines in Fig. 1?

Changed the "-" to "Table 3". The ranges are the absolute minimum and maximum values for the parameters in these simulations. The initial locations for each IS sampler is uniformly sampled from within this range of values. The black lines correspond to the initial distribution of the IS sampler initial locations.

1-8. P6 Lines 7-8: "However, coniferous evergreen trees do not ... following spring." I do not understand the connection before and after this sentence.

The spring event in the model determines when leaves can start to grow. Conifers do not shed all of their leaves (needles), so this "spring event" timing is not that important for them. They already possess considerable amount of leaves in spring when temperature rises and plants start to photosynthesise. In addition, the start of the photosynthetically active season for conifers in JSBACH has been observed to occur too early. So we try to amend the situation by adding the delayed effect for photosynthesis. We will amend the sentences to be clearer.

1-9. P7 (2.4 Parameter estimation) – P11 (2.9 Cost functions): I do not understand the procedure for the parameter optimization clearly. I think it is better to add the section for "overview of the experimental settings" between section 2.3 and section 2.4. Then, Section 2.7 may better to be included in this overview section. The procedures for initial parameter settings, spin-up, parameter optimization, and validation should be easily understood. A process diagram may help the readers to understand.

Especially, I do not understand the relations between the "parameter estimation" by adaptive population importance sampler (section 2.5) and "parameter optimization" by the simple custom stochastic optimizer (section 2.6). - In my understanding, the APIS is used to estimate model parameters roughly. Then using the estimated distributions by APIS as the initial state, parameter optimization is done. Is that correct? The overview section should include more detailed explanation for this point. Introduction (P1 Lines2-3) also needed to be corrected so that the readers can understand the procedure. - How many years observations are used for these two different optimization methods? Are the observations for the two optimization methods same? - I do not understand how to merge the optimizes parameters as shown in Table4, because in my understanding,the parameters are optimized separately at each study site. It is also necessary to clarify the role of "2.8 Simulation analysis" and "2.9 Cost function" at the overview section. In my understanding, the first half of 2.8 indicate the parameter sensitivity, and the latter half of 2.8 indicates the validation of the result using the observation. Cost function (2.9) is used both for

parameter optimization (APIS and a simple stochastic optimizer) and for validation of the results (e.g. Table 4, Table 6). These descriptions also should be included in the overview.

We will add an overview section comparing APIS procedure to a general MCMC chain with a bit more in detail than the one at the start of this document. Essentially you have understood the process correctly – APIS is used to estimate the parameter posterior distribution and the stochastic optimiser starts from there. Both methods use the same data for driving and optimising the model. In the general optimization, we take the mean of the site level cost function values and use this for APIS and the optimiser (for each conductance model separately; the parameters reflecting these are reported in Table 4). The dry period optimisation uses only data from Hyytiälä 2006 and the corresponding values are presented in Table 6.

1-10. P8 Line 23: What is “M=40 proposals”?

The number of IS samplers in APIS. We will add this information to the overview.

1-11. P8 Line 30 – P9 Line 3: I do not understand the procedure (see, 1-9 of my comment).

At each iteration, we run the spin-up for the model only for the IS sampler mean (and use the same spin-up for the other 49 samples). This is done to speed up the process. This information will be added to the overview.

1-12. P9 Lines 30-33, P10 Line2: I do not understand these sentences (see, 1-9 of my comment).

This is essentially the same thing as above. The spin-up is needed to get suitable starting conditions for the model by driving some state variables (e.g. soil water content and LAI) it into a (semi) steady state. Since in APIS we are mostly interested in the means of each IS sampler, we use the same spin-up for the other samples.

1-13. P11 Line 29: What is “sampling limits set”?

The quotation should be “sampling limits set for each parameter”. These are the limits (the range) given in Table 2. We have modified the sentence to better reflect this.

