https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

On the spatial calibration of imperfect climate models

Saloua Peatier[1], Benjamin M. Sanderson[2], and Laurent Terray[1]
[1]CERFACS/CECI, Toulouse, France
[2]CICERO, 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 which arise from model structural assumptions makes it impossible to find model versions fully consistent with historical observations, with the potential for multiple plausible configurations which make different tradeoffs between skill in different variables or spatial regions. In this study, we lay out a formalism for making different assumptions about how ensemble variability in a Perturbed Physics Ensemble (PPE) relates to model error, proposing an empirical but practical solution for finding diverse near-optimal solutions. We argue that the effective degrees of freedom in model performance response to parameter input (the 'parametric component') is, in fact, relatively small, illustrating why manual calibration is often able to find near-optimal solutions. Comparison with a perturbed initial condition ensemble reveals that internal variability associated with this parametric component of model error is negligible. Finally, there is a potential for comparably performing parameter configurations making different trade-offs in model errors. These alternative configurations can inform model development and could potentially lead to significantly different future climate evolution.

1 Introduction

Earth System Models are subject to a challenging calibration 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.

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; Bodman and Jones, 2016; Nauels et al., 2017; Dorheim et al., 2020; Meinshausen et al., 2011). However, more complex models such as the General Circulation Models used in successive generations of the Coupled Model Intercomparison Project (CMIP, (Eyring et al., 2016)) present a number of difficulties for objective calibration which have resulted in a *status quo* in which manual calibration remains the default approach (Mauritsen et al., 2012; Hourdin et al., 2017). Such approaches have proven remarkably robust (in that they have not yet been operationally replaced by objective calibration approaches), but leave large intractable uncertainties associated with the potential existence of comparably performing alternative

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

configurations with potentially significantly different future climate evolution (Hourdin et al., 2023; Ho et al., 2012). 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 projections 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 (Sellar et al., 2019; Nan et al., 2014; Price et al., 2006). Approaches to date with GCMs have mainly relied on an initial stochastic sample of the parametric response of the model, followed by the construction of meta-models to emulate the parametric response of model, either through 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 allow for the definition of implausible or "not ruled out yet" (NROY) 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 NROY 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 referred as an "iterative refocussing" approach) (Williamson et al., 2017), but the choice of which region to initially focus on is itself subject to error in emulation.

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 grid scale performance in order to be useful to development. However, whereas in a low dimensional space defined by global mean quantities, it is possible to find model versions which are consistent with observations (Williamson et al., 2015), this is not true for the assessment of a climate model's error integrated over a large number of pixels or variables where 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 the addition of additional variables to a spatial metric (Sanderson et al., 2008). As such, the potential structural error component is implicitly related to the dimensionality of the space in which the cost function is constructed.

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

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

tation, however, has some apparent drawbacks for optimization. An orthogonal space constructed from the dominant modes of variability in a perturbed parameter ensemble 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 spatially 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 an emulator 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 lay out an alternative formalism which makes different assumptions about how ensemble variability in a Perturbed Physics Ensemble (PPE) relates to structural error and how it can thus inform model development. This formalism allows the empirical decomposition of the model error into one component depending on the parameter values, and a component arising from structural inaccuracies. The approach 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).

2 Methods

The model used in this study is ARPEGE-Climat, the atmospheric component of the CNRM-CM6 climate model (Roehrig et al., 2020), referred bellow as f, the climate simulator. A perturbed parameter ensemble (PPE) using the simulator f is created, containing 102 AMIP simulations representing the period 1979-1981 (3 years), with pre-specified Sea Surface Temperatures (SSTs). Thirty model parameters are perturbed with a Latin Hypercube sampling strategy, producing a variety of simulated climate states in the experiment : $F = (f(\theta_1), \ldots, f(\theta_n))$ based on a space-filling parameter design $X = (\theta_1, \ldots, \theta_n)$ (Peatier et al., 2022). For the present study, only climatological annual means of the 1979-1981 historical period are considered. 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 and l the number of grid points. Elements are weighted according to their corresponding area.

2.1 EOF analysis

In order to build emulators of the simulated 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; Sexton et al., 2012; Wilkinson, 2010; Salter et al., 2019), which produces n eigenvectors that can be used as basis vectors. Given $n << l$, the reconstruction of F is exact and reduces the complexity of the emulator required.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

Variability in F is explained in descending order of eigenvectors, such that a truncation to the first q modes yields a basis $\Gamma_q = (\gamma_1, \ldots, \gamma_q)$ 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).

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

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

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

$$f(\theta_i) - \mu = \Gamma_q c(\theta_i) + r_f, \tag{2}$$

with r_f a residual that includes the ensemble mean μ and 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 y_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 the anomaly of the observation $(y_j - \mu)$ (which is also the error of the ensemble mean μ) onto the basis Γ_q using Eq. 1 :

$$y_i - \mu = \Gamma_q c_y + r_y \tag{3}$$

where r_y is a residual, it represents the sum of ensemble mean μ and the part of the observation y_j that can not be projected on the basis Γ_q. This residual r_y 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 y_j (Salter et al., 2019).

2.2 Model error partitioning

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

$$E_j(\theta_i) = \underbrace{\Gamma_q[c_y - c(\theta)]}_{Parametric} + \underbrace{r_y - r_f}_{Non-parametric} \tag{4}$$

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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

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

$$= \frac{1}{l}\left(\Sigma(\Gamma_q[c_r - c(\theta_i)] + r_y - r_f)^2\right) \tag{6}$$

Furthermore, because $(r_y - r_f)$ is orthogonal by construction to the basis Γ_q, the interaction terms in Eq.5 are zero. As 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:

$$e_j(\theta_i) = \frac{1}{l}\Sigma\left(\Gamma_q[c_r - c(\theta_i)]^2\right) + \frac{1}{l}\Sigma(r_y - r_f)^2, \tag{7}$$

$$= p_j(\theta_i) + u_j \tag{8}$$

2.3 The discrepancy term

We consider, following (Rougier, 2007; Salter et al., 2019), that an observation y can be represented as a sum of an optimal calibration $\theta*$ of the climate simulator f and a term (initially unknown) representing discrepancy η.

$$y = f(\theta*) + \eta \tag{9}$$

The discrepancy effectively represents how informative the climate model is about the true climate, and it measures the difference between the climate model and the real climate that cannot be resolved by varying the model parameters (Sexton et al., 2012). Such differences could arise from processes which are entirely 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 parameterisation scheme or parameters not included in the calibration process. The discrepancy η can be defined as the integrated error associated with the optimal calibration $\theta*$. Considering a variable j, the univariate discrepancy term η_j is defined as :

$$\eta_j = \frac{1}{l}\Sigma((y_j - f_j(\theta*))^2), \tag{10}$$

$$= e_j(\theta*) \tag{11}$$

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 :

$$\eta_j = \frac{1}{l}\Sigma\left(\Gamma_q[c_r - c(\theta*)]^2\right) + \frac{1}{l}\Sigma(r_y - r_f)^2, \tag{12}$$

$$= p_j(\theta*) + u_j \tag{13}$$

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

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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 by a mean and variance (Sexton et al., 2012; Rougier, 2007).

The calibration $\theta*$ 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 $\theta*$, hereafter named $\hat{\theta}$, and multiple 'best analogues' to this $\hat{\theta}$ exist (Sexton et al., 2012). In this work, we intend to select diverse optimal model candidates $(\hat{\theta}_1, ..., \hat{\theta}_m)$ sampling the discrepancy term distribution η. In this study, a uni-variate application is presented and the optimal input settings $\hat{\theta}$ will be defined as the calibrations minimizing the skill score $e_j(\hat{\theta})$ or, considering Eq. 12, the parametric component of this skill score : $p_j(\hat{\theta})$. We discuss optimization using a simple emulator design in Section 2.3 and candidates selection in Section 2.4.

