

Responses to Reviewer #2

We would like to express our gratitude once more to Reviewer #2 for their comprehensive and precise review, as well as to the editor for their dedication to this manuscript. We agreed to most of the minor comments and technical corrections and ensured that the colour schemes used in our maps and charts allowed readers with colour vision deficiencies to correctly interpret the findings.

Please, find our responses to Reviewer #2 comments below. The comments from Reviewer #2 are in black, our answers are in blue and the line numbers correspond to the track-change document.

Summary

This is the second review of the manuscript by Peatier et al. 2023.

In this study, the authors propose a method to calibrate a climate model, taking into account the model uncertainty due to input parameters. For this purpose, they propose the use of emulator (statistical model) to sample the input parameters space and select sets of parameters values that lead to similar results than a reference. In order to build the emulator, they reduce the dimensionality of the model output by using Empirical Orthogonal Function analysis. To select the candidates, a k-medians clustering method is employed. Furthermore, to better analyse the model error, a decomposition according to the parametric and non-parametric error is analysed spatially.

The manuscript has been improved from the last version. The last version really suffered of its structure. But, this new version is better structured and the main major comments have been answered. Furthermore, the methodology used is more convincing. For this reason, I accept the publication but only, if the following minors comments are taking into account. They mainly concern the structure or the explanation, not the scientific results.

Minor comments

Title

I find that the title of the manuscript is still very vague. It's not about diverse solutions for calibration, but rather about a single method allowing for diverse calibrations of the same model, taking into account the inherent uncertainty of the model. I suggest a title such as 'Addressing discrepancy in the calibration of a climate model'.

The title has been changed following reviewer's suggestion.

Abstract

I would add few sentences to describe the method succinctly. For example, 'A meta-model, simulating the outputs of a climate model, reduced through principal component analysis, is used to sample the degrees of freedom of the model. Thus, a

subset of input parameter values yielding results similar to a reference simulation is identified.'

Agreed

Introduction

- L. 14: Maybe add references for your sentences 'It is an important [...] quantify their uncertainties'
 - [Agreed, see L. 16](#)
- L. 15: 'atmospheric convection or clouds' is quite redundant as, generally, atmospheric convection refers to convective clouds. I suggest 'radiation or turbulence or clouds'.
 - [Agreed, see L.17](#)
- L. 29-30: 'In particular, the potential [...] future climate evolution'. I think this sentence is not completed.
 - [Agreed, "... is rarely considered" L.32](#)
- L.50: I don't agree. There are different studies that have conducted sensitivity analyses to quantify uncertainty even for atmospheric model. See for example (for limited area model):
Di, Z., Q. Duan, W. Gong, C. Wang, Y. Gan, J. Quan, J. Li, C. Miao, A. Ye, and C. Tong (2015), Assessing WRF model parameter sensitivity: A case study with 5 day summer precipitation forecasting in the Greater Beijing Area, *Geophys. Res. Lett.*, 42, 579–587, doi:10.1002/2014GL061623.
Wimmer, M., Raynaud, L., Descamps, L., Berre, L. and Seity, Y.(2022) Sensitivity analysis of the convective-scale AROME model to physical and dynamical parameters. *Quarterly Journal of the Royal Meteorological Society*, 148, 743, 920– 942, doi:10.1002/qj.4239
And even for ARPEGE-Climat, see works done by Laurent Descamps (not published, personal documentation).
 - [We agree that "impossibility" was strong word and we agree that work has been done to quantify uncertainty with PPE, we removed that part of the sentence.](#)
- L. 89: Add a reference to section 2: "The approach, presented in Section 2, is used as a practical..."
 - [Agreed, L.91](#)

Methods

- L. 97: I would suggest to create a new paragraph to present the PPE.
 - [Agreed, L.101](#)
- L. 99-102: This is a long sentence to describe the LHS. Please reformulate to make it clearer.
 - [Agreed, L.103-106](#)
- L. 102: I suggest: 'we consider the annual means averaged over the whole 1979-1981 period as model outputs'
 - [Agreed, L.107](#)
- L. 106: 'the simulated spatial climatology' → 'ARPEGE-Climat' or 'GCM'
 - [Agreed, L.110](#)
- L. 191: Maybe define $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$
 - [We removed the this part of sentences, as the vectors \$\hat{\theta}\$ are defined L.185](#)
- L. 182: 'objective function' → 'multi-variate spatial error'
 - [Agreed, 187](#)

- L. 183: Maybe define here θ_0 .
 - [Agreed, L.189](#)
- L. 183: It is not necessary to define again the reference simulation here.
 - [Agreed, L.186-187](#)
- L. 198-202: It is redundant to write twice 'The selection of candidate [...] are shown in Section'. Maybe write only: 'The selection of candidate calibrations is detailed in Section 2.6, the results for the application to surface temperature are shown in Section 3 and for the multi-variation application, in Section 4.'
 - [Agreed, L.205-208](#)
- L. 204: I suggest to keep a general view and to consider $\hat{\theta}$ instead of θ^* and $\hat{\theta}$ for the whole section. To make it clearer, I suggest: 'we aim to identify k solutions among m configuration, which explore...'.
 - [Agreed to keep a general view](#)
- L.204-226: I think the explanation in Appendix is quite good and should be put in the manuscript. Furthermore, there are redundancy in this section. This section must be reformulated.
 - [Agreed, L.210-216 and L.231-239](#)

First application: surface temperature error

- Figure 1: As previously, I still suggest to switch the left and right panel, as you firstly describe the right panels in the manuscript.
 - [Agreed, see Figure 1](#)
- L. 261: I would add: 'averaged over the PPE members, according to Figure 1b.'
 - [Agreed, L.277](#)
- L. 263: I suggest to better clarify that $q = 5$ is an example and add 'For example, for a truncation of $q = 5$, ...'.
 - [Agreed, L.279](#)
- All along the Section 3.2 section, you refer to θ , c , ... and not θ_{tas} , c_{tas} , ... I suggest to use the 'tas' abbreviation in the whole section.
 - [Agreed](#)
- L. 300: delete 'includes 10 atmospheric simulations [...] but different initial conditions'
 - [Agreed](#)
- L. 301: I suggest this: 'to compute their associated parametric errors (yellow in Figure 2.b)'.
 - [Agreed, L.318-319](#)
- L. 308: 'This emulator is then optimized to find an example' → 'This emulator is then used to optimized and thus find an example ...'.
 - [Agreed](#)
- L. 343: 'have to have' → 'must have' ?
 - [Agreed](#)
- L. 376: Can you justify these 4 candidates ? Why these 4 ones and not the others ?
 - [This sub-set include the best \(candidate 3\) and the worst \(candidate 7\) model, as well as some models with an interesting error pattern or emulated parametric error \(candidate 10\). Presenting all the 12 candidates on the same figure would make it impossible to read, but all of the candidates are visible in the appendices of the paper.](#)

Second application: multi-variate error

- Section 4.3: Why do you present candidate 6 in Figure 8 and not candidate 9 that you discuss more in the section? I would suggest to discuss and present only the best and worst candidate (1 and 5). This is sufficient. You can still discuss to the other candidates, but I think it is better to put candidates 6 and 10 in appendix.
 - We think it is useful to be able to see a bit of diversity in the spatial patterns of the model errors : candidate 10 has a very different SW biases than the other candidates, for example. Being able to compare these 4 candidates on the same figure helps to convey the message of the paper.

Conclusion

- I would suggest to review the structure of the conclusion. I mean it is better to summarise the study firstly (paragraph L. 540-547), and then present the main results (paragraph L.530-539).
 - Agreed
- L. 578: A missing word ? 'we did not [consider?] the observational uncertainty' ?
 - Agreed

Technical corrections

- We have made all of the technical corrections.
- L.73: "perturbed parameter ensemble" → " PPE"
- L.84: order in the references: "Peatier et al. 2022, Hourdin et al. 2023"
- L. 97: 'A PPE of this model f is created', correct the typology of 'f'.
- L. 98: 'amip' → 'AMIP'
- L.146: add a space between 'Rougier (2007)' and 'and'
- L. 233: '(Salter et al. 2019)' → 'Salter et al. (2019)'
- L. 234: 'yj' → 'zj'
- L. 243: '(Salter et al. 2019)' → 'Salter et al. (2019)'
- L. 270: 'Section 14' ?
- L.309: 'for' repeated twice
- L. 350; add space in '2,4 and 5'.
- L. 505: 'an d' → 'and'
- L. 540: 'Rougier et al. (2007)' → '(Rougier et al., 2007)'
- L. 570: 'Williamson et al. (2013), Hourdin et al (2023)' → '(Williamson et al. (2013), Hourdin et al (2023))'
- L. 575: '(Salter et al. 2019)' → 'Salter et al. (2019)'
- L. 577: '(Howland et al, 2022)' → 'Howland et al. (2022)'

Exploration of diverse solutions for ~~Addressing discrepancy in the calibration of imperfect a climate models~~ model

Saloua Peatier¹, Benjamin M. Sanderson², and Laurent Terray¹

¹CERFACS/CECI, Toulouse, France

²CICERO, Oslo, Norway

Correspondence: Saloua Peatier (peatier@cerfacs.fr)

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 arising 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. A meta-model is used to predict the outputs of a climate model, reduced through principal component analysis. Then, a subset of input parameter values yielding results similar to a reference simulation is identified. 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.

1 Introduction

General Circulation Models (GCM) and Earth System Models (ESM) are the primary tools for making projections about the future state of the climate system. It is an important goal of climate science to continually improve these models and to better quantify their uncertainties ([Balaji et al., 2022](#)). Constraints on computational resources limit the ability to resolve small-scales mechanisms, and sub-grid parametrizations are used to represent processes such as atmospheric ~~convection-radiation,~~ turbulence or clouds. These parametrizations are based on numerous unconstrained parameters that introduce uncertainty in climate simulations. Therefore, climate models are subject to a challenging calibration (or 'tuning') problem. When used as tools of projection of future climate trajectories, they cannot be calibrated directly on their performance. Instead, assessment of performance and skill arises jointly from confidence in the understood realism of physical parametrizations of relevant climatological processes, along with the fidelity of model representation of historical climate change. Practical approaches to model calibration are subject to both data, time and computational constraints.

25 For the simplest models (zero or low dimensional representations of the climate system), model simulations are sufficiently
cheap with sufficiently few degrees of freedom that Bayesian formalism can be fully applied to estimate model uncertainty
(Ricciuto et al., 2008; Meinshausen et al., 2011; Bodman and Jones, 2016; Nauels et al., 2017; Dorheim et al., 2020). However,
more complex models such as GCM present a number of difficulties for objective calibration which have resulted in a status quo
30 in which manual calibration remains the default approach (Mauritsen et al., 2012; Hourdin et al., 2017). 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
(Ho et al., 2012; Hourdin et al., 2023) is rarely considered. Failing to explore alternative model configurations can result in
model ensembles which may not adequately sample projection uncertainty. For example, some of the CMIP6 model projec-
35 tions were 'too hot' when compared with other lines of evidence and using all of these models without statistical adjustment
(in a simple 'model democracy' approach) could lead to an overestimate of future temperature change (Hausfather et al., 2022).

Although manual calibration remains by far the most common practice, objective calibration methods have been developed
and tested in climate models (Price et al., 2006; Nan et al., 2014; Sellar et al., 2019). Approaches to date with GCMs have
mainly relied on Perturbed Parameter Ensembles (PPE) of simulations, allowing an initial stochastic sample of the parametric
40 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). Each of these meta-modeling
approaches offers different advantages in terms of accuracy, flexibility and speed (Lim and Zhai, 2017), but often require prior
assumptions on how smooth the parameter response surface might be, how noisy the samples themselves are. Such approaches
45 allow for the definition of plausible or "not ruled out yet" spaces when using a low dimensional output space (such as global
mean quantities, Bellprat et al. (2012); Williamson et al. (2015)), potentially allowing for additional ensemble generations
which sample in the "not ruled out yet" space (Williamson et al., 2015). 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 re-
ferred as an "iterative refocussing" approach, Williamson et al. (2017)). However, the choice of which region to initially focus
50 on depends on advice from model developers and is itself subject to error in emulation. Finally, one of the strongest limitation
when developing GCM automatic tuning approach is the high computational cost of the PPE, ~~leading to the impossibility of
running ensemble large enough for uncertainty quantification~~. Dunbar et al. (2021) relies on calibrate-emulate-sample methods
to generate probability distributions of the parameters at a fraction of the computational cost usually required to obtain them,
allowing for climate predictions with quantified parametric uncertainties.

55

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),

60 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. For example, McNeall et al. (2016) found that
land-surface parameters which were optimal for the simulation of the representation of the Amazon forest fraction were not
optimal for other regions. In another case, structural errors in an atmospheric model were found to increase significantly with
65 the addition of variables to a spatial metric (Sanderson et al., 2008). As such, the potential structural error component is im-
plicitly related to the dimensionality of the space in which the cost function is constructed. For example, Howland et al. (2022)
demonstrated that the use of seasonally averaged climate statistics, rather than annually averaged, could narrow the uncertainty
of climate model predictions.

70 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, e.g. Higdon et al. (2008);
Wilkinson (2010); Sexton et al. (2012)). Notably, for some spatial applications, this dimensional reduction may be insufficient
to resolve certain important climatological features such as extreme precipitation frequency (Jewson, 2020). This PCA repre-
sentation, however, has some apparent drawbacks for optimization. An orthogonal space constructed from the dominant modes
75 of variability in a ~~perturbed-parameter-ensemble-PPE~~ may not be able to describe some components of the spatial pattern of
model error (O’Lenic and Livezey, 1988). Salter et al. (2019) proposed an approach to global optimization of a model with spa-
tially complex output by a rotation of principal components such that model errors were describable in reduced dimensionality
basis set by including some aspects of higher order modes in the rotated, truncated basis set in order to better describe the error
patterns of ensemble members. The method, however, makes some significant assumptions about the ability of a statistical
80 model to predict the parametric response of high order modes and does not allow for an exploration of structural trade-offs
between different variables, such as those found by McNeall et al. (2016).

In this study, we argue that considering a sub-set of plausible candidate calibrations sampling the diversity of model error
spatial patterns can help better understand the model biases. Such approach could also help to better understand model uncer-
85 tainty in climate projections, as previous studies highlighted the possibility for several calibrations of a single climate model
to present very different future climates ([Hourdin et al. \(2023\)](#), [Peatier et al. \(2022\)](#)[Peatier et al. \(2022\)](#), [Hourdin et al. \(2023\)](#)
) . In this sense, we are not searching for an optimal parameter configuration, rather for model configurations which perform
comparably to a reference model version. We lay out an alternative formalism which makes different assumptions about how
ensemble variability in a PPE relates to structural error and how it can thus inform model development. This formalism allows
90 the empirical decomposition of the model error into one component depending on the parameter values, and a component
arising from structural inaccuracies. The approach, [presented in Section 2](#), is used as a practical solution for finding diverse
near-optimal solutions exploring key model error trade-offs. We start by illustrating the method using a simplified uni-variate
case focusing on surface temperature errors (Section 3), before applying it to a more generalised multi-variate tuning case
using 5 climatic fields (Section 4). Finally, we discuss and summarize the main results (Section 5).

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).

100

A PPE of this model ~~f~~ is created, containing 102 ~~amip~~-AMIP simulations Eyring et al. (2016) differing by their parameter values, representing the period 1979-1981 (3 years), with pre-specified Sea Surface Temperatures (Peatier et al., 2022). Thirty model parameters (see Appendix A1) are perturbed with a Latin Hypercube Sampling (LHS) strategy based on a space-filling maximin design $\theta = (\theta_1, \dots, \theta_n)$, producing a variety of simulated climate states ~~in the experiment~~: $F = (f(\theta_1), \dots, f(\theta_n))$ ~~based on a space-filling maximin design $\theta = (\theta_1, \dots, \theta_n)$~~ (Peatier et al., 2022), with $n = 102$ and θ_i a vector of 30 parameter values (Peatier et al., 2022). For the present study, we consider the annual means averaged over the whole 1979-1981 period as model outputs. We write the model output $f(\theta_i)$ as a vector of length l , such that F has dimension $l \times n$, where n is the number of ensemble members ($n = 102$) and l the number of grid points ~~$l = 32768$~~ ($l = 32768$).