1-14. P12 Lines 1-18: This may be better explained in introduction (see my comment 1-4). The parallel mode is not used in this study, therefore this advantage (parallel simulation) is not suitable for this study setting.

These paragraphs have been moved to the introduction. The optimisation has been run in parallel mode and we consider this critique to be unqualified.

1-15. P12 Lines 29-31: “The actual soil moisture ... unreliable and even unrealistic.”:Then, what is the recommended setting for the future study? Is it OK for this experiment?

These sentences refer to the optimisation of the parameter q in the general setting. It takes effect only with very low soil moisture values that occur rarely in the optimisation, hence the amount of data affecting the optimisation of q is very limited. Therefore any values of q should be viewed with reservations. The situation is slightly improved during the dry period optimisation, but the differences in q with different conductance formulations are quite large.

1-16. P14 Line 25 – P15 Line 2: Some descriptions are redundant. Improvement of description is needed so that the readers understand the Fig. 2 clearly.

We will separate the “Results and discussion” into two sections, which should improve the readability.

1-17. P16 Line 9: “We optimized the model for individual (calibration) sites as well.”: I do not understand “as well”. I thought the model was optimized at the individual sites.

The general optimisation was done for all sites simultaneously in order to preserve the generality of the model (on Boreal coniferous forests). We will try to make this explicitly clear.

1-18. P16-P19: I do not understand the arguing point in this section (3.3). I think it is better to explain Table 6, Fig.3 and Fig. 4 first, and then more detailed discussion should be done. - P16 Lines 26-32: Too much detailed and complicated. First, the categorization of the optimized (fixed) parameters and the parameters for further optimization (for WUR) should be explained using Table 4 and Table 6. What is the most important different between these parameter groups? The detailed settings for the fixed parameter may better be explained in Appendix. - P16 Line 33 - P17 Line 4: This paragraph is important to describe the parameter optimization for the drought event. How many years WUR optimization was done? Is the optimization procedure different only for cost function calculation? Are the observations for year 2006 repeatedly used?

We will take these into consideration when we rewrite the discussion. The order will be affected by the division of “Results and discussion” into two separate sections.

1-19. P17 Line 16-18: The parameters are just optimized in this experimental setting, and the “true value” is not known. Therefore, I think “optimal value” should not be used here. Authors can just say that “the optimized parameter set for WUE greatly improved the simulation results (Fig. 3)”.

The “true” value is likely never known. Every optimisation experiment is situational and must be considered in the context of the optimisation. We will take the revised sentence into consideration.

1-20. P17 Line 14 – P 18 Line 18: The detailed explanations for each parameter are too much complicated and I do no understand. The paragraph of P18 Line 19-26 should be placed before these paragraphs. Then, the relationship between the results in Fig.3 and the estimated parameters should be discussed as below: - Which parameters are the important to control WUE in this experiment? - How do these parameters affect WUE? - Are the estimated parameters reasonable compared to the previous studies?If not, why?

The latter paragraph concerns the plant water use efficiency (WUE) during the drought. We do not calibrate the model with WUE, but with ET and GPP. Hence the results concerning WUE cannot precede the discussion on parameter values or of the ET and GPP fluxes.

1-21. P18 Lines 22-23: I do not distinguish “the actual drought” in Fig. 3. I think it is better to add the period of the drought in the Fig. 3.

This will be added.

1-22. P18 Lines 27-33: I do not understand this paragraph because I do not understand Fig. 4. Does the lower panel show USO results? What is “Medlyn”? I also do not understand how is the β-function for the observation calculated.

Unfortunately this image contains the surname of the first author of the USO paper. This will be fixed. The lower panel are the USO model results. For the observations the coloring is mentioned in the brackets (Bethy dry or Medlyn dry) and in the image text. We have used the same intensity as in the middle column, applied for the corresponding day.

1-23. P18 Lines 34- P19 Line 5: The authors did not show the experimental setting and result explicitly, therefore I do not understand the purpose of this experiment. What is the difference between this experiment and the parameter optimization in section 3.3?

