Semi-Automatic Tuning of Coupled Climate Models With Multiple Intrinsic Timescales: Lessons Learned From the Lorenz96 Model

Redouane Lguensat, Julie Deshayes, Homer Durand, Venkatramani Balaji

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003367

Special Section:
Machine learning application to Earth system modeling

Semi-Automatic Tuning of Coupled Climate Models With Multiple Intrinsic Timescales: Lessons Learned From the Lorenz96 Model

Redouane Lguensat[1], Julie Deshayes[2], Homer Durand[2], and Venkatramani Balaji[3,4,5]

[1]Institut Pierre-Simon Laplace, IRD, Sorbonne Université, Paris, France, [2]LOCEAN-IPSL, CNRS, Sorbonne Université, Paris, France, [3]Laboratoire des Sciences du Climat et de l'Environnement, CEA Saclay, Gif Sur Yvette, France, [4]Princeton University, Program in Atmospheric and Oceanic Sciences, Princeton, NJ, USA, [5]NOAA/Geophysical Fluid Dynamics Laboratory, Ocean and Cryosphere Division, Princeton, NJ, USA

Abstract The objective of this study is to evaluate the potential for History Matching (HM) to tune a climate system with multi-scale dynamics. By considering a toy climate model, namely, the two-scale Lorenz96 model and producing experiments in perfect-model setting, we explore in detail how several built-in choices need to be carefully tested. We also demonstrate the importance of introducing physical expertise in the range of parameters, a priori to running HM. Finally we revisit a classical procedure in climate model tuning, that consists of tuning the slow and fast components separately. By doing so in the Lorenz96 model, we illustrate the non-uniqueness of plausible parameters and highlight the specificity of metrics emerging from the coupling. This paper contributes also to bridging the communities of uncertainty quantification, machine learning and climate modeling, by making connections between the terms used by each community for the same concept and presenting promising collaboration avenues that would benefit climate modeling research.

Plain Language Summary Climate models are computer simulation codes that incorporate centuries of human knowledge of the physics of planet Earth. They are used to understand the past, the present and make projections about the future of our climate. To validate a climate model, scientists tune a number of its parameters so that it yields a simulated climate resembling real-life observations as much as possible. The main challenge in this tuning task is the extreme cost of climate models which limits a lot the number of tuning experiments scientists can run. In this paper we are interested in a technique that uses artificial intelligence in order to replace the expensive climate model with a cheaper surrogate. We experiment on a simplified model to assess the strengths and weaknesses of this semi-automatic technique, and show that it can be more efficient when combined with human expertise.

Key Points:

- The History Matching method is explained in detail then used for tuning a toy coupled model: the Lorenz 96 model
- The importance of several design choices is demonstrated, especially when considering forced experiments such as Atmospheric Model Intercomparison Protocol and Ocean Model Intercomparison Project
- We argue that this tuning method is semi-automatic & highlight the importance of human expertise when considering it for real coupled models

Correspondence to:
R. Lguensat,
redouane.lguensat@ipsl.fr

Citation:
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. https://doi.org/10.1029/2022MS003367

Received 25 AUG 2022
Accepted 6 APR 2023

Author Contributions:
Conceptualization: Julie Deshayes, Venkatramani Balaji
Funding acquisition: Venkatramani Balaji
Methodology: Redouane Lguensat, Julie Deshayes, Venkatramani Balaji
Project Administration: Venkatramani Balaji

1. Introduction

Climate models, or Earth system models (ESMs), have become a primary means of exploration of our changing climate. Numerical models of the Earth system were among the earliest applications of digital computing (see Platzman, 1979, for a participant's account of early attempts at numerical modeling of the atmosphere). This soon gave rise both to numerical weather prediction, and studies of the climate, what John von Neumann called the "infinite forecast" (Smagorinsky, 1983), the statistics of weather fluctuations over long time periods. The inclusion of the ocean circulation into the climate system, starting with Manabe and Bryan (1969), also led to our first attempts to understand the radiative and thermal balance of the planet under changes in CO_2 concentration (Manabe & Wetherald, 1975). This eventually leads to the series of reports issued by the Intergovernmental Panel on Climate Change, the latest one being the sixth, where the human footprint on climate and biodiversity is now starkly visible, with fossil fuels being a primary cause.

There are by now about 116 models from 44 institutions around the planet building ESMs (according to the ESGF data statistics portal) to understand the climate and work out the implications of various climate policy choices. The construction of these ESMs is generally described in "model documentation" papers, that detail the various scientific and technical choices that went into the building of these models. We take here as examples the documentation of the most recent model versions from two of the world's premier modeling institutions, the Geophysical Fluid Dynamics Laboratory (Held et al., 2019), and the Institut Pierre-Simon Laplace (IPSL) (Boucher et al., 2020). While these articles describe complete coupled models, they typically begin

Resources: Julie Deshayes
Software: Redouane Lguensat, Homer Durand
Supervision: Julie Deshayes, Venkatramani Balaji
Validation: Redouane Lguensat, Venkatramani Balaji
Writing – original draft: Redouane Lguensat, Julie Deshayes, Homer Durand, Venkatramani Balaji
Writing – review & editing: Redouane Lguensat, Julie Deshayes, Venkatramani Balaji

by describing the choices made in the "fast" component, the atmosphere (e.g., Zhao et al., 2018a, 2018b) for GFDL's model (Hourdin, Rio, et al., 2020), for IPSL. Typically this component is first constructed, by building or refining representations of the various physical processes contributing to the atmospheric circulation, applying the most recent physical knowledge and observational constraint. A frequent method to validate the atmospheric component is to run experiments conforming to the Atmospheric Model Intercomparison Protocol (AMIP: Gates, 1992). AMIP experiments typically run about 30 simulated years (SY). Subsequently this model is coupled to the "slow" component, the ocean. (Other components of the climate, e.g., changes in the biosphere are also "slow" relative to the atmosphere: in this paper, we use the ocean to stand in as the canonical "slow" component.) While the ocean component can be validated to a certain extent following the Ocean Model Intercomparison Project (OMIP: Griffies et al., 2016), several key phenomena of interest such as the El-Niño Southern Oscillation (ENSO), or simply sea-ice cover in northern and southern hemispheres, are emergent properties of the coupled ocean-atmosphere system, as noted by Held et al. (2019) and Adcroft et al. (2019) for and Mignot et al. (2021) for IPSL.

We draw attention to the appearance of the word "tuning" in the title of several of the articles cited above (Hourdin, Rio, et al., 2020; Mignot et al., 2021; Zhao et al., 2018b). The tuning, or calibration as termed by statisticians, of ESMs is itself an object of study in its own right (Hourdin et al., 2017; Schmidt, Bader, et al., 2017; Schmidt, Lan, et al., 2017; Schmidt, Teixeira, et al., 2017). Broadly speaking, this takes place in two stages, a first stage where we vary a few parameters in each process representation to bring them individually within observational constraints, and a second stage where the whole integrated system is calibrated to bring it within global constraints such as the Earth's radiative balance as observed from space by satellites. In the "traditional" approach to model calibration, described above, the model is run forward for a sufficient time for a given choice of parameters to validate it against observations. For the full coupled system including the "slow" physics of the ocean, doing this evaluation over a range of parameter choices can be expensive (Adcroft et al., 2019), 50,000–150,000 SY, limiting the exploration.

The stalling of arithmetic speed in computing in recent years has led to a similar slowdown in the addition of detail in ESMs (Balaji, 2021), compounding the problem. On the other hand, new computing hardware is well-suited to methods of machine learning (ML), leading to a reawakening of interest in ML methods first pioneered in the 1950s and 1960s (Sonnewald et al., 2021, for a recent review on the subject). We are interested in particular by the use of ML techniques for *surrogate modeling* where the broad approach is to use ML to *emulate* the expensive forward models. The emulator is then used to perform the explorations of parameter space for calibration, minimizing the number of instances of the forward model that need to be run for calibration. One recent approach applied for climate modeling is the *Calibrate-Emulate-Sample* method of the CliMA group (Cleary et al., 2021) where a limited set of runs of an atmospheric model is used for a broad characterization of an attractor in the parameter space of the model, by comparison with a reference, typically a model higher up "Charney's ladder" (Balaji, 2021), such as a large-eddy simulation (LES) model (Couvreux et al., 2020; Dunbar et al., 2020). Such models are very high resolution and very expensive and cannot be run on climate timescales (Schneider, Teixeira, et al., 2017). Emulators are then used to refine the landscape of the attractor, which can then be sampled to yield optimal values of parameters, as well quantify the uncertainty bounds on these values.

In an alternate approach pioneered by the HighTune Group (Couvreux et al., 2020; Hourdin, Williamson, et al., 2020), a version of the *History Matching* (hereinafter noted HM) method developed by D. Williamson et al. (2013) is employed. In this method, the emulator is used in successive waves not to find optimal parameter values, but to eliminate implausible regions of parameter space, according to a chosen set of metrics (distance between model outputs and observations). The forward model is then used to sample only the region of parameter space not ruled out yet (NROY). If the NROY space is a null space (i.e., no plausible parameter values that satisfy constraints to within tolerance), the method identifies "structural error" in the model, which is unable to satisfy observational constraints (D. Williamson et al., 2015).

While these methods show tremendous promise, their application so far has been in the tuning of a small component of the model, the physics of shallow clouds calibrated against an LES model where shallow cloud dynamics and physics are resolved. It has been shown by Hourdin, Williamson, et al. (2020) that these methods can accelerate atmosphere model tuning relative to the conventional methods. An open question is one of how these methods will fare in the presence of the multiple timescales of the ocean-atmosphere system. The challenges posed to some pioneering ML methods for the ocean's timescales have been outlined in Sonnewald et al. (2021).

Some of these questions are already under investigation in a coupled model, and the challenges are becoming clear. In this article, we use a canonical idealized model of the climate, the two-scale Lorenz96 (L96, described below in Section 2.1) system to explore the question of tuning in the presence of multiple timescales. The L96 system has often been used for studying climate models with a dynamical-systems lens, for example, Schneider, Lan, et al. (2017), Christensen and Berner (2019).

In the original description by Lorenz, this model represents westward propagating waves, the slow variable, amidst a noisy background, the fast variable, as a simplification of atmospheric flows. We envision wider implications of our results for model tuning. Tuning a coupled ocean-atmosphere model requires the encompassing of conjointly a slow variable, the ocean, and a faster one, the atmosphere. Alternatively, tuning an ocean model alone (as forced by prescribed atmospheric conditions), requires the tuning of parameterizations of fast meso and submesoscale processes, and larger-scale slower dynamics such as horizontal gyres or the meridional overturning circulation, with the risk of introducing compensating parameter errors that offset each other to provide approximately optimal solutions. These two illustrations are not unique, as one can easily find other examples of multi-scale dynamics in the climate system that require specific *ad hoc* tuning to respect direct and inverse energy cascades across scales. Here we use the L96 system where we interpret the slow and fast layers of the system as ocean and atmosphere, respectively. This paper explores one of the methods for automatic tuning outlined above, the HM method, in the presence of multiple timescales.

The paper is organized as follows: the two-scale L96 model and the History Matching algorithm are presented in Section 2. In Section 3, dedicated to the results of various experiments in a perfect model framework, we start with direct applications of HM to the full L96 model, and then consider AMIP and OMIP-like experiments that is, experiments where only one of the two variables is explicitly resolved. We finally discuss, in Section 4, the lessons learned from applying HM to L96 and the open research avenues that could lead to an efficient application of HM to coupled climate models.

2. Materials and Methods

2.1. The L96 Model

Introduced by Edward Lorenz in a ECMWF workshop on predictability (Lorenz, 1996), the L96 model is still one of the most used toy models in geosciences. It serves as a simple test bed to investigate the performance of algorithms related to dynamical system forecasting, parameterization, data assimilation and more (Chattopadhyay et al., 2020; Gagne et al., 2020; Lguensat et al., 2017; Lorenz & Emanuel, 1998; Rasp, 2019).

The two-scale Lorenz-96 model consists of K slow variables $X_k (k = 1, \ldots, K)$, each of which is coupled to J fast variables $Y_{j,k} (j = 1, \ldots, J)$. The ODE system is described by the following equations:

$$\frac{dX_k}{dt} = \underbrace{-X_{k-1}(X_{k-2} - X_{k+1})}_{\text{Advection}} \underbrace{-X_k}_{\text{Diffusion}} \underbrace{+F}_{Forcing} \underbrace{-\frac{hc}{b}\sum_{j=1}^{J} Y_{j,k}}_{Coupling} \tag{1}$$

$$\frac{dY_{j,k}}{dt} = \underbrace{-cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k}}_{\text{Advection}\quad\quad\quad\quad\quad\text{Diffusion}} + \underbrace{\frac{hc}{b}X_k}_{\text{Coupling}} \tag{2}$$

The two types of variables have periodic boundary conditions and are arranged in a cyclic way (Figure 1). The system is integrated using a Runge–Kutta fourth order scheme with a time step of $\Delta t = 0.001$ (note that L96 is fully non-dimensional). We based our implementation of the L96 on the Python code accompanying the paper of Rasp (2019) https://github.com/raspstephan/Lorenz-Online, with the difference being in using the original formulation of the L96 as stated in Equations 1 and 2.

