JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2020MS002225

This article is a companion to Couvreux et al. (2020), https://doi.org/10.1029/2020MS002217.

Key Points:

- We use an automatic tool to calibrate the parameterizations of a global climate model
- We show the benefit for global climate tuning of a preconditioning in single column mode
- We show how this approach allows us to revisit a parameterization of boundary layer convection

Correspondence to:
F. Hourdin,
frederic.hourdin@lmd.ipsl.fr

Citation:
Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., et al. (2021). 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. https://doi.org/10.1029/2020MS002225

Received 26 JUN 2020
Accepted 12 OCT 2020

Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global

Frédéric Hourdin[1], Daniel Williamson[2,3], Catherine Rio[4], Fleur Couvreux[4], Romain Roehrig[4], Najda Villefranque[4], Ionela Musat[1], Laurent Fairhead[1], F. Binta Diallo[1], and Victoria Volodina[3]

[1]LMD-IPSL, Sorbonne-Universités, CNRS, Paris, France, [2]Department of Mathematical Sciences, University of Exeter, Exeter, UK, [3]The Alan Turing Institute, British Library, London, UK, [4]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Abstract We demonstrate a new approach for climate model tuning in a realistic situation. Our approach, the mathematical foundations and technical details of which are given in Part I, systematically uses a single-column configuration of a global atmospheric model on test cases for which reference large-eddy-simulations are available. The space of free parameters is sampled running the single-column model from which metrics are estimated in the full parameter space using emulators. The parameter space is then reduced by retaining only the values for which the emulated metrics match large eddy simulations within a given tolerance to error. The approach is applied to the 6A version of the LMDZ model which results from a long investment in the development of physics parameterizations and by-hand tuning. The boundary layer is revisited by increasing the vertical resolution and varying parameters that were kept fixed so far, which improves the representation of clouds at process scale. The approach allows us to automatically reach a tuning of this modified configuration as good as that of the 6A version. We show how this approach helps accelerate the introduction of new parameterizations. It allows us to maintain the physical foundations of the model and to ensure that the improvement of global metrics is obtained for a reasonable behavior at process level, reducing the risk of error compensations that may arise from over-fitting some climate metrics. That is, we get things right for the right reasons.

Plain Language Summary In view of the importance of global numerical models for the anticipation of future climate changes, their improvement is often considered too slow. We present a new approach that we believe could boost model improvement significantly. This approach promotes the use of machine learning techniques developed by the "uncertainty quantification" community for the adjustment of model free parameters, or tuning. These techniques are applied to physics improvement at process scale, represented through parameterizations. In this approach, the tuning of the global atmospheric model is preconditioned by calibration of the model free parameters on a series of well documented cloud scenes for which explicit very high resolution simulations are available. We demonstrate on a real example how the reduction of the parameter space with this approach allows us to save a large amount of computer resources and detract from the long and tedious by-hand phase of model tuning. By automating part of the tuning process, the approach enables climate modeler expertize to focus on understanding and improving the model physics through parameterization.

1. Introduction

Given the high expectation on global circulation models, both for numerical weather prediction and anticipation of climate change, their improvement is often considered too slow. Among the main reasons, one finds the poor job done by convective parameterizations in summarizing convective motions that cannot be resolved with grid meshes larger than 300 m for boundary-layer convection, or 2 km for deep convection. A parameterization can be seen as a mathematical function $\mathcal{P}_p$ that expresses the effect on the model state variables $\boldsymbol{x}$ of the collective behavior of unresolved processes, which at the end appears as a source term $S_{\boldsymbol{x}} = \mathcal{P}_p(\boldsymbol{x}, \lambda_p)$ in the discretized form of the fluid dynamic equations. The different parameterizations are often connected to each other. For instance, a first one computes convection from the vertical profile of

potential temperature and humidity, then a second one deduces the fractional cover of clouds and cloud water content, which are finally integrated in a radiative calculation (third parameterization) to provide a vertical heating profile. Each parameterization depends on a set of free parameters $\boldsymbol{\lambda}_p$, some of which have a physical meaning (e.g., fall speed of ice crystals), some others resulting from the simplifications inherent to any parameterization (e.g., representing an ensemble of plumes by a single plume for example). Convective and cloud parameterizations are often developed in a single column model (SCM) framework by comparison with large eddy simulations (LES) of the same atmospheric column, in which convective motions are explicitly resolved. This SCM/LES comparison is used both to inspire parameterization development and to choose, calibrate or "tune" the model free parameters $\boldsymbol{\lambda}_p$ at process level. Once integrated in operational models, those parameterizations are active in each atmospheric column of the model, influencing both the global radiation budget and the large-scale circulation.

The development of a reference configuration of a climate model, as those involved in the Coupled Model Intercomparison Program (Taylor et al., 2012, CMIP), requires an intense phase of adjustment—including grid choice, bug corrections, activation of some parameterizations or code modifications—in which the tuning or calibration of free parameters is key (Mauritsen et al., 2012; Schmidt et al., 2014). A survey on climate model tuning revealed rather standard priorities, which consist of targeting the radiative forcing of the atmospheric circulation, thereby using model free parameters that most affect radiation, that is, cloud parameters (Hourdin et al., 2017). The complexity of the tuning process, given the large number of free parameters, the large number of possible targets and the computational cost of global climate simulations, probably partly explain the slowness of climate model improvements. One promising avenue is the use of more automatic and objective methods for tuning. However, although specific applications of such methods have been proposed for numerical weather forecast (Duan et al., 2017) or regional climate modeling (Bellprat et al., 2012), their direct use for global climate models remains challenging and most CMIP-class models are indeed hand-tuned so far. Typically, the tuning phase of the IPSL coupled model configuration for CMIP6 (IPSL-CM6A-LR) took more than 2 years, with repeated tuning phases targeting improvement of the radiative forcing of the circulation: global radiation, decomposed in terms of short-wave (SW) and long-wave (LW), clear-sky and cloud radiative effect (CRE), and some spatial variations of those fluxes like contrasts between mid-latitude and tropics, or between convective and subsiding regimes in the tropics. Such a tuning was done in practice each time a new version of the coupled model with significant changes was proposed. In total, 15 successive versions were tuned this way. For each version, systematic sensitivity experiments to 3–10 parameters were done with the stand-alone-atmospheric model forced by imposed sea surface temperature (SST) on a couple of years, changing the parameters one by one. Then diagnostics were computed and, by trial and error, a new radiative tuning was proposed and tested. Each of the 15 versions of the global model typically needed 1–5 iterations of this tedious sensitivity analysis. This later approach is done only by local perturbation around the previous tuning and explores independently the dependency to each individual parameter, hiding any compensating effects between them. During all of these processes, a series of SCM test cases were run and compared with LES in order to ensure that the model tuning was not pushed too far, at the risk of deteriorating the model behavior at process level.

To help accelerate this phase of model tuning and tackle model development and tuning together, Hourdin et al. (2017) identified at least three different levels of calibration in a model development: a first calibration at the level of individual parameterizations, then a calibration of each component of the Earth system model and eventually a calibration of the full Earth system model. In line with this proposal, we advocate in the first part of this paper (Couvreux et al., 2020, referred to as Part I hereafter) that a systematic comparison between LES and SCM simulations on a series of benchmark cases, making use of state-of-the-art machine learning techniques issued from the Uncertainty Quantification community may help accelerate model development and tuning at process scale. The history matching approach, used in this systematic comparison, consists in reducing iteratively the space of acceptable parameters by conserving parameter vectors for which the SCM results match LES values to a given tolerance error. The parameter space is explored using an "emulator," a statistical tool capable of estimating the value of some SCM metrics (with uncertainty) in the full parameter space, based on sampling with the true SCM.

Part I presents the rationale for the proposed approach and places it in the context of other approaches for model calibration and climate model tuning. The review of existing literature on the subject is not repeated

here. Part I also provides the mathematical basis and technical details for the particular method used for calibration, and therefore only the information necessary to understand the results is repeated here. The objective of this second part is to demonstrate how this framework can be used to speed up the process of model development, from the inspiration of new process-based parameterizations to the full development of a 3D General Circulation Model (GCM). Beyond streamlining and accelerating the tuning process, and helping to avoid some of the compensating errors that can result from over-adjusting the climate metrics, we illustrate, using state-of-the-art boundary layer and cloud parameterizations, how the method can inform us about the functioning of the climate model and the link between its climate performance and its physical content. We revisit more specifically choices made during the development phase of the so called "thermal plume model" (Hourdin et al., 2002), a parameterization of the convective boundary-layer transport and associated cumulus clouds (Rio & Hourdin, 2008), based on a mass flux representation of a mean thermal plume coupled to a bi-modal representation of the subgrid scale distribution of the saturation deficit (Jam et al., 2013). This thermal plume model was developed over a number of years using LES to inspire new pieces of parameterizations, to assess the proposed formulations and to propose acceptable values of the free parameters. Successive versions of this thermal plume model were introduced in the global LMDZ atmospheric model, giving rise in particular to the recent LMDZ 6A version (Hourdin et al., 2019, 2020a; 2020b) used as the atmospheric component of the Institut Pierre Simon Laplace Coupled Model, IPSL-CM6A-LR, which participated to the recent sixth phase of CMIP (CMIP6). With the increasing complexity of this parameterization suite, it became clear that further sophistication leading to demonstrable improvement was not possible without somewhat automatic tools to explore the parametric dependency of the results. In order to prove that a new parameterization suite $\mathcal{P}_1(\boldsymbol{x},\boldsymbol{\lambda}_1)$ behaves better than an old version $\mathcal{P}_0(\boldsymbol{x},\boldsymbol{\lambda}_0)$, one should show in principle that there exists at least one vector $\boldsymbol{\lambda}_1$ for which $\mathcal{P}_1$ gives globally better results than $\mathcal{P}_0$, whatever the value retained for $\boldsymbol{\lambda}_0$.

In this study, we illustrate the deployment of a well-defined calibration strategy based on two steps. The first step consists of a process-oriented calibration of the free parameters using SCM/LES comparisons combined with the "*High-Tune Explorer*" described in Part I (Couvreux et al., 2020). This SCM calibration is able to reduce the domain of acceptable values and this information is used in step 2 for the calibration of the global 3D configuration. A great advantage of history matching indeed is that it can be used to iteratively reduce the parameter space, taking new constraints into account. This saves important resources as the SCM/LES comparison is relatively computationally inexpensive, and does not require supercomputer time. With this new approach, we revisit here the parameter values involved in the formulations of lateral entrainment and detrainment that control the mass flux computation (Rio et al., 2010), and hence the convective transport as well as the bi-Gaussian cloud scheme (Jam et al., 2013).

After a description of the LMDZ model and cloud parameterizations in Section 2, we present a first illustration in Section 3, in which we revisit the calibration of three of the parameters systematically used for the 3D GCM tuning. They all concern the representation of boundary layer convection and clouds. We show that using systematic SCM/LES comparisons on a few contrasted test cases makes it possible to find a setting of the parameters very close to the one obtained after a long and tedious phase of manual tuning, demonstrating the capability of the tool in saving time and resources. In Section 4, we show an example of model retuning after some modifications are introduced in the model, here the increase of the vertical resolution in the first kilometers above surface. By doing this, we explore the impact of changing some key parameters of the mass-flux scheme, which were kept fixed so far, in view of the difficulty to explore a multi-dimensional space. Section 5 summarizes the main results and discusses the gain obtained from this revisiting of 15 years of model development.

2. Shallow Convection Parameterization in LMDZ

The representation of boundary layer convection, shallow cumulus and stratocumulus clouds is unified in the LMDZ model by using a combination of eddy diffusion and a mass flux scheme to parameterize the boundary layer transport. This approach is often referred to as an EDMF approach (see e.g., Köhler et al., 2011), for eddy-diffusivity and mass-flux. In LMDZ, the mass flux scheme is coupled to a bi-Gaussian representation of the sub-grid scale distribution of the saturation deficit, from which cloud

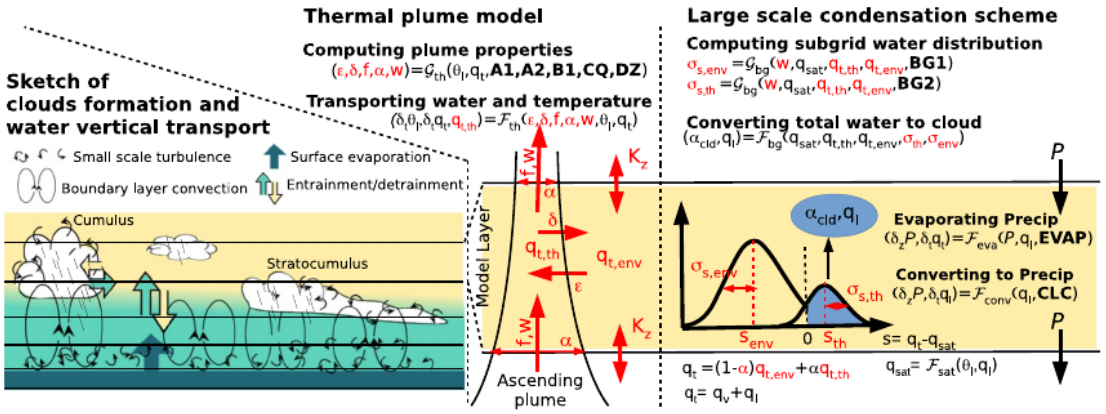


Figure 1. Sketch of the parameterizations and tuning parameters used in the present study. The sketch on the left-hand side presents the view of the boundary layer clouds and transport of water by boundary layer turbulence and convection, as well as the entrainment and detrainment at the boundaries of clouds and top of the boundary layer. These processes are represented in a model layer from the interplay between the thermal plume model (combining vertical diffusion with a mass flux scheme), a bi-gaussian representation of subrgid scale water distribution and the so-called "large scale" condensation scheme. The scheme internal variables are shown in red and the tuning parameters as bold fonts. $\delta_t = \delta t\partial_t$ is an increment over one-time step of a state variable and $\delta_z P$ the vertical variation of precipitation P over the depth of the layer. The complete formulas and notations are given in the text.

cover and condensed water are deduced. The mass flux scheme and bi-Gaussian scheme, the two targeted parameterizations of the parameter exploration presented in this study, are detailed hereafter. We identify the free parameters, which are used for the parametric exploration with bold font in the text. A sketch of the main elements of the parameterizations and associated free parameters is given in Figure 1.

2.1. The Thermal Plume Model

The "thermal plume model" under consideration in the present study summarizes the collective behavior of a population of thermal plumes (or cells, or rolls) through a unique bulk thermal plume. Each atmospheric column is divided into a mean ascending thermal plume of mass flux $f = \rho\alpha w_{th}$ (where ρ is the air density, α is the fractional cover and w_{th} is the vertical velocity of the plume), and a compensating subsidence in the environment of mass-flux $-f$. The value of a model state variable ψ within the thermal plume ψ_{th} is computed using the stationary plume conservation equation:

$$\frac{\partial f\psi_{th}}{\partial z} = e\psi - d\psi_{th} + \rho S_\psi \tag{1}$$

where e and d are the lateral entrainment and detrainment of air toward and away from the plumes (the quantity is assumed to enter the thermal plume with its large scale value ψ). For variables conserved by the convective transport, such as liquid potential temperature θ_l or total water q_t, the source term is set to $S_\psi \equiv 0$. The plume vertical velocity w_{th} is computed with the same equation with a source term that includes buoyancy and a drag term. The fraction of the horizontal surface covered by plumes at altitude z is then deduced as $\alpha = f/(\rho w_{th})$.

The total boundary layer vertical transport of ψ is

$$\overline{\rho w'\psi'} = f(\psi_{th} - \psi) - K_z\frac{\partial\psi}{\partial z}, \tag{2}$$

where $K_z = l_{\text{mix}}S(Ri)\sqrt{\text{TKE}}$ is the eddy diffusivity, l_{mix} being a turbulent mixing length and $S(Ri)$ a stability function that depends upon the local gradient Richardson number Ri. The turbulent kinetic energy TKE is integrated in time from a local prognostic equation, following Yamada (1983). The technical implementation details are given by Vignon et al. (2017). Given this framework, the mass flux part is entirely defined by the specification of e and d from which f is deduced from the continuity equation for the plume

$$\frac{\partial f}{\partial z} = e - d \tag{3}$$

In the original version of the thermal plume model (Hourdin et al., 2002) the plume is fed laterally by warm air from the surface boundary layer, with $e > 0$ when $\partial_z \theta_v > 0$ in the first unstable layers above the surface. Above this surface layer, entrainment is null and detrainment is viewed as a shedding due to lateral mixing. It consists in reducing the width of the thermal plume with height, compared to the width that would correspond to a conservative thermal plume ($\partial f/\partial z = 0$). Those formulations were inspired by physical considerations and tested a posteriori on a series of LES cases of dry convection proposed by Ayotte et al. (1996).

2.2. Entrainment and Detrainment Derived From LES Sampling

The subsequent versions of the entrainment and detrainment formulations were largely inspired and adjusted in the SCM/LES framework. In order to use LES to inspire the development of mass flux convective parameterizations, one has to identify and sample the thermal plumes in the LES, in a way that matches with the EDMF framework. The classical approach consists in applying a combination of thresholds on water vapor or condensed water in clouds, vertical wind or a virtual tracer emitted at the surface for that specific purpose (Couvreux et al., 2010). Once the plume region is identified, the plume vertical velocity, fractional cover and mass flux can be computed as well as the composite value $\psi_{\rm th}$ of any conserved quantity ψ inside the plume. Knowing f, ψ and $\psi_{\rm th}$, one can then invert the conservation equation of the mass flux (Equation 3) and ψ (Equation 1 with $S_\psi = 0$) to deduce e and d.

