Surrogate modelling based History Matching for an Earth system model of intermediate complexity

Maya Janvier
Centrale Supélec
Université Paris-Saclay
Gif-sur-Yvette, 91190 France
maya.janvier@student-cs.fr

Julie Deshayes
LOCEAN-IPSL, Sorbonne Université-
CNRS-IRD-MNHN
Paris, France

Redouane Lguensat
IPSL, IRD
Paris, France

Aurélien Quiquet, Didier M. Roche
LSCE-IPSL, CEA-CNRS-UVSQ,
Université Paris-Saclay,
Gif-sur-Yvette, France

V. Balaji
Schmidt Futures, IPSL

Abstract

Climate General Circulation Models (GCMs) constitute the primary tools for climate projections that inform IPCC Assessment Reports. Calibrating, or tuning the parameters of the models can significantly improve their predictions, thus their scientific and societal impacts. Unfortunately, traditional tuning techniques remain time-consuming and computationally costly, even at coarse resolution. A specific challenge for the tuning of climate models lies in the tuning of both fast and slow climatic features: while atmospheric processes adjust on hourly to weekly timescales, vegetation or ocean dynamics drive mechanisms of variability at decadal to millenial timescales. In this work, we explore whether and how History Matching, which uses machine learning based emulators to accelerate and automate the tuning process, is relevant for tuning climate models with multiple timescales. To facilitate this exploration, we work with a climate model of intermediate complexity, yet test experimental tuning protocols that can be directly applied to more complex GCMs to reduce uncertainty in climate projections.

1 Introduction

General Circulation Models (hereafter GCMs), widely used to produce simulations of the past, present and future climate, which subsequently inform the IPCC [1,2], include parameters that are more or less constrained by theory or observations (for example snow albedo). Some of these parameters are introduced as the physical equations are discretized to enable numerical computation, for example related to unresolved or ill-resolved processes which effects are parameterized (for example horizontal diffusion in the ocean). Even when direct measurements of these parameters are available, they may not be optimal for a given model as the latter remains one imperfect mathematical representation of the Earth system.

Calibration of climate model parameters, aka *tuning*, is driven by two motivations. First, it allows increased performances on specific prediction tasks: it is the Fitness-for-Purpose paradigm [2]. Climate model simulations will gain importance in the years to come with the need for accurate predictions of climate change to guide adaptation, which constitute *climate services* [3]. Secondly, tuning the models can also help quantifying their uncertainties [2], in the process of improving Earth modelling as well as understanding the causes of climate change to inform mitigation policies.

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2023.

GCMs are state-of-the-art in climate modelling: the 2021 Nobel prizes in Physics even honored some of the pioneering work done to support the development of GCMs [4]. However they come with a computational price: running a GCM for a hundred years prediction can take up to several weeks on the top High-Performance Computing centers. This makes traditional tuning-by-hand techniques too costly. Extensive grid search [5] techniques are also near impossible in practice which call for more intelligent techniques based on surrogate modeling. History Matching (HM) is a well established technique [6,7] that can improve and accelerate this tuning, and that recently attracted attention in the climate modeling community. HM makes use of an emulator to replace the climate model in the exploration of the parameters' space. For example, Lguensat et al. [8] applied HM to tune a classical toy model, Lorenz–96, coupling slow and fast variables as a simple analogy to an ocean-atmosphere model, they highlight the challenges induced by the multiple timescales. In this work, we consider iLOVECLIM [9], an Earth system Model of Intermediate Complexity (EMIC), which presents way more complex physical processes than a simple toy model. Following Loutre et al. [10], Shi et al. [11], we explore the parameters' sensitivity of iLOVECLIM for present-day climate simulations, we then perform a HM-based tuning of this model and present our conclusions.

2 Materials and Methods

Model, choice of parameters and ground truth We consider the iLOVECLIM Earth system model, derived from the LOVECLIM model by Goosse et al. [9]. It consists of several coupled components: ocean and sea ice (CLIO) with a 3° × 3° resolution, atmosphere (ECBilt) with a 5.6° × 5.6° resolution and a module for vegetation (VECODE). iLOVECLIM closes the carbon cycle through these three components. For our general tuning setup, we take nine land, ocean or atmosphere related parameters to calibrate with the HM algorithm. Based on [10, 11] in addition to our experts' intuition, we set for each parameter an a priori range interval that constitutes the search space (Table 4 in A.4). We select 15 yearly atmospheric and 9 monthly oceanic variables from which we derive mean metrics for further use and analysis (Table 5 in A.4).

