Response to Reviewer #1

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

Major comments

Method deficiencies (i) the emulator: The choice and use of emulator are not discussed, but in several places in the results it is indicated that there is room for improvement here. E.g. in Figure 3 - often the emulated values (that determine also the concept of "diversity" above) did not lead to close or hugely diverse simulations. I came to this conclusion as (1) the dots and triangles are often far from each other, and (2) the dots are typically more spread than the triangles, (3) one dot lead to a simulation crash for its corresponding triangle. Likewise in Figure 7: 4 diverse candidates were selected with emulated p < 1, yet all in practice most were > 1. Though this was noted in the document it was not attributed to the emulation stage being unable to well-represent the underlying model. Particularly in the results, the shortcomings due to the choice of emulator should be attributed clearly, and additional discussion added to the conclusions in terms of how to address this in future investigation.

As a better discussion about the choice and performance of the emulators was also requested by the Reviewer #2, we have decided to add an appendix showing the performance of our statistical model to emulate the parametric components of the individual and total model errors. We also compare their performances with other statistical models : a Random Forest and a LASSO regression.

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. However, results suggest that there is room for improvement, especially in the prediction of the LW errors, 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.

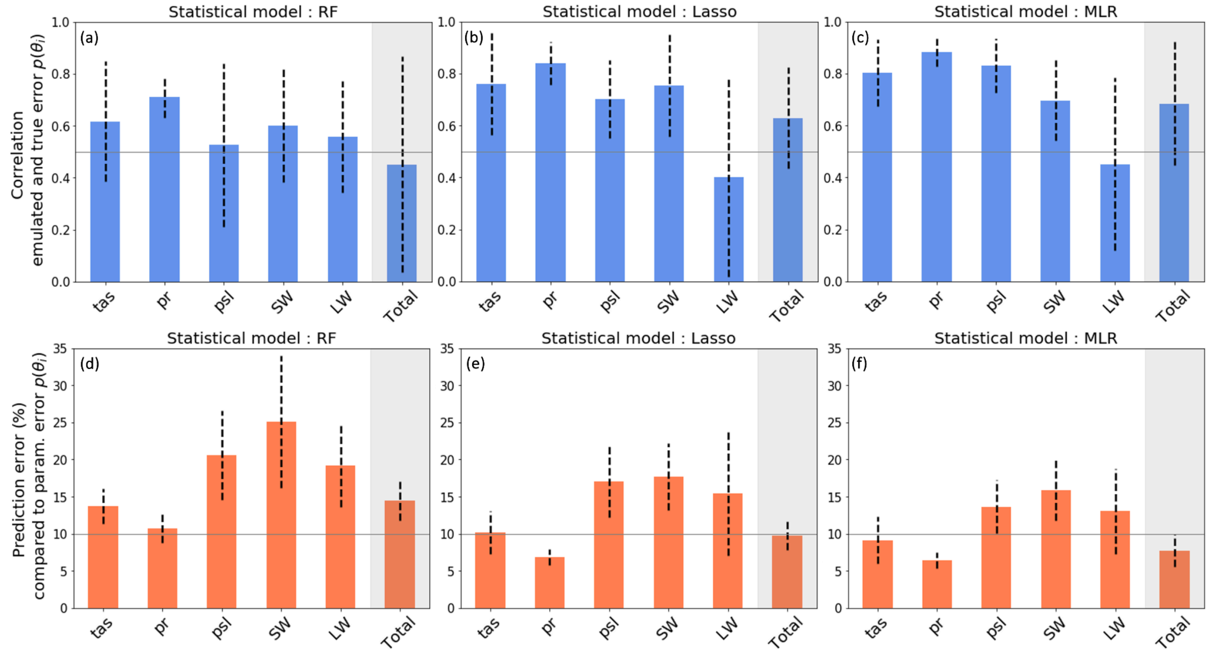

Fig 1. Correlations and RMSE (in % compared to the true values) between emulated and true parametric components of the errors within a test set representing 10% of the dataset. The evaluation is repeated 10 times with random sampling of training and test sets and the mean and standard deviation among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

Indeed, Figure 3 of the paper was highlighting a loss of diversity in the actual runs compared to the statistical predictions, but this might be linked to our previous method of candidate selection. The clustering analysis seems to show a bit less diversity in the emulated candidates, but a gain in diversity in the climate runs (Figure 2). There is still a prediction error (the dots and triangles can be far from each other in Figure 2 and we expect around 8% of prediction error, as presented in Figure 1), but the use of clustering analysis and the sampling of 10 instead of 5 candidates allow a better exploration of the optimal sub-set of emulations.

