

# Process-based climate model development harnessing machine learning: II. model calibration from single column to global

Frédéric Hourdin<sup>1</sup>, Danny Williamson<sup>2</sup>, Catherine Rio<sup>3</sup>, Fleur Couvreur<sup>4</sup>, Romain Roehrig<sup>5</sup>, Najda Villefranque<sup>6</sup>, Ionela Musat<sup>7</sup>, Fatoumata Bint Diallo<sup>8</sup>, Laurent Fairhead<sup>9</sup>, and Victoria Volodina<sup>10</sup>

<sup>1</sup>LMD

<sup>2</sup>University of Exeter

<sup>3</sup>Centre national des recherches météorologiques (CNRM), Université de Toulouse, Météo-France, CNRS

<sup>4</sup>Université Toulouse, CNRM, Météo-France, CNRS

<sup>5</sup>CNRM, Université de Toulouse, Météo-France, CNRS

<sup>6</sup>Centre National de Recherches Météorologiques

<sup>7</sup>IPSL/UPMC/CNRS

<sup>8</sup>Laboratoire de Météorologie Dynamique

<sup>9</sup>LMD/IPSL/CNRS

<sup>10</sup>The Alan Turing Institute

November 24, 2022

## Abstract

We demonstrate a new approach for climate model tuning in a realistic situation. Our approach, described in detail in Part I, systematically uses a single-column configuration of a global atmospheric model on a series of 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 that are consistent with the metrics computed on large eddy simulations within a given tolerance to error. The approach is applied to the recently designed 6A version of the LMDZ model, itself the result of 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. The approach allows us to automatically reach a tuning as good as that of the 6A version, after some improvements are done at process scale. This approach helps accelerate the introduction of new parameterizations, by avoiding a tedious manual tuning process and preventing some of the error compensations that could occur if calibration was carried out directly with the full atmospheric model. This way of using machine learning techniques 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. That is, we get things right for the right reasons.

# Process-based climate model development harnessing machine learning: II. model calibration from single column to global

Frédéric Hourdin<sup>1</sup>, Daniel Williamson<sup>3,4</sup>, Catherine Rio<sup>2</sup>, Fleur Couvreur<sup>2</sup>,  
Romain Roehrig<sup>2</sup>, Najda Villefranque<sup>2</sup>, Ionela Musat<sup>1</sup>, Laurent Fairhead<sup>1</sup>, F.  
Binta Diallo<sup>1</sup>, Victoria Volodina<sup>4</sup>

<sup>1</sup>LMD-IPSL, Sorbonne-Universités, CNRS, 4 pl Jussieu, Paris, 75005, France

<sup>2</sup>CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

<sup>3</sup>Department of Mathematical Sciences, University of Exeter, Exeter, UK.

<sup>4</sup>The Alan Turing Institute, British Library, London, UK

## 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.



## Abstract

We demonstrate a new approach for climate model tuning in a realistic situation. Our approach, described in detail in Part I, systematically uses a single-column configuration of a global atmospheric model on a series of 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 that are consistent with the metrics computed on large eddy simulations within a given tolerance to error. The approach is applied to the recently designed 6A version of the LMDZ model, itself the result of 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. The approach allows us to automatically reach a tuning as good as that of the 6A version, after some improvements are done at process scale. This approach helps accelerate the introduction of new parameterizations, by avoiding a tedious manual tuning process and preventing some of the error compensations that could occur if calibration was carried out directly with the full atmospheric model. This way of using machine learning techniques 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. 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 free model 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 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 the part of the tuning process that can be, the approach enables climate modeler expertise 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 can not 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  $\mathbf{x}$  of the collective behavior of unresolved processes, which at the end appears as a source term  $S_{\mathbf{x}} = \mathcal{P}_p(\mathbf{x}, \boldsymbol{\lambda}_p)$  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 heat profile. Each parameterization depends on a set of free parameters  $\boldsymbol{\lambda}_p$ , some of which have a physical meaning (maxi-

68 mum water content of clouds, fall speed of ice crystals), some others resulting from the  
 69 simplifications inherent to any parameterization (representing an ensemble of plumes  
 70 by a single plume for example). Convective and cloud parameterizations are often  
 71 developed in a single column model (SCM) framework by comparison with large eddy  
 72 simulations (LES) of the same atmospheric column, in which convective motions are  
 73 explicitly resolved. This SCM/LES comparison is used both to inspire parameteri-  
 74 zation development and to choose, calibrate or “tune” the model free parameters  $\lambda_p$   
 75 at process level. Once integrated in operational models, those parameterizations are  
 76 active in each atmospheric column of the model, influencing both the global radiation  
 77 budget and the large-scale circulation.

78 The development of a reference configuration of a climate model, as those in-  
 79 volved in the Coupled Model Intercomparison Program (Taylor et al., 2012, CMIP),  
 80 requires an intense phase of adjustment including – grid choice, bug corrections, activa-  
 81 tion of some parameterizations or code modifications in which the tuning or calibration  
 82 of free parameters is key. A survey on climate model tuning revealed rather standard  
 83 priorities, which consist of targeting the radiative forcing of the atmospheric circula-  
 84 tion, thereby using model free parameters that most affect radiation, i. e. cloud  
 85 parameters (Hourdin et al., 2017). The complexity of the tuning process, given the  
 86 large number of free parameters, the large number of possible targets, and the lack of  
 87 specific research in this area, probably partly explain the slowness of climate model  
 88 improvements. Typically, the tuning phase of the IPSL coupled model configuration  
 89 for CMIP6 (IPSL-CM6A-LR) took more than two years, with repeated tuning phases  
 90 targeting improvement of the radiative forcing of the circulation: global radiation,  
 91 decomposed in terms of short-wave (SW) and long-wave (LW), clear-sky and cloud  
 92 radiative effect (CRE), and some spatial variations of those fluxes like contrasts be-  
 93 tween mid-latitude and tropics, or between convective and subsiding regimes in the  
 94 tropics. Such a tuning was done in practice each time a new version of the coupled  
 95 model with significant changes was proposed. In total, 15 successive versions were  
 96 tuned this way. For each version, systematic sensitivity experiments to 3–10 param-  
 97 eters were done with the stand-alone-atmospheric model forced by imposed sea surface  
 98 temperature (SST) on a couple of years, changing the parameters one by one. Then di-  
 99 agnostics were computed and, by trial and error, a new radiative tuning was proposed  
 100 and tested. Each of the 15 versions of the global model typically needed one to five  
 101 iterations of this tedious sensitivity analysis. Among the limitations of the approach,  
 102 it can be done only by local perturbation around the previous tuning and it explores  
 103 independently the dependency to each individual parameter, hiding any compensating  
 104 effects between them. During all of these processes, a series of SCM test cases were  
 105 run and compared with LES in order to ensure that the model tuning was not pushed  
 106 too far, at the risk of deteriorating the model behaviour at process level.

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

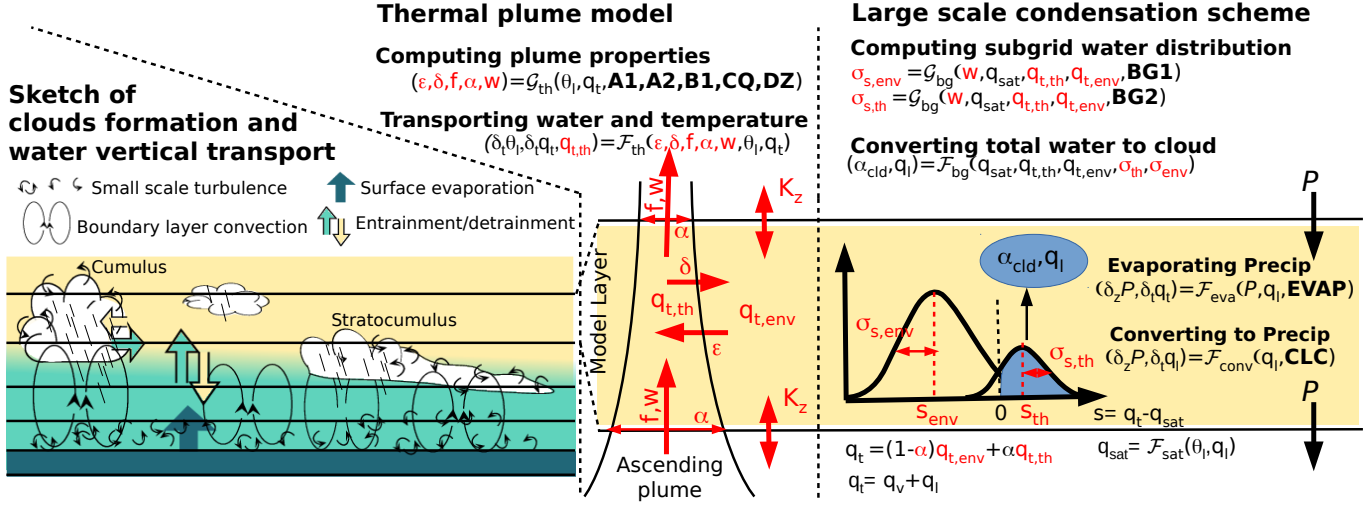
In Part 1, we presented the rationale and technical details of the approach with a simple illustration. The objective of this second part is to demonstrate how this framework can be used to accelerate the process of model development, from the process-based inspiration of new parameterizations to the full 3D GCM tuning. 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; Hourdin, Rio, Jam, et al., 2020; Hourdin, Rio, Grandpeix, et al., 2020) 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  $\mathcal{P}_1(\mathbf{x}, \boldsymbol{\lambda}_1)$  behaves better than an old version  $\mathcal{P}_0(\mathbf{x}, \boldsymbol{\lambda}_0)$ , 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. 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



**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 subgrid 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.

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 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 Fig. 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  $\rho$  is the air density,  $\alpha$  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  $\psi$  within the thermal plume  $\psi_{th}$  is computed using the stationary plume conservation equation:

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

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  $\psi$ ). For variables conserved by the convective transport, such as liquid potential temperature  $\theta_l$  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  $\psi$  is

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

where  $K_z = l_{\text{mix}} S(Ri) \sqrt{\text{TKE}}$  is the eddy diffusivity,  $l_{\text{mix}}$  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

$$\frac{\partial f}{\partial z} = e - d \quad (3)$$

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 (Couvreur 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  $\psi_{th}$  of any conserved quantity  $\psi$  inside the plume. Knowing  $f$ ,  $\psi$  and  $\psi_{th}$ , one can then invert the conservation equation of the mass flux (Eq. 3) and  $\psi$  (Eq. 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.

Following a proposition by Simpson and Wiggert (1969), most convective parameterizations use a momentum equation which assumes that subplume turbulent fluctuations and nonhydrostatic pressure perturbations reduce buoyancy and act as a drag term proportional to entrainment (see de Roode et al., 2012, for a discussion of the validity of this approach for shallow convection). Here, we simply consider turbulence by reducing the buoyancy term and pressure perturbations by adding a constant drag term. It appears as a source term in Eq. 1 for  $\psi_{th} = w_{th}$  and  $\psi = 0$ . It is 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 ( $\theta_v$  being the virtual potential temperature) that accelerates the plume and the second term a drag effect, with  $\mathbf{A1} = 2/3$  and  $\mathbf{A2} = 0.002 \text{ m}^{-1}$ .