Prior to launching experiments, and because iLOVECLIM simulations are distant from climate observations, we consider a reference simulation from iLOVECLIM to use as our ground truth, and perform a 5000-yr-long stationary simulation reflecting present-day climate. All of the simulations then start from this stabilized state of iLOVECLIM, obtained with default parameters. We place our study in a "perfect model" setting as in [3]. Knowing the ground truth parameters allows us to evaluate better the performances of HM on iLOVECLIM.

Reducing the parameters' space with HM HM uses reference data (our ground truth here) to rule-out any parameter settings which are implausible, because they are expected to produce simulations that are too different from the reference. Using few runs of iLOVECLIM simulations, HM trains a Gaussian Process based emulator (RBF kernel, see [8]) that can interpolate the search space at lower cost. An implausibility score is used to rule out implausible portions of the search space. The reduced search space, also called the Not Ruled Out Yet space or NROY, is the result of each HM iteration or wave. Figure 1 summarizes this process.

The implausibility score (see A.1) is then based on metrics of the model: here we select time average atmospheric and oceanic metrics (see previous section). Following Lguensat et al. [8], a major difference in our application of HM from D. B. Williamson et al. [12] and Hourdin, Williamson, et al. [13] relies on the use of Principal Component Analysis to reduce the dimensionality of the metrics vector: the Gaussian Process is fitted on these reduced metrics. As we want to tune 9 parameters, we run about 10 simulations per parameter, i.e 90 simulations in each wave. We stop iterating when the NROY is sufficiently reduced (e.g. less than 0.5% of the initial space).

Selection of candidates and quality assessment At the end of the HM algorithm, we are left with a final NROY. If the HM procedure leads to only few candidates, we select them all. Otherwise, an option could be to use a k-means algorithm on the last NROY, as in [8]. The optimal number of clusters is determined via the silhouette score. We then make sure the k-means centroids are in the NROY before using them for evaluation. To assess the quality of the candidates, we compare how close the variables' distributions from their corresponding iLOVECLIM simulations (in Table 5) are to the reference, using the Kullback–Leibler divergence (KL-div) following [8]: the lower the score, the closer the distribution is to the reference.

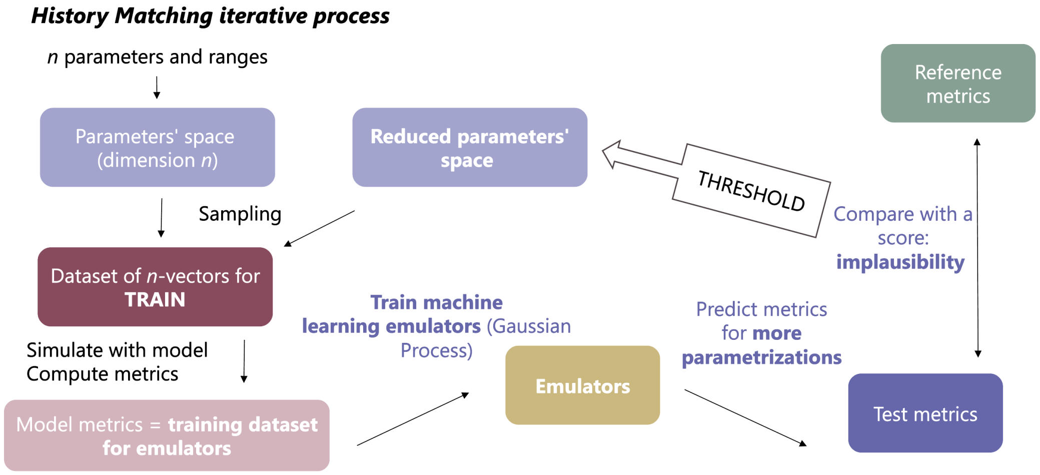

History Matching iterative process
n parameters and ranges
Parameters' space (dimension n)
Sampling
Dataset of n-vectors for TRAIN
Simulate with model
Compute metrics
Model metrics = training dataset for emulators
Train machine learning emulators (Gaussian Process)
Emulators
Predict metrics for more parametrizations
Test metrics
Reference metrics
Compare with a score: implausibility
THRESHOLD
Reduced parameters' space

Figure 1: Schematic of the iterative process in History Matching tuning procedure.