Such a sampling was used to estimate the vertical profiles of entrainment and detrainment in LES for standard cases of continental and marine cumulus (Rio et al., 2010). The analysis of the results showed that the entrainment was strong in regions of positive buoyancy, and that detrainment was dominating in regions of negative buoyancy of the plume. This would be the case for a plume with a value of $\rho\alpha$ that would not vary vertically (almost constant fractional cover), which would entrain air where it accelerates and detrain where it decelerates. From the LES sampling, it appears that the entrainment and detrainment values lie in between the plume obtained with the constant fractional cover approximation and a conservative plume ($\partial f/\partial z = 0, e = 0, d = 0$). A parameter **B1**, assumed to range between 0 and 1, was therefore included as a scaling factor of the entrainment and detrainment computed with the constant fractional cover approximation.

Like most convective parameterizations, we use a momentum equation which assumes that subplume turbulent fluctuations and non-hydrostatic pressure perturbations reduce buoyancy and act as a drag term proportional to entrainment (de Roode et al., 2012; Simpson & Wiggert, 1969). The plume vertical velocity $w_{\rm th}$ is obtained by solving Equation 1 for $\psi_{\rm th} = w_{\rm th}$ and $\psi = 0$, with a source term specified as $S_{w_{th}} = \mathbf{A1}\, B - \mathbf{A2}\, w_{th}^2$ where $B = g\,(\theta_{v,th} - \theta_v)/\theta_v$ is the buoyancy (θ_v being the virtual potential temperature) that accelerates the plume and the second term a drag effect, with $\mathbf{A1} = 2/3$ and $\mathbf{A2} = 0.002\ \mathrm{m}^{-1}$.

The entrainment rate $\epsilon = e/f$ depends on the plume buoyancy and vertical velocity:

$$\epsilon = \max\left[0., \frac{\mathbf{B1}}{1 + \mathbf{B1}}\left(\mathbf{A1}\frac{B}{w_{th}^2} - \mathbf{A2}\right)\right] \tag{4}$$

where $\mathbf{B1} = 0.9$, a value consistent with previous studies (Gregory, 2001; Siebert & Frank, 2003). The plume is mainly entraining in regions of positive buoyancy. It is the opposite for the detrainment rate $\delta = d/f$ which is favored in regions where buoyancy is negative, as suggested by observations (Bretherton & Smolarkiewicz, 1989). A satisfactory correlation is obtained between LES results and parameterization with the following definition of δ:

$$\delta = \max\left[0., -\frac{\mathbf{A1} \times \mathbf{B1}}{1 + \mathbf{B1}}\frac{B}{w_{th}^2} + \mathbf{CQ}\left(\frac{\Delta q_t / q_t}{(w_{th} / w_0)^2}\right)^D\right], \tag{5}$$

where Δq_t is the contrast in humidity between the plume and its environment, with **CQ** = 0.012 m^{-1} (the vertical velocity being normalized by $w_0 = 1$ m s^{-1}) and $D = 0.5$. The first term corresponds to the buoyancy contribution to the detrainment rate while the second term accounts for the fact that evaporation around the clouds can reinforce the negative buoyancy of extracted air parcels, a mechanism enhanced when Δq_t increases.

2.3. Modification for Stratocumulus Clouds

A recent modification of the scheme targeted the representation of stratocumulus clouds (Hourdin et al., 2019). Indeed, the previous version of the mass flux model was destroying stratocumulus clouds, by overshooting too far above the strong inversion at the stratocumulus cloud top.

Based on a combination of numerical and physical arguments, this deficiency was overcome by computing the plume buoyancy as the difference of the virtual potential temperature within the thermals at an altitude z with the virtual potential temperature in the environment at a higher altitude $z + \delta z$ (rather than at the same level), so that buoyancy reads:

$$B' = g\frac{\theta_{v\,th}(z) - \theta_v(z + \delta z)}{\theta_v(z + \delta z)}. \tag{6}$$

With this modification, the detrainment is "aware" of the inversion before reaching it, and starts to detrain below it.

In the current version, $\delta z = \mathbf{DZ} \times z$, **DZ** being considered as a new adjustable parameter. Based on a systematic sensitivity analysis to this single parameter in both SCM and 3D configurations, we identified a range of acceptable parameter values between 0.06 and 0.15. The value was finally fixed to 0.07 in the 6A version of LMDZ. One objective of the present paper is to revisit the value of this parameter whilst simultaneously adjusting the other parameters. This has not been possible previously, and can now be done systematically using the *High-Tune Explorer* described in Part I.

2.4. The Cloud Scheme for Boundary-Layer Clouds

In order to compute the cloud fraction and in-cloud condensed water, we use a probability distribution function for the sub-grid scale saturation deficit, s. This distribution $F(s)$ is approximated by a bi-Gaussian distribution. Thanks to a tracer-based sampling of LES results, Jam et al. (2013) demonstrated that one mode corresponds to the contribution from the thermal plumes and the second one to contribution from their environment. Based on these findings, a statistical cloud scheme was derived using five variables: the plume fraction α, the mean saturation deficits within environment, $s_{\rm env}$, and plumes, $s_{\rm th}$ (which are directly given by the thermal plume model), and their associated standard deviations, $\sigma_{s,\rm env}$ and $\sigma_{s,\rm th}$, for which a parameterization was proposed. Considering that the major contribution to both standard deviations of s is the exchange of air between the plume and its environment and that the dispersion of s values is enhanced when the contrast $s_{\rm th} - s_{\rm env}$ increases, standard deviations are parameterized as follows:

$$\sigma_{s,th} = \mathbf{BG2}\ (\alpha + 0.01)^{-\gamma 1}\ (\overline{s}_{th} - \overline{s}_{env}) + b\,\overline{q}_{tth} \tag{7}$$

and

$$\sigma_{s,env} = \mathbf{BG1}\,\frac{\alpha^{\gamma 2}}{1-\alpha}(\overline{s}_{th} - \overline{s}_{env}) + b\,\overline{q}_{t_{env}}, \tag{8}$$

where b, **BG1**, **BG2**, γ_1, and γ_2 are free parameters, and the last term, $b\overline{q}_{t,th}$ or $b\overline{q}_{t\,env}$, is a minimum width of the distribution introduced for a value of $\alpha \approx 0$. It was shown in preliminary tests that the three parameters, b, γ_1, and γ_2 do not have a dominant role and their values were kept fixed in the results presented here.

The values of $b = 2 \times 10^{-3}$, **BG1** = 0.92, **BG2** = 0.09, $\gamma_1 = 0.4$, and $\gamma_2 = 0.6$ were chosen using LES results by fitting independently the in-thermal and environment Gaussian distributions.

The thermal plume model is activated before the cloud scheme. The condensation is taken into account in the computation of liquid potential temperature (considered as conserved variable in Equation 1) and virtual potential temperature involved in the buoyancy computation. Once e, d, and f are determined, Equations 1 and 2 are applied to the total water and liquid potential temperature to compute tendencies associated with the boundary-layer transport. From the thermal plume model computation, the parameters of the bi-Gaussian sub-grid scale distribution, $F(s)$, for the saturation deficit can be estimated as explained above. From this distribution, the cloud fraction $\alpha_{cld} = \int_0^\infty F(s)ds$ and cloud liquid content $q_l = \int_0^\infty sF(s)ds$ at the grid scale are finally computed. Note that the same cloud scheme is applied with a single mode of width $\sigma_{s,env} = b\,\overline{q}_{t_{env}}$ when the thermal plume model is not activated (for stratiform clouds for instance) while a different scheme is used for deep convection. Equations and details on the cloud scheme are given in Hourdin et al. (2013).

The computation of the conversion from cloud water to rainfall follows Sundqvist (1978): rainfall starts to precipitate significantly above a critical value **CLC** for the in-cloud liquid water q_l, fixed to 0.65 g/kg in the 6A configuration, with a time constant τ of half an hour. The associated sink for liquid water is

$$\frac{dq_l}{dt} = -\frac{q_l}{\tau}[1 - e^{-(q_l/\mathbf{CLC})^2}] \tag{9}$$

Following (Sundqvist, 1988), a fraction of the precipitation is re-evaporated in the layer below and added to the total water of this layer before the statistical cloud scheme is applied. The associated reduction of the precipitation flux P with altitude z is given as

$$\frac{\partial P}{\partial z} = -\mathbf{EVAP}[1 - q_t / q_{sat}]\sqrt{P} \tag{10}$$

where q_t is the total water mixing ratio, q_{sat} the water mixing ratio at saturation and **EVAP** a free parameter.

A summary of the parameters finally retained as free parameters in the present study are given in Table 1.

3. Experimental Setup

3.1. The 6A Version of LMDZ

The parameterizations described here are a crucial piece of the physical parameterizations of the LMDZ atmospheric global model. The recent modification of the detrainment formulation presented above produced a major improvement in the 6A version, the atmospheric component of the IPSL-CM6A-LR used for CMIP6. This version is extensively described by Hourdin et al. (2020a). Beyond controlling boundary layer clouds, the thermal plume model provides a lifting energy and lifting power to a mass flux parameterization of deep convection, which itself can be self-maintained through its coupling with a parameterization of the cold pools created below cumulonimbus by rainfall evaporation (Grandpeix & Lafore, 2010). Deep convection and cold pools only indirectly affect the boundary layer convection and shallow cumulus, by modification of their environment. They are not active at all in the SCM test cases considered in the present study.

As explained in the introduction, the development and tuning of the 6A version of LMDZ resulted from a long iterative process. The final adjustment of the top-of-atmosphere (TOA) net radiation was based for a large part on the adjustment of the conversion rate of cloud liquid water to rainfall **CLC**. This parameter very efficiently modifies the net balance because it affects only liquid (thus essentially low) clouds and has thus a much larger impact on the SW than on the LW radiation at TOA.

Two vertical discretizations are used in the present study. The first one, based on 79 layers (L79) corresponds to the standard vertical grid in the 6A version of LMDZ. In the first 3 km, the layer thickness is typically $\Delta z \simeq 0.12z$. A L95 grid is defined for the present study to refine the vertical resolution in the first few km

Table 1
Parameters Involved in the Iterative Refocusing

Name	Min	Max	Ref	Sampling	Controls
A1	0.5	1.2	2./3.	Linear	Contribution of buoyancy to the plume acceleration
A2	1.5e-3	4.e-3	2.e-3	Linear	Drag term in the plume acceleration
B1	0.	1.	0.95	Linear	Scaling factor for entrainment and detrainment
CQ	0.	0.02	0.012	Linear	Influence of humidity contrast on detrainment
DZ	0.05	0.2	0.07	Linear	Environmental air altitude shift for buoyancy computation
BG1	0.4	2.	1.1	Linear	Width of the environment subgrid scale water distribution
BG2	0.03	0.2	0.09	Linear	Width of the plume subgrid scale water distribution
EVAP	5e-5	5e-4	1e-4	Log	Reevaporation of rainfall
CLC	1e-4	1e-3	6.5e-4	Linear	Autoconversion of cloud liquid water to rainfall

Note. The minimum and maximum values explored are given as well as the reference value used in the 6A configuration of LMDZ, the information on whether the parameter is explored with a linear or logarithmic sampling and the meaning of each parameter.

above surface. The layer thickness is typically $\Delta z \simeq 0.067z$. The dependency of layer thickness upon altitude is given in Figure 2.

The motivation for using these two vertical grids here is to illustrate the approach both on a revisit of previous results and on a predicted evolution for the next model generation. The vertical resolution is key for the representation of boundary layer clouds which are often not much thicker than one or a few model layers. It also allows us to illustrate the significance of the structural error in the simulation of the cloud altitude and its link with the model vertical resolution.

3.2. SCM/LES Test Cases and Associated Metrics

For the SCM calibration, we consider four test cases among the cases listed in Part I, including one that consists of three sub-cases.

The first case, IHOP/REF, corresponds to an almost cloud-free convective boundary layer observed during the International H_2O Project (IHOP) field-experiment. This case is derived from observations collected on June 14, 2002 over the Southern Great Plains (Couvreux et al., 2005).

The second case, ARMCU/REF, is derived from observations collected on June 21, 1997 at the Atmospheric Radiation Measurement site in Oklahoma, US (Brown et al., 2002). This idealized case is typical of the diurnal cycle of shallow convection over land with well-developed fair weather cumulus.

The RICO (Rain In Cumulus over the Ocean, van Zanten et al., 2011) experiment focuses on precipitation processes at play in the trade-wind shallow cumulus. During RICO, significant precipitation was frequently observed, offering a unique opportunity to study the dynamics of shallow cumuli and precipitation.

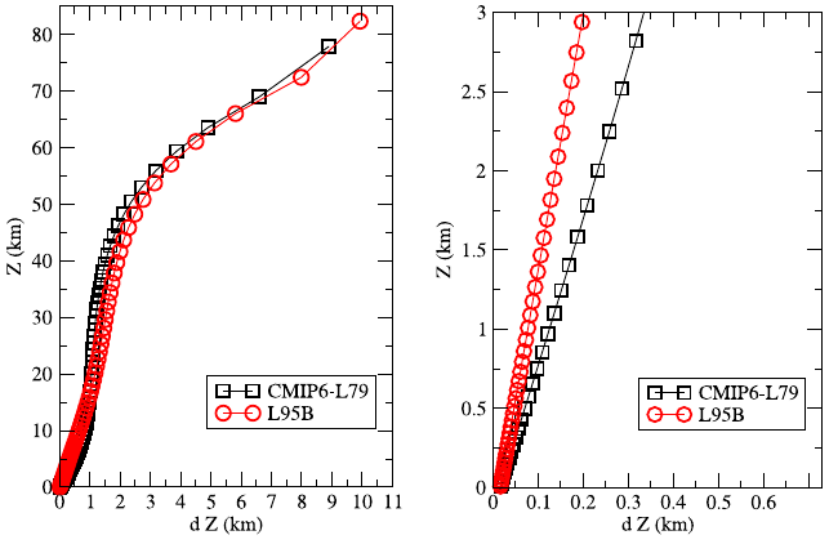


Figure 2. Vertical discretization: standard L79 grid of the 6A version and refined L95 discretization. The figure shows the layer thickness (*x*-axis) as a function of altitude (*y*-axis). The left panel shows the whole atmospheric column and the right panel is focused on the first 3 km above surface.

We finally use the composite stratocumulus-to-cumulus transition case discussed by Sandu and Stevens (2011). This case was built by compositing the large-scale conditions sampled along a set of individual Lagrangian 3-days trajectories that occurred over the northeastern Pacific during the summer months of 2006 and 2007. The stratocumulus deck presents a pronounced diurnal cycle and begins to break-up during the second day while the boundary layer deepens. Two variations of this SANDU/REF case, corresponding to a slower and a faster transition in cloud fraction were derived in a similar manner by compositing over the trajectories experiencing the fastest and the slowest decrease in cloud fraction over the first two days respectively (FAST and SLOW hereafter). The setup of the REF, FAST, and SLOW cases and the LES simulations are described in more detail in Sandu and Stevens (2011).

The ARMCU/REF and RICO/REF cases were used extensively for the inspiration, development and assessment of the thermal plume model and bi-gaussian cloud scheme (Couvreux et al., 2010; Jam et al., 2013; Rio et al., 2010). The SANDU cases were at the heart of the work on the modification of the thermal plume model to represent stratocumulus clouds (Hourdin et al., 2019).

Various metrics were tested and considered during preliminary experiments. Here, we retain metrics directly linked to the mean thermodynamical conditions targeted, as the mixed layer potential temperature and humidity, indicative of the mixing efficiency of the EDMF scheme. For all the cloudy cases, we retain either the total cloud cover ($\alpha_{cld,\max}$, computed as a maximum on the vertical) or the height of clouds. For the latter, two diagnostics are used: an average height $z_{cld,ave} = \int_0^\infty \alpha_{cld} z dz \,/\, \int_0^\infty \alpha_{cld} dz$ and a height that better emphasizes the height of the maximum cloud fraction, computed as $z_{cld,max} = \int_0^\infty z\alpha_{cld}{}^4 dz \,/\, \int_0^\infty \alpha_{cld}{}^4 dz$. This choice is rather arbitrary and was shown to work well in practice. Such integral metrics are less dependent on the model vertical resolution than maximum cloud height for instance. The metrics are averaged in time over a few hours in order to smooth out possible numerical oscillations. The choice of a particular set of metrics is rather arbitrary and thus critically relies on the modeler's expertize and objectives. The particular set of metrics retained here is given in Table 2.

As will be highlighted by the ensemble of simulations run with the *High-Tune Explorer*, two aspects are particularly critical and are thus targeted by the retained metrics. The first one concerns the RICO case which, depending on the parameter values, can have a maximum cloud fraction at 3 km varying from a few to 100%. This altitude corresponds to a second maximum, while the cloud fraction at cloud base is much

Table 2
Metrics Retained for the SCM/LES Tuning

Case	IHOP	ARMCU	RICO	SANDU	SANDU	SANDU
Subcase	REF	REF	REF	REF	SLOW	FAST
Time	7–9	7–9	19–25	50–60	50–60	50–60
$\theta_{400-600\ m}$	X	**X**	-	-	-	-
$q_{v,400-600\ m}$	-	**X**	-	-	-	-
$\alpha_{cld,max}$	-	**X**	**X**	-	-	-
$z_{cld,ave}$	-	X	-	X	-	-
$z_{cld,max}$	-	X	-	**X**	X	X

Note. The time retained for time average is given in hours from the beginning of the simulation. The X with bold fonts corresponds to the sub-set of metrics used in Section 4.