105

2.2 EOF analysis

In order to build emulators of ~~the simulated a~~ GCM spatial climatology, general practice is to reduce the dimensionality of the emulated response, and a common strategy is an EOF (Empirical Orthogonal Function) analysis (Higdon et al., 2008; Wilkinson, 2010; Sexton et al., 2012; Salter et al., 2019), which produces n eigenvectors that can be used as basis vectors. Given $n \ll l$, the reconstruction of F is exact and reduces the complexity of the emulator required.

110

Variability in F is explained in descending order of eigenvectors, such that a truncation to the first q modes yields a basis $\Gamma_q = (\gamma_1, \dots, \gamma_q)$ which produces an approximate reconstruction of the initial data, thus further reducing the scale of the emulation problem. Truncation length is often chosen such that a given fraction of ensemble variance (often 90-95%) is preserved (Higdon et al., 2008; Chang et al., 2014), but some authors have argued that higher order modes may need to be included to allow resolution of optimal configurations (Salter et al., 2019). We discuss choices of q in the first application (Section 3).

115

The EOF basis Γ_q is based on the centered ensemble $(F - \mu)$, with μ the ensemble mean. As a result, each anomaly $(f(\theta_i) - \mu)$ is associated with a coefficient $c(\theta_i)$ (or Principal Component, PC) such as:

$$c(\theta_i) = (\Gamma_q^T \Gamma_q)^{-1} \Gamma_q^T (f(\theta_i) - \mu) \quad (1)$$

120

Given an orthogonal basis, the full spatial field of length l can be approximately reconstructed as a function of the q coefficients:

$$f(\theta_i) - \mu = \Gamma_q c(\theta_i) + r_f, \quad (2)$$

with r_f a residual that depends on the choice of q . Considering a variable j (for example, the air surface temperature, as in the first application - section 3.1), such that z_j is the observed field for the variable and $f_j(\theta_i)$ is the model simulated field for that variable, for a given parameter input θ_i . As for F , we can subtract the ensemble mean μ from the observation and project
130 the anomaly of the observation ($z_j - \mu$) (which is also the error of the ensemble mean μ) onto the basis Γ_q using Eq. 1 :

$$z_i - \mu = \Gamma_q c_z + r_z \quad (3)$$

where r_z is a residual representing the part of the observation z_j that can not be projected on the basis Γ_q . This residual r_z will, as r_f , depend on the choice of q but will never (even when $q = n$) equals 0, as the basis Γ_q explains the maximum amount of variability in F but does not guarantee to fully represent the spatial pattern of the observation z_j (Salter et al., 2019).

135

2.3 Model error decomposition

The model error pattern of a given parameter sample, $E_j(\theta_i) = z_j - f_j(\theta_i)$, can be expressed in the basis Γ_q and becomes the sum of a term that depends on the vector of parameter θ_i (here called parametric component), and a term unsolvable in the basis Γ_q (here called non-parametric component) :

$$140 \quad E_j(\theta_i) = \underbrace{\Gamma_q [c_z - c(\theta)]}_{\text{Parametric}} + \underbrace{r_z - r_f}_{\text{Non-parametric}} \quad (4)$$

We could consider a skill score defined by the Mean Square Error (MSE) of the spatial error pattern $E_j(\theta_i)$:

$$e_j(\theta_i) = \frac{1}{l} \sum ((E_j(\theta_i))^2), \quad (5)$$

$$= \frac{1}{l} (\sum (\Gamma_q [c_r - c(\theta_i)] + r_z - r_f)^2) \quad (6)$$

Furthermore, because $(r_z - r_f)$ is orthogonal by construction to the basis Γ_q , the interaction terms in Eq.5 are zero. As
145 a result and using Eq. 4, the integrated model error $e_j(\theta_i)$ becomes a linear sum of a parametric component $p_j(\theta_i)$ and a non-parametric component u_j :

$$e_j(\theta_i) = \frac{1}{l} \sum (\Gamma_q [c_z - c(\theta_i)]^2) + \frac{1}{l} \sum (r_z - r_f)^2, \quad (7)$$

$$= p_j(\theta_i) + u_j \quad (8)$$

2.4 The discrepancy term

150 We consider, following Rougier (2007) and Salter et al. (2019), that an observation z can be represented as a sum of a simulation using the 'best' set of parameter θ^* of the climate model f and a term (initially unknown) representing discrepancy η .

$$z = f(\theta^*) + \eta \quad (9)$$

The discrepancy effectively represents the difference between the climate model and the measured climate that cannot be resolved by varying the model parameters (Sexton et al., 2012). Such differences could arise from processes which are entirely

155 missing from the climate model, or from fundamental deficiencies in the representation of processes which are included :
 through limited resolution, the adoption of an erroneous assumption in the parameterization scheme or parameters not included
 in the tuning process. The discrepancy η can be defined as the integrated error associated with the optimal calibration θ^* .
 Considering a variable j , the discrepancy term η_j is defined as :

$$\eta_j = \frac{1}{l} \sum ((z_j - f_j(\theta^*))^2), \quad (10)$$

$$160 \quad = e_j(\theta^*) \quad (11)$$

In this case and following Eq. 4, η_j can also be expressed as a linear sum of a parametric component $p_j(\theta^*)$ and a non-parametric component u_j :

$$\eta_j = \frac{1}{l} \sum (\Gamma_q [c_z - c(\theta^*)]^2) + \frac{1}{l} \sum (r_z - r_f)^2, \quad (12)$$

$$= p_j(\theta^*) + u_j \quad (13)$$

165 The irreducible error component of the climate model is represented by the η term, known as the discrepancy. To make this
 statement, Sexton et al. (2012) have to assert that the climate model is informative about the real system and the discrepancy
 term can be seen as a measure of how informative our climate model is about the real world. Sexton et al. (2012) think of the
 discrepancy by imagining trying to predict what the model output would be if all the inadequacies in the climate model were
 removed. The result would be uncertain and so discrepancy is often seen as a distribution, assumed Gaussian, and described
 170 by a mean and variance (Rougier, 2007; Sexton et al., 2012).

The calibration θ^* is usually defined as the 'best' input setting, but it is hard to give an operational definition for an imperfect
 climate model (Rougier, 2007; Salter et al., 2019). In practice, we can only propose an approximated θ^* and multiple 'best
 analogues' to this approximation exist (Sexton et al., 2012). In this work, we intend to select m near-optimal model candi-
 175 dates $(\hat{\theta}_1, \dots, \hat{\theta}_m)$ approximating θ^* and sampling the discrepancy term distribution η . We discuss optimization using a simple
 emulator design in Section 2.5 and candidates selection in Section 2.6.

2.5 Statistical model and optimization

Optimization requires the derivation of a relationship between the model input parameters θ and the PC coefficients $c(\theta)$. In
 the following illustration and as in Peatier et al. (2022), we consider a multi-linear regression:

$$180 \quad c_{em}(\theta_i) \approx \beta \theta_i + c_0, \quad (14)$$

where β is the least-square regression solution derived from F , and c_0 is the ensemble mean coefficients. The regression
 predictions are used in Eq. 12 to predict the model MSE as a function of input parameters θ_i . More details on the choice and
 performance of the statistical model can be find in the Appendix C.

185 In this study, the objective of optimization is to look for non-unique solutions $(\hat{\theta}_1, \dots, \hat{\theta}_m)$ whose performances are lower or comparable to that of a reference model, yet sampling possible trade-offs in the **objective function**. ~~The reference model will be the model CNRM-CM6-1, tuned by the model developers for the CMIP6 exercise (Roehrig et al., 2020).~~ multi-variate spatial error. This reference model has been validated by the experts and can serve as a threshold to define whether a model calibration is near-optimal. The vector of parameter values associated with this reference model will be named θ_0 .

190

We can then consider a 10^5 member Latin Hypercube sample of the model parameter space and produce a distribution of predicted $p_{em}(\theta_i)$ values. The parametric error associated with the reference model, hereafter named $p(\theta_0)$, is considered as a threshold to define the near-optimal candidates. For a given climatic field j , we consider the subset of m emulated cases $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$, where model error is predicted to be lower than the reference model error :

$$195 \quad p_{em,j}(\hat{\theta}_i) < p_j(\theta_0) \quad (15)$$

For operational use, ESM developers generally attempt to minimize a multi-variate metric (Schmidt et al., 2017; Hourdin et al., 2017), considering n_j different climatic fields. In this case, all the individual errors $e_j(\theta_i)$ and $p_j(\theta_i)$ need to be aggregated in a single score. The simplest way to obtain such multi-variate skill score is to normalize each uni-variate parametric errors $p_j(\theta_i)$ relatively to the reference model error such as :

$$200 \quad p_{em,tot}(\hat{\theta}_i) = \frac{1}{n_j} \sum_{j=1}^{n_j} \frac{p_{em,j}(\hat{\theta}_i)}{p_j(\theta_0)} \quad (16)$$

In this case, the condition for the near-optimal sub-set is :

$$p_{em,tot}(\hat{\theta}_i) < 1 \quad (17)$$

~~For an application to surface temperature, we can consider, within the optimized subset of emulated cases $\hat{\theta}_{tas} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$, a selection of candidate calibrations producing pattern error as diverse as possible while minimizing the aggregated metric $e_{tas}(\hat{\theta}_i)$.~~ The selection of candidate calibrations is detailed in Section 2.6, the results for the application to surface temperature are shown in Section 3. ~~For a and for the multi-variate application, we consider the near-optimal sub-set $\hat{\theta}_{tot} = (\hat{\theta}_1, \dots, \hat{\theta}_m)$, which minimize the metric $e_{tot}(\hat{\theta}_i)$. The selection of candidate calibrations is detailed in Section 2.6, the results are shown in Section in Section 4.~~

2.6 Selection of diverse candidate calibrations

210 Given the subsets of plausible model configuration $\hat{\theta}_{tas}$ or $\hat{\theta}_{tot}$, we aim to identify k solutions which explore different trade-offs. This is obtained through a k-medians clustering analysis, ~~which separate samples in~~. Clustering is a data mining technique that divides a dataset into different categories, based on the similarity between data. The k-medians analysis is a centroid-based algorithm which divide the data into k groups of equal variance, minimizing a criterion known as the inertia, computed as the sum of the minimal distances within the clusters (Hastie et al., 2009; Pedregosa et al., 2011) categories in order to maximize
 215 the similarity of data within a same cluster (Hastie et al., 2009; Pedregosa et al., 2011). Here, the index to measure similarity

between data is the Euclidean distance.

As a first step, we apply the k-medians clustering to the surface temperature Principal Components of the plausible model configuration sub-set $\hat{\theta}_{tas}$, the coefficients $c_{tas}(\hat{\theta}_{tas})$. The medians of the samples in each clusters are called the cen-
220 troids. The centroids are points from the original dataset, therefore we know their associated vector of parameters θ and can use them to sample the sub-set of diverse and plausible configurations. These calibration candidates are tested in the climate model and results are presented in the Section 3.

In a multi-variate context, the candidates should reflect the model error diversity among both the different climatic fields j and the different EOF modes of each field. ~~Considering the subset of plausible configurations $\hat{\theta}_{tot}$, we~~ We apply the k-medians
225 clustering analysis to ~~all~~ the data coefficients $c_j(\hat{\theta}_{tot})$, normalized by the reference model coefficients $c_j(\theta_0)$, for $j=1, \dots, n_j$ climatic fields. ~~In this case, applying the clustering analysis within the EOF space reduce the computational cost and allows to look for candidates representing the diversity of error patterns in a multi-variate context. The clustering will minimize the distances between the coefficients of a same cluster, while maximizing the inter-cluster distances.~~ As for the uni-variate application, the k centroids will be retained as candidates to sample the represent the diversity of error patterns in the plausible subset of
230 configurations. In the Section 4, we propose an application considering 5 climatic variables ($n_j = 5$, Table 1).

The k-medians analysis is sensitive to the choice of cluster numbers k , which depends on the dataset being classified. The inertia can help to estimate how well a dataset was clustered by k-medians. It 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
235 the k-means performance curve, where the decrease in inertia begins, to find the good trade-off : a good model is one with low inertia and low number of clusters k (Cui et al., 2020). Another criteria we can look at is the Dunn index : 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). For both cases, we tested the sensitivity of the analysis to the number of clusters.
240 Following the Elbow method applied the the inertia and the maximization of Dunn's index, we have decided to keep $k = 12$ clusters for both applications. More detailed about the sensitivity of the analysis to the cluster number and the choice of k are given in the Appendix B.

3 First application : surface temperature error

We consider an example problem where the objective is to propose diverse candidates minimizing the Mean Squared Error of a
245 single climatic field, the surface air temperature, when compared with observational estimates. Here we use the BEST dataset (Rohde and Hausfather, 2020), over the simulated period (1979-1981). Observations have been interpolated onto the model grid for a better comparison.

In this example, the first key question will be to select the truncation length of the basis Γ_q . (Salter et al., 2019) Salter et al. (2019)
250 define two main requirements for an optimal basis selection : being able to represent the observations $y_j - z_j$ within the chosen
basis (a feature not guaranteed by the EOF analysis of the PPE), and retaining enough signal in the chosen subspace to enable
accurate emulators to be built for the basis coefficients. Our objectives here are a bit different, as we want to conserve our
ability to identify the trade-offs made by candidates calibrations in the non-parametric components of the model performance.
We argue that the original basis Γ_q is representative of the spatial degrees of freedom achievable through perturbations of the
255 chosen parameters. As such, the degree to which the observational bias projects onto it is itself meaningful and can be used as
a tool to identify components of model error which are orthogonal to parameter response patterns (and therefore not reducible
through parameter tuning).

Furthermore, we want, as (Salter et al., 2019) Salter et al. (2019), to be able to build accurate emulators for the basis coeffi-
260 cients. In this sense, the basis should not include variability modes poorly represented by the emulator. Sections 3.2 and 3.3
discuss the choice of q , the truncation length.

3.1 Assessing meaningful number of degrees of freedom

We first consider how modes of intra-ensemble variability relate to the representation of model integrated Mean Square Error
of surface temperature $e_{tas}(\theta_i)$. Following Section 2.2, by projecting the spatial anomalies of models and observations onto the
265 basis defined by the truncated EOF set, the mean-squared error can be partitioned into a parametric component (the projection
 $p_{tas}(\theta_i)$) and non-parametric component (the residual u_{tas}). Figure 1 considers examples of the full model errors associated
with the PPE simulations and its decomposition for different numbers of EOF modes retained, $q = 102$ being the perfect re-
construction of the full error $e_{tas}(\theta_i)$.

270 While retaining a relatively small number of modes ($q = 5$), the correlation between the full model error and its parametric
component is already really strong among the PPE members, with a Pearson correlation coefficient of $r = 0.982$ (Figure 1ba).
This correlation does not improve a lot when considering higher modes : $r = 0.998$ for $q = 20$ and $r = 0.999$ for $q = 50$. This
implies that only a relatively small number of modes is required to reproduce the ensemble variance in $e_{tas}(\theta_i)$. The variance
in ensemble spatial error pattern could be described by a small number of degrees of freedom.

275 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, according to Figure 1. 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 example, for a truncation of $q = 5$, the non-parametric component of the error u_{tas} is 53% of the
280 total $e_{tas}(\theta_i)$, 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.

