Author responses to reviewer comments for:

McNeall, D., Robertson, E., and Wiltshire, A.: Constraining the carbon cycle in JULES-ES-1.0, Geosci. Model Dev. Discuss. [preprint], https://doi.org/10.5194/gmd-2022-280

Reviewer comments are in black text, with author responses in blue.

Reviewer 1 comments

This is a convincing case study of calibration and sensitivity analysis of a land-surface model, using Gaussian process emulation, experiment design, history matching, and different techniques of sensitivity analysis.

Main comment

Editorial suggestions aside (naming conventions, graphics, clarifications, detailed below) my main comment of substance is about the validation of the emulator. The process is quite well explained, with a focus on leave one out analysis. But:

- Some variables are not so well emulated (shurbFrac_lnd_mean, fLuc_lnd_sum). The authors are aware and seem to have done their best, but the expected implications on the output of sensitivity analysis are perhaps not quite well outlined

Response:

We have generated two figures for the supplementary material, and explained this process in appendix E, “Impact of uncertainty on sensitivity analysis”. We visually show the effect of emulator uncertainty on one-at-a-time sensitivity analysis, for both “absolute” and “modern value”, allowing the reader to make judgements about the impact of uncertainty on the sensitivity analysis.

- Little is said about the validation of emulator variance. Visual inspection of Figures C1 and C2 helps a bit, but quantitative diagnostics could be helpful here.

Response:

We have generated rank histogram reliability diagrams for all 14 main emulators, and added a description of this process and its implications to Appendix D. For each ensemble member, 1000 predictions were simulated from the prediction distribution, ranked, and the rank of the observation was compared. If the prediction probabilities are well calibrated, we should expect them to be approximately uniform. A “domed” shape indicates that uncertainty is overestimated (i.e. we more often have a middle-ranking prediction, and we do better in prediction than we think), and a “U” shape indicates that uncertainty is underestimated (i.e. the observation is more often ranked lower or higher, towards the edge of our distribution) . We see overestimated uncertainty in some of the emulators, particularly fLuc_Ind_sum and cVeg. We don’t see any underestimates of uncertainty, which would be a considerably larger problem, as we are mainly using the emulator central estimate (mean) as a predictor, rather than the full uncertainty distribution. In that case, we can be more confident that the central estimate lies within the uncertainty estimate than uncertainty estimates suggest.

Page per page comment

----

page 2: I would suggest to move away from the very idea that there is a "best" configuration and uncertainty around it. Because how you define the "best" configuration depends on a metric that is not unambiguously defined. Note also that what is meant by the "value" of a "configuration" is not clear. You want to delineate a space of configuration that is fit-for-purpose.

Response:

Amended on page 2 (introduction) to talk about “valid” configurations and associated probability distributions.

A nominal “best Input” still has an important role in the formal statistical structure of History matching, but it is probably not wise to use it in the more everyday language of an introduction, as the reviewer says.

We have amended the “History matching” section to indicate that the “best input” approach is a statistical convenience, and does not necessarily exist in an easily defined way.

p. 3 line 76: edit (repeated "and then")

Response: Edited

p. 7 l. 173 : it could be helpful to give the parameter numbers of fo_io and b_wl_io, for easier match between the text and the figures.

Response: Edited.

p. 7 l. 173 : I disagree with the interpretation. Given the nature of the projection, what we can visualise are thresholds beyond which there is no chance of having a zero carbon (we don't see green points). So the absence of intersection further suggests that the 'no chance zone' of having a 'zero carbon' is increased.

Response: This is a good point, paragraph amended to reflect.

Figure 3 and Figures C1 and C2. Labels are quite clear on Figure 3, and hardly legible on Figs C1 and C2 (at least on the print copy, you need to look a the PDF. Now I realise that the set simulated variables are different so the whole affair is rather confusing. Why not stick to the same set of variables throughout the study ?