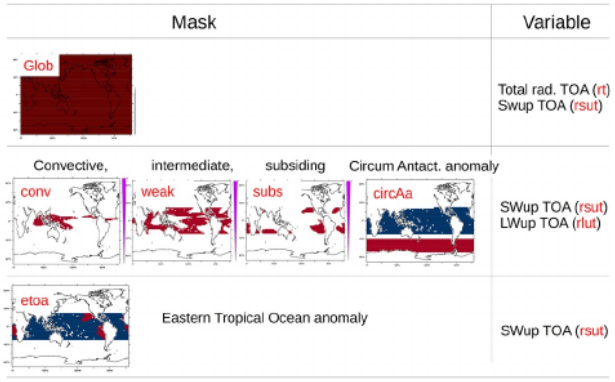
less sensitive to the tuning. The second aspect targeted by the metrics is the vertical development of the boundary layer in the transition cases. It was shown in particular in Hourdin et al. (2019) that this growth is very sensitive to the **DZ** parameter, introduced on purpose to improve the representation of stratocumulus clouds. In particular, it was more difficult to represent correctly the SANDU/SLOW case. For those cases, the height of the maximum cloud fraction, which is located just below the boundary-layer top, was used.

3.3. Setup of GCM Simulations and Associated Metrics

For the global simulations, we used stand-alone atmospheric simulations forced by SST and Sea Ice Cover (SIC) mean seasonal cycle, following the "AMIP" protocol (12 SST and SIC maps, one per month, interpolated in time with splines). Simulations are run on the standard low resolution (LR) horizontal grid made up of 144 points in longitude and 143 in latitude.

The metrics retained for the GCM simulations are typically those which were prioritized during the effective tuning of the 6A version of IPSL-CM6A-LR. They consist of radiation at top-of-atmosphere computed in annual mean and averaged over spatial masks as illustrated in Figure 3, using as a reference the CERES-EBAF L3b observational dataset (Loeb et al., 2009).

The global total radiation (imbalance between SW and LW) is of course a priority target. Note that the global radiative balance is not constrained by observations. It is assumed that it should be zero in a climate which would have reached an equilibrium (or quasi equilibrium). Because the climate is currently warming under the effect of green-house gas increase, it is assumed that there is in fact currently an imbalance in the global top-of-atmosphere radiation of about 0.5–1 W/m^2, which is equal to the "oceanic heat uptake," a downward net flux at the atmosphere-ocean interface, associated with the slow oceanic warming. Those values are, however, not observed; the typical uncertainty on the global SW and LW top-of-atmosphere fluxes being of the order of 4 W/m^2 (Loeb et al., 2009). In fact, rather than tuning the global radiation to the theoretical value of 0.5–1 W/m^2, we rather tuned it to a global imbalance of about 2.5 W/m^2. We know indeed that, for our particular model, an imbalance of 2.5 W/m^2 in forced-by-SSTs stand-alone atmospheric simulations leads to a global mean SST in the coupled model that matches present-day observation. The inconsistency between the tuning in stand-alone and coupled simulations may be due in part to some global energy leak in the model (typically of the order of 0.7 W/m^2 in the current IPSL-CM model) and changes in the mean climate



Metrics	target W m^{-2}	error W m^{-2}
glob.rt	2.5	0.2
glob.rsut	99.6	5
circAa.rsut	24.0	5
circAa.rlut	-48.6	5
subs.rsut	84.9	5
weak.rsut	81.8	5
conv.rsut	103.2	5
subs.rlut	274.6	5
weak.rlut	264.3	5
conv.rlut	235.8	5
etoa.rsut	11.0	5

Figure 3. Metrics retained for the GCM tuning consisting in radiative fluxes at top-of-atmosphere averaged over a mask, shown in red on the left hand side of the figure, or a difference between a red and blue mask (anomalies). The target and σ error retained for the history matching are shown in the table on the right hand side. The target values are computed from the CERES-EBAF L3b observational data set (Loeb et al., 2009). GCM, General Circulation Model.

that may induce changes in the global balance (like a different location of the mid-latitude jet, which may modify the latitudinal distribution of the CRE).

In addition to the global radiative balance, we also consider the global TOA SW upward radiation, assuming that the downward one is well constrained, and that the global LW outgoing radiation will be constrained automatically by the constraint on the SW and total radiation.

Additional constraints are considered by defining masks on the top-of-atmosphere outgoing LW and SW radiation, considering separately convective, subsiding and intermediate regimes in the tropics (defined by a threshold on the mean vertical velocity in ERAI reanalysis) and a contrast in latitude between the roaring forties and tropical oceans. These last metrics target a classical circum Antarctic warm bias in coupled ocean-atmosphere simulations. Similarly, a specific metric is dedicated to the SW contrast between Eastern Tropical Oceans and mean tropics: the ETO Anomaly, defined by Hourdin et al. (2015), in relation with the East Tropical Ocean classical warm biases.

3.4. Setup for the History Matching

The history matching sequence consists in the following steps which are described in detail in Part I (Couvreux et al., 2020).

1. The **metric selection and references** were just detailed for both SCM (Section 3.2) and GCM (Section 3.3) simulations.
2. The **selection of model parameters** to be adjusted and the a priori parameter ranges were presented in Section 2.
3. The **experimental design** then consists of defining the ensemble of SCM or GCM experiments on which metrics are effectively computed. The goal is to optimally sample the parameter space with a set of parameter values as small as possible (in practice a few tens to hundreds).
4. An **emulator** or surrogate model is then built for each metric, based on a Gaussian Process (GP). The emulator gives a statistical estimate of the corresponding metric value at any point of the full parameter space, without running the SCM or GCM, providing both the expectation of the metrics and an estimate of the uncertainty associated with the fact that only part of the parameter space was effectively sampled.
5. By comparing the reference metrics and those inferred with the emulators, **history matching** then rejects parameter values that lead to unacceptable model behavior (too large distance from the reference) and thus defines a not-ruled out yet (NROY) space, the model parameter space that cannot be further reduced given the sources of uncertainty.
6. **Iterative refocusing** finally consists in sampling the NROY space thus obtained and rerunning steps 3 to 5, constructing a refined emulator with smaller associated uncertainty inside this previous NROY. This new emulator is used to reduce the NROY space iteratively, each iteration being called "wave."

Note that the NROY space is not a well-defined geometrical object. It can only be defined by sampling the hypercube (with a much larger sample than the one used for the experimental design) and running the emulators to select which parameter vector is acceptable or not. The NROY at wave #N is defined in practice by applying sequentially all the emulators computed during the N waves, which have thus to be stored along the iterative procedure. The sample used for experimental design at wave #N+1 is a sub-sample, chosen randomly inside this selection.

Mathematically, the definition of the NROY space of parameters is based on implausibility derived from Gaussian process emulators fitted to each metric, as detailed in Part I. The implausibility itself (Williamson et al., 2013), $I(\boldsymbol{\lambda})$, is defined as the absolute difference between the observed metrics (target) and expectation of the emulator for the same metrics, divided by the standard deviation of this difference, comprising observational uncertainty, model structural uncertainty and uncertainty associated to the emulator (cf. Part I for a complete presentation). A point of the parameter space is kept in the NROY space when the implausibility is smaller than a threshold or cutoff. In all the applications presented below, a series of iterations or waves is done, keeping the same list of metrics at each iteration. The cutoff on implausibility defining the NROY space is progressively reduced from 3 for the first 4 waves, to 2.5 in the following 3 and finally 2 for wave number larger or equal to 8. Reducing the implausibility cutoff along the consecutive waves,

accompanying the progressive reduction of the emulator uncertainty, is a normal part of the sequential calibration procedure (see Williamson et al., 2017, for discussion). After a series of waves based on SCM simulations, additional waves are optionally completed with full 3D GCM simulations, adding the 3D GCM metrics to the SCM ones.

The iterative refocusing is applied here first on 20 or 30 waves in SCM mode, as described in Part I using the automatic *High-Tune Explorer* tool. For SCM/LES comparisons, the observational error is estimated from the intra-model spread in an ensemble of LES simulations. This variability is generally much smaller than the discrepancy (structural error) between LES and SCM simulations. The discrepancy error is not known, and so we use history matching whilst prescribing a "tolerance to error" as presented in Part I (and in Williamson et al., 2015, 2017). This tolerance determines the existence of a non-empty NROY space. As we move through the waves, tolerance to error can be reduced when we see that the model is capable of getting to within previous tolerances of target metrics, if there is a good physical reason for the model being able to reduce target metrics (for example, there may be inherent limitations with the vertical resolution of the SCM that would prevent a metric from being as close to a reference LES at some altitude without compromising the performance elsewhere in the column and hence getting the metric "right for the wrong reasons"; our tolerance to error should reflect those cases when they are understood). Four numbers are used to characterize the tolerance to error in the SCM experiments presented here. For the potential temperature and specific humidity in the mixed layer, we directly prescribe the tolerance in terms of an absolute tolerance Σ_T and Σ_q while a relative error is prescribed on the height of clouds $\Gamma_z = \Sigma_z/z$ and cloud fraction $\Gamma_{\alpha_{cld}} = \sum_{\alpha_{cld}} / \alpha_{cld}$. For the height of clouds, the choice of relative rather than absolute error specification is motivated by the fact that the layer thickness depends almost linearly upon altitude, so that a relative error in terms of altitude is an absolute error in fraction of layer thickness.

For a subset of experiments, a couple of waves of iterative refocusing are run with the full 3D GCM, starting from a sampling of the model parameters, inside the NROY space obtained at wave 20 or 30 of the iterative refocusing in SCM mode. The GCM tolerance to error is fixed to the values given in Figure 3.

4. Revisiting the Tuning of Low Clouds in LMDZ6A

In this section, we revisit the tuning of the 6A version of LMDZ without modifying the parameters that control detrainment and entrainment, except for the coefficient **DZ**, the only one that was used as a free parameter during the tuning phase of this model configuration. The two other parameters used for this first illustration are the threshold value for the auto-conversion of in-cloud water into rainfall, **CLC**, and the factor put on the re-evaporation of rainfall coming from layers above, **EVAP**, two parameters which were extensively used as well during the 3D tuning of this version. Succinctly, we automatically retune three of the model free parameters assuming that all the others are fixed to the values of the standard LMDZ6A configuration. This example is thought as a first proof of concept of our approach, and to illustrate on a simple case the added value of preconditioning 3D GCM tuning with SCM simulations. It is also an opportunity to revisit the choice of the **DZ** parameter which was tuned by hand, as documented in Hourdin et al. (2019). It was shown in that study with both a L79 and L95 vertical grid configurations (the adjustment of the altitudes of this L95 configuration being slightly more refined in the first kilometers than the one used here, which is more refined in the upper atmosphere, anticipating a use in the 3D global model) that there was an optimal value of parameter **DZ**, somewhere between 0.05 and 0.15. A value of 0.07 was finally retained in the 6A version.

4.1. 1D History Matching

For this first example, we use five metrics, the ones shown with bold crosses in Table 2. 20 waves are run iteratively following the protocol described in Section. 3.4. 0.56% of the parameter space is retained at wave 20 and the history matching appears to converge.

Figure 4 shows the "implausibility matrices" obtained for wave 1, 5 and 20 from left to right. Implausibility matrices constitute an attempt to visualize a n-parameter NROY space (here $n = 3$). The matrix itself is divided into 2D sub-matrices, each one being a restriction to two parameters, the names of which are given in

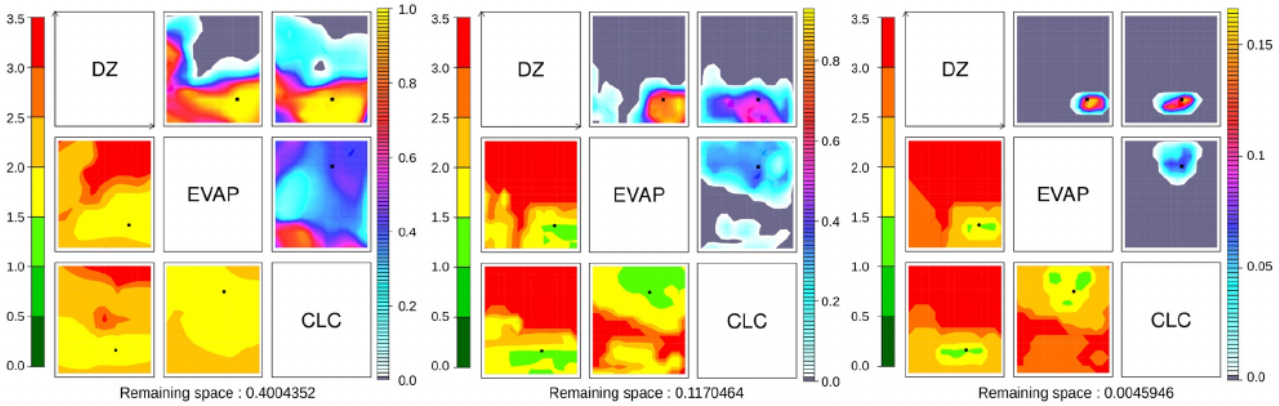


Figure 4. Implausibility matrices for wave 1, 5, and 20 of an history matching exploration, run with the L79 vertical grid and $\Gamma_z = 0.2$. The upper-right triangle is made of sub-matrices that display the fraction of points with implausibility lower than the chosen cutoff while the sub-matrices of the lower-left triangle show the minimum value of the implausibility when all the parameters are varied except those used as *x*- and *y*-axis, the name of which are given on the diagonal of the main matrix (additional details given in the text).

the diagonal of the main matrix. To fix ideas, the *x*-axis in the upper-right sub-matrix corresponds to **CLC** and the *y*-axis to **DZ**. Each axis spans the initial [min, max] range for the parameter considered. Each axis of the sub-matrix is divided into 15 sub-intervals (this number is adjustable within the tool), so that the matrix is made of 225 pixels. From a random sampling of (here) 10^6 vectors $\boldsymbol{\lambda}$, we compute the fraction of points with implausibility lower than the cutoff, when varying the $n-2$ (=1 here) other parameters. This fraction is displayed on the sub-matrices of the upper-right triangle. The total fraction of the volume of the NROY space relative to the initial *n*-dimension hypercube corresponding to the a priori [min, max] values of the parameters is the average over the sub-matrix, which should be the same for all the sub-matrices of the upper-right triangle and which is also indicated in text below the figure. A dark gray color means that there is no way to fit the observations by varying the *n*-2 unfixed parameters while a value of 100% means that values of the two parameters in x and y axis can be retained whatever the values of these *n*-2 parameters.

The sub-matrices of the lower-left triangle are displaying for each pixel the minimum implausibility obtained when varying the $n-2$ other parameters. They are orientated the same way as those on the upper-right triangle, for easier visual comparison, so that the labeling of the axis should be inverted for this lower-left triangle, compared to the names given on the diagonal (i.e., **CLC** corresponds to the x-axis and **DZ** to the y-axis for the lower-left sub-matrix as for the upper-right sub-matrix).

We note that, though we have performed 20 waves, here, the objective is not to find a single good simulation, which could be done using a Bayesian procedure within NROY space (Salter & Williamson, 2016), but to identify all good matches in order to use this subspace for the tuning of the 3D GCM.

The values of the three parameters retained for the 6A version of LMDZ6A, shown as dots in the figure, lie within the final NROY space. This result suggests that the long and slow expert tuning process of the 6A version was successful, at least for boundary-layer clouds and regarding the chosen metrics. It gives us confidence that in this case we did not miss a different tuning which could have significantly improved the results.

The size and shape of the final NROY space of course depends on the subjective choice of metrics and associated model tolerance, as well as on the vertical resolution. In the example shown here, we tested in particular the sensitivity of the NROY space to the addition of the slow and fast varying transition cases, to the resolution and to the tolerance error of the metrics associated with the height of clouds. Figure 5 compares the evolution with wave number of the size of the NROY space relative to the initial hyper-cube size with two values for the tolerance on the cloud height metrics, $\Gamma_z = 0.12$ and 0.2, for vertical resolution L79 and L95. In both cases for L95 resolution, the initial tuning of the 3 parameters lies in the NROY space. For the L79 grid, the NROY space becomes empty after 12 waves indicating that it is not possible to match the metrics with the lower resolution vertical grid for $\Gamma_z = 0.12$. For the L79 resolution, the error given by

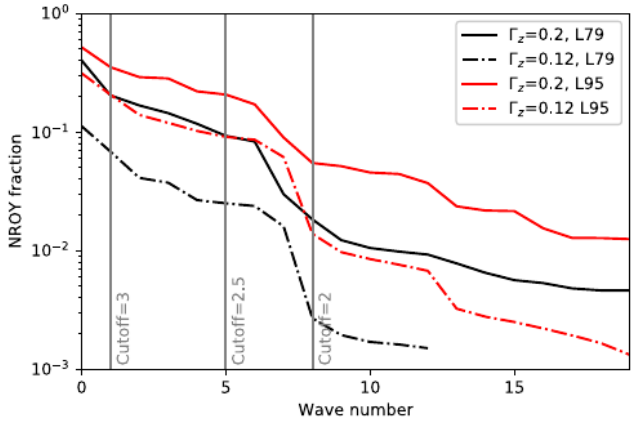


Figure 5. Reduction of the volume fraction of the NROY space (compared to the full initial hypercube volume, *y*-axis) remaining after *N* waves of history matching (*x*-axis) for the L79 and L95 vertical grids and with a relative tolerance to error on the cloud height of $\Gamma_z = 0.12$ and 0.2. The cutoff for implausibility is progressively reduced from 3 to 2.5 at wave 5 and 2 at wave 8, as indicated on the figure. NROY, not-ruled out yet.