As in Lorenz' seminal paper, we use $K = 36$, $J = 10$, a chaotic behavior is ensured by using the parameters $h = 1$ and $F = c = b = 10$ (Lorenz, 1996). In such setting, the temporal-scale ratio c implies that the Y (fast) variables fluctuate 10 times rapidly as the X (slow) variables, and b the spatio-scale ratio implies that their amplitude is 1/10 of the slow variables.

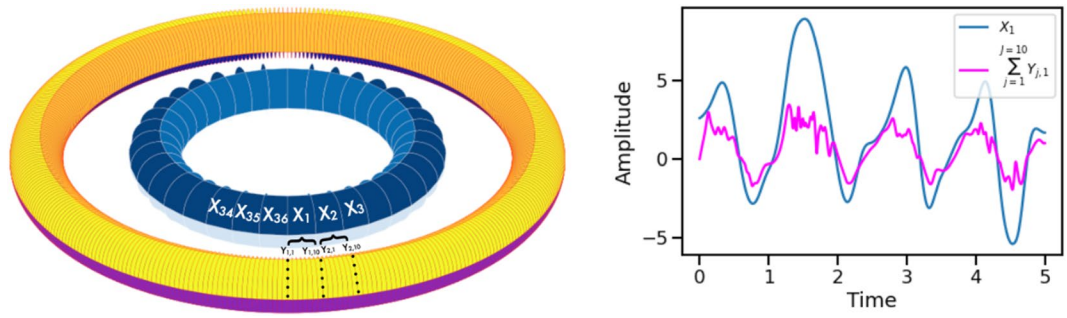

Figure 1. The two-scale Lorenz 96 system consists of a non-linear "slow" ODE of X variables coupled to a non-linear "fast" ODE of the Y variables. In our case, the number of X is 36, where each X is coupled to 10 Y variables as shown in this schematic. Time is expressed in MTUs (Model Time Units) where 1 MTU = $1000\Delta t$.

As noted above in Section 1, this article intends to reproduce the process of tuning of a system with multiple temporal scales, in order to help us anticipate what might transpire when the HM method is applied to a full coupled climate model. We underline that this is a gross simplification, as the real system has many interlocking feedbacks and more than just two timescales. The forcing F is also stationary here, whereas the external forcing in the real system has both short-lived (e.g., volcanic eruptions) and slowly-evolving (e.g., greenhouse gas emissions) components. Nevertheless, we will demonstrate that there are important lessons to be learned regarding the coupling of multiscale systems, and this article will walk through the steps of automatic tuning to illustrate the strengths of the objective approach, as well as aspects requiring caution.

Concretely, we use time-averaged statistics of the variables produced by such a system, with the settings mentioned above, as observations and consider the History Matching technique as a means to tune the four parameters $\{F, h, c, b\}$. As this is an abstract model with no observational counterpart we set ourselves in a *perfect model* framework, and consider a preliminary realization of the L96 model as the trajectory from which we can extract an analog of the climatological quantities used to tune climate models, that is, the "truth." Note that we are interested in approximating the model's *climate* (attractors in its parameter landscape), not the actual trajectory itself (its "weather"). Afterward, we pretend that we do not know which values of the parameters is the most appropriate to replicate the truth and examine to what extent the History Matching technique is capable of finding values of parameters $\{F, h, c, b\}$ that produce a suitable approximation of the true model climate.

2.2. The History Matching Method

The term "History Matching" (HM) stems from the oil engineering community, where the use of statistical "reservoir simulators" instead of complex and expensive fluid flow models serves to find inputs (reservoir geology e.g.,) for which the outputs closely match historical hydrocarbon reservoir production (Craig et al., 1997; Sacks et al., 1989). HM is a well published and established method used in several science and engineering applications such as galaxy formation models (Bower et al., 2010) and infectious disease models (Andrianakis et al., 2015), to name a few. HM is also closely related to precalibration (Edwards et al., 2011) used to reduce parameter search space before performing a Bayesian calibration.

The uptake of these methods in climate model development is perhaps not as high as it could have been, and it is still an open question whether this is simply due to conservatism, or whether there are indeed good reasons why these methods do not scale to full-complexity models (Salter et al., 2019). However there are recent forays into the use of HM in atmospheric model development, as noted above in Section 1, notably via the HighTune project (Couvreux et al., 2020; Hourdin, Rio, et al., 2020; Villefranque et al., 2021).

In the conventional method of tuning, the parameter space is explored by trial and error, where a number of (expensive) runs of the forward model are examined by experts to arrive an at "optimal" set of parameter values. In the HM approach, few runs of the expensive model are used to construct a "training" data set used by fast emulators or surrogates such as Gaussian Processes (GP) (Gramacy, 2020; Kennedy & O'Hagan, 2001) which mimic the parameterized model and extrapolate on the whole parameter space. Successive "waves" iteratively eliminate implausible regions of parameter space, and the not-ruled-out-yet (NROY) space presents a smaller region of parameter space to be explored, using the full (expensive) model rather than a surrogate.

Algorithm 1. History Matching With Iterative Refocusing

Require: the (expensive) simulator S + a priori information about the range of P parameters + M predefined metrics of interest

Step 1: Select a number of parameter sets

- Define the hypercube formed by the ranges of the P parameters.
- Sample $N = 10 \times P$ training points from that parameter space using maximin Latin Hypercube Sampling (or another space-filling design method).

Step 2: Calculate the metrics

- Evaluate the simulator at each of the N points. Construct the training data $D_{Train} = \{\mathbf{p}_n, S(\mathbf{p_n})\}_{n=1}^{N}$.

Step 3: Train the emulators

- Train a GP-based (or another ML-based) emulator for each of the M metrics (following a single GP per single output strategy).

Step 4: Calculate Implausibilities

- Evaluate the implausibility $I(\mathbf{p}')$ over a large number (usually hundreds of thousands) of parameter samples $\mathbf{p}'$ from the initial hypercube, using the emulators.
- Identify NROY space as points where the implausibility is less than 3 for each of the M metrics (conservative strategy). Alternatively, a finite number of metrics may be allowed to have implausibility higher than 3.

Step 5: Iterative refocusing

- Resample the NROY space and repeat steps 2–4 unless
  - the emulator's uncertainty V_c is smaller than the other uncertainties,
  - NROY is empty (all space is implausible),
  - the computational allowance for running the simulator has exhausted.

In the ML community, the process of using cheap emulators with the aim of derivative-free calibrating of costly black-box models, belongs to the field of surrogate modeling or meta-modeling (Gramacy, 2020). HM thus benefits from many of the advancements in this field, and is affected by the ongoing revolution in ML techniques and tools.

It is important to note that many similarities arise when comparing Bayesian calibration and HM, they however have fundamental differences in their strategies. In fact, calibration by construction will always result in a posterior distribution over all the input space (because the distribution needs to sum up to one) and then will always find a solution to the calibration problem. History matching, in contrast, focuses on ruling out bad solutions, and might even in some cases lead to an empty set of solutions.

Algorithm 1 summarizes the overall steps of the HM method, which are detailed in the following subsections.

2.2.1. Space Filling Design

The need for collecting small informative samples that are cheap and at the same time permit a good representation of a whole search space, made sampling methods at the heart of Design of Experiments (DoE or DoX), a research area on its own.

In surrogate modeling in general and HM in particular, starting from a guess interval for the set of parameters, we attempt to "fill" parameter space as uniformly as possible, and, if possible, aim for minimal correlation between the parameters in the design. Two families of sampling methods exist in the literature: space-filling designs and low-discrepancy sequences (see Steponavičė et al., 2016, for a review of these methods). As in most recent literature and in particular in D. B. Williamson et al. (2017), we use a popular space-filling design technique: maximin Latin Hypercubes Sampling (LHS) (McKay et al., 2000; Morris & Mitchell, 1995). Note that the space of ranges of all parameters describes a hyper-rectangle rather than a hypercube, until the parameters are normalized, a procedure applied later in the algorithm to help training the GP.

Chapman et al. (1994), a study from the sea ice modeling community, introduced the rule of thumb to determine the number of samples, and suggested that a number of $10 \times P$ samples where P is the number of model parameters, is usually a good starting point. Loeppky et al. (2009) ran a number of experiments to further support this rule of thumb.

2.2.2. The Metrics

Metrics here refer to the quantities chosen by the modeler to assess the performance of the model. Their role in tuning is crucial since they constrain the parameter search. We note that in the ML community, the term "metrics" refers usually to mathematical functions used for evaluating the performance of ML models (such as MSE, cross-entropy, etc.). Here it is the performance of black-box simulator, S, that is, the climate model in practice, that the metrics are expected to assess.

When tuning a climate model, the metrics are physical quantities that reflect the distance between climate model outputs and observations (e.g., the shortwave cloud radiative effect as in Hourdin, Williamson, et al., 2020).

A major difference in our application of HM from D. B. Williamson et al. (2017) and Hourdin, Williamson, et al. (2020) relies on the dimension of the metrics vector. With as many as 180 metrics (and presumably as soon as $M > 10$), it becomes relevant to reduce the dimensionality of the metrics vector using for example, Principal Component Analysis (hereafter PCA, also known as Empirical Orthogonal Functions in the geoscience literature, applied onto the matrix from model outputs at the different parameters). In brief, PCA aims at computing the eigenvectors and eigenvalues of the data covariance matrix. A classical and numerically stable approach to perform PCA is through the Singular Value Decomposition algorithm (SVD), we refer the reader to Shlens (2014) for more details. This algorithm is implemented within the `scikit-learn` Python library we use for PCA in this work (Pedregosa et al., 2011).

As a consequence, the GP will be fitted during step 3 in the reduced space and not in the M-dimensional space. Observations are also projected in the reduced space before calculating the implausibilities. Reducing the dimensionality of metrics seems to be relevant for metrics that are not scalar, for example, geographical maps of geophysical variables. However, due to the simplicity of L96 model, we could not test such procedure in the present study.

2.2.3. The Emulator

In the surrogate modeling literature, GP have been the method of choice when dealing with tuning or calibration problems. GP are simple regressors that have the advantage of yielding not only mean estimates but also uncertainties. As in D. B. Williamson et al. (2017), GP are considered in this work as a viable candidate to emulate the metrics in Equation 3.

Let us denote p, a set of tunable parameters as a *configuration*, and $f(p)$ the vector of PCA-transformed metrics associated with the result of numerical simulation using this configuration. The mathematical model used for the regression is based on the following formula:

$$f_i(p) = \sum_j \beta_{ij} g_j(p) + \epsilon_i(p) \tag{3}$$

$$\epsilon_i(p) \sim GP(0, C_i(.,.;\phi_i)) \tag{4}$$

where $g(p)$ contains a library of basis functions in p (polynomials e.g.,), β is a set of trainable coefficients, C_i denotes pre-specified covariance functions for the GP and ϕ_i are their parameters. Note that the two equations can be summarized into one by taking the linear regression part as a mean function of the GP. Training the overall regression model resorts to fitting all the parameters of the GP.

In this work, we take a blind approach to the fitting of the mean function, that is, no human expertise is available on the choice of basis functions. We however note that if such expertise is available it must be prioritized. The regression can then be performed using classical ordinary least squares or sparsity enforcing techniques such as LASSO; here we follow, as in D. B. Williamson et al. (2017), a forwards and backwards stepwise regression (Draper & Smith, 1998) approach. Regarding the covariance function, again several choices are available (Radial Basis Function (RBF), Matérn, Periodic, etc.) but throughout this paper we use the RBF kernel for its simplicity.

It must be noted that using an overly complex mean function can easily lead to overfitting, especially when being in a small data set regime (tens of samples) as in the experiments conducted in this paper. To address this

issue several procedures are employed following Salter and Williamson (2016): first, the maximum number of terms allowed for the mean function regression step is 10th of the sample size. Second, a nugget term is added to the RBF kernel to avoid interpolating the training design points. Last but not least, a proper validation of our GP-based emulators must be conducted, here we use a classical Leave-One-Out approach (L1O). In the L1O procedure we run a series of our emulator refitting by leaving out one sample each time from the design points and then testing if that sample lies within the 95% confidence interval of the emulator. We would ideally then expect that a 95% ratio of the left out points lie inside the confidence intervals if the emulator is good. While it is an ideal case, we consider in this work that the emulator is good if the ratio is at least over 90%.

