Response to Reviewer #3

First of all, we would like to thank the reviewers for the comments on our manuscript. We appreciate all valuable suggestions, which helped us improve the quality of the paper. Following the reviewers' suggestions, we have improved the method to select optimal and diverse candidates. It is now based on a clustering analysis, and sensitivity tests regarding the number of clusters to consider is presented in an appendix of the paper. Given this new method, we found the need to consider around 10 clusters to explore the diversity of error patterns in the ensemble. Therefore, 10 candidate calibrations have been tested in the climate model and are presented in the paper. Most of the figures related to the candidate selection and their associated comments will be updated based on these modifications. Another appendix will be added regarding the choice and evaluation of the statistical models, and the paper will be slightly restructured to present all the methodological details in the Method section. Note also that, as requested by the reviewers, the title will be modified to : "Exploration of diverse solutions for the calibration of imperfect climate models".

Our responses to the reviewer #3 comments are described below. In the following, the comments from reviewers are in black, our answers are in blue.

General comments

1) The Introduction is rather terse and could be more welcoming for a broader audience. It would also be an improvement if the authors would discuss which fields are generally used in the calibration of climate models. Is 'calibration' the same process as what is often referred to as 'tuning'?

We will add a first paragraph in the Introduction that should help the paper to be more welcoming to a broader audience, as well as better explain the concept of model calibration (or 'tuning', as it is often referred to).

New paragraph in the Introduction : "General Circulation Models (GCM) and Earth System Models (ESM) are the primary tools for making projections about the future state of the climate system. It is an important goal of climate science to continually improve these models and to better quantify their uncertainties. Constraints on computational resources limit the ability to resolve small-scales mechanisms, and sub-grid parametrizations are used to represent processes such as atmospheric convection or clouds. These parametrizations are based on numerous poorly constrained parameters that introduce uncertainty in climate simulations. Therefore, climate models are subject to a challenging calibration (or 'tuning') problem. When used as tools of projection of future climate trajectories, they cannot be calibrated directly on their performance. Instead, assessment of performance and skill arises jointly from confidence in the understood realism of physical parametrizations of relevant physical processes, along with the fidelity of model representation of historical climate change.

Practical approaches to model calibration are subject to both data, time and computational constraints."

It would be nice if the authors would define the concept 'emulator'. It seems from the examples that this is here understood as statistical tools to interpolate between the parameters of which you have climate model experiments. However, it is my understanding that emulators also can be simple physical models.

We consider emulators to be a computationally efficient approximation of a more complex model. In the literature, these can be either purely statistical (as is the case here), or quasi-physical (as for simple climate models). In either case, there are degrees of freedom which can be adjusted to reproduce certain aspects of complex model response. The term 'emulator' has long been used in PPE analysis (Sanderson et al. (2008)), where mapping is from the perturbed parameter values to a selection of outputs of the climate model .

However, given the increased use of the term 'emulator' to represent simple climate models in, for example, IPCC applications, we propose to replace the term 'emulator' in the Introduction with 'statistical model' or 'statistical predictions'. Moreover, the Section **2.4 Emulator design and optimization** will be re-named **2.4 Statistical model and optimization** and will clearly define the concept of "emulations", that refers to the statistical model predictions.

Reference : Sanderson, B. M., Knutti, R., Aina, T., Christensen, C., Faull, N., Frame, D. J., ... & Allen, M. R. (2008). Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *Journal of Climate*, *21*(11), 2384-2400.

Where does the word 'spatial' in the title come from? I guess the method is rather general although you apply it here to spatial fields.

We agree that the title of the paper was not the most appropriate and we changed it to : "Exploration of diverse solutions for the calibration of imperfect climate models".

2) It seems to be assumed that observations are perfect. This is not the case and often there are also errors originating from the finite sampling: the estimated climatology in both observations and models depends on the length of the time-series. How will these errors impact your results?