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

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

where **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  $\delta$ :

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

where  $\Delta q_t$  is the contrast in humidity between the plume and its environment, with  $\mathbf{CQ} = 0.012 \text{ m}^{-1}$  (the vertical velocity being normalized by  $w_0 = 1 \text{ m s}^{-1}$ ) 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  $\Delta q_t$  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:

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

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

In the current version,  $\delta z = \mathbf{DZ} \times z$ ,  $\mathbf{DZ}$  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  $\alpha$ , the mean saturation deficits within environment,  $s_{env}$ , and plumes,  $s_{th}$  (which are directly given by the thermal plume model), and their associated standard deviations,  $\sigma_{s,env}$  and  $\sigma_{s,th}$ , 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  $s_{th} - s_{env}$  increases, standard deviations are parameterized as follows:

$$\sigma_{s,th} = \mathbf{BG2} (\alpha + 0.01)^{-\gamma_1} (\bar{s}_{th} - \bar{s}_{env}) + b \bar{q}_{t,th} \quad (7)$$

and

$$\sigma_{s,env} = \mathbf{BG1} \frac{\alpha^{\gamma_2}}{1 - \alpha} (\bar{s}_{th} - \bar{s}_{env}) + b \bar{q}_{t,env}, \quad (8)$$

where  $b$ ,  $\mathbf{BG1}$ ,  $\mathbf{BG2}$ ,  $\gamma_1$  and  $\gamma_2$  are free parameters, and the last term,  $b\bar{q}_{t,th}$  or  $b\bar{q}_{t,env}$ , 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,  $b$ ,  $\gamma_1$  and  $\gamma_2$  do not have a dominant role and their values were kept fixed in the results presented here.

The values of  $b = 2 \times 10^{-3}$ ,  $\mathbf{BG1} = 0.92$ ,  $\mathbf{BG2} = 0.09$ ,  $\gamma_1 = 0.4$  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 Eq. 1) and virtual potential temperature involved in the

name	min	max	ref	sampling	controls
<b>A1</b>	0.5	1.2	2./3.	linear	contribution of buoyancy to the plume acceleration
<b>A2</b>	1.5e-3	4.e-3	2.e-3	linear	drag term in the plume acceleration
<b>B1</b>	0.	1.	0.95	linear	scaling factor for entrainment and detrainment
<b>CQ</b>	0.	0.02	0.012	linear	influence of humidity contrast on detrainment
<b>DZ</b>	0.05	0.2	0.07	linear	environmental air altitude shift for buoyancy computation
<b>BG1</b>	0.4	2.	1.1	linear	width of the environment subgrid scale water distribution
<b>BG2</b>	0.03	0.2	0.09	linear	width of the plume subgrid scale water distribution
<b>EVAP</b>	5e-5	5e-4	1e-4	log	reevaporation of rainfall
<b>CLC</b>	1e-4	1e-3	6.5e-4	linear	autoconversion of cloud liquid water to rainfall

**Table 1.** Parameters involved in the iterative refocusing. The minimum and maximum values explored are given as well as the reference value used in the 6A configuration of LMDZ, and the information on whether the parameter is explored with a linear or logarithmic sampling.

278 buoyancy computation. Once  $e$ ,  $d$  and  $f$  are determined, Eq. 1 and Eq. 2 are applied  
 279 to the total water and liquid potential temperature to compute tendencies associated  
 280 with the boundary-layer transport. From the thermal plume model computation, the  
 281 parameters of the bi-Gaussian sub-grid scale distribution,  $F(s)$ , for the saturation  
 282 deficit can be estimated as explained above. From this distribution, the cloud fraction  
 283  $\alpha_{cld} = \int_0^\infty F(s)ds$  and cloud liquid content  $q_l = \int_0^\infty sF(s)ds$  at the grid scale are  
 284 finally computed<sup>1</sup>.

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

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

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

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

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

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

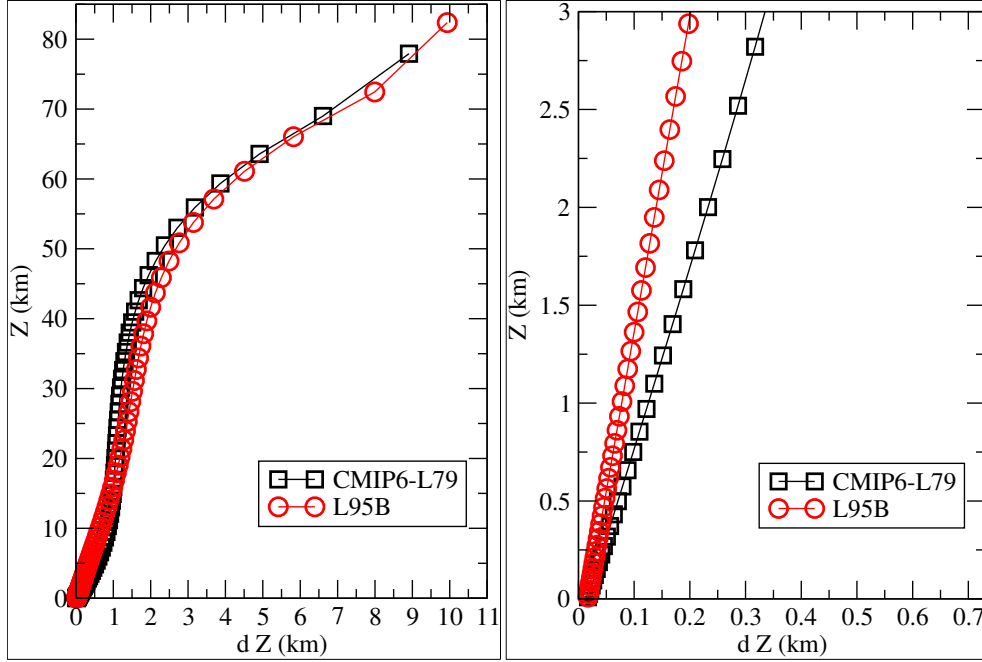
### 289 3 Model setup

#### 290 3.1 The 6A version of LMDZ

291 The parameterizations described here are a crucial piece of the physical pa-  
 292 rameterizations of the LMDZ atmospheric global model. The recent modification  
 293 of the detrainment formulation presented above produced a major improvement in

<sup>1</sup> Note that the same cloud scheme is applied with a single mode of width  $\sigma_{s,env} = b \bar{q}_{t,env}$  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).





**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 three km above surface.

the 6A version, the atmospheric component of the IPSL-CM6A-LR used for CMIP6. This version is extensively described by Hourdin et al. (2020, accepted in James, DOI:10.1029/2019MS001892). 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 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 above surface. The layer thickness is typically  $\Delta z \simeq 0.067z$ . The dependency of layer thickness upon altitude is given in Fig. 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<sub>2</sub>O Project (IHOP) field-experiment. This case is derived from observations collected on 14 June 2002 over the Southern Great Plains (Couvreur et al., 2005).

The second case, ARMCU/REF, is derived from observations collected on 21 June 1997 at the Atmospheric Radiation Measurement site in Oklahoma, U.S.A. (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) 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.

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-day 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).


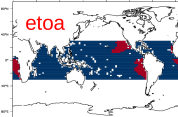
The ARMCU/REF and RICO/REF cases were used extensively for the inspiration, development and assessment of the thermal plume model and bi-gaussian cloud scheme (Couvreur et al., 2010; Rio et al., 2010; Jam et al., 2013). The SANDU cases were at the heart of the work on the modification of the thermal 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 maximum cloud fraction height, 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 expertise and objectives. The particular set of metrics retained here is given in Tab. 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

Case subcase	IHOP REF	ARMCU REF	RICO REF	SANDU REF	SANDU SLOW	SANDU FAST
time	7-9	7-9	19-25	50-60	50-60	50-60
$\theta_{400-600hPa}$	<b>X</b>	X	<b>X</b>			
$q_{v,400-600hPa}$		X				
$\alpha_{cld,max}$		X	X			
$z_{cld,ave}$		<b>X</b>		<b>X</b>		
$z_{cld,max}$		<b>X</b>		X	X	X

**Table 2.** Metrics retained for the SCM/LES tuning. The time retained for time average is given in hours from the beginning of the simulation.

Mask	Variable	Metrics	target $W m^{-2}$	error $W m^{-2}$
	Total rad. TOA (rt) Swup TOA (rsut)	glob.rt glob.rsut circAa.rsut	2.5 99.6 24.0	0.2 5 5
Convective, intermediate, subsiding, Circum Antact. anomaly	SWup TOA (rsut) LWup TOA (rlut)	circAa.rlut subs.rsut weak.rsut conv.rsut subs.rlut	-48.6 84.9 81.8 103.2 274.6	5 5 5 5 5
 Eastern Tropical Ocean anomaly	SWup TOA (rsut)	weak.rsut conv.rsut etoe.rsut	264.3 235.8 11.0	5 5 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  $\sigma$  error retained for the history matching are shown in the table on the right hand side. The target values are computed from the EBAF observational dataset.

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 less sensitive to the tuning. The second aspect targeted by the metrics is the 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 (twelve SST and SIC maps, one per month, interpolated in time with splines). Simulations are run on the standard horizontal grid made up of 144 points in longitude and 143 in latitude (Low resolution or LR).

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 Fig. 3.

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  $0.5\text{--}1\text{ W/m}^2$ , 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  $4\text{ W/m}^2$  (Loeb et al., 2009). In fact, rather than tuning the global radiation to the theoretical value of  $0.5\text{--}1\text{ W/m}^2$ , we rather tuned it to a global imbalance of about  $2.5\text{ W/m}^2$ . We know indeed that, for our particular model, an imbalance of  $2.5\text{ W/m}^2$  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  $0.5\text{ W/m}^2$  in the current IPSL-CM model) and changes in the mean climate 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 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 initial (original) input space is progressively reduced to obtain the Not-Ruled-Out Yet (NROY) space of parameters based on implausibility derived from Gaussian process emulators fitted to each metric, as detailed in Part I. The implausibility itself (Williamson et al., 2013),  $I(\lambda)$ , 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.

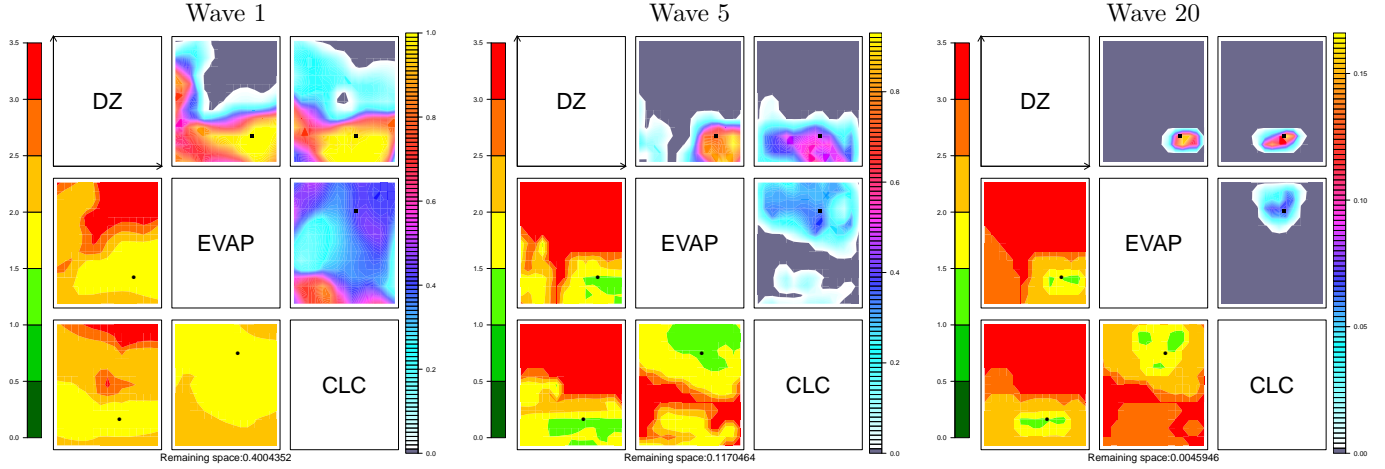
Part I for a complete presentation). A point of the parameter space is kept in the NROY space when the implausibility is smaller than a threshold or cutoff. In all the applications presented below, a series of iterations or waves is done, keeping the same list of metrics at each iteration. The cutoff 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.