2.2.4. The Implausibility

A natural way to find the configuration (i.e., the set of parameters p) that allows the model to get as close as possible to the real state of the climate system, resorts to defining a distance measure between the PCA-transformed metrics $f(p)$ and the PCA-transformed observations of the real system z. Note that these observations are eventually biased, due to instrumental inaccuracies for example, hence the metrics calculated from observations are eventually distinct from the actual true metrics, denoted y below.

Using the emulator to find the configuration that minimizes the distance between the model output and the system state, comes to solving the following optimization problem:

$$p^* = argmin_p \|z - f(p)\|,$$

where $\|\cdot\|$ is a norm taking into account different uncertainty sources. The mathematical formulas hereinafter take place within each component of the transformed space. Here we use for $\|\cdot\|$ the Mahalanobis distance

$$\|z - f(p)\| = (z - f(p))^T \mathrm{Var}\,[z - f(p)]^{-1}(z - f(p)).$$

Because we only run a selected number of simulations and use the emulators to complete the hypercube of parameters, we do not have access to the entire distribution of $f(p)$ but only to the expectation $\mathrm{E}[f(p)]$ and to the variance $\mathrm{Var}[f(p)]$. Following notations in D. B. Williamson et al. (2017), we denote V_e the error variance of observations and V_η the error variance due to the simulator uncertainties that we suppose to be independent. We can then reformulate the distance to observations, using the prediction of the emulator, and we refer to it as the *implausibility*:

$$\begin{aligned} I(p) &= \|z - \mathrm{E}[f(p)]\| \\ &= (z - \mathrm{E}[f(p)])^T \mathrm{Var}\,[z - \mathrm{E}[f(p)]]^{-1}(z - \mathrm{E}[f(p)]) \\ &= (z - \mathrm{E}[f(p)])^T \mathrm{Var}\,[(z - y) + (y - f(p)) + f(p) - E[f(p)]]^{-1}(z - \mathrm{E}[f(p)]) \\ &= (z - \mathrm{E}[f(p)])^T (V_e + V_\eta + \mathrm{Var}[f(p)])^{-1}(z - \mathrm{E}[f(p)]) \end{aligned}$$

With this definition, we factor in the implausibility for the different uncertainties associated with the model itself (L96 in our case, the climate model otherwise), with observations and with the predictions of the emulators. Thus the small values of implausibility appear in two cases only: the distance between the prediction of the model and the real state of the system is small or one of the three uncertainties is too high.

Having calculated the implausibility for a given set of parameter p, deciding to keep it or rule it out is based on a threshold value T for the implausibility that is, p is ruled out if $I(p) > T$ and the NROY region is $\{p: I(p) \leq T\}$. For d-dimensional $\mathbf{z}$, Bower et al. (2010) use $T = \chi^2_{d,0.995}$, the 99.5 th percentile of the χ^2 distribution with d degrees of freedom.

If z is one-dimensional, as will be considered in this work, the implausibility is written as:

$$I(p) = \frac{|z - \mathrm{E}[f(p)]|}{\left[V_e + V_\eta + \mathrm{Var}[f(p)]\right]^{1/2}} \tag{5}$$

and the Pukelsheim rule is generally used and sets $T = 3$. This general rule states that any continuous unimodal distribution contains at least 95% of its probability mass within a distance of 3 standard deviations from its mean.

2.2.5. Iterative Refocusing

When the emulators are trained, they can be applied in inference mode on a high number of configurations from the initial hypercube as mentioned in Algorithm 1, the implausibility score and the threshold T define the NROY

as explained in the previous section. Once the first NROY space is produced, it seems obvious to apply the HM procedure, on this reduced space. This has the advantage of training better GP-based surrogates only on samples from the new NROY, because the parameter space is smaller than before, therefore reevaluating the implausibilities. We refer to this task by the term *refocusing* (D. B. Williamson et al., 2017), and each iteration of steps 2–4 in Algorithm 1 is called a wave.

The iterative aspect of the HM provides a certain flexibility that other approaches may not. After having significantly reduced the parameter space with a set of metrics describing well the general tendencies of the system, one could try to reduce it further by using metrics describing some more local aspects. Notwithstanding, the multi-wave HM comes with some challenges. Many flavors of HM exist in the literature, where some specific choices are made for some steps, notably, the stopping criteria and the NROY resampling strategy. Overall, the stopping criteria are problem-dependent, and the limitation of computational resources most often comes to play. In a pragmatic manner, the iterative process must be halted when it is expected that performing one additional wave would not reduce the parameter space sufficiently compared to the computational cost that it would require (as a reminder, the simulator is run at each step 2 of the algorithm). Also, as stated in D. B. Williamson et al. (2017), "*when the emulator variance is largely smaller than the denominator in the implausibility calculation, then it is unlikely that further waves will change the implausibility very much*" and it may be unreasonable to perform a new wave.

In addition, a difficulty arises with multi-wave design right after the first wave: LHS cannot sample the NROY space as it is in general not a hyper-rectangle anymore and may contain several disconnected regions. Several strategies were followed in the literature: Yeh et al. (2016) used clustering to find medoids that act as representatives of each cluster, Andrianakis et al. (2015) used Gaussian random variables centered at the mean from NROY points to sample design points for the subsequent wave. This is an open research question and recent advances in ML can lead to promising avenues (Garbuno-Inigo et al., 2020). In the present application of HM to L96, we follow a rejection sampling approach to sample subsequent waves' NROY as in reference papers employing HM, such as Bower et al. (2010): we apply LHS on the entire initial parameter space with enough samples to retain approximately the desired $10 \times P$ configurations after rejecting those with an implausibility score greater than 3 for each emulator. Concretely, we introduce in this work this simple method: after wave number w, we estimate the search space reduction ratio

$$r = \frac{\text{NROY}_w \text{volume}}{\text{Initial search space volume}},$$

we then generate a maximin LHS sampling of size $\lceil \frac{10\times 4}{r} \rceil$ on the initial search space, where $\lceil . \rceil$ is the ceiling function. The samples of wave w are checked if they are ruled out by a previous wave emulator, only the samples that were not ruled out by all the emulators of the previous waves are kept. If the number of the remaining samples is lower than 10×4 we repeat this procedure with a new LHS sample and concatenate the results. In most cases, we end up with a number of samples that is higher than 10×4 from which we select randomly 10×4 samples to respect the computational budget allowed for each wave.

If r becomes too small (which happens naturally when HM effectively reduces a lot the search space), the maximin LHS becomes computationally heavy and we then switch to a random LHS (without maximin optimization).

2.2.6. The Code

As stated before, the L96 model is written in Python, special attention was given to parallel computing in order to run several experiments efficiently, the JOBLIB Python library was used for this task. The main HM routine is written in R (LHS, GP-fitting, L1O, visualization). The code is based on the EXETERUQ_MOGP library that acts as an R interface for the Python written MOGP_EMULATOR library for training GPs, also available online. Again, to speed-up the calculation of implausibilities, we make use of the FUTURE R library for parallel computation. Please note that the code requires minimal R knowledge, which would not constitute a barrier for pure Python practitioners. All the scripts for this paper are preserved at Zenodo https://doi.org/10.5281/zenodo.7384270 (Lguensat, 2022).

3. Results

In this section we demonstrate the relevance of the HM algorithm in a perfect model setting that is, ($V_e = V_\eta = 0$). Our aim is to outline new insights regarding the influence of several design choices when applying HM and to evaluate its applicability to models with different time scales. The general goal of this study is to answer several

Table 1
Prior Intervals for the Parameters Along With the Ground Truth Values, for the History Matching Experiment With No Physical Based Prior

Params	Prior	True
F	[–20,20]	10
h	[–2,2]	1
c	[0,20]	10
b	[–20,20]	10

questions about the choices usually made when applying HM for climate modeling applications. Specifically, for the current paper, we aim to "history match" the metrics in Equation 6 to reduce the search space for the four L96 parameters $p = \{h, F, c, b\}$.

As the L96 model is only a toy version of a climate model, we chose as metrics the first and second order momenta of both variables, as well as their covariances:

$$f(p) = \begin{pmatrix} \langle X \rangle_\tau \\ \langle \bar{Y} \rangle_\tau \\ \langle X^2 \rangle_\tau \\ \langle X\bar{Y} \rangle_\tau \\ \langle \bar{Y^2} \rangle_\tau \end{pmatrix} \quad (6)$$

where X represents the $K = 36$ X_k variables, and $\bar{Y}$ represents the means of the $J = 10$ $Y_{k,j}$ variables associated to each X_k, hence the metrics vector $f(X, Y)$ includes 36*5 = 180 metrics. Note that the carets denote the time average of a function $\phi(t)$ over the time interval $[t_0, t_0 + \tau]$,

$$\langle \phi \rangle_\tau = \frac{1}{\tau} \int_{t_0}^{t_0+\tau} \phi(t) \mathrm{d}t,$$

which we calculate using discrete sums instead of integrals.

3.1. Data Generation

Using the "true" set of parameters $p_{true} = \{10, 1, 10, 10\}$ we run a short simulation (10 MTUs) to reach a state in the L96 attractor, save that state, then run again a longer simulation starting from the attractor for $\tau = 100$ MTUs. This latter simulation yields a specific trajectory of the L96 model that we will consider as the ground truth for our tuning experiments. As written above, in our chosen set-up, evaluating our metrics against the latter ground truth, yields a vector of 180 dimensions (Equation 6) for every set of parameters p.

For the maximin LHS at the first HM wave, we generate 10×4 (10 times the number of parameters to tune) = 40 set of parameters for the first wave. Subsequent waves follow the rule introduced in Section 2.2.6.

The emulation training problem calls for the use of multi-input multi-output GP (inputs: 4 dimensional, outputs: 180 dimensional). It is common to find in the literature the use of multiple single-output GP instead (one per output so 180 single-output GP in our case) due to their better scaling with memory and computational time. Obviously, this would disregard correlations between the outputs and may lead to lower performance since some information is discarded. A straightforward way to tackle both problems is through the use of dimensionality reduction algorithms. In this study, we use PCA and keep a number of principal components that retain 99% of the total variance. Emulating the metrics in the reduced space with independent GP often performs as well as using a multi-output emulator, especially if the size of the training samples is large compared to the dimension of the reduced space (Wilkinson, 2010).

3.2. History Matching With No Physical Prior

We start our experiments with a standard setup where we suppose we have access to the simulator, here the L96 code, and to the ground truth output metrics, but have priors uninformed by physical bounds on the parameters. Hence we start with arbitrary uniform priors except for parameter c that is positive because we know from the equations that c represents a time-scale ratio (Table 1).

We perform HM as described in the previous sections, and carry out a total of 6 waves which sums up to 240 simulations of the L96 model (Table 2). PCA reduces the problem to fitting 8 metrics instead of 180, an illustration of

Table 2
Evolution of the Not Ruled Out Yet Ratio r for Each Wave of the History Matching Experiment With No Physical Based Prior

Wave	NROY (%)	NbSim
1	16.96	40
2	7.91	40
3	5.55	40
4	2.31	40
5	0.94	40
6	0.02	40
		Total = 240

Note. Right hand side column specifies the number of explicit simulation with L96 model carried out at each wave.

the projection space of the two leading PCs we are fitting with GP along with the projected observations (ground truth) is shown in Figure 2. The NROY ratio r moves from 16.96% at wave one to 0.02% at wave six, this sharp reduction might be explained by the large priors. This situation is rather preferable than choosing narrow priors at the risk of missing the good parameters which would lead to an empty NROY from wave one.

More specifically, the NROY matrix suggests that most of the reduction of the parameter space is tied to the F parameter (Figure 3, we refer the reader to Appendix A for an explanation of NROY visualizations): nearly all the negative values are ruled-out (remember that the NROY matrix shows parameters normalized on the [–1, 1] interval, hence projecting back to the original space, this means that values of F on the interval [–20, 0] are ruled-out). NROY plots reveal also some covariance between parameters h and b. In wave 1 this is shown as a broad diagonal stripe in the NROY plot. By wave 6 the lowest and highest values of both parameters are ruled out, but the diagonal feature remains. The coupling terms in Equations 1 and 2 both have $\frac{h}{b}$, so the covariance identified is valid.

At wave 6, we are left with 0.02% of the initial search space. The associated NROY plots show that we successfully narrowed the search space for parameters h, b, F, but interestingly parameter c presents multiple NROY regions (Figure 3, right hand side). This suggests possible metastability phenomena on time scales longer than the integration time used to calculate the metrics, which is also encountered in the Ensemble Kalman Inversion based calibration used in Schneider, Teixeira, et al. (2017). The effect of parameter c seems to be weaker than that of the other parameters, the wave 6 NROY suggests that several satisfactory solutions can be found with very different values of c.

It is worth highlighting again that the goal of HM is not finding *exactly* the ground truth solution, but ruling out the wrong ones. Thus, we consider a HM experiment to be successful if: (a) the ground truth solution is not ruled out, (b) the parameter search space is largely reduced. For this experiment with no physical prior (other than on the sign of c), running additional waves does not change the NROY space further, and becomes tedious because of the rejection sampling of an extremely small region.