Within these lines we explain, that we examined the ET and GPP cycles with all conductance models and all sites with the generally optimised parameter set (all sites simultaneously) and the dry period set (only Hyytiälä dry period optimisation). So the optimised parameters are the same, but we expand the examination to all sites (so we check the ET and GPP cycles of other sites with Hyytiälä dry period calibration). This paragraph

highlights, that in general, the site level optimisation is poor, when applied to other sites and compared to the more general optimisation. We will clarify this at the beginning of the paragraph.

1-24. P19 Lines 17-20: I do not understand these sentences.

The parameter q only affects the model output, when soil moisture is below the fraction θ tsp. Because this fraction was lowered, q is practically ineffective (this relates to comment 1-15). The dry period optimisation raised the fraction θ tsp, so q was again effective. We will amend the sentences.

2-1. P17 Lines 7-13: The authors explain that the setting for USO is not appropriate. Then the results should not be used for further discussion after this paragraph. If authors would like to use the result, they should perform the experiment again with the appropriate settings.

We disagree with this comment, as we have explained above when answering the general comment concerning this topic.

Technical corrections:

1. P1 L11: “correctly time and replicate” -> “correctly reproduce”

Changed.

2. P2 Line 27: Abbreviation “APIS” should be placed here (this is the first appearance).

Added.

3. P9 Line 5: “from 2006” -> “in 2006” (only one-year optimization).

This is an incorrect correction, but we restructured the sentence.

4. P10 Lines 17-18: “high”, “average”, or “low” effectiveness value: this explanation should be the same as Table 4.Added “change in the parameter values”.

Modified.

5. P11 Lines 19-20: Description of the Supplemental materials is needed at under each figure.

These will be added.

6. P15 Table5: some values are different form the Fig. 2 (i.e., r2 of Bethy).

We checked and noticed, that this is indeed so. The values in the images are automatically calculated, so they are correct. These will be corrected.

7. P18 Line13: “disregardin” -> “disregarding”

Corrected.

8. P17 Line 14 “The most noteworthy” what? (modified word is needed)

Added “change in the parameter values”.

Comments by Reviewer 2

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), aswell as a number of smaller issues that need to be cleared up before publication.

- 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.

One of the main reasons, why we have focused on the boreal evergreen forests, is to better understand the model deficiencies in replicating drought conditions. The stomatal conductance models formulate the plant response to these conditions. Because the models differ from one another, it is also important to calibrate the parameters accordingly. Therefore it seemed redundant to separate these aspects into two different papers and more beneficial to compare the model performance under drought conditions to the more general setting.

We understand the criticism towards the case study, but disagree with the implication. Contrary to the reviewers view, we have the experience that most models can replicate the Hyytiälä site flux characteristics quite well. The "case study" was made at Hyytiälä, because of the exceptional drought of 2006. Regionally it resulted in visible discoloration of needles, vegetation dying etc. that have been reported in Muukkonen et al (2015). Furthermore, the drought is visible in the Hyytiälä eddy covariance data in a manner we have no known record from any other Boreal coniferous EC site. The reason for examining the drought event, is because general vegetation models tend to run into problems when replicating droughts. These events are important to examine for the benefit of model improvement. It has also been speculated, that these events are likely to increase in frequency in the future. However, we will give more emphasis on general model improvements.

- 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)

This comment has a lot to do with the actual formulation of APIS and how it works, which apparently has not been conveyed well enough in the paper (as noted at the start of this document). We will clarify the descriptions. We do agree on the interpretation, as we have systematically called the objective function a "cost function". The word likelihood only appears when we are referring to the general Bayesian framework. We will reformulate the wording regarding the posterior pdf's. We will also add convergence diagnostics of the APIS parameter expected values. Benchmarking the algorithm against other algorithms is, however,

beyond the scope of this paper. The reasons for the use of an optimiser has been presented at the start of the document.