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  $\Sigma_T$  and  $\Sigma_q$  and a relative error on the height of clouds  $\Gamma_z = \Sigma_z/z$  and cloud fraction  $\Gamma_{\alpha_{cld}} = \Sigma_{\alpha_{cld}}/\alpha_{cld}$ . 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. The GCM tolerance to error is fixed to the values given in Fig. 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 3 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



**Figure 4.** Implausibility matrices for wave 1, 5, and 20 of an history matching exploration, run with the L79 vertical grid and  $\Gamma_z=0.2$ . Explanation of the building of the figures is given in the text with additional details in Part I.

global model) that there was an optimal value of this parameter, somewhere between 0.05 and 0.15. A value of 0.07 was finally retained in the 6A version.

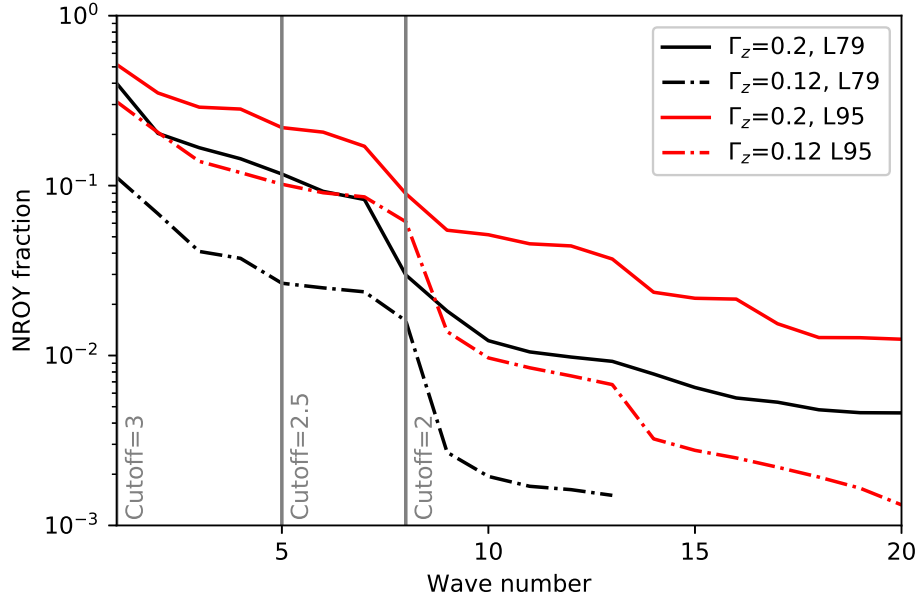
#### 4.1 1D history matching

For this first example, we use five metrics, the ones shown with bold crosses in Tab. 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.

The building of the implausibility matrices shown in Fig. 4 for wave 1, 5 and 20 from left to right is explained in Part I. Each 2D sub-matrix in Fig. 4 is a restriction to two parameters, the names of which are given in the diagonal of the main matrix. Each axis of the sub-matrix is divided into 15 subintervals (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  $\lambda$ , we compute the minimum implausibility and the proportion of points with implausibility lower than the cutoff within each pixel (and so in the dimensions behind it). The latter values are displayed in the sub-matrices of the upper right triangle. The total fraction of the volume of the NROY space relative to the initial space is the average of the matrix, which should be the same for all the sub-triangles. A dark grey colour means that there is no way to fit the observations by varying the third parameter (or N-2 unfixed parameters in a general case) while a value of 100% means that values of the two parameters in x and y axis can be retained whatever the values of the third parameter. In the lower-left triangle, the minimum value of the implausibility is shown. These plots are orientated the same way as those on the upper triangle, for easier visual comparison, so that the labelling of the axis should be inverted for this lower left triangle, compared to the names given on the diagonal.

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.



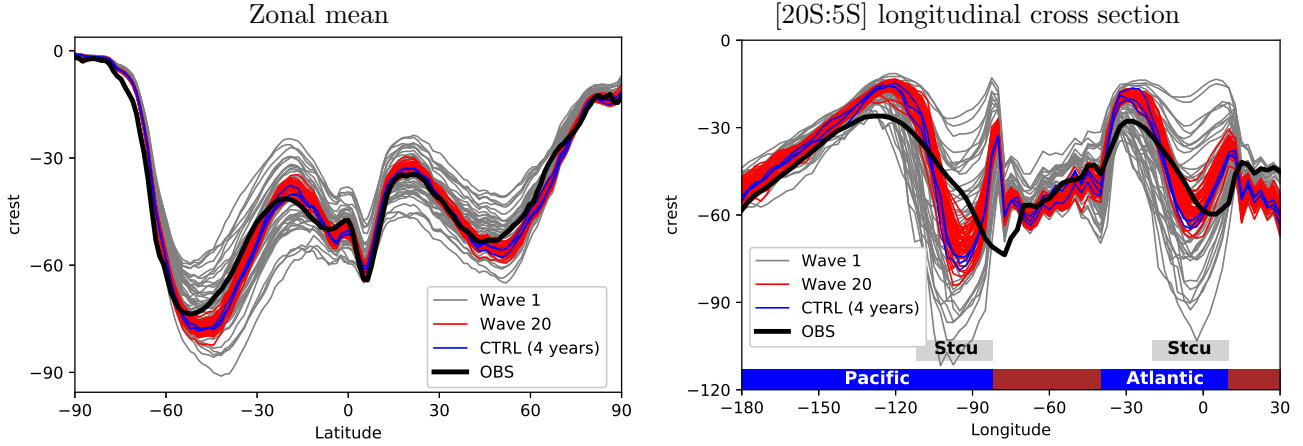


**Figure 5.** Reduction of the volume fraction of the NROY space (compared to the full initial hypercube volume, y-axis) remaining after  $N$  waves of history matching (x-axis) for the L79 and L95 vertical grids and with a relative tolerance to error on the cloud height of  $\Gamma_z=0.12$  and  $0.2$ . 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.

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. Fig. 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  $\Gamma_z = 0.12$  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



**Figure 6.** Zonally average latitudinal variation (left) and latitudinally averaged (between 20S and 5S) zonal variation (right) of the SW Cloud Radiative Effect (CRE) at TOA for 45 simulations run with the sample of parameters used for wave 1 (grey) and a sampling of the NROY space remaining at wave 20 of the SCM history matching (red). The blue curves correspond to year 1 to 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.

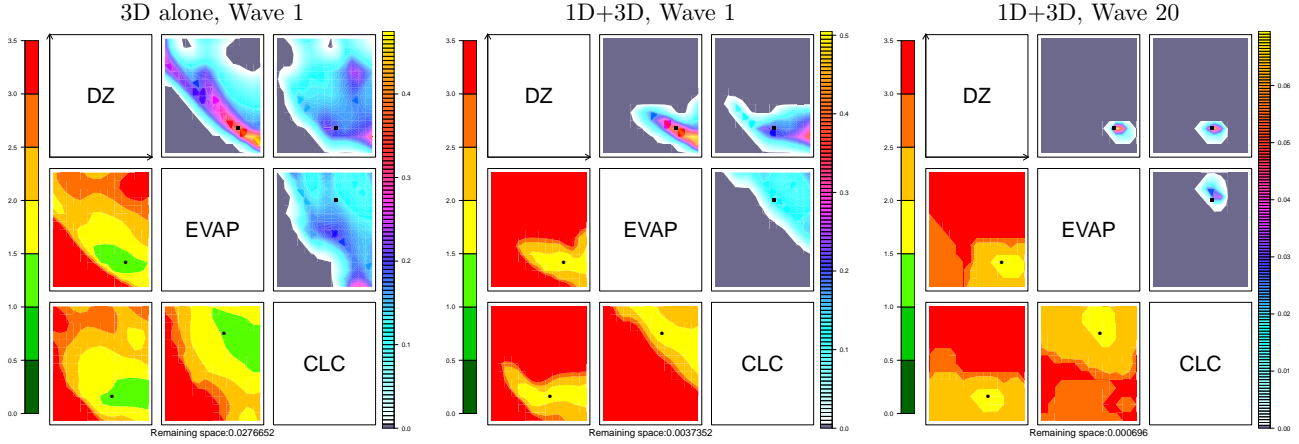
with vertical resolution. However, the sensitivity to the tolerance to errors for the other variables would deserve investigation as well.

#### 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 Fig. 6 shows the mean latitudinal variations of the TOA SW CRE averaged both zonally and annually. While the spread across models is of  $30 \text{ W/m}^2$  before NROY selection, it reduces to a few  $\text{W/m}^2$  at wave number 20. 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 Fig. 6 shows the longitudinal variation of the same SW CRE in the southern tropics. This diagnostic underlines the contrast between a weak cooling in the regions of trade winds cumulus, at around  $130^\circ\text{W}$  in the Pacific ocean and  $40^\circ\text{W}$  over the Atlantic, and strong cooling in the regions of stratocumulus, at  $100^\circ\text{W}$  over the Pacific and at Greenwich longitude over the Atlantic. The large range of SW CRE explored (from  $-20$  to  $-110 \text{ W m}^{-2}$ ) in the stratocumulus regions before any parameter selection (wave 1, grey curves) is consistent with the strong impact of the value of **DZ** (Hourdin 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).





**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, i. e. adding 3D GCM metrics after 20 waves run with the SCM only (right).

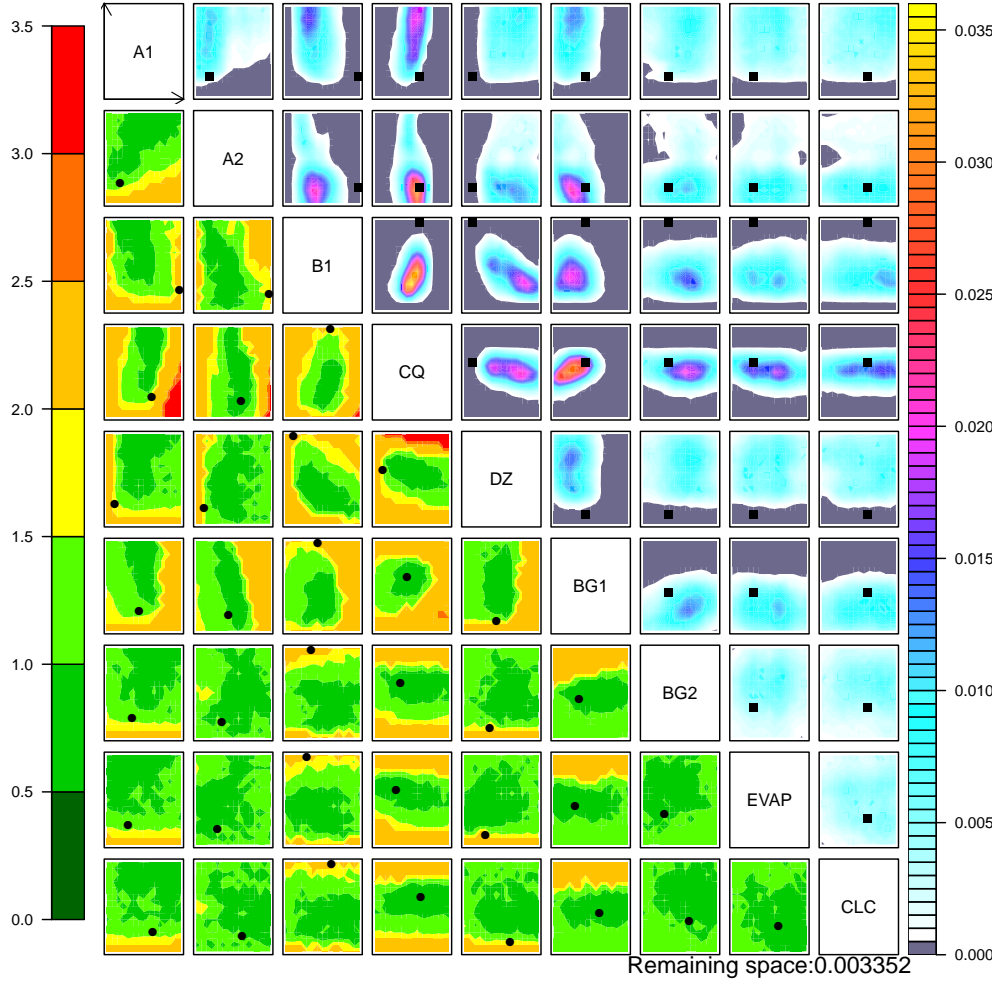
We present in Fig. 7 the implausibility statistics obtained after considering 3D simulations using the 3D metrics presented in Fig. 3. The left panel shows the implausibility matrix, which would be obtained with one single wave without preconditioning by 1D tuning. In this simple case, the selection 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 Fig. 4 reduces the NROY to a small and well defined space which includes the tuning finally retained for the LMDZ6A version.