From the remaining configurations in the NROY of wave 6, the modeler needs to make a final choice of the configurations they plan to run using the real (and expensive) simulator. Sampling a representative low number of configurations from the last NROY is not in itself part of the HM algorithm, but remains an open question from a practical point of view. We propose here to use a clustering algorithm, namely K-means, as a low cost solution to find an ensemble of representative points from the HM final NROY.

By applying the K-means algorithm we find six clusters (according to the silhouette score (Appendix D) whose centers are considered as candidate configurations. Before running the L96 model (our "expensive" simulator in

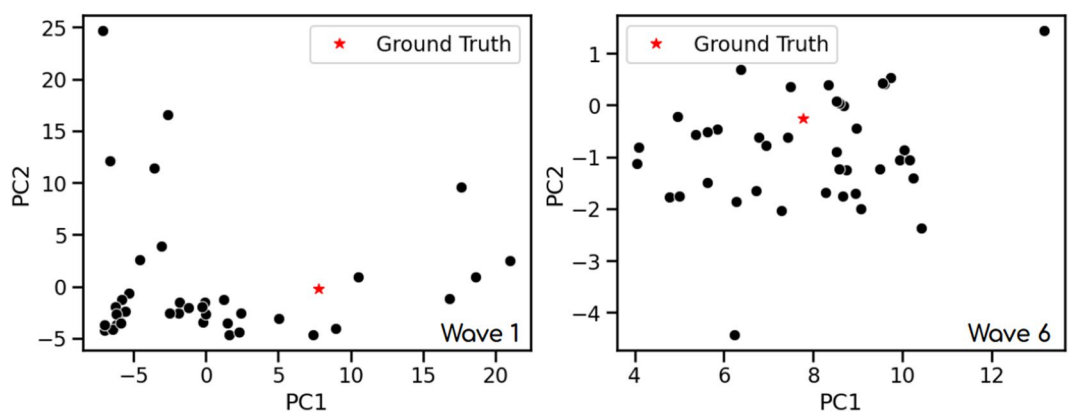

Figure 2. Samples used to train the Gaussian Process emulators projected onto the two leading PCs, at wave 1 (left hand side) and at wave 6 (right hand side).

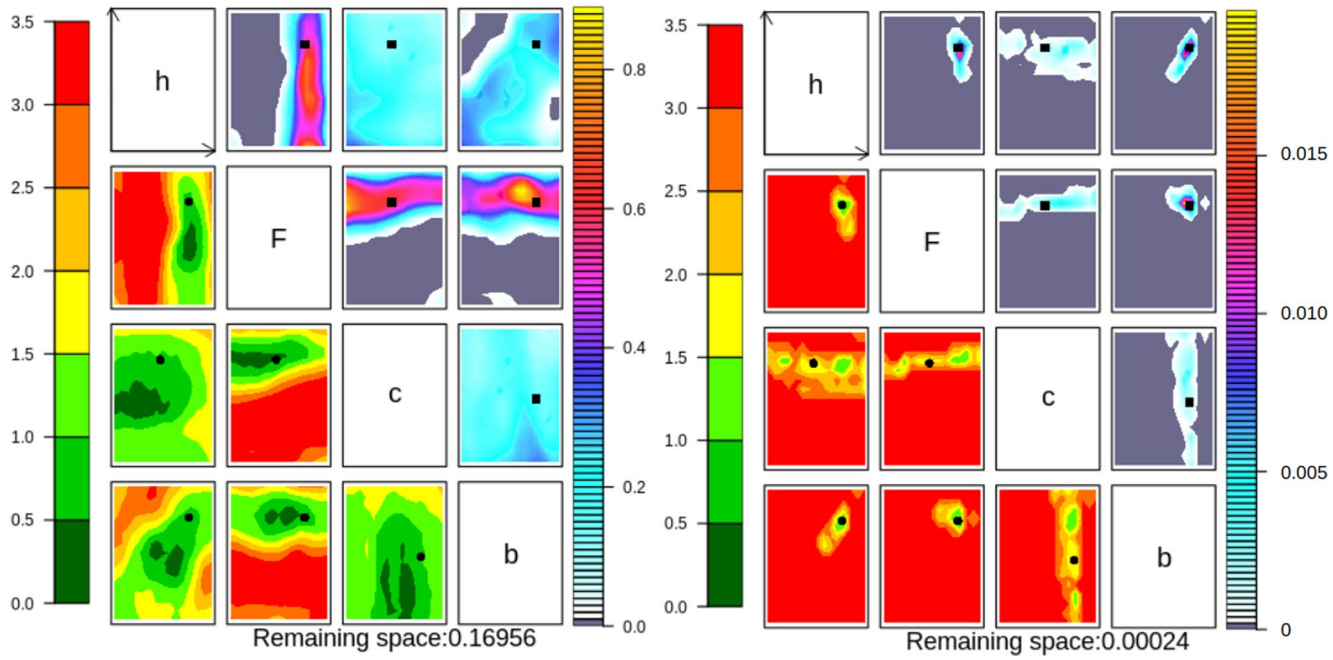

Figure 3. Not ruled out yet plots for the History Matching experiment with no physical based prior, at wave 1 (left hand side) and at wave 6 (right hand side).

this idealized study) with these sets of parameters, we verify that they belong to the last NROY effectively: it is not granted that a K-means cluster center is included in the NROY. Indeed, one of the centers did not respect this condition hence we finally end up with 5 configurations (Table 3).

Finally, we run L96 simulations with the 5 sets of parameters identified by the HM procedure (Table 3), calculate the metrics and overlay their distribution onto that of observed metrics (i.e., our initial L96 simulation, considered as ground truth, Figure 4). Note that we evaluate the realism of metrics with statistical distributions, similar to climate modelers evaluating simulated climate against observations. Looking at these distributions, one configuration clearly stands out as a bad candidate (Figure C1), although it was not ruled out by HM ! This configuration is an example illustrating that HM can retain a low-implausibility configuration not because the metrics are close to observations (low nominator of implausibility), but because the GP were uncertain (high denominator of implausibility). The reader can notice in Table 3 that *h* and *c* have values that are too close to the ranges boundaries, which might explain why the GPs are uncertain.

Table 3
Final Sets of Parameters Identified by the K-Means Procedure and Consistent With History Matching Results, in the Experiment With Unphysical Based Prior, Alongside the Ground Truth (GT) Values (Bottom Line), and the Corresponding Median KL-Div Between Their Simulations and Observations (Right Hand Side Column)

	h	F	c	b	
1	0.99	11.90	16.21	9.17	
2	1.15	8.94	3.94	10.33	
3	0.58	10.31	11.90	6.89	
4	0.92	10.31	10.28	9.35	
5	1.87	11.52	19.03	15.65	
GT	1	10	10	10	

As a conclusion from the HM experiment with no physical based prior, we demonstrate that the HM procedure can effectively reconstruct the parameters of the initial L96 simulation, with no physically based prior on the value of these parameters. Notwithstanding, we identify two major challenges that the user is left with. First, the NROY space reduces, to the most, at nearly 2000 configurations of parameters, despite the simplicity of this experiment (which only has 4 tunable parameters, as a reminder). Applying additional statistical techniques and evaluating explicitly some NROY configurations against observations, reduces the number of good candidates for reconstructing the initial simulation to 4. Second, the HM may retain low-implausibility configurations, not because the metrics are close to observations, but because the metrics were not sufficient to constrain the results. Additionally, when applying to ocean and atmosphere data, uncertainties that feed into the

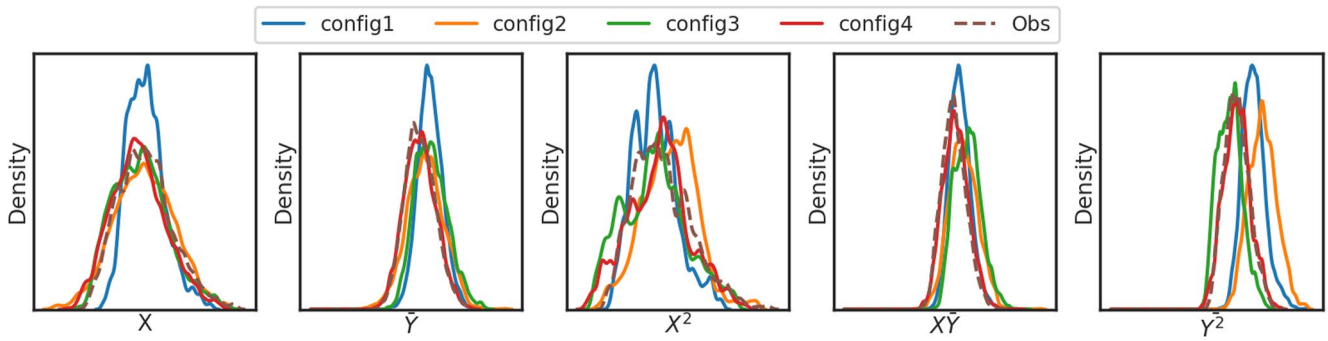

Figure 4. Histograms of metrics on the final 4 configurations identified by the K-means procedure and consistent with History Matching results, in the experiment with unphysical based prior, along with the histogram of the observed metrics.

implausibility may be too large, which we cannot test with a model as simple as L96. Incorporating more physical expertise in HM procedure, as illustrated in the section below, helps resolving these challenges.

3.3. Incorporating Domain Expertise in History Matching

In this section, we discuss how important is reducing the prior space, especially if the HM user has a physics-informed initial guess on the value of the parameters. For example, when using the two-scale L96 model, usually the X variables are slower than the Y variables, and their amplitude is higher than those of Y. We assume then c and b to be higher than 1. In addition, we can also introduce knowledge from the domain of application, for example, assuming that F and h are positive (which is the case in all studies where L96 model is used as a toy model for climate). The advantage of HM is that if these new priors do no contain the ground truth solution, the first wave will result in an empty NROY, and the modeler will be enforced to review their model, metrics, uncertainties and also the priors.

Performing the HM procedure with physically-informed priors (as listed in Table 4), required only 5 waves and a total of 200 simulations to reach the same reduction in parameter space as with no physical-based prior (Table 5). Actually, the NROY space becomes less than 1% of the initial space at wave 4, while it took one additional wave in the previous experiment. While running two waves ($40 \times 2 = 80$ L96 simulations) is computationally cheap in our study, we highlight that each wave that we would run for a coupled climate model would be very expensive. This illustrates how beneficial it is to chose the prior intervals as carefully as possible, incorporating as much as possible the knowledge from experienced modelers in the domain of application.

3.4. AMIP-Style Experiment

Our ambition when applying HM to L96 model is to mimic the calibration of parameters in ocean-atmosphere models, which couple a fast media, the atmosphere, where variables generally adjust within days to months, and the ocean, the slow media, where some variables require up to hundred years to reach equilibrium. Such calibrating exercise, commonly begins with preliminary calibration of the parameters of each media in uncoupled experiments, which explicitly simulate only one of the two media. In this section, we start with exploring the potential of HM for calibrating the parameters of L96 fast variables alone, that is, Y. As parameter F is not involved directly in the Y variables equation (Equation 2), only the three parameters $\{h, c, b\}$ are tuned here. Hence at each wave of the HM algorithm, we decide to run 30 simulations only. The prior intervals are taken to be similar to those of the previous section.

Table 4
Prior Intervals for the Parameters for the History Matching Experiment With Domain Expertise

Params	Prior	True
F	[0,20]	10
h	[0,2]	1
c	[1,20]	10
b	[1,20]	10

Those simulations only resolve the equation for Y variables, with X information extracted from the initial simulation, considered as ground truth (see Equation 7):

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,k}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b}\underbrace{X_k}_{\substack{\text{Retrieved}\\ \text{from initial}\\ \text{simulation}}} \quad (7)$$

Table 5
Same as Table 2 for the History Matching Experiment With Prior Intervals for the Parameters Limited by Domain Expertise

Wave	NROY (%)	NbSim
1	25.57	40
2	3.18	40
3	0.59	40
4	0.04	40
5	0.02	40
		Total = 200

where the parameters to be tuned are {h,c,b} and are highlighted in blue color.

In AMIP-style experiment, the metrics reduce to those concerning Y variables only, that is,

$$\boldsymbol{f}(p) = \begin{pmatrix} \langle \bar{Y} \rangle_\tau \\ \langle \bar{Y}^2 \rangle_\tau \end{pmatrix}, \tag{8}$$

Following PCA, the problem resorts to fitting 2 metrics instead of 72. Only 5 waves are necessary to reduce the NROY from 71.86% to less than 0.01% of the initial parameter space (Table 6). Still, at wave 5, there remains some uncertainty in the calibration for parameter c, as indicated by the band-shape features in $c-$ related panels (Figure 6).