$\Gamma_z = 0.12$ corresponds to one-layer depth. It is to say that, for a coarser grid the tolerance to errors has to be larger. Although not a surprise, this point is quantified here by our approach. Adding the SANDU/SLOW case to this history matching sequence with the L79 grid results in an empty NROY before convergence, for both $\Gamma_z = 0.12$ and 0.2 (results not shown). This is the reason why the SANDU/SLOW case was not included in this first sequence.

Note that only the sensitivity of the history matching sequence to the tolerance to errors on cloud height metrics was tested because of the rather straightforward link with vertical resolution. However, the sensitivity to the tolerance to errors for the other variables would deserve investigation as well.

4.2. 3D Test of the SCM-Based Tuning

The reduction of the NROY space based on a series of SCM simulations for four test cases is a very interesting result in practice, as it may save both time of scientific experts and computer resources needed for the full 3D global tuning.

In order to illustrate this point further, we run two sets of 45 2-year long experiments with the 3D GCM with the samples of the parameter space used for wave 1 (before any reduction) and for wave 20. The left panel of Figure 6 shows the mean latitudinal variations of the TOA SW CRE averaged both zonally and annually. While the spread across models is of 30 W/m^2 before NROY selection (gray), it reduces to less than 10 W/m^2 at wave number 20 (red). All the simulations using wave 20 parameters are close to the nominal 6A model configuration (blue) and in reasonable agreement with EBAF observation (black). This shows that a very similar tuning to the final one would have been obtained by tuning in 1D only, once the other model parameters are fixed. The right panel of Figure 6 shows the longitudinal variation of the same SW CRE in the southern tropics. This diagnostic underlines the contrast between a weak reflection of SW radiation (weak negative CRE) in the regions of trade winds cumulus, at around 130W in the Pacific Ocean and 40W over the Atlantic, and strong reflection in the regions of stratocumulus, at 100W over the Pacific and at Greenwich longitude over the Atlantic. The large range of SW CRE explored (from −20 to −110 W m^{-2}) in the stratocumulus regions before any parameter selection (wave 1, gray curves) is consistent with the strong impact of the value of **DZ** (Hourdin

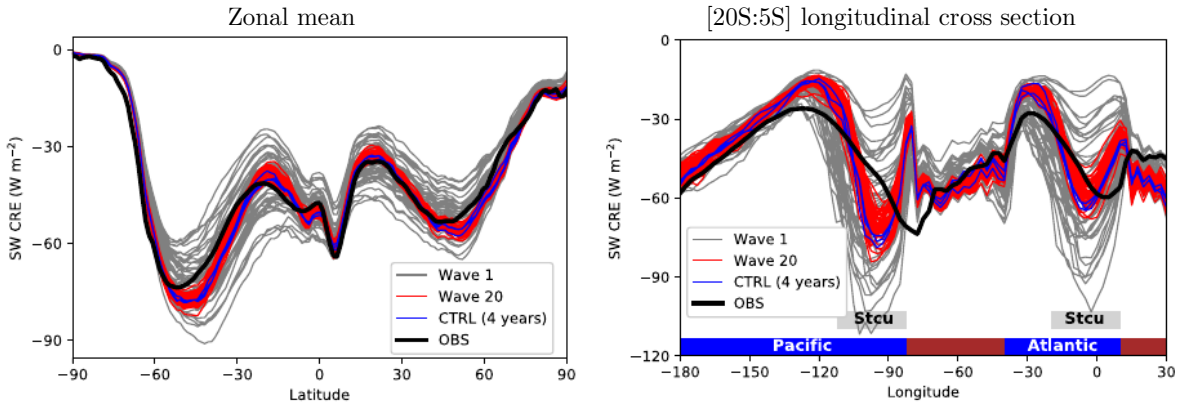


Figure 6. Zonally average latitudinal variation (left) and latitudinally averaged (between 20 and 5S) zonal variation (right) of the SW cloud radiative effect (CRE) at TOA for 45 L79 GCM simulations run with the sample of parameters used for wave 1 (gray) and a sampling of the NROY space remaining at wave 20 of the SCM history matching (red). The blue curves correspond to year 1–10 of a simulation run with the nominal values of the 3 parameters. The EBAF observations are superimposed in black. The location of continents, oceans and stratocumulus (Stcu) regions are indicated on the bottom of the right figure. NROY, not-ruled out yet. GCM, General Circulation Model.

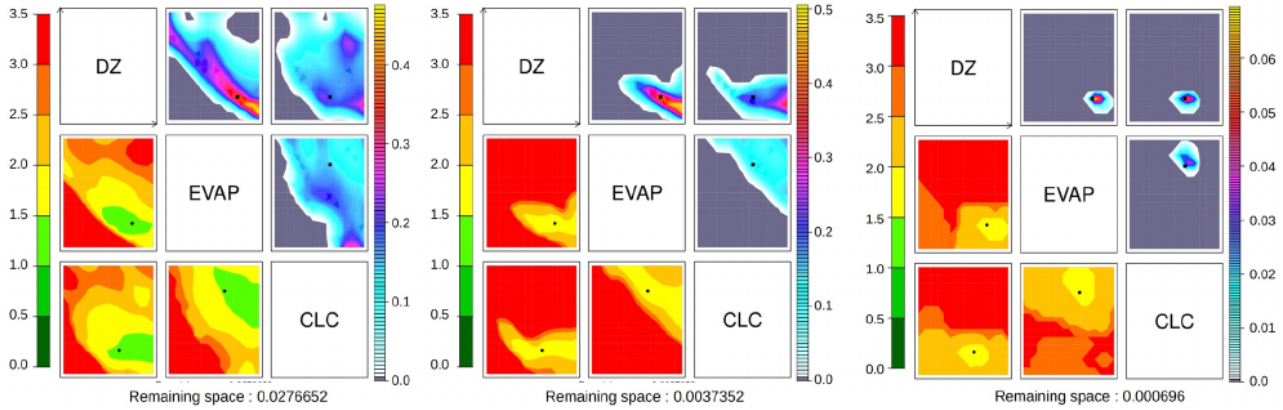


Figure 7. Implausibility matrices for wave 1 using only the 3D GCM simulations and metrics (left), wave 1 using both SCM and GCM metrics (middle) and wave 20 with both SCM and 3D, that is, adding 3D GCM metrics after 20 waves run with the SCM only (right). Both the SCM and GCM use the L79 vertical grid. SCM, Single column model.

et al., 2019) on the thickness of the stratocumulus clouds or even its disappearance. All the simulations using wave 20 parameters (red curves) produce results consistent with the control simulation (blue).

We present in Figure 7 the implausibility statistics obtained after considering 3D simulations using the 3D metrics presented in Figure 3. The left panel shows the implausibility matrix, which would be obtained with one single wave without preconditioning by 1D tuning. In this simple case, the selection is already quite efficient. The second panel shows the combination, on this first wave, of 1D and 3D metrics (using 45 parameter vectors used in parallel in 1D and 3D simulations), illustrating the significant gain of adding 1D metrics in the 3D tuning. However, in this case, the cost is essentially the same (the 45 GCM simulations). Finally, the last panel shows how adding one wave with the 45 3D simulations performed on wave 20 of the 1D multi-wave tuning shown in Figure 4 reduces the NROY space to a small and well defined region which includes the tuning finally retained for the LMDZ6A version.

5. Improving the Representation of Boundary-Layer Clouds

In this second example, we setup and tune a new version of the global model after modifications have been done to improve the representation of boundary-layer clouds at process level. The modification of the model consists here in both increasing the model vertical resolution and varying internal parameters of the thermal plume model that were kept fixed so far. The sensitivity of the parameterization behavior to the value of those parameters was partly explored during this development phase, by comparing SCM and LES results (Jam et al., 2013; Rio et al., 2010). However, without the tools presented here, it was not possible to fully explore the parameter space and some arbitrary values were finally retained, which have not been modified since. Indeed, even in the SCM framework, and even for a subset of parameterizations, exploring the full parameter space without tools such as those presented here is not practicable.

Here we explore the sensitivity to parameters **A1**, **A2**, **B1**, **CQ**, **BG1**, **BG2** (see Table 1). The tuning process is applied by varying these parameters together with those used in the previous section: **DZ**, **EVAP**, and **CLC**.

5.1. SCM History Matching With 9 Parameters

We first perform a 30-wave SCM history match with the extended set of parameters. Note that 20 or 30 waves may sound like a large number, though this has been done in epidemiological studies (Andrianakis et al., 2017), and is inexpensive using the SCM. The NROY matrices are shown in Figure 8 for $\Gamma_z = 0.12$ and Figure 9 for $\Gamma_z = 0.03$. The decrease of the NROY fraction with increasing wave number is shown in Figure 10 for three values of Γ_z (0.12, 0.06, and 0.03) and the two vertical grids.

The following lessons can be drawn from this new history matching sequence:

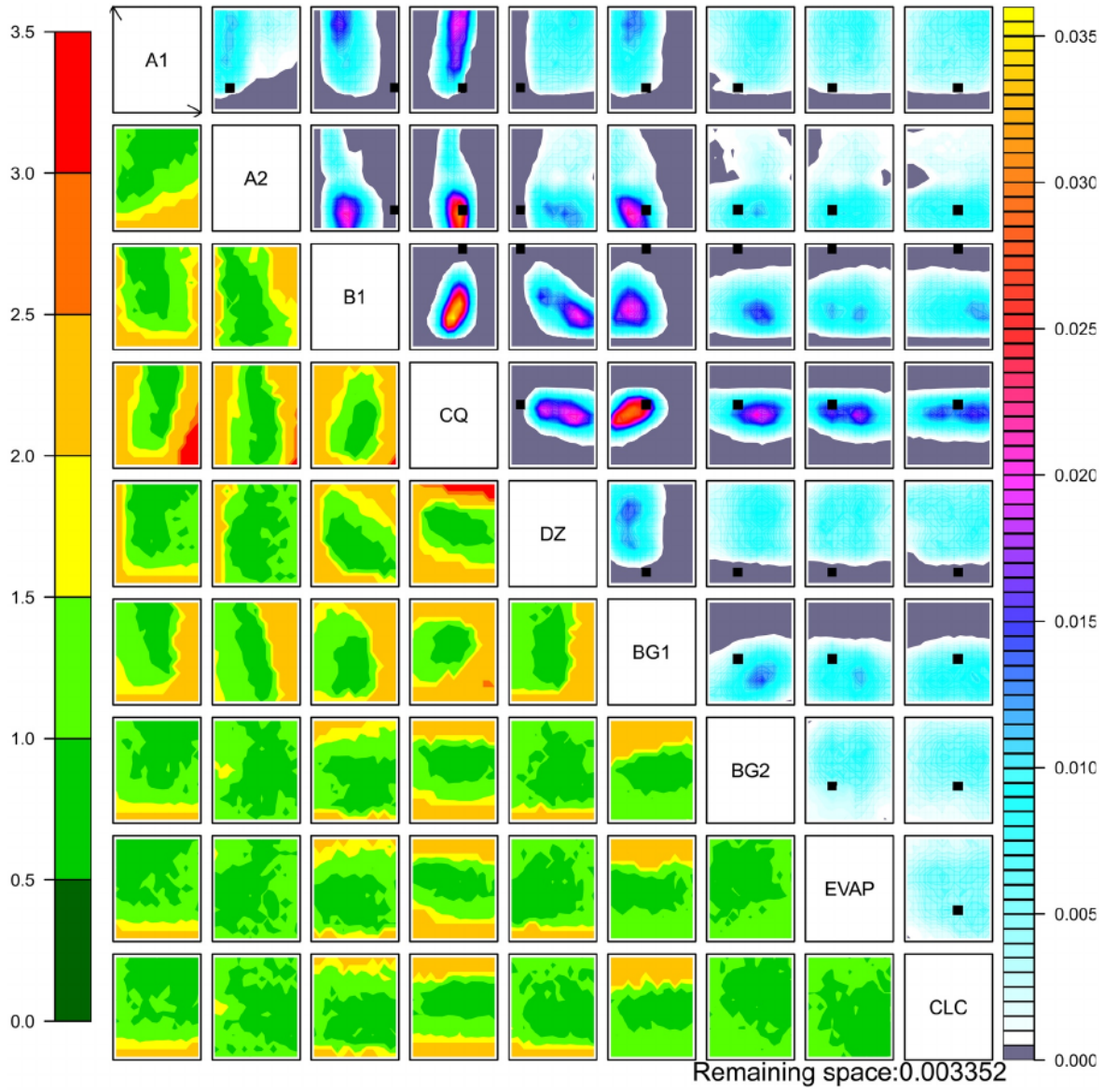


Figure 8. Implausibility matrix for the 9-parameter history match after 30 waves, vertical grid L95 and with a relative tolerance to error on the cloud height $\Gamma_z = 0.12$.

1. The history matching seems to converge and to produce a rather smooth and consistent picture of the NROY space
2. Due to the freedom given by the additional parameters, it is now possible to keep a significant NROY even with $\Gamma_z = 0.03$ for the L95 resolution. With this value of Γ_z, the $\pm 2\Sigma$ tolerance to error is $0.06 \times z$, which is close to the layer thickness
3. For the coarser grid, L79, only the $\Gamma_z = 0.12$ case is able to maintain a nonzero NROY space after 30 waves, that is, for a $\pm 1\Sigma$ tolerance to error close to the layer thickness
4. The NROY is obtained for values of the **B1** parameter much smaller than initially assumed, compensated by a larger value of **A1** and of **DZ**. So, in this case the tuning retained for CMIP6 was probably sub-optimal. The physical interpretation of this different tuning will be discussed later on
5. In particular, the value retained for CMIP6 of the **DZ** parameter is now out of the final NROY space. This is due to the fact that the tolerance has been reduced and the number of metrics increased. In particular,

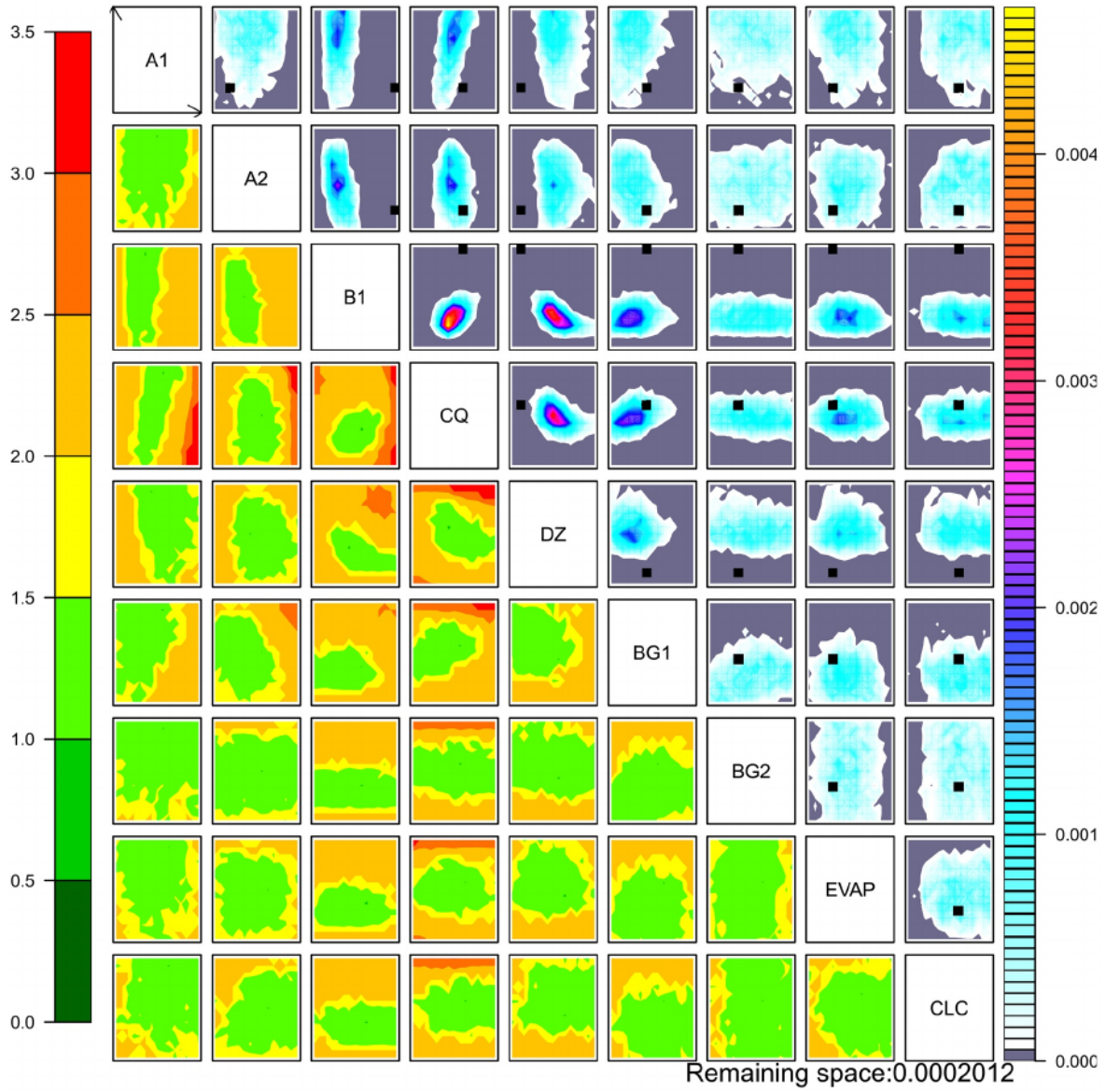


Figure 9. Same as Figure 8 (wave #30, vertical grid L95) but with a relative tolerance error on the cloud height of $\Gamma_z = 0.03$.