2.4 Emulator design 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:

$$c(\theta_i) \approx \beta\theta_i + c_0, \tag{14}$$

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.

In this study, the objective of optimization is to consider non-unique solutions $(\hat{\theta}_1, ..., \hat{\theta}_m)$ comparably performing, yet sampling possible trade-offs in the objective function. In this case, we consider a diversity of candidate calibrations with comparable integrated error metric values. An empirical solution considers a second emulated sample of parameter space, a 10^6 member Latin Hypercube sample of the model parameter space, producing a distribution of predicted $p(\theta_i)$ values. The parametric error associated with the reference calibration of the CNRM-CM model, hereafter named $p(\theta_0)$, is considered as a threshold to define the optimal candidates. For a given climatic field j, we consider the subset of m emulated cases $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$, where model error is predicted to be lower than the reference model error :

$$p_j(\hat{\theta}_i) < p_j(\theta_0) \tag{15}$$

For a uni-variate application, we can now consider, within the optimized subset of emulated cases $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$, a selection of candidate calibrations producing pattern error as diverse as possible while minimizing the common aggregated metric $e_j(\hat{\theta}_i)$. The selection of candidate calibrations is detailed in Section 2.5, the results are shown in Section 3.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

2.5 Selection of diverse candidate calibrations

Given the subset of plausible model configurations $\hat{\theta} = (\hat{\theta}_1, ..., \hat{\theta}_m)$, we then aim to identify a subset of k solutions which explore different trade-offs. For a given variable j, the Root Mean Square distances reconstructed in the EOF basis are computed for each pair of configurations $\hat{\theta}_i$ and $\hat{\theta}_k$ such as :

$$d_j(\hat{\theta}_i, \hat{\theta}_k) = \sqrt{\frac{1}{l}\Sigma(\Gamma_q[c_j(\hat{\theta}_k) - c_j(\hat{\theta}_i)]^2)} \tag{16}$$

The nearest-neighbor pair $(\hat{\theta}_i, \hat{\theta}_k)$ is then identified by the minimum value of $d_j(\hat{\theta}_i, \hat{\theta}_k)$, and one of the pair is removed at random. Near neighbors are then removed until n_k solutions remain, representing the diverse set of plausible solutions. Illustrations of this selection is given in Section 3 for a uni-variate metric ($n_j = 1$, the surface temperature) and five model candidates selected ($n_k = 5$).

For operational use, ESM developers generally attempt to minimize a multi-variate metric (Schmidt et al., 2017; Hourdin et al., 2017), and perspectives for multi-variate application of the discrepancy distribution sampling and partitioning (with $n_j = 5$) is presented in Section 4.

3 First application : surface temperature error

We consider an example problem where the objective is to propose diverse candidates minimizing the Root Mean Squared Error of a 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) define two main requirements for an optimal basis selection : being able to represent the observations y_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 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), to be able to build accurate emulators for the basis coefficients. 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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 **??**, by projecting the spatial anomalies of models and observations onto the 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 partitioning for different numbers of EOF modes retained, $q = 102$ being the perfect reconstruction of the full error $e_{tas}(\theta_i)$.

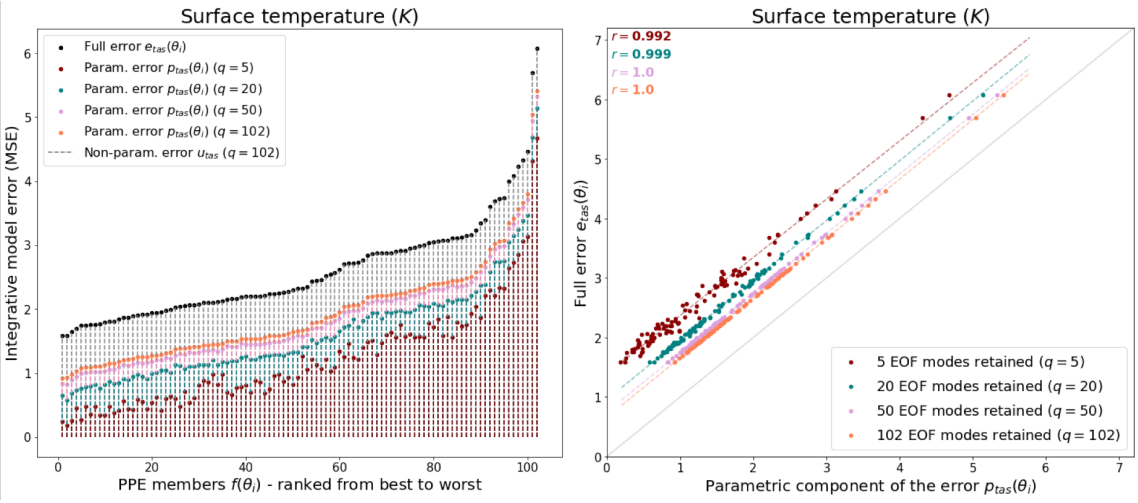

Figure 1. Figure showing the full model error e_{tas} and its parametric component $p_{tas}(\theta_i)$ for different truncation length : $q = 5$ (red dots), $q = 20$ (blue dots), $q = 50$ (pink dots), $q = 102$ (orange dots). For these different truncation lengths, the left plot shows the error partitioning in parametric and non-parametric components in the PPE members $f(\theta_i)$ ranked from lowest to highest error. The right plot shows the correlation between the full error and its parametric component within the PPE.

A number of features are notable in Figure 1. Firstly, with 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 $r = 0.982$. 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 variation in ensemble spatial error pattern could be described by a small number of degrees of freedom.

However, even for the perfect reconstruction of the model error (when $q = 102$), a non-null non-parametric component exists, and its ratio corresponds to 26% of the full model errors averaged over the PPE members. This ratio increases when

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

retaining less EOF modes, and a large fraction of the model error pattern is not represented within the parametric component. Using $q = 5$, for example - the non-parametric component of the error u_{tas} is 53% of the total $e_{tas}(\theta_i)$ in average. Together, this implies that the variation in model error seen in our ensemble can be explained by a small number of degrees of freedom, but a significant fraction of this error is not a function of the perturbed parameters.

3.2 Truncation and parametric emulation

In section 3.1, it was evident that the majority of variance in model MSE could be described as a function of a small number of spatial modes, but how does this relate to parametric dependency? We follow Section 14 to build a linear emulator relating the model parameters θ to the PC coefficients $c(\theta)$. 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 $c(\theta)$ anc true $c(\theta)$).

Figure 2 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 $x = 5$, 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, for example, in (Salter et al., 2019)) - indicating that high order spatial modes 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 poitn 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.

Figure 2 also shows the ratios between the PPE parametric and non parametric errors and the total error , 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 example 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 the PPE (represented by the standard deviation) is constant irrespective 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 $q = 18$.