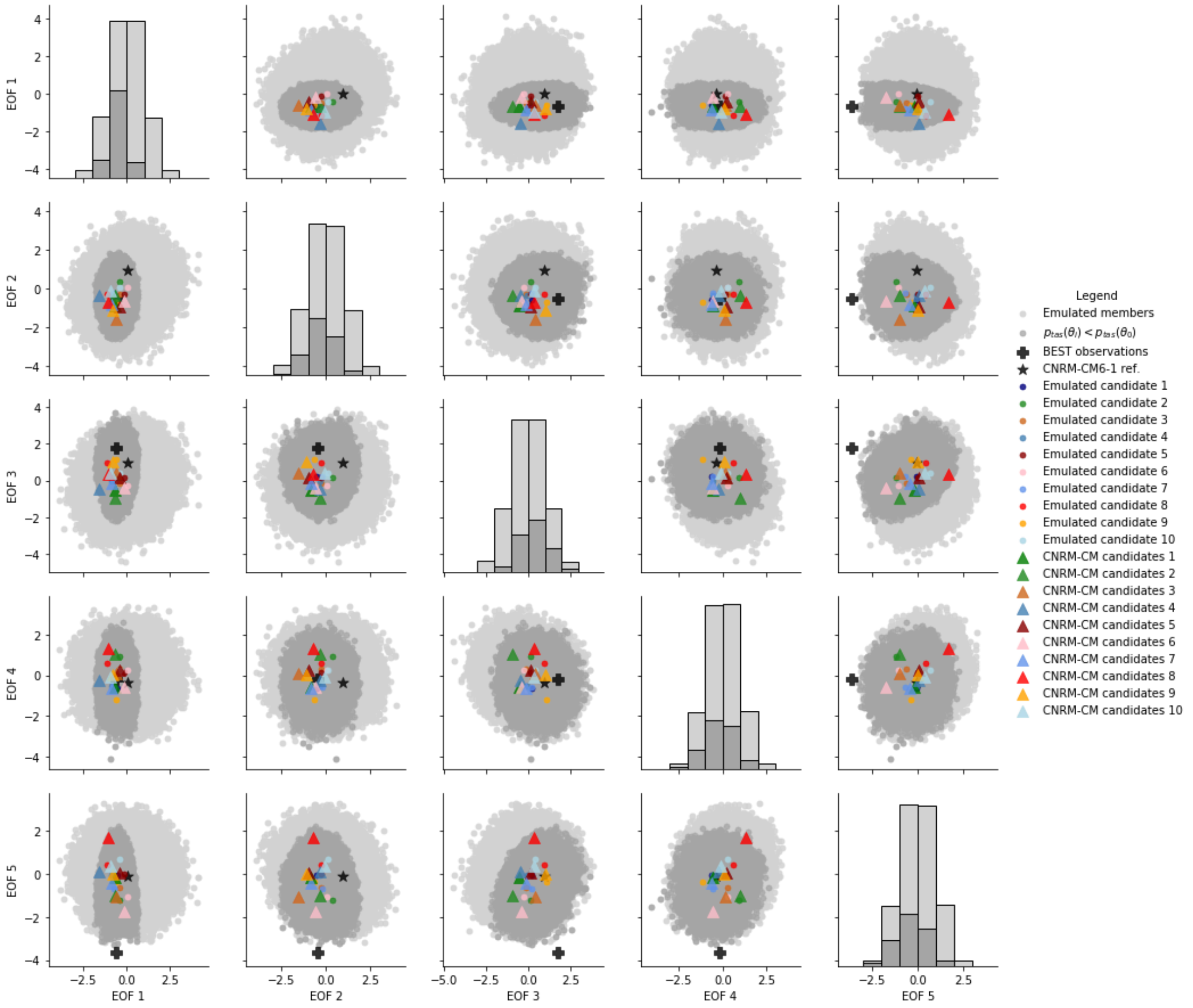

Fig 2. Standardised principal components associated with the 5 first modes of the surface temperature EOF basis computed on the 102 members of the PPE. The Figure presents the projections of the 100 000 emulated members (light gray), the 'optimal' emulated members (with a parametric error lower than the reference CNRM-CM6-1, in dark grey), the 10 emulated candidates (colored dots) and the 10 candidates CNRM-CM (colored triangles). This Figure will replace Figure 3 in the main paper.

Finally, we agree that the fact that the candidates were selected with emulated parametric error p<1 but shows, in practice, higher parametric error should be attributed to the limitations of the emulators and we will add a sentence clearly stating that in the main paper. Table 1 is showing the values of the parametric errors for the new candidates selected with the clustering analysis : only 2 of them have lower aggregated error than the reference.

Candidate 1	1.68	Candidate 6	1.03
Candidate 2	1.0	Candidate 7	1.09
Candidate 3	0.88	Candidate 8	1.7
Candidate 4	1.05	Candidate 9	1.22
Candidate 5	0.91		

Table 2. Parametric component p of the error in RMSE for the 9 candidates of the multivariate application. A candidate has a lower error than the reference when the parametric error is lower than 1. Note that 10 candidates were sampled originally, but one of them crashed.

Method deficiencies (ii) exploring diversity: Perhaps the weakest part of the investigation involved the exploration of "diversity of solutions". Though I agree that the approach removes solutions that have similar parametric components, it is mentioned L284 that the authors "maximize" the diversity in solutions, or L177 "a selection ... as diverse as possible", yet their approach is not a seemingly well defined optimization procedure. In fact the approach the authors use appears to be non-standard for such exploration, likewise with the termination of "5" final candidates is completely unmotivated. Finally the sensitivity to the choice of the threshold was