## 5 Improving the representation of boundary-layer convection

In this second example, we illustrate how tuning can be used together with model development and improvement in a more realistic situation. We now consider revisiting the representation of boundary-layer convection by both increasing the model vertical resolution and re-tuning the thermal plume model internal parameters.

As already explained, during the tuning of the 6A version, the parameters that control the mass flux in the thermal plume model were fixed to values retained during the course of the development of the parameterization. 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 (Rio et al., 2010; Jam et al., 2013). 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 Tab. 1). The tuning process is applied by varying these parameters together with those used in the previous section: **DZ**, **EVAP**, and **CLC**.



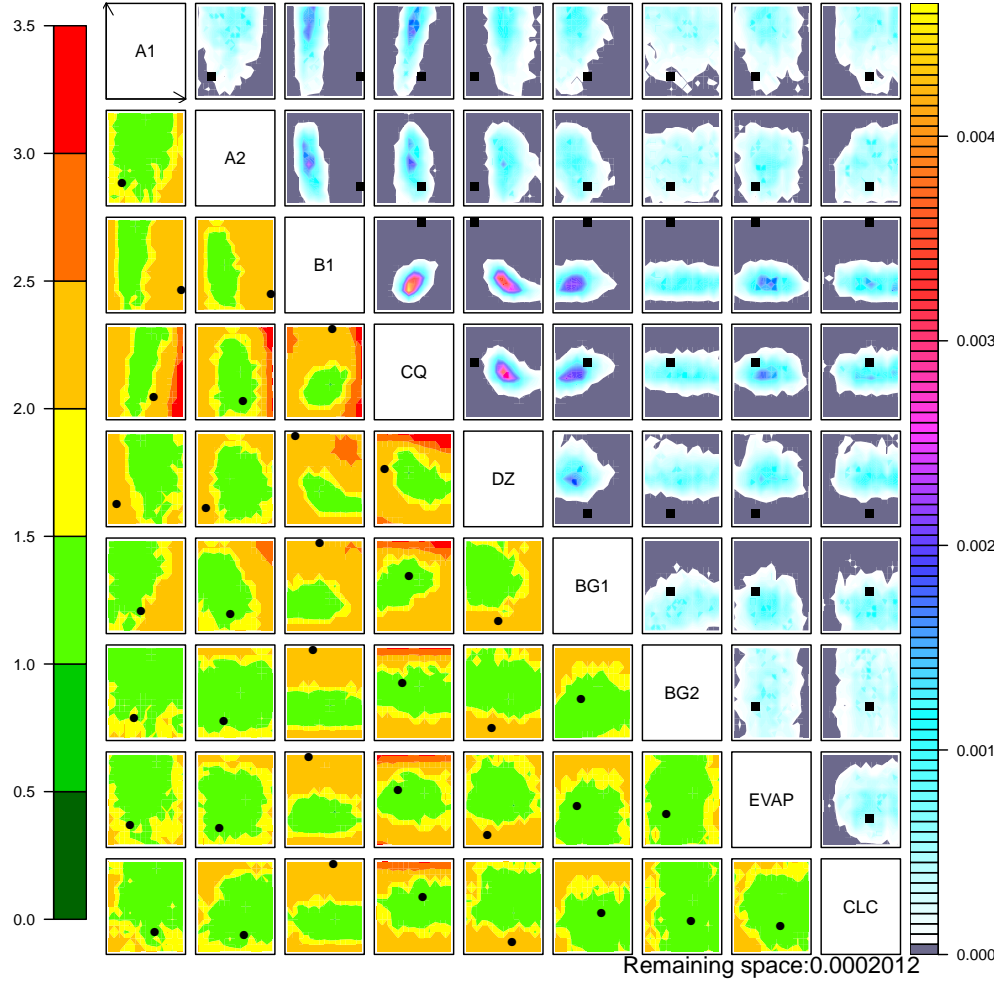
**Figure 8.** Implausibility matrix for the 9-parameter history match after 30 waves, vertical grid L95 and with a relative tolerance to error on the cloud height  $\Gamma_z=0.12$

### 5.1 SCM history matching with 9 parameters

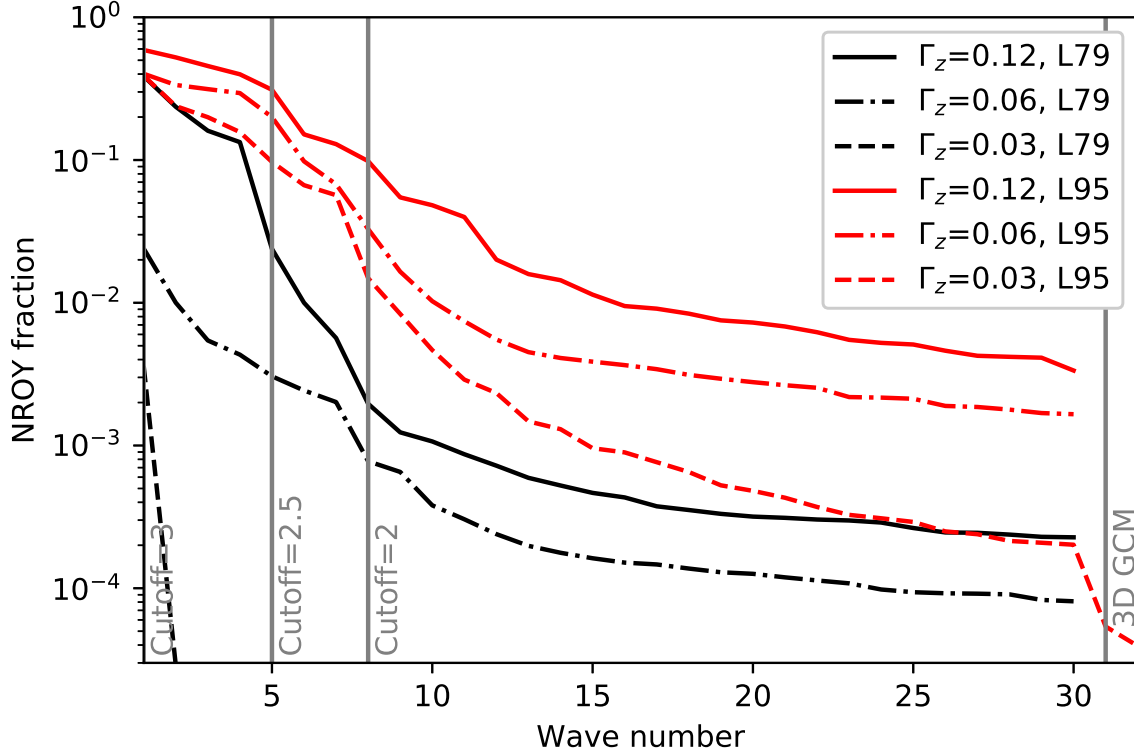
We first perform a 30-wave SCM history match with the extended set of parameters. Note that 20 or 30 waves may sound like a large number, though this has been done in epidemiological studies (Andrianakis et al., 2017), and is inexpensive using the SCM. The NROY matrices are shown in Fig. 8 for  $\Gamma_z=0.12$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

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

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  $\Gamma_z$ , the  $\pm 2\sigma$  tolerance to error is of  $0.12 \times z$ , which is about 1.8 time the layer thickness.
3. For the coarser grid, L79, only the  $\Gamma_z=0.12$  and  $\Gamma_z=0.06$  cases are able to maintain a non zero NROY space after 30 waves.



**Figure 9.** Same as Fig. 8 but with a relative tolerance error on the cloud height of  $\Gamma_z=0.03$ .



**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 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.

4. 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, it is now possible to include the SANDU/SLOW case, which was too badly represented to be considered in the previous section.
5. The NROY is also 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.
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.

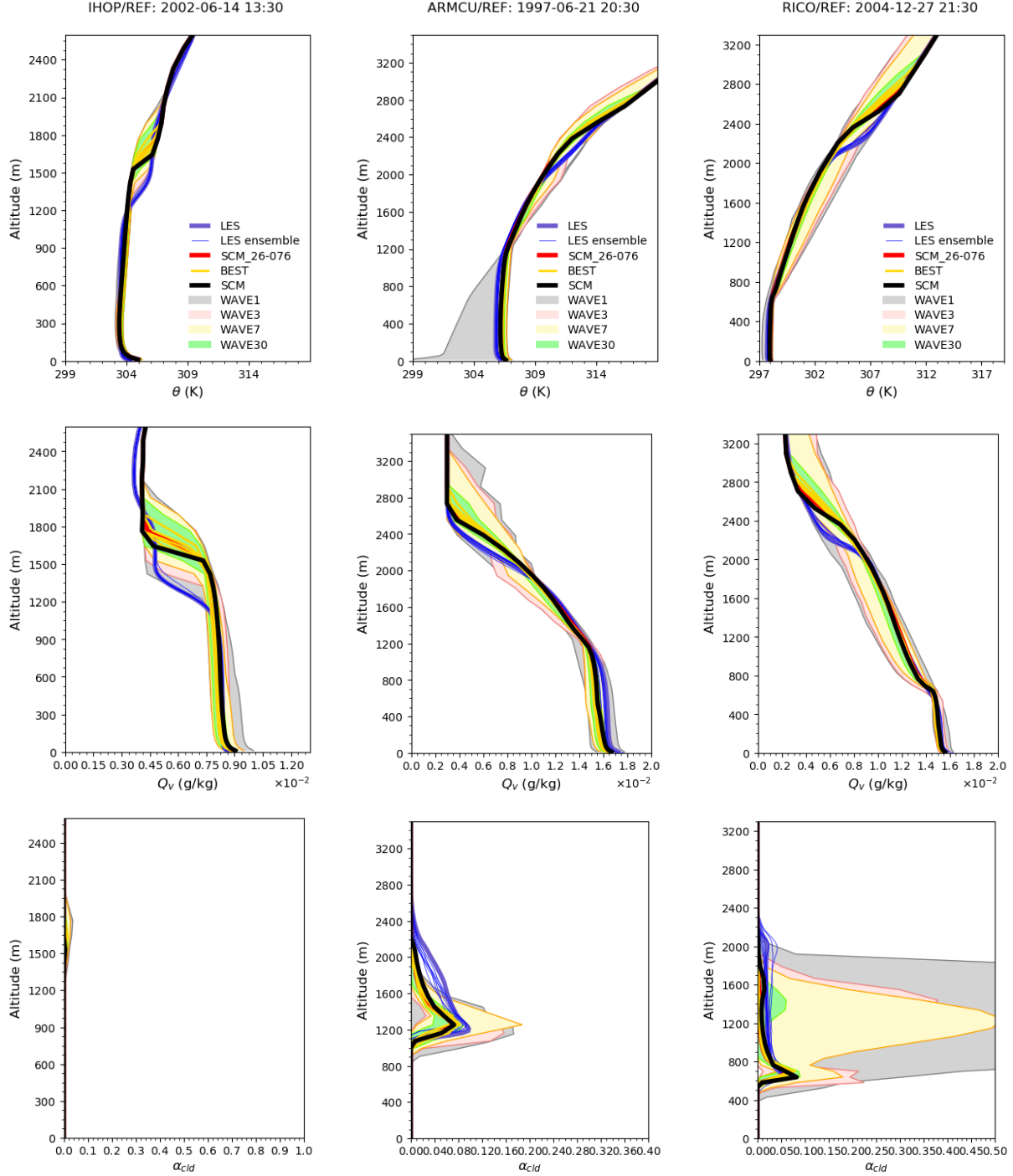
We show in Fig. 11 and Fig. 12, for waves number 1 (grey), 3 (pink), 7 (yellow) and 30 (green), the envelope of the vertical profiles of potential temperature, specific humidity and cloud fraction for the 90 SCM simulations run to build the emulator with the L95 configuration and smallest tolerance to error. For the cumulus cases (Fig. 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 (Fig. 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 was obtained as the 76th element of wave 26 (named SCM-26-076 on the graph).