- 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)

We do so in the revised manuscript.

Detailed Comments:

Title: says nothing about model improvements

1. 1.3 Name algorithm

We have added the abbreviation here. The modification (shape parameter adaptation) itself is not enough to warrant a new name for the algorithm.

2. 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.

This relates to both the cost function values (of optimised parameter sets), but also to site specific correlation and bias. The differences are too small to make definite statements about the best stomatal conductance formulation, but the individual metrics indicate better performance for some. We will improve the abstract in this regard.

3. 1.8 why would the improvement in the Finnish site be important?

The improvements are not important, because the site is in Finland, but due to the specific drought conditions we are interested in.

4. 1.10 This seems a completely unconnected question that is suddenly introduced here at the end of the abstract. See main comment 1

Yes this does seem to be a disconnected addition and we will improve the motivation.

5. 2.12 logical gap here – not clear why the problems named before call for species / zone specific parameterizations.

The drought responses are (usually) extensively generalised processes utilising bulk parameters. Soil moisture drought and vapour pressure deficit (VPD) affect different plants differently under various environmental conditions. In this paper we examine the drought effect on one plant functional type (PFT, not a species). It is much easier to examine the effects without the added complications of many PFTs.

6. 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

This is broadly the same argument as ours and we will make this more explicit.

7. 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

Both of these questions are important, when we consider model caveats and deficiencies. Especially the droughts represent conditions, where many land-surface or ecosystem models fail to correctly replicate the observations.

8. 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.

The end state of the APIS algorithm reflects all the modes of the target, so there is no contradiction here. The latter remarks are true but could not be directly used with some of the simulations (as explained at the start of the document).

9. 2.30 I'm not sure why you provide this information at this point – an overview about the methods would have been more logical

Will be moved.

10. 3.3 by aggregates you mean means?

Usually yes, but this depends on the units of the variable. In some instances, rainfall can be given in just mm without specified time (so in this case the values would be summed).

11. 3.8 The temporal split is of course a less independent validation than a new site, but OK, why not…

The independent validation sites are always valuable, but in many cases we are dealing with a limited amount of available data. In these cases it is usually more beneficial to split the timeseries into separate sections.

12. 4.3 What do you mean by "observational meteorological dataset" – where the weather stations on the flux sites? If not, how far away, and is that a problem?

All of the measurements are taken at the sites themselves (not including the specific data we mention in the manuscript). By meteorological we mean that these are site level measurements of the (current) meteorological conditions. Usually eddy flux sites have instrumentation that is at least on the level of typical weather station.

13. 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

We are pointing here to model deficiencies, which may affect the results. This is a known problem with JSBACH (uncoupled) simulations especially when run with prescribed meteorological data. In certain winter conditions the model tries to balance the energy flux by condensing water. Additionally "climate" and "meteorological "measurements" are different things – climate is typically a statistical value (sum, mean or deviation) of meteorological conditions of a longer time period e.g. 30 years.

14. 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?

This contrasts more to the general difference of running the model on a grid (and what problems we get there) and a site level simulation. Some other papers require these types of distinctions. We will likely remove the references and simply state that the simulations are done on site-level.

15. 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 intotiles…

Agreed.

16. 4.17 site-level

Corrected.

17. 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.

The "right" LAI is a tricky question that also depends on how the model is applied. We are considering boreal coniferous forests, where light penetration is deep. The light conditions themselves are more homogenous than for deciduous trees and therefore we can also assume more homogenous leaf distribution. JSBACH also takes into account leaf clumping. One of the more difficult aspects is the leaf shape (cylindrical) and orientation – these we assume similar throughout the study sites. We will consider adding some discussion about this to the paper. Optimisation, focusing on radiative effects, might be interesting, although JSBACH only utilises the two-stream approach.