also not explored. Typically for sourcing representative simulations a practitioner might use a clustering algorithms (k-means/medians/mediods etc.) over the data, and draw representative samples from the cluster centers to illustrate diversity. Such approaches are scalable and robust, and one can quantitatively assess sensitivity to e.g. the number of clusters the data appears, thresholding etc. The random removal methodology presented here appears untested. For example, if the authors ran the method over 100 trials, would they expect to find the same nodes each time, or if not, would they expect the selections of 5 display similar diversity, how would they quantitatively measure this? How would they compare the performance of terminating at different numbers of solutions. Some references and/or computational validation experiments for the approach taken are needed, otherwise, again some clear admission in the methods and final discussion that this is a proof of concept, and suggest other approaches that readers can try in future to explore diversity.

From a naive inspection of the samples in Figure 3, it does not seem out of the question that there are more representative samples one could find. I also think the use of emphatic words such as "large" diversity (L452 and elsewhere), "the key" trade-offs (L507 and elsewhere) should really be removed when one is not able to really quantify how representative these differences are, or how many trade-offs are observed due to the selection procedure.

We agree that the selection of diverse candidates was done through a non-standard approach and that the sensitivity of the results to the number of candidates was not tested in the submitted version of the paper. 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 variables, normalized by the principal components of the reference model. We 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, as stated by the reviewer, the clustering analyses are sensitive to the choice of cluster number *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*.