## 5.2 3D history matching

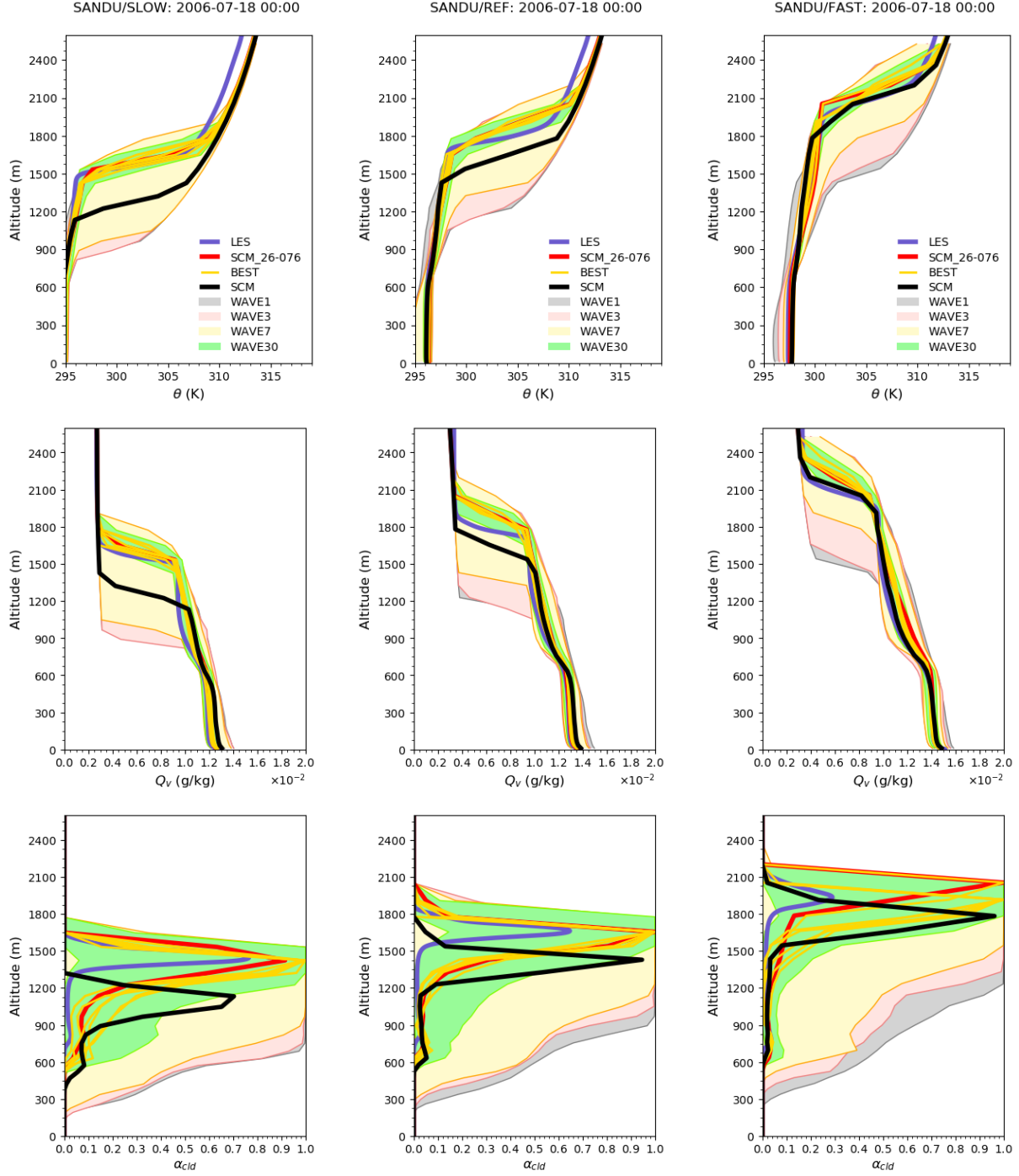
We present here the results of two subsequent waves of history matching with the 3D GCM, starting from wave 30 of the SCM history matching, with the L95 vertical grid and  $\Gamma_z = 0.03$ . For waves 31 and 32, both the previous 12 SCM metrics and the 11 3D GCM metrics presented in Fig. 3 are used. The implausibility graph of wave 32 is shown in Fig. 13. The fraction of the NROY space compared to the initial parameter hyper-cube is reduced from  $2 \cdot 10^{-4}$  at wave 30 to  $4 \cdot 10^{-5}$  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 (Fig. 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 Fig. 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 (grey curves in Fig. 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.

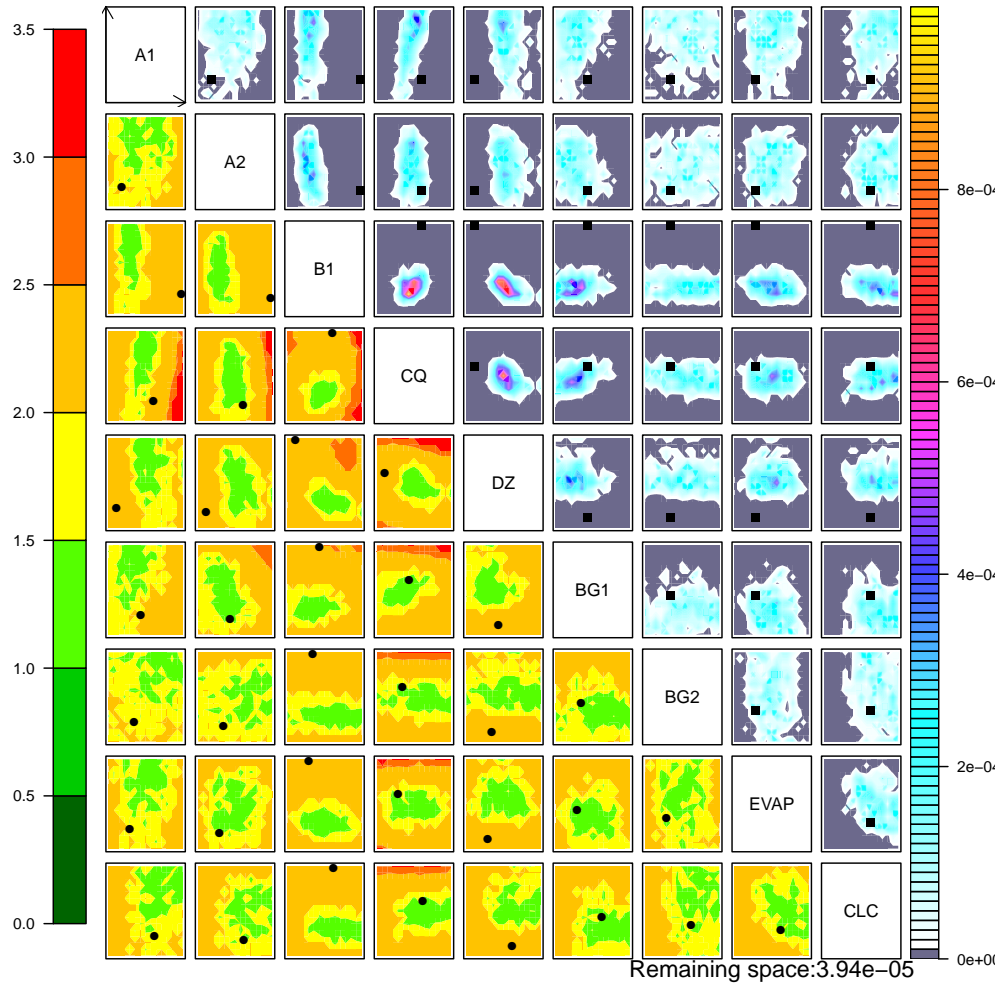
We show in Fig. 15 the normalized (by the tolerance to error) error for the GCM metrics for the 90 GCM simulations run for wave 32. The simulations are ranked according to the maximum value of this normalized error. For most of the simulations, the global net radiative balance 'glob.rt' dominates the error, which is of course partly attributable to the fact that we took an arbitrarily small error of  $0.2 \text{ W/m}^2$  for this particular metrics (targeting a  $0.2 \text{ K}$  in coupled simulations). After the global radiative balance, some metrics are particularly difficult to get within the tolerance to errors,



**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 super-imposed 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).

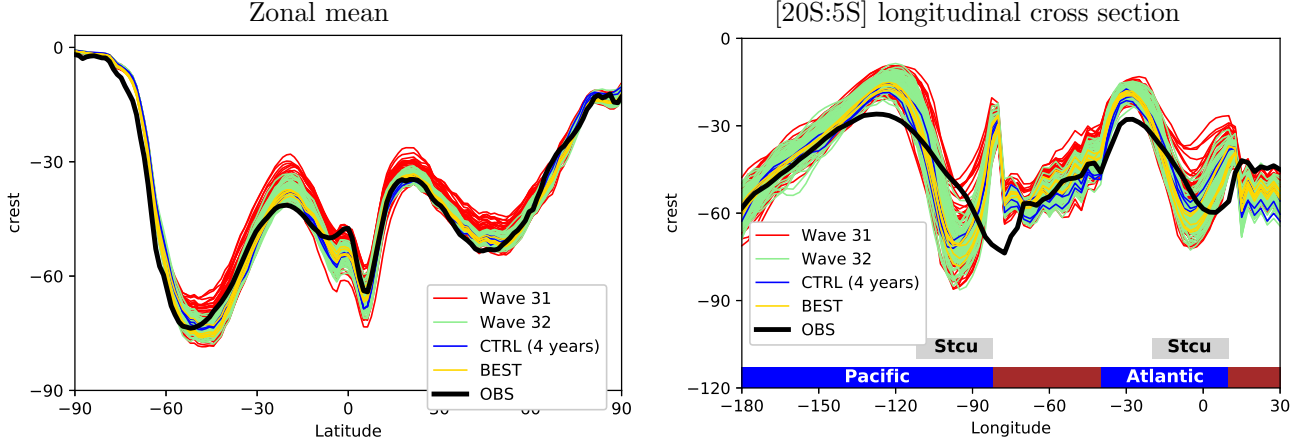


**Figure 12.** Evolution of envelopes of the vertical profiles of potential temperature (first row), specific humidity (second row) and cloud fraction (third row) for the three SANDU transition sub-cases. Same conventions as in Fig. 11.