3 Experiments and Results

Sensitivity analysis In a sensitivity analysis of iLOVECLIM by Shi et al. [11], they concluded that all of the ocean parameters were only a little sensitive in the time scale of thousands of yr. This surprisingly disagrees with the role played by the ocean in climate variability at centennial to millennial time scales, which has earned it the title of the *slow* component of climate.

We investigated the sensitivity of the parameters when using only atmospheric metrics (Shi setup) and in coupled mode when adding the oceanic metrics, with 5 and 20-yr means. To assess the level of sensitivity of a parameter, we considered the following characteristics of the 2D-NROYs panels (Figure 3 in A.2): the size (the smaller the NROY for a given parameter, the more sensitive iLOVECLIM is to this parameter), the noisiness (the more discretized the NROY, the less sensitive a parameter, sort of connectedness), the optical depth or specificity (the higher the optical depth, the more plausible parametrizations there is in a given area).

Table 1: Qualitative analysis of the sensitivity of iLOVECLIM parameters, as compared to Shi et al. [4] and our own experiments.

Parameter	Domain	Shi's sensitivity	Shi setup's sensitivity	Test setup's sensitivity
ampwir	atm	Very sensitive	Sensitive	Sensitive
expir	atm	Very sensitive	Sensitive	Sensitive
relhmax	atm	Very sensitive	Very sensitive	Highly sensitive
cwdrag	atm	Sensitive	Sensitive	Very sensitive
alphd	land	Sensitive	Sensitive	Sensitive
cgren	land	Not very sensitive	Not very sensitive	Sensitive
ai	ocean	Not very sensitive	Not very sensitive	Very sensitive
aitd	ocean	Not very sensitive	Not very sensitive	Sensitive
avkb	ocean	Not tested	Very sensitive	Highly sensitive

We obtained similar sensitivity results for [Shi setup] as for the original study. This analysis reveals that oceanic metrics are crucial to tune not only the oceanic parameters but can help improve tuning of land and atmospheric parameters as well (Table 1), by reducing more efficiently the parameters' space than with atmospheric metrics only, as [11] suggested. We also highlight that the temporality of the metrics is important for the tuning, as computing 5-yr-means is the best when considering atmospheric metrics alone, while computing 20-yr-means seems optimal when considering atmospheric and oceanic metrics (Figure 3 in A.2).

Tuning of iLOVECLIM But are 20-yr-long means enough to tune oceanic parameters? Following the steps described in Section 2, we performed HM with mean metrics over 20 and 100-yr-long simulations. We ended up with 3 candidates after wave 2 for 20-yr-means (M20d), and 7 after wave 1 with k-means selection (M20k). We select 6 candidates with k-means for 100-yr-means (M100k). For consistency, we compute the KL-divergence using 100-yr-simulation for all the setups. It also ensures the KL-divergence is defined for all metrics, as we only have yrly points for atmospheric data. Table 2 gathers the characteristics as well as the performances of these three experiments. For more details on the individual KL-div medians of the candidates and the NROYs, see Table 3 and Figure 4 in the appendix.

Experiment	**M100k**	**M20k**	**M20d**
Mean computed over	100 yr	20 yr	20 yr
Candidate selection method	k-means	k-means	direct
Selection performed after wave number	3	1	2
NROY size after step 2	0.0044%	0.0307%	0.0003%
Optimal number of clusters $>$ 2 for k-means	6	7	3
Number of candidates in NROY	6/6	6/7	3/3
Number of candidates in NROY with finite KL-div	6/6	3/6	3/3
Best candidate KL-div median	0.1467	0.1153	0.1910
Mean of 3 best candidates KL-div median	0.1746	0.2053	0.2332

Table 2: Summary of the three tuning experiments.

Overall, the experiments produce candidates with similar KL-div median. As no strategy stands out from an individual candidate perspective, it seems best to consider ensemble modelling techniques: instead of searching for the best candidate, we evaluate strategies on several candidates. This is a very common technique, used in the machine learning community that helps improve the prediction performances and also quantify uncertainties, in the spirit of the IPCC Coupled Model Intercomparison Project or the Perturbed Parameter Ensemble technique (PPE) [2]. Also, k-means have been proven efficient to select parameters' sets for iLOVECLIM. Looking at 2, using 100-yr-means provides more reliable candidates than 20-yr-means for ensembling, but the latter can also find better individual candidates (best performance $M20_{k6}$, see 3). The 20-yr-mean failing on KL-div test (M20k) reveals that some low frequency behaviour is missed as compared to 100-yr-means: perhaps the NROY was not reduced enough. However this strategy is more discriminant in the first waves (faster reduction of the NROY for M20d) which could be interesting in the acceleration of the tuning.