We can consider also how internal variability impacts the parametric errors. The GMMIP dataset includes 10 AMIP simulations of CNRM-CM6 with the reference calibration but different initial conditions, that can be projected into the PPE-derived

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

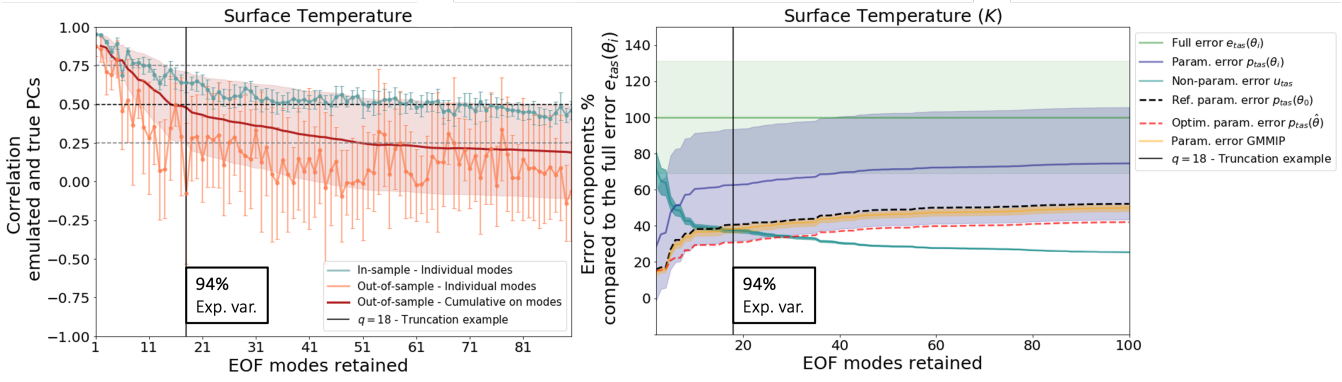

Figure 2. Truncation choice based on parametric emulation and error decomposition. The left plot shows the correlation between the emulated and true PCs of the surface temperature EOF, for the different modes of variability. The correlation is showed within the training set (blue curve) and the test set (orange curve). The evaluation is repeated 10 times with random sampling of training and test sets, the mean and the standard deviation among these 10 evaluation are represented by the dots and the error bars, respectively. The red curve and shading is the mean correlation averaged over the modes cumulatively. The right plot 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.

EOF basis to compute their associated parametric errors. 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 2 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 to find an example of 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 the right plot of Figure 2, the parametric component of the 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. This parallel evolution between the optimal calibration and the PPE mean indicates that 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 with a rank higher than 7.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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 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 2.5, 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 optimal calibrations. From this sub-set, five candidates have been identified in order to maximize the diversity of model errors. The five selected parameter calibrations were then used in the ARPEGE-Climat model to produce actual atmospheric simulations. One of the calibrations leads to a crash in the mode, the four others produced atmospheric simulations of the annual mean surface temperature that could be 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 provides some confidence in both the emulation skill and the method used for the selection of 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 optimal sub-set of calibrations for the 5 first EOF modes, thus respecting the condition for optimal calibration. Moreover, for the 4 first modes, the candidates seem to explore a range of principal component values as wide as the optimal sub-set of calibrations, meaning that we achieve the diversity expected in terms of model errors. Regarding the fifth EOF mode, a lack of diversity within the candidate sub-set appears. The fifth mode is also the only EOF in which 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 allows to see the constraints due to optimisation on the principal components of the optimal sub-set of calibration. Indeed, the principal components associated with the first EOF mode of the optimal sub-set of calibrations (in dark gray on the Figure) 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 modes 3 to 5. In other words, the candidates have to 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

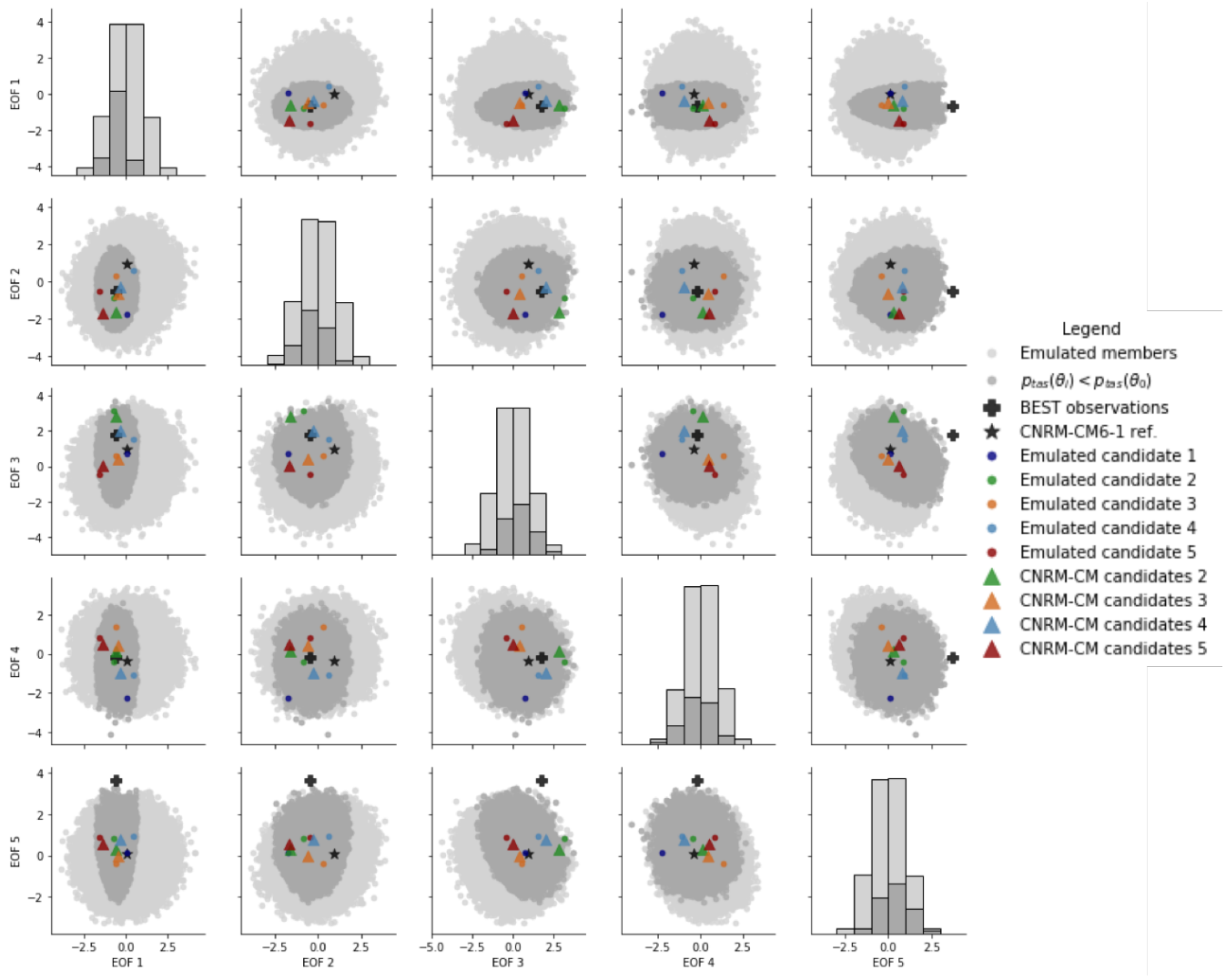

Figure 3. Standardised principal components associated with the 5 first modes of the EOF basis computed on the 102 members of the PPE. The Figure presents the projections of the 100,000 emulated members (light gray), the 'optimal' emulated members (with a parametric error lower than the reference CNRM-CM6-1, in dark grey), the 5 emulated candidates (colored dots) and the 4 candidates CNRM-CM (colored triangles).

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. Looking at candidate 2 for example, it represents very well the modes 1 and 4, with values of principal components almost equal to those obtained by projecting the observation on the EOF basis. For the modes 2,3 and 5 however, candidate 2 is further from the observations. In the same way, candidate 3 performs well for the modes 1, 2 and 4 but not for modes 3 and 5. Candidate 4 is the best candidate,

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

fitting the observations on all modes 1 to 4. Considering mode 5, where the observation is outside of the emulated ensemble, none of the candidates show good performance (but it also, by construction, accounts for the smallest fraction of error variance).

All of the 5 candidates have comparable values of their integrated temperature errors (both $p(\hat{\theta}_i)$ and $e(\hat{\theta}_i)$) 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 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 optimal representation of this field : the temperature will be equally well represented in all the candidates when considering an integrated score (like an RMSE), but their spatial error patterns will differ.