Response: We have unified the set of variables used in the study by removing landCoverFrac, which was included in error (and as a multivariate categorical output, is unsuitable for giving a mean value).

We have increased the font size for labels on figures C1 and C2.

Section 2.5:

- Technical question: how do you calibrate the emulator in the parameter space where experiments fail ? Emulator will be unreliable there, so how do you safeguard interpretations ? (e.g. : variance analysis coming from running the emulator).

During the process of generating the input design points for the second wave, inputs with a run failure were removed from the training set used to build the emulator. This was to avoid numerical problems building the emulator. Unlike “zero carbon” ensemble members from section 2.1 “Failure analysis”, there were no clear thresholds in input space leading to consistent failure, and a clean identification and removal of failure regions proved difficult. As the reviewer points out, this led to an imperfect History Matching process, as there is a potential for candidate design points that should be ruled out as possible failures, to be accepted.

The consequences of this problem during the iterated constraint process are limited, as any design point accepted in this way simply led to a failure in the second wave. As we had a relatively large run budget, and a fast-running and easy to set up simulator, the consequences of a large number of failures could be countered by simply running more simulations, or running another wave. In addition, not removing input space that might have a failure is consistent with the aims of History Matching, where we remove input space which is statistically likely to prove uninformative to our overall problem.

As it turned out, there was a smaller proportion of failures - only 10 from 400 members, and those of a different nature to the original ensemble - in the second wave design than the first.

In the situation where the simulator is more computationally expensive, or during the final calibration phase of iterated constraint, this problem might lead to an over-estimation of “NROY” space. To counter this, we propose a separate emulator be trained to predict model failures. This could be deployed on short (and therefore computationally cheap) model runs, as in the atmospheric sciences. It might also be deployed after the history matching process, or during. In any case, this would be a significant research problem in itself, and therefore left to further work.

We have amended the emulator description text to discuss this issue.

- What a "wave" is must be clearly defined at one point, and then the word must be used consistently. A wave is a plan of experiments with the full model (simulator). If you take that definition, then you can reformulate a few sentences: 'as we have seen from the first wave --> as we have seen by inspecting the output of the first wave'.

Response: We have clarified this throughout the text.

Can we have a more legible scheme (why the leading zero in wave00 and wave01). Perhaps more explicit terms would help.

Response. Clarified waves in section “Experiment setup”, to avoid confusion and inconsistency in definition.

- At places you refer to "Our modeller"; later "modeller Wilshire" -> is it really necessary to name the co-author

Response: references to co-author removed.

p. 13 l. 264 : 'if history matching was perfect, 95% of ensemble members would have output'. I am not sure. I believe this is wrong. Suppose that the hyper-surface delineating I=3 is convex. In that case, 100 % of the experiments inside this surface must have I<3. Maybe this region of I=3 is very complex and scattered, with many islands. But in any case, the link between I=3 being a 95% error measure, and the rate of success in the emulator delineating the I=3 region is not straightforward.

Response: The reviewer is correct here, and this has a bearing on a similar subject, brought up by reviewer 2. We have amended the text, to clarify that there is no clear expectation for a certain proportion of ensemble members to be within the constraints and added a reference (Williamson & Vernon, 2013), which discusses the difficulty of generating designs in small or unusual input spaces.

l. 269: "is much higher" ... than what ?

Response: Than in our initial ensemble. Text amended to reflect this.

l. 276 : "numerical problems" : are these numerical problems caused by the specific choice of parameters, or could numerical problems occur quasi-stochastically for any parameter value ? Whether it is one or the other will have implications about whether we could be allowed to simply ignore failures in the emulator design, or whether we somehow need to keep track of the fact that one parameter region is dangerous / irrelevant.

Response: A sample of only 10 failures in the second ensemble is too small to make any conclusive statements about whether they are caused by parameter settings, or occur stochastically. We suggest this is a matter for further studies. We have added text to this effect.

Figure 5 : what is the dark-red zone ?