Acceleration of the tuning What do we lose exactly with 20-yr-long simulations? The NROY after wave 1 for 100-yr-means has 1.8488% of the original space left, against only 0.0307 % with 20-yr-means. It may indicate that if a simulation is too far from the ground truth in the beginning, it is safe to rule it out immediately. We then compare M20 and M100 on the simulations obtained from parameters' sets still considered plausible after the first wave (i.e in the first NROY), versus simulations from which the parameters are not in any of the experiments' first NROYs considered (blue lines, see Figure 2). This experiment is really close to PPE [2], as we perturb the initial system of one model by changing the parameters' set.

We observe again that the simulations in the first NROY are better constrained, more centered around the truth and less dispersed than the simulations outside of the NROYs (blue lines). Oceanic variables present more diversity in their convergence patterns, such as a *compensation effect* in AABex: the mean metrics on 100 yr is not adapted here as it compensates the extreme values, allowing the selection of oscillating series around the ground truth, when 20 yr directly select steadier series. We observe a *short-term drift* in Fc30A: the M100 series start to converge to the truth only after 20 yr, here 100-yr-means are more interesting than 20-yr-means. Finally, S1mo shows overfitting behaviour for both M20 and M100 as some simulations start diverging a lot from the truth after 50 yr of simulation. Atmospheric variables converge quickly. They are pretty sensitive to the initial perturbation, making M20 strategy more discriminant, even if a few simulations can still diverge in the long term (pp panel in Figure 2). These observations make both short (first years) and long-term metrics significant for the NROY reduction, and should be considered when designing a HM tuning strategy.

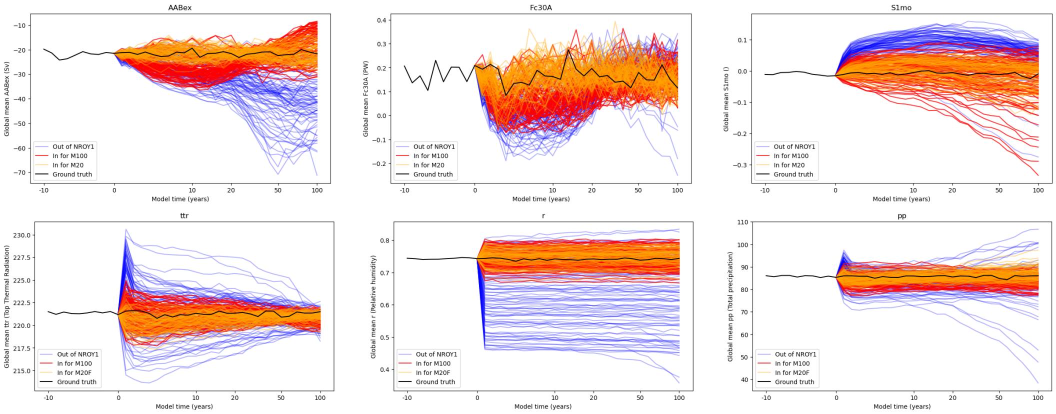


Figure 2: Evolution of mean metrics over time for M20 (orange) and M100 (red). The year 0 is the start of the simulations, before we plot the ground truth only. The time scale is non-linear to represent more precisely the convergence patterns in the first years of simulation. The first row displays oceanic variables , the second atmospheric ones.

4 Conclusions and Future work

In this work we showed that tuning simultaneously the atmospheric and oceanic parameters is advantageous, as adding oceanic metrics improve the constraints on parameters from land, atmosphere and ocean domains at the same time. Comparing metrics calculated over 20 yr and 100 yr long simulations for the tuning, we found the M100 strategy safer from an ensemble perspective, but M20 strategies remain promising for the acceleration of the tuning in the first waves. We strongly recommend to take these results into account when designing semi-automatic tuning protocols for more complex GCMs, although it remains to move from this *perfect model setup* towards tuning against real observations, which is not straightforward in a model of intermediate complexity.