3.4 Examples of temperature discrepancy term partitioning

Considering, as described in the Section 2.3, that the error associated with an optimally calibrated 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 2.3. 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 in any case not predictable from parameters. The truncation chosen here is $q = 18$ and Figure 4 presents the decomposition of the optimal candidates errors, based on this EOF basis $\Gamma_{q=18}$.

The candidates 2 to 4 shows full temperature RMSEs $e(\hat{\theta}_i)$ of $1.31K$, $1.34K$ and $1.27K$, so below the RMSE of the reference of $e(\theta_0) = 1.41K$. The candidate 5 is the least good, and shows a full error slightly above reference : $e(\hat{\theta}_5) = 1.43K$ and $e(\theta_0) = 1.41K$. Overall, the statistical emulations of the parametric component perform acceptably : the emulated patterns of the parametric components are quite close to the actual patterns. We note an exception for the candidate 4 (which is also the best candidate in terms of integrated score), where the statistical emulation overestimates the positive biases everywhere and fails to represent the strong negative biases over central Africa and South Europe. For most of the candidates, the linear regressions were able to emulate the spatial pattern of the parametric error, probably thanks to the choice to truncate the EOF basis after the first few modes ($q = 18$), allowing for skillful emulations (Figure 2).

As stated before, optimal candidate errors are our best estimate of the discrepancy term diversity. The full errors shown in Figure 4 display features common to the 4 candidates and the reference : positive biases over the mountain regions (Himalaya, Andes and North American mountains) and a negative bias over central Africa. However, the magnitude and position of these

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

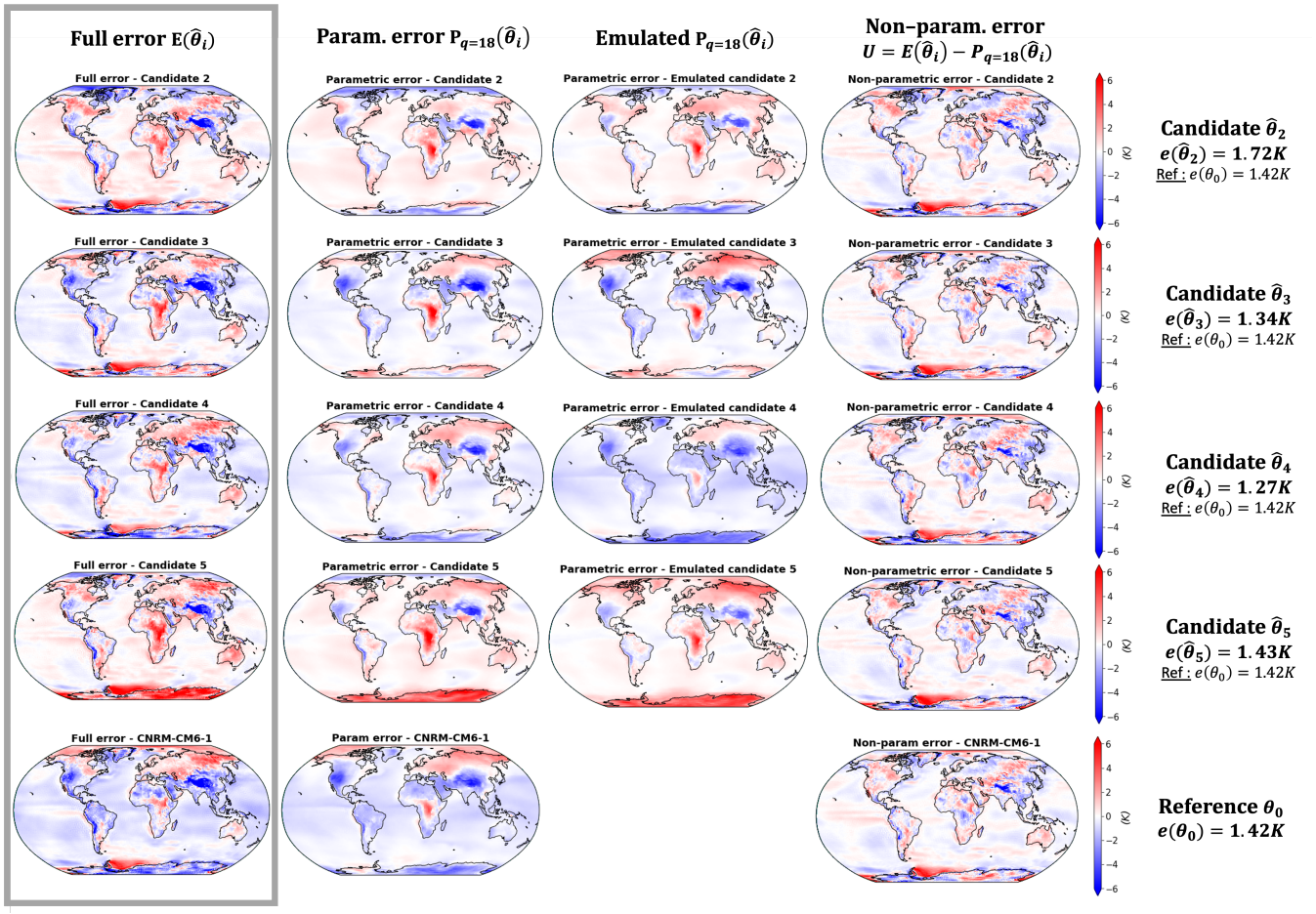

Figure 4. Differences between the simulations of temperature and the observations BEST (Rohde and Hausfather, 2020), for the 4 model candidates and the reference. Decomposition of model errors in parametric and non-parametric components using the methodology described in the Section 2, with an EOF basis 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.

biases vary from a model to another, with a particularly strong positive bias over north America in candidate 3 and a strong negative bias over central Africa in candidate 5, 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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 optimal 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

4.1 Variables, EOF analysis and truncations

The uni-variate analysis conducted in Section 3 illustrates quantitatively the potential for trade-offs and multiple optimal solutions of the climate model optimization problem. It is based on the minimization of a single uni-variate metric, allowing to select 5 optimal candidates maximizing the diversity of spatial error patterns and trade-offs among the different EOF modes.

Field	Symbol	Units	Citation	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.

In an operational GCM tuning application, the minimized metric must encompass multiple climate fields and the optimization results in trade-offs between different uni-variate metrics, with 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 weighting, 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 the annual means of the different climatic fields and the EOF truncation choices depends on the parametric emulation skill and 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

Based on this result, it is clear that, as for the uni-variate application, the optimization should only retain the first few modes. 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. 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 should be considered for the sea level pressure in order to keep satisfying statistical predictions. Moreover, some variables require more EOF modes than other in order to explain most of the ensemble variance. e.g. 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 4 EOF modes, as suggested by the correlations between reconstructed and full errors (Figure 6). Considering these 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 modes are retained for the reconstruction but is never equal to 0 (even for a full reconstruction of $q = 102$), 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 correlations between the reconstructions and the actual full model errors among the PPE, with an offset due to the non-parametric 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.