it is now possible to include the SANDU/SLOW case, which was too badly represented to be considered in the previous section

6. In the final NROY, the range of some parameters is quite narrow, as that of **B1**, **DZ**, or **CQ**, but others like **CLC** give room for a further tuning of the radiative balance in the full 3D global model

We show in Figures 11 and 12, for waves number 1 (gray), 3 (pink), 7 (yellow) and 30 (green), the envelope of the vertical profiles of potential temperature, specific humidity and cloud fraction for the 90 SCM simulations run to build the emulator with the L95 configuration and smallest tolerance to error. For the cumulus cases (Figure 11), the history matching converges to a narrow envelope (green) which contains the nominal 6A configuration (black). The improvement compared to the original profile is significant for the transition cases (Figure 12). Allowing the thermal plume parameters to vary allows the boundary layer to grow higher, in particular for the SANDU/SLOW case. The red curve on these figures is the best of the simulations run to build the emulators for the 30 waves, best in the sense that the maximum (across metrics) value of the ratio of the distance to observations divided by the tolerance to error is the smallest. This best simulation

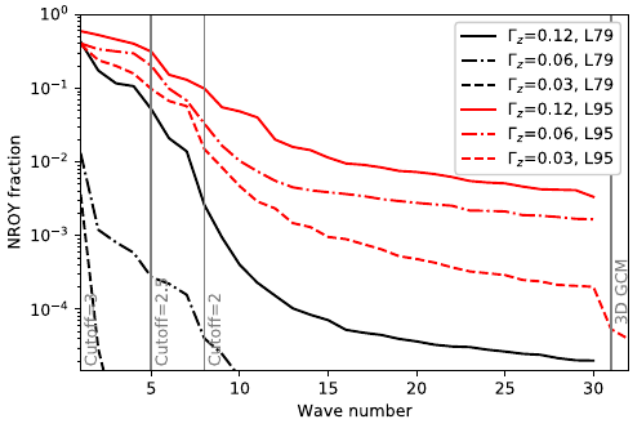


Figure 10. Reduction of the NROY volume fraction (compared to the full initial hypercube volume, *y*-axis) remaining after *N* waves of history matching (*x*-axis) for the L79 and L95 vertical grid and relative tolerance error on the cloud height $\Gamma_z = 0.03$, 0.06, and 0.12. The cutoff for implausibility is progressively reduced from 3 to 2.5 at wave 5 and 2 at wave 8, as indicated on the figure. For the case with the L95 grid and $\Gamma_z = 0.03$, two additional waves are added with 3D GCM simulations. NROY, not-ruled out yet.

was obtained as the 76th element of wave 26 (named SCM-26-076 on the graph). Note that the best simulation is not in wave #30 which is not a surprise. Because the iterative refocusing converges with a weak decrease of the NROY space in the last waves, the probability of sampling good simulations is not very different for these last waves.

5.2. 3D History Matching

We present here, the results of two subsequent waves of history matching with the 3D GCM. As explained in Section 3.4, the experimental design of the first wave in 3D is taken as a sub-sample of the sampling of the final NROY space obtained from the 30-wave history matching with the SCM, here with the L95 vertical grid and $\Gamma_z = 0.03$. For the experimental design of waves 31 and 32, 90 SCM and GCM simulations are run with the same sets of model parameters, from which the previous 12 SCM metrics and the 11 3D GCM metrics presented in Figure 3 are computed. The implausibility graph of wave 32 is shown in Figure 13. The fraction of the NROY space compared to the initial parameter hyper-cube is reduced from 2 10^{-4} at wave 30 to 4 10^{-5} at wave 32. Some parameters known to control the global radiative balance seem to contribute to this space reduction as seen for instance by a slight reduction of the NROY space in the (**EVAP**, **CLC**) subspace. As for the previous set of 3D GCM experiments (Figure 6) we first illustrate the GCM behavior in terms of mean latitudinal variations of the SW CRE averaged both zonally and annually (left panel of Figure 14), and of longitudinal variations in the southern tropics (right panel) of the same SW CRE.

The spread across models of wave 31 is not reduced as much as for wave 21 in the previous experiments where the sensitivity to three parameters only was explored. The gain compared to no preconditioning by SCM tuning (gray curves in Figure 6 gives an underestimation of the dispersion with no preconditioning since only three parameters were varied) is however significant, as is the reduction in the spread in the latitudinal variation when going from wave 31 to wave 32.

We show in Figure 15 the normalized (by the tolerance to error) error for the GCM metrics for the 90 GCM simulations run for wave 32. The simulations are ranked according to the maximum value of this normalized error. For most of the simulations, the global net radiative balance "glob.rt" dominates the error, which is of course partly attributable to the fact that we took an arbitrarily small error of 0.2 W/m^2 for this particular metrics (targeting a 0.2 K in coupled simulations). After the global radiative balance, some metrics are particularly difficult to get within the tolerance to errors, such as the LW circum Antarctic anomaly. It is interesting since this metric was introduced on purpose, targeting classical warm biases in coupled ocean-atmosphere models.

Five "BEST" simulations were selected from the ranking of Figure 15. By doing so, we go further than theoretically authorized by the history matching philosophy, that is, not going beyond the constraints imposed by the predefined tolerance in order to avoid over-fitting and subsequent compensating errors. It is done here to accelerate the tuning process and be sure to select simulations with a well-balanced global net radiation, in order to run one of them in coupled atmosphere-ocean mode. The five simulations are superimposed with gold color in Figures 11, 12, and 14.

The agreement with observations is at least as good for those BEST simulations as it is for the standard LMDZ6A configuration. In order to characterize further the behavior of these selected simulations, we show in Figure 16 for the SW CRE (left), the LW CRE (middle), and the precipitation (right) the mean bias and root-mean-square error computed on the mean seasonal cycle. The CMIP5 and CMIP6 multi-model ensembles are displayed (first two rows from bottom) in order to contextualize those results with respect to the state-of-the-art. The 5A, 5B, and 6A versions of the IPSL model (based on LMDZ for the atmosphere) are identified in blue, violet and red respectively. A general improvement is visible from CMIP5 to CMIP6,

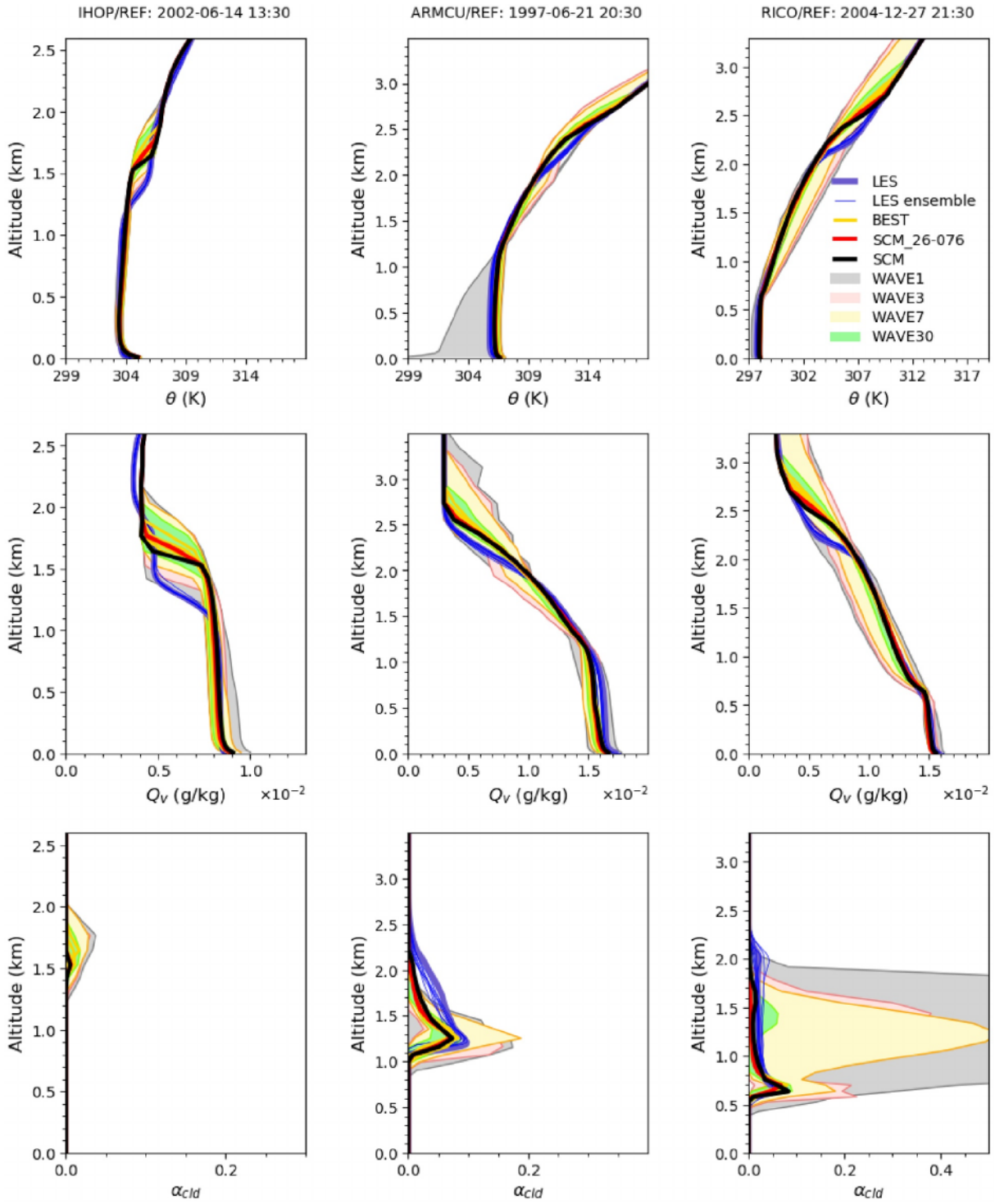


Figure 11. Evolution of envelopes of the vertical profiles of potential temperature (first row), specific humidity (second row) and cloud fraction (third row) for the IHOP, ARMCU, and RICO cumulus cases obtained with the L95 vertical grid and $\Gamma_z = 0.03$. Individual curves are superimposed for: LES (blue), LMDZ6A with nominal values of the parameters (black), the best simulation obtained with SCM tuning (red, the 76th simulation of wave #26 named SCM-26-076) and the BEST cases retained after subsequent 3D GCM tuning (gold). LES, large eddy simulations; SCM, Single column model.

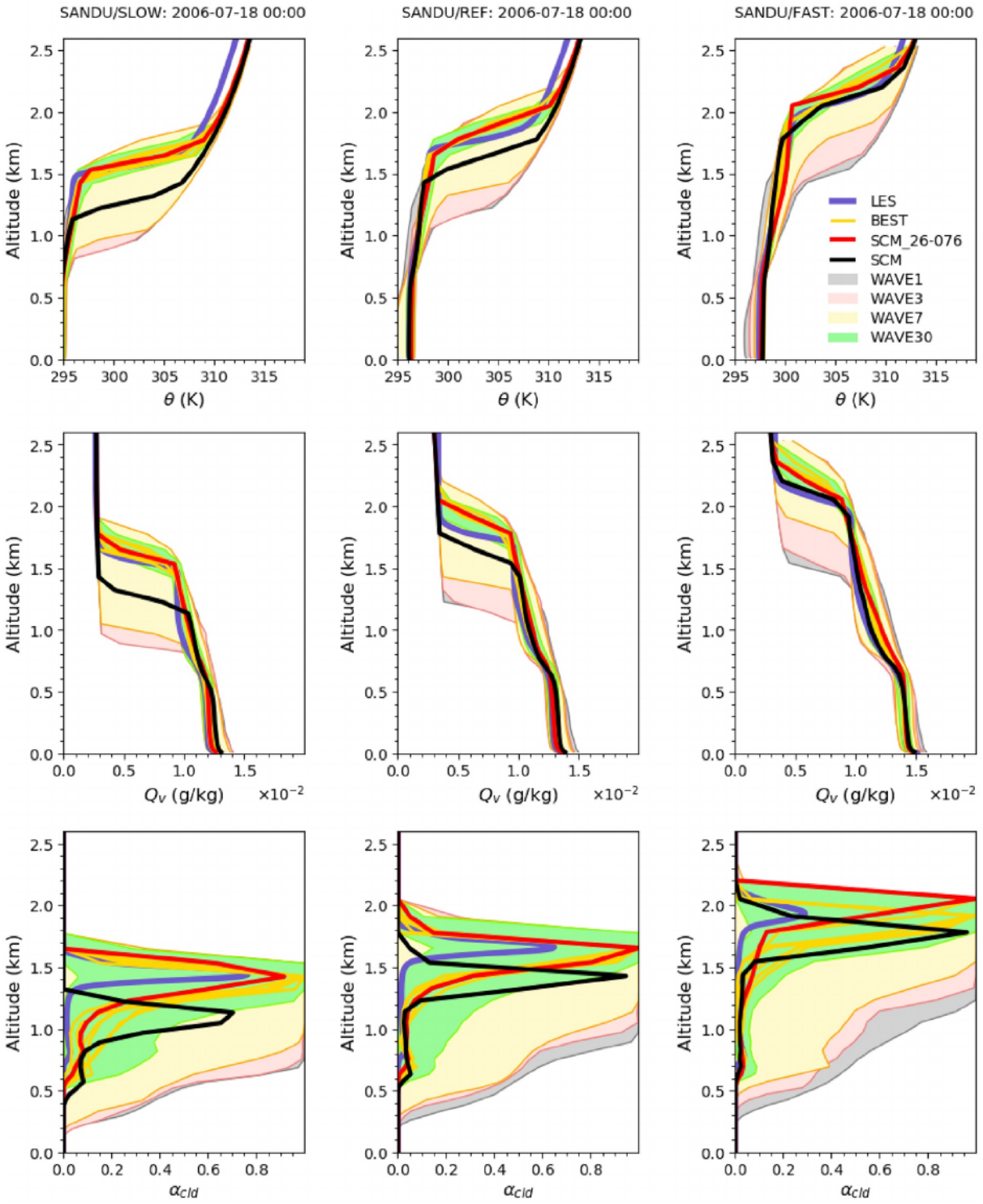


Figure 12. Evolution of envelopes of the vertical profiles of potential temperature (first row), specific humidity (second row) and cloud fraction (third row) for the three SANDU transition sub-cases. Same conventions as in Figure 11.