Gaussian Process based History Matching is on its way to become a reference tool for tuning climate models, although many details in the implementation into the tuning procedure of climate models are left for exploration, such as the tuning of parameters that induce anomalies at different time scales. Applying the method to a climate model of intermediate complexity, illustrates a few caveats that may emerge when tuning more complex GCMs against observations. We hope that advances in machine learning based surrogate modeling can benefit to HM and pave the way for a faster and less costly tuning of GCMs, as compared to current practice [14], in order to reduce uncertainties in their climate predictions and related societal impacts.

References

[1] IPCC Working Group 1 Second Assessment Report (1996)

[2] Chen, D. & M. Rojas & B.H. Samset & K. Cobb & A. Diongue Niang & P. Edwards & S. Emori & S.H. Faria & E. Hawkins & P. Hope & P. Huybrechts & M. Meinshausen & S.K. Mustafa & G.-K. Plattner & A.-M. Tréguier (2021). Framing, Context, and Methods. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* pp. 147–286

[3] World Meteorological Organization (2013) What Do We Mean by Climate Services ? Bulletin n° : Vol 62 (Special Issue)

[4] The Nobel Committee for Physics, *Scientific Background on the Nobel Prize in Physics* 2021

[5] LaValle S. M. & Branicky, M. S. & Lindemann, S. R. (2004). On the relationship between classical grid search and probabilistic roadmaps. *The International Journal of Robotics Research*, 23(7–8), 673–692.

[6] Craig, P. S. & Goldstein, M. & Seheult, A. H. & Smith, J. A. (1997). Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. *Case studies in bayesian statistics*, pp. 37–93

[7] Sacks, J. & Welch, W. J. & Mitchell, T. J. & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical science 4*, pp. 409–423.

[8] Lguensat, R. & Deshayes, J. & Durand, H & Balaji, V (2023) Semi-Automatic Tuning of Coupled Climate Models With Multiple Intrinsic Timescales: Lessons Learned From the Lorenz96 Model, *Journal of Advances in Modeling Earth Systems, 15*, e2022MS003367.

[9] Goosse, H. & Driesschaert, E. & Fichefet, T. & Loutre, M.-F. (2007) Information on the early Holocene climate constrains the summer sea ice projections for the 21st century. *Climate of the Past 3*, pp. 609–616.

[10] Loutre, M. F. & Mouchet, A. & Fichefet, T. & Goosse, H. & Goelzer, H. & Huybrechts, P. (2011) Evaluating climate model performance with various parameter sets using observations over the recent past. *Climate of the Past 7*, pp. 511–526.

[11] Shi, Y & Gong, W & Duan, Q & Charles, J & Xiao, C & Wang, H (2019) How parameter specification of an Earth system model of intermediate complexity influences its climate simulations. *Progress in Earth and Planetary Science 6*, pp. 46.

[12] Williamson, D. B. & Blaker, A. T., & Sinha, B. (2017). Tuning without over-tuning: parametric uncertainty quantification for the nemo ocean model. *Geoscientific 983 Model Development, 10 (4)*, 1789–1816.

[13] Hourdin, F. & Williamson, D. & Rio, C. & Couvreux, F. & Roehrig, R. & Villefranque, N. & Musat, I. & Fairhead, L. & Binta Diallo, F. & Volodina, V. (2020). Process-based climate model development harnessing machine learning: II. model calibration from single column to global. *Journal of Advances in Modeling Earth Systems, 13*, e2020MS002225.

[14] Mignot, J. & Hourdin, F. & Deshayes, J. & Boucher, O. & Gastineau, G. & Musat, I. & Vancoppenolle, M. & Servonnat, J. & Caubel, A. & Chéruy, F. & Denvil, S. & Dufresne, J.-L. & Ethé, C. & Fairhead, L. & Foujols, M.-A. & Grandpeix, J.-Y. & Levavasseur, G. & Marti, O. & Menary, M. & Rio, C. & Rousset, C. & Silvy, Y. (2021). The tuning strategy of IPSL-CM6A-LR. *Journal of Advances in Modeling Earth Systems, 13*, e220MS002340.

A Appendix

A.1 Implausibility score

Following Lguensat et al. [8], we define the implausibility as a distance measure between the PCA-tranformed metrics $f(p)$ and the PCA-transformed observations of the real system z (different to the actual metrics y, due to biases in the instrumental measurement). Our emulator is minimizing $||z - f(p)||$ (we choose $||.||$ as the Mahalanobis distance). In reality, with our limited number of simulations, we only have access to the expectation $\mathbb{E}(f(p))$ and the variance $\mathbb{V}(f(p))$. We then define the implausibility as follow:

$$I(p) = ||z - \mathbb{E}(f(p))|| = (z - \mathbb{E}(f(p))^T (V_e + V_\eta + \mathbb{V}(f(p))^{-1} (z - \mathbb{E}(f(p)) \qquad (1)$$

with V_e the error variance of observations and V_η the error variance due to the simulator uncertainties (see [12]). The NROY region is $\{p : I(p) \leq T\}$, with $T = 3$ using the Pukelsheim rule.

