Response to Reviewer #2

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 #2 comments are described below. In the following, the comments from reviewers are in black, our answers are in blue.

Major comments

This study is really interesting and brings a new view of the calibration problem that is in the scope of ESD. The authors described the state-of-the art and contextualise their own method. The new method is described, tested and analysed. This seems to be completed. I really appreciate the fact that author went in details in the method section. Figures show sufficient results to support the interpretation and conclusions. However, I think some points must be addressed before to consider a publication.

Structure : the paper is quite well-structured, but sometimes, some methodology details are presented in the results section or even in the figure captions. I think, the author must put all the method elements in the method section.

Agreed - we will restructure the paper to put all the methodology elements in the method section. The definition of the near-optimal subset of emulations and the selection of diverse candidates will be presented for both the univariate and multivariate applications in the method section, instead of later in the paper for the multivariate case.

Method

1. To built the emulator, only 102 simulations are considered for a uncertainty due to 30 parameters. I think 102 simulations is clearly not enough to built a robust emulator. Generally, we consider 10 simulations by number of parameters. So here, the authors needs something like 300 simulations, that is very far away from the 102 simulations considered. Even if the emulator shows very good

results, the fact that an emulated optimized candidate crash when the author use the real model suggests that there is a lack of data in the training process that leads the emulator to be not enough constraint. I think the author should better evaluate the metamodel in order to demonstrate that in their case 102 simulations is enough.

2. The cross validation method is not clearly presented and the method seems to change between the univariate and multivariate case: the training data is 90% of the data in the univariate case, while 80% in the multivariate case. Furthermore, using 90% to train the emulator, means using 92 simulations. As already said, I am not sure it is enough. Authors should be consistent and use the same percentage to define the training and testing datasets.

3. The evaluation of the emulator is also problematic. The emulator is evaluated by analysing spatial patterns or correlation between predicted and true data. A emulator must be evaluated by calculating an error (not a correlation).

As a better discussion about the choice and performance of the emulators was also also requested by the Reviewer #1, we have decided to add an appendix showing the performance of our statistical model to emulate the parametric component of the individual and total model errors. We also compare their performances to other statistical models : a Random Forest and a LASSO regression. Finally, as suggested by reviewer #2 , we have added an evaluation of the emulators based on the root-mean square error (RMSE) between the predictions and the true values of the parametric components of the model errors, in addition to the evaluation based on the correlation.

First of all, the ensemble size of the PPE is very limited (102 simulations) and our capacity to train emulators is fundamentally limited by the sample size available. The emulators used in this study are Multi Linear Regressions (MLR) taking the model parameters as input and predicting the Principal Components (PC) used to reconstruct the 3D variables and the parametric model errors when comparing with observations. 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.

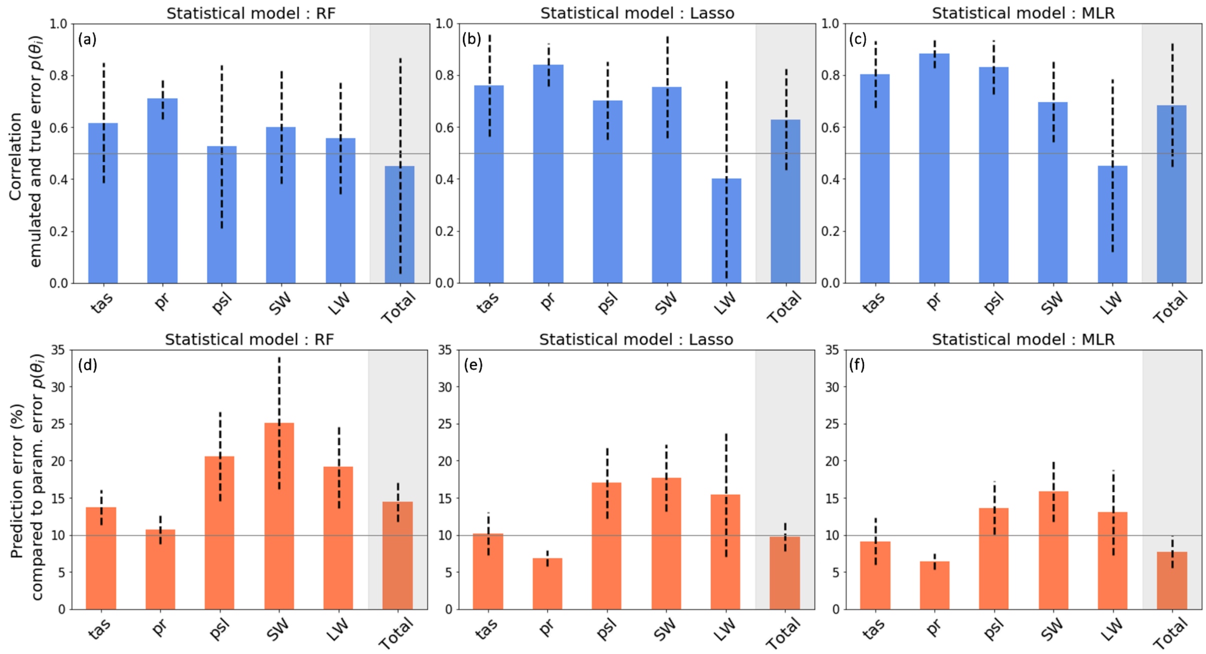

Fig 1. Correlations and RMSE (in % compared to the true values) between emulated and true parametric components of the errors within a test set representing 10% of the dataset. The evaluation is repeated 10 times with random sampling of training and test sets and the mean and standard deviation among 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.

These results suggest that there is room for improvement, especially in the prediction of the LW error, and that another model could potentially 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.

In conclusion, it seems that using a non-linear emulator could improve certain aspects of the predictions, though enhancing the size of the ensemble would be a necessary prerequisite to try to improve our statistical predictions. Figure 1 will be added in an appendix and we will add sentences in the main paper clearly stating the limitations of the emulators. Gaussian Processes are statistical 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. To define the truncation of the EOF modes, the author used a threshold on the correlation. I think it is notrobust. I suggest to define a threshold on the variance.

The idea of a threshold based on the correlation between true and predicted values is to rule out the EOF modes poorly emulated by our statistical models. We argue that these higher modes will introduce noise in the analysis, and are not strongly affected by the change in parameter values anyway, as they explain a very small part of the PPE variance. However, we also want to make sure that we consider enough EOF modes to represent a large fraction of the PPE variance. There was already an implicit threshold, since all of the truncations chosen were explaining more than 85% of the PPE variance. In order to make this point clearer in the paper, we added another y axis on Figures 2 and 5 to show the % of explained variance depending on the truncation, and we draw the threshold of 85% for each variable to illustrate that our chosen truncation is explaining more than this threshold. An example of the modifications is shown here for Figure 2, which is the Figure 2 of the paper (Figure 5 will also be updated accordingly).

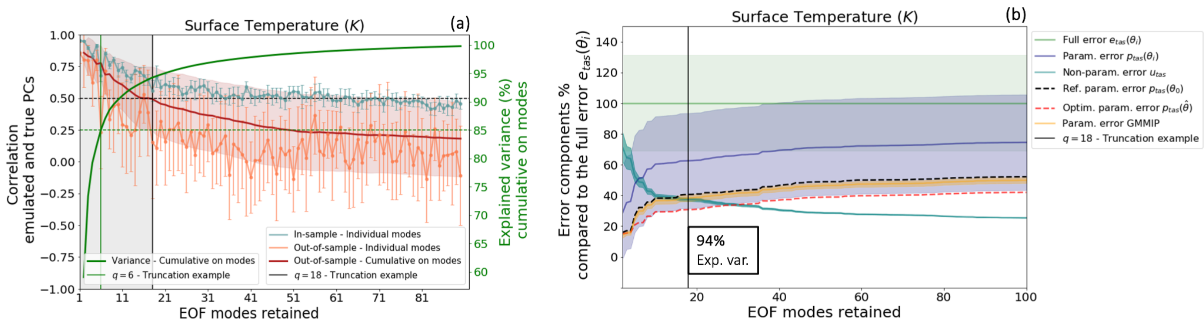

Fig 2. Truncation choice based on parametric emulation and error decomposition. The left plot (a) shows the correlation between the emulated and true PCs of the surface temperature EOF, for the different modes of variability. The correlation is shown within the training set (blue curve) and the test set (orange curve). The evaluation is repeated 10 times with random sampling of training and test sets, the mean and the standard deviation among these 10 evaluations are represented by the dots and the error bars, respectively. The red curve and shading is the mean correlation averaged over the modes cumulatively. The solid green curve represents the percentage of variance explained when retaining up to x modes of the EOF. The dashed green horizontal line shows a threshold of 85% of explained variance and the solid vertical line is the truncation length needed to satisfy this threshold. The right plot (b) shows the ratio of the error components compared to the full error $e_{tas}(\theta_i)$ (in green) as a function of the number of modes of variability retained. The lines are the ensemble means and the shadings represent the standard deviations. The plot shows the ratios of the PPE parametric error (dark blue), the PPE non-parametric error (light blue), the reference calibration parametric error $p_{tas}(\theta_0)$ (red dotted curve) and the GMMIP parametric error (orange). An example of truncation at q = 18 is represented on both plots by the black vertical line.

