Responses to Reviewers' 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 and the reviewers for the valuable comments which help improving the quality of our manuscript. We have carefully addressed your comments and the reviewers' comments. A summary of main modifications and a detailed point-by-point response to your comments and to the comments from Dr. Yang (Reviewer 1, following the order in the open review page) and Prof. Hourdin (Reviewer 2) are given below.

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

Sincerely,

On behalf of all authors,

Pauline Bonnet

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

Authors' Response to the Editor

General Comments. Dear authors,

in my role as Executive editor of GMD, I would like to bring to your attention our Editorial version 1.2:

https://www.geosci-model-dev.net/12/2215/2019/

This highlights some requirements of papers published in GMD, which is also available on the GMD website in the 'Manuscript Types' section:

http://www.geoscientific-model-development.net/submission/manuscript_types.html

In particular, please note that for your paper, the following requirements have not been met in the Discussions paper:

"The main paper must give the model name and version number (or other unique identifier) in the title." "If the model development relates to a single model then the model name and the version number must be included in the title of the paper. If the main intention of an article is to make a general (i.e. model independent) statement about the usefulness of a new development, but the usefulness is shown with the help of one specific model,the model name and version number must be stated in the title. The title could have a form such as, "Title outlining amazing generic advance: a case study with Model XXX (version Y)"." "Code must be published on a persistent public archive with a unique identifier for the exact model version described in the paper or uploaded to the supplement, unless this is impossible for reasons beyond the control of authors. All papers must include a section, at the end of the paper, entitled "Code availability". Here, either instructions for obtaining the code, or the reasons why the code is not available should be clearly stated. It is preferred for the code to be uploaded as a supplement or to be made available at a data repository with an associated DOI (digital object identifier) for the exact model version described in the paper. Alternatively, for established models, there may be an existing means of accessing the code through a

particular system. In this case, there must exist a means of permanently accessing the precise model version described in the paper. In some cases, authors may prefer to put models on their own website, or to act as a point of contact for obtaining the code. Given the impermanence of websites and email addresses, this is not encouraged, and authors should consider improving the availability with a more permanent arrangement. Making code available through personal websites or via email contact to the authors is not sufficient. After the paper is accepted the model archive should be updated to include a link to the GMD paper." All these rules apply to your paper. As you use ICON as a test case please expand the title along the lines: "Tuning a Climate Model with Machine-learning based Emulators and History Matching: a case study with ICON X.y"

Furthermore, your code availability section is insufficient. All code used for the publication needs to be available already at the time of the review / public discussion. Thus a promise into the future as issued in your code availability section is not acceptable. Please provide asap. the code and data you used. This includes the ML-algorithm used, the training data and the exact version of the ICON code used.

Yours,

Astrid Kerkweg (GMD executive Editor)

Response: We appreciate your handling of the review process. To address the first comment, we changed the title as suggested: "Tuning the ICON-A 2.6.4 Climate Model with Machine-learning based Emulators and History Matching". To address the second comment, we updated the code availability section with a doi link to the code used in this work.

Authors' Response to Reviewer 1

General Comments. Summary:

This work uses Gaussian Process emulator together with history matching to estimate the parameters of the atmospheric component of the ICON model. It is well-written with high-quality tables and figures. The following aspects are considered and analyzed in the ML-based tuning approach: (1) The optimal set of parameters to tune; (2) different weights on the target variables to tune (a priority is given for radiation). These two points and the discussions on them together with the fact that very limited ensemble members are available are the innovative parts of this manuscript, in my opinion. The work shows that (1) one iteration of the method could converge relatively well to a model configuration that are generally consistent with observations, and (2) the temporal variability in each model run could be at the same magnitude as the variability from the varied parameters for the dynamics outputs, which have implications for future studies on climate model PPEs. The analysis in this work is based on limited ensemble members, which is a challenge in building emulators, but I find the analysis to evaluate the performance of the emulator and to identify the sensitive parameters robust and convincing.

The confusing part of this manuscript is how the parameter sets are determined (e.g., Pp1, Pp2, and Ppd), which seems a bit arbitrary for those who do not know much about the model. The corresponding part of the text also could benefit from a better organization. I recommend moderate to minor revision for this manuscript. Please see the comments below for more detailed questions.

Response: We thank the reviewer very much for the summary and the very positive feedback on our work, and for recognizing its novelty in the context of parameter tuning in climate models. We have taken the comments regarding the choice and explanation of the parameters very seriously and addressed them in the paper as well as in the answers below.

Comment 1

Main comments

1. It is difficult to understand why and how parameter sets Pp1, Pp2, and Ppd are selected. It seems that Pp2 is selected partially because it has a strong influence on global cloud cover (also pointed out in Line 260), which is proved in Appendix C, but the way it is described seems to be Pp2 is first selected, and then proved to do better for cloud cover. What about other possible parameter combinations?
I wouldn't consider Pp2 an extension of Pp1, because some parameters in Pp1 that are varied are kept fixed in PPE3 and PPE4, and some are varied in these two PPEs. What is the justification in this? Figs 2 and 3 provide some explanations, but more discussion is needed. For example, the black points (PPE2) cluster around a certain area in Fig. 3c, which seems to suggest that Pr0 and crt do not need to be varied (hence fixed in PPE3 and PPE4), but they are included and varied in Pp2. Another example could be csatsc, it belongs to the cloud cover scheme, but why it is not varied in Pp1 and Pp2, and its value is chosen to be fixed while other cloud cover parameters are varied? It is varied in PpD but only briefly mentioned in Appendix C. Justifications on why some parameters are included in Pp2 and Ppd need to be pointed out more explicitly (references, analysis, knowledge on the simulated processes, or limited computational resources?).