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 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 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

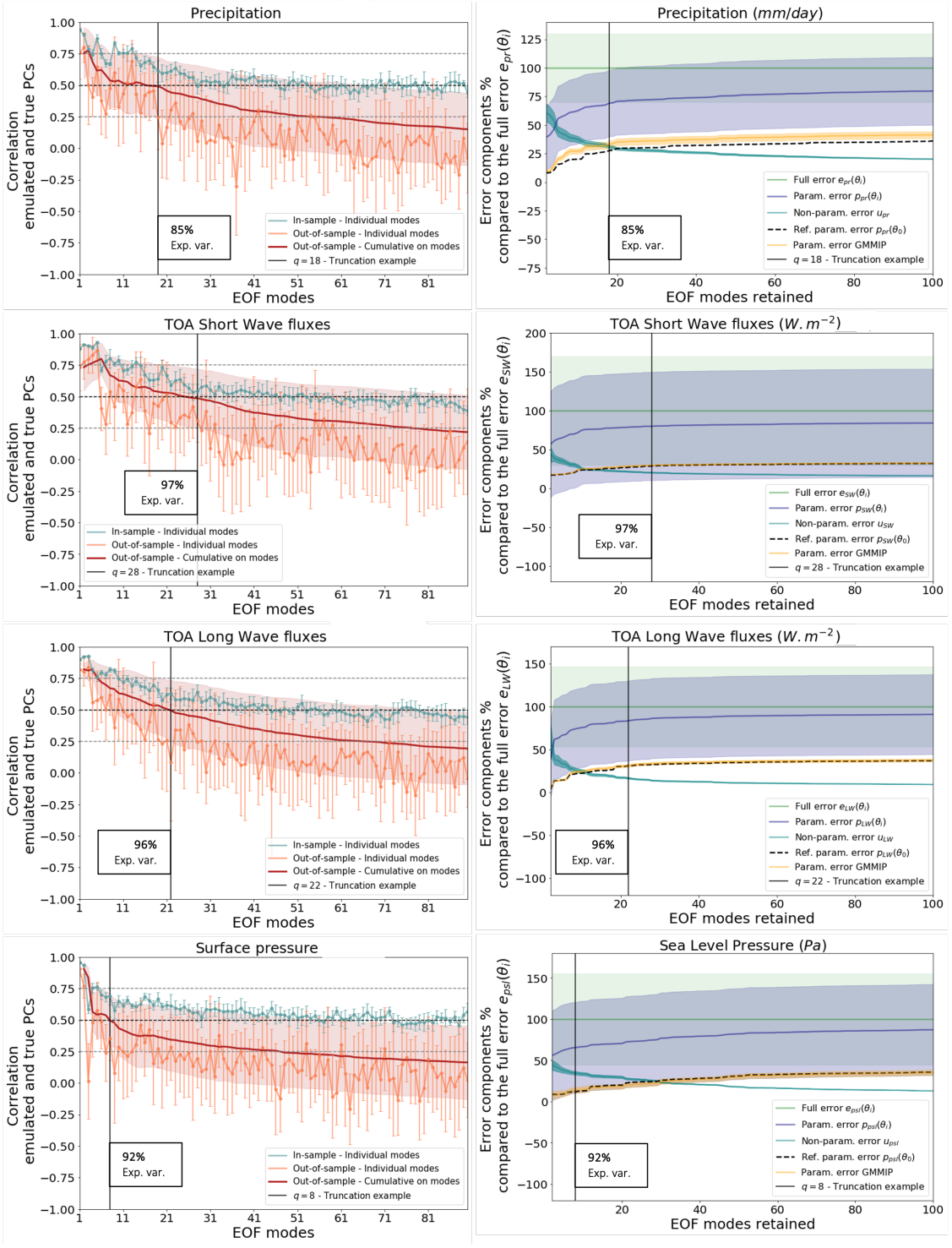

Figure 5. Truncation choice based on parametric emulation and error decomposition for 5 climatic fields : surface temperature, precipitations, TOA SW fluxes, TOA LW fluxes and surface pressure. Same legend as Figure 2, the observations used are listed in Table 1.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

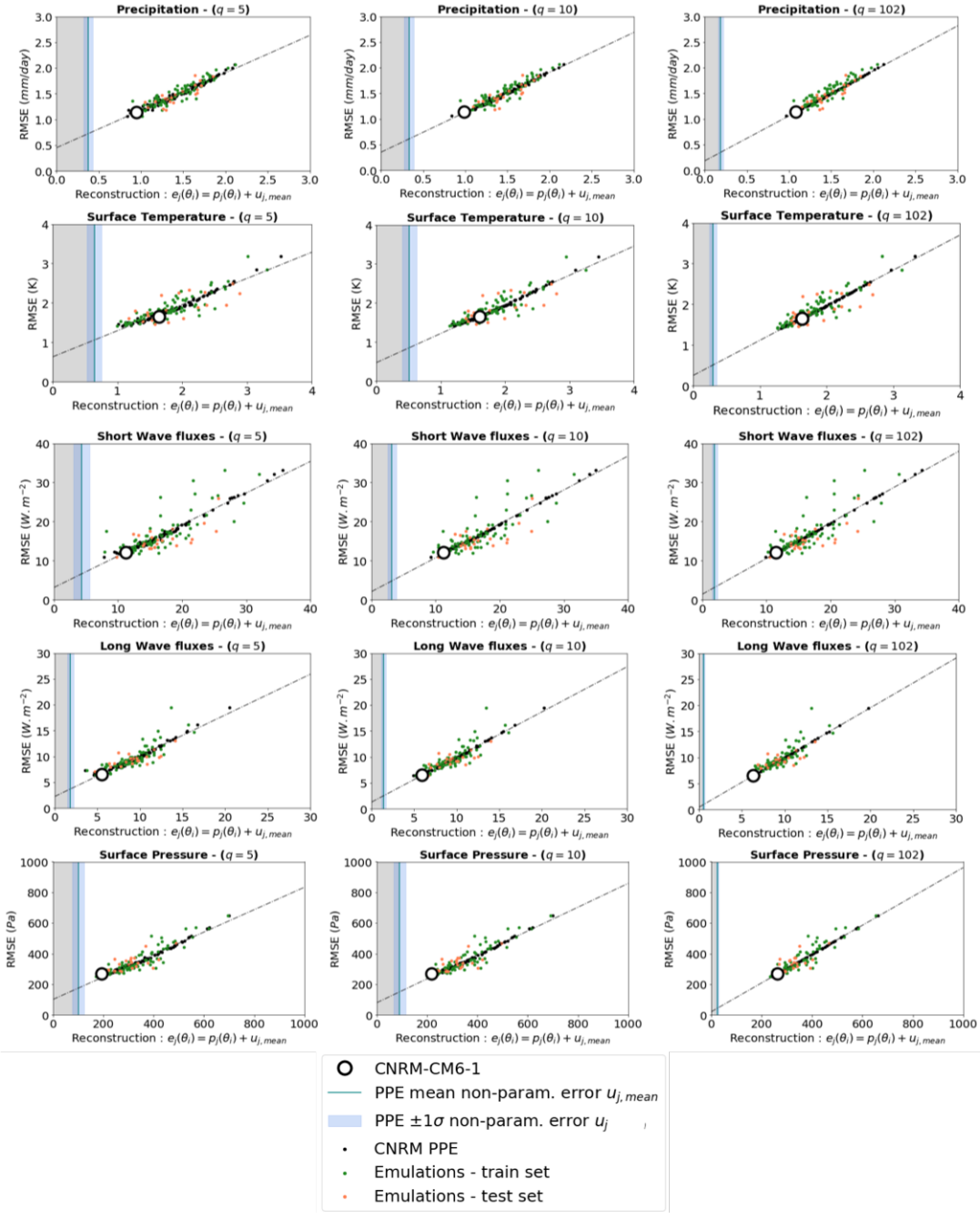

Figure 6. Correlations between full errors (y axis) and EOF-based reconstructions of these errors (x axis) using different truncation examples : retaining 5 (left column), 10 (center column) and 102 EOF modes (right column). Results are presented for the CNRM PPE (black dots) and for statistical predictions 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).