18. 5.2 You give no reasons, but I assume the modifications were done to facilitate the calibration?

Yes, otherwise the model would have to be recompiled each time a parameter values is changed.

19. 5.7 Why give numbers for the groups and not a name?

Group names would take more space in the tables.

20. 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.

Parameter ranges are not priors. This is explained at the beginning of this document.

21. 6.1 "The" lacking. In general, you are very economic regarding the use of articles.

Added.

22. 6.3 heatsum sum

Corrected.

23. 6.6 ready

We have modified the sentence as per request if the other reviewer.

24. 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)

Yes. We will be falling back to the standard notation.

25. 7.11 Likewise, observations don't have to be continuous, thus not element R

As above.

26. 6.15 What do you mean by directly assessed?

This statement was made to reflect that we do not a have a readily available distribution that would also correctly reflect different observational sets (and more widely, different PFTs, soil conditions etc.). We will amend the sentence.

27. 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

The reason was stated in the sentence, following this equation, where we noted that we are following the notation of Martino et al. (2015). This was done, so it would be easier for the reader to refer to this paper. We will be reverting back to the standard notation.

28. 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

Will be amended.

29. 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).

Agreed.

30. 8.30 This entire procedure remained nebulous to me. First of all, if you favor the proposal 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?

APIS uses blind adaptation, so there are no acceptance probabilities. Each time the IS estimators are adapted, we generate new spin-ups (for each IS using the location of that estimator). We use the locations as the first draw, since the spin-up was generated with these parameter values and in APIS we are mostly interested in the proposal means. This means that we do favour the mean (as 1 of 50 draws is predetermined). In practice, we are (slightly) diminishing the rate of convergence (how "far" is the next IS location).

The purpose of the spin-up is to define the initial state for the model, and it mainly affects the "reservoirs" of soil water content and LAI. The difference in using the "correct" spin-up and one generated with the mean value, is typically small (cost function values within 1% of one another). This estimate is based on previous work, where we have dealt with the issue as well. The scaling (decrease of weights) is used to reflect our confidence in the draws. This procedure, and the one described in the paragraph above, both mainly affect the adaptation of the location parameters (the scaling is not used in the global estimates).

31. 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)

The method itself is more closely related to HMC as we are not estimating the gradient (or even trying to). We did not want to revert to the more usual gradient based methods, since we (somewhat) criticize these in the paper and it would have been a bit hypocritical. If it is required, we can test the stability of the optimisation (starting from the optimised values) with some common algorithm. The dislike of a method itself, is not a good enough reason to rerun all of the optimisations.

32. 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.

This question closely relates to the one above, as does the answer. One of the reasons to use new methods, is to test how they perform. In this sense, it is not the most important thing if they are the best in the field, rather if they can get the job done.

33. 10.4 The spin-up procedure seems to favor the proposal mean and could thus distort the posterior. Please discuss

This was discussed above in the answer to 30.

34. 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.

The KDE is used purely for visualization and we do not draw samples from this. We use KDE to estimate the distribution behind the location parameters (snapshots at specified iterations) of the IS samplers.

35. 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.

The reason to add this "effectiveness" to the paper was to give the reader a sense of which parameters affect most the situation in JSBACH/APIS. It was not meant to be an exact measurement of anything (and we do not present these values in the manuscript). The equifinality of (two hypothetical) parameters is indeed something this measure would not capture (or rather it would capture the "effectiveness" but it would not be constrained as the reviewer suggests). For this we do not have a better answer other than to state that we do not seem to have encountered this type of situation. Our previous work on the subject (Makela et al. 2016) did not reveal considerable correlations (linear or otherwise) between the parameters and there is no indications that the processes (that are controlled by these parameters) themselves would support this.

36. 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.

We will discuss this in detail when answering the comment 40.

37. 10.25 If this is a likelihood, correct interpretation would be that likelihood is normal

Yes if indeed it is interpreted as a likelihood, it would be normal.