Response: We agree with the reviewer that the explanation on the choice of the tuning parameters could be clearer. We have added further explanations throughout the manuscript, and especially in the results section, to better clarify and motivate our choices. To answer the reviewer's questions, the selection of the parameter set $\mathcal{P}_{\text{p2}}$ was motivated by expert knowledge on the simulated processes, i.e., by choosing the additional parameters that in the ICON manual tuning history have been deemed most influential for cloud cover, after private communications with authors from (Giorgetta et al. 2018). Due to the computational resources available we have decided to investigate only the reported parameter set $\mathcal{P}_{\text{p2}}$, without looking into other combinations. As the

reviewer points out, $\mathcal{P}_{p2}$ is not an extension of $\mathcal{P}_{p1}$, and we apologize if our wording may have caused confusion on that. Rather, what we adopted was a sequential approach where (most of) the parameters in $\mathcal{P}_{p1}$ have been fixed to their best value for PPE_3 and PPE_4. Specifically, the parameters associated to a higher first order Sobol index, crt and pr0, have been kept from $\mathcal{P}_{p1}$ to $\mathcal{P}_{p2}$. Following the reviewer's suggestions (and in particular the comment 3 below), we have integrated the appendix on the Sobol indices in the main text, for better clarity on this point. The same strategy we adopted in moving to $\mathcal{P}_{pd}$, with the addition of crs and csatsc after further advice from ICON experts.

We have clarified our criteria for selecting the different parameter sets throughout the manuscript. In particular, our strategy is now made transparent already in the introduction and thoroughly described in Section 2.4. There, we clarifyied that in our tuning experiment we integrate history matching in a sequential approach where at each step we may add or remove tuning parameters based on the history matching results and the subsequent sensitivity analysis. We also modified section 3.1 adding a clearer explanation on our specific choices for the parameters for our experiments. Furthermore, to corroborate these choices and for improving the presentation of our results, we decided to move the Appendix C on sensitivity analysis to the main text, with the title "Sensitivity analysis for the physics parameters and outputs", now Section 3.2.2.

Comment 2

2. In addition to emulator construction and history matching, this work does another level of analysis/comparison, which is to compare which parameter set (Pp1 or Pp2) is better to tune. This seems more like a part of a method. Regardless I think this level (i.e., the parameter set selection between Pp1 and Pp2) should be more explicitly pointed out somewhere in the manuscript. It would also help with the organization of the manuscript.

Response: We thank the reviewer for this suggestion. In fact, our aim in choosing different parameter sets for tuning was not to perform a direct comparison between those, but rather to develop a sequential approach for where different sets of parameters are subsequently tuned with the help of ML emulators. We have clarified this point in the manuscript accordingly, by explaining our sequential approach in the introduction as well as in the methods and results section.

We added a clearer explanation of our sequential approach together with the criteria for selecting parameters in the steps therein, in the introduction and in Section 2.4, together with an updated Figure 1 for a schematic view of our method. The technical details of our method for selecting new parameters (the sensitivity analysis based on Sobol indices) is now in Section 3.2.2, "Sensitivity analysis for the physics parameters and outputs".

Comment 3

3. Some content in the Appendices, especially Appendices C and F, is very tied to the logic and flow of the manuscript. This is most pronounced at Lines 263-266, which is very strongly supported by Fig D1. and Fig. F1. Similarly, statements in the last few sentences of the abstract seem to be also supported by contents in the Appendices. I think moving some contents in the Appendices to the main text (e.g., presenting a subset of Fig F1 for the parameters that are mentioned in the main text) would make the manuscript flow better.

Response: We thank the reviewer for this suggestion. To improve clarity on our choices and steps, we moved the Appendix C on sensitivity analysis (with Sobol indices) to the main text, with the title "Sensitivity analysis for the physics parameters and outputs", now Section 3.2.2. We also adapted the section 3.2.3 so that the new Fig. 7 is now combining a subset of the previous Fig. 7 (parameter-to-output map with parameter set $\mathcal{P}_{\text{p1}}$) and the previous Fig. F1 (parameter-to-output map with parameter set $\mathcal{P}_{\text{p2}}$). We

also moved the previous Fig. 7 (now Fig. E1) to Appendix E, with the previous Fig. F1 (now Fig. E2).

Comment 4

Minor comments

1. Because of the main comments, I think Figure 1 and the workflow described in Lines 7-10 and 135-165 a bit oversimplify (or undersell) the work done in this manuscript.

Response: We agree with the reviewer, and have substantially improved the presentation of our method and its difference from 'traditional' history matching in our revised version. The description in the methods section (2.4) and Figure 1 therein now provides a better overview of the detailed steps of our method.