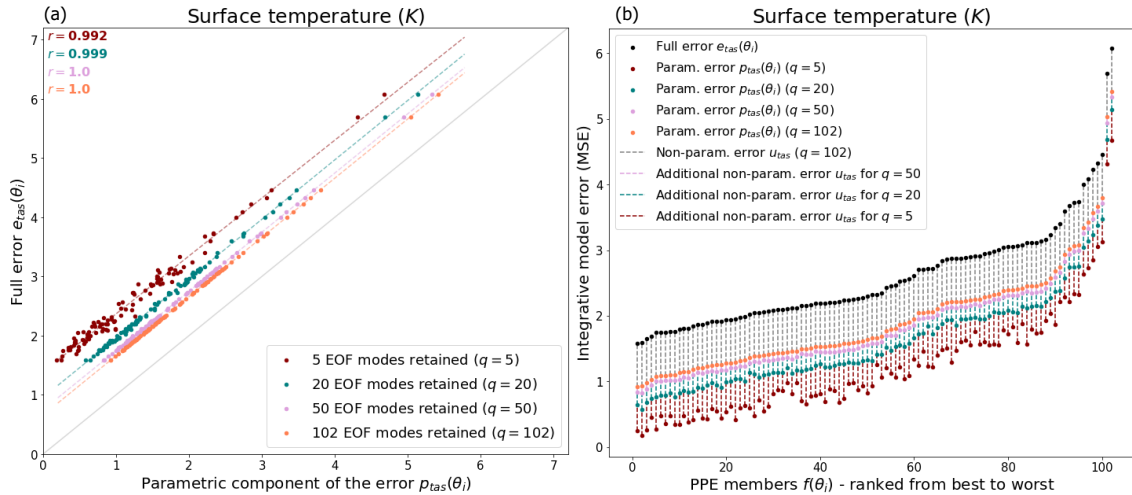


Figure 1. Full model error e_{tas} and its parametric component $p_{tas}(\theta_i)$ for different truncation length : $q = 5$ (red dots), $q = 20$ (blue dots), $q = 50$ (pink dots), $q = 102$ (orange dots). (a) Correlation between the full error e_{tas} and its parametric component $p_{tas}(\theta_i)$ within the PPE. (b) 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.

3.2 Truncation and parametric emulation

In section 3.1, we demonstrated that the majority of variance in model MSE can be described as a function of a small number of spatial modes. In an operational model tuning exercise, we want to make sure that we explain most of the ensemble variance within the truncated EOF basis, so we decided to use a subjective minimum of 85% of explained variance when deciding on the truncation length. Now, how does this relate to parametric dependency? We follow Section 4.2.5 to build a linear emulator relating the model parameters θ to the PC coefficients $e(\theta)c_{tas}(\theta)$. Out of a total of 102 simulations, 92 are randomly selected to form the training set. This training set is used to compute the EOF analysis and to derive least-square regression coefficients of the emulator. The out-of-sample emulator performance is then assessed on the remaining 10 simulations, after projection onto the EOF basis. This process is repeated 10 times with random samples of F used as training sets, to assess the predictive performance of the regression model (i.e. the correlation between out-of-sample predicted $e(\theta)$ and true $e(\theta)c_{em,tas}(\theta)$ and true $c_{tas}(\theta)$).

Figure 2a shows both in-sample and out-of-sample skill scores, in terms of mean and standard deviation across the 10 repetitions. The average of out-of-sample performance cumulative on modes is also represented by the red curve (ex : when $q = 5$, the red curve is the average of the orange curve over the modes 1 to 5 included). We find that out-of-sample emulation skill declines rapidly when the number of modes increases. This result challenges the utility of including high-order modes in the high order modes in spatial emulator of parametric response (as in Salter et al. (2019)), indicating that high order spatial modes

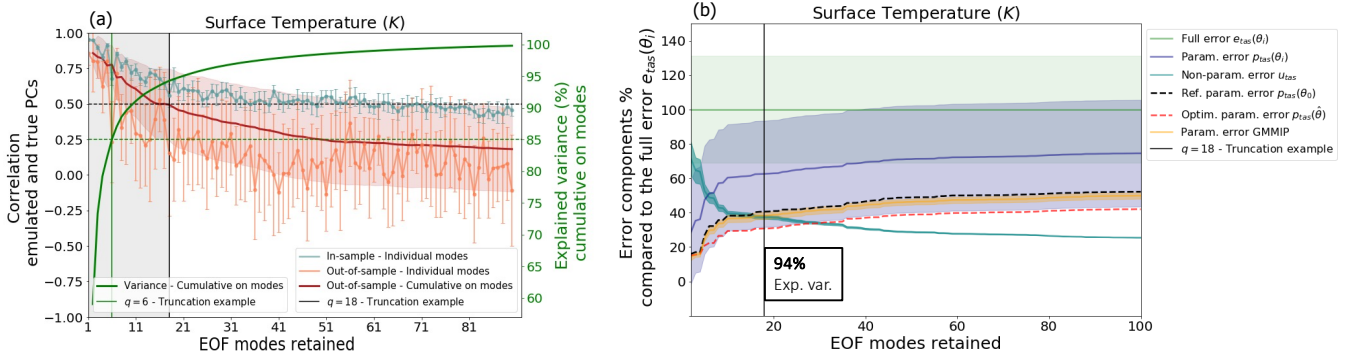


Figure 2. Truncation choice based on parametric emulation and error decomposition. (a) Shows the correlation between the emulated and true PCs of the surface temperature EOF, for the different modes of variability. The correlation is showed within the training set (blue curve) and the test set (orange curve). 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 q modes of the EOF. The dashed green horizontal line shows a threshold of 85% of explained variance and the solid green vertical line is the truncation length needed to satisfy this threshold. (b) Shows the ratio of the error components compared to the full error $e_{tas}(\theta_i)$ (in green) as 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.

300 may be too noisy to represent any parametric signal in the ensemble and emulator design considered here. Here we consider an example of truncation at $q = 18$ that will be used in the rest of the study. It corresponds to the point when the average of out-of-sample performance cumulative on modes reach the arbitrary threshold of 0.5 and explains 94% of the ensemble variance (respective our condition of at least 85% of explained variance).

305 Figure 2b shows the ratios between the PPE parametric (dark blue), non-parametric errors (light blue) and the total error (green), as a function of the number of EOF modes retained. We see that for a EOF basis retaining 1 to 5 modes, each component represents around 50% of the total error on average. For the truncation of $q = 18$, the parametric error represents 63% of the full error on average, and the non-parametric error 37%. This ratio evolves slowly when adding higher modes, and for a perfect reconstruction ($q = 102$), $\frac{p(\theta_i)}{e(\theta_i)} = 74\%$ and $\frac{u}{e(\theta_i)} = 26\%$. But we also note that the large variability of $p_{tas}(\theta_i)$ across
 310 the PPE (represented by the standard deviation) is constant irrespectively of the number of EOF modes retained, highlighting the strong dependency of this error component on the parameter values. On the other hand, the variability of the residual error u within the PPE decreases when retaining more EOF modes, and is already very small for our truncation example of $q = 18$.

In the context of the Global Monsoons Model Inter-comparison Project (GMMIP, Zhou et al. (2016)), an ensemble of 10
 315 atmospheric-only simulations of the CNRM-CM6-1 was run. In this ensemble, the reference model calibration was used, the

SST was forced with the same observations as the PPE and the members differ by their initial conditions only. This dataset can be used to consider the effect of internal variability on the error decomposition, and will be referred to as the GMMIP dataset. The GMMIP dataset ~~includes 10 atmospheric simulations of CNRM-CM6-1 with the reference calibration but different initial conditions, that~~ can be projected into the PPE-derived EOF basis to compute their associated parametric errors (yellow in Figure 2b). The variability of the parametric component of the error across the GMMIP dataset is very small and does not depend on the truncation length. The fact that, for $q = 18$ or higher, the variability of u is even smaller than the internal variability of the parametric component confirms that this part of the error is not dependent on the parametric values anymore.

Another point to note from Figure 2b is that the reference calibration of the model performs well and shows a near-minimal value of parametric error in the ensemble. Following Eq. 14, we use a multi-linear regression that emulates the parametric component of the model error from the calibration values. This emulator is then optimized-used to find an example of near-optimal calibration $\hat{\theta}$ that minimizes the parametric component of the error. The optimization is done for ~~for~~ all the different truncation lengths. As shown on Figure 2a, the parametric component of the near-optimal calibration $p(\hat{\theta})$ is a bit lower than the parametric error of the reference calibration when retaining 5 or more modes and starts evolving in parallel of the PPE mean when retaining 7 or more modes. The difference between the PPE mean and this example of optimal calibration becomes constant when $q = 7$ or more, suggesting that there is no improvements of the optimization when adding modes higher than 7.

These results suggest that the EOF basis Γ_q truncated at a relatively small number $q = 18$ is a good representation of the parametric component of the model error pattern. Therefore, the truncation can be used to identify the residual u , that does not depend on the perturbed parametric values. Adding further modes has limited impact on the representation of ensemble variation in integrated error and does not improve the ability to find near-optimal candidates because of the poor skill of the higher modes regression prediction. In the following, we will only use a truncation at rank $q = 18$.

3.3 Trade-offs in model candidates

Following the methodology discussed in the Section 2.6, all emulated members with a parametric error lower than the reference are selected from a 100,000 LHS set of emulations and considered as a sub-set of near-optimal calibrations. From this sub-set, 12 candidates have been identified in order to maximize the diversity of model errors. The 12 calibrated set of parameters were then used in the ARPEGE-Climat model to produce actual atmospheric simulations. One of the calibrations leads to a crash in the model and 11 others produced the complete atmospheric simulations. The annual mean surface temperature of these 11 candidates were projected onto the EOF basis computed from the 102 members of the PPE, to obtain the principal components. Figure 3 presents the representation of the first five EOF modes by the principal components of the projected model candidates : the closer the candidates are to the observation in the different modes, the lower their parametric error.

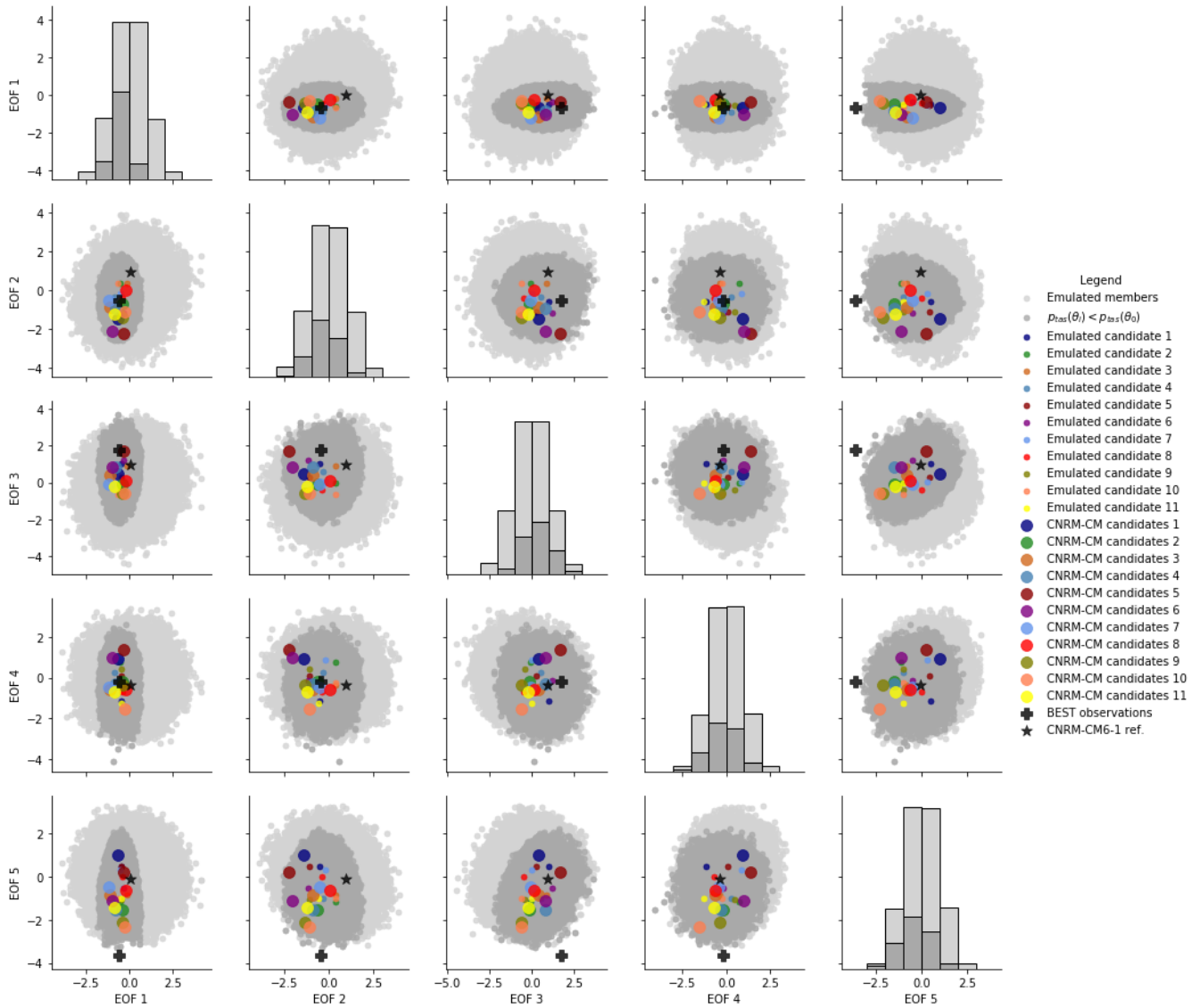


Figure 3. Correlation between the different standardised PC (obtained from the 102 memberPPE EOF) for the 100.000 emulated simulations (light gray), the near-optimal emulated members (dark grey, parametric error lower than the reference CNRM-CM6-1, in dark grey), the 11 emulated candidates (colored dots) and the 11 actual CNRM-CM simulations (colored disks), the reference simulation (star) and the observation (cross).

Figure 3 provides some confidence in both the emulation skill and the method used for the selection of near-optimal and diverse candidates. Although some differences exist between the emulations of the candidates and their actual atmospheric simulations, all of them show principal components within the near-optimal sub-set of calibrations for the 5 first EOF modes, thus respecting the condition for near-optimal calibration. Moreover, the candidates seem to explore a range of principal com-

ponent values as wide as the near-optimal sub-set of calibrations, meaning that we achieve the diversity expected in terms of model errors. In the fifth mode, the projected observations are outside of the emulated ensemble, illustrating that all ensemble members have a non-zero error for this mode, highlighting the existence of a structural bias preventing us from tuning the model to match observations on this axis.

Figure 3 also illustrates the constraints due to optimisation on the principal components of the near-optimal sub-set of calibration. Indeed, the principal components associated with the first EOF mode of the near-optimal sub-set of calibrations (in dark gray on Figure 3) span a very reduced range of values compared to the full emulated ensemble. This result highlights a strong constraint on the first mode of the EOF, stronger than on the other modes. In other words, the candidates ~~have to have~~ must have a representation of the first EOF mode close to the projected observations in order to achieve a parametric error below the reference. This is an expected result, knowing that the first mode explains most of the PPE variance and that the amount of variance explained by each mode individually decreases in higher modes.

Finally, Figure 3 illustrates that it is impossible for the model candidates to perform equally well on all modes and fit observations perfectly. Tradeoffs exist even in this space where the variability is driven by the calibration.

Candidate 5, for example, represents very well the modes 1 and 3, with values of principal components almost equal to those obtained by projecting the observation on the EOF basis, but is further from the observations in the modes 2, 4 and 5. In the same way, candidate 10 performs well for the modes 1, 2 and 5 (being the candidate the closest to observation in mode 5, with the observation outside of the emulated ensemble), but not for modes 3 and 4. Candidate 3 is the best candidate, close to the observations on all modes 1 to 4.