5. The authors do not explain how they get the number of optimized candidates n_k. It seems to be completely subjective.

We agree that the number of optimized candidates was completely subjective, which was a design choice. Following reviewers comments, we have changed our methodology and replaced the previous algorithm by a clustering method, applied to the Principal Components of the different fields, normalized by the Principal Components of the reference model. WWe have chosen the Euclidean distance as a measure of similarity in the clustering analysis and we have selected the centroids as our set of "diverse candidates". As a result, we rewrote the paper Section "**2.5 Selection of diverse candidate calibrations**" to present the clustering analysis.

Moreover, the clustering analyses are sensitive to the choice of cluster numbers *k*, which depends on the dataset to be classified. Figure 3 presents a sensitivity test of the k-means analysis to the number of clusters for the univariate and multivariate application. The inertia is defined as the sum of the squared distances between each data point and the centroid within a same cluster. The Elbow method consists in finding the inflexion point in the k-means performance curve (Figure 3 (a), (c)), where the decrease in inertia begins, to find the good trade-off for the number of clusters. Another criteria we looked at is the Dunn index (Figure 3 (b), (d)) : the ratio between the minimal inter-cluster distances and the maximal intra-cluster distances. A higher Dunn index represents a higher distance in between the centroids (clusters are far away from each other) and a lower distance in between the data points and the centroid of a same cluster (clusters are compact).

Even though it is not as pronounced in the multivariate application, the inertia sensitivity test suggests that we should choose a value of *k* in between 8 and 18 to be in the Elbow of the curve. Then, it appears that we should not take a value of *k* too high, as the Dunn index tends to decrease for both applications for a number of clusters higher than 10. Based on these two criteria, we have decided to keep 10 clusters for the analysis and we carried out the analyses for 10 optimal and diverse candidates. Figure 3 will be part of an appendix detailing the choice of *k*.

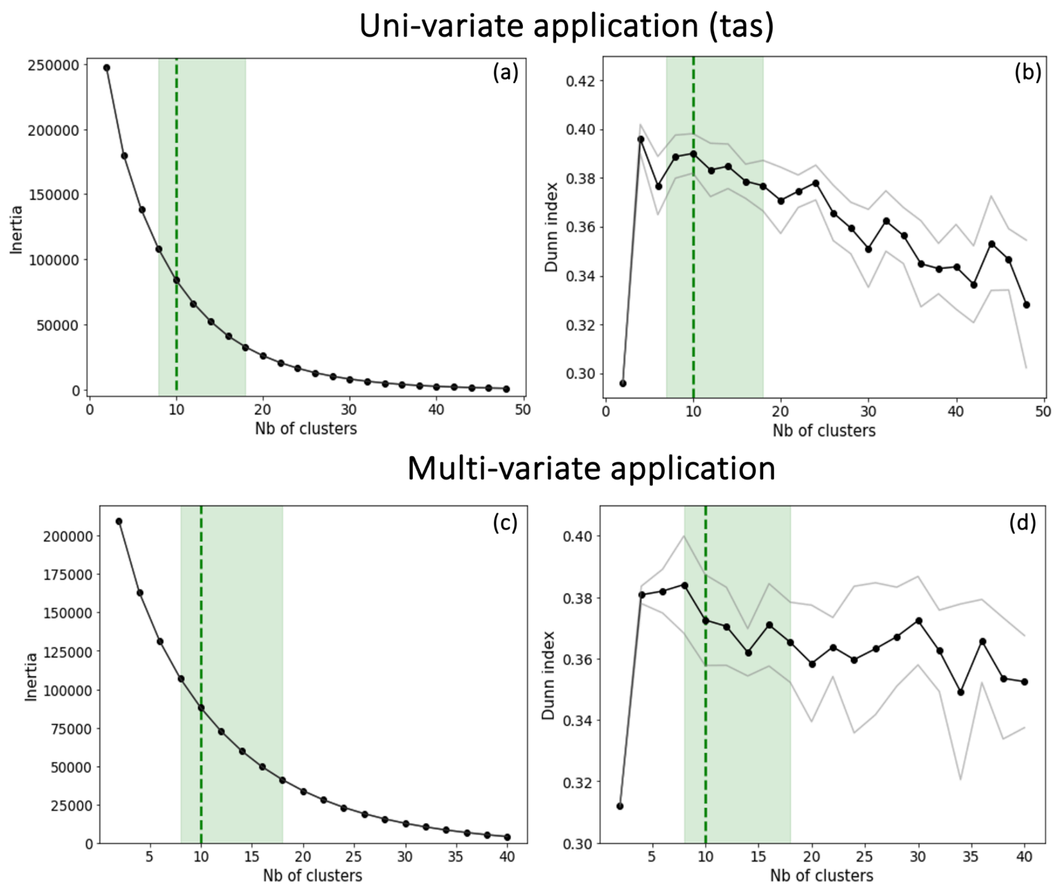

Fig 3. Sensitivity test of the clustering analyses for the uni-variate (first row) and multi-variate (second row) applications. The inertia criteria ((a), (b)) and the Dunn indexes ((b), (d)) are shown depending on the number of clusters (x axes). The green shaded areas present the acceptable number of clusters following the Elbow method applied to the inertia. The green dashed line shows the number of clusters retained for our analyses : k=10 in both applications. The grey thin lines show the $\pm$ 1σ range for a repetition of 10 evaluations with random selection of initial centroids (the range is negligible for the inertia).

6. Some inconsistencies have been found. For example, at some points, the authors used 10^6 emulated simulations and other points, they used 100.000.

Yes, this was a mistake in the text, we are using 100 000 everywhere, the mistake will be corrected.

7. The reference simulation is not presented.

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.

Title: This study does not really deal with spatial calibration. It is more a calibration of CNRM-CMIP6 parameters taking into account a spatial uncertainty. I suggest to reformulate the title. For example: "Considering an ensemble of calibration in CNRM-CMIP6 in order to represent spatial model uncertainty"

We agreed and changed the title for : "Exploration of diverse solutions for the calibration of imperfect climate models".

Abstract: The abstract do not fully synthesise the paper. I mean, some results (concerning the comparison with perturbed initial condition) are mentioned in the abstract while their appear in few sentences in the paper and are even not mentioned in the conclusions. While, some main results are not present in the abstract.

Agreed - the abstract will be updated considering the new results and the sentence about the comparison with the initial condition ensemble will be added in the conclusion as well.

Abstract : "The calibration of Earth System Model parameters is subject to both data, time and computational constraints. The high dimensionality of this calibration problem, combined with errors which arise from model structural assumptions, makes it impossible to find model versions fully consistent with historical observations. Therefore, the potential for multiple plausible configurations presenting different tradeoffs between skills in various variables and spatial regions remains usually untested. In this study, we lay out a formalism for making different assumptions about how ensemble variability in a Perturbed Physics Ensemble relates to model error, proposing an empirical but practical solution for finding diverse near-optimal solutions. We argue that the effective degrees of freedom in model performance response to parameter input (the 'parametric component') is, in fact, relatively small, illustrating why manual calibration is often able to find near-optimal solutions. The results explore the potential for comparably performing parameter configurations making different trade-offs in model errors. These model candidates can inform model development and could potentially lead to significantly different future climate evolution."

Clarity of the paper: Some sentences of the paper must be clarified and some inconsistency have been found (see Minor Comments). Furthermore, as it is not clearly specify if they analyse the results of the emulator or the results of the real model, it is quite confusing.

We will answer each Minor Comment in the next section. Moreover, we will note the emulated scores $p_{em,j}(\theta_i)$ and the actual scores $p_j(\theta_i)$ in the paper, in order to differentiate whether we discuss statistical predictions of climate model outputs.

Minor comments

Abstract

1. L. 3: "it impossible to find one model version"

Ok

2. L. 5: Delete the PPE abbreviation (not necessary in the abstract)

Agreed

3. L. 9: Results on initial conditions perturbation are clearly not a main result of this study. I suggest to do not putit in the abstract, but instead, to present the two applications (uni-variate, multi-variate cases)

Agreed

Introduction

1. L. 21 - 24: Sentence too long. Please, rewrite it.

Agreed - new sentence : "However, more complex models such as GCM present a number of difficulties for objective calibration which have resulted in a *status quo* in which manual calibration remains the default approach (Mauritsen et al., 2012; Hourdin et al., 2017)."