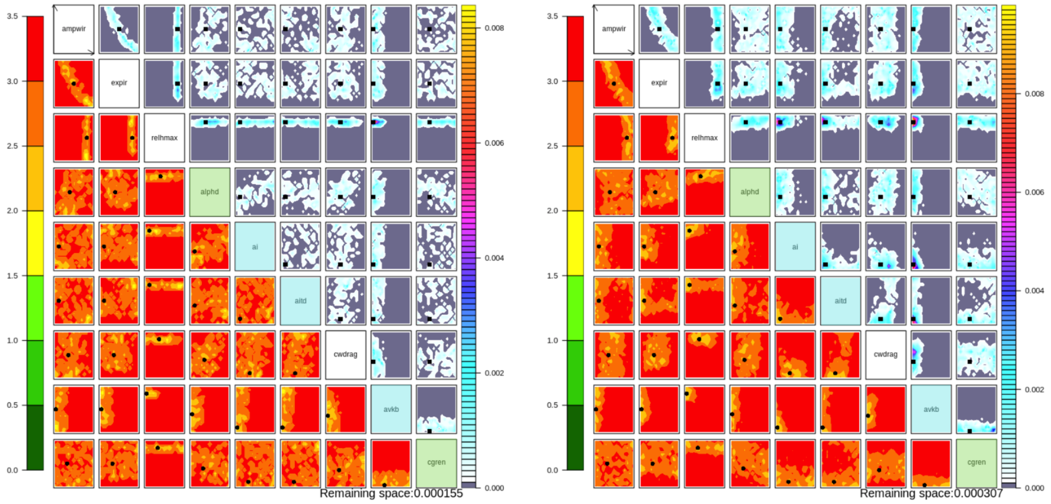
A.2 Sensitivity analysis of iLOVECLIM

Analysing time variability of a few oceanic and atmospheric metrics reveals that the reference simulation has a mainly decadal variability, so we work with **20-yr-long simulations**, and want to compare the History Matching performances with:

- **[Shi setup]**: Mean metrics computed on 15 atmospheric variables from [4]
- **[Test setup]** We add our 9 oceanic variables to the precedent setup (see Table 5).

We perform their means over the last 5, 10 and 20 yr of the simulations for both setups.

From all the experiments, the best outcomes, that is to say the most reduced NROY parameter space, were obtained with the mean metrics calculated over the last 5 yr for **[Shi setup]** (Figure 3 , left) and over the last 20 yr for **[Test setup]** (Figure 3, right). In both cases, parameter values used to produce the reference simulation are included in the final NROY.



[Shi setup] Wave 2

[Test setup] Wave 1

Figure 3: HM NROY: Using atmospheric mean metrics only, computed over the **last 5 yr (left)** and using atmospheric and ocean mean metrics, computed over the **last 20 yr (right)**. The left-hand-side color bar is the implausibility score (the greener the more plausible, subplots to the bottom left of the diagonal), the right-hand-side color bar is the optical depth (the higher the more plausible parametrizations there is in a given area), and the black point locates the parameter values used to produce the reference. The background color of the diagonal indicates the component of the climate system directly affected by the parameter : blue for ocean parameters, white for atmosphere parameters and green for land parameters.

To assess the level of sensitivity of a parameter, we consider the following characteristics of the 2D-NROYs panels:

- the **size**: The smaller the NROY for a given parameter (the larger red/grey areas in the two respective subplots), the more sensitive iLOVECLIM is to this parameter. Indeed, a large NROY means we can choose whatever value for this parameter and it will not affect much the model solution (for the metrics considered in the sensitivity assessment).
- the **noisiness**: Intuitively, the noisiness is a perturbation of our main signal, here the discretization of the NROY considered. The noisier the NROY, the less sensitive a parameter, as the algorithm cannot clearly identify a single plausible area. This characteristic can be defined more rigorously with the mathematical terminology of *connectedness* in topology. A connected space is a topological space that cannot be represented as the union of two or more disjoint non-empty open subsets. For example, the 2D-NROY is a connected space in the relhmax/alphd panel, but not in the relhmax/ai panel, in which there are several disjoint connected spaces called *connected components*. We interpret a space as "noisier" the more connected components it has, and " more compact" when it is a connected space (only one connected component) or has fewer connected components.
- **the optical depth or specificity**: The optical depth is the fraction of configurations with implausibility smaller than the predefined threshold. In other words, in a 2D-NROY, the higher the optical depth, the more plausible parametrizations there is in a given area, in the other parameters dimensions. Having high optical depth scores means the HM is specific, as it precises regions more plausible than others within the same NROY.