Indeed, in this study, we did not consider the uncertainty associated with observations, but additional analyses suggest that our results are sensitive to the observational dataset used. As a rough illustration, we present here (Figure 1), the projection of different surface temperature observations and reanalyses (BEST, ERA5, NCEP, NOAA) on the PPE EOF basis. We also selected candidates minimizing the error associated with each one of the projected observations. It appears that the distances between candidates in the EOF space is as large as the spread of the 5% best performing

models when considering a single observation dataset (BEST here), suggesting that considering a different observational dataset would affect our results to some degree. But the uncertainty range represented in the top 5% of models is representative of what we should expect from observational uncertainty.

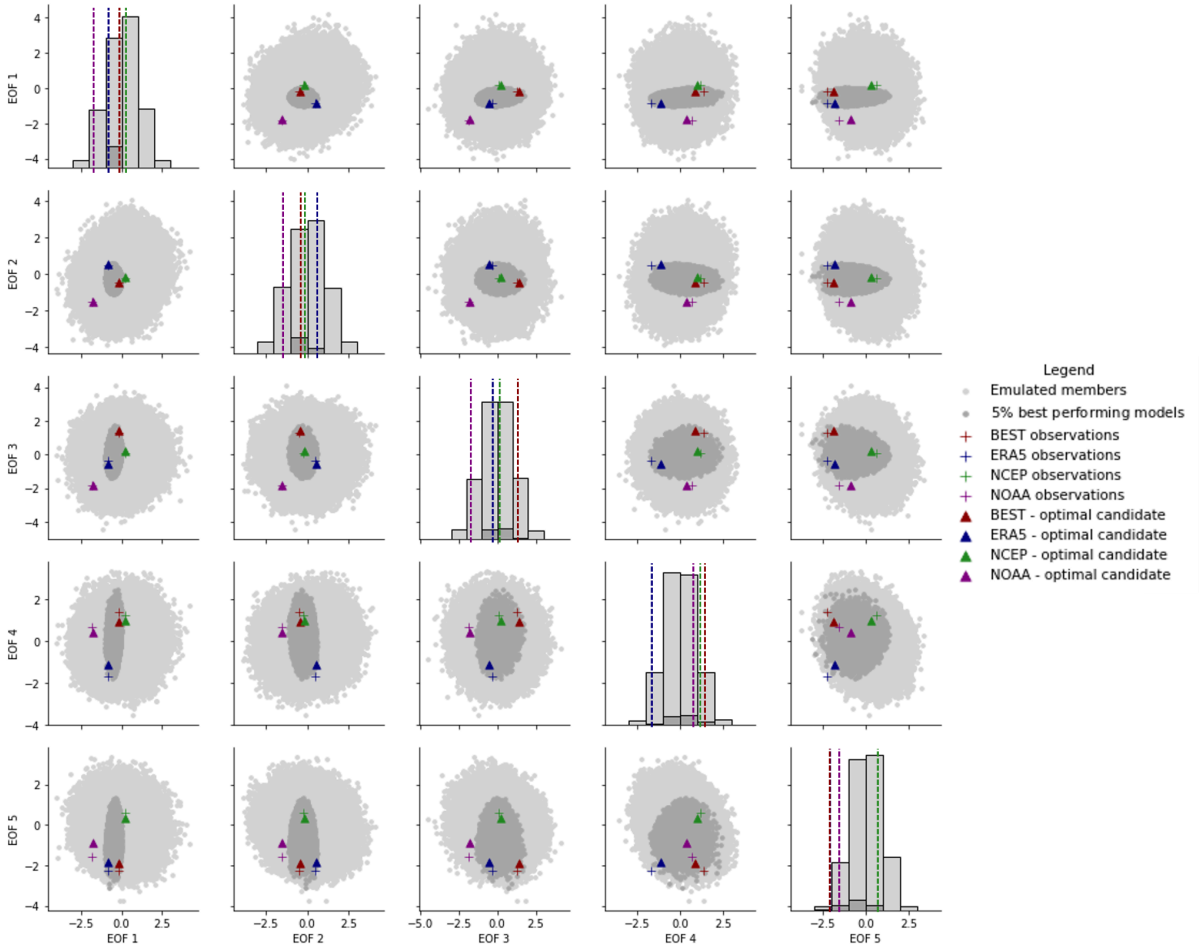

Fig 1. Standardised principal components associated with the 5 first modes of the surface temperature EOF basis computed on the 102 members of the PPE. The Figure presents the projections of 1 000 000 emulated members (light gray), the 5% best emulated members compared to the BEST observations (dark grey), the projections of different surface temperature observations (colored crosses) and the 4 candidates CNRM-CM minimizing the error compared to a given observational dataset (colored triangles). Note that this analysis uses a different and larger LH sampling of the parameter space than what is presented in the paper, therefore the grey dots will not match the Figures of the paper.

Finding a formal way to include observational uncertainty in our method for candidate selection is beyond the scope of this paper. But we will add a paragraph in the conclusion to highlight this limitation and suggest the consideration of observational uncertainty as a perspective to improve the method.