Comment 5

2. How many samples are generated from the emulator for history matching?

Response: We generate 300,000 samples from the emulator at one iteration of the history matching technique. We added it to bullet 3 in section 2.4 as follows:

"(typically ranging from 10^5 to 10^6)" -> "(typically ranging from 10^5 to 10^6, here $N = 3 \times 10^5$)"

Comment 6

3. If something is mentioned earlier (as suggested in the Main comment 1.), it would be easier to justify (or understand) Lines 224-235.

Response: As stated in the answer to Comment 1, we have clarified our explanation in Section 3.1 and added the supporting Section 3.2.2 on the sensitivity analysis, and

we believe that the manuscript is now more understandable concerning our choices and their justifications.

Comment 7

4. Fig. 3: it is interesting that Fig. 3a and b has two clusters of PPE2 points. This should be pointed out?

Response: We thank the reviewer for this observation. Since this is not the focus of our analysis, and we haven't made further investigations for those, we prefer not to add speculation on the nature of their origin.

Comment 8

Details

Line 70-75: A bit confusing here because here it seems to say either Pp1 + Ppd or Pp2 + Ppd, the comparison between Pp1 and Pp2 is not mentioned here.

Response: We thank the reviewer for this observation. As stated before, our aim was not to perform a direct comparison between different sets of tuning parameters, but rather to change the parameter sets sequentially based on the history matching results and on sensitivity analysis. We think that the confusion arising here has now been addressed after our rewriting of the introduction and methods sections.

Comment 9

Line 75-77: the second tuning targets the dynamics outputs, but Lines 76-77 (starting from where ...) seem to suggest that the criterion is just based on achieving a nearly balanced global annual net radiation flux at TOA. Please clarify (although it is clarified later in Line 175). Maybe something like "keep the highest priority ... meanwhile trying to match ...".

Response: We thank the reviewer for the comment. We changed the text accordingly as:

"puts the highest priority on achieving a nearly balanced global annual net radiation flux at top of the atmosphere (TOA)." -> "keep the highest priority on achieving a nearly balanced global annual net radiation flux at top of the atmosphere (TOA) meanwhile trying to achieve a high performance on the dynamics outputs."

Comment 10

Line 111: Why the average period is different (1980 in Line 105 and 1980-1989 in Line 111)?

Response: The average period is different for physics and dynamics metrics due to the different variability and equilibration times of the associated variables. In particular, as substantiated in Section 3.3.1, the physics metrics have a lower year-to-year variability compared to the dynamics ones, which means that one simulated year is sufficient to obtain a representative value for the annual average. Conversely, for dynamics metrics the annual averages need to be estimated from multi-year simulations due to their larger variability and sensitivity to geographic patterns. We have made this clearer in Section 2.2.

We added the following paragraph in Section 2.2: "We use different averaging periods for physics and dynamics outputs because of the different year-to-year variability and equilibration times of the associated variables. As substantiated in Section 3.3.1, the physics outputs have lower year-to-year variability compared to the dynamics ones, meaning that one simulated year is sufficient to obtain a representative value for the annual averages. Conversely, for dynamics metrics the annual averages need to be estimated from multi-year simulations due to their larger variability and sensitivity to geographic patterns."

Comment 11

Line 155: the symbol n is used too many times (here and the number of ensemble members before, and in point 5, too).

Response: Since the symbol n refers to the same number, that is the size of the PPE, which is kept constant throughout each wave of history matching for a given set of tuning parameters, we decided not to change our notation.

Comment 12

Eq. 3: can you explain or provide reference on why these are perturbed? (similar question to Main comment #1).

Response: This set of parameters for $\mathcal{P}_{\text{p1}}$ are the tuning parameters that were used by Giorgetta et al. (2018). We added the reference in the text, after Eq. (3).

Comment 13

Line 192: the "previous PPEs" here refer to from previous studies or PPE1 and PPE2 done in this work? It seems to be PPE1 and PPE2 here, but not clear. Please clarify. This is related to Main comment #2.

Response: It referred to $\mathcal{PPE}_1$ and $\mathcal{PPE}_2$ in our study. We have clarified this in the revised version of Section 3.1.

Comment 14

Line 215 and Lines 219-223: I assume what Line 192 refers to is Lines 219-223? I think it might be better to put Lines 219-223 to Line 215?

Response: Yes, this is correct. We moved the following paragraph (Line 219-223) to Line 215 as suggested:

However, Fig.2 panel (b) shows that global cloud cover still remains lower than the observational data (of approximately 1% compared to CLARA-AVHRR, and 3% compared to ESACCI), despite PPE_2 yielding a slightly higher cloud cover (closer to the observed range) than PPE_1.

Comment 15

Section 3.3.1 helps explain why some observations corresponding to the dynamics output cannot be matched. I think a sentence or two pointing this out would make the logic flow more nicely. For example, at the end of Line 295, add a sentence saying that the output variability is also a factor.

Response: We thank the reviewer for this comment. We added the following sentence before Section 3.3.1:

"In the next section we analyze the variability of the dynamics outputs, and we identify in it a possible explanation for the difficulty of matching them in our tuning."

Comment 16

Line 347: I recommend deleting the sentence "since the required size of the PPEs .... The tuning parameter space.", as this is only practically true. We know that ideally to fill in the parameter space with points, one more parameter means significantly more points (e.g., the difference of 10^7 and 10^8).