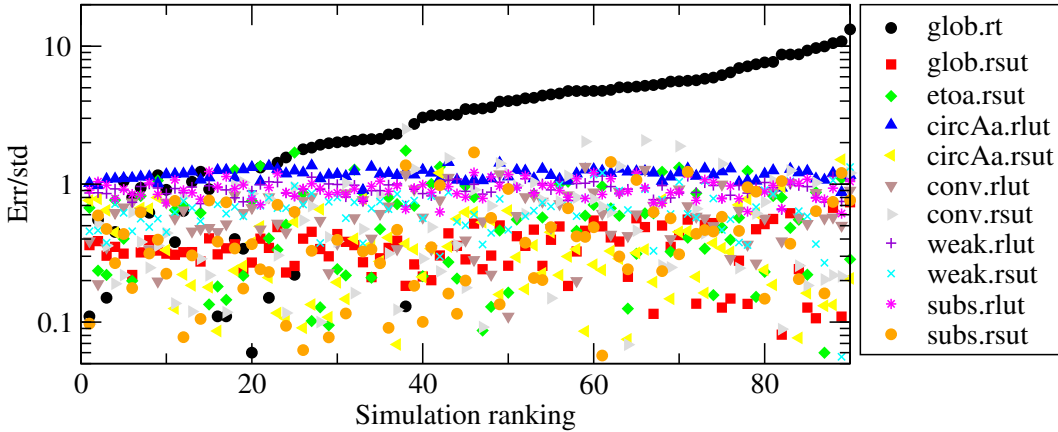


**Figure 13.** Impausibility matrix for the 9-parameter history match, at wave 32, build by adding 2 iterations with SCM and GCM metrics after 30 waves of SCM history matching.





**Figure 14.** Zonally average latitudinal variation (left) and latitudinally averaged (between 20S and 5S) zonal variation (right) of the SW cloud Radiative Effect (CRE) at TOA for 90 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 year 1 to 10 of a simulation run with the nominal values of the 9 parameters. The gold curves correspond to the 5 BEST simulations (see text for details). The EBAF observations are superimposed in black.



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

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 this ranking. By doing so, we go further than theoretically authorized by the history matching philosophy, i.e. not going beyond the constraints imposed by the predefined tolerance in order to avoid overfitting 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 Fig. 11, Fig. 12 and Fig. 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 Fig. 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, 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), we decided to abandon this target for the 6A version.

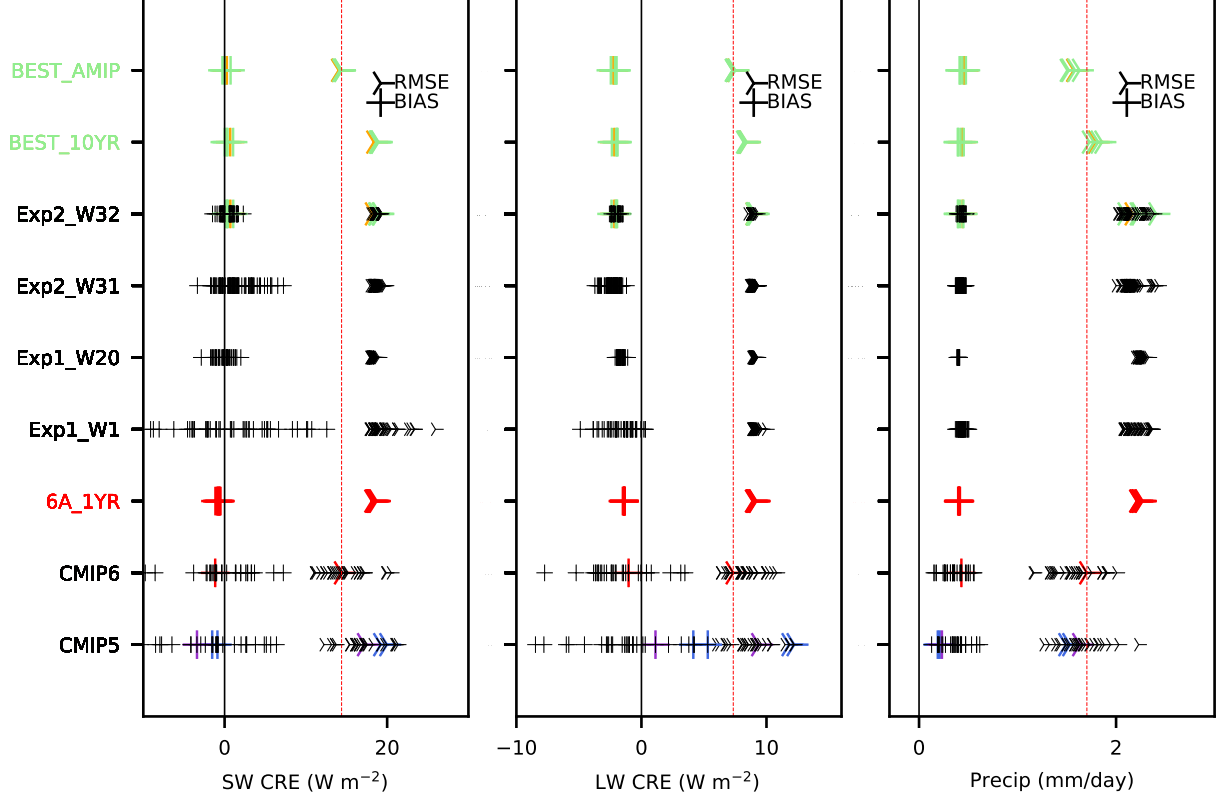
For the 6A version, we show as well 10 consecutive years run on climatological SSTs in order to illustrate the error and dispersion that come from this different setup (the CMIP diagnostics correspond to the mean seasonal cycle over the period 1979-2005). The mean bias is not significantly affected by the different setup, and its interannual 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 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.

Fig. 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. Note that this result is obtained without further tuning of the parameters involved in the representation of high-level clouds.

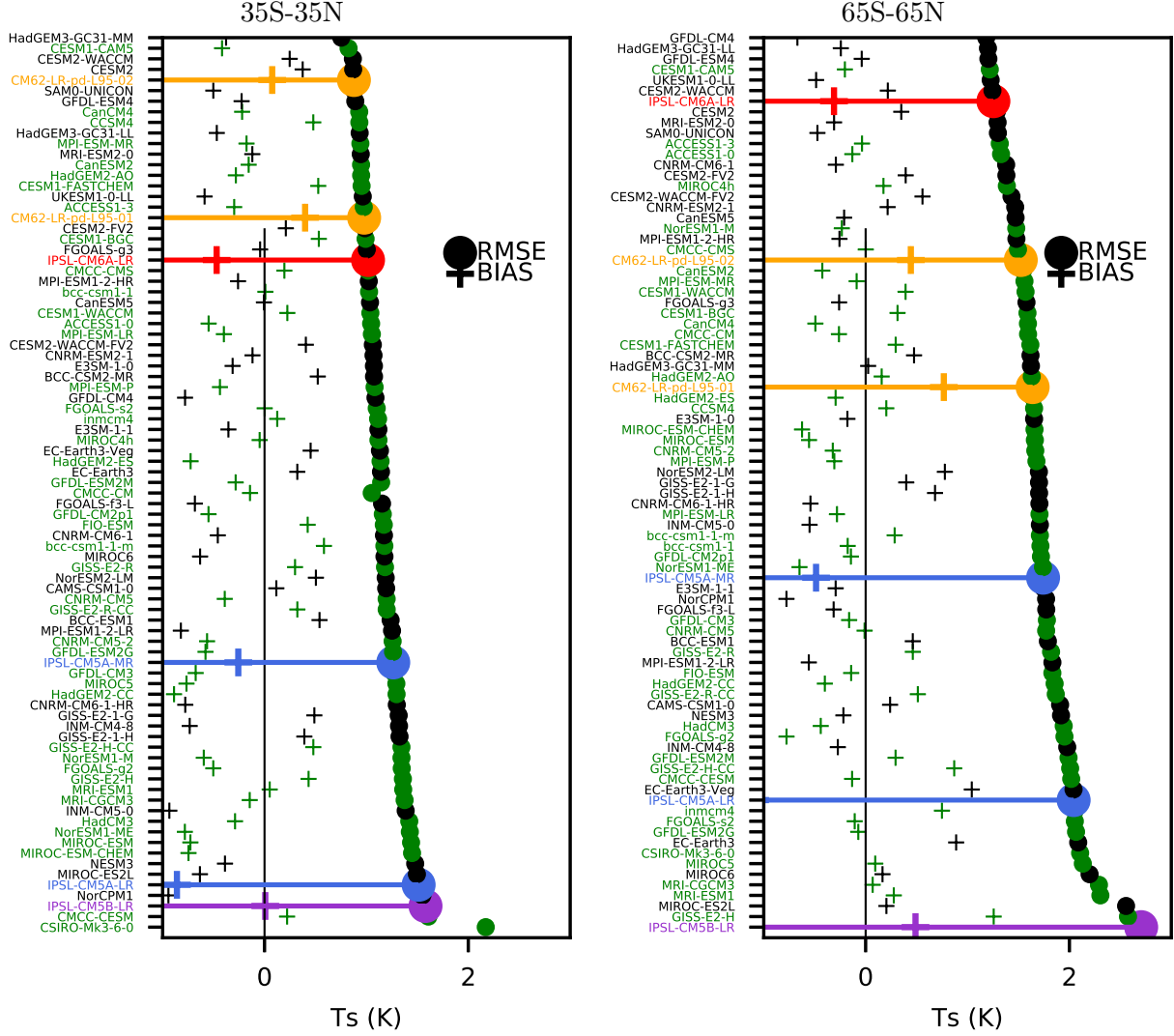
### 5.3 Test in coupled atmosphere-ocean configuration

Finally, the “BEST1” simulation 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.