The dark red zone is where the semi-transparent histograms overlap. We have amended the figure caption to state this.

l. 306 : my remark about how the emulator behaves in the 'unsafe' zone where the model is failing comes again. What are the possible implications on the sensitivity analysis ?

Response: All sensitivity analyses in the paper are completed after renormalising input space to X_level1a.

“We use the 751 members from both wave00 and wave01 that conform to the level1a constraints, being as wide an input space as possible without significant model failure.”

This is defined as where f0_io < 0.9 and b_wl_io > 0.15. Using these thresholds removes a majority of the “failures”, “zero carbon cycle” and “timeseries error” runs from the ensemble. The remaining errors are so rare that it would take a considerable effort - probably a specifically targeted ensemble - and considerable computational resources to remove them. As mentioned before, this could be achieved with a separate “failure” emulator, but we have made the judgement in this paper that the remaining areas that might be affected by run failures are too small to make a substantial difference to the results of the sensitivity analysis.

Figures 7, 8 and 10 : what is the right-hand-side part of the Figure ? It does not seem to bring much information.

Response: This is the summary part of the sensitivity analysis, the mean of the rows, by which the inputs are ranked in terms of sensitivity. We have added this to the main text and to the individual figure captions.

p. 27, l. 410 : It's -> It is

Response: Corrected.

Perhaps one more point worth mentioning in the discussion is that the whole process is influenced by the way constraints are selected and aggregated. This is a place where non-epistemic value judgement may be injected (again, the myth of 'a' best parameter configuration).

Hopefully the aim is to be conservative, so that using more data to constrain sequentially leads to further reductions in space, rather than bringing back space.

Final thoughts

- Posterior variance is an important aspect of the GP approach, but it is used little in this study, except (crucially) in equation 3. Again, this variance does not seem to be thoroughly assessed. Does it matter ?

Response: This is addressed with extra verification for the emulator in appendix D.

- Graphics: make sure labels are legible, and where possible, use more self-explicit labels.

Response: We have increased label size in a number of plots, where it is important to read specific numbers.

- Graphics (again): the pdf size is big, and displays slowly; consider rasterizing some figures.

Response: We will confer with the editorial office to make this a better experience.

Citation: https://doi.org/10.5194/gmd-2022-280-RC1

Reviewer 2 comments

RC2: 'Comment on gmd-2022-280', Anonymous Referee #2, 15 Mar 2023 reply

The manuscript “Constraining the carbon cycle in JULES-ES-1.0” by McNeal et al. examines the use of Gaussian process emulation for exploring the parameter space relevant to the global carbon cycle in the land surface model JULES. It first uses a large perturbed parameter ensemble of JULES runs to train emulators for each of the desired output variables, and then exploits these to carry out history matching and sensitivity analyses. Overall, this is a thoughtful and well written paper, which may well pave the way for better understanding of how we parameterise our land surface models.

However, the paper does not deliver on its initial promise. No concrete results on the extent to which the carbon cycle has been constrained are reported. The closest the authors come to this are Figure 3, 4 and 5, which summarise ensembles of various model diagnostics under different levels of constraint, but I did not find these figures particularly informative, and if they have enabled the authors to make useful decisions about the parameters in JULES that does not come across.

It is very telling that that the discussion and conclusion of the paper focus on the technical aspects of what was done with the ensemble generation and emulation and how this could be improved. Discussion of the what was learned about the JULES carbon cycle and the relevant parameters is largely absent from these sections. Do the authors recommend changing any of the JULES-ES parameters for future configurations based on this work? Great if so, but this is not mentioned. If the outcome of a paper claiming to constrain the carbon cycle in a model does not put forward any proposed changes to the model parameters and/or process then can it really claim to have constrained the model? Overall, I was left wondering if the failed emulation had essentially limited the author's ability to comment on the carbon cycle.