2. L. 25: "Robust". I am not sure this a right way to say it.
3. L. 24 - 27: Sentence too long. Please, rewrite it.

Agreed - new sentences : "Such approaches have not yet been operationally replaced by objective calibration approaches, but leave large intractable uncertainties. In particular, the potential existence of comparably performing alternative configurations with significantly different future climate evolution (Hourdin et al., 2023; Ho et al., 2012)."

4. L. 28: As the ensemble model approach is inserted here, maybe it is better to insert here also the notion of PPEmodel.
5. L. 35: Put the list of emulators in another sentence.

We inserted a sentence line 34-35 to introduce the notion of PPE and rewrite the list of emulators :

"Approaches to date with GCMs have mainly relied on Perturbed Parameter Ensembles (PPE) of simulations, allowing an initial stochastic sample of the parametric response of the model. The construction of meta-models is then needed to emulate this parametric response and enhance the number of samples. The meta-models can be quadratic (Neelin et al., 2010), logistic regression (Bellprat et al., 2012), Gaussian process emulators (Salter and Williamson, 2016) or neural networks (Sanderson et al., 2008)."

6. L. 40: Can you clarify the "low dimentionnal output space" ?

It is already written L40 "low dimentionnal output space (such as global mean)".

Indeed, the emulators can not predict the high dimensionality of a model output, with the numerous grid points and time steps. The example of global mean given L.40 is rarely used in PPE analysis and the EOF (or PCA analysis) presented L.57 is the most common way to reduce the output dimensionality for such applications.

L.57 : "In order to reduce the complexity of the emulation problem, and to preserve the covariance structure of the model output, it is common to reduce the dimensionality of the output through Principal Component Analysis (PCA) (Higdon et al., 2008; 60 Sexton et al., 2012; Wilkinson, 2010)."

7. L. 44: Maybe add a comment on the fact that it depends on subjective advises from physical experts.

Agreed - L.44 : "Emulators can be improved in promising sub-regions of the parameter space by running a new PPE in a reduced parameter space to increase the ensemble density (sometimes referred as an "iterative refocussing" approach) (Williamson et al., 2017). However, the choice of which region to initially focus on depends on advice from model developers and is itself subject to error in emulation."

8. L. 39: Does the "NROY" abbreviation really need to be defined as it is only used once ?

No, agreed

9. L. 48 - 52: Sentence too long. Please, rewrite it.
10. L. 48: I do not understand the "grid scale". I suppose you want to say "represent spatial performance"
11. L. 48: Clarify that it is for the development of the model.
12. L. 50: Use "grid points" instead of "pixels"
13. L. 46 - 56: Please, clarify here that you speak about spatial calibration.

New sentences L.46 - 52 : "Climate models produce high dimensional output across space, time and variable dimensions. Performance is often addressed by integrated output spanning these dimensions (Gleckler et al., 2008; Sanderson et al., 2017) and so calibration techniques must be able to represent spatial performance in order to be useful to development. In a low dimensional space defined by global mean quantities, it is possible to find one model version which is consistent with observations (Williamson et al., 2015), but this is not true when considering the high dimensionality of climate model outputs. When considering an assessment of model error integrated over a large number of grid points and variables, structural trade-offs may arise between model outputs which cannot be simultaneously optimized by adjusting model parameters."

14. L. 54: "addition of additional", please rephrase.

Agreed - L.54 : "In another case, structural errors in an atmospheric model were found to increase significantly with the addition of variables to a spatial metric (Sanderson et al., 2008)."

15. L. 53: "structural". What do you mean by this word ?

We refer here to the impossibility to "tune down" all the variables simultaneously, and to the fact that the discrepancy term associated with an optimal parameterization increases when adding variables to the multi-variate score considered.

Citation from (Sanderson et al. (2008)) : "Requiring models to match all observations simultaneously proved a more difficult task for all of the ensembles. The ANN simulated ensemble suggested that model parameters could at best be tuned to a compromise configuration with a finite error from the observations. This “best model discrepancy” was found to increase with the inclusion of increasing numbers of separate observations, and was not itself a strong function of S.”

16. L. 59: ”(PCA; e.g. Hidgon et al., 2008; Sexton et al., 2012; Wilkinson et al. 2010; ...)

Agreed

17. L. 63: Define the PPE abbreviation here (if nor defined before)

It is now defined L.34

18. L. 72: Delete the definition of PPE abbreviation. Can you clarify the ”structural error” ? Do you mean ”spatial error” ?

No, we mean the part of the error that does not relate to the model tuning. This is the part of error arising because of processes not or poorly represented in the model. But we do not need to use this world here and we are interested in model error in general.

New sentence : “In this study, we lay out an alternative formalism which makes different assumptions about how ensemble variability in a Perturbed Parameter Ensemble relates to model error and how it can inform model development.”

19. L. 71: ”it can inform model development”. Please reformulate

We think this sentence is clear enough : the model error decomposition presented here can help identify the part of error that is not affected by model tuning, which can bring information about the limitations of a climate model. Additionally, identifying model candidates presenting diverse regional biases can be of interest for better understanding the representation of certain mechanisms in the model, or for considering parameter uncertainty in studies of climate change impact.

20. L. 77: Add also a mention to the conclusions in section 5.

Agreed

Methods

1. L. 79 - 86: Put that paragraph in a subsection named ”Used model”, for example. Ok
2. L. 79: Here, you refer the model as CNRM-CM6. Later in the paper, you use the name ”CNRM-CMIP6”, or ”CNRM-CM”. Please, use the same name everywhere and clarify with ARPEGE-Climat. If you consider that the model used is the full Earth System model, use ”CNRM-CMIP6”, while if you consider that you run only the atmospheric component, please use ”ARPEGE-Climat”.

Ok

3. L. 79: Specify that CNRM-CM6 is one of the CMIP models.

   Ok

4. L. 80: To avoid confusion with emulator, I would say ”climate model” instead of ”climate simulator”.

   Ok

5. L. 80: Definition of PPE abbreviation not necessary Ok
6. L. 81: Can you define AMIP abbreviation ?

   Ok (we added a reference)

7. L. 81: I would say ”102 simulations differing by their parameters value”. Ok
8. L. 82: That could be interesting to give some example of parameters. Do they all appears in physical parameterization ? A table with the whole list of parameter could be added in appendix.

   Ok - see Table 1, added in Appendix

Name	Minimum	Maximum	Reference	Description	Units
AKN	0.06	0.28	0.126	Strength of the turbulent mixing	-
ALPHAT	0.5	3.0	1.13	Strength of the turbulent mixing for temperature (Prandtl number)	-
ALD	0.5	3.0	1.18	Strength of the turbulent kinetic energy dissipation	-
ALMAVE	0	30	10	Lower bound of the mixing length	m
AGREF	-0.5	-0.01	-0.36	Parameter in the boundary-layer-top entrainment parameterization	-
AGRE1	0	10	5.5	Parameter in the boundary-layer-top entrainment parameterization	-
AGRE2	0	10	0	Parameter in the boundary-layer-top entrainment parameterization	-
RAUTEFR	0.5×10^{-3}	1×10^{-2}	1×10^{-3}	Inverse timescale for liquid autoconversion	s^{-1}
RQLCR	0.5×10^{-4}	1×10^{-3}	2×10^{-4}	Critical liquid water content for liquid autoconversion	kg kg^{-1}
RAUTEFS	0.5×10^{-3}	1×10^{-2}	5.2×10^{-3}	Inverse timescale for ice autoconversion	s^{-1}
RQICRMIN	0.1×10^{-5}	0.1×10^{-7}	0.1×10^{-6}	Critical ice content for ice autoconversion at low negative temperatures	kg kg^{-1}
RQICRMAX	0.05×10^{-4}	1×10^{-4}	0.21×10^{-4}	Critical ice content for ice autoconversion at high negative temperatures	kg kg^{-1}
TFVL	0.001	0.2	0.02	Falling speed of cloud water droplets	m s^{-1}
TFVI	0.001	0.2	0.04	Falling speed of cloud ice crystals	m s^{-1}
TFVR	0.1	6.0	3.0	Falling speed of rain	m s^{-1}
TFVS	0.1	6.0	0.6	Falling speed of snow	m s^{-1}
RKDN	3×10^{-5}	7×10^{-5}	5e-05	Minimum drag for the convective updraft vertical velocity	Pa^{-1}
RKDX	8×10^{-5}	6×10^{-4}	1×10^{-4}	Maximum drag for the convective updraft vertical velocity	Pa^{-1}
TENTR	2×10^{-6}	1×10^{-5}	4×10^{-6}	Minimum turbulent entrainment in the convective updraft	Pa^{-1}
TENTRX	3×10^{-5}	1×10^{-4}	6×10^{-5}	Maximum turbulent entrainment in the convective updraft	Pa^{-1}
VVN	-1	-5	-2	Critical convective updraft Vertical velocity for maximum entrainment and drag	Pa s^{-1}
VVX	-25	-50	-35	Critical convective updraft Vertical velocity for minimum entrainment and drag	Pa s^{-1}
ALFX	0.01	0.1	0.04	Maximum convective updraft area fraction	-
FNEBC	0	20	10	Parameter for computing the convective cloud fraction	-
RLWINHF_ICE	0.5	1.0	0.9	Ice cloud heterogeneity coefficient in the longwave spectrum	-
RLWINHF_LIQ	0.5	1.0	0.9	Liquid cloud heterogeneity coefficient in the longwave spectrum	-
RSWINHF_ICE	0.5	1.0	0.71	Ice cloud heterogeneity coefficient in the shortwave spectrum	-
RSWINHF_LIQ	0.5	1.0	0.71	Liquid cloud heterogeneity coefficient in the shortwave spectrum	-
RELFCAPE	0.2	10.0	2.0	Parameter used in the convection scheme Convective Available Potential Energy closure	-

Table 1. Description of the 30 parameters.