Following this methodology, we synthesize the sensitivity of iLOVECLIM to land, ocean and atmosphere parameters in Table 1.

Looking at Table 1, we can see that we obtained similar sensitivity results for [Shi setup] as for the original study [11]. The diminution of global sensitivity for *ampwir* and *expir* can be explained by our method to evaluate sensitivity: with History Matching we are studying the sensitivities of all the parameters taken together as a set, and not one by one. Indeed in the [Shi setup] plot (Figure 3, left), the 2D-NROYs are noisy and large in the ocean and land related panels, but more compact and reduced in the atmospheric ones. All in all, our [Shi setup] reproduces Shi et al.'s conclusions, allowing us to use it as the reference for our experiments.

When comparing the two NROYs in Figure 3, we see that the parameters' space is more reduced for the [Shi setup] (0.0155% of the initial space) than for the [Test setup] (0.0307% of the initial space), even if comparable. However, two waves were needed to obtain this results when using only atmospheric metrics, instead of one with both atmospheric and oceanic metrics.

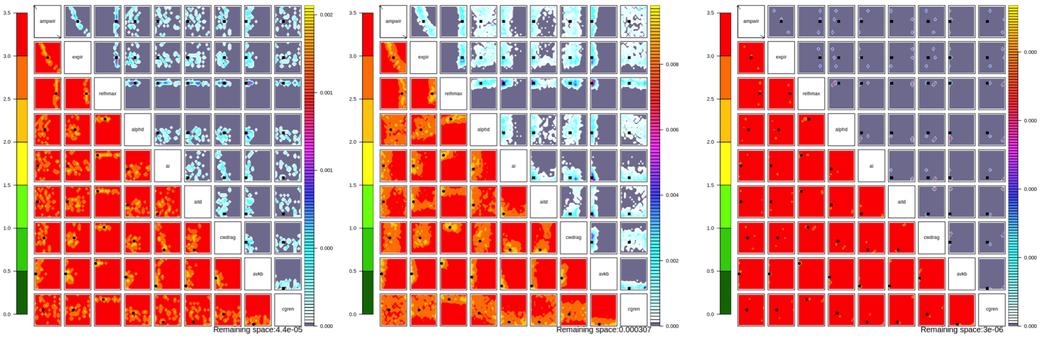
However, when looking at the [Test setup] plot (Figure 3, right), we see that adding the oceanic metrics reduces the noisiness of the NROY for all the parameters compared to the [Shi setup]. Furthermore, adding the oceanic metrics increases the sensitivities of not only oceanic parameters like *ai*, *aitd* or *avkb*, but also of the land parameter *cgren* or the atmospheric one *cwdrag*, that have less noisy and more reduced NROYs. This is a new result from [4], that did not consider oceanic metrics at all, but discussed that they may be important to the tuning of atmospheric parameters.

Furthermore, two already sensitive parameters, *relhmax* (atm) and *avkb* (ocean) are even more specified by the History Matching algorithm with atmospheric and oceanic metrics. Indeed when looking at the optical depth, the values are mostly around 0.002 in the Shi setup, except for the *relhmax/avkb* panel. In the Test setup, we have larger and numerous regions with an optical depth greater than 0.006, almost all centered in the ground truth (in *relhmax/avkb* but also in the *ai* and *aitd* planes for both *relhmax* and *avkb*). We thus have more privileged plausible regions than without oceanic metrics. This behaviour was replicated in the experiments with different temporality of metrics, confirming that it is the presence of oceanic metrics that is at stake here.

Finally, our experiments tell us about the importance of temporality of the metrics. In general, the more years are used in the computation of the means, the more reduced the global NROY is at wave 1. However at wave 2, when using atmospheric metrics only, the best outcome is with 5-yr-means. It is coherent with the "fast" variability of the atmosphere. For atmospheric and oceanic metrics, computing 20-yr-means is still optimal for both waves, which is also coherent with the "slow" variability of the ocean discussed in 3.a.

