Review of "On the spatial calibration of imperfect climate models" by Peatier et al. (2023)

1 Summary

In this study, the authors propose an approach of calibration that take into account the spatial parameters uncertainty. Indeed, Earth System Models deals with a lot of parameters those values are uncertain and required to be tuned. Generally, a manual approach is done to calibrate a model by minimising a metric. This lead to unique parameters values. However, different sets of parameter values can lead to comparably model skills. Here, the originality of this work is that the objective is to find not one but n_k sets of parameter values that lead to similar model skills. But, instead of selecting randomly n_k sets of parameter values that optimised a metric, they propose these n_k sets must be as various as possible in order to be representative to the whole parametric uncertainty. To that purpose, the authors propose the following method:

1. generate a training data set to built an emulator: for each members (102 members in this study) of a Perturbed Parameters Ensemble, compute the Empirical Orthogonal Function (EOF) coefficients in order to reduce the dimensionality of the problem
2. built a multi-linear emulator, that predicts EOF coefficients directly from parameter values.
3. with the emulator, generate 100.000 emulated EOF coefficients from a 100.000 sampling of parameters values using a Latin Hypercube Sampling
4. from EOF coefficients, compute the simulation error (the metric to optimise) and decompose it as parametric and non-parametric ones.
5. select as optimised candidates, all emulated simulation that have a smaller parametric error than a reference
6. reduce the number of the optimised candidates by selecting only n_k ones: maximise the distance, in the EOF basis, between the n_k candidates in order to get n_k candidates as various as possible in terms of spatial error patterns

This method is applied on the CNRM-CMIP6 model, with a parametric error due to 30 parameters and on 2 cases: a uni-variate case on surface temperature and a multi-variate one. The method is evaluated by comparing the emulated EOF results and real EOF results coming from real simulations using the parameter values obtained from the optimisation process and by comparing spatial pattern diversity obtained from the optimisation process. A deep analysis of the spatial patterns of the different errors (parametric, non-parametric, full) between the n_k optimised candidates is realised.

2 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.
- 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 uni-variate and multi-variate case: the training data is 90% of the data in the uni-variate 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).
4. To define the truncation of the EOF modes, the author used a threshold on the correlation. I think it is not robust. I suggest to define a threshold on the variance.
5. The authors do not explain how they get the number of optimized candidates n_k. It seems to be completely subjective.
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.
7. The reference simulation is not presented.

- 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"
- 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.
- 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.

3 Minor comments

3.1 Abstract

1. L. 3: "it impossible to find **one** model version"
2. L. 5: Delete the PPE abbreviation (not necessary in the abstract)
3. L. 9: Results on initial conditions perturbation are clearly not a main result of this study. I suggest to do not put it in the abstract, but instead, to present the two applications (uni-variate, multi-variate cases)

3.2 Introduction

4. L. 21 - 24: Sentence too long. Please, rewrite it.
5. L. 25: "Robust". I am not sure this a right way to say it.
6. L. 24 - 27: Sentence too long. Please, rewrite it.
7. L. 28: As the ensemble model approach is inserted here, maybe it is better to insert here also the notion of PPE model.
8. L. 35: Put the list of emulators in another sentence.
9. L. 40: Can you clarify the "low dimentionnal output space" ?
10. L. 44: Maybe add a comment on the fact that it depends on subjective advises from physical experts.

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

12. L. 48 - 52: Sentence too long. Please, rewrite it.

13. L. 48: I do not understand the "grid scale". I suppose you want to say "represent spatial performance"

14. L. 48: Clarify that it is for the development of the model.

15. L. 50: Use "grid points" instead of "pixels"

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

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

18. L. 46 - 56: Please, clarify here that you speak about spatial calibration.

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

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

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

22. L. 71: "it can inform model development". Please reformulate

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

3.3 Methods

24. L. 79 - 86: Put that paragraph in a subsection named "Used model", for example.

25. 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".

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

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

28. L. 80: Definition of PPE abbreviation not necessary

29. L. 81: Can you define AMIP abbreviation ?

30. L. 81: I would say "102 simulations differing by their parameters value".

31. 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.