Response: We agree with the reviewer that the linear scaling with the number of parameters is a practical criterion for obtaining a uniform-looking LHC sampling, but we also point out that for the purpose of emulation it may not be necessary to densely sample the whole parameter space (for which case the scaling would be exponential). We therefore decided to keep the sentence, but to remove the adjective 'linear' to avoid possible confusion or misunderstanding.

Comment 17

Tables and Figures

Table 5: fixed parameters in PPE1 and PPE3 are for different reasons (PPE1 for default, but PPE3 seems to be from PPE2). Maybe worth pointing that out? Besides that, I think this is a great table that summarizes what is done in this work.

Response: We thank the reviewer for the comment. We have pointed this out explicitly in the revised Table 5, adding "(fixed from default configuration)" for the parameters fixed in PPE_1, and "(fixed from best conf. in PPE_2)" for the parameters fixed in PPE_3 and PPE_5.

Comment 18

Fig. 2: there is only 29 squares (PPE2; including the selected run). Is one outside the extent or overlapping?

Response: We thank the reviewer for noticing this. This is correct, only 29 runs were included in that PPE, and we updated the table 5 accordingly.

Comment 19

Fig. 4: there are less than 30 points for PPE3 and PPE4. Please specify why they are not plotted here (maybe outside the extent).

Response: The remaining points were outside the extent of the plot. The extent was enlarged slightly, to include all points of PPE4.

Comment 20

Line 233: I recommend adding boxes in Fig. 4 showing the extent of Fig. 2 to highlight the point in this line.

Response: We thank the reviewer for this comment. We highlighted this in Fig. 4, in the caption of Fig. 4 and in the text (last paragraph before Section 3.2.1).

Comment 21

Fig. 5: I think this is a figure with robust results but I recommend putting the y-axis limit to be 0-1 or -0.1-1. The way it is presented now, i.e., lower y-axis limit being -0.5, seems a bit unnecessary.

Response: We thank the reviewer for the comment. We implemented the suggested changes.

Comment 22

Fig. 7a: there is another two points that are just right above the yellow triangle. They seem to fit well to Fig. 7a, too. Maybe it would help highlighting where they are in Fig. 7b-d (which I assume that are far from the observations), such that the selection of the two runs is more convincing?

Response: We thank the reviewer for the suggestion. These two points have been highlighted in the new Fig. 8 a-d and the caption has been updated. We also added these points to the Fig. C2 (parameter space).

Authors' Response to Reviewer 2

General Comments. The paper entitled "Tuning a Climate Model with Machine-learning based Emulators and History Matching" describes a tuning protocol applied to the Icon global climate model.
Although the paper is very clear, well written and easy to read, and probably useful for the community as well, I have major concerns which make it not suitable for publication at this stage.

Response: We thank the reviewer for the thorough and constructive comments on our manuscript. We have taken those very seriously, and addressed them in the points below. The criticisms raised by the reviewer helped us in identifying points in which our manuscript may have been unclear and to better fit our work in the existing literature on the topic. We are confident that after our revisions it is now suited for publication.

Comment 1

Major concerns. My first major concern is about the novelty of the work and the way it is presented in the frame of the ongoing literature on the subject. Since this point concerns in particular scientific papers I was involved in (most of which are cited, this is not an issue), and to avoid any ambiguity, I decided to sign my review, although I usually prefer not to. The title, the abstract and the introduction, are suggesting or saying that the originality of the paper is to use History Matching and Emulators to tune a climate model (for instance in the abstract, line 6, "Here, we develop a MB based tuning method ...). However, this method is exactly the one that was proposed by Daniel Williamson, and first applied to an oceanic model in Williamson et al., 2017. Proof of concept of the potential of the method to tune a global climate model were given in two papers (Hourdin et al., 2021, 2023). In the first paper, we have shown how, with a combination of single column simulations and global climate simulations, using History Matching with Gaussian

Process (GP) based emulators, we were able to automatically retune the model's free parameters and automatically reach a tuning as good as that of the previous 6A version of the model (used for CMIP6 production). The second paper, which is cited here in a general sentence about uncertainty quantification, presents a successful automatic tuning for the IPSL global climate model: 18 parameters are varied and we go so far as to show that 2 of the finally selected simulations could have been used as reference configurations for CMIP6 in place of the IPSL-CM6A configuration, which was obtained after a long and fastidious phase of manual tuning. If we go into even more details, the fact that the "physics tuning" is done considering the second year of 2-year long forced-by-SSTs simulations (line 94-95) is exactly the protocol we already published (Hourdin et al. 2021, 2023). In fact the authors mention that it was a protocol already used for manual tuning (which is the case as well for the IPSL model). But it is interesting to underline that it seems to be a relevant and shared protocol.