A.3 History Matching tuning experiment

Figure 4 displays the NROYs after step 2, from which we select our candidates. Table 3 gathers the individual KL-div median of all the candidates. The best candidate for each experiment is in bold, showing us that a bad method for ensemble techniques can lead to good candidates individually.



NROY 3: M100k **NROY 1: M20k** **NROY 2: M20d**

Figure 4: NROYs before candidate selection (Step 3)

Table 3: KL-divergence of candidates

Candidate ID	**Median of KL-divergence mean metrics**
$M100_{k1}$	0.1715
$M100_{k2}$	0.4446
$M100_{k3}$	0.5568
$M100_{k4}$	**0.1467**
$M100_{k5}$	0.2351
$M100_{k6}$	0.2058
$M20_{k1}$	Infinite KL-div for 3 variables
$M20_{k2}$	0.2548
$M20_{k3}$	Infinite KL-div for 2 variables
$M20_{k4}$	Infinite KL-div for 1 variable
$M20_{k5}$	0.2458
$M20_{k6}$	**0.1153**
$M20_{d1}$	0.2075
$M20_{d2}$	**0.1910**
$M20_{d3}$	0.3011

A.4 Parameters and metrics

This section provides the setup for reproducing our History Matching experiments: the parameters to tune in Table 4, the variables from which we derive our mean metrics in Table 5.

Table 4: Selection of parameters to study and their characteristics

Parameter	Module	Definition	Default	Range	Range origin	Shi's Sensitivity
ampwir	ECBilt-atm	Scaling coefficient in the longwave radiative scheme	1	[0.5,1.5]	[4]	Very sensitive
expir	ECBilt-atm	Exponent in the longwave radiative scheme	0.4	[0.2,0.6]	[4]	Very sensitive
relhmax	ECBilt-atm	Precipitation also occurs if the total precipitable water below 500hPa is above this relevant threshold	0.83	[0.5,0.9]	[4]	Very sensitive
cwdrag	ECBilt-atm	Drag coefficient to compute wind stress	2.1e-3	[1.0e-3,4.0e-3]	[4]	Sensitive
alphd	ECBilt-land	Albedo of snow	0.72	[0.6,0.9]	[4]	Sensitive
cgren	ECBilt-land	Increase in snow/ice albedo for cloudy conditions	0.04	[0.01,0.10]	[4]	Not very sensitive
ai	CLIO-ocean	Coefficient of isopycnal diffusion ($m^2 s^{-1}$)	300	[200,1000]	[4] modified	Not very sensitive
aitd	CLIO-ocean	Gent-McWilliams thickness diffusion coefficient ($m^2 s^{-1}$)	300	[200,1000]	[4] modified	Not very sensitive
avkb	CLIO-ocean	Minimum vertical diffusivity for scalars	1.5e-5	[1.0e-5, 1.0e-4]	[3] modified	Not tested

Table 5: Output variables (24) of LOVECLIM from which we derived our metrics for History Matching

Name	Domain	Definition
q	atm	Specific humidity
ts	atm	Surface temperature
bm	atm	Bottom moisture
shf	atm	Surface sensible heat flux
lhf	atm	Surface latent heat flux
r	atm	Relative humidity
alb	atm	Surface albedo
ssr	atm	Surface solar radiation
tsr	atm	Top solar radiation
str	atm	Surface thermal radiation
ttr	atm	Top thermal radiation
evap	atm	Surface evaporation
pp	atm	Total precipitations
sp	atm	Surface pressure
snow	atm	Total snow fall
ADPro	ocean	Maximum of the meridional overturning streamfunction in the North (Sv)
AABex	ocean	Maximum of the meridional overturning streamfunction in the bottom cell (Sv)
Fc30A	ocean	Meridional heat flux in the ocean at 30°S (PW)
T1mo	ocean	Difference of sea surface temperature between model and observation (°C)
S1mo	ocean	Difference of sea surface salinity between model and observation (°C)
VOLN	ocean	Sea ice volume in the Northern Hemisphere
VOLS	ocean	Same in Southern Hemisphere
A15N	ocean	Sea ice extent (15%, i.e calculated as the total area (km2) of grid cells with sea ice concentrations (sic) of at least 15%) in the Northern Hemisphere
A15S	ocean	Same Southern Hemisphere