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

4.2 Candidate selection in a multi-variate context

The method to select optimal but diverse model candidates is similar to what is described in Section 2.5. However, we will now minimize a multi-variate metric 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 :

$$p_{tot}(\hat{\theta}_i) = \frac{1}{n_j} \Sigma_{j=1}^{n_j} \frac{p_j(\hat{\theta}_i)}{p_j(\theta_0)} \tag{17}$$

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

$$p_{tot}(\hat{\theta}_i) < 1 \tag{18}$$

The model candidates should reflect the model error diversity among both the different climatic fields and the different EOF modes of each field. So we will consider a selection based on an the average of the inter-point distances $d_j(\hat{\theta}_i, \hat{\theta}_k)$, normalized by the PPE mean inter-point distance $\overline{d_j}$.

$$d(\hat{\theta}_i, \hat{\theta}_k) = \sum_{j=1}^{n_j} \frac{d_j(\hat{\theta}_i, \hat{\theta}_k)}{\overline{d_j}} \tag{19}$$

As for the uni-variate application, an iterative process will identify the pair of emulations with the smallest inter-point distance and randomly remove one of the calibration parameter sets, until a chosen value of n_k optimal candidates remain in the set. We propose here an application considering the 5 chosen climatic variables ($n_j = 5$, Table 1) and retaining 4 optimal candidates ($n_k = 4$).

The results in terms of integrated multi-variate skill scores $e_{tot}(\theta_i)$ are presented in the Figure 7. Among the 4 selected candidates, only one shows a multi-variate skill-score lower than the CNRM reference model. However, all of them have a lower error than the PPE mean and 3 of them are in the low tail of the PPE distribution. Moreover, most of the CMIP6 models have undergone a tuning process and are considered to represent the control 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 grid before computing the error. It appears that 3 CNRM candidates selected here have a lower error than the mean of the 40 CMIP6 models. These 3 CNRM candidates are part of the interval of plus or minus one standard deviation of the CMIP6 error centered around the error of the CNRM reference model, indicating that they can be considered "as good as" the CNRM reference model given the tolerance considered here. The fourth candidate is above this interval, but is still very close to the CMIP6 ensemble mean and better performing than several CMIP6 models.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

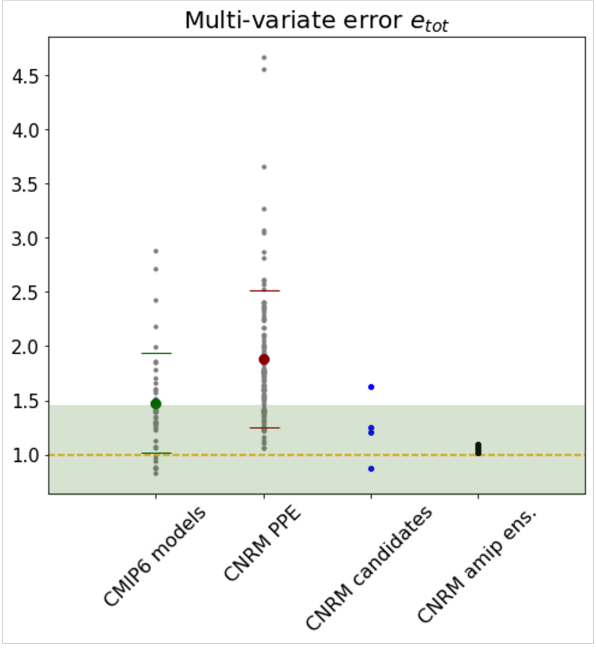

Figure 7. Multi-variate error e_{tot} for the available CMIP6 models, the CNRM PPE members, the selection of 4 diverse CNRM candidates and the 10 members of CNRM reference model with different initial conditions. Each small dots correspond to a model, the bigger dots correpond to the ensemble means and the dashes are the standard deviations. Following equation 17, each RMSE e_j has been normalized by the CNRM reference model error and average across the 5 variables considered (Table 1), therefore e_{tot} has no units and the orange dasehd line at 1.0 represents the CNRM reference model error. The green area indicates the interval of plus or minus one standard deviation of the CMIP6 errors, centered arounr the CNRM reference model error. Both the observations and the CMIP6 models have been interpolated onto the CNRM model grid before computation.

The Figure 7 is also presenting the multi-variate error among an ensemble of 10 CNRM reference model simulations, starting at different initial conditions. The 4 perturbed parameter candidates are much more diverse in terms of integrated model error that the 10 perturbed initial conditions members. When considering a multi-variate score, it is clear that the effect of internal variability is very small compared to the effect of varying the model parameters.

4.3 Diversity of error patterns among candidates

As described in Figure 7, the 4 CNRM candidates present a satisfactory multi-variate error compared to the CMIP6 ensemble, with 3 of them performing comparably to the CNRM reference model, while showing a significant diversity compared to the CNRM internal variability. We are now interested to see how this diversity translates in terms of spatial patterns of the univariate errors and trade-offs among the variables.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

The Figure 8 presents the full pattern errors of the 4 CNRM candidates for all of the 5 variables considered for the selection. We can see a large diversity of patterns among the candidates. Even between the Candidates 2 and 3, the closest in terms of integrated multi-variate score, some major differences can be observed in between the uni-variate error patterns. Both 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, some important differences exist between them. Candidate 2 proposes a better representation of the precipitation and the surface pressure (even better than the CNRM reference model). However, additional biases appear in the representation of tropical outgoing fluxes : positive in the SW and negative in the LW. The simulation of tropical clouds in Candidate 2 seems biased, even though it apparently improved tropical precipitation. On the other hand, Candidate 3 shows a less biased representation of tropical outgoing fluxes, with a SW pattern closer to the observation than the CNRM reference model. However, this is the worst candidate in terms of sea level pressure and tropical precipitation, with a candidate that accentuate the usual features of a model too wet over the ocean and too dry over the continents. The large positive LW biases over south America, central Africa and Indonesia could indicate that the model does not simulate enough clouds in these regions, which translates into dry areas on the precipitation map and a warm bias over central Africa.

Candidate 4 is the best performing candidate in terms of multi-variate score and shows errors lower than the CNRM reference model for all the variables except the sea level pressure (Figure 8). Candidate 4 shows a SW error map very similar to Candidate 3, slightly improved over the continents. But unless Candidate 3, this is associated with a really good representation of the outgoing LW pattern, that does not include strong positive biases over the tropic. We can assume that the model is better representing tropical clouds, which translates into the best representation of tropical precipitation within the selection. In contrast, Candidate 1 is the worst performing model of the selection, with too much outgoing SW radiation everywhere (Figure 8). The LW pattern is similar to Candidate 2, as well as the precipitation pattern, that is not aberrant and not the worst of the selection. Interestingly, this is also the only model that is not showing the usual negative SW biases over the mid-latitude eastern border of the oceans (except in a small area, very close to the continental borders), but they are replaced with important positive biases almost everywhere on the.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