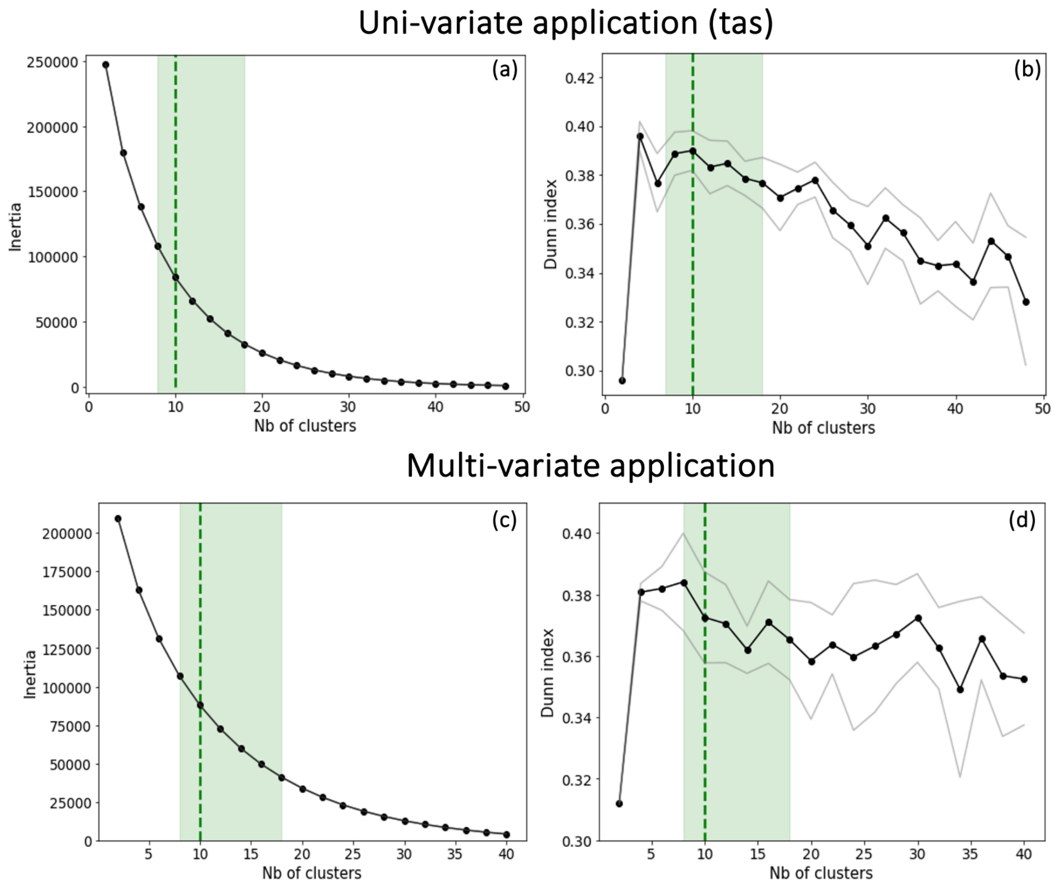

Fig 3. Sensitivity test of the clustering analyses for the uni-variate (first row) and multi-variate (second row) applications. The inertia criteria ((a), (b)) and the Dunn index ((b), (d)) are shown depending on the number of clusters (x axes). The green shaded areas present the acceptable number of 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).

Optimality and near-optimality

As a reader, this is the walkthrough of optimality in this manuscript :

- An optimal solution is defined to be an exact quantity f(theta*) eqn(9).
- This is then redefined in L158 where optimal is an input that minimized e_j.
- Then in L168 the optimal solution is considered to be the input that minimizes an emulated objective (maybe this should be called e_{em,j}).
- In L172 The introduction of a black-box “reference calibration” CNRM-CM (6?) is introduced as a threshold for optimal candidates (this is the first time this is mentioned in the document and it is unreferenced), it is used to rule out all candidates with greater error. It is not clear if this is calculated with real or emulated coefficients. Is the intention to have a “typical” level of accuracy, arising from current “tuning procedures”?
- In L520 there is mention of near-optimal configurations
- In L528 optimality and an optimal space are then unified with near optimality?

I request that the authors solidify these notions and be consistent throughout. I also request that the authors use concrete notation for where they are using emulated vs actual scores. Finally, the threshold should be presented in a more clear fashion and explanation as to why this choice should define optimality.

Agreed - we will better define the notion of “optimal” in the paper by 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.”