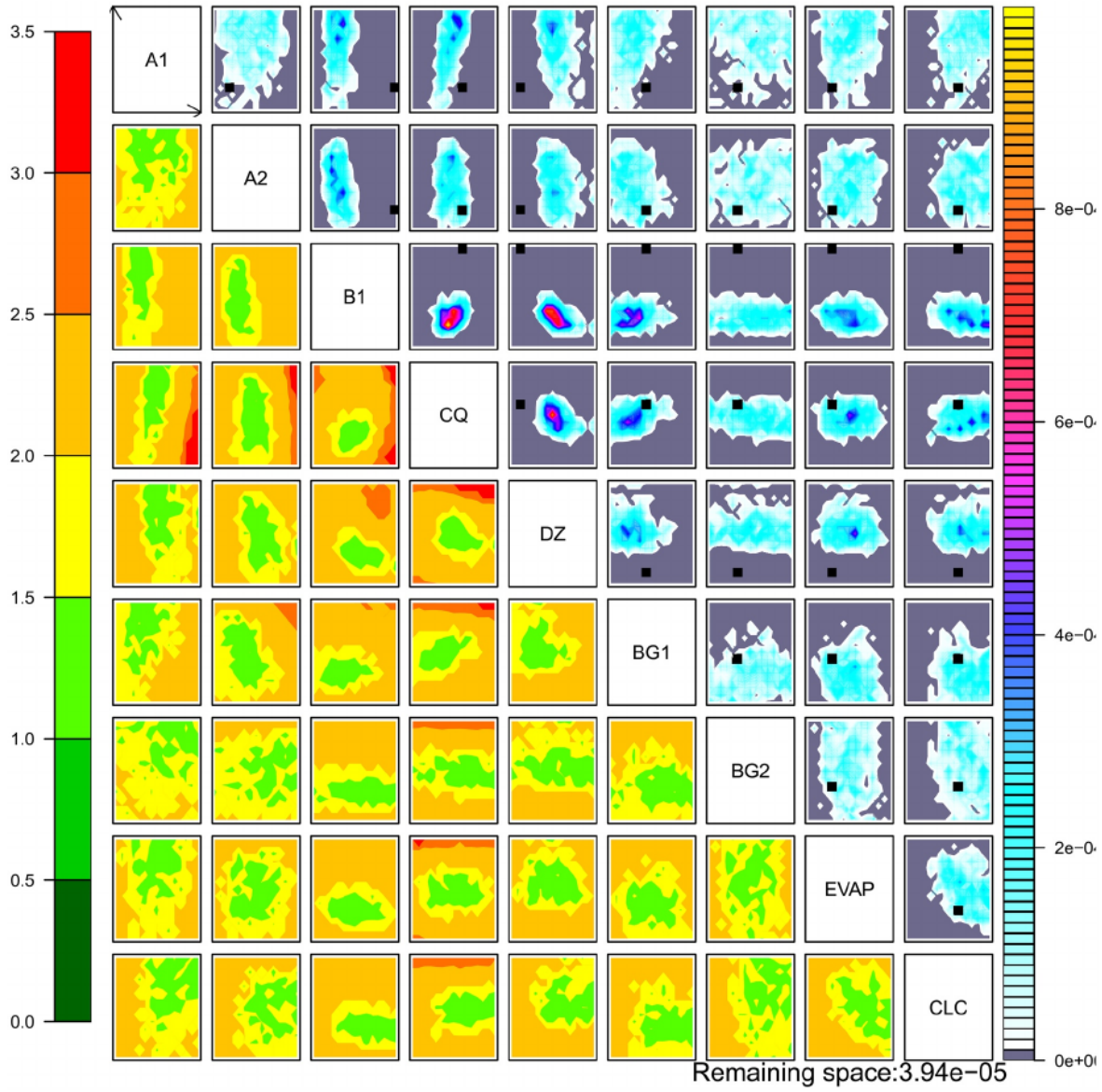


Figure 13. Implausibility matrix for the 9-parameter history match, at wave 32, built by adding two iterations with SCM and GCM metrics after 30 waves of SCM history matching, obtained with the L95 vertical grid. SCM, Single column model.

from the narrowing of the bias distribution and reduction of the mean RMSE. For the IPSL model, the 6A version behaves much better than the 5A and 5B versions, except for the rainfall. For rainfall, this has to be related to the fact that we struggled to reduce the mean rainfall in the 5A and 5B versions to compensate for a tendency of global models to overestimate the mean rainfall. Because it is not clear whether this mean bias is outside the observational errors (the observed mean rainfall may be significantly underestimated, see e.g., Berg et al., 2010; Stephens et al., 2012), we decided to abandon this target for the 6A version.

For the 6A version, we show as well 10 consecutive years run on climatological SSTs in order to illustrate the error and dispersion that come from this different setup (the CMIP diagnostics correspond to the mean seasonal cycle over the period 1979–2005). The mean bias is not significantly affected by the different setup, and its inter-annual variability is weak, a very important point for the tuning strategy adopted here. The root-mean-square error, on the opposite is significantly degraded when considering 1-year long simulations

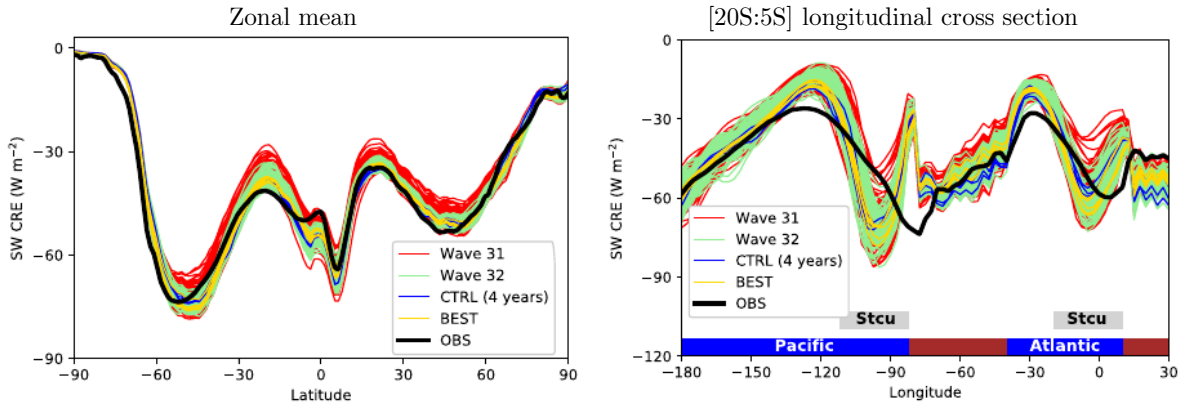


Figure 14. Zonally average latitudinal variation (left) and latitudinally averaged (between 20 and 5S) zonal variation (right) of the SW cloud radiative effect (CRE) at TOA for 90 L95 GCM simulations run with the sample of parameters used for wave 31 (red, i.e., after selection based on SCM/LES comparisons only) and wave 32 (green). The blue curves correspond to years 1–10 of a simulation run with the nominal values of the nine parameters. The gold curves correspond to the five BEST simulations (see text for details). The EBAF observations are superimposed in black. LES, large eddy simulations; SCM, Single column model.

on climatological SSTs. It is why we decided to rerun the BEST simulations on AMIP SSTs as well (upper row in the graphs). The scores of the SW and LW CRE is very similar as for the standard LMDZ6A configuration, and even better for the root-mean-square error for rainfall, without clear explanation for it so far.

Figure 16 also shows the results of wave 1 and 20 for the first 3-parameter tuning and wave 31 and 32 for the 9-parameter tuning. The reduction of the dispersion in the mean bias is clearly visible in this graph.

5.3. Test in Coupled Atmosphere-Ocean Configuration

Finally, the "BEST1" simulation (the first one in the ranking of Figure 15) is run in coupled mode, over 50 years, starting from initial conditions with present day forcing. A trick is used in this simulation to compensate the global oceanic heat uptake (of about 0.5–1 K in the present-day warming climate). It consists in increasing of the oceanic albedo by 0.007.

The seasonal cycle of SSTs is almost stabilized at the fifth decade. Figure 17 shows the mean bias and root-mean-square error of SST computed on a mean seasonal cycle of the BEST1 simulation (gold), compared to the other CMIP5 (green) and CMIP6 (black) simulations with IPSL simulations highlighted with different colors. The BEST1 simulation itself is a bit too warm. A second simulation is then run by just readjusting the **CLC** parameter by hand, by running one sensitivity experiment in forced mode to estimate the sensitivity of the global mean radiative balance to the parameter (without worrying about whether all the parameters

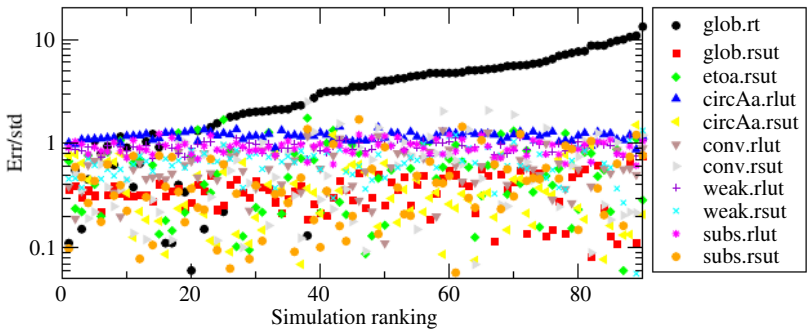


Figure 15. For each 3D GCM metrics, the ratio error/σ is shown, where σ is the tolerance to error used for history matching. The 90 L95 GCM simulations of wave 32 are ranked according to the maximum value of error/σ.

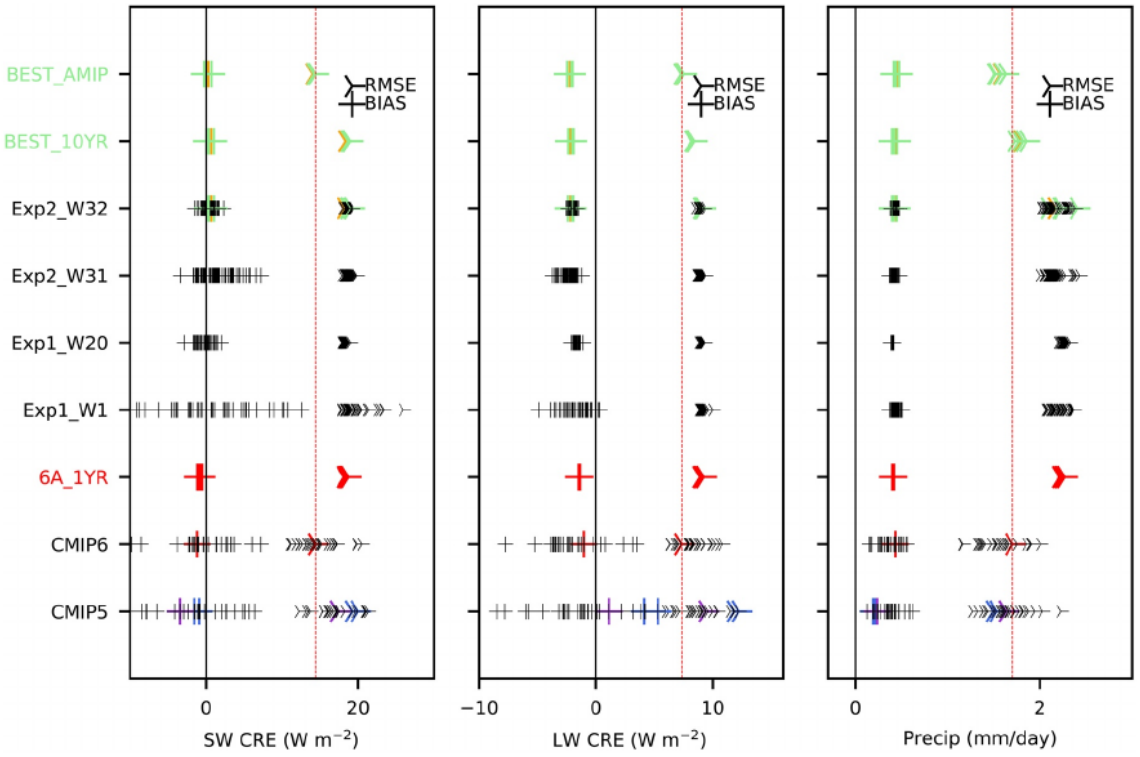


Figure 16. Mean bias and root-mean-square error (RMSE) of the SW CRE (left), LW CRE (middle) and rainfall (right) in LMDZ and CMIP simulations. The RMSE is computed on the mean seasonal cycle (i.e., from 12 monthly values on each grid cell after interpolation on a common $2^o \times 2^o$ longitude latitude grid). On each graph, from bottom to top, we show: the CMIP5 and CMIP6 multi-model ensembles (AMIP simulations over the period 1979–2005). For CMIP5 simulations, the blue and violet colors correspond respectively to the 5A and 5B versions of LMDZ (the 5A version was run with two different resolutions) and red color is used for the 6A version of the LMDZ model. The line labeled "6A_1 YR" shows 10 individual years with the standard LMDZ6A (L79 vertical grid) configuration run on climatological SSTs. The lines with label starting with "Exp" show the second year of a 2-year long simulation run on climatological SSTs for waves 1 and 20 of the first set of experiments (L79 vertical grid) and wave 31 and 32 of the second set (L95 vertical grid). The 5 "BEST" simulations are identified with green color. The two upper lines show the results of simulation obtained with the BEST configurations, when run over 10 years with climatological SSTs ("BEST_10YR") or over the 1979–2005 period with annually varying SSTs (AMIP protocol as for CMIP simulations, "BEST_AMIP"). The orange color corresponds to the "BEST1" simulation. The vertical lines correspond to a zero bias (black) and RMSE of the CMIP6 IPSL-6A-LR configuration (red dashed). The EBAF observations are used for the CRE and Global Precipitation Climatology Project (GPCP, Huffman et al., 2001) data set for precipitation.

are in the NROY space). For both simulations, the results are quite close to the 6A simulation. The results are better in the tropics (35 S:35 N) than for the full globe (65 S:65 N, removing latitude beyond 65° to avoid questions related to the sea-ice mask). This better performance when focusing on the tropics is probably due to the fact that the East Tropical Ocean warm bias is rather reduced in the BEST simulation compared to the 6A version while the circum-Antarctic warm bias is somewhat increased as illustrated in Figure 18.

6. Discussion

Both in the 3-parameter and 9-parameter history matching, a multi-wave tuning in SCM configuration is enough to partly constrain the radiative fluxes. It provides an avenue for process-based improvement of climate models, from SCM to global coupled model, following a systematic and rigorous approach.

6.1. Benefit for 3D GCM Tuning

Though the 9-parameter history matching with increased vertical resolution does not significantly improve the agreement with observations of the top-of-atmosphere distribution of radiative fluxes in a 3D GCM, it should be kept in mind that we did not include any parameters affecting the high clouds in the tuning

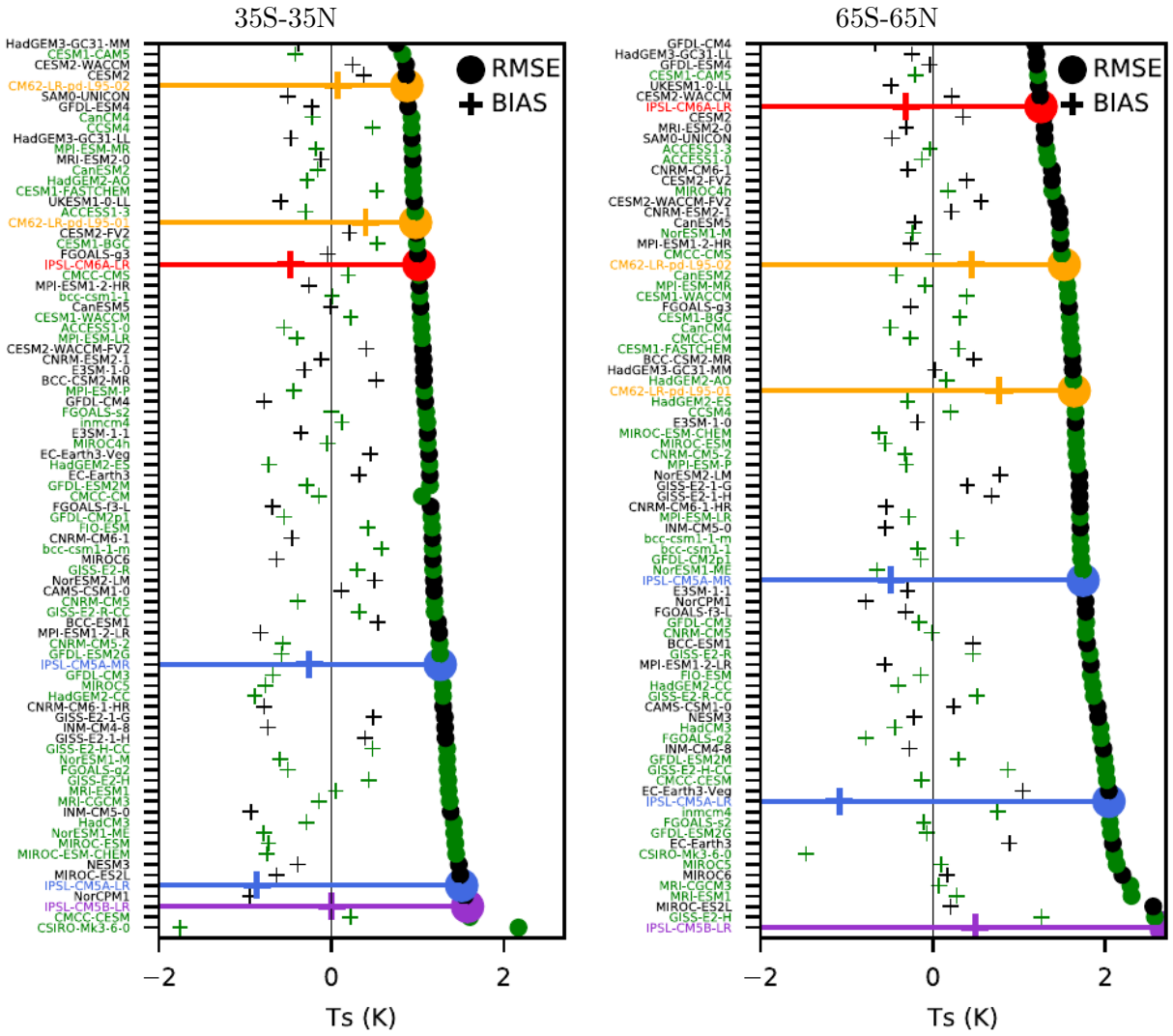


Figure 17. SST mean bias and root-mean-square error computed from the mean seasonal cycle (12 monthly means) after interpolation on a 120 × 90 regular longitude-latitude grid. The diagnostics are shown for tropical latitudes (left, 35 S:35 N) and for the global ocean (latitudes 65 S:65 N). All the CMIP5 (green) and CMIP6 (black) models available to us are shown. The color code for the IPSL CMIP configurations is: 5A (blue), 5B (violet), 6A (red), BEST (gold). The two gold points correspond to the best tuning (simulation CM62-LR-01 corresponding to simulation 35 of wave 32) and a second one with the parameter **CLC** slightly increased (simulation CM62-LR-02, after a by-hand tuning) to cool the simulations.

procedure, which of course would make the retuning easier by benefiting from a reasonable tuning of the high clouds. It could be, for example, that there are some compensating errors in the 6A configuration between high and low clouds, in mid and high latitudes. In addition, the control simulation considered here was the product of a long phase of a careful tuning of the global model, in which the metrics used here were explicitly high priority targets. Though we can be confident in the processes resulting from our tuning (for low clouds), additional parameters may need to be exposed to tuning for the full 3D model (or similar strategies for process based tuning with relevant parameters for other processes) to work around existing compensating errors and to fully benefit from our strategy.

