Responses to Editors' Comments for Manuscript EGUSPHERE-2024-2508

Tuning the ICON-A 2.6.4 Climate Model with Machine-learning based Emulators and History Matching

Addressed Comments for Publication to

GMD

by

Pauline Bonnet, Lorenzo Pastori, Mierk Schwabe, Marco A. Giorgetta,

Fernando Iglesias-Suarez, Veronika Eyring

Dear Editor Dr. Caldwell,

We would like to thank you for the additional valuable comments which helped us to further improve the quality of our manuscript. We have carefully addressed your comments. A summary of main modifications and a detailed point-by-point response to your comments are given below.

We hope that our paper can now be accepted for publication, and we are looking forward to your reply.

Sincerely,

On behalf of all authors,

Pauline Bonnet

Note: To enhance the legibility of this response letter, all the editor's comments are typeset in boxes.

Authors' Response to the Editor

General Comments. Public justification (visible to the public if the article is accepted and published): I think this paper meets the bar for publication in GMD even though I found the analysis to be superficial and the experimental protocol to be haphazard. In particular, history matching is meant to be an iterative procedure but in this work the first round converged in one iteration, the authors only allowed one iteration for the second round due to a lack of computer time, and decadal variability was deemed so large that history matching was scrapped entirely in round 3 in favor of just choosing the best ensemble member. I understand that model tuning is a messy business, but it is hard for readers to learn much when the study is dominated by judgement calls. Also, the unique aspect of this paper is its sequential application of history matching, but since the authors seem to have just given up on the method in 2 out of 3 rounds there's not much that readers can apply from this work to their own history matching efforts.

Despite this, I think the topic is important and understudied and the authors did a good job of addressing reviewer concerns (particularly the more negative comments of reviewer 2, who declined to re-review). Thus I recommend publication after one more opportunity to fix some minor issues.

Comment 1

L83: you provide the resolution for ICON, but just the grid size for other models. Please provide resolution for other models.

Response: We would like to point out that we did indicate the number of grid points on top of indicating the grid size, for the two references mentioned l.83, "For instance, Williamson et al. (2013) used a resolution of 96 x 73 grid points in latitude and longitude (approx. 417 km x 278 km at the equator), while Hourdin et al. (2021, 2023) utilized 144 x 143 grid points (approx. 160 km at the equator)." We hope this answers the question of the editor.

Comment 2

It is weird to use capital letters for some variables representing total counts of things but lower cases for others (e.g. L164). Please use more rational nomenclature.

Response: Thank you for this suggestion. We now use capital letters for all variables representing total counts.

See changes in Section 2.4. and Fig. 1:

- we changed the size n of the ICON PPE to N
- we changed the size N of the emulator metric ensemble to M

In Appendix B, we changed the covariance matrix K (already the number of parameters in set P in section 2.4) to C.

Comment 3

How did you decide how many members to include in each ensemble? I agree with reviewer 2 that the value of this paper is to add to the body of knowledge about how to use history matching and rules of thumb like how many ensemble members are needed is an obvious example of valuable information. Perhaps you can say that your first ensemble was too large since it converged in one step? Could you see what the impact is of using less PPE members?

Response: We thank the editor for this comment. As explained in Section 2.4 (point 1. of the algorithm), a typical rule of thumb is to have a size of the PPE ten times the number of tuning parameters, in order to have an approximately uniform sampling of the parameter space. In our case, due to computational constraints, we used five times the number of tuning parameters, i.e., 30 members for our first physics tuning step. We would avoid claiming that this number was too large. Rather, the PPE size was sufficient to train an emulator that was predictive enough to allow the method to converge to a region where at least the TOA radiation balance was consistent with observations. Note that this did not imply convergence with respect to the total cloud cover metric, for which we needed to include further parameters in a second phase of our tuning approach (as explained in the manuscript, Section 3.2). The impact of the number of PPE members on the quality of the emulator can be seen in Figure 5: a denser sampling of the parameter space generally leads to a more accurate emulator.