3.5. OMIP-Style Experiment

We then explore the OMIP-style experiment and calibrate the parameters affecting L96 slow variables alone, that is, X. We are left with only two independent parameters to tune, F and $G = \frac{hc}{b}$ (Equation 9).

$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - G \underbrace{\sum_{j=1}^{J} Y_{j,k}}_{\text{Retrieved from initial simulation}} \tag{9}$$

The prior intervals for both parameters is the same as for previous experiments, [0,20], and we initially include all metrics that concern the X variables, that is,

$$\boldsymbol{f}(p) = \begin{pmatrix} \langle X \rangle_\tau \\ \langle X^2 \rangle_\tau \\ \langle X\bar{Y} \rangle_\tau \end{pmatrix}. \tag{10}$$

PCA reduces the problem to fitting 7 metrics instead of 108. With such definition of metrics, the NROY at wave 1 is empty, which seems to be due to mismatches between the initial simulation and the OMIP-style outputs for $\langle X\bar{Y} \rangle_T$ metrics primarily (Figure 7). This is presumably due to the lack of coupling between X and Y variables when running L96 in OMIP-style experiment (cf. Equation 9). Actually, $\langle X\bar{Y} \rangle_T$ metrics, which associates the slow and fast variables of L96 model, stands for any properties emerging from the coupling of an ocean with an atmosphere, for example, sea-ice related quantities.

When removing the $\langle X\bar{Y} \rangle_T$ metric from the evaluation against ground truth in OMIP-style experiment, then a single wave of HM reduces the parameter space to 1.8%. The final NROY space includes a single cluster centered on values $[F, G] = [10.23, 1.06]$, which are very close to the ground truth (Figure 8). As a conclusion, HM is efficient in OMIP-style experiments only if metrics reduce to those affecting the slow variable exclusively. This is another lesson we learned that we discuss in the next section.

Table 6
Same as Table 2 for the AMIP-Style History Matching Experiment, Where Only the Fast Component of L96 Is Calibrated

Wave	NROY (%)	NbSim
1	71.86	30
2	12.19	30
3	1.24	30
4	0.12	30
5	0.01	30
		Total = 150

4. Discussion

The Lorenz96 system, since it was first articulated by Lorenz, has been used to understand general circulation models from a dynamical systems perspective. In its typical use (such as by Lorenz himself in Lorenz (1996)), the X layer stands in for the resolved-scale flow, and the Y for subgrid-scale unresolved flows, to study the response of the full system to the aggregate

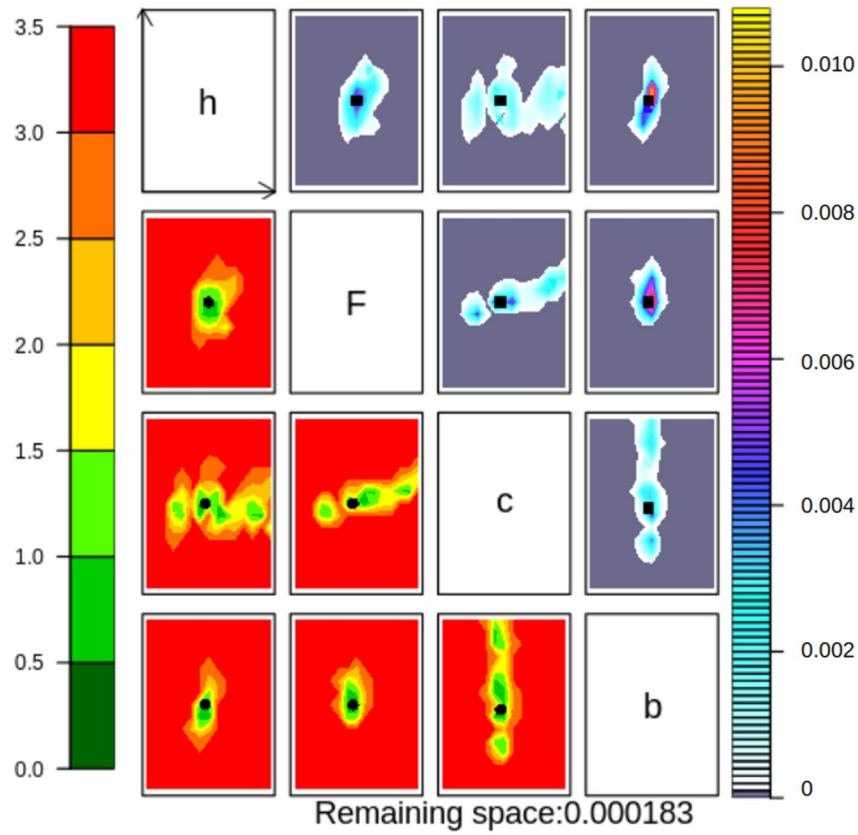

Figure 5. Not ruled out yet plots for the History Matching experiment with prior intervals for the parameters limited by domain expertise, at wave 5.

behavior of the unresolved scales. In this manuscript we take the same L96 system to study a different aspect of the coupled climate system, with a fast (Y, "atmosphere") and slow (X, "ocean") component, with the coupling representing the surface exchanges.

Our objective is to revisit the problem of model calibration in the presence of distinct timescales in separate components of the coupled system. As noted in studies of model calibration or "tuning" (e.g., Hourdin et al., 2017), this is generally done by first separately calibrating individual components of the climate system to conform to theoretical and observational constraints, and then in a separate step, applying global constraints on emergent properties of the coupled climate system, such as top-of-atmosphere radiative balance, or ENSO behavior (Held et al., 2019; Mignot et al., 2021). The initial stages are done with fixed surface exchanges: for example, in AMIP experiment, Zhao et al. (2018a) calibrates the atmosphere under prescribed sea surface temperatures. As the intent is not to "retune" the atmosphere after coupling, there is a careful assessment of how the ocean might react, by looking at diagnostics such the "implied" ocean heat transport (Zhao et al., 2018b). The ocean is likewise first calibrated using fixed atmospheric forcings, following the OMIP protocol described in Griffies et al. (2016). The ocean tuning in then refined a coupled setting (Adcroft et al., 2019), to adjust the tunings to match the differences between the standard atmosphere forcings of AMIP and the actual atmospheric model to be used in the GCM. This carries the risk, of course, of compensating for atmospheric model biases by adjusting ocean parameters. When tuning the IPSL climate model for CMIP6, it was actually intentional to compensate for biases in the atmosphere, that initially limited dense water formation, by tuning the sea-ice leads fraction and finally enhance ventilation of the deep ocean (Mignot et al., 2021).

As GCMs move toward more objective methods of model calibration, we must examine these nuances introduced by coupling. Here we have emulated the HM-based model calibration process in the L96 system. An initial "truth" run of L96 serves as the HM target for a given set of values of the parameters $\{F, h, c, b\}$: the aim is to discover optimal tuning values of these parameters from a given prior distribution of possible values. This is thus

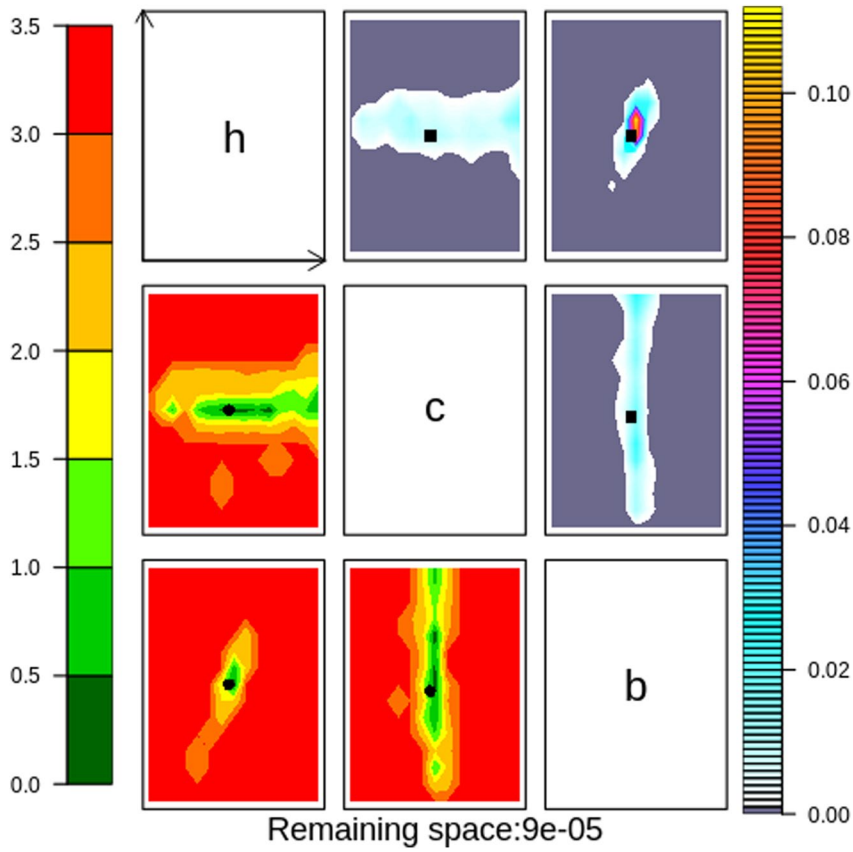

Figure 6. Not ruled out yet plots for the AMIP-style History Matching experiment, at wave 5.

a "perfect model" setting, and only parametric error can be studied, not structural error, since the same equations are used in the test model, with only the parameters to be discovered. In an ocean-atmosphere coupled model, as well as in a more complex climate model, there are structural biases inherent to the grid resolution, the choice of algorithms for discretization, and/or poorly known physics, in particular at the interface between the different

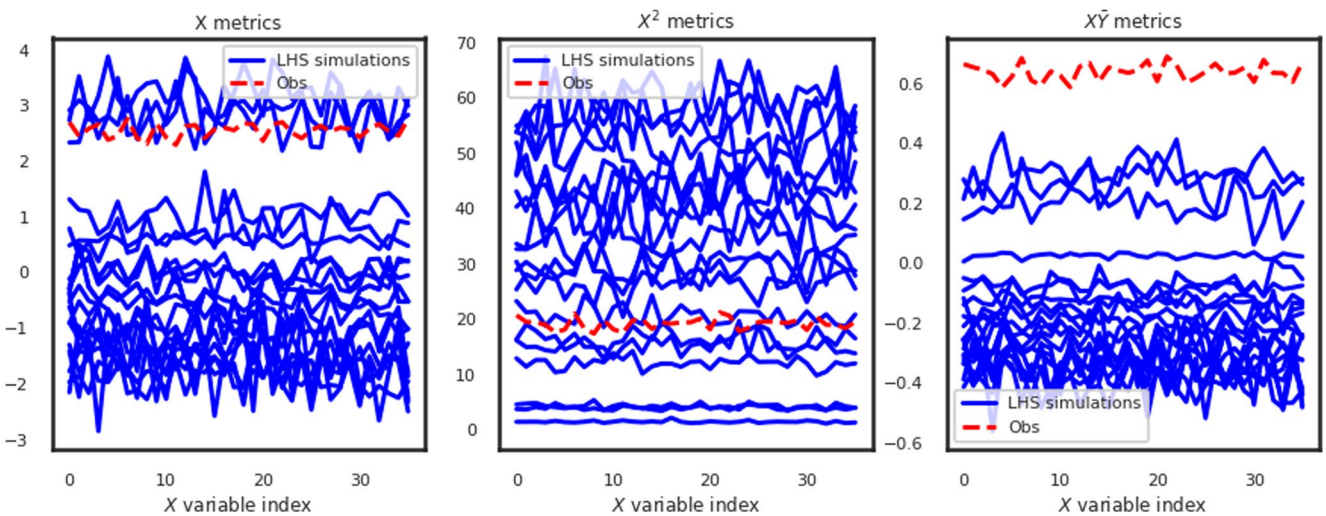

Figure 7. Non-dimensional metrics for the OMIP-style experiment: for the ground truth (red lines) and for a selection of parameters (blue lines).