3) The emulator used in this study seems to be linear multiple regression. Furthermore, the methodology of the analysis is based on EOF/PC analyses. But other emulators are non-linear. The output of such emulators are therefore not limited to the linear space spanned by the EOFs. So how would your analyses change with a non-linear emulator and how will it change your conclusions?

The emulators used in this study are Multi Linear Regressions (MLR), which take the model parameters as input and predict the Principal Components of the different climatic fields, used to reconstruct the 3D fields and the parametric component of the model error compared to observations. This choice was made because, given the modest sample size of perturbed ESM runs available, the linear model outperformed nonlinear techniques - implying that higher order parameter effects are not robustly sampled.

As a better discussion about the choice and performance of the emulators was requested by the Reviewers #1 and #2, we have decided to add an appendix showing the performance of our statistical model and comparing it to linear regression with penalization (LASSO model) and a non-linear emulator (a Random Forest).

In 10 random selections of out-of-sample test sets, we obtain an average correlation of 0.7 between the predictions and the true values of total error (Figure 1 (c)), with a RMSE between predictions and true values representing 8% of the total parametric error (Figure 1 (f)), which is sufficient to validate the use of this model for our study. However, results suggest that there is room for improvement, especially in the prediction of the LW error, and that another model could improve the predictions, as it is the case of the Random Forest model. The error bars associated with the prediction of the total error suggests that the MLR performance is sensitive to the test set selected and that the model will perform unevenly across the parameter space. Thanks to variable selection and regularization, the Lasso model seems a bit less sensitive to the test set selection for the prediction of total error, but the prediction of LW error is still a limitation. Figure 1 will be added in an appendix and we will add sentences in the main paper clearly stating the limitations of the emulators.