9. L. 81: Why do you choose a period of 3 years ? And why this period ? Is there a reason ?

The choice of the period was detailed in Peatier et al. (2022), it was the time needed for the feedback to stabilize

10. L. 82: I am not sure it is enough to use 102 simulations for 30 perturbed parameters. Some later point of your paper make me questioning this design of experiment.

    We agree that 102 members is really restrictive. Enhancing the PPE size would be required in order to improve our approach, and we will make sure to highlight this point in the discussion.

11. L. 82: Did you use a classic Latin Hypercube Sampling ? Or did you use a maximin method with a centered design ? In Peatier et al (2022), I understand you used, a maximin method. However as 68 simulations crashed, the design of experiment may not sample some part of the parameters. That could explains why you got a crash with Candidate #1 (in uni-variate case). Please, comment this.

    We use exactly the same PPE as in Peatier et al. (2022). This is the same LHS with the maximin method, and we indeed had some crashes in the simulations. We agree that we are not sampling all of the parameter space, because of these crashes, and we expect to have some crashes also in our candidates (as some areas of the parameter space seem to lead to crashed simulations). This design could be improved by using sensitivity analysis to better understand which sub-space of the parameter led to the crashes, re-sampling the space avoiding these areas, and running another PPE better exploring the newly designed parameter space.

12. L. 83: Please define θ and define n earlier.

    Ok

13. L. 84: Can you clarify the ”climatological annual means” ? I understand that it is the mean over the period of 3 years and not the mean for each year. Because you need to reduce your calibration to 2 dimensions (parameters and spatial dimension) and thus, delete the temporal dimension.

    Ok - yes, this is the annual means average over the 3 years

14. L. 86: ”Elements.. area”: which elements ? Which area ? What are these ”weights” ?

    This sentence was out of place

New paragraph L.79 - 84 :

“ 2.1 Model and Perturbed Parameter Ensemble (PPE)

The model used in this study is ARPEGE-Climat, the atmospheric component of the CNRM-CM6 climate model, referred to as f, the climate model. The reference configuration of this model will be referred to as CNRM-CM6-1 and has been tuned by the model developers for the CMIP6 exercise (Roehrig et al., 2020). A PPE of this model f is created, containing 102 amip simulations (Eyring et al. (2016)) differing by their parameter values, representing the period 1979-1981 (3 years), with pre-specified Sea Surface Temperatures (SSTs) (Peatier et al. (2022)). Thirty model parameters (see Appendix A1) are perturbed with a Latin Hypercube Sampling (LHS) strategy, producing a variety of simulated climate states in the experiment : F = (f(θ1),...,f(θn)) based on a space-filling maximin design θ = (θ1,...,θn) (Peatier et al., 2022), with n = 102 and θi a vector of 30 parameter values. For the present study, we consider the annual means averaged over the whole 1979-1981 period. We write the model output f(θi) as a vector of length l, such that F has dimension l×n, where n is the number of ensemble members (n=102) and l the number of grid points (l=32 768)."

New reference : Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

1. EOF Analysis

1. L. 95: What is γ ?

   These are the different modes of variability (eigen vectors) of the EOF basis

2. L. 100: Can you clarify which mean μ is ? Ensemble and temporal mean over the 3 years period ? Is there a spatial mean ? I suppose no.

   It is the ensemble mean of the annual means averaged over the 3 years period (temporal means). \mu is a 2D matrix (lat x lon), without spatial mean.

3. L. 101: Introduce the "PC" abbreviation here.

   Ok

4. L. 105: If r_f includes μ, why does μ appears in the equation ? Is there a mistake ?

   Yes, there is a mistake. Written as it is, \mu is not part of the residual rf. We will remove \mu from the equation and write that rf and contains \mu

5. L. 107: Generally y is used for the result of a function (for example a model). For observation, I would suggest o.

   (Salter et al. (2019)) and (Rougier (2007) use z to refer to the observations. We will replace "y" with "z" in the paper to match their notations.

2. Model error partitioning

1. L. 116: Replace ”partitioning” to ”decomposition”

   Ok

2. L. 119: θ_i is not a calibration, it is a value of parameters. Delete the word ”calibration”.

   Ok

3. L. 122: Can you argue the use of MSE instead of RMSE or bias, or other error definition ? Furthermore, in the rest of the paper, it is mentionned sometimes MSE, some other times RMSE. Is it possible to be consistent and say RMSE if RMSE have been used in the two applications ?

   Ok

4. L. 123: What is c_r ?

   This is a typo, this should be c_y

4. The discrepancy term

1. L. 132: Correct the ∗ position: θ^* - OK
2. L. 132: I think the formulation is wrong. θ^* is not an optimal calibration. It is the optimal set of parameter θ. Am I wrong ? OK
3. L. 134: Delete ”how informative the climate model is about the true climate, and it measures” OK
4. L. 135: Replace ”real” by ”measured”. Even observations are not perfect. The reality is between models and observations. OK
5. L. 138: There is also the fact that in numerical models, the Earth system is discretized and not continuous. The fact that is is discretized is also a source of error. OK
6. L. 138: Generally, to address this issue (parameters not included in calibration process) and in order to include all parameters that must be calibrated, sensitivity analyses are performed (Saltelli et al, 2004). They can estimate the influence of parameters on model outputs and thus give a list a parameter that must be calibrated.

   Not entirely, because we would still be limited by the choice of the initial parameter space to sample. A sensitivity analysis such as the one described in (Saltelli et al., 2004) requires an emulator mapping from the parameter space to the climate model outputs, and such emulators have to be trained on a PPE. Given the huge number of parameters used in complex models such as GCM, it seems unlikely that we would be able to run a PPE perturbing all of the model parameters. Subjective choices, informed by expert advice, have to be made when selecting the parameters to perturb in the first place, and a sensitivity analysis will never be able to go beyond this choice. We would gain information about the effect of these selected parameters on the model output variance, but we would not be able to know whether perturbing a parameter not included in this initial sampling would have a greater effect.

7. L. 139: See comments for L. 132 ("calibration") OK
8. L. 139: Why do you present the uni-variate example and not the multi-variate one ? The uni-variate one is just a specific case of the multi-variate one.

   We do not present the uni-variate example, j could be any variable. In a multi-variate context, there would be as many discrepancy terms as variables considered - or we would have a general discrepancy term of 3 dimensions (lat, lon, nj), with nj the number of variables considered for the performance assessment. For more clarity, we will change the sentence L.139 "Considering a variable j, the discrepancy term nj is defined as : "

9. L. 147: This sentence repeats what have been already said in L.134

   OK

10. L. 154: Can you clarify the word "operational" ?

    OK

11. L. 155 - 156: reformulate the sentence. For example, "in this work, we propose several $\hat{\theta}$ that approximate θ^*".

    OK

12. L. 156 - 157: What is *m* ? I suggest : "to select *m* optimal model candidates"

    OK

<u>5. Emulator design and optimization</u>

1. L. 162: PC abbreviation not defined OK
2. L. 163: Does the linearity assumption is right ? Can you justify such choice of emulator. Did you compare to other emulators ?

   Yes - see answer to major comment and Appendix

3. L. 168: Can you clarify "comparably performing" ?

   Their emulated parametric errors are bellow the parametric error of the reference model.

4. L. 169: Can you clarify "objective function" ?

   We mean "trade-offs in the multi-variate spatial errors"

5. L. 171: Please specify the characteristic of this new LHS sampling. Is it a centered one ? with maximin space-filling method ?