Main corrections:

1) Despite the above comments, I believe there is much value in this paper, but the way it is framed is misleading. The simplest thing for the authors to do is to change the title to better reflect the content of the paper (something like "Emulation of the carbon cycle in JULES-ES to explore input parameter distributions" could be appropriate, but I do not intend this to be prescriptive). However, if the authors feel that I am wrong, then I think there needs to be a much more in-depth discussion of what has been found about the JULES carbon cycle. Ideally this should include some quantification of key global stocks and fluxes and/or discussion of what parameters or process need to be changed in the model and there should be ample discussion of this. As an aside, I note that the abstract of the current manuscript does not mention the carbon cycle or relevant variables.

Response: We feel that the paper title is suitable, as the aim of the paper was to constrain both input parameter space and output space of the ensemble of carbon cycle simulations. Both are materially constrained by our process, leaving an ensemble of historical simulations and an input space that are consistent with observations. We have endeavoured to clarify this result in various places the text, in order to justify the title's validity. However, the reviewer is correct, in that this is primarily a technical paper, providing techniques and information to climate and carbon cycle modellers about how to constrain and ensemble, rather than the last word on the history of the carbon cycle.

Our headline result is that the input parameter space is 88% smaller under our final constraint than in the initial ensemble, and indeed this figure is in the abstract and conclusions. To clarify and enhance the discussion on the constraint of the carbon cycle, we have added two figures, and extended discussion of the constraint of the model outputs (and simulated historical carbon cycle) in section 2.7.1. We have extended the discussion of the meaning for the carbon cycle of the various constraints, and the implications of the position of the "standard" ensemble member in terms of picking better (or other valid) input parameter settings.

We have added considerable discussion of the constraint of input space in section 2.7.2, including clarifying the difference between marginal constraint and joint constraint of the input parameters.

We have added to figure captions for both time series outputs and input space, in order to clarify the results of our constraints.

We have added a table of JULES-ES-1.0 outputs (appendix B) in order to clarify the relationship between the figure labels and JULES-ES-1.0 outputs.

2) It is not clear to me that the results are trustworthy. To what extent does the poor performance of the emulators influence the results? In Figure 5, apparently only 32% of ensemble members conform to the constraints, compared to the theoretical ideal of 95%. I accept that the 95% level

is unobtainable given emulators will never be perfect but 32% seems, to me, to throw doubt over the validity of most of the results in the paper that follow. The cVeg emulator appears not to work, but yet the analysis in the paper includes it in the formulation of the constraint despite that.

The reviewer points out that only 32% of wave01 ensemble members produce output within the limits tolerated by the modellers, whereas ideally this should be closer to 95%. As pointed out by reviewer 1, and subsequently amended by us, there is no expectation that 95% of proposed NROY ensemble members in a new wave of history matching will fall into NROY space when the model is run. The actual number depends strongly on the nature of the behaviour of the model across input space, the uncertainty of the emulator, observations and the tolerance to error of the modeller. The new input space is small, and, as pointed out in other responses, it can be difficult to sample this space efficiently using Monte Carlo techniques. An important feature of any ensemble design where an aim is to build emulators, is that the emulator is not forced to extrapolate beyond the design. It is useful therefore to span a wider parameter space than strictly consistent with constraints therefore, in order that any sampling with the emulator is an interpolation not extrapolation.

We demonstrate a good degree of accuracy in predicting whether ensemble members will be “in” or “out” of the constraints using a leave-one-out analysis in appendix S section D2. Of the 361 members, we successfully predicted the class of 333 (92%) of them. After this second wave is used to re-train the emulators, they become more accurate, so that the final accuracy is better than the accuracy of the first-wave emulator.