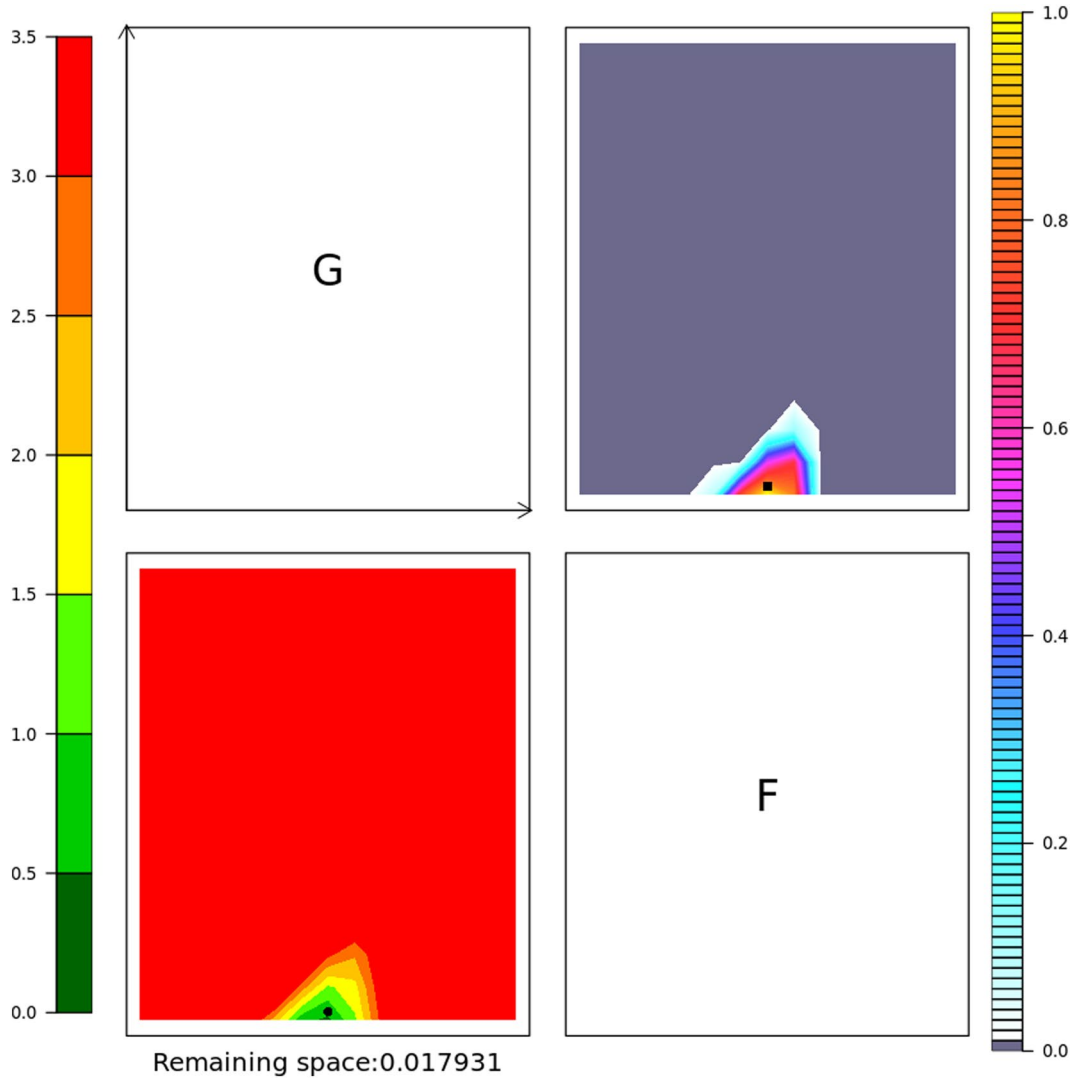

Figure 8. OMIP tuning experiment: not ruled out yet plot after one wave.

components. The HM procedure enables us to factor those in through uncertainty in the metrics, but an accurate quantification of those uncertainties is not straightforward as it often combines with parametric uncertainties (Balaji et al., 2022). Also when tuning a climate model, metrics refer to observations, that may encompass intrinsic uncertainty which should also be taken into account, as well, this limitation cannot be envisioned in a perfect model setting.

When tuning the L96 model, we use the set of metrics shown in Equation 6: they are first- and second-order metrics covering the slow and fast components separately, as well as a cross-term metric for the emergent properties of the coupled system. In calibrating a real GCM, we of course consider more complex and higher-order metrics, but broadly they will fall into one of these categories: fast, slow and emergent-coupled. It must be noted however that the L96 system is extremely simple compared to real GCMs, leading to some limitations in the applicability of our conclusions. Notwithstanding, our conclusion that ocean-only experiments cannot be used to calibrate ocean parameters if the emergent-coupled metrics are included in the HM procedure, seems to be robust. For example, tuning sea-ice quantities in IPSL-CM6A-LR had to be done in coupled mode (Mignot et al., 2021). This advocates for running OMIP experiments to calibrate ocean quantities related to the ocean interior dynamics, and calibrating upper-ocean variables, including sea-ice, in coupled mode.

A second aspect of objective calibration that we consider, is the independence of metrics. As there are many measures of validation, objective methods can rapidly become expensive if they are all to be taken into account. However there is considerable covariance between metrics. We explore here the use of PCA to find a minimal optimal set of metrics that capture the variance across metric space, similar to the approach in Baker et al. (2015). Even in the case of L96 model with a relatively small number of metrics as compared to a climate model, it appears efficient to reduce dimensionality of the metrics, hence we strongly recommend to proceed similarly with climate models, in line with Salter et al. (2019). Another approach that can be useful consist of selecting a small number of metrics that are directly relatable (i.e., understandable by the modeler) and for which there exists good observations with known uncertainties, these can prove valuable as a sanity check. It is important to mention here a limitation of this study which is the idealized setting of the experiments. While our use of PCA was successful in all the experiments, Salter et al. (2019) warn from a usual risk coming with the use of PCA in HM studies where misspecification of model discrepancy can lead the HM algorithm to yield an empty NROY. They called it the terminal case and present a methodology for defining calibration-optimal bases that help avoid this issue.

Finally, a benefit of HM is that it can capture equally plausible but nonunique parameter regimes that arise when learning from GCM climate statistics. This appears clearly when applying HM in AMIP-style experiments of L96 model: HM procedure identifies a region of parameter space within which realizations of the model would likely result in comparable solutions to the "ground truth." The non-uniqueness of parameters, for a given model code and specific metrics, is an emergent property of GCMs that must be discussed in the context of CMIP exercises.

5. Conclusion and Future Work

We find that the L96 system can indeed shed light on certain aspects of the calibration of coupled models with two intrinsic timescales, providing lessons for the use of HM or other objective calibration methods in GCM development. The key findings of this paper can be summarized as follows.

First, we find that the HM method does indeed allow calibration of models close to optimal values, with minimal errors relative to the true values of $\{F, h, c, b\}$. Simulations with the tuned values of these parameters produce simulations whose "climate"—time-averaged metrics—is statistically close to the initial simulation considered as ground truth. Three waves of tuning efficiently reduces the NROY space to a fraction $\mathcal{O}\left(10^{-3} - 10^{-4}\right)$ of the prior volume of parameter space to be explored.

In the meanwhile, we show that we can reduce the space of metrics considerably by finding the eigenfunctions in metric space that explain most of the variance. We conducted preliminary experiments using nonlinear PCA and ML-based methods of dimensionality reduction (autoencoders) and did not find much benefit relative to standard PCA, at least in the L96 system. We cannot infer from this that nonlinear PCA might not be of use for reducing the dimensions of the space of metrics in a real GCM; it has been used for related purposes in other contexts, such as estimating the similarity of climate across simulations (Baker et al., 2015).

Second, we illustrate how much the number of waves hence explicit simulations of the L96 model can be drastically reduced if physical expertise is introduced prior to the initiation of HM algorithm. Additionally, as shown in Appendix B, the remaining NROY includes a smaller number of clusters, hence plausible sets of parameters to mimic ground truth. This advocates to merging as much physical expertise as possible when setting the prior intervals for each parameters. It's important to underline here that the goal of HM here is not to find the true value of parameters used for the initial experiment. Rather, it is to find acceptable values that produce a similar climate. Hence it has to be verified that each cluster from the final NROY is fed into a simulation, from which the climate is compared to the ground truth.

Third, we demonstrate that conducting independent component tuning (AMIP and OMIP) is indeed capable of introducing compensating errors into the coupled system, due to metrics emerging from the coupling between the slow and the fast components, such as sea-ice for example, The compensation was revealed by the empty NROY resulting from the first wave of the OMIP experiment with all the metrics. Additionally, we can sometimes find a non-compact NROY with two "clusters" with different values of two parameters (Figure 6), that yield equally satisfying metrics (Figure B1). Here, physical insight might play a key role in deciding which NROY cluster to select. One should then be careful when using clustering techniques, K-means has its shortcomings such as the lack of flexibility in cluster shape. For complex (e.g., non-elliptic) clusters, DBSCAN or other density-based

clustering techniques can be explored, but then the idea of a representative centroid of the cluster is not relevant anymore.

An issue with a potential use of HM for real coupled climate models is the difficulty, or even the impossibility, of running hundreds of high cost simulations. Active learning techniques can help overcome the greedy sampling done in the iterative refocussing step. It resorts to choosing the next design point by solving an optimization problem (Craig et al., 1997), recent work by Garbuno-Inigo et al. (2020) delves into this question and presents three active learning criteria. We believe that advancements in this direction might help lower the cost of a global coupled model HM.

As a final word, as we learned many lessons from applying HM to L96 model in the context of climate model tuning, it appears clearly that HM should not be seen as a fully automatic stand-alone solution for climate model tuning. Rather, it should be seen as a semi-automatic tool that, combined with physical expertise on climate model sensitivities, can provide efficient reduction of parameter search space. Also, as compared to manual tuning, it offers a unique opportunity to reduce the number of climate model simulations yet exploring a larger diversity of parameter sets, which helps transitioning climate modeling to a more sustainable science. Hence we encourage climate modelers to start incorporating it in the tuning road-map of their models, as a first step toward quantifying the overall uncertainty of climate projections.

Appendix A: Visualizing NROY Plots

In order to visualize the NROY space, two types of information are usually considered in HM literature: *minimum implausibility* and *optical depth*. An example using 3 tunable parameters (a,b,c) is shown in Figure A1, where the minimum implausibility is represented below the diagonal of the NROY "matrix" while the optical

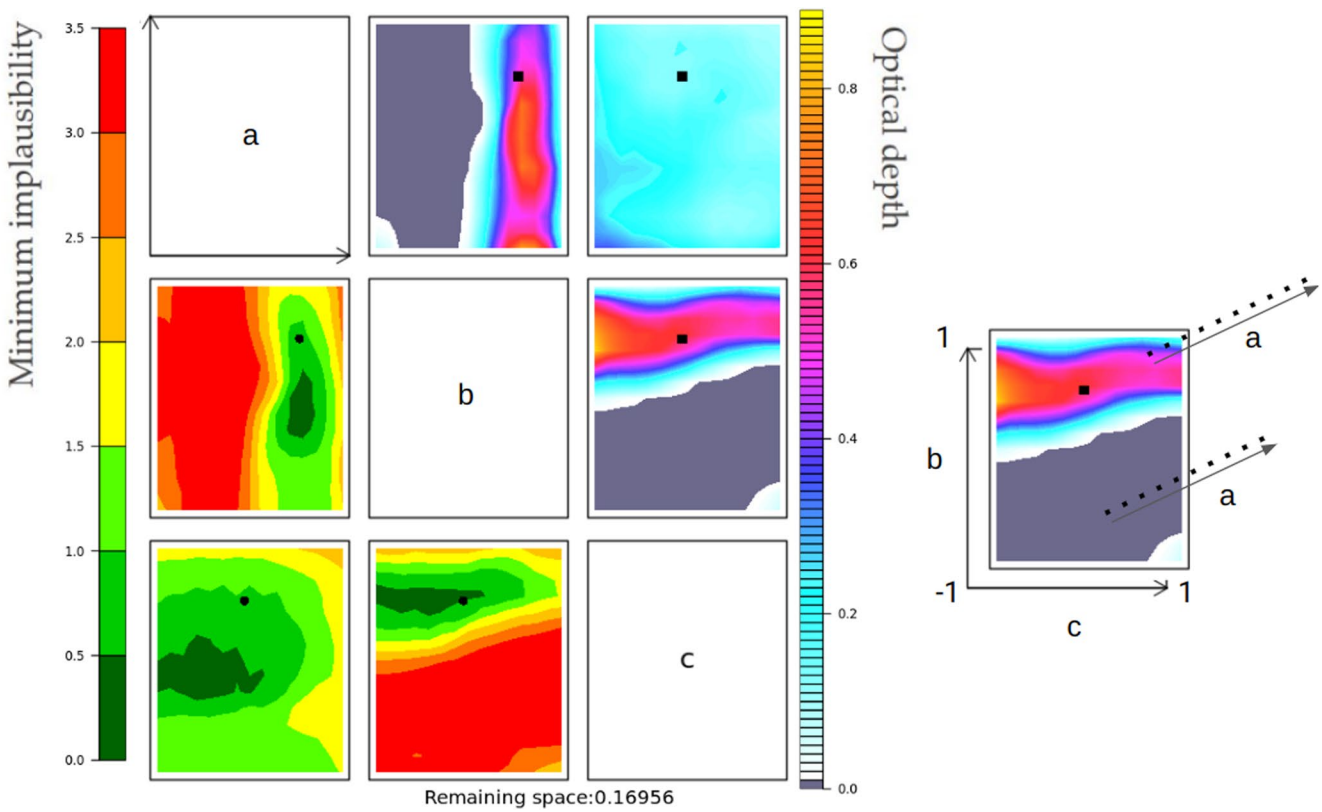

Figure A1. Not ruled out yet plots for the theoretical case of applying History Matching to tune a 3-parameter model. Colors associated to the left-hand-side bar indicate the minimum implausibility, while those associated to the right-hand-side bar show the optical depth, for every grid point of each 2-dimensional subplot. Subset to the right hand side provides further details on how to read each subplot.

depth is shown above it. Before explaining these two information in detail, we note that given the difficulty of visualizing a d–dimensional space with $d > 3$, the NROY space visualization is done using a composition of $\frac{d(d-1)}{2}$ two-dimensional subplots.