   It is a maximin space-filling, this will specified

6. L. 171: instead of ”distribution”, use ”ensemble of ” OK
7. L. 172: What is the reference calibration ? I suppose it is the model using the default value of parameters that have been calibrated manually.

   Yes it is, we will detail this in the new Section **2.1 Model and Perturbed Parameter Esnemble (PPE)**

8. L. 172: This lays on the fact that you suppose the reference as calibrated and also to the fact that the error difference between the optimal simulation and the reference is smaller than the emulator error. Did you verify this second point ?

   Yes, the reference model version has been validated by experts and we assume its calibration to be near-optimal. We did not include the emulator error in this threshold, explaining also why we end up with actual candidate simulations showing higher error than the reference. This is a design choice aiming at preserving diversity in the sub-set of candidates, allowing them to show slightly higher integrated error than the reference.

9. L. 177: ”pattern error”. Do you mean ”spatial error” ? Yes

<u>6. Selection of diverse candidate calibrations</u>

This section has been completely rewritten to present the clustering analysis. We have better justified the choice of nk (or k) and presented the method for both the univariate and the multivariate applications.

1. L. 180: ”plausible optimal model configurations” We will call them “near-optimal”
2. L. 185: How do you define n_k? What is the difference with k defined in L.180 ? No difference, we will just call them k
3. L. 185: Why don’t you fix n_k = 5 for the two applications ? Why don’t you select randomly 5 candidates, 100 times, and select the group of 5 candidates amongst the 100 groups which have the highest variance ? See answer to Major comments
4. L. 185: Maybe clarify that ”diverse” means the selected candidates have a high variance See answer to Major comments
5. L. 186: What is n_j? The number of variables considered

<u>First application : surface temperature error</u>

1. L. 192: At equation (5), you consider a minimization of MSE, while here, you consider RMSE. Please, if you really use the RMSE, use the RMSE in equation (5)

We use the MSE, we will change L.192

<u>1. Assessing meaningful number of degrees of freedom</u>

1. L. 210: At line 192, you consider RMSE, while here MSE. Please clarify.

   This is MSE

2. L. 213 - 215: This is not coherent with lines 113-114. Even if $q = n = 102$, because of the observations, the non-parametric component (residual) is not null, so the reconstruction of the full error e is not perfect. Therefore, it is the best that you can get not the perfect one. (The same for L. 224 and L.252)

   We call “perfect reconstruction” the reconstruction that uses all of the 102 EOF modes, and Figure 1 illustrates the fact that, even retaining all of the EOF modes, the non-parametric error is indeed, not null. The “perfect reconstruction” does not reach the amplitude of the full error, because of the projection of observation.

3. Figure 1: I suggest this caption: ”Full model error e_{tas} and its parametric component $p_{tas}(\theta_i)$ for different truncation length : $q = 5$ (red dots), $q = 20$ (blue dots),$q = 50$ (pink dots), $q = 102$ (orange dots). a: Full error partitioning in parametric and non-parametric components in the PPE members $f(\theta_i)$ ranked from lowest to highest error. b: correlation between the full error e_{tas} and its parametric component $p_{tas}(\theta_i)$ within the PPE.”

   OK

4. Figure 1: I would recommend to add a) and b) in the upper part of the figure in order to differentiate the left andright panels and to refer to the ”Figure 1a” and ”Figure 1b”, in the text. And I would do the same suggestion for all figures.

   OK

5. Figure 1: The non-parametric error is symbolised with dashed line. Logically, the non-parametric error should bethe difference between the full error and the parametric error. So, the dashed line should be between the black dots and the colors dots, not between the full error and 0. Am I wrong ?

   Yes, this figure was confessing. We simplified it :

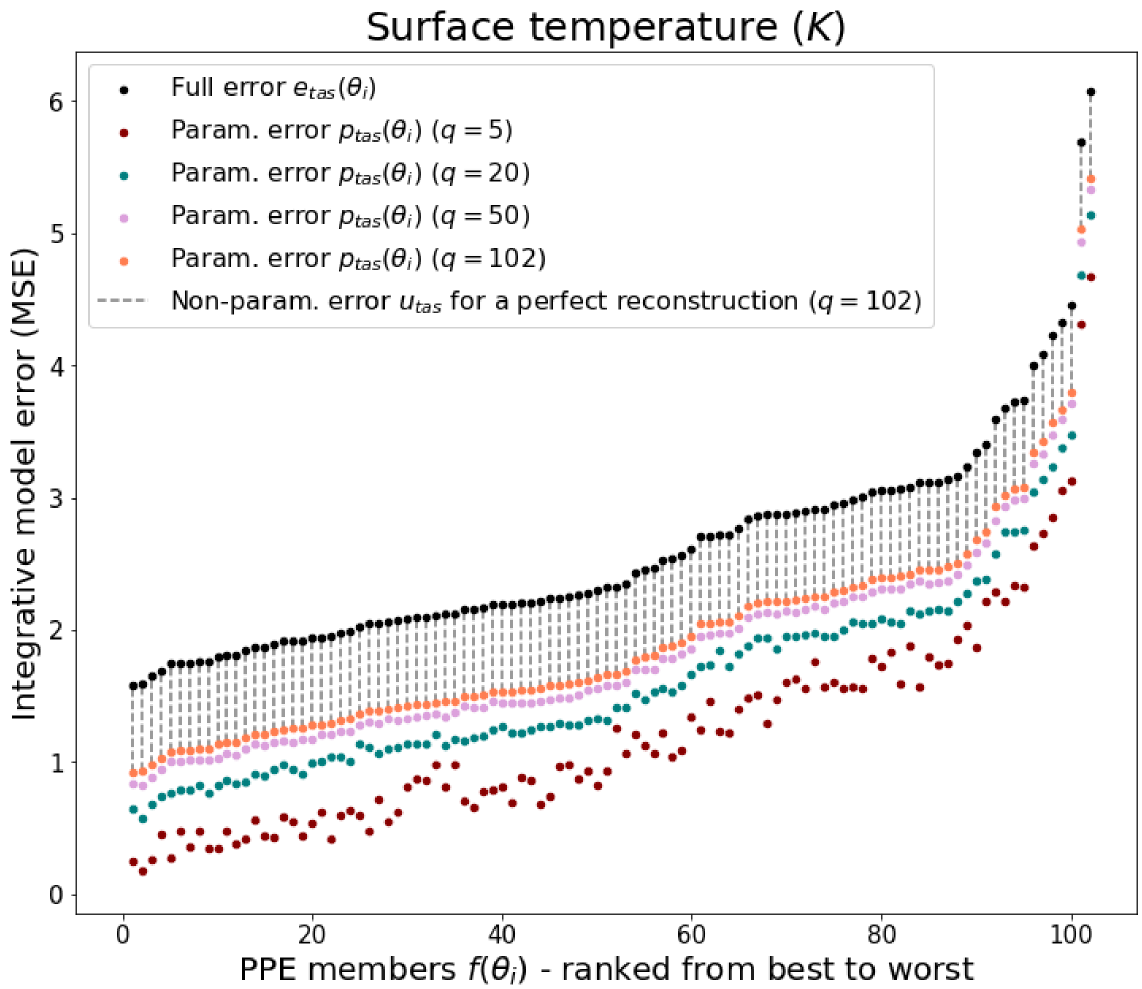

Fig 4. Full model error e_{tas} and its parametric component $p_{tas}(\theta_i)$ for different truncation length : q = 5 (red dots), q = 20 (blue dots), q = 50 (pink dots), q = 102 (orange dots). Full error partitioning in parametric and non-parametric components in the PPE members $f(\theta_i)$ ranked from lowest to highest error.

6. Figure 1: In the rest of the paper, you choose q = 18. Why don't you present the result for q = 18 in this figure ?

   Because the choice of q=18 is based on results presented in Figure 2, there is no reason to highlight this particular choice of q=18 in Figure 1.

7. L. 217: "A number of feature are notable in Figure 1." Not sure that this sentence is really pertinent.

   Agreed

8. L. 217 - 222: Why do you begin to analyse the right panel ? If you want to analyse it first, put the right panel on the left.

   OK

9. L. 219: Instead of using a minus (-) symbol use a colon (:)

   OK

10. L. 221: Instead of ”variation”, I would use ”spread” or ”variance”. (The same for L. 228)

    OK

11. L. 221: Maybe add a comment to say, that you will focus in the rest of the paper, on the 5 first modes.

    We do not focus on the 5 fist modes, we chose a truncation at q=18 for the analysis. This is only for graphical purposes that Figure 3 shows the 5 first modes (to avoid overloading the figure with 18 by 18 pairwise plots).

12. L. 225: The percentage is 26% in average over all PPE. However, it can be drastically different between the best and worst PPE. Can you give a range of this ratio ? Is it ok if the major part of the error is non-parametric for the best PPE, particularly if q = 5 ? Can you comment it ?