This is not to say that the work itself is not interesting and does not deserve publication. I am actually quite convinced that we need much more publications of this type and I am glad that this paper was submitted. As we have written in the conclusions of several papers on the subject, this History Matching approach for model tuning is not the end of the story. Rather, and this is something that has gained in strength and depth as we keep working with it, it is opening a new area for climate modeling, one with a lot of room and questions to investigate. Indeed, the approach only provides a framework. There are so many possible ways to implement it, concerning for instance the choice of metrics and model configurations. Depending on these choices, the approach may be more or less efficient or successful, in ways we do not know or understand very well yet. I think that this is the interesting part of the work, and therefore that the paper should really focus on the specificities of the protocol. The approach is not novel but as it is quite recent, every new implementation of it brings new insights into

modeling and models' behavior. The authors could for instance reflect on: What is common with or different from previous studies? Why did they make a particular choice instead of another? For instance, from my perspective, one specificity is to propose successive phases of tuning in which some parameters are set to their "best values" and new parameters are varied. This might be an interesting choice and it would be very interesting to discuss it in more depth. It may be cheaper than varying all the parameters at once. Also it probably makes tuning experiments easier to interpret, by separating questions from the beginning. On the other hand, it clearly forbids some compensation between parameters of two phases as clearly seen when looking at the TOA global net radiative budget in the various experiments. One thing I would wish to find in a revised version of the manuscript is the authors' view on these questions, and more generally, I wish that this paper and other publications on tuning would be thought and written as a contribution into building this new science, beyond just reporting results — which is of course also an important aspect of publication.

Response: We start by thanking the reviewer for recognizing the value of our work and its contributions to the existing literature on climate model tuning. As an application of history matching to tuning the ICON model, we also see ours as an important contribution shedding light on different aspects and implementations of this method. We also recognize that our original formulations may have caused misunderstanding on the points of originality of our work. We carefully went through the manuscript and changed our formulations accordingly. In particular, also addressing some of the next comments by the reviewer, we have now given more details on the work done in the citations in the introduction, and made explicit what differentiates our work from what is already present in the literature. As correctly pointed out by the reviewer, the main distinctive feature of our approach is its sequential nature, which we have chosen to reduce the number of climate model runs to construct the PPEs, given the relatively high resolution (approximately 80 km) targeted here. These reasons the way we implemented

our approach are now much more thoroughly clarified in the introduction as well as in the methods and results sections. Furthermore in the revised version of the manuscript we also discuss the possible criticalities associated with this approach, especially given our concrete results in ICON, including recommendations on when it may be advisable not to use it. We are convinced that thanks to the reviewer's comments our work is now much clearer, especially with regards to its specificity, and can be better framed within the existing literature on automatic parameter tuning.

We made the following changes throughout the manuscript for addressing this and some of the following comments:

- We have modified some of our formulations in the abstract, clarifying what parts of our approach are shared with already proposed methods and what parts are instead unique to our case.
- In the introduction, we have added a more thorough description of the work done in the cited literature (now lines 67 to 78).
- The aspects that distinguish our implementation from the ones reported in the literature are discussed in the introduction (now lines 81 to 89).
- The Methods Section 2.4 now provides a much clearer explanation of our sequential approach where history matching is integrated.
- The presentation of the results in Section 3 has been adapted to better reflect the structure of our approach. We also added a technical Section (3.2.2) on sensitivity analysis, which is part of our approach, for better clarity on our choices of the parameters in subsequent phases of our method.
- In the concluding section (now Section 5) we further discuss benefits and criticalities of our sequential approach (now lines 444 to 457).

Comment 2

My second major concern is about the way history matching is presented. It really is a major concern since one important aspect of this particular moment in the history of climate modeling science is to clarify the concepts, establish a common vocabulary and so on. There are two points I wish to make. First, the fact that the approach is presented as an optimization problem (see e. g. line 42 p2 , line 54 p2 , line 31 p7). Daniel Williamson, when promoting history matching, insisted rather heavily on the fact that it is not an optimization approach. After years of working with this approach, I am convinced that it is one of the most important aspects of his proposal. The approach consists in finding the region of the free-parameter space (a hypercube defined as the product of [min,max
segments for each free parameter) that is compatible with observations for a series of chosen metrics. "Compatibility" is defined through a set tolerances to error which should in principle include at least the uncertainty on the target (often observational uncertainty) and the model structural errors (generally unknown). A state that was reached by optimizing farther than this tolerance should not and theoretically can not, be preferred to another one. Of course, in practice, climate modelers may (and may have to) choose their "best" setup (that can be for instance selected using metrics not already used in the tuning procedure). But they should clearly motivate the choice and be conscious that this is outside the history matching philosophy. The second point is tightly related to the first one. It concerns the definition of Implausibility. When formulating history matching in a Bayesian framework, the Implausibility should include at the denominator the uncertainty of the emulator (as proposed here, Eq 1) but also the tolerance to error associated with the other sources of uncertainty. One of the goals of the iterative refocusing usually associated with history matching is to reduce the uncertainty of the emulator and reduce the denominator to the a priori tolerance to error. Is the choice of not including tolerance to error at the denominator related to the

idea of seeing history matching as an optimisation problem ? If yes, this should be discussed much more in depth and the mathematical foundation presented. It should be acknowledged that it is not what is proposed in the papers cited on History Matching. If it is a bad understanding, I recommend modifying the text to correct this and to interpret the results with this in mind. In fact, in practice it is probably possible to discuss the results presented including the idea of tolerance to error.

Response: We thank the reviewer for these comments on history matching. We do agree that establishing a common terminology throughout the literature is of great importance, and therefore we have taken these very seriously. Regarding the first point made by the reviewer, we have modified our statements throughout the manuscript on the presentation of history matching as an optimization algorithm. In particular, we made clear that automatic tuning approaches based on history matching aim at shrinking the parameter space to regions for which outputs are consistent with observations, where consistent means within the observational range. Regarding the second point, we have corrected the text accordingly. In particular, we specified that while we use an implausibility measure which does not contain the observational uncertainty, the standard definition does contain it, and that in our case the observational uncertainty is instead accounted for when selecting the potentially tuned model configurations and comparing them, which we do in the new Section 4 of the revised manuscript.