Let us consider one subplot such as the one shown to the right of Figure A1. The ordinate and abscissa represent respectively b and c parameter intervals projected into [–1, 1], a normalization that is done to help training the GP. We remind the reader that the NROY plots are obtained at test time (in ML terminology), rather than at training time. In fact, we apply our trained GP to a large variety of configurations, usually in the order of hundreds of thousands of parameter sets (in this work we initially use one million samples from a random LHS of the search space for parameters). Since the 2-dimensional subplots are restricted to combinations of 2 variables, each subplot is divided into bins in the 2 directions of the plot. In the aforementioned example, each bin has a restricted space for b and c but contains all the possible values of a. The latter group of configurations is then qualified by the minimum implausibility (i.e., the minimum implausibility amongst all configurations of that group) and the optical depth, that is the fraction of configurations with implausibility smaller than the pre-defined threshold. The rationale behind the minimum implausibility is that if the minimum is already higher than the threshold considered for implausibility (here 3), then whatever the choice for the other parameter(s), the metrics would be far from observations.

For easier visualization, the orientation of all subplots follow the arrows of the top-left one. The black dots in each subplot represent the ground-truth configuration, if already known.

Appendix B: K-Means Clustering Results

B1. Incorporating Domain Expertise in History Matching Experiment

We apply K-means on the final NROY (Figure 5) and find 3 clusters whose centers also belong to the final NROY (Table B1). In this case, all configurations produce metrics whose distributions are similar to observations (Figure C2). Interestingly, the remaining 3 configurations have median KL-div that are lower than 0.09 which is the value for the second best configuration in the very first experiment shown in this paper. This confirms that narrowing the prior helps HM reducing faster the search space and leads to better configurations.

B2. AMIP-Style Experiment

The K-means algorithms identifies two solutions, which have a very low KL-div (Table B2). We conclude that both sets of parameters can be considered as acceptable configurations. Indeed, running simulations with these two sets of parameters yields distributions of metrics that are very close to observations, equally (Figure B1). Table B2 shows also that values of both h and b are identified close to the "ground truth," whilst the value of c seems to be almost unimportant. The very lowest values of c are slightly less plausible, but anything between 5 and 20 looks to be acceptable.

Table B1
Same as Table 3 for the History Matching Experiment With Domain Expertise

	h	F	c	b	KL-div
1	1.05	9.87	9.06	10.52	0.03
2	0.81	10.06	11.46	9.30	0.05
3	1.11	12.13	17.10	10.48	0.07

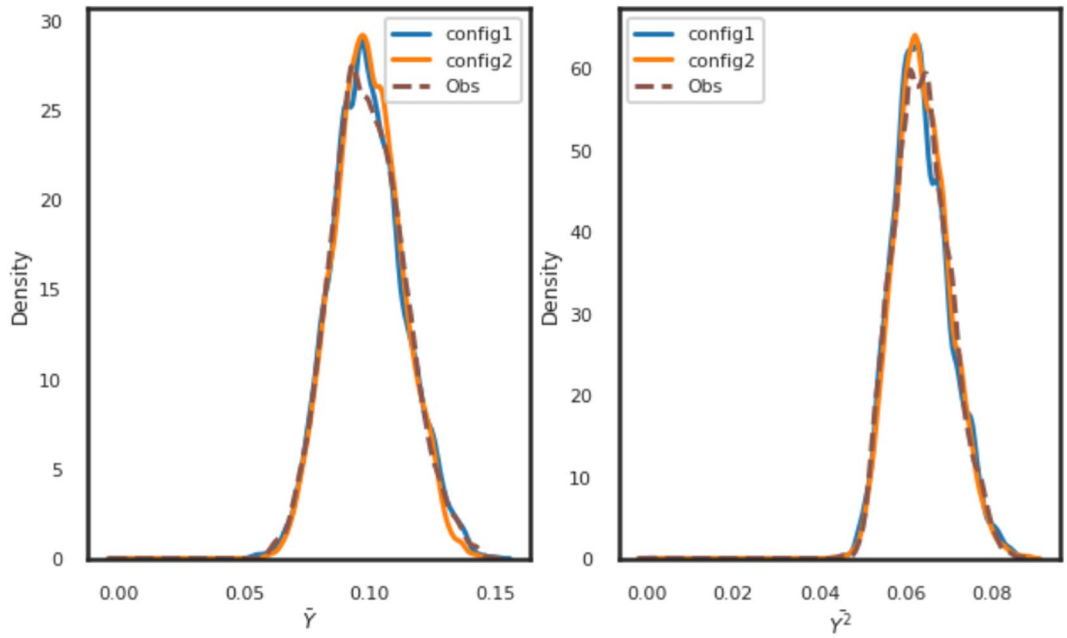

Figure B1. Histograms of metrics on the final 2 configurations identified by AMIP-style History Matching, and the observed metrics.

Table B2
Same as Table 3 for the AMIP-Style History Matching Experiment

	h	c	b	KL-div
1	1.00	10.26	10.03	0.0018
2	1.05	16.98	10.41	0.0024

Appendix C: Metrics for the Resulting Configurations Found by the K-Means

C1. Experiment 1: Tuning L96 With Unphysical Prior

See appendix Figure C1

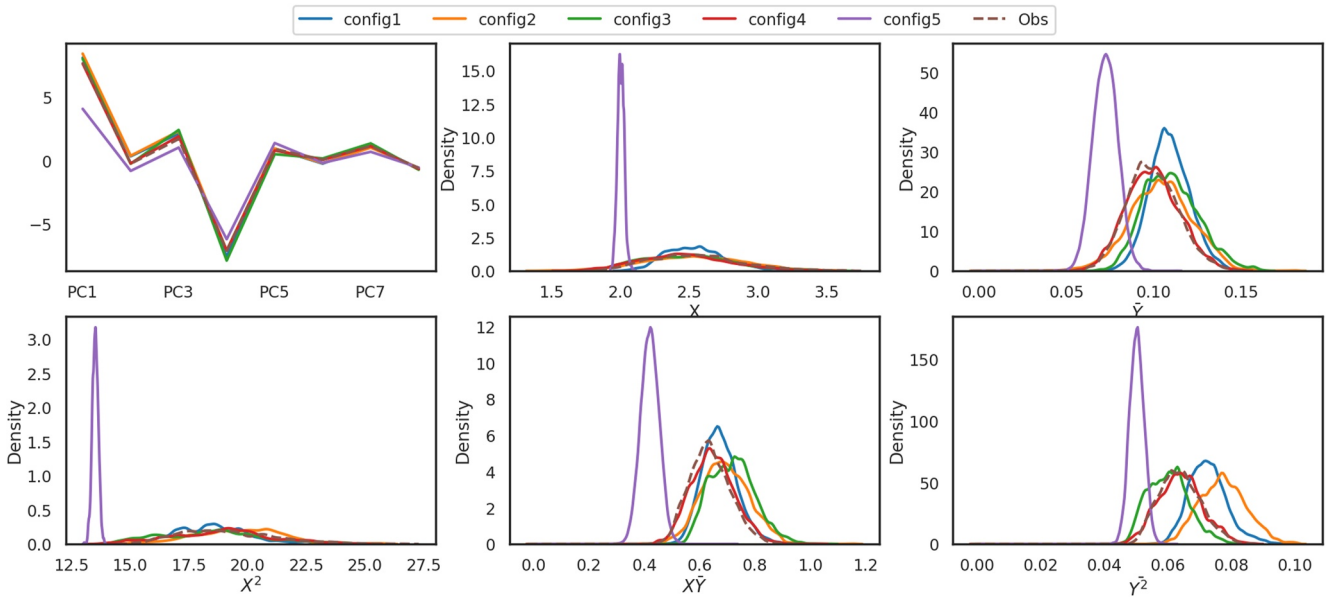

Figure C1. Histograms of the metrics on the configurations (straight lines) and the observed metrics (dashed line).

C2. Experiment 2: Tuning L96 With Physical Prior

See appendix Figure C2

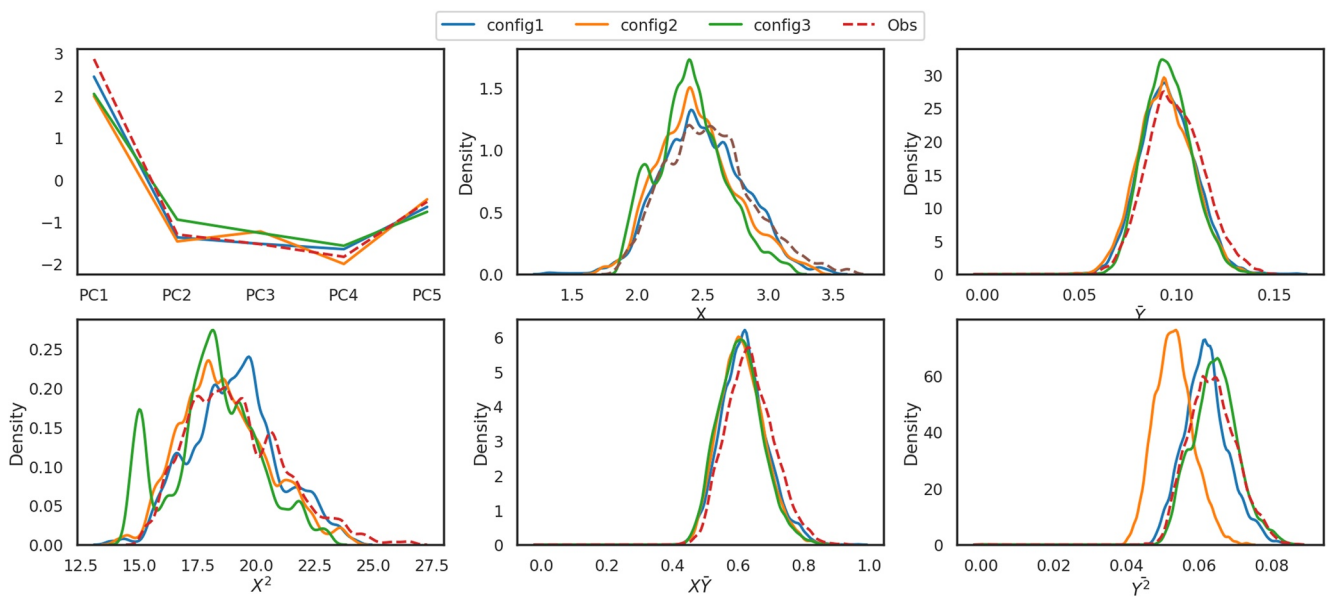

Figure C2. Histograms of the metrics on the configurations (straight lines) and the observed metrics (dashed line).

Appendix D: Silhouette Scores

Presented in this appendix are silhouette scores used for deciding the number of clusters in the K-means clustering Figure D1.

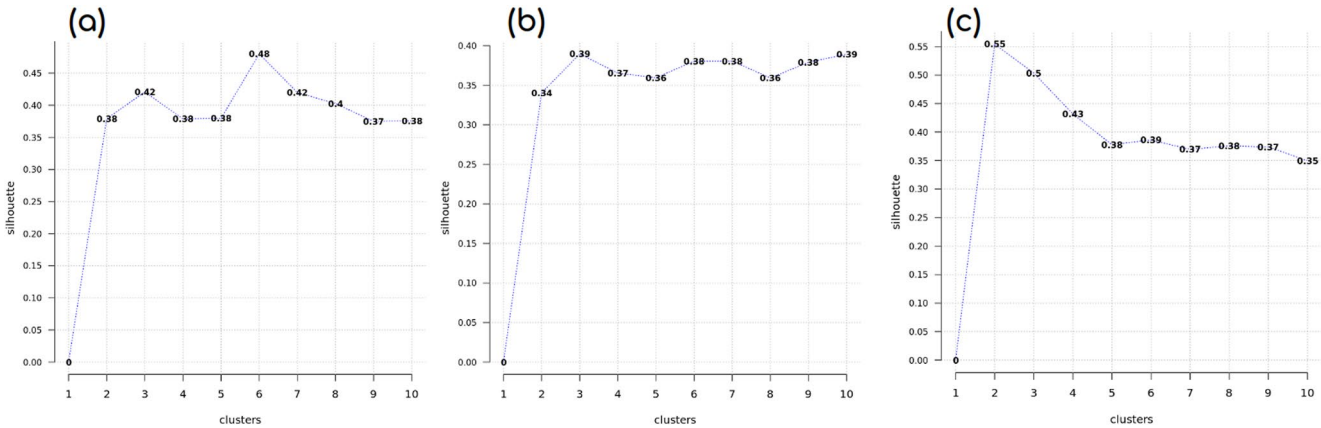

Figure D1. (a) L96 tuning experiment without physical prior: silhouette score, here optimal number of classes is 6 (b) L96 tuning experiment with physical prior: optimal number of classes is 3 (c) AMIP experiment: optimal number of classes is 2.