All of the 11 candidates have comparable values of their integrated temperature errors (and all lower than the reference values $p(\theta_0)$ and $e(\theta_0)$) and Figure 3 is a good representation of the trade-offs they have to make in order to minimize this metric. This is a good illustration of the main issue of model tuning : the existence of structural error, which is illustrated here by mode 5, makes impossible the perfect fitting to the observation and candidates are making trade-offs to achieve the metric minimization. This is well known when considering a classic model tuning approach, where multiple climatic variables are considered and the near-optimal calibrations are better representing certain fields at the expense of others in order to minimize a multi-variate metric. Figure 3 is illustrating the problem at the scale of a single field (surface temperature, in this case), highlighting the existence of trade-offs within the near-optimal representation of this field : the temperature will be equally well represented in all the candidates when considering an integrated score (like an MSE), but their spatial error patterns will differ.

3.4 Examples of temperature discrepancy term decomposition

Considering, as described in the Section 2.4, that the error associated with a near-optimal model is an approximation of the discrepancy term magnitude, the candidates selected here illustrate that near-optimal solutions can be obtained with a diversity

of spatial trade-offs that can be made for a minimization problem, even for a single variable output. Moreover, the discrepancy terms can be decomposed in parametric and non-parametric components as seen in Section 2.4. Given the results of Section 3.3, there is a good practical case for choosing a low dimensional basis for calibration - with evidence that it is sufficient to describe the majority of ensemble error variability and that higher modes are not predictable from parameters. The truncation chosen here is $q = 18$ and Figure 4 presents the decomposition of the near-optimal candidates errors, based on this EOF basis $\Gamma_{q=18}$.

For practical reason, only 4 of the 11 candidates are presented in Figure 4, the rest of the candidates can be seen in Appendix (Figure D1). All of the candidates show full temperature MSEs $e(\hat{\theta}_i)$ between $1.62K$ and $1.99K$, so below the MSE of the reference of $e(\theta_0) = 2.01K$. The candidate 7 is the least good, with $e(\hat{\theta}_7) = 1.99K$ and $p(\hat{\theta}_7) = 0.98K$ and the candidate 3 is the best performing model, with $e(\hat{\theta}_3) = 1.62K$ and $p(\hat{\theta}_3) = 0.62K$. The quality of the statistical emulations of the parametric component varies depending on the candidate and the biases over Antarctica are poorly captured by the emulations. We note that the emulation of candidate 3 shows a rather different parametric error than the actual pattern, with an opposite sign of the biases over Antarctica, Australia, India and Argentina, as well as a strong under-estimate of the positive bias over central Africa. For candidate 10, the statistical emulation of the parametric error's spatial pattern is really closed to the truth. We discuss the uneven performance of the statistical predictions in Appendix C and argue that the emulator skill is mostly limited by the size of the training set.

~~Differences between the simulations of temperature and the observations BEST (Rohde and Hausfather, 2020), for the 4 model candidates and the reference. Decomposition of model errors in parametric and non-parametric components using the methodology described in the Section 2, with an EOF basis truncated after the mode 18. The left column shows the full differences between simulations and observations, the second column shows the parametric component of this difference, the third column presents the emulation by the linear regression of this parametric component and the last column is the non-parametric component estimated as the difference between the full error and its parametric component.~~

As stated before, near-optimal candidate errors are our best estimate of the discrepancy term diversity. The full errors shown in Figure 4 display features common to 4 candidates and the reference : negative biases over the mountain regions (Himalaya, Andes and North American mountains) and a positive bias over central Africa. However, the magnitude and position of these biases vary from a model to another, with a particularly strong negative bias over north America in candidate 1 and a strong positive bias over central Africa in candidate 3, for example. This diversity is highlighted when looking at the parametric components of the candidate errors, showing a diversity of error signs and patterns over the poles (especially Antarctica), the South of Europe, India, North Africa and Canada.

The non-parametric components of the errors are smaller and qualitatively similar among the candidates, confirming that they are not strongly controlled by the parameter values. In other words (as expected by the method), the first few modes of the EOF analysis are enough to represent the diversity of model error spatial trade-offs among a sub-set of near-optimal

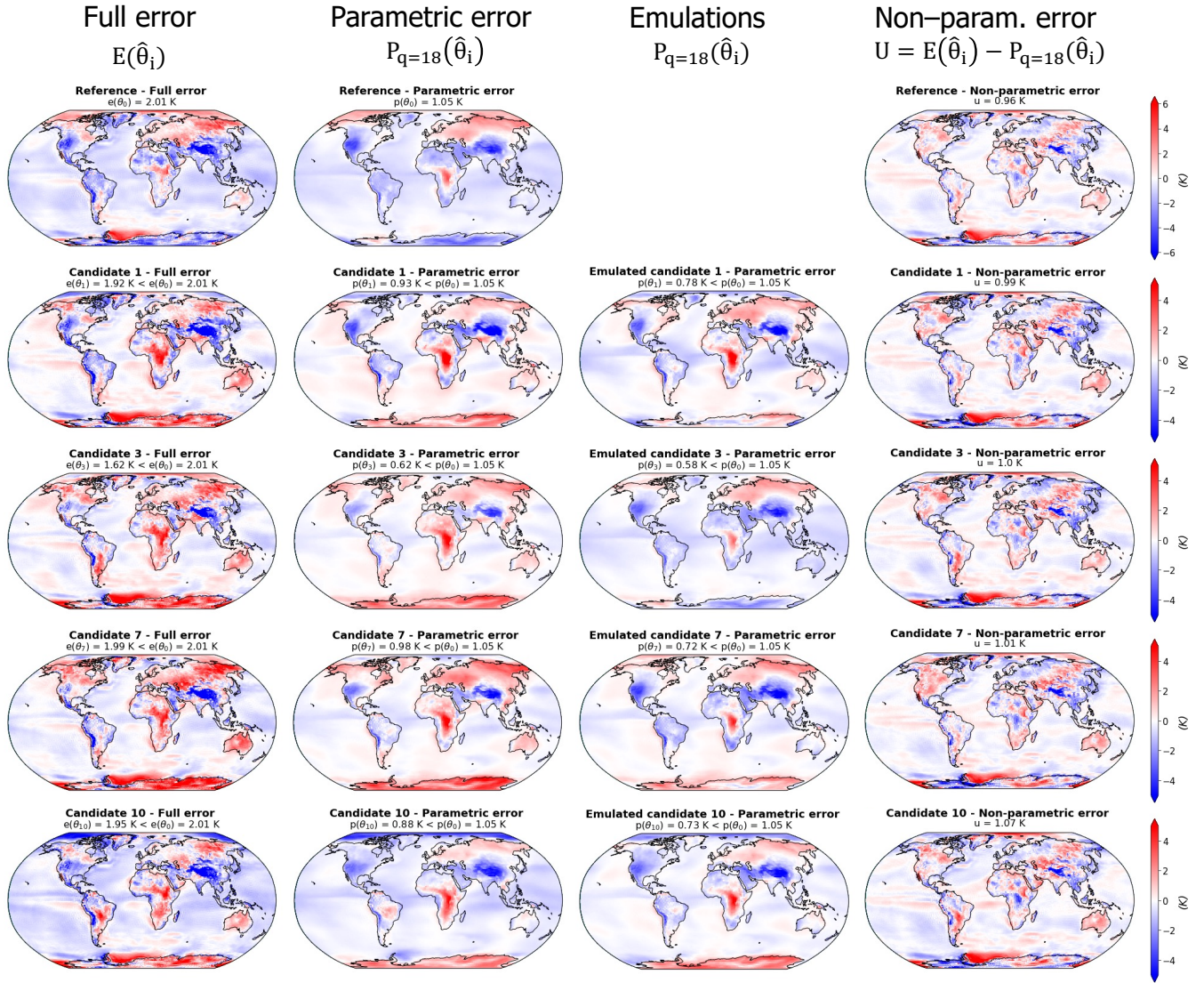


Figure 4. Differences between the simulations of temperature and the observations BEST (Rohde and Hausfather, 2020), for the 4 model candidates and the reference. Decomposition of model errors in parametric and non-parametric components using the methodology described in the Section 2, with an EOF basis truncated after the mode 18. The left column shows the full differences between simulations and observations, the second column shows the parametric component of this difference, the third column presents the emulation by the linear regression of this parametric component and the last column is the non-parametric component estimated as the difference between the full error and its parametric component.

candidates. Moreover, the method allows to visualize and compare these trade-offs through the spatial representation of the parametric component (Figure 4).

4 Second application : multi-variate error

425 4.1 Variables, EOF analysis and truncations

The uni-variate analysis conducted in Section 3 illustrates qualitatively the potential for trade-offs and multiple near-optimal solutions of the climate model optimization problem. In this Section, we considered a single uni-variate metric, allowing to select 12 near-optimal candidates maximizing the diversity of spatial error patterns and trade-offs among the different EOF modes.

Observable variables	Symbol	Units	Data product reference	Years
Surface Temperature	tas	K	(Rohde and Hausfather, 2020)	1979-1981
Precipitation	pr	mm/day	(Huffman et al., 2009)	1979-1981
Sea Level Pressure	psl	Pa	(Saha et al., 2010)	1979-1981
SW flux, TOA	SW	$W.m^{-2}$	(Loeb et al., 2018)	2000-2002
LW flux, TOA	LW	$W.m^{-2}$	(Loeb et al., 2018)	2000-2002

Table 1. Table of observable variables used in this study, plus citations for the data-products used.

430 In an operational GCM tuning application, the metric considered must encompass multiple climate fields and the optimization results in trade-offs between different uni-variate metrics, with near-optimal models better representing some fields at the expense of others. The general solution to model calibration for operational use requires consideration of a wide range of climatological fields spanning model components, including both mean state climatologies, assessment of climate variability and historical climate change. This is inherently more qualitative - requiring subjective decisions on variable choices and weight-
435 ing, which are beyond the scope of this study. However, we can consider an illustration of a multi-variate application, based on 5 climatic fields : the surface temperature (*tas*), the precipitation (*pr*), the short wave (*SW*) and long wave fluxes (*LW*) at the top-of-atmosphere and the surface pressure (*psl*). The model errors will be defined as the MSE between model simulations and the observational datasets lists in Table 1. As for the uni-variate application, EOF analysis of the PPE variance is computed for
440 the error decomposition.

Figure 5 presents the performances of multi-linear regressions in the prediction of the principal components for the 5 fields and we note a strong decrease in out-of-sample prediction skills as we move toward higher EOF modes for all climatic fields. Based on this result, it is clear that, as for the uni-variate application, the optimization should only retain the first few modes.
445 The truncation lengths should be different from a climatic field to another as the linear regressions perform the best for the SW fluxes but have rather poor out-of-sample skill in terms of sea level pressure, for example. Examples of EOF truncations are given on Figure 5, based on an arbitrary threshold of 0.5 for the averaged correlation coefficient of predicted and true out-of-sample PCs. We also ensured that the truncated basis explained at least 85% of the ensemble variance. These examples suggest

that it is possible to retain up to 28 EOF modes for the TOA SW fluxes uni-variate metric, whereas no mode higher than 8
450 should be considered for the sea level pressure in order to keep satisfying statistical predictions. Moreover, some variables re-
quire more EOF modes than other in order to explain most of the ensemble variance. For precipitation, we need to keep 18 EOF
modes in order to explain 85% of variance, whereas for sea level pressure, the first 8 EOF modes explain 92% of the variance.
However, for every climatic field considered, the variance of model errors within the PPE is already very-well represented by
the first 5 EOF modes, as suggested by the correlations between reconstructed and full errors (Figure 6). Considering these
455 truncation lengths, the PPE mean parametric component represents 80% of the full PPE mean error for the TOA SW fluxes,
but only 66% for the sea level pressure.

The error reconstructions presented on Figure 6 are the sums of the parametric components of the errors $p_j(\theta_i)$ and the PPE
mean non-parametric components $u_{j,mean}$. As expected, the PPE mean non-parametric components decrease as higher EOF
460 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). As presented before,
the parametric component $p_j(\theta_i)$ can be emulated with multi-linear regressions, and the PPE mean non-parametric component
 $u_{j,mean}$ can be used as an approximation to reconstruct the full error $e_j(\theta_i)$. This method succeeds to produce high correla-
tions between the reconstructions and the actual full model errors among the PPE, with an offset due to the non-parametric
465 component variability across the PPE, which decreases when retaining more EOF modes. Even though higher EOF modes
are not well predicted by the emulators (Figure 5), they also explain small fractions of the model error variances. As a result,
the performances of the emulators to predict model errors are much more sensitive to the climatic field considered than to the
number of EOF modes retained.

470 On the other hand, the reference calibration CNRM-CM6-1 remains one of the best models of the PPE for most of the
climatic fields and can be considered as near-optimal in the ensemble. Therefore, its model bias can be seen as a representative
of the CNRM-CM discrepancy term. Indeed, the reference CNRM-CM6-1 is the best model for surface pressure and one of
the best for precipitation and TOA fluxes, but several PPE members outperform it for surface temperature. 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
475 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). In the next Section, we
will attempt to identify some of them, in order to illustrate the different choices that could be made when tuning a climate
model.

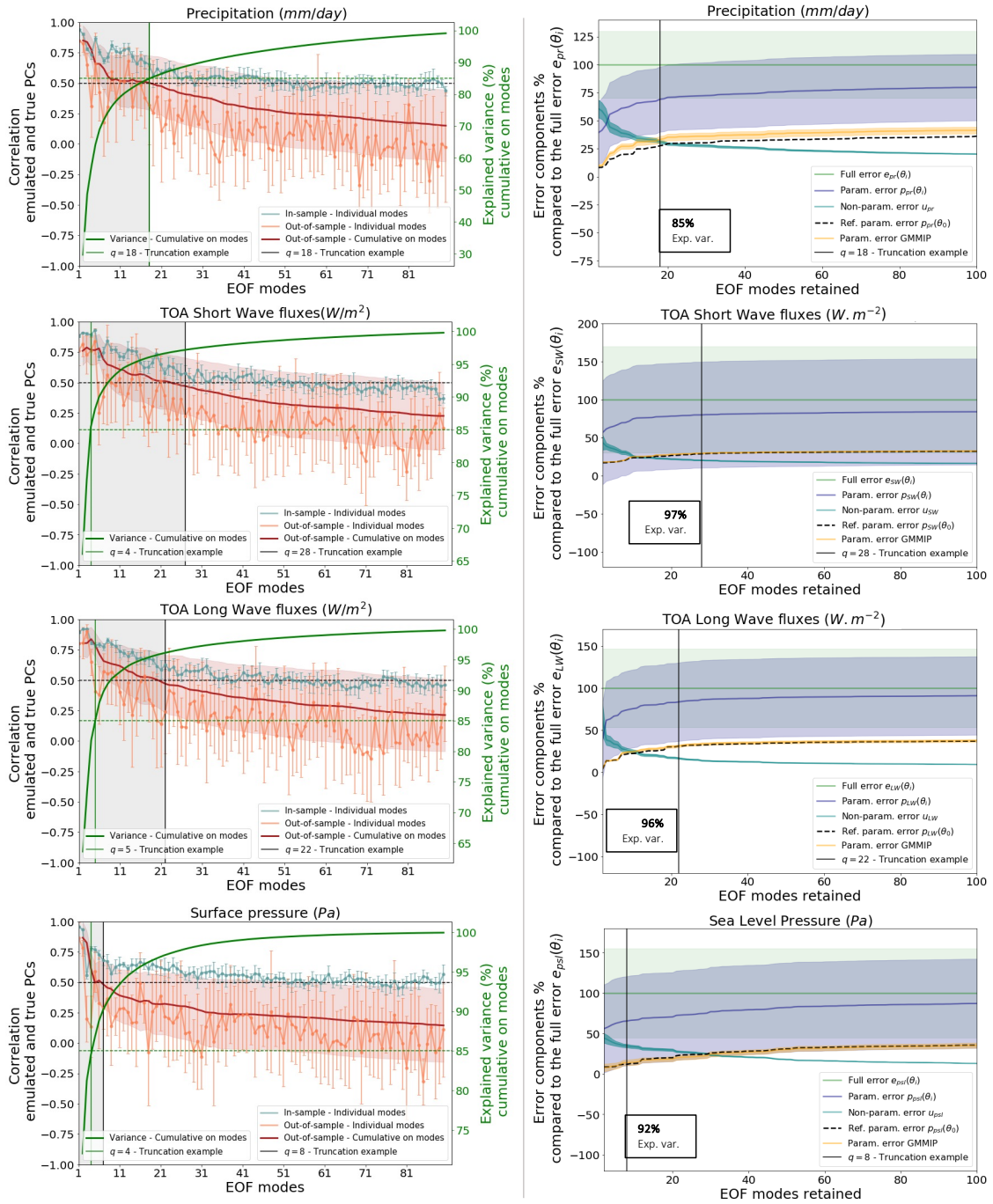


Figure 5. Truncation choice based on parametric emulation and error decomposition for 5 climatic fields : surface temperature, precipitations, TOA SW fluxes, TOA LW fluxes and surface pressure. Same legend as Figure 2, the observations used are listed in Table 1.