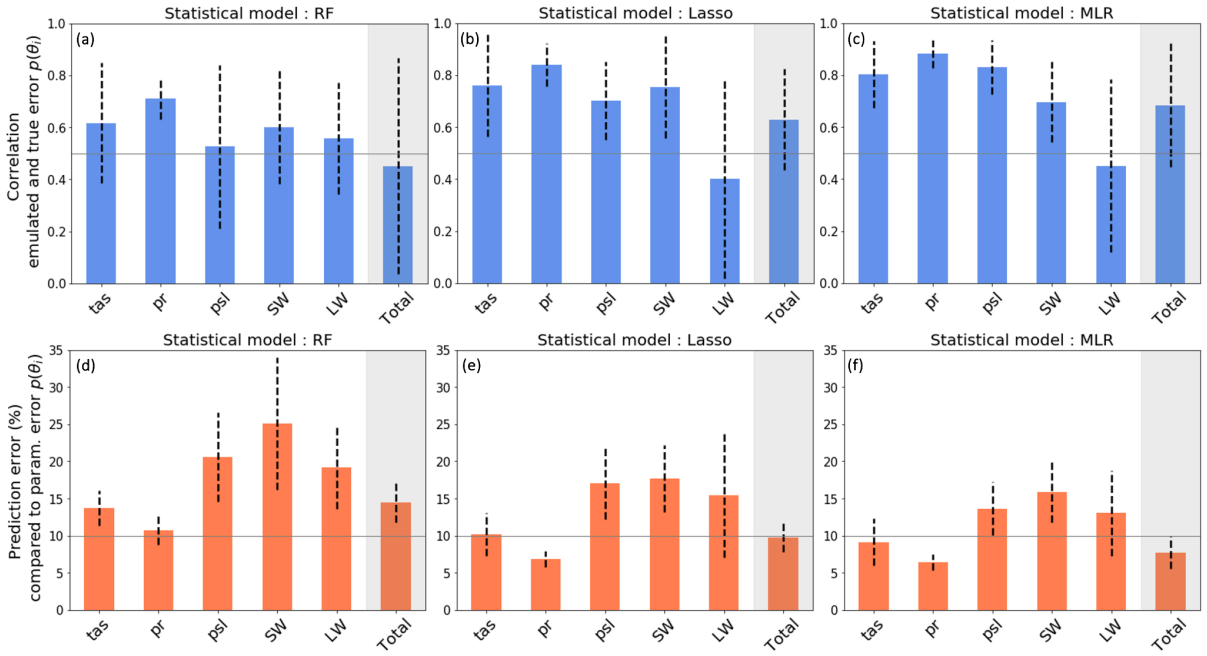

Fig 1. Correlations and RMSE (in % compared to the true values) between emulated and true parametric components of the errors within a test set representing 10% of the dataset. The evaluation is repeated 10 times with random sampling of training and test sets and the mean and standard deviation among these 10 evaluations are represented by the bars and the dashed lines, respectively. Performances are shown for (a), (d) a Random Forest, (b), (e) a LASSO regression and (c), (f) the Multi Linear Regression used in this analysis and. The EOF truncation lengths used to compute the parametric error are presented in Figure 2 and 5 of the paper.

In conclusion, it seems that using a non-linear emulator could improve certain aspects of the predictions, though our capacity to train such emulators is fundamentally limited by the sample size available in the dataset, and enhancing the size of the ensemble would be a necessary prerequisite to try to improve our statistical predictions. Gaussian Processes are non-linear models often used in PPE analysis and even though we did not test them in this study, we will add sentences in the conclusion suggesting it as a potential perspective to this work.

4) The analytical deviations can in places be hard to follow. I would suggest that a notation is used that differentiates between scalars, vectors, and matrices.

Thank you for the suggestion. After taking into account the feedback from all the reviewers, the point that seems the most confusing when trying to follow the equations is to identify whether we are talking about emulated scores or outputs of the climate models. To improve this aspect of the manuscript, we have decided to keep the notations as they are, but to use $p_{em,j}(\theta_i)$ for the emulated scores and $p_j(\theta_i)$ for the actual scores, in order to differentiate whether we discuss statistical predictions or climate model outputs.

5) I am also confused about the optimisation described in section 2.4 and 2.5. As far as I can see you don't apply a minimizing method but generates a lot of emulations with different parameters. Then you find a set of parameters with error smaller than the error from a reference model (what is this). And then you prune that set to a much smaller set (2.5). Is this correct? So the set you find are then not local minima?

Reviewer #1 also requested that we clarify the notion of 'optimality' in the paper, so we will be using the term "near-optimal" to refer the vectors of parameter values $\hat{\theta}_i$ associated with emulated parametric errors lower than the reference model configuration CNRM-CM6-1.

The reference model configuration CNRM-CM6-1 results from a tuning by the developers for the CMIP6 exercise. This tuning was done following the historical common practices for tuning a climate model (Hourdin et al. (2017), Schmidt et al. (2017)) and has been validated by model developpers. This reference model will be better defined and cited in the paper : "The reference model will be the model CNRM-CM6-1, tuned by the model developers for the CMIP6 exercise (Roehrig et al.