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

33. 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.

34. 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.

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

36. 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.

37. L. 86: "Elements.. area": which elements ? Which area ? What are these "weights" ?

3.3.1 EOF Analysis

38. L. 95: What is γ ?

39. 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.

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

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

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

3.3.2 Model error partitioning

43. L. 116: Replace "partitioning" to "decomposition"

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

45. 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 ?

46. L. 123: What is c_r ?

3.3.3 The discrepancy term

47. L. 132: Correct the $*$ position: θ^*

48. 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 ?

49. L. 134: Delete "how informative the climate model is about the true climate, and it measures"

50. L. 135: Replace "real" by "measured". Even observations are not perfect. The reality is between models and observations.

51. 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.

52. 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.

53. L. 139: See comments for L. 132 ("calibration")

54. 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.

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

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

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

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

3.3.4 Emulator design and optimization

59. L. 162: PC abbreviation not defined

60. L. 163: Does the linearity assumption is right ? Can you justify such choice of emulator. Did you compare to other emulators ?

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

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

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

64. L. 171: instead of "distribution", use "ensemble of "

65. L. 172: What is the reference calibration ? I suppose it is the model using the default value of parameters that have been calibrated manually.

66. 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 ?

67. L. 177: "pattern error". Do you mean "spatial error" ?

3.3.5 Selection of diverse candidate calibrations

68. L. 180: "plausible **optimal** model configurations"

69. L. 185: How do you define n_k ? What is the difference with k defined in L.180 ?

70. 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 ?

71. L. 185: Maybe clarify that "diverse" means the selected candidates have a high variance

72. L. 186: What is n_j ?

3.4 First application : surface temperature error

73. 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)

3.4.1 Assessing meaningful number of degrees of freedom

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

75. 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)

76. 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."

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

78. Figure 1: The non-parametric error is symbolised with dashed line. Logically, the non-parametric error should be the 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 ?

79. 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 ?

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

81. 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.

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

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

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

85. 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 ?

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

3.4.2 Truncation and parametric emulation

87. 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.

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

89. L. 240: What is x ? The abscissa ?

90. 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.

91. 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.

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

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

94. L. 253: Give the equation $\frac{p(\theta_i)}{e(\theta_i)}$ since line 251.

95. L. 257: Introduce GMMIP

96. 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.

97. L. 271 - 273: Please reformulate

98. Figure 2: I suggest to do not present the method on the caption : "Truncation choice based on parametric emulation and 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."

3.4.3 Trade-offs in models candidates

99. 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 ?

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

101. 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 ?

102. 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.

103. 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.

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

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

106. 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.

107. 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 ?

108. 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 ?

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

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

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

112. 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).

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

114. Figure 3: I suggest this caption: "Correlation between the different standardised PC (obtained from the 102 member PPE 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).

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

3.4.4 Example of temperature discrepancy term partitioning

116. 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.

117. 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 ?

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

119. 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 ?

120. 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.

121. 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" ?

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

123. Figure 4: I suggest to change the order of the column : emulated parametric error in column 1, parametric error in column2, non parametric error in column3 and full error in column4. But it is just a suggestion.

3.5 Second application : multi-variate error

3.5.1 Variables, EOF analysis and truncations

124. 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 ?

125. Table 1: Instead of "field", use "observable variables" as in the caption and instead of "citation", use "data product reference" for example.

126. L. 366: "spanning model components": can you clarify ?

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

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

129. 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.

130. 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.

131. 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.

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

133. L. 388: "66%" Is it enough ?

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

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

136. 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 ?

137. 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)"

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

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

140. 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 ?

141. L. 405: ". This is a simple illustration of a complex tuning problem, and based on the results we obtained in the uni-variate application**. I**t 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)"

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

143. Figure 6: What is the grey shading ?

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

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

146. Figure 6: As you should sampling the training and testing datasets 10 times (according to L.236), the trained and testing 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.

147. Figure 6: Please, do not present the training dataset (80%) and testing dataset (20%) in the figure caption.

148. 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 ?

3.5.2 Candidates selection in a multi-variate context

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

150. 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.