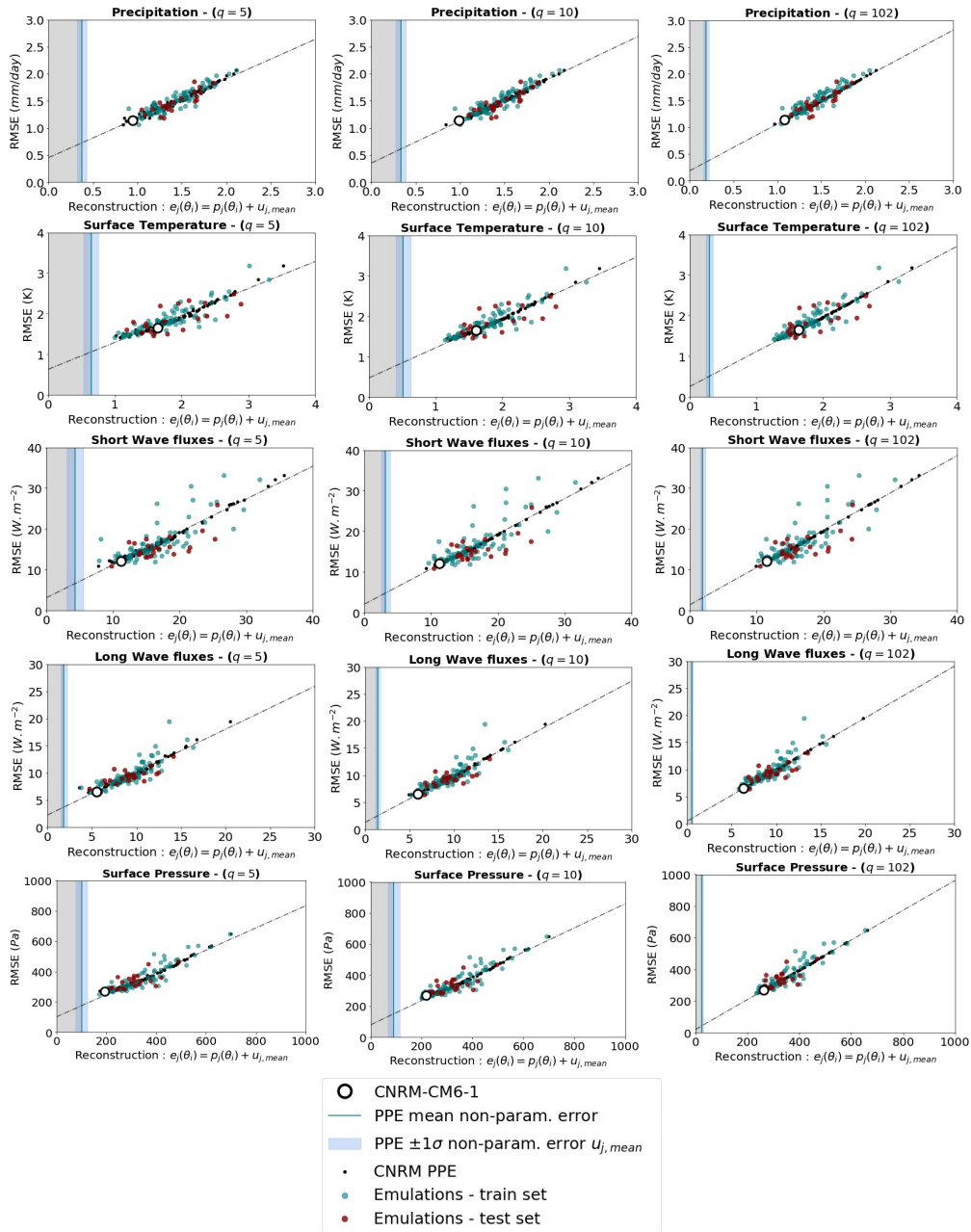


Figure 6. Correlations between full errors (coordinate) and EOF-based reconstructions of these errors (abscissa) using different truncation examples : retaining 5 (left column), 10 (center column) and 102 EOF modes (right column). Results are presented for the CNRM PPE (black dots) and for statistical predictions of the PPE using linear regressions trained on 80% of the data (green dots) and tested on the other 20% (orange dots). For each PPE member or emulation, the error reconstruction is the sum of the parametric component of the errors $p_j(\theta_i)$ and the PPE mean non-parametric component $u_{j,mean}$ (blue line). The variability of u_j among the PPE is represented by the standard deviation σ and the range $\pm 1\sigma$ (light blue shading).

4.2 Candidate selection in a multi-variate context

480 The results in terms of integrated multi-variate skill scores $e_{tot}(\theta_i)$ are presented in the Figure 7. Among the 12 selected candidates, 2 lead to an incomplete simulation and will not be presented here. None of the 10 remaining candidates (light blue dots) show a multi-variate skill-score lower than the CNRM-CM6-1 reference model (orange dashed line). However, all of them have a lower error than the PPE mean (red disk) and 3 of them are in the low tail of the PPE distribution (bellow the red dash). Moreover, most of the CMIP6 models have undergone a tuning process and are considered to represent the control
 485 climate satisfactorily. We can therefore use the CMIP6 ensemble as an indicator of the tolerance that can be given to this multi-variate error. Here we considered the outputs of 40 CMIP6 models that have been interpolated onto the CNRM-CM grid before computing the error. It appears that 9 CNRM-CM candidates selected here have a lower error than the mean of the 40 CMIP6 models (green disk). These 9 CNRM-CM candidates are part of the interval of plus or minus one standard deviation of the CMIP6 error centered around the error of the CNRM-CM6-1 reference model (orange area), indicating that they can be
 490 considered "as good as" the CNRM-CM6-1 reference model given the tolerance considered here. The 10th candidate is above this interval, but is still very close to the CMIP6 ensemble mean and better performing than several CMIP6 models.

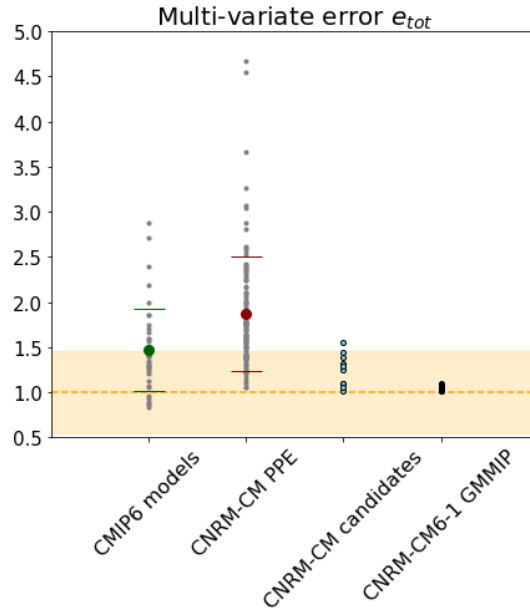


Figure 7. Multi-variate error e_{tot} for the CMIP6 models, the CNRM-CM PPE members, the selection of 10 CNRM-CM candidates and the GMMIP dataset. Each small dots correspond to a model, the bigger dots correspond to the ensemble means and the dashes are the standard deviations. The orange dashed line at 1.0 represents the CNRM-CM6-1 reference model error. The orange area indicates the interval of plus or minus one standard deviation of the CMIP6 errors, centered around the CNRM-CM6-1 reference model error.

The Figure 7 is also presenting the multi-variate error among the 10 simulations from the GMMIP dataset. The 10 perturbed parameter candidates are much more diverse in terms of integrated model error that the 10 perturbed initial conditions mem-

bers. When considering a multi-variate score, it is clear that the effect of internal variability is very small compared to the effect
495 of varying the model parameters.

4.3 Diversity of error patterns among candidates

As described in Figure 7, the 10 CNRM-CM candidates present a satisfactory multi-variate error compared to the CMIP6
ensemble, with 9 of them performing comparably to the CNRM-CM6-1 reference model, while showing a significant diversity
500 compared to the CNRM-CM GMMIP ensemble. We are now interested to see how this diversity translates in terms of spatial
patterns of the uni-variate errors and trade-offs among the variables.

Here again, for practical reason, a sub-set of 4 candidates is presented in Figure 8, the rest of the candidates can be seen
in Figure E1. Within the 4 candidates presented in Figure 8, we have selected the best performing model (Candidate 5) and
505 the worst performing model (Candidate 1). All the candidates have features common to the CNRM reference model (Roehrig
et al., 2020) : an over-estimate of the tropical precipitation and large SW fluxes biases over the mid-latitude eastern border of
Atlantic and Pacific oceans. However, they all have a better representation of surface temperature than the reference model.
None of the candidate show a better representation of precipitation, sea level pressure or LW outgoing fluxes than the reference
and Candidates 1, 2, 5 and 9 provide a better representation of the SW outgoing fluxes.

510

Moreover, some differences exist between the spatial patterns of the candidate errors. Candidate 10 is the model configura-
tion with the lowest MSE of precipitation ($e_{pr}(\hat{\theta}_{10}) = 2.18 \text{ mm/day}$, still higher than the reference), but is also showing
important tropical biases in the radiative fluxes (positive in SW and negative in LW), in the same regions were the model over-
estimate the tropical precipitation, suggesting a biased representation of tropical clouds. Candidate 5 on the other hand, have
515 a better representation of the radiative fluxes in these regions, with a better representation of SW fluxes than the reference and
the best representation of LW among the candidates, suggesting a better representation of tropical clouds. However, candidate
5 is presenting the same biases in precipitation as Candidate 10, with an even higher MSE.

Candidate 1 in the worst performing model of the whole selection, with a total MSE of $e_{tot}(\hat{\theta}_1) = 1.56$. This is mostly
520 due to important biases in precipitation, sea level pressure and LW fluxes representations. Candidate 1 presents strong positive
tropical biases in LW fluxes, over the northern part of South America, central Africa and Indonesia. These areas corresponds
to dry biases in the map of precipitation. Over the tropical oceans, it is one of the candidate that is not showing the negative
LW and positive SW tropical biases, other examples can be found in candidates 3, 8 and 9 (Figure E1). These candidates all
have positive LW biases over the tropical continents, and less biases over the tropical oceans. Interestingly, this is one of the
525 candidate with the best representation of SW fluxes, with candidate 9 (Figure E1), which has a lower MSE. The SW fluxes
biases over the mid-latitude eastern border of Atlantic and Pacific oceans seem to be reduced in these 2 candidates compared
to the other models and the reference.

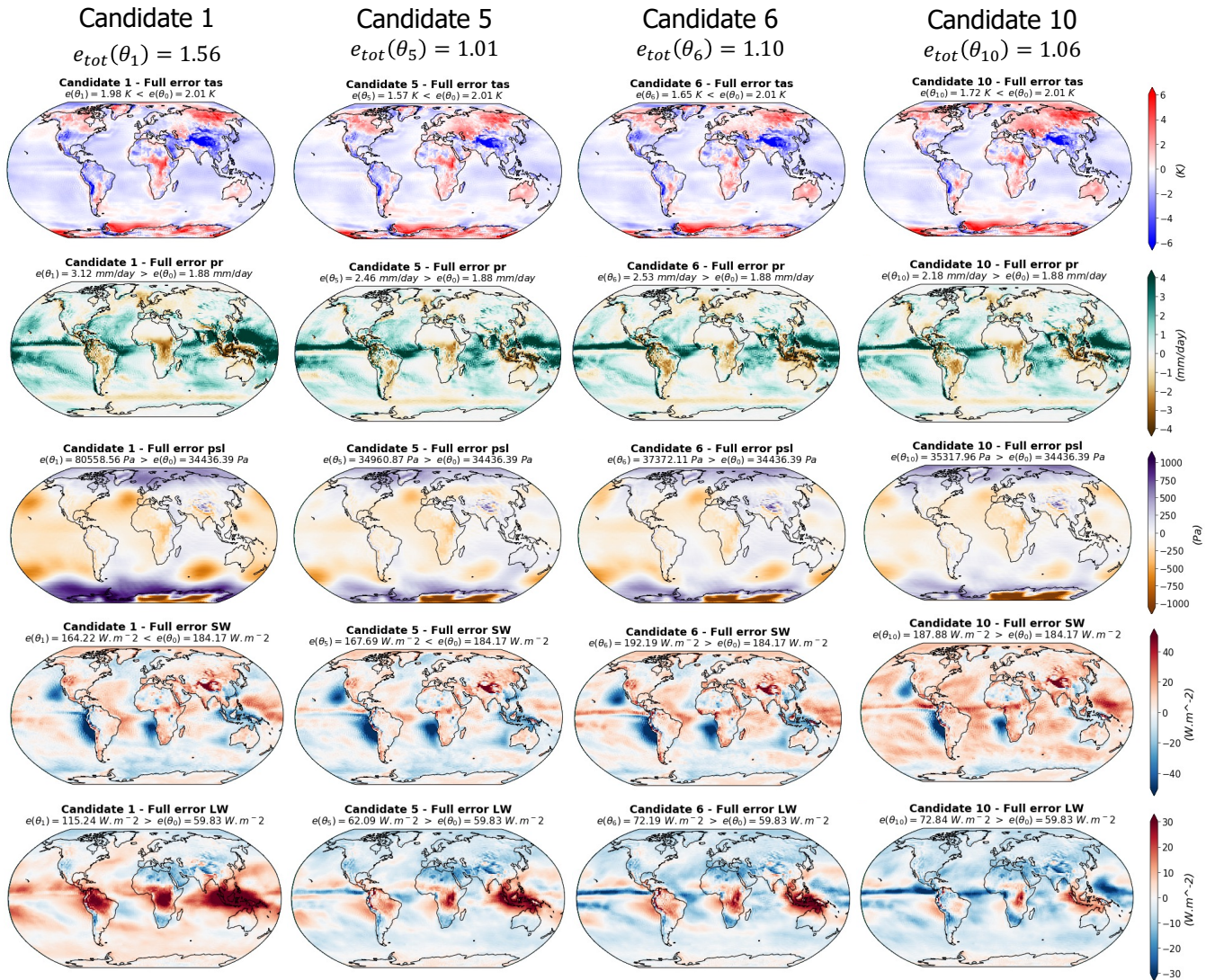


Figure 8. Differences between the simulations and the observations (Table 1), for the 4 model candidates and the 5 variables considered : surface temperature, precipitation, sea level pressure, short and long waves top-of-atmosphere fluxes. Each column represents a model candidate and each row corresponds to a variable. The green dots highlights the cases for which the RMSE is lower than the CNRM reference model.

Candidate 5 is the best performing candidate in terms of multi-variate score and shows errors lower than the CNRM reference model for surface temperature and SW fluxes (Figure 8). Candidate 5 shows a LW error map similar to candidate 6 and 10 but with a reduction of the biases amplitudes. We can assume that the model is better representing tropical clouds, but this does not translate into the best representation of tropical precipitation within the selection.