    The percentage of non-parametric error is not drastically different from one model to another. When considering the perfect reconstruction (q=102), as it is the case in the sentence L.225, the range is actually negligible. It can be seen on Figure 5 (Figure 2 in the paper), where the range of non-parametric error depending on truncation length is represented in teal. For q=102, there is almost no range, and the non-parametric error represents 26% of the full error (as stated L.225). However, we agree that for a truncation at q=5, the percentage of non-parametric error does vary a little bit from one model to another (see range FIgure 5), and Figure 4 (Figure 1 in the paper) suggests that the best models have the biggest percentage of non-parametric error. This might indicate that a truncation at q=5 will not be adequate in order to fully capture the parametric error, and that the best models tend to reduce errors in the first 5 modes of the EOF.

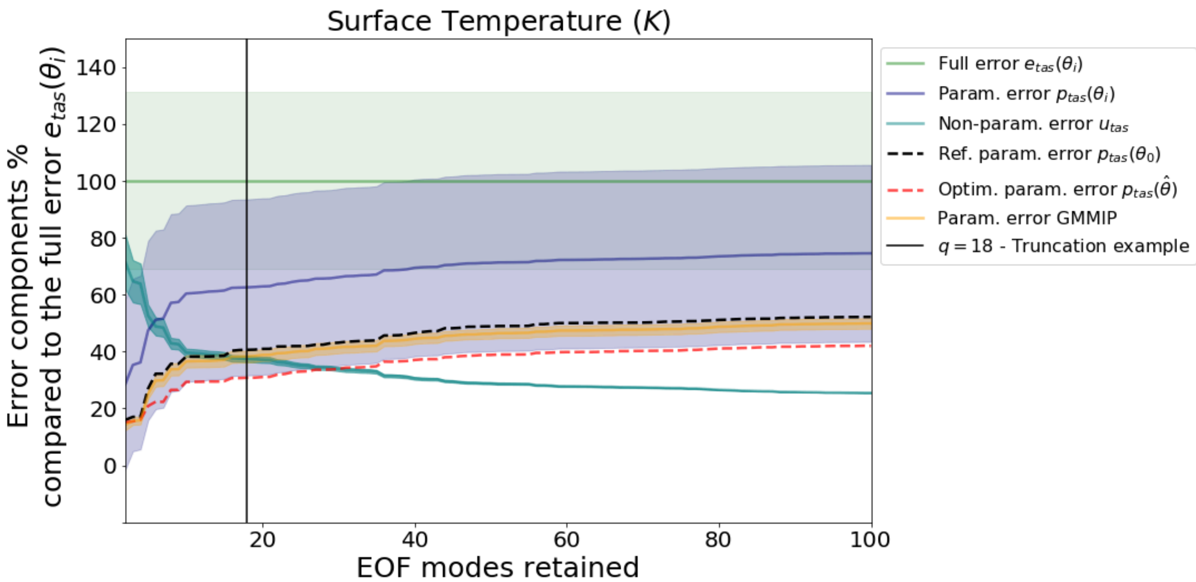

Fig 5. Truncation choice based on parametric emulation and error decomposition. The plot shows the ratio of the error components compared to the full error etas(θi) (in green) as a function of the

number of modes of variability retained. The lines are the ensemble means and the shadings represent the standard deviations. The plot shows the ratios of the PPE parametric error (dark blue), the PPE non-parametric error (light blue), the reference calibration parametric error ptas(θ0) (red dotted curve) and the GMMIP parametric error (orange). An example of truncation at q = 18 is represented on both plots by the black vertical line.

13. L. 227-229: Can you better explain your reasoning ?

    Agreed, new sentence L. 227-229 : "However, even for the perfect reconstruction of the model error (when q = 102), a non-null non-parametric component exists, and its ratio corresponds to 26% of the full model errors averaged over the PPE members. This ratio increases when retaining less EOF modes, and a large fraction of the model error pattern is not represented within the parametric component. For a truncation of q = 5, the non-parametric component of the error u_tas is 53% of the total e_tas(θi), in average over the PPE. Together, this implies that the variance in model error seen in the PPE can be explained by a small number of modes, but a significant fraction of this error is not represented within the parametric component of the error decomposition."

<u>2. Truncation and parametric emulation</u>

1. L. 233: Why don't you use the Leave-One-Out Cross validation method ? Maybe add a subsection in the Method section to present the validation method you used for validating the emulator.

   Agreed - see answer to major comments

2. L. 237: No, you cannot assess the predictive performance of the emulator properly by looking at the correlation between prediction and true data.

   See answer to major comments

3. L. 240: What is *x* ? The abscissa ?

   Yes, but this should be "q" in order to stay consistent with previous notation

4. L. 246: The chosen threshold is quite subjective. Maybe it is better to fix a variance at 95% and deduce a threshold in correlation, instead of fixing, arbitrary, a correlation at 0.5.

   See answer to major comments

5. L. 249: Is this the ratios for the emulated PPE ? Is there a link between the left and right panels ? If not, maybe consider 2 different figures for these two plots.

   No, it is the ratio within the actual PPE. The link between the two figures is that they are both used to detail and justify the choice of truncation length.

6. L. 249: Help the reader and add reference to the colour of the line : "PPE parametric (blue line) ..."

   Agreed

7. L. 251: Delete the word "example"

   Agreed

8. L. 253: Give the equation since line 251.

   Agreed

9. L. 257: Introduce GMMIP

   Agreed

10. L. 267: "This emulator is then optimized...". Not sure it is an optimization of an emulator, but more an optimization of the chosen set of calibrated parameter.

    No, here it is an actual example of optimization, where we used an optimizer to look for a minimum. This is not a method that is presented in the rest of the paper and this is not what is used for the selection of candidates (where we use a clustering method). This is just an illustration of the effects the addition of higher modes can have on an optimization.

11. L. 271 - 273: Please reformulate

    New sentence : "The difference between the PPE mean and this example of optimal calibration becomes constant when q = 7 or more, suggesting that there are no improvements of the optimization when adding modes higher than 7."

12. Figure 2: I suggest to do not present the method on the caption : "Truncation choice based on parametric emulationand error decomposition. a: Correlation between the emulated and true PCs coefficient of the surface temperature EOF, for the different modes of variability and for : the training set (blue curve), test set (orange curve). Mean is represented by dots and standard deviation by error bars. Averaged correlation over the modes cumulatively is shown by the red curve and the standard deviation by the red shading. b: Ratio as function of the number of modes of variability retained, of the error components compared to : the full error $e_{tas}(\theta_i)$ (in green), the PPE parametric error (dark blue), the PPE non-parametric error (light blue), the reference calibration parametric error (red dotted curve) and the GMMIP parametric error (orange). The lines are the ensemble means and the shadings represent the standard deviations. The black vertical line represents the truncation at q = 18."

    Agreed

<u>3. Trade-offs in models candidates</u>

1. L. 283: at line 171, you precise that you used 10^6 emulated simulations, while here, it is indicated 100.000. Is there an error ?

   Yes it is, we use a 100.000 LHS throughout

2. L. 284: "The five selected parameters calibrations were then used..." -> "The five calibrated set of parameters were then used..."

   Agreed

3. L. 286: The fact that there is a crash by using the real model, suggests a problem in the emulator building, maybe a lack of data in the training phase of the metamodel. Can the author comment this crash ?

   We had some crashes in the simulations of the PPE. Because of these crashed, we might not be sampling all of the parameter space, and we expect to have some crashes also in our candidates (as some areas of the parameter space seem to lead to crashed simulations). This design could be improved by using sensitivity analysis to better understand which sub-space of the parameter led to the crashes, re-sampling the space avoiding these areas, and running another PPE better exploring the newly designed parameter space.

4. L. 286: Concerning the other simulation using the 4 calibrated set of parameters, it can be interesting to quantify the emulation error (relative error between emulation and real results). That could be also, another way to verify the ability of the metamodel to emulate ARPEGE-Climat.

   See answer to major comments

5. L. 291: "provides some confidence in [..] the emulation skill". I am not convinced by this conclusion, particularly because the emulated and real simulation do not give similar results. For EOF3 according to EOF1 (1st columns, 3rd line in Figure 3), I agree. But for EOF2 according to EOF1 (1rst column, 2nd line), the emulated set #3 (orange dot) is closer to the real set #4 (blue triangle) than the real simulation #3 (orange triangle)... Maybe a calculation of distance between all pairs of emulated and real simulations can justify your conclusion better. Verify that, for example, the mean (average over all EOF couple) distance between the emulated simulation #1 and the real simulation #1 is smaller than with the other real simulations. Do that calculation for each emulated candidate simulation, and each EOF couple.