It is to the author’s credit that they have been explicit about what has gone wrong, but quantification of the implications for the onward analysis in absent. There is a statement at line 300 that says “we judge that the emulator is accurate enough” but, despite having looked at Appendix C in detail, I am unable to arrive at the same conclusion. Perhaps I am just missing something, and in fact the results are robust despite the poor emulator skill, but if so, I think that needs to come across in a quantitative fashion in the manuscript, ideally in the main text. Maybe an easier option would be to drop the cVeg constraint, be explicit that that has been done, and then re-do the analysis in the paper with just the other three constraints.

In summary, I think one of the following is required: (a) quantification of the errors in the results due to the poor emulation or (b) removal of the cVeg emulator and re-calculation of the analysis in the paper.

We believe that a combination of factors may have (understandably) lead to over-expectation on behalf of the reviewer as to the performance of the emulator, particularly the cVeg emulator, which is highlighted by the reviewer as appearing not to work. We wish to reassure the reviewer that the emulator, while it could be improved, is fit for purpose for this study.

In a leave-one-out validation exercise (appendix C), the cVeg emulator error measure, the “proportional mean absolute error” (PMAE) is 4.45%. This is smaller than the error for NBP (another of our constraints) at 4.93%, for example. It is also considerably smaller than the error for other emulated outputs such as shrubFrac (8.25%), landCoverFrac (6.75%) and C4PftFrac (6.61%) that are obviously well correlated and adequate. It may be that the appearance of inadequacy stems from the fact that some of the higher values of cVeg are predicted low, and this has increased visual weight in the plot, despite there being a much larger cluster of overlapping and well-predicted ensemble members nearer the centre of the range. The four “constraint outputs” are the most important to emulate well, as these determine which input parameter settings are judged NROY, and influence the constraint. All of these outputs are among the lowest-error emulators. Emulation of the other outputs (including anomaly outputs) is important, but a secondary consideration, perhaps more important to the sensitivity analysis section.

The process of constraint is conservative in that we only rule out parts of input space that we are sure are not informative about the real world. An analysis of the variance estimates of the emulators (now included in appendix D section d1.1) concludes that our uncertainty estimates are either well calibrated, or - as in the case of the cVeg emulator - overestimates of uncertainty. This means that cVeg has more uncertainty than necessary in the history matching process which means that the process will be more conservative than necessary. It is therefore more likely to keep an input parameter setting (while it should be removed) than it is to reject it (when it should be kept). This is in keeping with the philosophy of history matching, in that an input parameter setting is conditionally kept if there is doubt, in order to be removed by other data when it becomes available. The result is that the cVeg emulator is simply less effective at removing parts of input space, and the overall analysis would not benefit from its removal as a constraint.

Typos and minor corrections:

L20: “linit”

Response: corrected

L63: “train a an emulator”

Response: corrected

L98: “CO2” should have a subscripted “2”

Response: corrected

L122: unclosed “(“

Response: corrected

Fig 3 and 4: two subplots are labelled “c3PftFrac” presumably one is for C4?

Response: corrected

L280: “also see a constraints of”

Response: corrected

L383: “modellers” needs a possessive apostrophe in this instance

Response: corrected

L399: “sensitivity analysis analyses share”

Response: corrected

L406: is “no-regrets” a specific term, or would “an exploratory analysis” suffice? I am not sure what is intended by “no-regrets”

Response: replaced “no-regrets” with “exploratory”

L452: “bwlio” and “f0io” presumably need underscores, which have instead been translated to subscripts.

Response: corrected

L460: “There (Level 2) using the emulator”

Response: Changed to “Constraining input space to level 2, using the four basic observational constraints and the emulator removed 88\% of initial input parameter space, but marginal ranges were hard to constrain”

L499: “kriging” should be capital “K”.

Response: corrected

L505: “give the modeller as far as possible about the” – missing word?

Response: corrected to “information about the …”

L540-542 and Table C1: The numbers in the text and table do not agree. Possible I have misunderstood and they are not meant to, but it seemed they should.

Response: Numbers from the table were correct. Text amended to reflect this.

Citation: https://doi.org/10.5194/gmd-2022-280-RC2