4.4 Examples of discrepancy term decomposition

Following the method described in Section 2.4, the full error patterns presented in Figure 8 can be decomposed into a parametric components (Figure 9) and non-parametric components (Figure 10). The EOF truncation lengths used for this decomposition
535 are based on the examples given in Figure 5 : 18 modes for *tas* and *pr*, 8 modes for *psl*, 28 for *SW* and 22 for *LW*.

As expected, the candidates parametric component error patterns resemble the full error patterns, with as much diversity in
between the candidates (Figure 9). The non-parametric components on the other hands, are more patchy, are smaller in terms
of amplitude and are common to all the candidates (Figure 10). This validates the method : we were able to select a set of
540 candidates with diverse error patterns and to isolate the error component that is unaffected by parameter variation from the
component that varies during model tuning.

A notable feature of these candidate error decompositions is the *SW* error patterns. The non-parametric component of the
SW error appears very patchy, but contains a small part of the negative biases over the oceanic mid-latitude eastern border that
545 we described in the full error patterns, directly at the continental border (Figure 10). The main part of these biases are pre-
sented in the parametric component of the error (Figure 9). This result suggests that such biases could be enhanced or reduced
by varying the model parameters, but part of them is non-parametric and directly linked to the physics of the model.

In conclusion, when considering error patterns and multi-variate illustration, the effective degrees of freedom in model
550 performance optimization might be smaller than expected. Our method allowed for an empirical exploration of the key trade-
offs that could be made during the tuning, providing interesting information about model non-parametric biases and examples
of alternative model configurations.

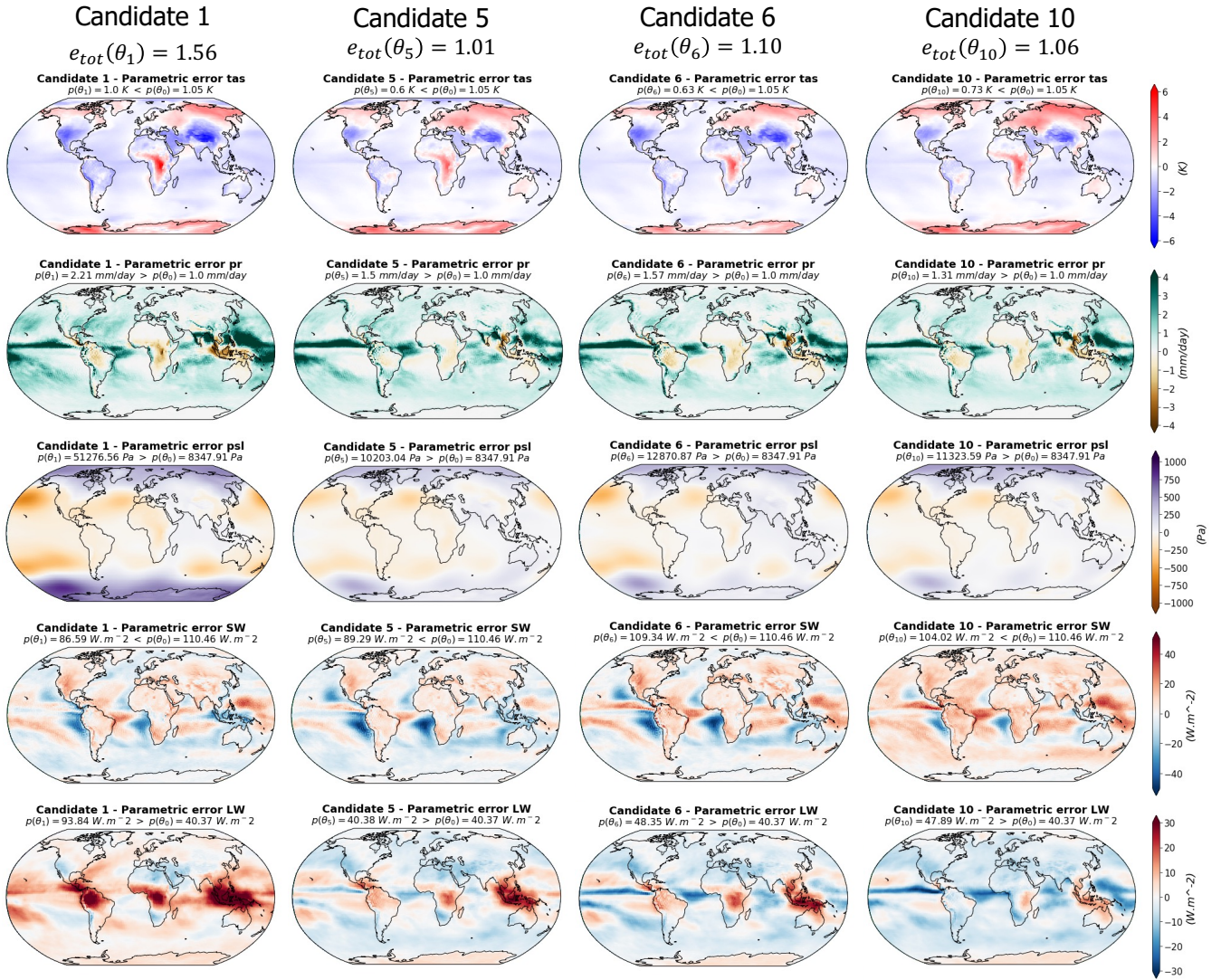


Figure 9. Parametric component of the differences between the simulations and the observations (Table 1), for the 4 model candidates and the 5 variables considered. Each column represents a model candidate and each row corresponds to a variable. The decomposition of model errors in parametric and non-parametric components is based on the methodology described in the Section 2.4, with the EOF bases truncated following examples given in Figure 5 : 18 modes for *tas* and *pr*, 8 modes for *psl*, 28 modes for *SW* and 22 modes for *LW*.

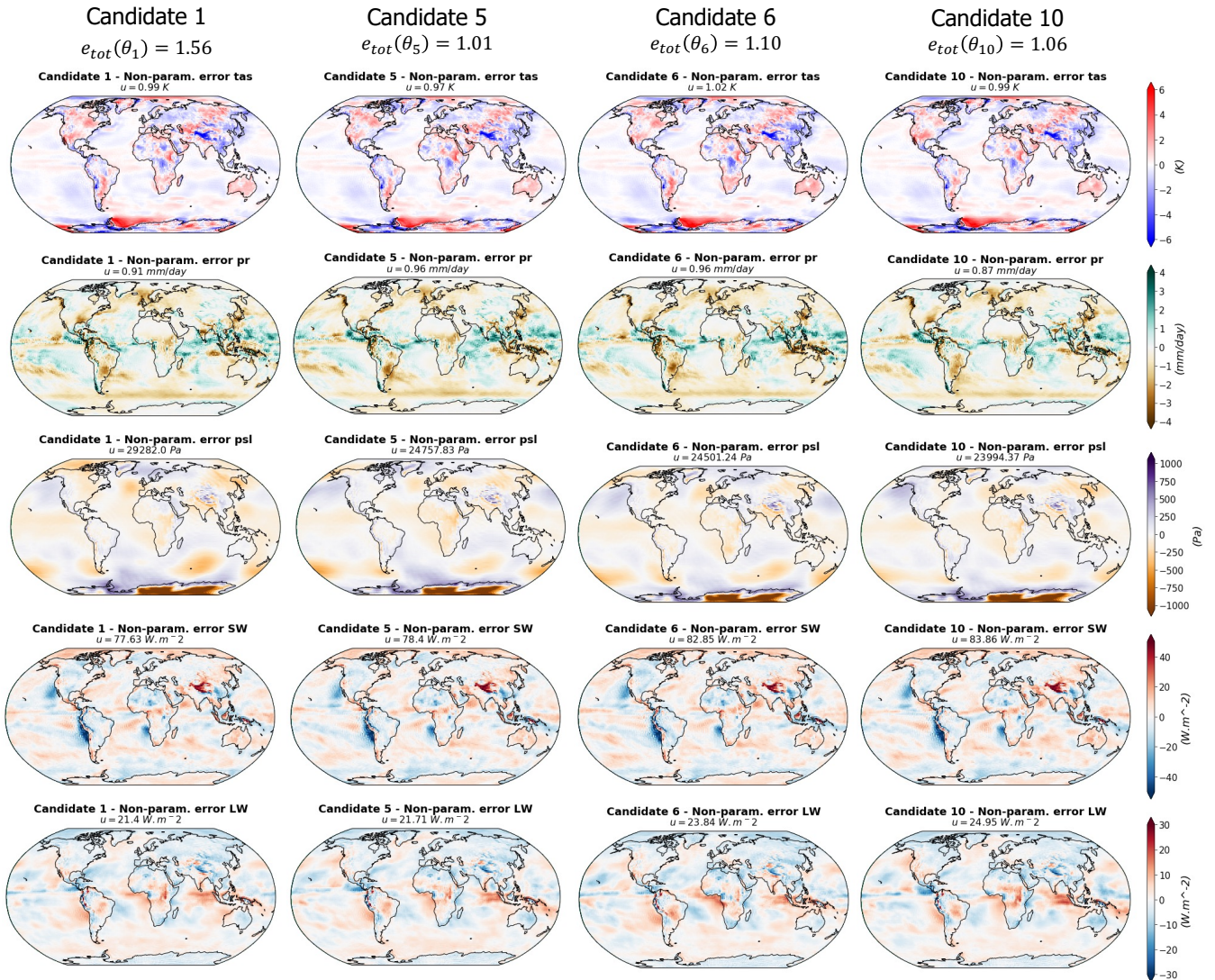


Figure 10. Non-parametric component of the differences between the simulations and the observations (Table 1), for the 4 model candidates and the 5 variables considered. These components are the differences between the Figure 8 and 9.

5 Conclusions

This study presented a new framework, based on a PPE of a CMIP6 General Circulation Model, allowing for the empirical selection of diverse near-optimal candidate calibrations. ~~We have demonstrated that this approach is practically useful for different reasons :-~~

- ~~1. The effective degrees of freedom in model performance response to parameter input are in fact relatively small, allowing a convenient exploration of key tradeoffs-~~
- ~~2. Higher modes of variability should not be included because they cannot be reliably emulated and they do not contribute significantly to the component of model error controlled by model parameters-~~
- ~~3. As such, reference model version shows the lowest integrated performance metric and historical common practices for parameter tuning could be more robust than often assumed-~~
- ~~4. However there remains potential for comparably performing parameter configurations making different model trade-offs~~

560

565

Using the 'best input' assumption ~~-, Rougier (2007)~~ ([Rougier, 2007](#)), we assume that these candidates sample the distribution of atmospheric model discrepancy term. These discrepancy term can be decomposed in parametric and non-parametric components using PPE derived EOF basis. The candidates are selected from a PPE of the CNRM-CM atmospheric model. The optimization is based on multi linear predictions of the parametric components of the model errors, from a 100,000 LH sampling of the perturbed parameters. The candidates are considered near-optimal when their emulated parametric components are lower than the reference parametric component and are selected to exhibit pattern errors as diverse as possible within this near-optimal sub-space using a k-medians clustering algorithm. As such, the sub-set of candidates offer a diversity of model errors sampling the CNRM-CM model discrepancy term distribution, while exploring different trade-offs.

570

The decomposition of the discrepancy terms depends on the truncation choice : the non-parametric component increases when retaining more EOF modes, at the expense of the parametric component. However, we argue that there are no particular benefits in retaining high-order EOF modes, for two reasons. Firstly, the performance of the predictions quickly decreases for the high rank EOF modes, which suggests that these modes are not very predictable from the parameter values. Then, the fact that the first few modes are sufficient to reconstruct the PPE variance of the model errors for the 5 climatic fields considered here and that high modes explain a very small fraction of the PPE variance. Therefore, retaining more EOF modes will increase the part of the model error represented by the EOF basis, called the parametric component, but will not improve the optimization.

580

In the first step, the method was validated for surface temperature error, revealing a diversity of trade-offs among different EOF modes when considering diverse but near-optimal candidates. These trade-offs indicate the presence of a parametric component in the discrepancy terms, which no candidates could eliminate completely. The non-parametric component, on the other hand, is independent of parameter choice and very similar from one candidate to another. These model candidate errors are considered to represent empirical examples of the model discrepancy term for temperature and can offer insights for model developers.

585

In the second step, the framework was applied in a multi-variate context. Trade-offs were observed in error patterns across climatic fields, with different candidates excelling in various aspects. All of the candidates were selected with an emulated

590

parametric error lower than the reference but showed, in practice, higher parametric errors. This result can be attributed to the limitations of the emulators. However, as discussed in Appendix C, our capacity to train emulators is fundamentally limited by the sample size available, which rather small in this study (102 simulations). The use of a non-linear emulator, such as a Gaussian Process, often used in automatic tuning applications [Williamson et al. \(2013\)](#); [Hourdin et al. \(2023\)](#) [\(Williamson et al., 2013; Hourdin et al., 2023\)](#), could help improve predictions, provided we can increase the size of the PPE. In summary, 9 candidates achieved integrated multi-variate scores within CMIP6 ensemble standards, but none of them performed better than the reference model.

We have demonstrated that this approach is practically useful for different reasons :

1. The effective degrees of freedom in model performance response to parameter input are in fact relatively small, allowing a convenient exploration of key tradeoffs
2. Higher modes of variability should not be included because they cannot be reliably emulated and they do not contribute significantly to the component of model error controlled by model parameters
3. As such, reference model version shows the lowest integrated performance metric and historical common practices for parameter tuning could be more robust than often assumed
4. However there remains potential for comparably performing parameter configurations making different model trade-offs

Though we do not attempt it here, the discrepancy estimate could be used in parallel with a history matching approach such as [\(Salter et al., 2019\)](#) [Salter et al. \(2019\)](#), or a Bayesian calibration (Annan et al., 2005) to yield a formal probabilistic result. Enhancing the PPE size would allow for better statistical predictions, maybe through the use of Gaussian Processes as statistical models. We could also consider seasonal metrics instead of annual average, as suggested in [\(Howland et al., 2022\)](#) [Howland et al. \(2022\)](#). Another important caveat of this study is that we did not consider the observational uncertainty. Indeed, additional analyses suggest that our results are sensitive to the observational dataset used. Therefore, defining a formal way to include observational uncertainty in our method for candidate selection would be a valuable improvement of the method. Finally, performing sensitivity analysis could help better understand the effect of each parameters on the biases we observed, potentially leading to a selection of a meaningful sub-set of parameter for a new wave of simulation, in an iterative process.

In summary, we argue that the model discrepancy term can be represented as a sum of two parts - a component which is insensitive to model parameter changes, and a component which represents parameter trade-offs, which manifest as an inability to simultaneously reduce different components of the model bias (e.g. in joint optimization of different regions or fields). We further argue that parameter calibration by hand could be more tractable than often assumed and the reference versions may often be the best model configuration achievable in terms of integrated multi-variate metrics. A feature we see evidenced here by the high performance of the reference simulation, but also reported in similar past PPE efforts (Sanderson et al., 2008; Li

et al., 2019). Finally, we demonstrate a practical method for utilising these concepts for the identification of a set of comparably
625 performing candidate models can inform developers on the diversity of possible trade-offs. The selection of diverse candidates
can help better understanding the limits of model tuning to reduce model error, identify non-parametric biases that are not
visible when looking at the full model error and help choose the model configuration best suited to the research interest.
Moreover, the diversity of model errors can reflect a diversity of future climate responses (Peatier et al., 2022; Hourdin et al.,
2023)) and selecting diverse candidates will help the quantification of uncertainty in climate change impact studies.

Table A1. Description of the 30 perturbed parameters

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	-
RELCAPE	0.2	10.0	2.0	Parameter used in the convection scheme Convective Available Potential Energy closure	-

Appendix B: Clustering analysis and sensitivity to the number of clusters

Clustering is a data mining technique that divides a dataset into different categories, based on the similarity between data. The k -medians analysis is a centroid-based algorithm which divide the data into k categories in order to maximize the similarity of data within a same cluster (Hastie et al., 2009; Pedregosa et al., 2011). Here, the index to measure similarity between data is the Euclidean distance. The k -medians analysis is sensitive to the choice of cluster numbers k , which depends on the dataset being classified.