   See answer to major comments

6. L. 301: "Figure 3 also allows to see", please rewrite this formulation.

   Agreed "Figure 3 illustrates ... "

7. L. 301: if the "PC" abbreviation is used and defined, you can replace "Principal Components" by its abbreviation

Agreed

8. L. 304: Mode 2 is not constraint according to Figure 3. Why the author says "stronger than on the modes 3 to 5" and not "stronger than the other modes" ? Please, clarify.

   Agreed - "stronger than the other modes"

9. L. 309: "perform equally well on all modes". I understood that is it not true for EOF5 (in line 296) so not "on all modes". Am I wrong ?

   We wrote "it is impossible for the model candidates to perform equally well on all modes and fit observations perfecty", so there is no contradiction with what was saif L.296 statting that a lack of diversity in the candidates appears in mode 5

10. L. 309-315: Is it possible to calculate the distance between calibrated simulation (real or emulated ones) with observations, in order to justify your comment ?

    See answer to major comments

11. L. 309-315: Clarify that if you consider here the emulated or the real candidates.

    We refer to real candidates - agreed

12. L. 310: Help the reader and add a "in green" after "candidate 2".

    Agreed

13. L. 312-313: Not sure it is pertinent to comment on mode 5 as observations is outside the emulated ensemble.

    We think that it is still interesting to note

14. L. 317: Do the 5 candidates really have a comparable $e(\hat{\theta}_i)$, while candidate #4 performs better in terms of p ? You give the e value in line 336. Maybe, you can discuss this RMSE in section 3.3 instead of 3.4 (in that case, delete the RMSE appearing in Figure 4).

    We discuss the values of p and e later on in the paper

15. Section 3.3: Add a comment to the fact that emulated candidate #1 is generally far away from the other candidates.

    This is the candidate that crashes, so we might think that he is in a poorly sampled part of the parameter space. Figure 3 also suggests that he might present outputs very different from the other candidates. This is an interesting point ...

    However the candidates will whange with the new methodology.

16. Figure 3: I suggest this caption: "Correlation between the different standardised PC (obtained from the 102 memberPPE EOF) for the 100.000 emulated simulations (light gray), the optimal emulated simulations (dark grey, parametric

error lower than the reference), the 5 selected emulated candidates (colored dots), the 4 real ARPEGE-Climat simulations (colored triangles), the reference simulation (star) and the observation (cross).

Ok

17. Figure 3: Please, clarify between CNRM-CM, CNRM-CM6-1 and ARPEGE-Climat. If you use one model, pleaseuse the same name all along your paper (see comment L.172.).

Agreed

4. Example of temperature discrepancy term partitioning

1. L. 333: At line 279, you have already mention that you will considered, in the rest of the paper, q = 18. It is not necessary to mention it again.

Agreed

2. L. 334: Specify that you consider the emulated or real calibrated simulation. I understand you consider here the emulated candidates, but why not the real one ?

Figure 4 presents both the emulated and the real ones next to each others

3. L. 337: Please, add a comment on Candidates #5 which has a smaller p than the reference but a higher e.

Candidates will whange with the new methodology

4. L. 339: Are you comparing the 3th line plots at 1st and 3rd columns ? I see an overestimation of the negative bias or a underestimation of positive bias (not an overestimation of positive bias as written) in the emulated simulation (3rd column compared to 1st). There is also a too large negative bias in Antarctica in the emulated candidate #4. Can you comment it ?

Candidates will whange with the new methodology

5. L. 346: According to the colorbar, positive bias is in red, and in the Figure, mountain regions are in blue : there is not a positive bias in mountain region. The same for Africa, the bias is positive, not negative as written. Is there a mistake or did I misunderstood the analysis ? Please clarify.

There is indeed a mistake in the text about the sign of the bias, but candidates will whange with the new methodology

6. L. 348: "vary from a model to another" : the model does not changed (it is still ARPEGE-Climat or CNRM-CMIP6, whatever the name you want to give). Maybe you mean "vary from a parameter set to another" ?

"A model" refer here to $f(\theta_i)$, a particular configuration of ARPEGE-Climat

7. L. 348: I see the opposite: strong negative bias on North America and strong positive bias on central Africa.

   There is indeed a mistake in the text about the sign of the bias, but candidates will whange with the new methodology

8. Figure 4: I suggest to change the order of the column : emulated parametric error in column 1, parametric error in column 2, non parametric error in column3 and full error in column4. But it is just a suggestion.

   Agreed

Second application : multi-variate error

1. Variables, EOF analysis and truncations

1. Table 1: It is quiet confusing to use radiative data from the 2000-2002 period for a study on the period 1979-1981. Can you explain this choice ?

   The CERES observational data starts in 2000

2. Table 1: Instead of ”field”, use ”observable variables” as in the caption and instead of ”citation”, use ”data productreference” for example.

   Agreed

3. L. 366: ”spanning model components”: can you clarify ?

   We mean that the atmospheric component is not the only one that needs to be tuned when considering a climate model. The ocean, the surface and the sea-ice models all need to be tuned, which makes the tuning process and the uncertainty quantification associated, even more complex.

4. L. 370: here, the author used the MSE and not the RMSE contrarily to the uni-variate case. Is it possible to justify ?

   Indeed, we always use the MSE

5. L. 372: Is it really the annual mean or is it the mean over the 3 years period ?

   It is the mean over the 3 years period

6. L. 377 - 410: As already said for section 3.2, do not consider only the correlation to validate your emulator but also the relative error.

   See answer to major comments

7. L. 380 -385: same comments as in L.246. To stay robust, you should fix the percentage of variance you want, and deduce to that percentage, the number of modes. A correlation of 0.5 does not mean the same for each variable.

   See answer to major comments

8. L. 386: You do not show results for q=4, but for q=5 in Figure 6. So the variance model error is already very-well represented by the first 5 modes, instead of 4.

   Agreed

9. L. 387: I suggest to present Figure 6 before to analyse it.

   Ok

10. L. 388: ”66%” Is it enough ?

    Here we are not talking about the variance explained by the truncated EOF, which is 92% for sea level pressure, we are referring to the part of error included in the parametric component.

11. L. 390: Do you consider here, the emulated simulations or the 102 real ones ? I understand the real one, but precise it.

    Yes, the real ones. We do not have access to the non-parametric components of the emulations, as we only emulate within the EOF space. So we are always talking about actual simulations when refering to non-parametric components.

12. L. 390: As already said, do not consider correlation only for your validation. Consider also the relative errors, here to validate your EOF truncation.

    See answer to reviewers

13. L. 390: Why does the non-parametric error is averaged but not the parametric component ? Is it the averaged according to the 5 climatological field ?

    To validate the fact that, with enough EOF modes retained, we can reconstruct the full error from the parametric component, when using the mean non-parametric component as an approximation. It is averaged separately on the different variables.

14. L. 393: ”As expected, the PPE mean non-parametric components decrease as higher EOF modes are retained for the reconstruction but is never equal to 0 (even for a full reconstruction of q = 102). This is due to the fact that observations can never be fully captured by their projections into the model EOF basis (Figure 6)”

    Agreed

15. L. 402-404 : Be consistent with model name all along the paper.

    Ok

16. L. 402: as already said, define your reference calibration.

    Ok

17. L. 405. Maybe, there is a known bias in surface temperature in the ARPEGE-Climat model, (logically a bias that appears due to the use of the SURFEX surface parameterization). Can you discuss about it ?

Candidates will change with the new selection method, so this point about the candidate models outperforming the reference in terms of surface temperature might not be true anymore. But even if there was a known bias in the reference model, what is pointed out here goes beyond that, noting that we could reduce the surface temperature bias with another set of model parameters.

18. L. 405: ". This is a simple illustration of a complex tuning problem, and based on the results we obtained in the uni-variate application. It seems likely that comparably performing parameter configurations potentially exist for a multi-variate tuning problem, making different model trade-offs among both climatic fields and EOF modes representations of uni-variate errors (Figure 3)"

    Agreed

19. Figure 5: You do not comment the 2nd column of the figure. If you did not need it to support your discussion, thesegraphs can be delete of your paper.

    We think they are interesting. We will add a comment about them

20. Figure 6: What is the grey shading ?

    The PPE means non-parametric component, the part that has been added to all the points in order to reconstruct the full error.

21. Figure 6: Instead of "(y axis)" and "(x axis), use "coordinate" and "abscissa".

    Agreed

22. Figure 6: Logically, all your 102 CNRM PPE data must be used for training and testing your emulator. So, all dotsmust be in green or orange, no one in black. I don't understand your figure.

    The black dots are the actual runs, the 102 PPE dots. The green dots are the emulations of the train set and the orange dots are the emulations of the test set.

23. Figure 6: As you should sampling the training and testing datasets 10 times (according to L.236), the trained andtesting datasets are not fixed: some point should appear in green for a sampling, but maybe in orange for another sampling. I still not understand the orange and green dots.