We made the following changes throughout the manuscript for addressing this comment:

- We have modified our formulations in the introduction regarding the presentation of automatic tuning and history matching. In particular, we write that the aim of automatic tuning methods is "to improve the accuracy and reproducibility of parameter tuning by giving it a mathematical formulation amenable to numerical treatment", and making clear that "the goal is to find

the regions of parameter space for which the model outputs are consistent with observation-based reference datasets" (now lines 45 to 50).

- In Section 2.4, while explaining in details our sequential approach based on history matching, we have again clarified that the goal is to find parameter regions for which the model outputs are compatible with observations, where compatibility means within the observational range (now lines 155-156).

- In Section 2.4, below Eq, (1), we provided a better explanation of the idea behind the definition of the implausibility measure ρ. Furthermore, we made clear that while we use an implausibility measure which does not contain the observational uncertainty, the standard definition does contain it. In our case, the observational uncertainty is accounted for during the comparison of the potentially tuned model configurations.

Comment 3

The last concern, less fundamental, is that I am missing an evaluation of whether the approach proposed was at the end conclusive or not. Would you choose your final best simulation instead of the previous ICON-aes-1.3 configuration ? The text and Figure 7 suggest that at least one metric is farther from observation for all the simulations of the ensemble. This question could be addressed by computing some classical RMS metrics (see for instance Fig 3 and SI in Hourdin et al. 2023), independent from the one used for tuning on both the ensemble and ICON-aes-1.3 configuration. The answer can be no, in which case it is still important to discuss the reason for this relative failure in the conclusion (which is partly done already).

Response: We thank the reviewer for this comment. We added an evaluation section, now Section 4, where we computed five additional metrics in the new Fig. 10. These metrics were not targeted during the tuning experiment, and consist of global and multi-year averages of surface temperature, total precipitation, sea-level pressure, vertically

integrated cloud water and cloud ice. Our tuning experiment successfully produced configurations largely comparable to ICON-aes-1.3, without showing signs of 'overtuning' to the chosen metrics. This is a positive result, since one of the aims is indeed to generate configurations that are at least as good as manually tuned ones, with less manual steps for the users. However, we acknowledge that it did not reduce biases compared to the previously tuned ICON-aes-1.3. Moreover, one should keep in mind that the manually tuned version ICON-aes-1.3 had a coarser resolution (160 km) than the one we are working on (80 km), thus being less computationally costly to tune.

Comment 4

Specific comments

Abstract line 6 : change the wording to better acknowledge that it is an application of already published work.

Response: We have reformulated the abstract to make this clear (lines 6 to 13).

Comment 5

line 37 p2 : should be good to have a citation here. For instance one concerning the GFDL results reproduced in Fig 3 of Hourdin et al. 2017.

Response: We thank the reviewer for the comment and the reference. We added in now line 41 p2:

"(see e.g. Fig. 3 of Hourdin et al., 2017)."

Comment 6

line 44 p2 : not sure to know what you have in mind when saying "most commonly used one in climate model tuning". Citations ?

Response: By that we mean that those are the methods that are most represented in the literature on climate model tuning with ML-based automatic approaches, at least to the best of our knowledge, for which we provide the list of works in the sentence afterwards. We decided however to omit this part of the sentence to avoid confusion, since it is not crucial for delivering the main message.

Comment 7

line 47 p2 : think you should give more details on the work done in the citations, to better position your work with respect to it.

Response: We thank the reviewer for the suggestion. We expanded the part on the existing literature in the introduction as follows (now lines 67 to 78):

"Several implementations of the ideas above have been proposed, for tuning models of different complexity. History matching has been implemented to constrain parameters in the coupled climate model (HadCM3) (D. Williamson et al., 2013) and to estimate parametric uncertainty in the NEMO ocean model (D. B. Williamson et al., 2017). It has also been used to tune parameters of the turbulence scheme of a single column model version of ARPEGE-Climat 6.3, using large-eddy simulations as reference (Couvreux et al., 2021). History matching in combination with single-column models was also employed to constrain convective parameters for their subsequent use in the LMDZ atmospheric model of the IPSL Earth System Model (Hourdin et al., 2021). Furthermore, (Hourdin et al., 2023) showed another successful application to the IPSL model, finding an ensemble of tuned parameter configurations as good as the manually tuned vesion IPSL-CM6A-LR used for CMIP6. Besides their use in history matching, ML-based emulators find applications in parameter tuning also in combination with ensemble methods (Cleary et al., 2021) (with test applications on Lorenz '63 and '96 models (Cleary et al.,

2021), convection schemes in idealized global circulation model (Dunbar et al., 2021), gravity waves parameterizations (Mansfield & Sheshadri, 2022)), and with approximate Bayesian computation (Watson-Parris et al., 2021)."

Comment 8

line 53 p2 : citation of Hourdin et al 2023 should also be listed in the examples of use of history matching with GP emulators for climate model tuning.