> We added these two sentences at the end of the first paragraph of section 3.1: "The sizes of the PPEs are chosen to be smaller than the typical value of ten times the number of parameters (six parameters in $\mathcal{P}_{\mathrm{p1}}$ and seven parameters in $\mathcal{P}_{\mathrm{p2}}$) (Loeppky et al., 2009). This size allows a lower computational cost while being large enough to train an emulator that allows convergence of the PPE towards reference observations, as explained in the next section 3.2.1.

Comment 4

> It is unclear to me whether your methodology is to quit when you get a single member which is good enough or whether you expect all members of the next round of the ensemble to be good enough. The paper seemed unclear on this (e.g. lowest text in Fig 1 versus L199).

Response: We thank the editor for pointing this out. Indeed, we stop when one or more runs are compatible with the observations. In general, when conducting history matching, especially at the earlier iterations of it, not all the members of the next round are expected to be compatible with the observational references. Our approach is to take the configurations we sample with our method to be representative of the space of plausible tuned parameters, and based on the subsequent evaluation of those representative configurations one can decide whether or not these can be used as tuned configurations.

We changed the text in Fig 1. The initial text: "When a compatible region is found, or if the PPEs are far from observational references, a new parameter set is chosen with the help of sensitivity analysis (C). The new parameter set (D) is used for a new phase of the tuning experiment."
was changed into: "If the PPEs are far from observational references, a new parameter set is chosen with the help of sensitivity analysis (C). The new parameter set (D) is used for a new phase of the tuning experiment. When one or more of the model configurations generated in the last PPE are compatible with observations, the iterations of this tuning approach stop. The model configurations compatible with observations are then evaluated. "
We added also a sentence at the new line 210: "In general, in the earlier iterations of history matching, not all the members of the next round are expected to be compatible with the observational references. The configurations that are found compatible with observations are considered representatives of the space of plausible tuned parameters, and are subsequently evaluated on additional evaluation metrics to assess their quality as tuned configurations (see Section 4)"

Comment 5

You don't explain *why* you change the methodology to include observational uncertainty in evaluation. I also didn't feel like I fully understood how you do include observational uncertainty after reading the manuscript.

Response: We thank the editor for pointing the attention to this aspect which required further clarifications from our side. We decided not to include the observational uncertainty in the denominator of Eq.(1) in order to have a more stringent criterion for selecting the parameter configurations in the next PPE and targeting directly the observational means. This stricter criterion makes it more likely to draw parameter configurations with output metrics closer to the observational means, and thus to draw configurations that can be interpreted as representatives for the space of plausible tuned parameters, in the limited number of iterations we did. We clarified this in the description of the method, together with pointing out that this is an important distinction compared to traditional history matching. Regarding the second point of the comment, we now better clarify in the text how the observational uncertainty is dealt with, namely, by assessing whether the output metrics for the parameter configurations chosen as potentially tuned are within the spread of the observational datasets used as reference.

> In the beginning of Section 2.4, we point out that the steps described there concern a method inspired by history matching.
>
> In point 3. of Section 2.4 we added: "This is an important distinction between traditional history matching and our implementation, which we motivate in the next point" referring to the absence of the observational uncertainty in the denominator of Eq.(1).
>
> In the same point, we added: "where we assess whether the outputs of the parameters configurations sampled with our procedure (see next points) are within the spread of the observational datasets used as reference" to clarify how the observational uncertainty is taken into account in our work.
>
> In point 4. of Section 2.4 we added: "Given that we are interested in drawing parameter configurations that are representative of the space of plausible tuned parameters in few iterations, our choice of the implausibility measure as in Eq. (1) provides stricter constraints on the selected parameters, with the observational means Y^0 being the target values for the corresponding metrics."
>
> In the beginning of the evaluation Section 4 we added: "Specifically, we assess whether the outputs of our selected parameter configurations are also compatible with the evaluation metrics, i.e., within the spread of the reanalysis and observational datasets used as reference."
>
> In the conclusions we added a further remark pointing to the differences between our approach and history matching: "Furthermore, we remark that our approach

> presents some differences to traditional history matching implementations. While it allowed us to draw some configurations with outputs compatible with observations for some metrics, a thorough characterization of the space of plausible parameters (the not-ruled-out-yet space (Williamson et al. 2013)) is beyond the scope of our work, and would require several iterations of standard history matching."