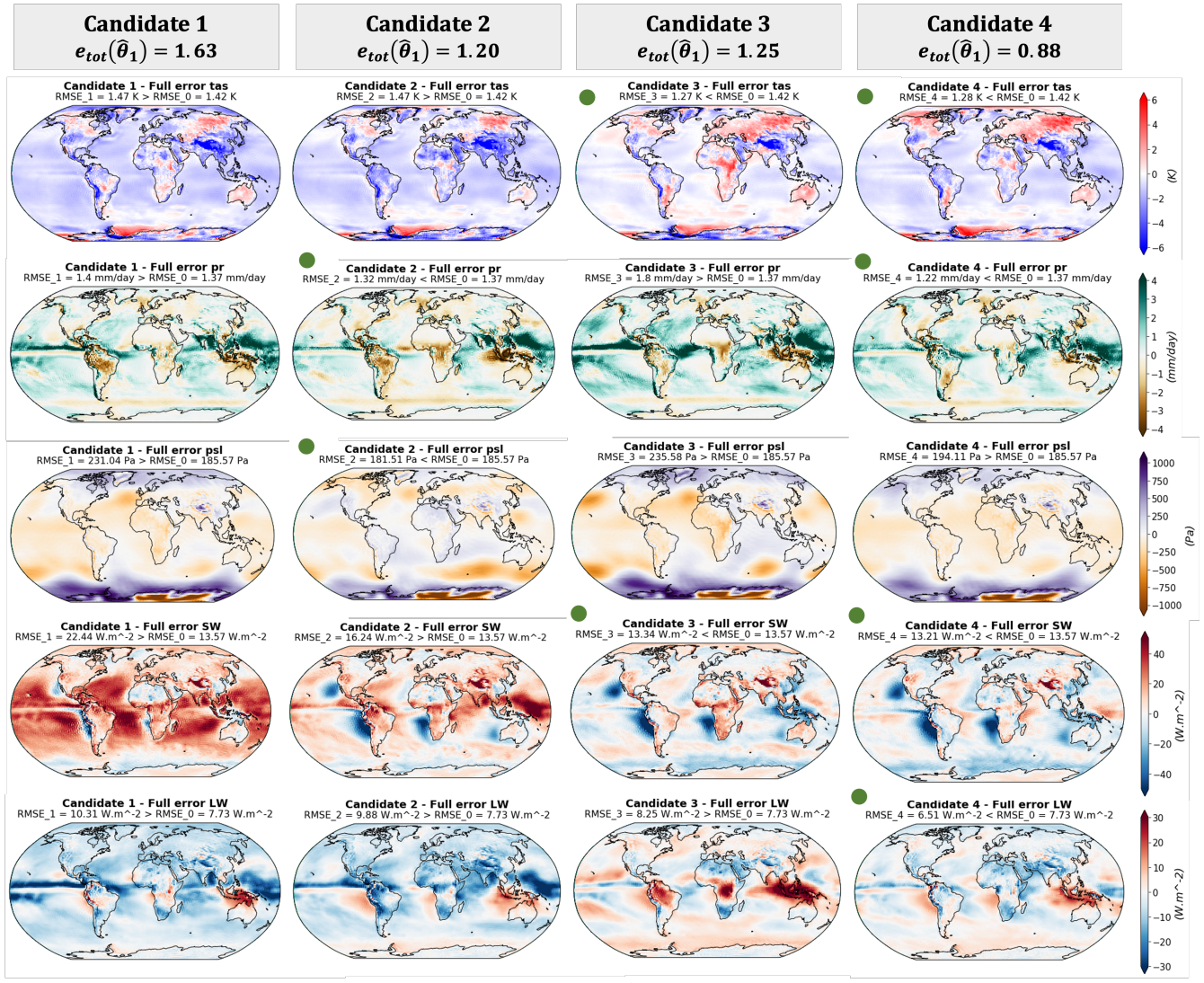

Figure 8. Differences between the simulations and the observations (Table 1), for the 4 model candidates and the 5 variables considered : surface temperature, precipitation, sea level pressure, short and long waves top-of-atmosphere fluxes. Each column represents a model candidate and each raw corresponds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

4.4 Examples of discrepancy term partitioning

Following the method described in Section 2.3, 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 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*.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

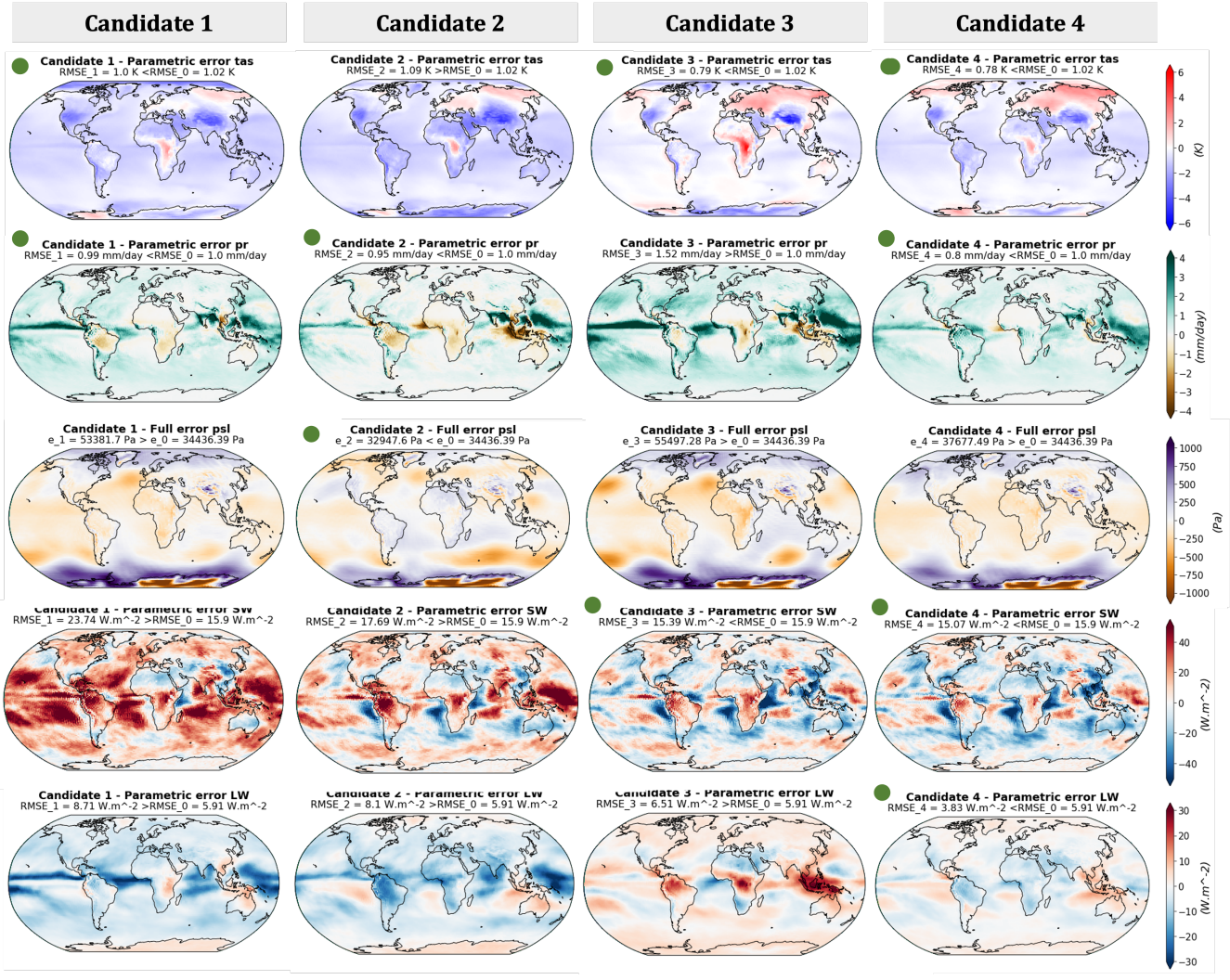

Figure 9. Parametric component of the differences between the simulations and the observations (Table 1), for the 4 model candidates and the 5 variables considered. Each column represents a model candidate and each raw corresponds to a variable. The decomposition of model errors in parametric and non-parametric components is based on the methodology described in the Section 2.3, 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*.