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

152. 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)**"

153. L. 433: "CNRM **model** grid before"

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

155. 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) ?

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

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

158. 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. "

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

3.5.3 Diversity of error patterns among candidates

160. L. 451: delete "for the selection"
161. L. 451 - 475: I suggest to present the results in the order of candidate number : discuss firstly candidate 1, then candidate 2, ...
162. L. 433: to support your analysis, cite the value of RMSE.
163. L. 451 - 475: Is it possible to attribute theses differences to particular parameters value ?
164. L. 471: "everywhere" -> "on the whole domain"
165. L. 472: Delete "(Figure8)"
166. L. 472: "not the worst of the selection" not coherent with "is the worst performing" at L.471
167. L. 475: Sentence not finished...
168. Figure 8: Candidate 1 has no one green dot : all RMSE are higher than the reference. Please, add a comment on it.
169. Figure 8: Why is it always $\hat{\theta}_1$ in the grey rectangle ? Why the 1 ?
170. Figure 5 - 6 - 8 - 9 - 10: Keep variables appearing in the same order for all figures

3.5.4 Examples of discrepancy term partitioning

171. Figures 9 and 10: add p and u in the grey rectangle, as you added e in the Figure 8
172. L. 483: delete some "are"
173. L. 484: I am not sure that it validates the method properly. But, at least, it shows that the objectives are achieved.
174. L. 488 - 505: The analysis is only conducted for SW. Do you have any comments on the other variables ?
175. L. 499 - 502: split this sentences in two different ones.
176. L. 506: please, explain better the link with the effective degrees of freedom

3.6 Conclusions

177. L. 511: "perturbed physics" -> "perturbed parameters"
178. L. 512: "diverse" : reformulate
179. L. 513: "a number of" -> "different"
180. L. 513: delete "which we illustrate ... (General Circulation Model)"
181. L. 523: delete "examples"
182. L. 525: "CNRM-CM", please use the same name for the same model
183. L. 525: Really 10^6 simulations ? Or is it 100.000 as written in the paper ?
184. L. 525: "of the perturbed parameter of the parametric components of the model errors", please reformulate
185. L. 527: Use a more appropriated vocabulary than "diverse"
186. L. 529: "CNRM discrepancy" -> "CNRM **model** discrepancy"
187. 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
188. Add some perspectives. For example, it could be interesting to analyse the parameters value between the different candidates, in order to explain which parameters lead to such biases.

4 Technical corrections

1. L. 21, 27, 33, 59-60, 91, 152, 188-189, 560: put the citation in chronological order
2. L. 23: correct the parenthesis "(CMIP; Eyring et al, 2016)"
3. L. 40: correct the parenthesis "(such as global mean quantities, Bellprat et al., 2012 ; Williamson et al., 2015)
4. L. 43: correct the parenthesis "(sometimes referred as an "iteratire refocusing" approach; Williamson et al, 2017) ...
5. L. 59: PCA not defined
6. L. 52, 64, 69, 131, 148, 149, 163, 196, 206, 243, in Table 1: correct the parenthesis in the citation (parenthesis around the date and not the name)
7. L. 159: Error in the section referring: "Section 2.4" and "2.5" instead of "Section 2.3" and "2.4"
8. L. 207: Error in the section referring: "Section 3.1" and "3.2" instead of "Section 3.2" and "3.3"
9. L. 211: error in the section referring number
10. L. 232: I think you mean "equation 14", not "Section 14".
11. L. 242: "high-order modes" repeated twice.
12. L. 245: "point"
13. L. 268: "for" repeated twice
14. L. 383 : "e.g.", belonging to which sentence ?
15. L. 420 : "an the" -> "an" or "the" but not the two ones.
16. L. 451: "The Figure" -> "Figure"
17. L. 483: "on the other ends" -> "on the other hands"
18. L. 484: remove the space before ":"
19. L. 490: (Figure 10) -> (Figure 8)
20. L. 525: LH abbreviation not defined
21. L. 533-536: Did you want to do a list as in lines 514-521 ?
22. L. 552: opimization -> optimization
23. Figure 7: "dasehd" -> "dashed"
24. Figure 7: "arounr" -> "around"