The inertia can help to estimate how well a dataset was clustered by k -medians. It 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, where the decrease in inertia begins, to find the good trade-off : a good model is one
640 with low inertia and low number of clusters k (Cui et al., 2020). Another criteria we can look at is the Dunn index : 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).-

Sensitivity test of the clustering analyses for the uni-variate (first row) and multi-variate (second row) applications. The
645 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 : $k = 12$ in both applications.

For both applications, the k-medians analysis is repeated 10 times for different values of k and the average of inertia and
Dunn indexes are presented in Figure B1. The inertia sensitivity test suggests that we could chose a value of k in between
650 10 and 20 to be in the Elbow of the curve, for both applications. Then, even though it is less obvious for the multi-variate
application, the results suggest that we should not take a value of k too high, as the Dunn index tends to decrease. Based on
these two criteria, we have decides to keep 12 number of clusters for the analysis : $k = 12$.

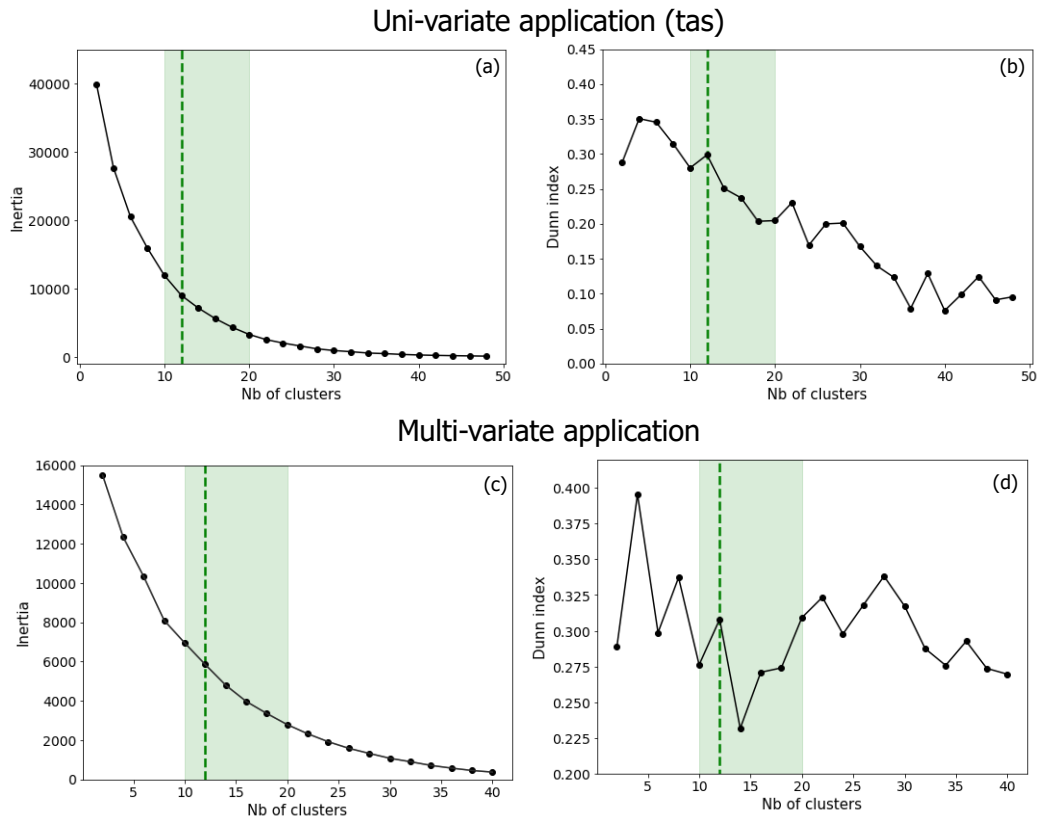


Figure B1. Sensitivity test of the clustering analyses for the uni-variate (first row) and multi-variate (second row) applications. The inertia criteria ((a), (b)) and the Dunn indexes ((b), (d)) are shown depending on the number of clusters (x axes). The green shaded areas present the acceptable number of clusters following the Elbow method applied to the inertia. The green dashed line shows the number of clusters retained for our analyses : $k = 12$ in both applications.

Appendix C: Evaluation of the statistical predictions

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. 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. However, 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 C1 (c)), with a RMSE between predictions and true values representing 8% of the total parametric error (Figure C1 (f)), which is sufficient to validate the use of this model for our study.

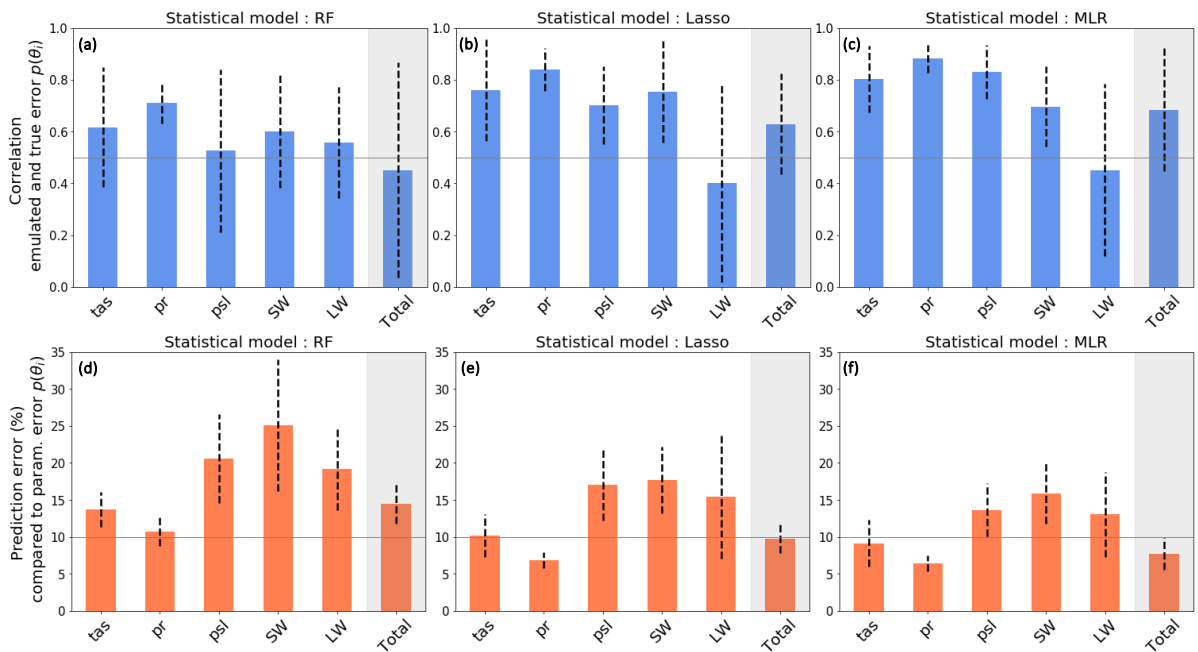


Figure C1. Correlations and RMSE (in % compared to the true values) between emulated and true parametric components of the errors within a test set representing 10% of the dataset. The evaluation is repeated 10 times with random sampling of training and test sets and the mean and standard deviation among 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.

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. 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.

665

Appendix D: First application : additional candidates

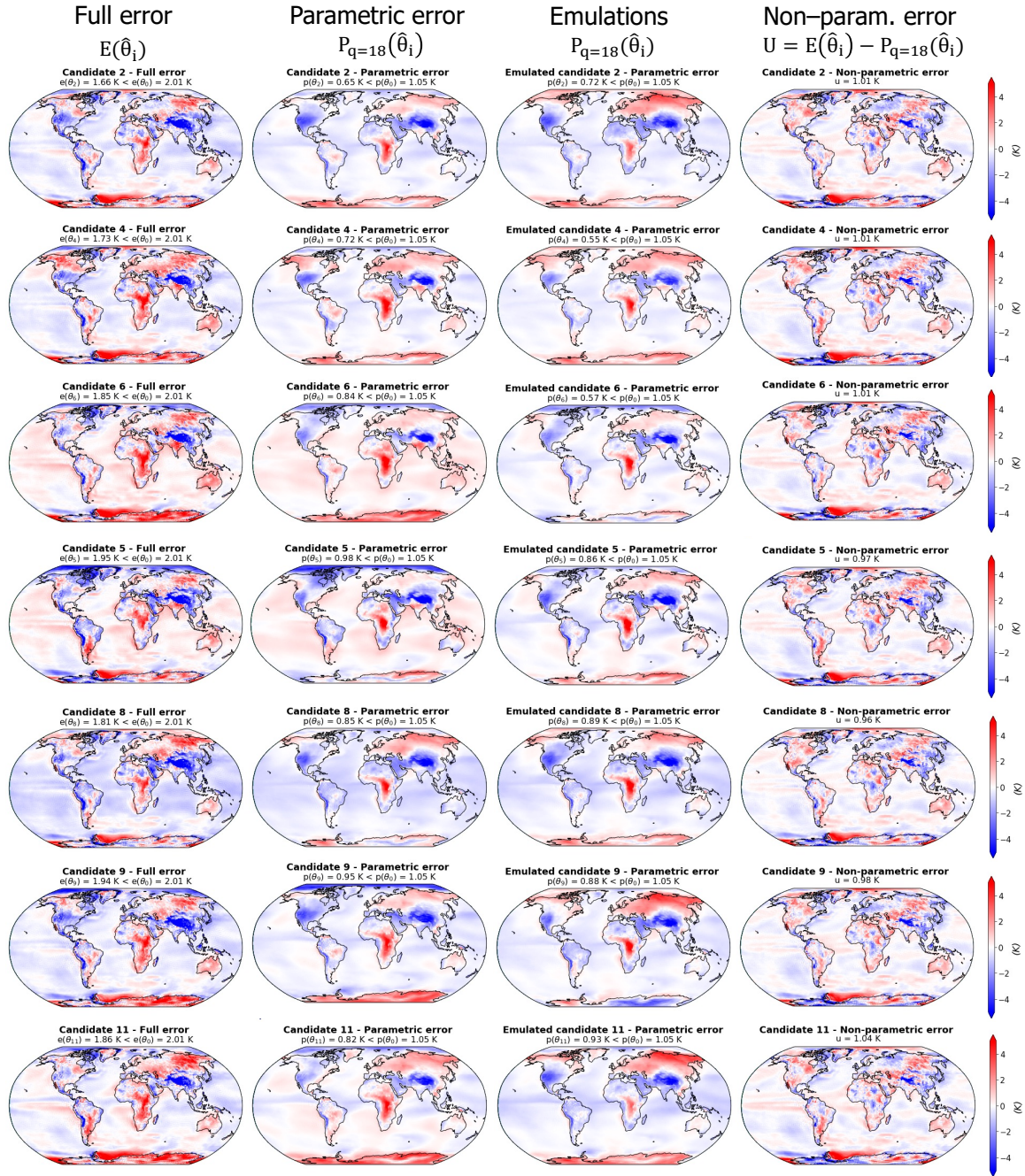


Figure D1. Decomposition of surface temperature error in the first sub-set of candidates. Same caption as Figure 4 for additional candidates.

Appendix E: Second application : additional candidates

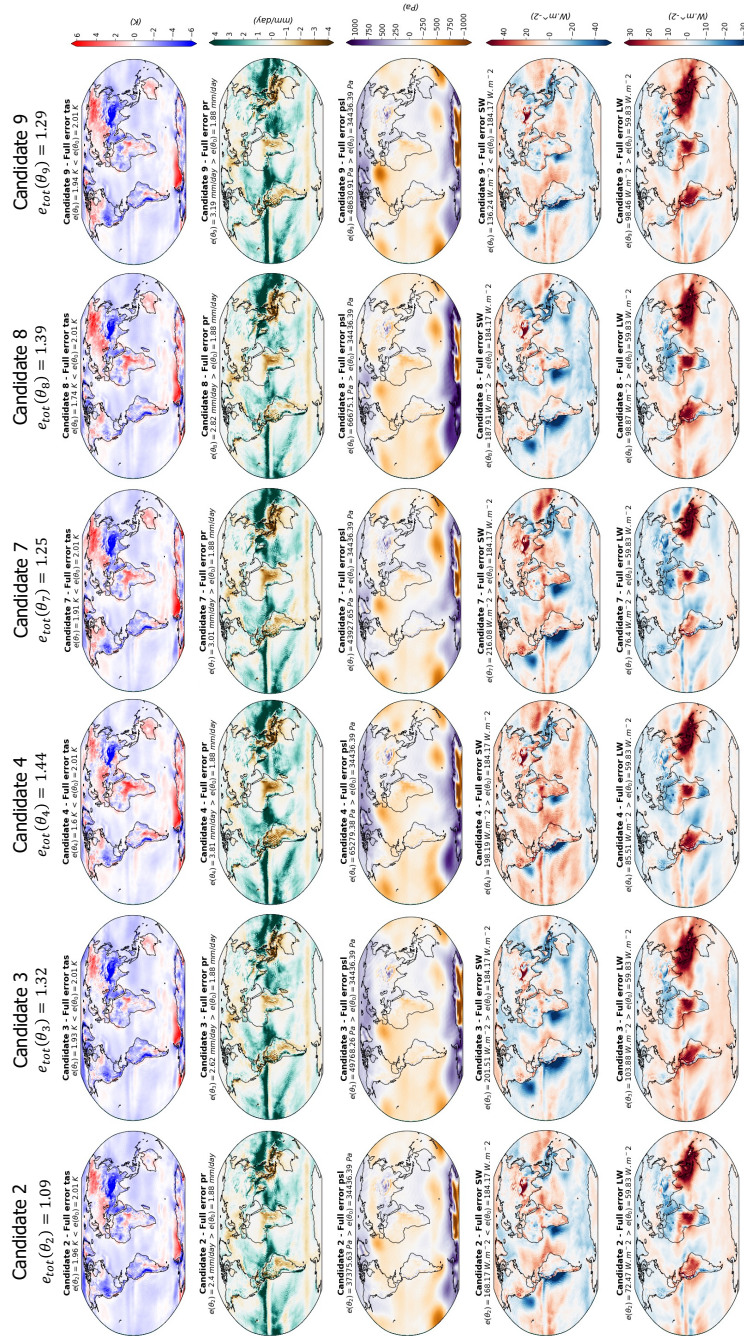


Figure E1. Full model errors in the second sub-set of candidates. Same caption as Figure 8 for additional candidates.