Comment 6

L244 - you explain how you generate PPEs 1,2,4 but not PPE 3.

Response: We thank the editor for this remark. To generate $\mathcal{P}_{\mathrm{p2}}$, we chose to keep parameters from PPEs 1 and 2 that had a high Sobol index, in addition we added other tuning parameters potentially allowing to reach a higher value of the cloud cover. To generate PPE$_3$, we conducted a Latin Hyper Cube sampling on this set of parameters $\mathcal{P}_{\mathrm{p2}}$. This is stated in L234 of the old manuscript (now L246 in the revised version), which is followed by the explanation for our choice of the parameter set $\mathcal{P}_{\mathrm{p2}}$. However, in order to make our manuscript clearer on this aspect, we added the following sentence in the revised version:

> At the beginning of Section 3.1, just before the sentence "PPE$_4$ is produced by applying history matching on the results of PPE$_3$.", we added: "The set $\mathcal{P}_{\mathrm{p2}}$ is used to generate PPE$_3$, consisting of 30 samples sampled with LHC sampling.".

Comment 7

In Fig 2, I don't understand why the default tuning is so far outside PPE1.

Response: We thank the editor for this remark. The resolutions of these runs are different. This could explain the difference in the cloud cover value between the icon-aes-1.3 (resolution R2B4, approximately 160 km) and the current ICON-A (R2B5, approximately 80 km) runs. This supports the fact that one has to retune a run when the resolution is changed. To make it clearer, we updated the introduction and Section 3.2 as indicated below.

We changed this paragraph in the introduction L.97: "Our results are compared to the manually tuned version of the ICON-A model that was presented in Giorgetta et al. (2018); Crueger et al. (2018)."
into: "Our results are compared to the manually tuned version of the ICON-A model that was presented in Giorgetta et al. (2018) and Crueger et al. (2018), with a grid size of approximately 160 km (*R2B4* grid), which is two times coarser than the resolution we focus on in this paper (grid size of approximately 80 km, *R2B5* grid)."
We also changed this part at the beginning of the 2nd paragraph of section 3.2.: "Therefore, in the next generation of PPEs (the second phase of our sequential approach), we select the parameter set $\mathcal{P}_{\mathrm{p2}}$ to contain parameters to which cloud cover is more sensitive, following the criteria explained in the previous section."
Into: "ICON-aes-1.3 exhibits a higher value of global cloud cover (orange triangle in Fig. 2.b) than our PPE_1 and PPE_2. The resolution of ICON-aes-1.3 (approximately 160 km) is coarser than PPE_1 and PPE_2 (approximately 80 km). The authors of (Giorgetta et al., 2018) have investigated the six tuning parameters used in $\mathcal{P}_{\mathrm{p1}}$. Here, with these six parameters, we are not able to reach a similar performance for the cloud cover metric. This supports the fact that one should repeat the tuning process when the model resolution is changed (Crueger et al., 2018). Moreover, in addition to the parameters in $\mathcal{P}_{\mathrm{p1}}$, the authors of (Giorgetta et al., 2018) explored other tuning parameters, and these results were not published because having a negligible influence on their tuning process (as explained in their Section 5). In the next generation of PPEs (the second phase of our sequential approach), we investigate the impact of some of these parameters. Therefore, the parameter set $\mathcal{P}_{\mathrm{p2}}$ contains parameters potentially having a stronger effect on cloud cover at the present resolution. "

Comment 8

L352: Are you saying that your 10 yr temporal variability is so large that you abandon history matching in favor of just picking the highest-skill member from your initial PPE?

Response: Yes, this is to some extent correct. The large temporal variability makes it hard to train an emulator to a good prediction skill, which then implies that history matching would potentially require very many iterations (with costly 10-year-long runs) to converge. Due to these large computational costs, we generate only one PPE and then adopt the method that was already used in the manual tuning of ICON in Giorgetta et al. 2018.

To make this clearer, we added this at the beginning of Section 3.3.: "we expect history matching to require a large number of iterations and costly ICON simulations."