Altogether, our results confirm that the proposed strategy is able to provide reasonable tuning of a coupled model, by applying a rather systematic procedure making use of machine learning techniques and starting from LES/SCM comparisons. This study shows how a 3D GCM can be retuned after some modifications with an automatic procedure, avoiding a long phase of by-hand retuning. The model evolution tested here

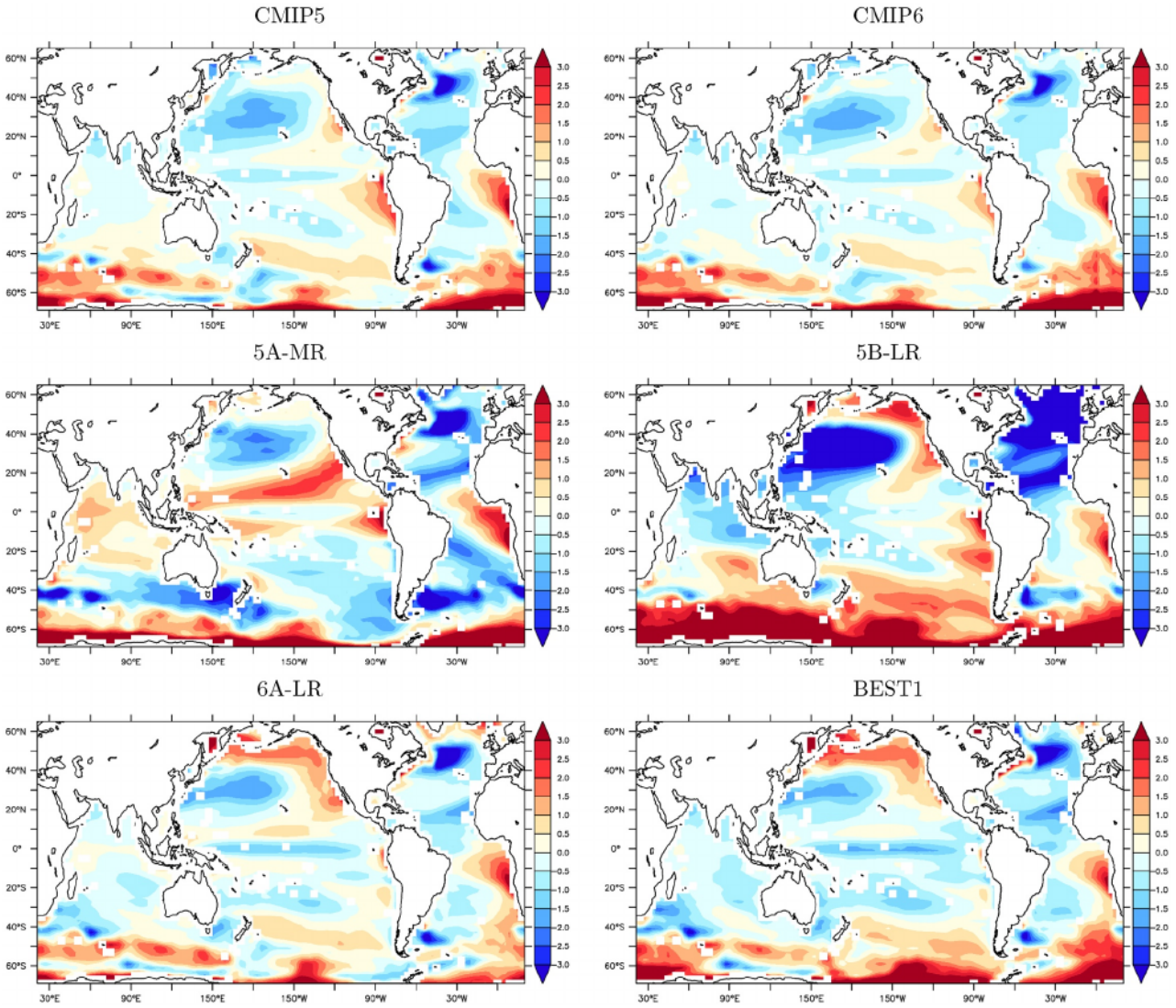


Figure 18. SST (K) mean bias for the CMIP5 and CMIP6 multi-model ensemble, for the 5A-MR, 5B-LR and 6A-LR and for the BEST1 simulation (with retuning of the **CLC** parameter). The global mean of the bias is removed to highlight the structure of the bias.

consists in increasing the vertical resolution together with allowing us to vary some additional free parameters. In this case, it was possible to improve the representation of clouds at process level, in particular by reproducing better the 1D "transition cases," with a 3D tuning at least as good as the previous one.

6.2. Enlightening the Representation of Cloud Processes

In order to interpret further the modification induced by this new tuning at the process scale, we show in Figure 19 the internal variables of the thermal plume model obtained with the ARM cumulus and SANDU/REF cases. The results are averaged on the same time period as that used for the metrics computations shown in Table 2: between hour 7 and 9 of the simulation for the ARMCU case which corresponds to 0030–0230 p.m. local time, and between hour 50 and 60 for SANDU, in the afternoon and evening of the third day. The vertical velocity is overestimated throughout the depth of the cloud for the control simulation, when compared to the plume velocity sampled in LES, and slightly underestimated near the surface. The retuned version amplifies the overestimation in the cloud. This could be seen as a degradation of the

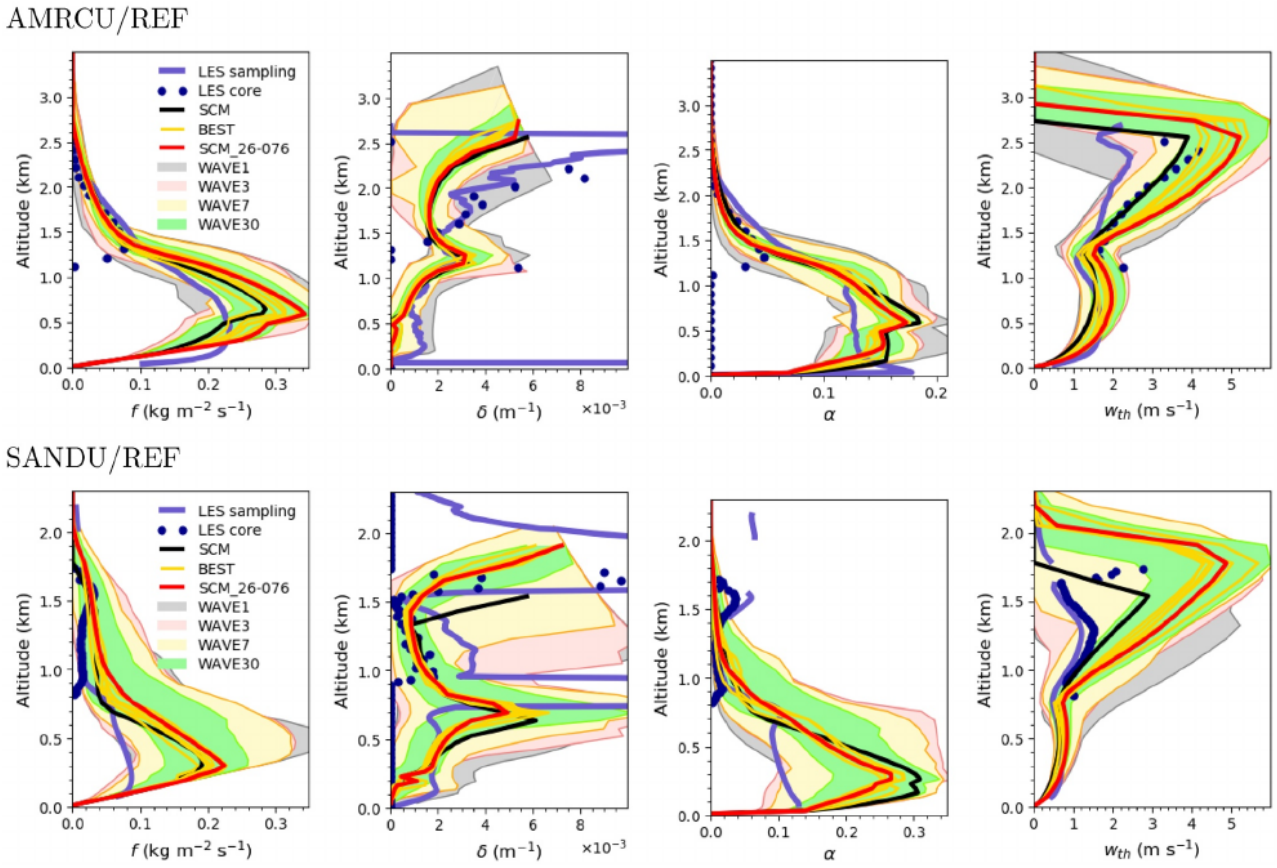


Figure 19. Vertical profiles of the internal variables of the mass flux scheme for the ARM cumulus simulation averaged between hour 7 and 9 of the simulation and for the SANDU/REF case, averaged between hour 50 and 60 of the simulation. As in Figure 11, we show both the evolution of envelopes of the vertical profiles obtained with the L95 vertical grid and $\Gamma_z = 0.03$ for successive waves as well as individual curves: LES (blue), LMDZ6A with nominal values of the parameters (black), the best simulation obtained with SCM tuning (red, the 76th simulation of wave #26 named SCM-26-076) and the BEST cases retained after subsequent 3D GCM tuning (gold). For the LES, we consider only one simulation and show for each case two ways of sampling the LES results. For the ARM case, we use the tracer-based sampling used for instance by Jam et al. (2013). For the SANDU case, in the absence of tracers in the simulations, we use the sampling retained by Hourdin et al. (2019). Compared to the standard sampling, the core sampling imposes that the sampled points show an excess of virtual potential temperature when compared to the horizontal average, retaining only points with positive buoyancy. LES, large eddy simulations; SCM, Single column model.

scheme or question the way thermals are sampled in LES. We could have selected more active parcels by using a more restrictive sampling threshold as illustrated by retaining only points with positive buoyancy (core sampling, blue dots). In the end, what really matters for the transport is the mass flux. It appears that the vertical velocity increase is in part compensated by a reduction of the fractional cover attributed to convective plumes leading to a very similar mass flux, constrained by the requirement to faithfully represent the clouds, as imposed through the history matching procedure.

We observe that the procedure tends to favor tuning with stronger velocity, which can be related to the use of values of coefficient **B1** much smaller than one. This coefficient enters in the definition of both entrainment and detrainment, and would be 0 for a plume with conserved mass flux, which would just accelerate without entraining air from the mixed layer (in which case the plume fractional cover decreases when the plume accelerates), and one for a plume that would entrain enough air to keep its fractional cover constant.

With this stronger vertical velocity, the plumes are able to overshoot a bit higher above inversion, helping the clouds to develop more efficiently on the vertical, without significantly affecting the other aspects.

A possible interpretation of the above result, therefore, is that the air parcels that really contribute to vertical transport and should then be targeted by the parameterization, are the core of the plumes, which are less subject to entrainment. This highlights the importance of being able to sample structures responsible for the vertical transport in LES but also raises the question about the degree to which the internal variables should be tuned against some equivalent diagnostic in the LES. As already explained, LES were used to inspire the parameterizations, that is, to identify the mathematical functions that relate internal variables to the large scale state variables, and then to compute the tendencies to be incremented on those state variables. The representation of this final tendency, and its dependency to input state variables may be seen as more important targets than the accurate representation of internal variables, suggesting not to push too far the procedure of fitting the details of those internal variables. However, a correct profile of vertical velocity or entrainment may be needed if these variables are used in other parts of the model, that is, parameterizations of micro-physics. The automatic tools presented here now permit us to address such questions in more detail.

6.3. Learning From the Various Configurations Tested

In order to check the importance of vertical resolution change versus the fact of varying parameters which were fixed so far, we superimpose in Figure 20 for the 9 parameters explored, the range of parameter values obtained at the end of the multi-wave history matching when the NROY space was not empty.

Consistently with the lesser reduction of the NROY space seen in Figures 5 and 10, using a finer grid (L95, red) reduces less the parameter range compared to the coarser grid (L79, black) when the same setup is used for the history matching, both for the 3-parameter tuning of Section 4 with $\Gamma_z = 0.2$ (dashed curves) or for the 9-parameter tuning of Section 5 with $\Gamma_z = 0.12$. For the 3 parameters which were varied in both setups on the other hand, allowing for varying the other parameters or not matters more than the vertical resolution. As said above, the acceptable values of the **B1** parameter are much smaller than the nominal value (less entraining plume) compensated by a larger value of **DZ** (to favor detrainment below inversion). Note also that, by giving a nonzero value of the **CQ** parameter, with a range which is relatively both narrow and consistent across configurations, the history matching done here demonstrates unambiguously the need for a dependency of the detrainment on the contrast of water between the cloud and its environment. Note also that the 1D test cases and associated metrics used here are much more constraining for the **BG1** parameter that controls the width of the sub-cloud distribution outside the plume that for **BG2** associated with the in-plume distribution.

The values of the BEST simulations are shown as well (gold markers). By construction, these values are inside the NROY space of the 30-wave history matching done with the SCM, that corresponds to the last full red line on the right of each panel (for $\Gamma_z = 0.03$) of Figure 20. However, it may happen that the value shown on the graph is slightly out of this range (for **DZ** and **BG1**). It is due to the fact that the range shown here are based on the 90-member experimental design used to run the SCM or GCM, which is too small a number to really explore the full parameter range. A bit surprisingly maybe, the BEST simulations do not seem to favor a particular sub-range of parameters. This may be related to the fact that the BEST simulations are those which have the good compensation to obtain the right global radiative balance at TOA.

6.4. Keeping Physics at the Model Heart

Note that having a reasonable representation of mass fluxes at the core of boundary-layer parameterizations is important to ensure the robustness of the parameterizations when exploring very different regimes from those which were explored in the SCM/LES machine learning sequence. It also allows us to transport any sort of tracer with the mass flux without needing an additional tuning of the tracer tendencies. On the other hand, a direct application of machine learning to predict the vertical profiles of heating, moistening and wind acceleration from the model state variables, as proposed by Brenowitz and Bretherton (2018), Gentine et al. (2018), and Krasnopolsky et al. (2013), would offer no guarantee that the model behavior would be at all physical for these “out of sample” situations, and would require an independent learning for any new combination of atmospheric constituents.

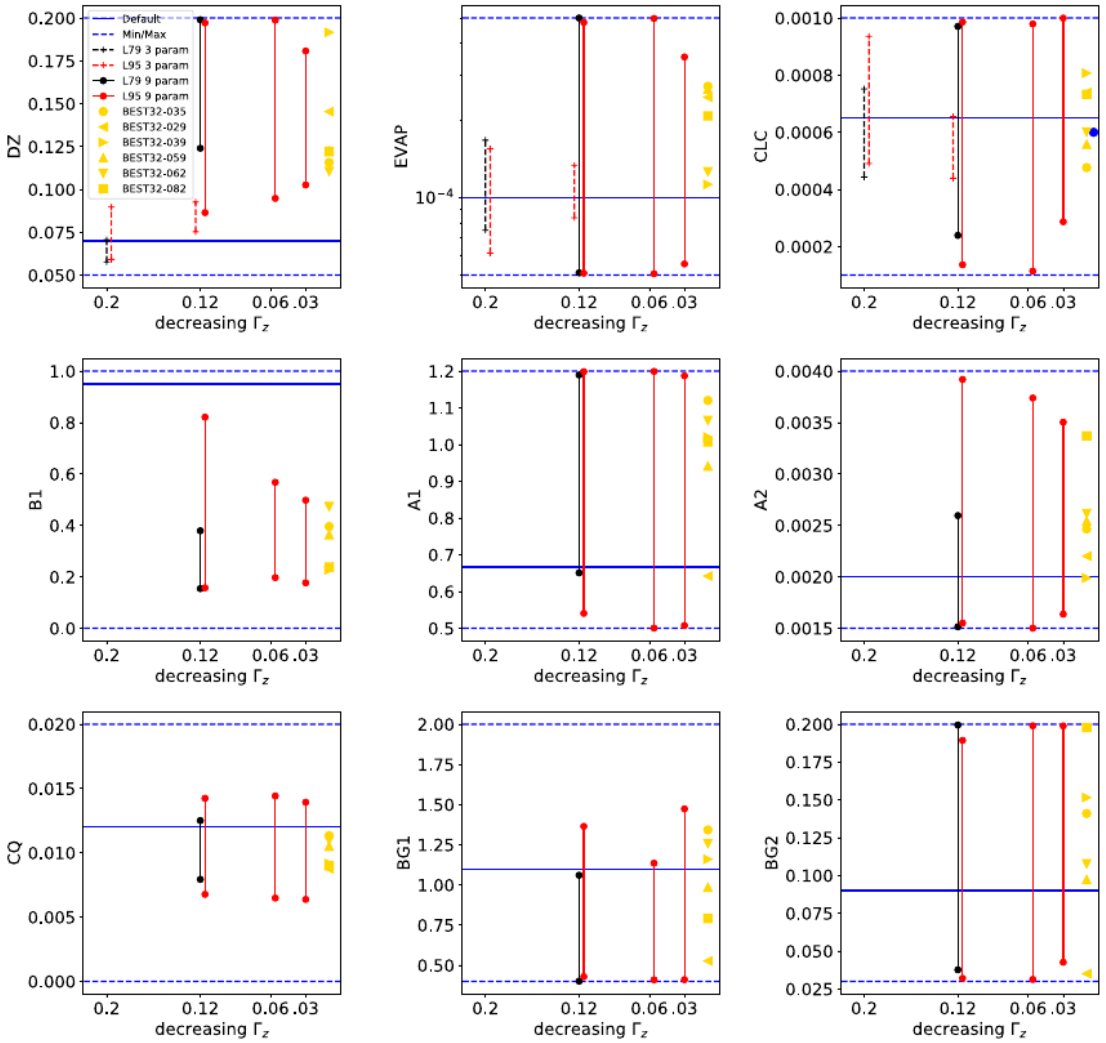


Figure 20. Range of parameters selected by history matching with iterative refocusing with the various configurations tested. For each of the 9 parameters varied in this study, we show: the nominal value (full blue line) and the a priori [min, max] range used for history matching; the range of parameters obtained at the end of the 20-wave 3-parameter history matching of Section 4 (dashed line, computed as the minimum and maximum values of the 45-member experimental design of wave #20) and at the end of the 30-wave 9-parameter history matching of Section 5 (full lines, computed from the 90-member experimental design), showing results for the L79 (black) and L95 (red) configurations when the NROY was not empty at the end of the process; the parameters of the BEST simulations (gold markers). For the BEST1 simulation (circles), the retuned value of the **CLC** parameter chosen for the coupled simulation is shown as well (blue). On each graph, the *x*-axis shows the Γ_z parameter and the *y*-axis the parameter value. NROY, not-ruled out yet.