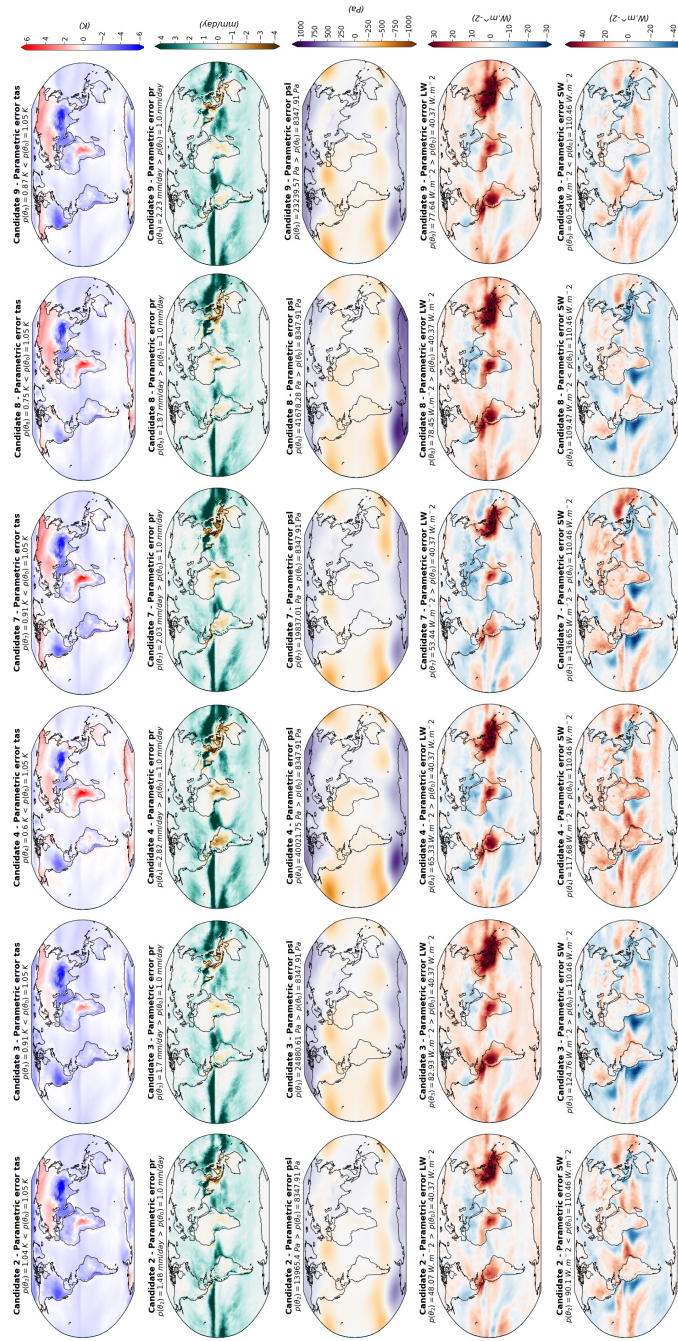


Figure E2. Parametric model errors in the second sub-set of candidates. Same caption as Figure 9 for additional candidates.

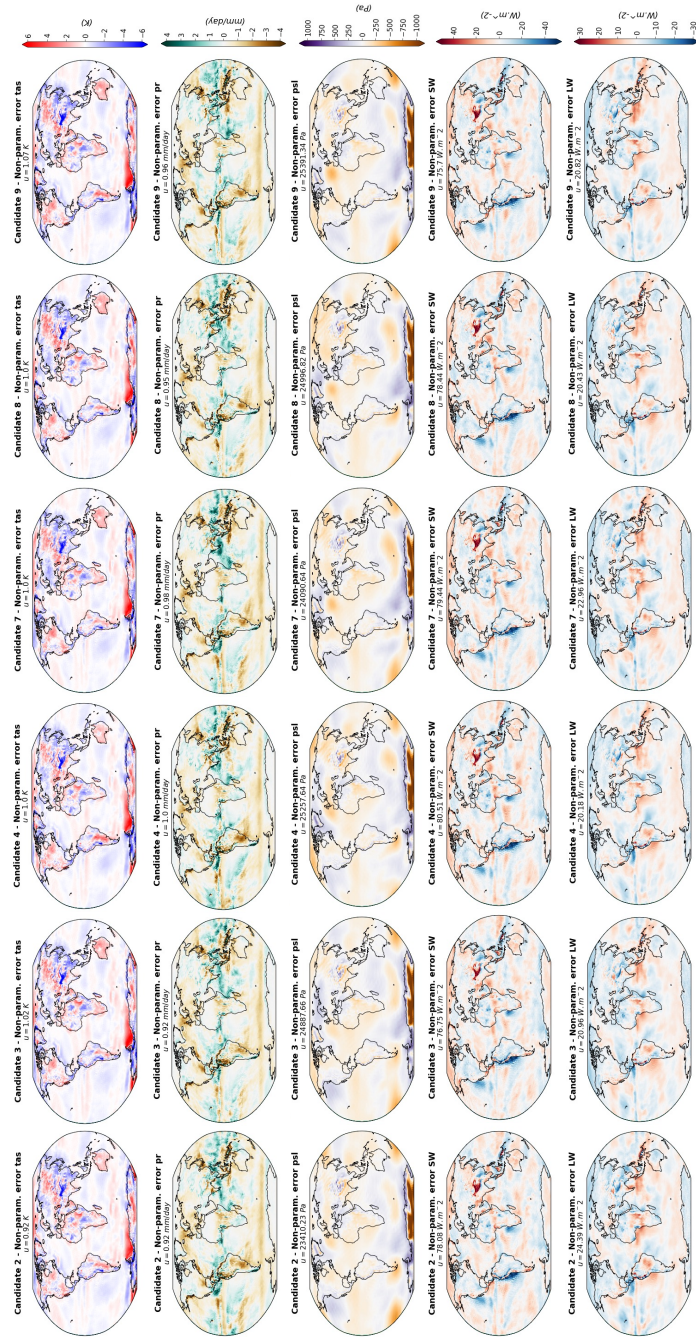


Figure E3. Non-parametric model errors in the second sub-set of candidates. Same caption as Figure 10 for additional candidates.

670 *Author contributions.* SP carried out the simulations and the analysis. SP prepared the manuscript with contributions from all co-authors. BS developed the initial theoretical formalism. SP, BS and LT conceived the analysis. LT supervised the findings of this work.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work is partly funded by the French National Research Agency, project no. ANR-17-MPGA-0016. BS, LT and SP are supported by the H2020 project ESM2025, (Grant ID: 101003536). The authors thank the CNRM-CERFACS modeling group for
675 developing and supporting the CNRM-CM6-1 model.

References

- Annan, J. D., Lunt, D. J., Hargreaves, J. C., and Valdes, P. J.: Parameter estimation in an atmospheric GCM using the Ensemble Kalman Filter, *Nonlinear Processes Geophys.*, 12, 363–371, <https://doi.org/10.5194/npg-12-363-2005>, 2005.
- Balaji, V., Couvreur, F., Deshayes, J., Gautrais, J., Hourdin, F., and Rio, C.: Are general circulation models obsolete?, *Proceedings of the National Academy of Sciences*, 119, e2202075 119, 2022.
- Bellprat, O., Kotlarski, S., Lüthi, D., and Schär, C.: Objective calibration of regional climate models, *Journal of Geophysical Research: Atmospheres*, 117, 2012.
- Bodman, R. W. and Jones, R. N.: Bayesian estimation of climate sensitivity using observationally constrained simple climate models, *Wiley Interdisciplinary Reviews: Climate Change*, 7, 461–473, 2016.
- 685 Chang, W., Haran, M., Olson, R., Keller, K., et al.: Fast dimension-reduced climate model calibration and the effect of data aggregation, *Annals of Applied Statistics*, 8, 649–673, 2014.
- Cui, M. et al.: Introduction to the k-means clustering algorithm based on the elbow method, *Accounting, Auditing and Finance*, 1, 5–8, 2020.
- Dorheim, K., Link, R., Hartin, C., Kravitz, B., and Snyder, A.: Calibrating simple climate models to individual Earth system models: Lessons learned from calibrating Hector, *Earth and Space Science*, p. e2019EA000980, 2020.
- 690 Dunbar, O. R., Garbuno-Inigo, A., Schneider, T., and Stuart, A. M.: Calibration and uncertainty quantification of convective parameters in an idealized GCM, *Journal of Advances in Modeling Earth Systems*, 13, e2020MS002 454, 2021.
- 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, *Geoscientific Model Development*, 9, 1937–1958, 2016.
- Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, *Journal of Geophysical Research: Atmospheres*, 695 113, 2008.
- Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H.: *The elements of statistical learning: data mining, inference, and prediction*, vol. 2, Springer, 2009.
- Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., and Zelinka, M.: Climate simulations: recognize the ‘hot model’ problem, *Nature*, 605, 26–29, 2022.
- 700 Higdon, D., Gattiker, J., Williams, B., and Rightley, M.: Computer model calibration using high-dimensional output, *Journal of the American Statistical Association*, 103, 570–583, 2008.
- Ho, C. K., Stephenson, D. B., Collins, M., Ferro, C. A., and Brown, S. J.: Calibration strategies: a source of additional uncertainty in climate change projections, *Bulletin of the American Meteorological Society*, 93, 21–26, 2012.
- Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., et al.: The art and science of climate model tuning, *Bulletin of the American Meteorological Society*, 98, 589–602, 2017.
- 705 Hourdin, F., Ferster, B., Deshayes, J., Mignot, J., Musat, I., and Williamson, D.: Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections, *Science Advances*, 9, eadf2758, 2023.
- Howland, M. F., Dunbar, O. R., and Schneider, T.: Parameter uncertainty quantification in an idealized GCM with a seasonal cycle, *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002 735, 2022.
- 710 Huffman, G. J., Adler, R. F., Bolvin, D. T., and Gu, G.: Improving the global precipitation record: GPCP version 2.1, *Geophysical Research Letters*, 36, 2009.

- Jewson, S.: An Alternative to PCA for Estimating Dominant Patterns of Climate Variability and Extremes, with Application to US and China Seasonal Rainfall, *Atmosphere*, 11, 354, 2020.
- Li, S., Rupp, D. E., Hawkins, L., Mote, P. W., McNeall, D., Sparrow, S. N., Wallom, D. C., Betts, R. A., and Wettstein, J. J.: Reducing climate model biases by exploring parameter space with large ensembles of climate model simulations and statistical emulation, *Geoscientific Model Development*, 12, 3017–3043, 2019.
- Lim, H. and Zhai, Z. J.: Comprehensive evaluation of the influence of meta-models on Bayesian calibration, *Energy and Buildings*, 155, 66–75, <https://doi.org/https://doi.org/10.1016/j.enbuild.2017.09.009>, 2017.
- Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., Liang, L., Mitrescu, C., Rose, F. G., and Kato, S.: Clouds and the earth's radiant energy system (CERES) energy balanced and filled (EBAF) top-of-atmosphere (TOA) edition-4.0 data product, *Journal of Climate*, 31, 895–918, 2018.
- Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., Haak, H., Jungclaus, J., Klocke, D., Matei, D., et al.: Tuning the climate of a global model, *Journal of advances in modeling Earth systems*, 4, 2012.
- McNeall, D., Williams, J., Booth, B., Betts, R., Challenor, P., Wiltshire, A., and Sexton, D.: The impact of structural error on parameter constraint in a climate model, *Earth System Dynamics*, 7, 917–935, <https://doi.org/10.5194/esd-7-917-2016>, 2016.
- Meinshausen, M., Raper, S. C., and Wigley, T. M.: Emulating coupled atmosphere-ocean and carbon cycle models with a simpler model, *MAGICC6-Part 1: Model description and calibration*, 2011.
- Nan, D., Wei, X., Xu, J., Haoyu, X., and Zhenya, S.: CESMTuner: An auto-tuning framework for the community earth system model, in: 2014 IEEE Intl Conf on High Performance Computing and Communications, 2014 IEEE 6th Intl Symp on Cyberspace Safety and Security, 2014 IEEE 11th Intl Conf on Embedded Software and Syst (HPCC, CSS, ICESS), pp. 282–289, IEEE, 2014.
- Nauels, A., Meinshausen, M., Mengel, M., Lorbacher, K., and Wigley, T. M.: Synthesizing long-term sea level rise projections-the MAGICC sea level model v2. 0., *Geoscientific Model Development*, 10, 2017.
- Neelin, J. D., Bracco, A., Luo, H., McWilliams, J. C., and Meyerson, J. E.: Considerations for parameter optimization and sensitivity in climate models, *Proceedings of the National Academy of Sciences*, 107, 21 349–21 354, 2010.
- O'Lenic, E. A. and Livezey, R. E.: Practical considerations in the use of rotated principal component analysis (RPCA) in diagnostic studies of upper-air height fields, *Monthly Weather Review*, 116, 1682–1689, 1988.
- Peatier, S., Sanderson, B., Terray, L., and Roehrig, R.: Investigating parametric dependence of climate feedbacks in the atmospheric component of CNRM-CM6-1., *Geophysical Research Letters*, p. e2021GL095084, 2022.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825–2830, 2011.
- Price, A. R., Voutchkov, I., Pound, G. E., Edwards, N., Lenton, T. M., and Cox, S. J.: Multiobjective tuning of grid-enabled earth system models using a non-dominated sorting genetic algorithm (NSGA-II), in: 2006 Second IEEE International Conference on e-Science and Grid Computing (e-Science'06), pp. 117–117, IEEE, 2006.
- Ricciuto, D. M., Davis, K. J., and Keller, K.: A Bayesian calibration of a simple carbon cycle model: The role of observations in estimating and reducing uncertainty, *Global biogeochemical cycles*, 22, 2008.
- Roehrig, R., Beau, I., Saint-Martin, D., Alias, A., Decharme, B., Guérémy, J.-F., Voldoire, A., Abdel-Lathif, A. Y., Bazile, E., Belamari, S., et al.: The CNRM global atmosphere model ARPEGE-Climat 6.3: Description and evaluation, *Journal of Advances in Modeling Earth Systems*, 12, e2020MS002 075, 2020.

- 750 Rohde, R. A. and Hausfather, Z.: The Berkeley Earth land/ocean temperature record, *Earth System Science Data*, 12, 3469–3479, 2020.
- Rougier, J.: Probabilistic inference for future climate using an ensemble of climate model evaluations, *Climatic Change*, 81, 247–264, 2007.
- Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., Tripp, P., Kistler, R., Woollen, J., Behringer, D., et al.: The NCEP climate forecast system reanalysis, *Bulletin of the American Meteorological Society*, 91, 1015–1058, 2010.
- Salter, J. M. and Williamson, D.: A comparison of statistical emulation methodologies for multi-wave calibration of environmental models, 755 *Environmetrics*, 27, 507–523, 2016.
- Salter, J. M., Williamson, D. B., Scinocca, J., and Kharin, V.: Uncertainty quantification for computer models with spatial output using calibration-optimal bases, *Journal of the American Statistical Association*, 114, 1800–1814, 2019.
- Sanderson, B. M., Knutti, R., Aina, T., Christensen, C., Faull, N., Frame, D., Ingram, W., Piani, C., Stainforth, D. A., Stone, D., et al.: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes, *Journal of Climate*, 21, 2384–2400, 760 2008.
- Sanderson, B. M., Wehner, M., and Knutti, R.: Skill and independence weighting for multi-model assessments, *Geoscientific Model Development*, 10, 2379–2395, 2017.
- Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., Molod, A., Neale, R. B., and Saha, S.: Practice and philosophy of climate model tuning across six US modeling centers, *Geoscientific Model Development*, 10, 3207–3223, 2017.
- 765 Sellar, A. A., Jones, C. G., Mulcahy, J. P., Tang, Y., Yool, A., Wiltshire, A., O’connor, F. M., Stringer, M., Hill, R., Palmieri, J., et al.: UKESM1: Description and evaluation of the UK Earth System Model, *Journal of Advances in Modeling Earth Systems*, 11, 4513–4558, 2019.
- Sexton, D. M., Murphy, J. M., Collins, M., and Webb, M. J.: Multivariate probabilistic projections using imperfect climate models part I: outline of methodology, *Climate dynamics*, 38, 2513–2542, 2012.
- 770 Wilkinson, R. D.: Bayesian calibration of expensive multivariate computer experiments, *Large-Scale Inverse Problems and Quantification of Uncertainty*, pp. 195–215, 2010.
- Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble, *Climate dynamics*, 41, 1703–1729, 2013.
- 775 Williamson, D., Blaker, A. T., Hampton, C., and Salter, J.: Identifying and removing structural biases in climate models with history matching, *Climate dynamics*, 45, 1299–1324, 2015.
- Williamson, D. B., Blaker, A. T., and Sinha, B.: Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model, *Geoscientific Model Development*, 10, 1789–1816, 2017.
- Zhou, T., Turner, A. G., Kinter, J. L., Wang, B., Qian, Y., Chen, X., Wu, B., Liu, B., Zou, L., and He, B.: GMMIP (v1. 0) contribution to 780 CMIP6: global monsoons model inter-comparison project, *Geoscientific Model Development*, 9, 3589–3604, 2016.