We also added this in the Conclusions and Discussions, Section 5 (second paragraph), as mentioned in the next comment 9:

"This suggests at the same time that metrics averaged over broader spatial regions may suffer less from these issues and be more amenable to emulator-based approaches, although too much averaging in space would make the tuning target less characteristic. For the case of the dynamics variable which are proxies for polar stratospheric vortices (zonal mean zonal wind, averaged at 60° North and 60° South, at 10hPa, 10-year average), a possible way to reduce the noise would be to increase the simulation duration and to average the field over only winter or summer months."

Comment 9

It makes sense to me that your dynamic variables would have much larger variance than your physics variables because the latter are all global averages while the former are sampled at a single latitude and height. It would probably help to sample over a broad area in latitudes and pressure levels. I worry that your dynamics variables may mostly be sampling noise.

Response: We thank the editor for this comment and suggestion. In this case study, we decided to select zonal mean zonal winds at 10hPa and 60°N and 60°S as proxies for the stratospheric polar vortices. This measures a wind structure which is not overly sensitive to the choice of the point, owing to the scale of the circulation. Hence these metrics are not just sampling noise, although more noisy than the physics metrics that are global averages. The zonal mean and seasonal (DJF, JJA) means of the zonal wind at 60° North and 60° South and 10hPa has quite often been used as a target for evaluating or comparing simulations of the polar jets in models resolving the stratosphere (e.g, Tripathi et al. 2014; Domeisen et al. (2020a, b); Rao et al. (2020); Baldwin et al. (2021)). We agree that sampling over broader areas in latitude and altitude would help reducing the variability and achieve better performances for the emulator, however too much averaging in space would make the tuning target less characteristic, i.e., average out the aspects (polar vortex) we want our model to capture. Another possible improvement would also be to increase the simulation duration (longer than 10 years) and to average the field over only winter months or summer months (DJF, JJA). We added comments on this in Section 2.2. and the Discussions and Conclusions section of the revised manuscript.

In Section 2.2, we added: "This is a widely used target for evaluating simulations of the polar jets in models resolving the stratosphere (e.g. as seasonal means in Tripathi et al. (2014); Domeisen et al. (2020a, b); Rao et al. (2020); Baldwin et al. (2021))"

As mentioned in the comment 8 above, we added this in the Discussions and Conclusions, Section 5 (second paragraph):

"This suggests at the same time that metrics averaged over broader spatial regions may suffer less from these issues and be more amenable to emulator-based

> approaches, although too much averaging in space would make the tuning target less characteristic. For the case of the dynamics variable which are proxies for polar stratospheric vortices (zonal mean zonal wind, averaged at 60° North and 60° South, at 10 hPa, 10-year average), a possible way to reduce the noise would be to increase the simulation duration and to average the field over only winter or summer months."

Comment 10

Typos/Grammar: L9: "similar to OTHER proposed ML-based tuning"

Response: Thank you for noticing this typo. We changed the text accordingly. (L9)

Comment 11

L16: "emulator allows US to identify..."

Response: We edited the text as suggested. (L16)

Comment 12

L18: "and the CONVERSION COEFFICIENT from cloud water..."

Response: Thank you, it was corrected. (L18)

Comment 13

L63: "(small MEANING close to...")

Response: We corrected it as suggested. Thank you. (L63)

Comment 14

L102: criticalities => "potential issues"

Response: It was corrected accordingly. (L103 in the new manuscript)

Comment 15

L124: output metrics => output observational targets (?)

Response: We clarified it in the text as follows.

> The text: "Table 1 reports the output metrics that we focus on in this study,"
> Was changed into: "Table 1 reports the output metrics, and the corresponding reference datasets and values, that we focus on in this study,". (L124)

Comment 16

L132: duration instead of period?

Response: We left "period", as it refers to the specific years of averaging.

Comment 17

L258: "in more detail" rather than details

Response: We removed "s" to "details". Thank you for this suggestion. (L274 in the new manuscript)

Comment 18

L342: "identified as MOST influential"

Response: Thank you, we changed "mostly" to "most". (L364 in the new manuscript)