We will also note the emulated scores $p_{em,j}(\theta_i)$ and the actual scores $p_j(\theta_i)$ in order to differentiate whether we discuss statistical predictions or climate model outputs.

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.

<u>Why explore the diversity of solutions?:</u> Given that the investigation is centered around exploring the diversity of these solutions of low parametric error, It was mentioned only in the final sentence as to why one wishes to do this beyond scientific curiosity. I would argue the authors should more strongly present their case throughout the text, particularly how this specific approach of selecting a few diverse candidate simulations can fit into the practitioners workflow, or how it relates to uncertainty quantification.

Agreed - we will add a paragraph in the introduction better motivating this exploration of diverse solutions.

Minor comments

Please change the title. Having read the document I still do not know what “Spatial Calibration” is. It is in fact this term is not once mentioned in this entire manuscript. One suggestion: “Exploration of diverse solutions from imperfect climate model calibration”?

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”.

My own background implies that “model error” = “structural error” = “non-parametric error” = “model discrepancy” here some of these are treated differently, please be clear in defining all of these terms and ensure consistency through the text.

Agreed - we will make sure all the terms are defined and consistent.

L19 The literature review missed the growing works of the CliMA group: Here, for example, Idealized GCM’s have been calibrated with Bayesian Formalism, using tools from data assimilation and accelerated samplers (https://doi.org/10.1029/2020MS002454, https://doi.org/10.1029/2021MS002735), the calibration approach has shown scalable to higher dimensional parameter spaces in different settings.

Agreed - the above references will be added in the literature review

L25 Stating that hand-tuning “has proven remarkably robust”, is easily misunderstood by the reader, the authors add a long caveat. Instead why not state something clear, such as “Such approaches remain popular in operational settings”

Corrected

L40 For history matching / NROY approaches also cite the studies of (https://doi.org/10.1029/2020MS002217, https://doi.org/10.1029/2020MS002225)?

Agreed - the above references will be added in the literature review

L72 This investigation explicitly explores the role of parametric error (whereas structural error often refers to an error incurred due to mis-specification of model structure, i.e. non-parametric error),

We replaced ‘structural error’ by ‘model error’

L84 Are there findings from (Peatier et al 2022) about the validity of this PPE also be provided - it seems that results are critically linked to the exploration of this ensemble.

My initial concern is that 100 members across 30 dimensions leads to poor exploration without very tight bounds and well chosen points. Perhaps the size of the non-parametric error obtained in this investigation can also shed some light here?

We agree that the small size of the ensemble is a limiting factor in this analysis and that a larger PPE will improve parameter space observation and might affect the amplitude of the non-parametric component … We will add a few words in the conclusion stating this caveat.

L86 what is an element, why are they weighted, what is their corresponding area..? Was this sentence misplaced?

Indeed, this was misplaced

L151 - in other applications it is also seen as a Gaussian Process (following https://doi.org/10.1111/1467-9868.00294)

Agreed

L360 “quantitatively”, should be “qualitatively”. No rigorous scoring was assigned to the amount of trade-offs, and although I am sure the “multiple optima” behavior is true, the authors did actually evaluate any optimal solutions in this investigation, rather solutions achieving below a user-chosen parametric error threshold.

Agreed

L519, L553 state that by-hand calibrations are “tractable”. This is not true, it may be the case that hand-tuned models can still be performant, but this is not the same as tractability - which additionally implies the procedure of hand-tuning to be easy, and straightforward practice, while the reality is that it is more a “dark art” of the climate modeling community.

We will remove the word ‘tractable’ and change the sentence to : “Reference model version shows one of the lowest integrated performance metric and historical common practices for parameter tuning is more robust than often assumed”

Typos etc.

L211 dead link - Ok
L241-2 repeated phrase “high-order modes” - Ok
L245 - “point” - Ok
single quotes backwards e.g. L7, L522., - Ok
Typically latex for “theta star” has a superscript \theta^*. - Ok

Final sentence of L16, was this meant to be here? Seems out of place - We have changed this whole paragraph