As expected, the candidates parametric component error patterns resemble the full error patterns, with as much diversity in between the 4 candidates (Figure 9). The non-parametric components on the other ends, 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 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

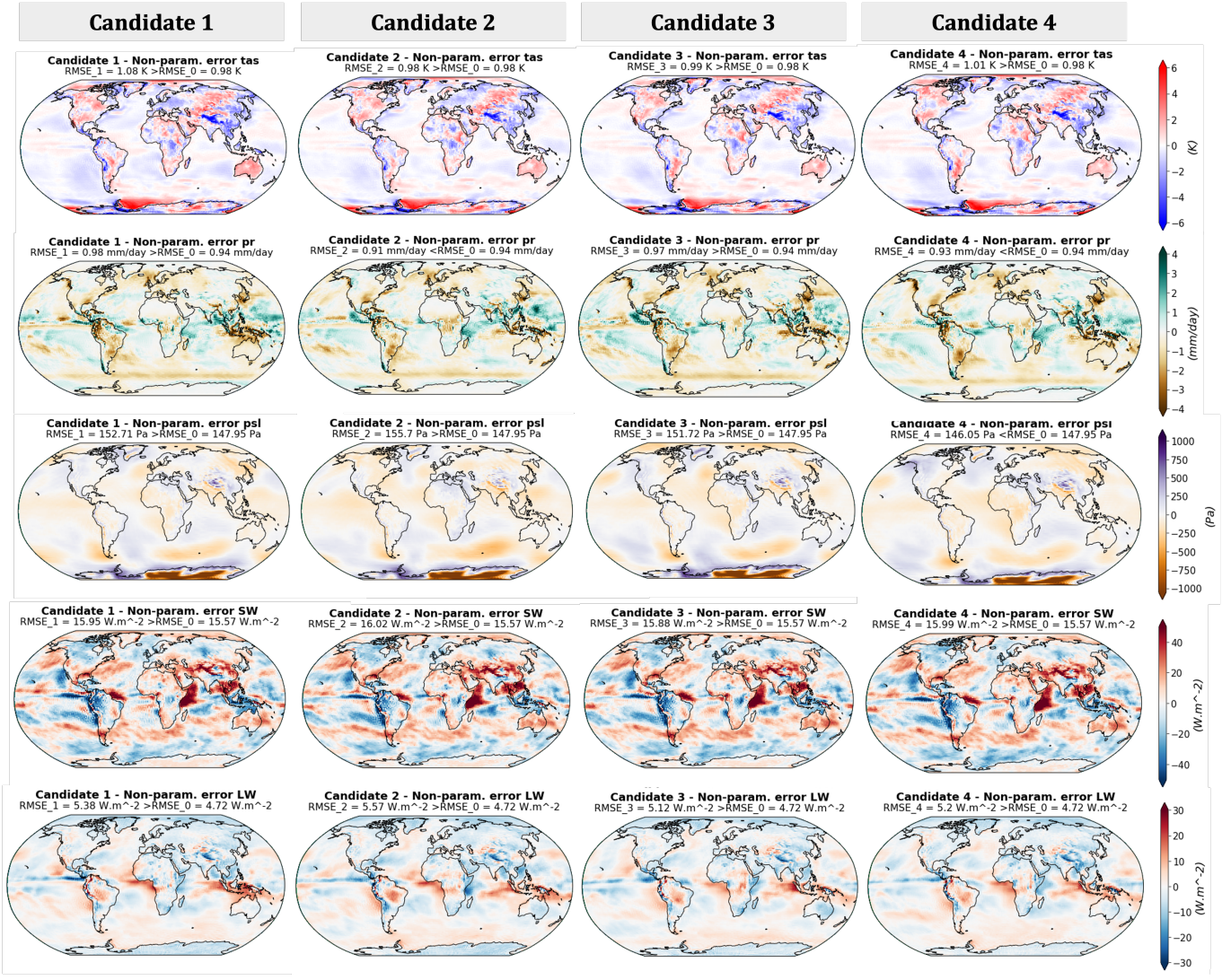

Figure 10. Non-parametric component of the differences between the simulations and the observations (Table 1), for the 4 model candidates and the 5 variables considered. These components are the differences between the Figure 8 and 9.

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 does not show clear sign of the negative biases over the oceanic mid-latitude eastern border
490 that we described in the full error patterns (Figure 10). This result suggests that such biases could be enhanced or reduced by varying the model parameters. However, important positive biases over Indian ocean and south-west Asia, less notable in the full error patterns, appear in the non-parametric component of the SW error, probably due to model inconsistencies that does not depend on the parameter choice (Figure 10). These non-parametric biases are compensated by negative biases over the same regions in the parametric component (Figure 9), suggesting that the model tuning tends to reduce these initial model inconsis-

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

tencies. But simultaneously, the correction of the Indian ocean and south-west Asia biases are associated with the emergence of negative biases over the oceanic mid-latitude eastern border in the SW parametric component (Figure 9). Looking back at the SW full errors (Figure 8), the Indian ocean and south-west Asia non-parametric positive biases are almost fully corrected for Candidate 3 and 4 (the best candidates in terms of SW error), but are still visible on Candidate 2, which is also showing slightly less important oceanic eastern border biases than Candidate 3 and 4. We can assume that these oceanic eastern border negative biases are sensitive to the parameter variability, but allow for the compensation of non-parametric biases elsewhere (Indian ocean and south-west Asia) and appear when it comes to optimizing an integrated metric because the model faces the impossibility of correcting all of these SW biases simultaneously. Candidate 1 proposes an alternative trade-off in which the oceanic eastern border biases are almost not visible, but ends up with important positive biases almost everywhere on the map, which is probably a less satisfactorily option.

In conclusion, when considering error patterns and multi-variate illustration, the effective degrees of freedom in model 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.

5 Conclusions

This study presented a new framework, based on a perturbed physics ensemble of a CMIP6 General Circulation Model, allowing for the empirical selection of diverse optimal candidate calibrations. We have demonstrated that this approach is practically useful for a number of reasons which we illustrate in a case study with an operational GCM (General Circulation Model) :

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, parameter configuration by hand is more tractable than often assumed, and the reference version may often be near-optimal in terms of integrated performance metrics
4. However there remains potential for comparably performing *near-optimal* parameter configurations making different model trade-offs

Using the 'best input' assumption (Rougier, 2007) we assume that these optimal candidates sample the distribution of atmospheric model discrepancy term. These discrepancy term examples can be partitioned in parametric and non-parametric components using PPE derived EOF basis. Optimal candidates are selected from a PPE of the ARPEGE-Climat model, the atmospheric component of CNRM-CM. The optimization is based on multi linear predictions from a 10^6 LH sampling of the

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

perturbed parameters of the parametric components of the model errors. The candidates are considered optimal when their parametric components are lower than the reference parametric component and are selected to exhibit errors as diverse as possible within this optimal space. As such, the sub-set of optimal candidates offer a diversity of model errors sampling the CNRM discrepancy term distribution while exploring different trade-offs.

The partitioning 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. (1) 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. and (2) 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.

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 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.

In the second step, the framework was applied in a multi-variate context. In this case, three candidates achieved integrated multi-variate scores within CMIP6 ensemble standards, with one performing slightly better than the reference model. Trade-offs were observed in error patterns across climatic fields, with different candidates excelling in various aspects.

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 opimization 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 near-optimal 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 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

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

diversity of model errors can reflect a diversity of future climate responses (Hourdin et al., 2023; Peatier et al., 2022)) and selecting diverse candidates will help the quantification of uncertainty in climate change impact studies.

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 and BS 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 developing and supporting the CNRM-CM6-1 model.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

References

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.

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.

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.

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, 113, 2008.

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.

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.

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.

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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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.

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.

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, 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, 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.

https://doi.org/10.5194/egusphere-2023-2269
Preprint. Discussion started: 16 October 2023
© Author(s) 2023. CC BY 4.0 License.

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.

Wilkinson, R. D.: Bayesian calibration of expensive multivariate computer experiments, Large-Scale Inverse Problems and Quantification of Uncertainty, pp. 195–215, 2010.

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.