Response: We thank the reviewer for pointing this out. We included the citation as follows (now lines 71 to 74):

"History matching in combination with single-column models was also employed to constrain convective parameters for their subsequent use in the LMDZ atmospheric model of the IPSL Earth System Model (Hourdin et al., 2021). Furthermore, (Hourdin et al., 2023) showed another successful application to the IPSL model, finding an ensemble of tuned parameter configurations as good as the manually tuned vesion IPSL-CM6A-LR used for CMIP6."

Comment 9

line 64-67 p3 : Worth making it clearer.

Response: We reformulated those lines to give a better explanation of history matching, also incorporating the previous comments from the reviewer. The new paragraph now reads as follows (now lines 60 to 66):

"History matching aims at minimizing the number of required model simulations in the search of optimal parameters, by balancing the sampling of unexplored parameter regions with the sampling close to configurations found potentially compatible with observations. This is achieved using a metric that weights both

the distance of the emulator predictions from the observational references (small close to observationally-compatible configurations), and the uncertainty of the emulator (high in unobserved parameter regions). The three steps described above are repeated until the model outputs used as tuning metrics converge to the corresponding observational range, thus yielding one or multiple tuned parameter configurations, or a distribution thereof (Watson-Parris et al., 2021)."

Comment 10

line 21 : is not 1W/m2 a very optimistic value for errors on TOA fluxes ?

Response: In this study, we adopted the reference value of a 1 W/m2 error range in TOA fluxes, as previously targeted by (Giorgetta et al., 2018) during the manual tuning of ICON (see Section 2.3, now lines 142 to 146). This threshold was selected by the authors of (Giorgetta et al., 2018) for the following reason: "This tuning goal is chosen to make the atmospheric model suitable for coupled climate simulations, for which large energetic unbalances at TOA would be detrimental."

Comment 11

line 80-82 p 8 are worth expanding a little bit.

Response: We have expanded the explanation in the main text as follows (now lines 213 to 217):

"Separating the tuning of physics-only metrics from that involving also dynamics outputs allows us to use different durations of the ICON-A simulations for the two steps, and to further reduce the computational costs. Specifically, as substantiated in Section 3.3.1, the physics outputs have lower year-to-year variability and shorter equilibration timescales compared to the dynamics outputs. This means that for

physics outputs shorter simulations are needed for obtaining a representative value for the annually averaged variables used as metrics."

Comment 12

Table 6 : is there some argument, statistical or physical, to say that a value of $R^2 > 0.75$ is enough ? Could be interesting to discuss this a little bit more.

Response: There is no general argument for deciding which value of R^2 is high enough, and in our work the goodness of an $R^2 > 0.75$ was justified *a posteriori* by the fact that the emulator predictions were informative enough to guarantee a convergence in the iteration of history matching.

Comment 13

Figure 5 : Why using only five samples ? It should be quite cheap to run more to have more robust estimates, no ?

Response: We thank the reviewer for the suggestion. We ran it again, with fifty samples instead of five, and updated the Figure 5 and the caption accordingly.

Comment 14

line 8 p48 : you say that the approach can guide the sensitivity analyses [... as we did with [...] Sobol indices. I would rather say that history matching is a way to make a much more complete sensitivity analysis than local linearization or Sobol indices.

Response: This sentence was referring not to the history matching approach in general, but rather to the visualization of the parameter-to-output maps specifically. However, we agree that history matching already inherently makes use of the information contained in these parameter-to-output relationships. Our point in presenting in this visualization

is that it helped in the selection of the parameter sets across the phases of our sequential approach, ultimately helping in keeping the dimensionality of the parameter space low. We clarified this point in the updated version of the manuscript, rewriting the paragraph as follows (now lines 321 to 325):

> "The previously trained emulator can also be used for the visualization of the parameter-to-output dependencies. These visualizations complement the sensitivity analysis presented in the previous section, and further helped us in the selection of the tuning parameters to be kept across the phases of our sequential tuning approach. Generally, such visualizations are very useful for informing the user of the effect of a parameter on the outputs: they can help selecting the most influential parameters and the corresponding plausible ranges, potentially reducing the computational costs of tuning exercises."

Comment 15

line 65-66 p14 : you could mention that cvtfall was identify as a tuning parameter widely shared among climate models in the Hourdin et al 2017 synthesis paper.

Response: We thank the reviewer for the suggestion. We added a sentence on that (now lines 344 to 346):

> "Note that parameters governing cloud microphysical processes (e.g. fall velocities such as cvtfall) were identified as tuning parameters widely shared among climate models in Hourdin et al. (2017) synthesis paper(see Table ES4 therein)".

Comment 16

Figure 6 : why starting numbering graphs from "b, c, d ..." rather than "a, b, c ..." ?

Response: We thank the reviewer for noticing this. We changed the numbering to "a, b, c,..." in the new Figure 7.

Comment 17

Figure 7 : other choices of marker color and thickness could make the figure easier to read.

Response: We changed the marker color for the reference values in new figures 8.c and 8.d in the revised manuscript and increased the size of the markers in Figure 8 d.

Comment 18

line 96 p16 : "the effects of" can be removed

Response: We implemented the change (now line 381):

"We now use PPE_5 to analyze the internal variability of the investigated output metrics and compare it to the parameters' effects."