    Here we do not say that we are sampling the training and testing datasets 10 times. That was the case for Figures 2 and 5, but in Figure 6 we are just showing a one time splitting of the dataset between train and test set, and a reconstruction of the full error.

24. Figure 6: Please, do not present the training dataset (80%) and testing dataset (20%) in the figure caption.
25. Figure 6: Why do you use a training dataset of 80% for the multi-variate case while 90% for the uni-variate case ?Did you also repeat the training 10 times ?

    This is a different analysis than in Figures 2 and 5, that brings a different message than a validation of the emulators. We are mostly interested in the reconstruction of the full error from the emulation of the parametric components depending on the truncation choice and the variables. We used another ratio of training/test and we did not repeat the analysis 10 times as it was the case for Figures 2 and 5.

2. Candidates selection in a multi-variate context

1. L. 412 - 426: This paragraph must be in the Method section.

   Agreed

2. L. 413 : For the subset candidate selection, you maximise the variance of the multi-variate metric. The minimisation of the multi-variate metric is not the process to select a set of optimal candidate.

   The candidate selection changed completely - see answer to major comments

3. L. 425 : Why $n_k = 4$ and not $n_k = 5$ as in the uni-variate application ?

   The candidate selection changed completely - see answer to major comments

4. L. 429 - 434: Help the reader by adding in parenthesis to which feature in the Figure it corresponds : "Among the 4 selected candidates (blue dots)", "than the reference model (yellow dashed line", "PPE mean (red dot)", "mean of the 40 CMIP6 models (green dot)"

   Agreed

5. L. 433: "CNRM model grid before"

   Agreed

6. L. 433: Precise here (not in the figure caption) that observation have been interpolated also.

   Agreed

7. L. 435: Is there a justification to apply the uncertainty (standard deviation) of CMIP6 model to CNRM reference model and not the uncertainty of CNRM model itself (CNRM-PPE) ?

   The PPE variance can not be used as an estimate of the tolerance we should have when considering the model performance, as the PPE members have not been tuned and many of them will be showing really bad performances. On the other hand, the CMIP6 models have all been tuned and validated by experts and they are all considered to show a plausible representation of the historical climate. We can use the variance of the error within the CMIP6 ensemble as a tolerance for the evaluation of our candidates.

8. L. 436: ",indicating ..." -> "This indicates"

OK

9. L. 440: Maybe present all the data used for this work in the Method section instead of presenting them in the results section.

   Thanks for the suggestion, but we think that the method section is already quite dense, and this dataset is only used once in the paper. So it makes sense to introduce it in the results section.

10. Figure 7: I suggest this caption: "Multi-variate error e_{tot} for the CMIP6 models, the CNRM-CMIP6 PPE members, the 4 optimal CNRM-CMIP6 candidates and the 10 members of CNRM reference model with different initial conditions. Each small dots correspond to a model, the bigger dots correspond to the ensemble means and the dashes are the standard deviations. The orange dashed line at 1.0 represents the CNRM reference model error. The green area indicates the interval of plus or minus one standard deviation of the CMIP6 errors, centered around the CNRM reference model error. "

    Agreed

11. Figure 7: "available CMIP6 model" : Does this mean that all models are not used ?

    We used all models that ran the amip-hist experiment and that made their outputs available on the ESGF platform.

<u>3. Diversity of error patterns among candidates</u>

1. L. 451: delete "for the selection"

   Agreed

2. L. 451 - 475: I suggest to present the results in the order of candidate number : discuss firstly candidate 1, then candidate 2, ...
3. L. 433: to support your analysis, cite the value of RMSE.

   Agreed

4. L. 451 - 475: Is it possible to attribute theses differences to particular parameters value ?

   We could run a sensitivity analysis to identify the effect of the different parameter perturbation on model performances and error patterns. However, this is beyond the scope of the present study.

5. L. 471: "everywhere" -> "on the whole domain"

   Agreed

6. L. 472: Delete "(Figure8)"

Agreed

7. L. 472: "not the worst of the selection" not coherent with "is the worst performing" at L.471

   The candidates will change with the new selection method, as will the associated comments. But it is the worst performing when looking at multi-variate score, but it is not the worst of the selection when looking at the radiative fluxes only.

8. L. 475: Sentence not finished...

   Agreed, thanks

9. Figure 8: Candidate 1 has no one green dot : all RMSE are higher than the reference. Please, add a comment on it.

   The candidates will change with the new selection method, as will the associated comments.

10. Figure 8: Why is it always $\hat{\theta}_1$ in the grey rectangle ? Why the 1 ?

    This was a mistake. It should be \theta_1, \theta_2, \theta_3, \theta_4

11. Figure 5 - 6 - 8 - 9 - 10: Keep variables appearing in the same order for all figures

    This is the case

<u>4. Examples of discrepancy term partitioning</u>

1. Figures 9 and 10: add p and u in the grey rectangle, as you added e in the Figure 8

   Agreed

2. L. 483: delete some "are"

   Agreed

3. L. 484: I am not sure that it validates the method properly. But, at least, it shows that the objectives are achieved.

   Agreed

4. L. 488 - 505: The analysis is only conducted for SW. Do you have any comments on the other variables ?

   The candidates will change with the new selection method, as will the associated comments.

5. L. 499 - 502: split this sentences in two different ones.

   The candidates will change with the new selection method, as will the associated comments.

6. L. 506: please, explain better the link with the effective degrees of freedom

   Agreed

Conclusions

1. L. 511: ”perturbed physics” -> ”perturbed parameters”

   Agreed

2. L. 512: ”diverse” : reformulate

   This is the term that has been used throughout the paper, we think it is more consistent to use it also in the conclusions

3. L. 513: ”a number of” -> ”different”

   Agreed

4. L. 513: delete ”which we illustrate ... (General Circulation Model)”

   Agreed

5. L. 523: delete ”examples”

   Agreed

6. L. 525: ”CNRM-CM”, please use the same name for the same model

   Agreed

7. L. 525: Really 10^6 simulations ? Or is it 100.000 as written in the paper ?

   Yes, 100.000

8. L. 525: ”of the perturbed parameter of the parametric components of the model errors”, please reformulate

   Agreed : “The optimization is based on multi linear predictions of the parametric components of the model errors, from a 10^5 LH sampling of the perturbed parameters”

9. L. 527: Use a more appropriated vocabulary than ”diverse”

   This is the term that has been used throughout the paper, we think it is more consistent to use it also in the conclusions

10. L. 529: ”CNRM discrepancy” -> ”CNRM model discrepancy”

    Agreed

11. L. 547: Add a precision about the fact that candidate performs better than the reference in terms of e , while it has been optimised according to p

    Agreed

12. Add some perspectives. For example, it could be interesting to analyse the parameters value between the differentcandidates, in order to explain which parameters lead to such biases.

    Agreed

Technical corrections

1. L. 21, 27, 33, 59-60, 91, 152, 188-189, 560: put the citation in chronological order

OK

2. L. 23: correct the parenthesis "(CMIP; Eyring et al, 2016)"

OK

3. L. 40: correct the parenthesis "(such as global mean quantities, Bellprat et al., 2012 ; Williamson et al., 2015)

OK

4. L. 43: correct the parenthesis "(sometimes referred as an "iteratire refocusing" approach; Williamson et al, 2017) ... 5. L. 59: PCA not defined

OK

6. L. 52, 64, 69, 131, 148, 149, 163, 196, 206, 243, in Table 1: correct the parenthesis in the citation (parenthesis aroundthe date and not the name)

OK

7. L. 159: Error in the section referring: "Section 2.4" and "2.5" instead of "Section 2.3" and "2.4"

OK

8. L. 207: Error in the section referring: "Section 3.1" and "3.2" instead of "Section 3.2" and "3.3"

OK

9. L. 211: error in the section referring number

OK

10. L. 232: I think you mean "equation 14", not "Section 14".

OK

11. L. 242: ”high-order modes” repeated twice.

OK

12. L. 245: ”point”

OK

13. L. 268: ”for” repeated twice

OK

14. L. 383 : ”e.g.”, belonging to which sentence ?

OK

15. L. 420 : ”an the” -> ”an” or ”the” but not the two ones.

OK

16. L. 451: ”The Figure” -> ”Figure”

OK

17. L. 483: ”on the other ends” -> ”on the other hands”

OK

18. L. 484: remove the space before ”:”

OK

19. L. 490: (Figure 10) -> (Figure 8)

OK

20. L. 525: LH abbreviation not defined

OK

21. L. 533-536: Did you want to do a list as in lines 514-521 ?

OK

22. L. 552: opimization -> optimization

OK

23. Figure 7: ”dasehd” -> ”dashed”

OK

24. Figure 7: ”arounr” -> ”around”

OK