(2020)). This reference model has been validated by the experts and can serve as a threshold to define whether a model calibration is near-optimal."

References :

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., ... & Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589-602.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J. C., Hannay, C., ... & Saha, S. (2017). Practice and philosophy of climate model tuning across six US modeling centers. *Geoscientific Model Development*, *10*(9), 3207-3223.

Specific comments

43: This is the first time PPE is used in the main text. It should be spelled out here.

Agreed

64: Salter et al. should not be in ().

Agreed

l85: So n=102 and each \Theta_i is a vector of length 30? As mentioned above it would be helpful if the notation separated vectors from scalars.

Yes - θ_i is a vector of length 30, which is an input of the climate model $f(\theta_i)$ and the PPE $F = (f(\theta_1), \dots, f(\theta_n))$ contains $n = 102$ simulations $f(\theta_i)$ that can have l grid points (finite elements).

l106: \mu seems to be on the left-hand side in Eq. 105, so does r_f include \mu here?

Agreed that, as said in the text, r_f should include \mu, which is not obvious looking at the equation. We will remove \mu from equations (2) and (3) and explain in the text that f_f and r_y include \mu.

Eqs. 6 and 7: c_f → c_y ?

Yes, this should be c_y

Eq. 14: So this is multi-linear regression. I am again confused about notation. Should \Theta_i just be the vector \Theta? Theta_i is the vector of parameters used in the i'th climate model?

With \theta = (\theta_1,...,\theta_n), \theta_i is a particular vector of 30 parameter values, that is used as input of the climate model f(\theta_i) or the emulator : c(\theta_i) = \beta \theta_i + c0 with \beta a vector of 30 coefficients of the multi linear regression.

l174: What is a reference model? Is there any reason to assume that this is better than any random selected model from the ensemble?

The reference model configuration CNRM-CM6-1 results from a tuning by the developers for the CMIP6 exercise. This tuning was done following the historical common practices for tuning a climate model (Hourdin et al. (2017), Schmidt et al. (2017)) and has been validated by model developpers. This reference model will be better defined and cited in the paper : "The reference model will be the model CNRM-CM6-1, tuned by the model developers for the CMIP6 exercise (Roehrig et al. (2020)). This reference model has been validated by the experts and can serve as a threshold to define whether a model calibration is near-optimal."

References :

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., ... & Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589-602.

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J. C., Hannay, C., ... & Saha, S. (2017). Practice and philosophy of climate model tuning across six US modeling centers. *Geoscientific Model Development*, *10*(9), 3207-3223.

Section 3, beginning: You consider the SAT from 3 years but what is more precisely used here? The annual means?

We use the annual mean averaged over the 3 years - this will be added in the paper

l211: Section ??

Ok

In Fig. 1 left: It is not clear to me what the hatched regions show. Is the non-parametric error only shown for q=102?

We agree that Figure 1 included too many hatched regions, making the legend a little bit confusing. We removed most of the hatched regions to simplify the figure. Here, we highlight the fact that a non-parametric component of the error exists for a perfect reconstruction of q=102 and that this non-parametric component does not evolve much when considering a lower truncation of q=50.