7. Conclusions

This study presents a first proof of concept of the use of history matching to go from an improvement at process level to a new model configuration applying a systematic and objective approach. It uses in particular the *High-Tune Explorer* tool that we intend to distribute freely to the community of climate modelers.

The availability of this tool does not in any way detract from the importance of the modelers expertize. It must be underlined indeed that the results presented here were obtained after significant work was done by the authors in tuning the 6A version of the LMDZ model by hand. So a good idea of the relevant metrics to be used and associated error was already there, a key ingredient for the success of the history matching procedure. We must, therefore, underline the following point: the tool is automatic and objective in the sense that, once one has specified physically relevant and useful metrics, their measurement errors and tolerance to model error, the procedure will locate the conforming parameter space automatically. The choice of those

metrics and tolerances is and will remain, however, a subjective expert judgment. The number of uses of a climate model is almost infinite (let's just consider so-called impact studies on any location over the globe), and so is the number of possible metrics. Discussing the advantages and rationale for the choice of particular sets of metrics and tolerance will not disappear. However, it is now possible to quantify the impact of such choices and to do so far more quickly than before.

A by-product of the present study is to suggest that the standard 6A version of the LMDZ model was probably rather well tuned, at least for the parameters considered here. Note that the 3D retuning presented here was obtained without varying the parameters that control convection and high clouds. Including such parameters in the tuning process may allow the 3D tuning to be pushed further. In parallel to the illustrations presented here, we have already run 20-parameter history matching experiments with the 3D GCM that show very promising results.

Altogether, this tuning process may seem quite costly. Each SCM simulation used here lasts between half a day and three days depending on the case (typically 1 s CPU time on an intel processor). Typically, 10 days altogether for one parameter choice. With 20 waves of 100 simulations, it is like running 1 day of simulation on a 200 × 100 grid (typically a lower bound of the current CMIP grid size). Even with a larger number of cases, days and parameter space, this step will remain cheap. The following 3D waves are much costlier. This cost is proportional to the required sample size, itself being typically proportional to the number of parameters. A lot can be done for radiative effect of clouds with 1-year long simulations forced by SST, which already means hundreds of simulations. Note however that those hundreds simulations can be run with a perfect scalability on large parallel computers. Note also that control coupled atmosphere-ocean simulations typically last 1,000 years to reach a quasi-steady state of the deep ocean. The tuning of the IPSL-CM6A configurations, including atmospheric tuning and long-term coupled simulations is equivalent to about 20,000 years run over the 2 years of the model preparation. In order to save computer time, various strategies are foreseen like using coarser grid for preconditioning the finer grid tuning, using short-term simulations with nudged winds, etc. The transition from forced-by-SST to coupled simulations will be an important practical issue as well.

One point to notice in terms of cost, is that more metrics than presented here can be applied to each wave, once a series of GCM or SCM simulations have been run. It was not that easy so far with the version of the *High-Tune Explorer* tool used for the present paper, but a much faster one (by orders of magnitude) is available now. However, by increasing the number of constraints, in particular issued from the increasing number of global satellite reference products, or the number of SCM test cases, it may become difficult to find parameter ranges that overlap enough to achieve agreement across the board. This issue, however, is not a limitation of the method. On the contrary, the proposed method makes it possible to start address it. Determining which tolerance to error is needed to find a not empty NROY space, knowing the other sources of errors, is a way to give access to a quantification of the model structural error concerning the metrics added in the process, that is, on the limits of the model physical content and its ability to match so many metrics. We intentionally limited the number of global metrics here, with a focus on radiation. Our belief is that the requirement we put on the radiative forcing of the circulation is a minimum prerequisite to get a reasonable distribution of SSTs in the coupled model, which in turn will condition many aspects of the climate. However, we are already experimenting with a different setup than the one presented here the addition of global metrics, in particular with respect to rainfall. Independently of finding a better configuration for our next generation model version, we would like to explore, within a NROY space constrained by SCM cases and global radiative metrics, the possible worlds that the GCM is able to produce in terms of rainfall distribution, tropical variability or climate sensitivity.

Without anticipating the research spaces thus opened, we can already see that the preconditioning of 3D GCM tuning by SCM simulations is extremely efficient and should be generalized. It requires a rigorous definition of the LES and SCM setups, to avoid compensating for setup errors during the tuning process, as well as testing the model in a configuration that creates some unwilled numerical problems specific to the 1D framework.

Extension of the set of LES test cases is an issue as well. In particular, it would be very important to share well-established and validated LES configurations with deep convection and high clouds if wanting to

obtain for the tuning of convection and high clouds a similar gain in efficiency as the one obtained here for boundary layer convection and associated clouds.

By carrying out this systematic work and sharing the tools with other teams, and by promoting this approach of tuning combining a series 1D cases with 3D simulations, we hope to achieve a faster and more efficient improvement of the climate models involved in the anticipation of climate change. We hope that, relieved of the burden of manual calibration, model developers will spend far more time proposing new ideas for physics-based parameterizations and testing them in global models.

Data Availability Statement

The *High-Tune Explorer* (htexplo) is available through the open source version control system "subversion" (svn) at http://svn.lmd.jussieu.fr/HighTune. A snapshot version of the source codes used for this study is available at DOI https://doi.org/10.14768/70efa07b-afe3-43a4-8334-050354f9deac. The data that supports this research, the results of the SCM simulations and history matching, as well as the scripts for visualization are available at DOI https://data.ipsl.fr/catalog/srv/eng/catalog.search#/metadata/29fb-fe70-a8e8-41db-914c-b14be9a6f90b. The corresponding DOIs will be provided during galley proofs by placeholder "IPSL data catalog." For the GCM simulations, only the preprocessed (netcdf format) climatologies are made available, both for the CMIP simulations taken from the The Earth System Grid Federation (ESGF) and the specific tuning simulations run with LMDZ.

Acknowledgments
This work received funding from grant HIGH-TUNE ANR-16-CE01-0010. It was supported by the DEPHY research network (a french GdR), funded by INSU/CNRS and MeteoFrance. The 3D simulations were granted access to the HPC resources of IDRIS under the allocation gencmip6 attributed by GENCI (Grand Equipment National de Calcul Intensif) and the resources of TGCC from a Prace allocation to the "QUEST" project. Daniel Williamson was funded by N0045RC grant: NE/N018486/1 and by the Alan Turing Institute project "Uncertainty Quantification of multi-scale and multiphysics computer models: applications to hazard and climate models" as part of the grant EP/N510129/1 made to the Alan Turing Institute by EPSRC.

References

Andrianakis, I., Vernon, I., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., et al. (2017). History matching of a complex epidemiological model of human immunodeficiency virus transmission by using variance emulation. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(4), 717–740. https://doi.org/10.1111/rssc.12198 Retrieved from https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12198

Ayotte, K. W., Sullivan, P. P., Andrén, A., Doney, S. C., Holtslag, A. A., Large, W. G., et al. (1996). An evaluation of neutral and convective planetary boundary-layer parameterizations relative to large eddy simulations. *Boundary-Layer Meteorology*, *79*, 131–175.

Bellprat, O., Kotlarski, S., Lüthi, D., & SchäR, C. (2012). Objective calibration of regional climate models. *Journal of Geophysical Research*, *117*(D23), D23115. https://doi.org/10.1029/2012JD018262

Berg, W., L'Ecuyer, T., & Haynes, J. M. (2010). The Distribution of Rainfall over Oceans from Spaceborne Radars. *Journal of Applied Meteorology and Climatology*, *49*(3), 535–543. https://doi.org/10.1175/2009JAMC2330.1

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018GL078510

Bretherton, C., & Smolarkiewicz, P. (1989). Gravity waves, compensating subsidence and detrainment around cumulus clouds. *Journal of the Atmospheric Sciences*, *46*, 740–759.

Brown, A., Cederwall, R., Chlond, A., Duynkerke, P., Golaz, J.-C., Khairoutdinov, M., et al. (2002). Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Quarterly Journal of the Royal Meteorological Society*, *128*, 1075–1093.

Couvreux, F., Guichard, F., Redelsperger, J. L., Kiemle, C., Masson, V., Lafore, J. P., et al. (2005). Water-vapour variability within a convective boundary-layer assessed by large-eddy simulations and IHOP_2002 observations. *Quarterly Journal of the Royal Meteorological Society*, *131*, 2665–2693.

Couvreux, F., Hourdin, F., & Rio, C. (2010). Resolved versus parametrized boundary-layer plumes. Part I: A parametrization-oriented conditional sampling in large-Eddy simulations. *Boundary-Layer Meteorology*, *134*, 441–458. https://doi.org/10.1007/s10546-009-9456-5

de Roode, S. R., Siebesma, A. P., Jonker, H. J., & de Voogd, Y. (2012). Parameterization of the vertical velocity equation for shallow cumulus clouds. *Monthly Weather Review*, *140*(8), 2424–2436.

Duan, Q., Di, Z., Quan, J., Wang, C., Gong, W., Gan, Y., et al. (2017). Automatic model calibration: A new way to improve numerical weather forecasting. *Bulletin of the American Meteorological Society*, *98*(5), 959–970. https://doi.org/10.1175/BAMS-D-15-00104.1

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock?. *Geophysical Research Letters*, *45*, 5742–5751. https://doi.org/10.1029/2018GL078202

Grandpeix, J., & Lafore, J. (2010). A density current parameterization coupled with Emanuel's convection scheme. Part I: The models. *Journal of the Atmospheric Sciences*, *67*, 881–897. https://doi.org/10.1175/2009JAS3044.1

Gregory, D. (2001). Estimation of entrainment rate in simple models of convective clouds. *Quarterly Journal of the Royal Meteorological Society*, *127*, 53–72.

Hourdin, F., Couvreux, F., & Menut, L. (2002). Parameterisation of the dry convective boundary layer based on a mass flux representation of thermals. *Journal of the Atmospheric Sciences*, *59*, 1105–1123.

Hourdin, F., Grandpeix, J.-Y., Rio, C., Bony, S., Jam, A., Cheruy, F., et al. (2013). LMDZ5B: The atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection. *Climate Dynamics*, *40*, 2193–2222. https://doi.org/10.1007/s00382-012-1343-y

Hourdin, F., Găinusă-Bogdan, A., Braconnot, P., Dufresne, J.-L., Traore, A.-K., & Rio, C. (2015, December). Air moisture control on ocean surface temperature, hidden key to the warm bias enigma. *Geophysical Research Letters*, *42*, 10. https://doi.org/10.1002/2015GL066764

Hourdin, F., Jam, A., Rio, C., Couvreux, F., Sandu, I., Lefebvre, M.-P., et al. (2019). Unified parameterization of convective boundary layer transport and clouds with the thermal plume model. *James*, (Vol. *11*), (9), 2910–2933. https://doi.org/10.1029/2019MS001666

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*, 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. (2020a). LMDZ6A: The atmospheric component of the IPSL climate model with improved and better tuned physics. *James*, *12*(7), e01892. http://dx.doi.org/10.1029/2019MS001892

Hourdin, F., Rio, C., Jam, A., Traore, A.-K., & Musat, I. (2020b). Convective boundary layer control of the sea surface temperature in the tropics. *James*, Vol. *12*(6), e01988. https://doi.org/10.1029/2019MS001988

Huffman, G. J., Adler, R. F., Bolvin, D. T., Curtis, S., Joyce, R., & Morrissey, M. M. (2001). Global Precipitation at One-Degree Daily Resolution from Multisatellite Observations. *Journal of Hydrometeorology*, *2*, 36–50. https://doi.org/10.1175/1525-7541

Jam, A., Hourdin, F., Rio, C., & Couvreux, F. (2013). Resolved versus parametrized boundary-layer plumes. Part III: Derivation of a statistical scheme for cumulus clouds. *Boundary-Layer Meteorology*, *147*, 421–441. https://doi.org/10.1007/s10546-012-9789-3

Köhler, M., Ahlgrimm, M., & Beljaars, A. (2011). Unified treatment of dry convective and stratocumulus-topped boundary layers in the ECMWF model. *Q. J. R. Meteorol. Soc*, *137*, 43–57. https://doi.org/10.1002/qj.713

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013). Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artifical Neural Systems*, *203*(3), 13. https://doi.org/10.1155/2013/485913

Loeb, N. G., Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., et al. (2009). Toward optimal closure of the earth's top-of-atmosphere radiation budget. *Journal of Climate*, *22*(3), 748–766. https://doi.org/10.1175/2008JCLI2637.1

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., et al. (2012). *Tuning the climate of a global model*, (Vol. 4). https://doi.org/10.1029/2012MS000154

Rio, C., & Hourdin, F. (2008). A thermal plume model for the convective boundary layer: Representation of cumulus clouds. *Journal of the Atmospheric Sciences*, *65*, 407–425.

Rio, C., Hourdin, F., Couvreux, F., & Jam, A. (2010). Resolved versus parametrized boundary-layer plumes. Part II: Continuous formulations of mixing rates for mass-flux schemes. *Boundary-Layer Meteorology*, *135*, 469–483. https://doi.org/10.1007/s10546-010-9478-z

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

Sandu, I., & Stevens, B. (2011). On the factors modulating the stratocumulus to cumulus transitions. *Journal of the Atmospheric Sciences*, *68*, 1865–1881. https://doi.org/10.1175/2011JAS3614.1

Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., et al. (2014). Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive. *Journal of Advances in Modeling Earth Systems*, *6*, 141–184. https://doi.org/10.1002/2013MS000265

Siebert, P., & Frank, A. (2003). Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode. *Atmospheric Chemistry and Physics Discussions*, *3*, 4515–4548.

Simpson, J., & Wiggert, V. (1969). Models of precipitating cumulus towers. *Monthly Weather Review*, *97*(7), 471–489.

Stephens, G. L., Li, J., Wild, M., Clayson, C. A., Loeb, N., & Kato, S. (2012). An update on Earth's energy balance in light of the latest global observations. *Nature Geoscience*, *5*(10), 691–696. https://doi.org/10.1038/ngeo1580

Sundqvist, H. (1978). A parameterization scheme for non-convective condensation including prediction of cloud water content. *Quarterly Journal of the Royal Meteorological Society*, *104*, 677–690. https://doi.org/10.1002/qj.49710444110

Sundqvist, H. (1988). *Parameterization of condensation and associated clouds in models for weather prediction and general circulation simulation physically-based modelling and simulation of climate and climatic change*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, *93*, 485–498. https://doi.org/10.1175/BAMS-D-11-00094.1

vanZanten, M. C., Stevens, B., Nuijens, L., Siebesma, A.P., Ackerman, A.S., & Burnet, F. (2011). Controls on precipitation and cloudiness in simulations of trade-wind cumulus as observed during RICO. *Journal of Advances in Modeling Earth Systems*, *3*(2), M06001. https://doi.org/10.1029/2011MS000056

Vignon, E., Hourdin, F., Genthon, C., Gallée, H., Bazile, E., Lefebvre, M.-P., et al. (2017). Antarctic boundary layer parametrization in a general circulation model: 1-D simulations facing summer observations at Dome C. *Journal of Geophysical Research*, *122*, 6818–6843. https://doi.org/10.1002/2017JD026802

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

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

Williamson, D., 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

Yamada, T. (1983). Simulations of nocturnal drainage flows by a q^2l turbulence closure model. *Journal of the Atmospheric Sciences*, *40*, 91–106.