38. 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

Yes, this should be more in the lines that this type of cost function has been used before in similar settings. We will remove the word "successfully". We stated that the cost function is "based on the mean squared error" as it is indeed not the MSE. The terminology used here will be revised.

39. 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.

This comment in the manuscript was a reflection to the standard Bayesian framework, where both error terms are used.

40. 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))ˆ2where 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.

These are all valid remarks and we will state the assumptions explicitly in the revised paper. As stated before, APIS works by estimating the expected value of the distribution. As per design, it works very much in the sense of a MAP estimate (which is unaffected by the scaling). The scaling (dividing by the number of unmasked data points) does indeed inflate/deflate the distribution, which is accounted for by the adaptation of the shape parameters in the APIS. The reason to include this divider, is that otherwise the cost function would be biased towards certain study sites. This formulations allowed us also to compare the single-site values and multi-site values directly etc.

Another remark we would like to make, is that we also tested a likelihood (although on a single site and I can't seem to find the results/chains) with the observational error of 20% of the flux value (and using the same scaling). We used the wintertime observations to estimate the precision for each variable. The difference in these results was small. We also wanted to be able to compare the results directly to Knauer et al. (2015) that contained previous work on the subject with JSBACH and utilized the same formulations. Hence, the benefits of translating (in this work) to the new formulation, seemed small when we considered the drawbacks.

The residuals that go to zero are of less importance in this type of work. Almost all of the near-zero values occur in wintertime which the model has some problems and we were initially debating whether to just mask all wintertime values. Additionally, on experience the model results in wintertime are near identical (this holds especially for GPP that should be zero in winter, the ET suffers from e.g. the lack of closure of the energy balance mentioned in the manuscript).

41. 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

Thank you for the comment, we will be separating these into two sections.

42. 11.13 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.

We will be adding the convergence test to the paper.

43. 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.

These are readily available. However, convergence can not be proven. What can be shown, is the lack of divergence.

44. 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.

In the sense, that the probability distribution is defined by the likelihood, this is absolutely true. Here we were referring more to the fact that in APIS, the draws are not evaluated one-by-one, but in groups of 50, so the characteristics of one parameter can be masked by the characteristics of another.

45. 11.20 What does reasonably stable mean? See comment about convergence diagnostics above

We will be adding the convergence test results.

46. 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

Answered at the start of this document.

47. 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.

There're both cases, but this is mostly in reference to the relative humidity fraction (which is at the limit).

48. 12.8 You can parallelize the chains in DREAM, which means that, assuming you run 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.

This characteristics of DREAM is mentioned in the next lines. In our setting for APIS, the two thousand (2000) draws (50 draws from 40 IS samplers) can be estimated and run simultaneously (we did not run it this way, as the resources are limited). The main difference here, is that APIS requires only a fraction of the amount of sequential draws (this is the point we are making). In this paper, we have demonstrated, that APIS can be used in these types of situations. It is clearly not the best in the case of unimodal (or even with few peaks) distributions, but it should be kept in mind when estimating more complicated targets.

49. 12.20 So, why not calibrate them right away?

The original idea was to restrict the calibration to g0 and g1 only so that all of the Ball-Berry variants would be on the same "line".

50. 12.20 / 12.26 Most of the info in these paragraphs are not results

Will be moved.

51. 16.22 The entire section reads like an independent case study with its own methods, results and discussion.

This will likely change as we separate the “Results and discussion” into two sections.

52. 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.

The convergence test results will be added to the paper. The comparison to other algorithms is beyond the scope of this paper.

53. 19.10 Define successful.

We will amend the sentence and make this more explicit.

54. 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.

We will divide the “Results and discussion” into two sections and add the convergence tests.

55. 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.

The code is under MPI-M License agreement and we cannot distribute it. The driving data (approximately 500Mb) and chains can be uploaded e.g. as supplements.