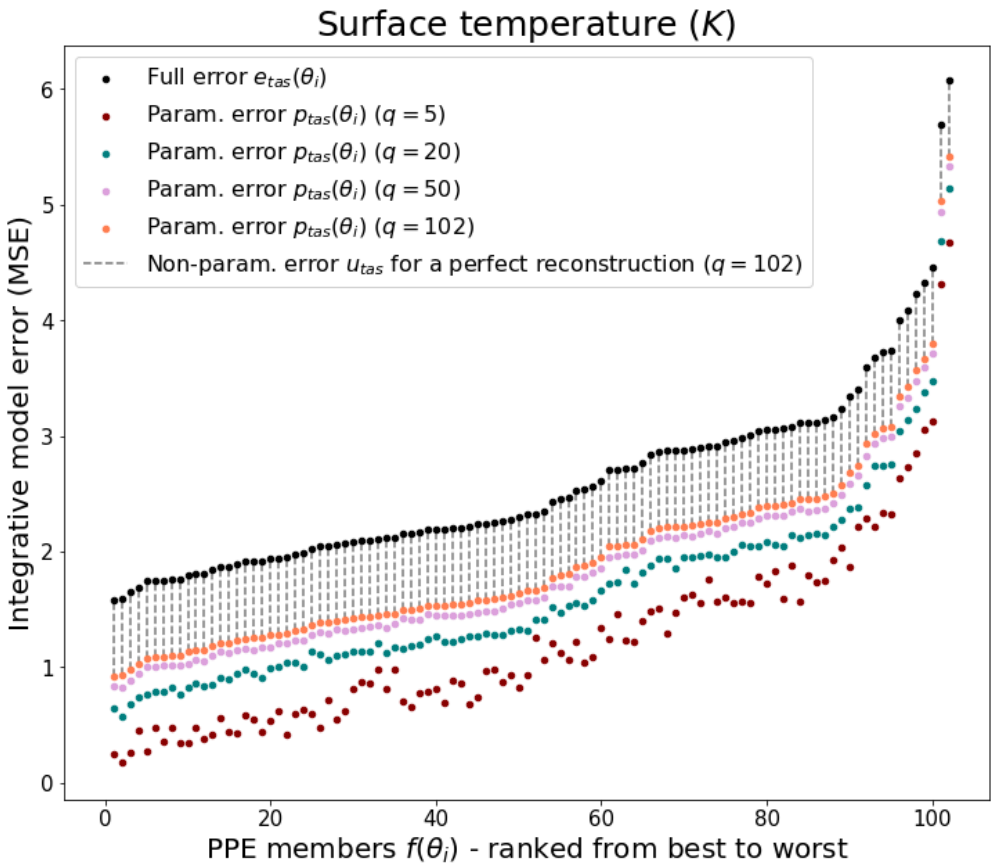

Fig 2. Figure showing the full model error $e_{tas}(\theta_i)$ and its parametric component $p_{tas}(\theta_i)$ for different truncation lengths : q=5 (red dots), q=20 (blue dots), q=50 (pink dots), q=102 (orange dots). The grey hatched region shows the non-parametric components of model errors for a perfect reconstruction of q=102.

l232: We follow section 14? Eq. 14?

Yes, it should be Eq. 14

l239: It is not clear to me what in-sample refers to here.

"In-sample" refers to the sample used for the training of the emulator and "out-of-sample" refers to a test set that was not used for the training. The out-of-sample test set is used to evaluate the prediction skill of the emulator.

l257: I think this is the first time in the paper GMMIP is mentioned. What is it?

Agreed - new sentences here : "In the context of the Global Monsoons Model Inter-comparison Project (GMMIP) (Zhou et al. (2016)), an ensemble of 10 atmospheric-only simulations of the CNRM-CM6-1 was run. In this ensemble, the reference model calibration was used, the SST was forced with the same observations as the PPE and the members differ by their initial conditions only. This dataset can be used to consider the effect of internal variability on the error decomposition, and will be referred to as the GMMIP dataset."

Reference : Zhou, T., Turner, A. G., Kinter, J. L., Wang, B., Qian, Y., Chen, X., Wu, B., Wang, B., Liu, B., Zou, L., and He, B.: GMMIP (v1.0) contribution to CMIP6: Global Monsoons Model Inter-comparison Project, Geosci. Model Dev., 9, 3589–3604, https://doi.org/10.5194/gmd-9-3589-2016, 2016.

l283: What does LHS mean?

"Latin Hypercube Sampling (LHS)" will be added line 82, when the term first appears

Fig. 3: The caption should also describe the plots along the diagonal.

Agreed