Data Availability Statement

Jupyter notebooks (Python and R) developed for this work are open sourced and can be used to generate the datasets and figures used in this paper. They can also be minimally modified in order to be used for another simulation model for further exploration. They can be accessed from the Github repository https://github.com/HRMES-MOPGA/L96HistoryMatching and are preserved at Zenodo https://doi.org/10.5281/zenodo.7384270 (Lguensat, 2022).

Acknowledgments

The authors would like to thank Frédéric Hourdin, Fleur Couvreux, Daniel Williamson, Najda Villefranque and Martin Vancoppenolle for the helpful discussions. The authors are also grateful for the constructive comments and valuable suggestions provided by two anonymous reviewers that highly improved the quality of this paper. This research was supported by the "Agence Nationale de la Recherche" through the HRMES ANR-17-MPGA-0010 project. This work was granted access to the HPC/AI resources of IDRIS under the allocation A0120107451 made by GENCI.

References

Adcroft, A., Anderson, W., Balaji, V., Blanton, C., Bushuk, M., Dufour, C. O., et al. (2019). The GFDL global ocean and sea ice model OM4.0: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, *1*(10), 1–45. https://doi.org/10.1029/2019MS001726

Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., et al. (2015). Bayesian history matching of complex infectious disease models using emulation: A tutorial and a case study on HIV in Uganda. *PLoS Computational Biology*, *11*(1), e1003968. https://doi.org/10.1371/journal.pcbi.1003968

Baker, A. H., Hammerling, D. M., Levy, M. N., Xu, H., Dennis, J. M., Eaton, B. E., et al. (2015). A new ensemble-based consistency test for the Community Earth System Model (pyCECT v1.0). *Geoscientific Model Development*, *8*(9), 2829–2840. https://doi.org/10.5194/gmd-8-2829-2015

Balaji, V. (2021). Climbing down Charney's ladder: Machine learning and the post-Dennard era of computational climate science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, *379*(2194), 20200085. https://doi.org/10.1098/rsta.2020.0085

Balaji, V., Couvreux, F., Deshayes, J., Gautrais, J., Hourdin, F., & Rio, C. (2022). Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, *119*(47), e2202075119. https://doi.org/10.1073/pnas.2202075119

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019MS002010

Bower, R. G., Goldstein, M., & Vernon, I. (2010). Galaxy formation: A Bayesian uncertainty analysis. *Bayesian Analysis*, *5*(4), 619–669. https://doi.org/10.1214/10-ba524

Chapman, W. L., Welch, W. J., Bowman, K. P., Sacks, J., & Walsh, J. E. (1994). Arctic sea ice variability: Model sensitivities and a multidecadal simulation. *Journal of Geophysical Research*, *99*(C1), 919–935. https://doi.org/10.1029/93jc02564

Chattopadhyay, A., Subel, A., & Hassanzadeh, P. (2020). Data-driven super-parameterization using deep learning: Experimentation with multiscale Lorenz 96 systems and transfer learning. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002084. https://doi.org/10.1029/2020ms002084

Christensen, H. M., & Berner, J. (2019). From reliable weather forecasts to skilful climate response: A dynamical systems approach. *Quarterly Journal of the Royal Meteorological Society*, *145*(720), 1052–1069. https://doi.org/10.1002/qj.3476

Cleary, E., Garbuno-Inigo, A., Lan, S., Schneider, T., & Stuart, A. M. (2021). Calibrate, emulate, sample. *Journal of Computational Physics*, *424*, 109716. https://doi.org/10.1016/j.jcp.2020.109716

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., et al. (2020). Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement (preprint). In *Earth and space science open archive section: Atmospheric sciences*. https://doi.org/10.1002/essoar.10503597.1

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. In *Case studies in Bayesian statistics* (pp. 37–93). Springer.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Dunbar, O. R. A., Garbuno-Inigo, A., Schneider, T., & Stuart, A. M. (2020). Calibration and uncertainty quantification of convective parameters in an idealized GCM. *Journal of Advances in Modeling Earth Systems*, *13*, e2020MS002454. https://doi.org/10.1029/2020MS002454

Edwards, N. R., Cameron, D., & Rougier, J. (2011). Precalibrating an intermediate complexity climate model. *Climate Dynamics*, *37*(7), 1469–1482. https://doi.org/10.1007/s00382-010-0921-0

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896. https://doi.org/10.1029/2019ms001896

Garbuno-Inigo, A., DiazDelaO, F. A., & Zuev, K. M. (2020). History matching with probabilistic emulators and active learning. arXiv preprint arXiv:2004.07878.

Gates, W. L. (1992). AMIP: The atmospheric model intercomparison project. *Bulletin of the American Meteorological Society*, *73*(12), 1962–1970. https://doi.org/10.1175/1520-0477(1992)073<1962:atamip>2.0.co;2

Gramacy, R. B. (2020). *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC.

Griffies, S. M., Danabasoglu, G., Durack, P. J., Adcroft, A. J., Balaji, V., Böning, C. W., et al. (2016). OMIP contribution to CMIP6: Experimental and diagnostic protocol for the physical component of the ocean model intercomparison project. *Geoscientific Model Development*, *9*(9), 3231–3296. https://doi.org/10.5194/gmd-9-3231-2016

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., et al. (2019). Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3691–3727. https://doi.org/10.1029/2019MS001829

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589–602. https://doi.org/10.1175/bams-d-15-00135.1

Hourdin, F., Rio, C., Grandpeix, J.-Y., Madeleine, J.-B., Cheruy, F., Rochetin, N., et al. (2020). LMDZ6A: The atmospheric component of the IPSL climate model with improved and better tuned physics. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS001892. https://doi.org/10.1029/2019MS001892

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., et al. (2020). Process-based climate model development harnessing machine learning: II. Model calibration from single column to global (preprint). In *Earth and space science open archive section: Atmospheric sciences*. https://doi.org/10.1002/essoar.10503845.1

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B*, *63*(3), 425–464. https://doi.org/10.1111/1467-9868.00294

Lguensat, R. (2022). HRMES-MOPGA/L96HistoryMatching: History matching of L96 (v0.2.0). [Software]. Zenodo. https://doi.org/10.5281/zenodo.7384270

Lguensat, R., Tandeo, P., Ailliot, P., Pulido, M., & Fablet, R. (2017). The analog data assimilation. *Monthly Weather Review*, *145*(10), 4093–4107. https://doi.org/10.1175/mwr-d-16-0441.1

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, *51*(4), 366–376. https://doi.org/10.1198/tech.2009.08040

Lorenz, E. N. (1996). Predictability: A problem partly solved. *Proc. seminar on predictability*, *1*.

Lorenz, E. N., & Emanuel, K. A. (1998). Optimal sites for supplementary weather observations: Simulation with a small model. *Journal of the Atmospheric Sciences*, *55*(3), 399–414. https://doi.org/10.1175/1520-0469(1998)055<0399:osfswo>2.0.co;2

Manabe, S., & Bryan, K. (1969). Climate calculations with a combined ocean-atmosphere model. *Journal of the Atmospheric Sciences*, *26*(4), 786–789. https://doi.org/10.1175/1520-0469(1969)026<0786:ccwaco>2.0.co;2

Manabe, S., & Wetherald, R. T. (1975). The effects of doubling the CO_2 concentration on the climate of a general circulation model. *Journal of the Atmospheric Sciences*, *32*(1), 3–15. https://doi.org/10.1175/1520-0469(1975)032<0003:teodtc>2.0.co;2

McKay, M. D., Beckman, R. J., & Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *42*(1), 55–61. https://doi.org/10.1080/00401706.2000.10485979

Mignot, J., Hourdin, F., Deshayes, J., Boucher, O., Gastineau, G., Musat, I., et al. (2021). The tuning strategy of IPSL-CM6A-LR. *Journal of Advances in Modeling Earth Systems*, *13*(5), e2020MS002340. https://doi.org/10.1029/2020MS002340

Morris, M. D., & Mitchell, T. J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, *43*(3), 381–402. https://doi.org/10.1016/0378-3758(94)00035-t

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Platzman, G. W. (1979). The ENIAC computations of 1950—Gateway to numerical weather prediction. *Bulletin of the American Meteorological Society*, *60*(4), 302–312. https://doi.org/10.1175/1520-0477(1979)060<0302:tecotn>2.0.co;2

Rasp, S. (2019). Online learning as a way to tackle instabilities and biases in neural network parameterizations. *Geoscientific Model Development*, *13*(5), 2185–2196. https://doi.org/10.5194/gmd-13-2185-2020

Sacks, J., Welch, W. J., Mitchell, T. J., & Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, *4*(4), 409–423. https://doi.org/10.1214/ss/1177012413

Salter, J. M., & Williamson, D. (2016). A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. *Environmetrics*, *27*(8), 507–523. https://doi.org/10.1002/env.2405

Salter, J. M., Williamson, D. B., Scinocca, J., & Kharin, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *Journal of the American Statistical Association*, *114*(528), 1800–1814. https://doi.org/10.1080/01621459.2018.1514306

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., et al. (2017). Practice and philosophy of climate model tuning across six us modeling centers. *Geoscientific Model Development*, *10*(9), 3207–3223. https://doi.org/10.5194/gmd-10-3207-2017

Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *1–1*(24), 12396–12417. https://doi.org/10.1002/2017GL076101

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*(1), 3–5. https://doi.org/10.1038/nclimate3190

Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100.

Smagorinsky, J. (1983). The beginnings of numerical weather prediction and general circulation modeling: Early recollections. In B. Saltzman (Ed.), *Advances in geophysics* (Vol. 25, pp. 3–37). Elsevier. https://doi.org/10.1016/S0065-2687(08)60170-3

Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., & Balaji, V. (2021). Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environmental Research Letters*, *16*(7), 073008. https://doi.org/10.1088/1748-9326/ac0eb0

Steponavičė, I., Shirazi-Manesh, M., Hyndman, R. J., Smith-Miles, K., & Villanova, L. (2016). On sampling methods for costly multi-objective black-box optimization. In *Advances in stochastic and deterministic global optimization* (pp. 273–296). Springer.

Villefranque, N., Blanco, S., Couvreux, F., Fournier, R., Gautrais, J., Hogan, R. J., et al. (2021). Process-based climate model development harnessing machine learning: III. The representation of cumulus geometry and their 3D radiative effects. *Journal of Advances in Modeling Earth Systems*, *13*(4), e2020MS002423. https://doi.org/10.1029/2020ms002423

Wilkinson, R. D. (2010). Bayesian calibration of expensive multivariate computer experiments. In *Large-scale inverse problems and quantification of uncertainty* (pp. 195–215).

Williamson, D., Blaker, A. T., Hampton, C., & Salter, J. (2015). Identifying and removing structural biases in climate models with history matching. *Climate Dynamics*, *45*(5), 1299–1324. https://doi.org/10.1007/s00382-014-2378-z

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, *41*(7), 1703–1729. https://doi.org/10.1007/s00382-013-1896-4

Williamson, D. B., Blaker, A. T., & Sinha, B. (2017). Tuning without over-tuning: Parametric uncertainty quantification for the nemo ocean model. *Geoscientific Model Development*, *10*(4), 1789–1816. https://doi.org/10.5194/gmd-10-1789-2017

Yeh, T., Uvieghara, T., Jennings, J., Chen, C., Alpak, F., & Tendo, F. (2016). A practical workflow for probabilistic history matching and forecast uncertainty quantification: Application to a Deepwater West Africa reservoir. In *Spe annual technical conference and exhibition*.

Zhao, M., Golaz, J., Held, I. M., Guo, H., Balaji, V., Benson, R., et al. (2018a). The GFDL global atmosphere and land model AM4.0/LM4.0: 1. Simulation characteristics with prescribed SSTs. *Journal of Advances in Modeling Earth Systems*, *10*(3), 691–734. https://doi.org/10.1002/2017MS001208

Zhao, M., Golaz, J., Held, I. M., Guo, H., Balaji, V., Benson, R., et al. (2018b). The GFDL global atmosphere and land model AM4.0/LM4.0: 2. Model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, *10*(3), 735–769. Early Online Release(0). https://doi.org/10.1002/2017MS001209