Comment 19

Figure 8 : It took me time to understand exactly this (relevant and interesting) figure. Changing "against annual mean (1980" by "against the mean of one particular year (here 1980" could help.

Response: We changed the caption as suggested (now Fig. 9 in the revised manuscript).

Comment 20

line 23 p19 : "aided by an emulator for the outputs" could be a little bit better phrased. "aided by building and using emulators for each output metrics"

Response: We thank the reviewer for the suggestion. We changed the formulation as follows (now lines 423 to 426):

> "This exploration is aided by building and using emulators, here Gaussian processes (GPs), for each of the considered output metrics. The emulator approximates the climate model simulation outputs for arbitrary values of the tuning parameters, and can be used to create large emulated metrics ensembles at a much cheaper computational cost."

Comment 21

line 24 p19 : I do not like the idea of using the wording PPE to describe an ensemble of metrics computed with the emulator. For me, a PPE is an ensemble of (real) GCM runs.

Response: We changed the wording to "emulated metrics ensemble" when referring to the ensembles generated with the emulators (Figure 1, now lines 177 and 425).

Comment 22

line 30 p19 : could be interesting to mention that these results may be strongly dependent on the setup used. With in particular, here, a small number of parameters.

Response: We added the following clarification in the end of the paragraph (now lines 432 and 433):

> "We remark that these results, in particular the speed of convergence of history matching, generally depend on the specific setup."

Comment 23

line 43-44 : this discussion is interesting and important. You may spend more on it and make the link with the spread on the global radiative metrics in PPE5 while the physics parameters were fixed previously.

Response: We thank the reviewer for the suggestion. We expanded the discussion as follows (now lines 449 to 454):

"Indeed, while with our analysis we are able to identify which parameters are influential for the chosen metrics (see Section 3.2.3), we cannot establish a clear hierarchy of which of these should be tuned in a sequential manner. This is exemplified by Figures 4 and 8, with the PPEs showing a large spread in the global radiative metrics despite some of the physics parameters being kept fixed. Furthermore, accounting for all parameter dependencies and feedbacks could be particularly important for tuning coupled models, e.g., for properly accounting for the interactions between atmosphere and ocean"

Comment 24

line 51 p19 : good to remind that the number of real simulations is still the limiting factor for tuning. But it would be worse with any other method available.

Response: We thank the reviewer for the suggestion. We changed the sentence as follows (now lines 458 to 460):

"We also note that even though history matching is constructed to minimize the number of climate model simulations for the PPEs, this number is still the major computational bottleneck in tuning, which gets worse when tuning models at resolutions higher than the one considered here."

Comment 25

End of conclusion : I partly disagree, but on a rather fundamental level, with the last paragraph "[...] the seamless integration of such methods within the specific climate modeling framework - to practically enable a largely automatic application - is an aspect that needs to be addressed in further studies. We foresee that incorporating the other tuning steps, such as sensitivity analysis and choice of tuning parameters, their exploration and the evaluation of the outcomes in an automated approach will lead to more accurate and potentially computationally cheaper model tuning, also making this important step in climate model development more objective and reproducible." Of course history matching allows us to make, in an objective, efficient and reproducible way, things which were very hard to conceive and formalize before. However, the choice of metrics has to and will remain subjective, given the dimension and complexity of the system. Using history matching with different metrics, either process oriented, or end-user oriented, not only may be relevant depending on the target applications, but also may help to go much more in depth into the link between physics content and climate simulations performances. This is why we are convinced that History Matching is opening a new area in climate research and model tuning. And being able to make subjective choices more objective or at least quantifiable (through parameters ranges, metrics choices, targets and tolerances) will allow sharing, improving, making tuning more efficient in the future. But I am convinced we should absolutely avoid starting to propose standardizing and automatizing those choices. We need diversity. We should promote different teams trying different ways of using it. Not only for improving simulations but also to understand the climate system better through numerical modeling.

Response: We agree with the reviewer in saying that the choice of the tuning metrics will remain subjective, given the differences among different climate models as well as to potentially achieve a deeper understanding. In fact, we have specifically not mentioned

any automatization of the choice of the tuning metrics in this outlook, for exactly the same reasons the reviewer pointed out. We still believe that model development must be a synergistic effort comprising sensitivity study, parameter tuning and model evaluation, and that several aspects of these steps (not all of course) could be automatized. This automatization or standardization will of course require more studies on the choice of the specific methods and hyper-parameters therein (e.g., which type of emulator is best suited for a given task, and so on), and our statement aims at motivating interest along this direction as well. To address the reviewer's concern and avoid ambiguities, we decided to reformulate this last paragraph as follows (now lines 466 to 474):

"Finally, while here we explored the feasibility of ML-based tuning approaches to improve the tuning of climate models, the seamless integration of such methods within the specific climate modeling framework - to practically enable an automatic application - is an aspect that needs to be addressed in further studies. Some aspects of model tuning, such as the choice of tuning metrics, will remain subjective, as highly dependent on the details and complexity of the model as well as on its intended uses. Other steps however, such as sensitivity analysis and selection of tuning parameters, their exploration and the evaluation of the outcomes could be incorporated, at least partly, in an automated approach. It is therefore important to understand which design choices are best suited for such automatic approaches, as we foresee that these will lead to more accurate and potentially computationally cheaper model tuning, also making this important step in climate model development more objective and reproducible."