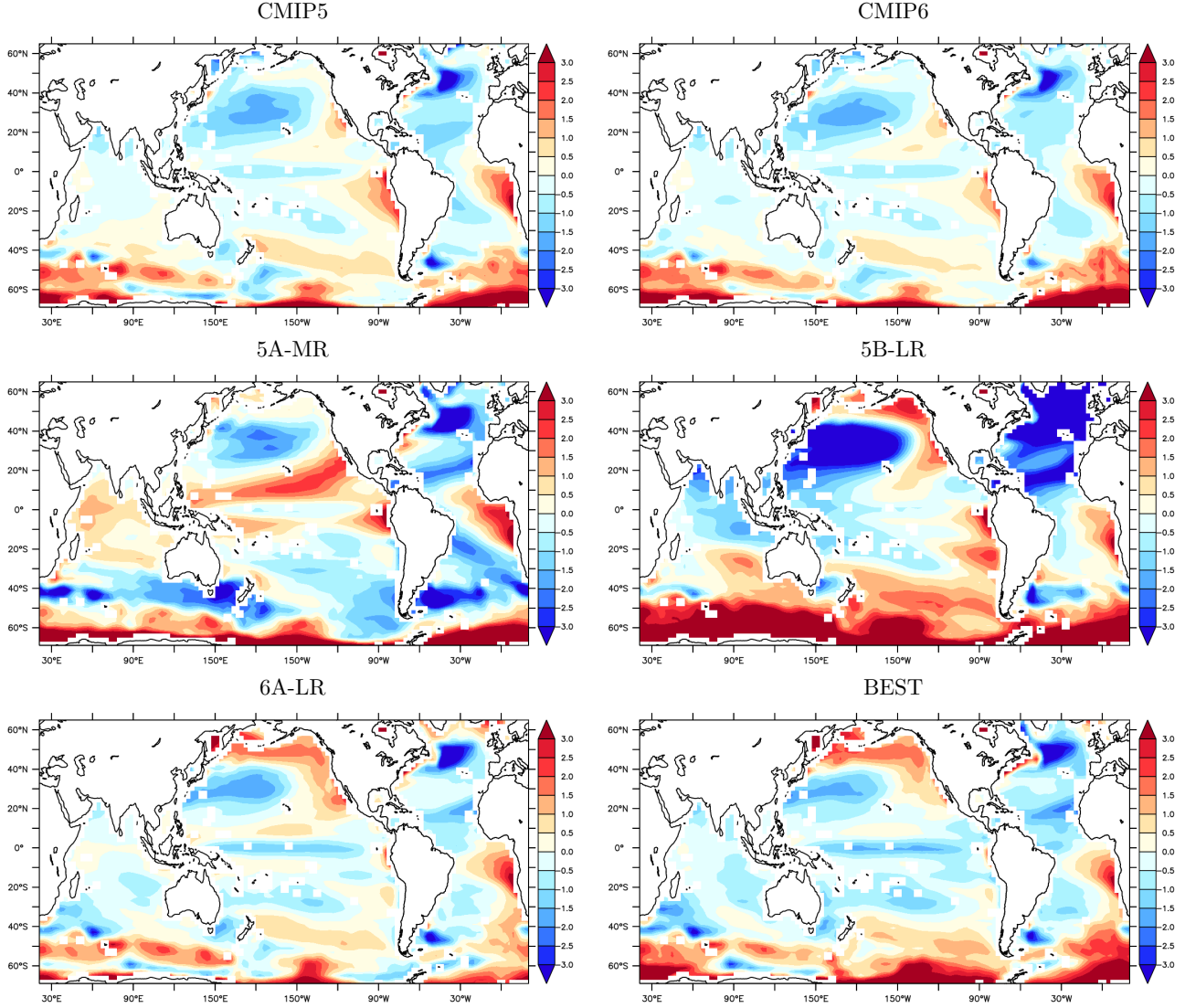
The seasonal cycle of SSTs is almost stabilized at the fifth decade. Fig. 17 shows the mean bias and root-mean-square error of SST computed on a mean seasonal cycle of the BEST simulation (gold), compared to the other CMIP5 (green) and CMIP6 (black) simulations with IPSL simulations highlighted with different colors.



**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 twelve monthly values on each grid cell after interpolation on a common  $2^\circ \times 2^\circ$  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), 10 individual years with the standard LMDZ6A configuration run on climatological SSTs, the results of the wave 1 and 20 of the first set of experiments and wave 31 and 32 of the second set (second year of a 2-year long simulation run on climatological SSTs), the 5 best simulations of wave 32 run over 10 years with climatological SSTs, and, at the top, the same 5 simulations run over the 1979-2005 period with annually varying SSTs (amip protocol as for CMIP simulations). Some simulations are highlighted with a colour code: for CMIP5 simulations, the blue and violet colours correspond respectively to the 5A and 5B versions of LMDZ (the 5A version was run with two different resolutions). The red colour is used for the 6A version of the LMDZ model, the green to the 5 best simulations and the orange to the best one. The vertical lines correspond to a zero bias (black) and RMSE of the CMIP6 IPSL-6A-LR configuration (red dashed).



**Figure 17.** SST mean bias and root-mean-square error computed from the mean seasonal cycle (12 monthly means) after interpolation on a  $120 \times 90$  regular longitude-latitude grid. The diagnostics are shown for tropical latitudes (left, 35S:35N) and for the global ocean (latitudes 65S:65N). 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.



**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 BEST simulations, without (left) and with (right) final retuning. The global mean of the bias is removed to highlight the structure of the bias.

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 are in the NROY space). For both simulations, the results are quite close to the 6A simulation. The results are better in the tropics (35S:35N) than for the full globe (65S:65N, removing latitude beyond 65 degrees 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.

## 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 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 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. Additionally, the bias in the zonal mean may be partly related to the shifted position of the mid-latitude jet which is particularly sensitive to the horizontal grid resolution, as seen on the left hand side of Fig. 6 and Fig. 14, in particular in the southern mid-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 workaround existing compensating errors and to fully benefit from our strategy.

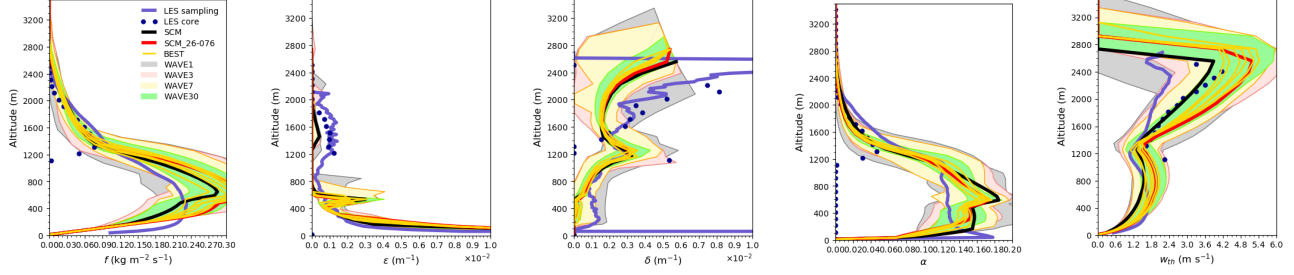
Altogether, our results confirm that the proposed strategy is able to provide reasonable tuning of a coupled model, by applying a rather systematic procedure making use of machine learning techniques and starting from LES/SCM comparisons and with only 9 parameters. This study shows how an improvement in the parameterization can be implemented in the full 3D GCM with an automatic tuning procedure, avoiding a long phase of by-hand retuning. The improvement tested here consists in the increase of the vertical resolution together with allowing us to vary some additional free parameters. As just shown, it is possible to better reproduce the 1D “transition cases” with the modified scheme.

### 6.2 Enlightening the representation of cloud processes

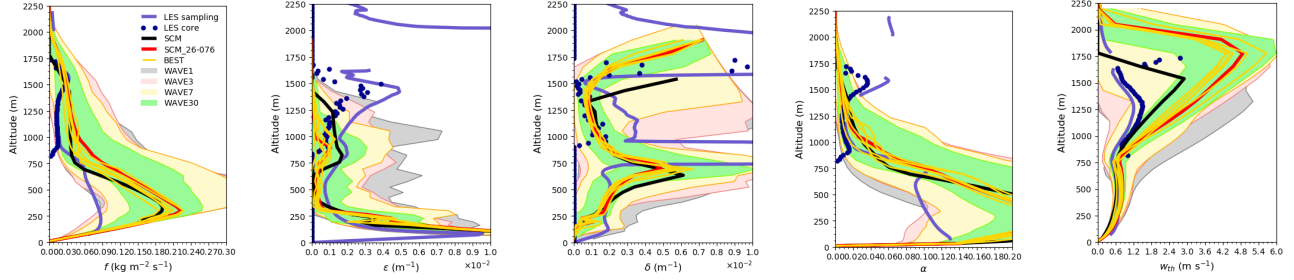
In order to interpret further the modification induced by this new tuning at the process scale, we show in Fig. 19 the internal variables of the thermal plume model obtained with the ARM cumulus case at 2 to 3 PM local time and in afternoon and evening of the third day of the SANDU/REF case. The vertical velocity is globally overestimated in the cloud for the control simulation, when compared to the plume



AMRCU/REF



SANDU/REF



**Figure 19.** Vertical profiles of the internal variables of the mass flux scheme for the ARM cumulus simulation averaged between 2 and 3 PM local time and for the SANDU/REF case, averaged before noon and midnight during the third day of simulation. As in Fig. 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.

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 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 favour 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 1 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, i. e. 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, e.g. parameterizations of microphysics. The automatic tools presented here now permit us to address such questions in more detail.

### 6.3 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 Krasnopolsky et al. (2013); Brenowitz and Bretherton (2018); Gentine et al. (2018), 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.



## 7 Conclusions

This paper presents a first proof of concept of the use of history matching to go from a process-based parameterization improvement to a new model configuration. More specifically, it presents a successful exercise of tuning of a global climate model with an automatic procedure after some improvement was introduced concerning the representation of boundary layer processes and associated clouds.

It should be noted that the availability of this tool is a necessary condition for the success of the exercise, but that it does not in any way detract from the importance of the modelers expertise. It must be underlined indeed that this result was 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.

Another 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. However, it is possible that the previous tuning was obtained thanks to compensating errors with high clouds which are not directly affected by the parameters selected in the present study. It is possible as well that the fine tuning of the parameters of the thermal plumes does not matter that much. So at least we obtained automatically a tuning as good as the previous one, after modification that improved the agreement at a process level. Possibly as well, the tuning could be even better if we had enabled retuning with other parameters. Note that the value retained for the **DZ** parameter is a bit larger when the 9-parameter tuning is used, probably suggesting a compensation with more penetrative plumes obtained when reducing the value of **B1**.

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 second 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 200x100 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 more costly. 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 1000 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.

In any case, 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 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.

## Acknowledgments

This work received funding from grant HIGH-TUNE ANR-16-CE01-0010. It was supported by the DEPHY2 project, funded by the French national program LEFE/INSU. The 3D simulations were granted access to the HPC resources of IDRIS under the allocation gencmip6 attributed by GENCI (Grand Equipement National de Calcul Intensif) and the ressources of TGCC from a Prace allocation to the “QUEST” project. The data that supports this research and the visualizations are available at <https://doi.org/10.14768/20190626001.1>. Daniel Williamson was funded by NERC 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., ... White, R. G. (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. Retrieved from <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssc.12198> doi: 10.1111/rssc.12198
- Ayotte, K. W., Sullivan, P. P., Andr n, A., Doney, S. C., Holtslag, A. A., Large, W. G., ... Wyngaard, J. C. (1996). An evaluation of neutral and convective planetary boundary-layer parameterizations relative to large eddy simulations. *Boundary-layer Meteorol.*, 79, 131-175.
- Brenowitz, N. D., & Bretherton, C. S. (2018, June). Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, 45(12), 6289-6298. (WOS:000438499100052) doi: 10.1029/2018GL078510
- Bretherton, C., & Smolarkiewicz, P. (1989). Gravity waves, compensating subsidence and detrainment around cumulus clouds. *J. Atmos. Sci.*, 46, 740-759.
- Brown, A., Cederwall, R., Chlond, A., Duynkerke, P., Golaz, J.-C., Khairoutdinov, M., ... Stevens, B. (2002). Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Q. J. R. Meteorol. Soc.*, 128, 1075-1093.
- Couvreur, F., Guichard, F., Redelsperger, J. L., Kiemle, C., Masson, V., Lafore, J. P., & Flamant, C. (2005). Water-vapour variability within a convective boundary-layer assessed by large-eddy simulations and IHOP\_2002 observations. *Q. J. R. Meteorol. Soc.*, 131, 2665-2693.
- Couvreur, F., Hourdin, F., & Rio, C. (2010, March). Resolved Versus Parametrized Boundary-Layer Plumes. Part I: A Parametrization-Oriented Conditional Sampling in Large-Eddy Simulations. *Boundary-layer Meteorol.*, 134, 441-458. doi:

- 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.
- Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018, May). Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.*, *45*, 5742–5751. doi: 10.1029/2018GL078202
- Grandpeix, J., & Lafore, J. (2010, April). A Density Current Parameterization Coupled with Emanuel’s Convection Scheme. Part I: The Models. *Journal of Atmospheric Sciences*, *67*, 881–897. doi: 10.1175/2009JAS3044.1
- Gregory, D. (2001). Estimation of entrainment rate in simple models of convective clouds. *Q. J. R. Meteorol. Soc.*, *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. *J. Atmos. Sci.*, *59*, 1105–1123.
- Hourdin, F., Grandpeix, J.-Y., Rio, C., Bony, S., Jam, A., Cheruy, F., ... Roehrig, R. (2013, May). LMDZ5B: the atmospheric component of the IPSL climate model with revisited parameterizations for clouds and convection. *Clim. Dyn.*, *40*, 2193–2222. doi: 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. *Geophys. Res. Lett.*, *42*, 10. doi: 10.1002/2015GL066764
- Hourdin, F., Jam, A., Rio, C., Couvreux, F., Sandu, I., Lefebvre, M.-P., ... Idelkadi, A. (2019). *Unified parameterization of convective boundary layer transport and clouds with the thermal plume model*. Accepted in JAMES.
- Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... Williamson, D. (2017, March). The Art and Science of Climate Model Tuning. *Bull. Am. Meteorol. Soc.*, *98*, 589–602. doi: 10.1175/BAMS-D-15-00135.1
- Hourdin, F., Rio, C., Grandpeix, J.-Y., Madeleine, J.-B., Cheruy, F., Rochetin, N., ... Ghattas, J. (2020). *LMDZ6A: the atmospheric component of the IPSL climate model with improved and better tuned physics*. James, accepted for publication.
- Hourdin, F., Rio, C., Jam, A., Traore, A. K., & Musat, I. (2020). *Convective boundary layer control of the sea surface temperature in the tropics*. James, accepted for publication.
- Jam, A., Hourdin, F., Rio, C., & Couvreux, F. (2013, June). Resolved Versus Parametrized Boundary-Layer Plumes. Part III: Derivation of a Statistical Scheme for Cumulus Clouds. *Boundary-layer Meteorol.*, *147*, 421–441. doi: 10.1007/s10546-012-9789-3
- Köhler, M., Ahlgrimm, M., & Beljaars, A. (2011, January). Unified treatment of dry convective and stratocumulus-topped boundary layers in the ECMWF model. *Q. J. R. Meteorol. Soc.*, *137*, 43–57. doi: 10.1002/qj.713
- Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013, March). 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 Artificial Neural Systems*, *203*(3), 13. doi: 10.1155/2013/485913
- Loeb, N. G., Wielicki, B. A., Doelling, D. R., Smith, G. L., Keyes, D. F., Kato, S., ... Wong, T. (2009, FEB). Toward Optimal Closure of the Earth’s Top-of-Atmosphere Radiation Budget. *J. Climate*, *22*(3), 748–766. doi: {10.1175/2008JCLI2637.1}
- Rio, C., & Hourdin, F. (2008). A thermal plume model for the convective boundary layer : Representation of cumulus clouds. *J. Atmos. Sci.*, *65*, 407–425.
- Rio, C., Hourdin, F., Couvreux, F., & Jam, A. (2010, June). Resolved Versus

- Parametrized Boundary-Layer Plumes. Part II: Continuous Formulations of Mixing Rates for Mass-Flux Schemes. *Boundary-layer Meteorol.*, *135*, 469–483. doi: 10.1007/s10546-010-9478-z
- Salter, J. M., & Williamson, D. (2016, December). A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. *Environmetrics*, *27*(8), 507–523. (WOS:000392948100005) doi: 10.1002/env.2405
- Sandu, I., & Stevens, B. (2011, September). On the Factors Modulating the Stratocumulus to Cumulus Transitions. *J. Atmos. Sci.*, *68*, 1865–1881. doi: 10.1175/2011JAS3614.1
- Siebert, P., & Frank, A. (2003). Source-receptor matrix calculation with a lagrangian particle dispersion model in backward mode. *Atmos. Chem. Phys. Discuss.*, *3*, 4515–4548.
- Simpson, J., & Wiggert, V. (1969). Models of precipitating cumulus towers. *Mon. Wea. Rev.*, *97*(7), 471–489.
- Sundqvist, H. (1978, July). A parameterization scheme for non-convective condensation including prediction of cloud water content. *Q. J. R. Meteorol. Soc.*, *104*, 677–690. doi: 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*. Kluwer Academic Publishers, Dordrecht, the Netherlands.
- Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012, April). An Overview of CMIP5 and the Experiment Design. *Bull. Am. Meteorol. Soc.*, *93*, 485–498. doi: 10.1175/BAMS-D-11-00094.1
- Vignon, E., Hourdin, F., Genthon, C., Gallée, H., Bazile, E., Lefebvre, M.-P., ... Van de Wiel, B. J. H. (2017, July). Antarctic boundary layer parametrization in a general circulation model: 1-D simulations facing summer observations at Dome C. *J. Geophys. Res.*, *122*, 6818–6843. doi: 10.1002/2017JD026802
- Williamson, D., Blaker, A. T., Hampton, C., & Salter, J. (2015, September). Identifying and removing structural biases in climate models with history matching. *Clim. Dyn.*, *45*, 1299–1324. doi: 10.1007/s00382-014-2378-z
- Williamson, D., Blaker, A. T., & Sinha, B. (2017, April). Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development*, *10*(4), 1789–1816. doi: 10.5194/gmd-10-1789-2017
- Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013, October). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim. Dyn.*, *41*, 1703–1729. doi: 10.1007/s00382-013-1896-4
- Yamada, T. (1983). Simulations of nocturnal drainage flows by a  $q^2l$  turbulence closure model. *J. Atmos. Sci.*, *40*, 91–106.

Figure 1.

## Sketch of clouds formation and water vertical transport



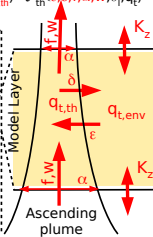
## Thermal plume model

### Computing plume properties

$$(\epsilon, \delta, f, \alpha, W) = \mathcal{G}_{th}(\theta_l, q_t, \mathbf{A1}, \mathbf{A2}, \mathbf{B1}, \mathbf{CQ}, \mathbf{DZ})$$

### Transporting water and temperature

$$(\delta_t \theta_l, \delta_t q_t, q_{t,th}) = \mathcal{F}_{th}(\epsilon, \delta, f, \alpha, W, \theta_l, q_t)$$



## Large scale condensation scheme

### Computing subgrid water distribution

$$\sigma_{s,env} = \mathcal{G}_{bg}(W, q_{sat}, q_{t,th}, q_{t,env}, \mathbf{BG1})$$

$$\sigma_{s,th} = \mathcal{G}_{bg}(W, q_{sat}, q_{t,th}, q_{t,env}, \mathbf{BG2})$$

### Converting total water to cloud

$$(\alpha_{cld}, q_l) = \mathcal{F}_{bg}(q_{sat}, q_{t,th}, q_{t,env}, \sigma_{th}, \sigma_{env})$$

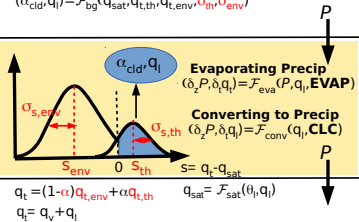


Figure 2.



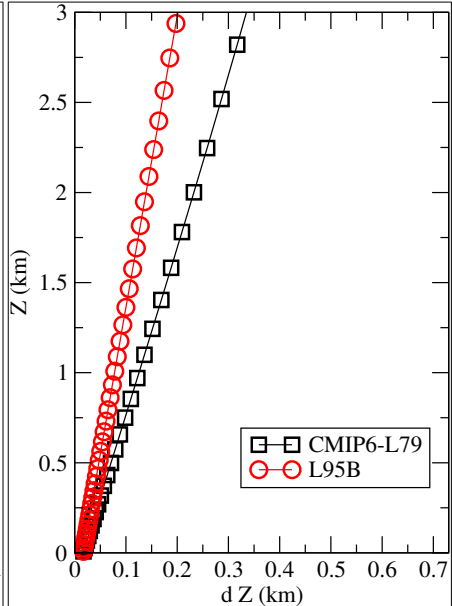
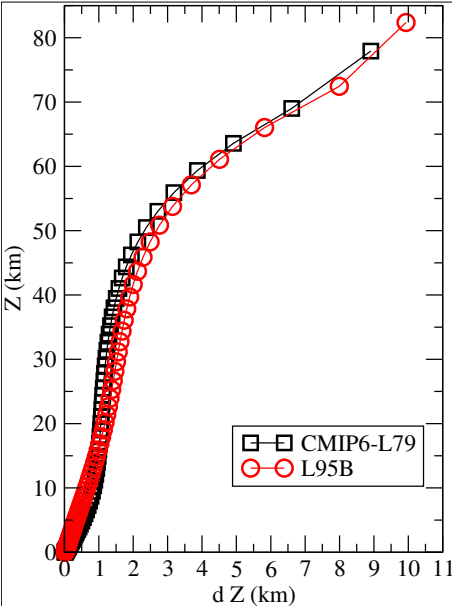


Figure 3.


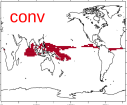
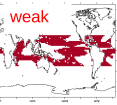
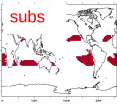
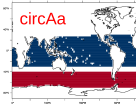
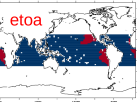
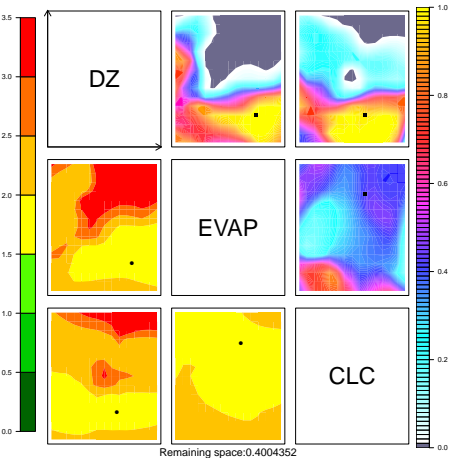
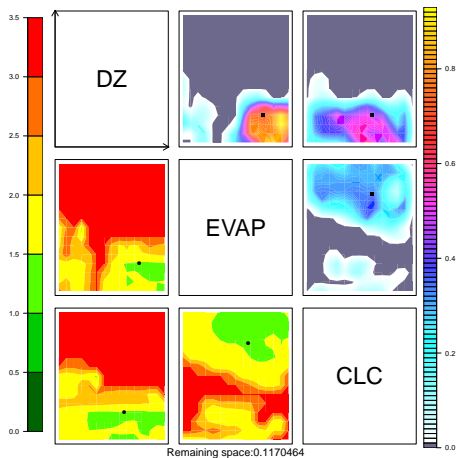
Mask	Variable	Metrics	target W m <sup>-2</sup>	error W m <sup>-2</sup>
 Glob	Total rad. TOA (rt) Swup TOA (rsut)	glob.rt	2.5	0.2
		glob.rsut	99.6	5
		circAa.rsut	24.0	5
		circAa.rlut	-48.6	5
 conv	SWup TOA (rsut) LWup TOA (rlut)	subs.rsut	84.9	5
 weak		weak.rsut	81.8	5
 subs		conv.rsut	103.2	5
 circAa		subs.rlut	274.6	5
 etoa	SWup TOA (rsut)	weak.rlut	264.3	5
		conv.rlut	235.8	5
		etoa.rsut	11.0	5

Figure 4.

Wave 1



Wave 5



Wave 20

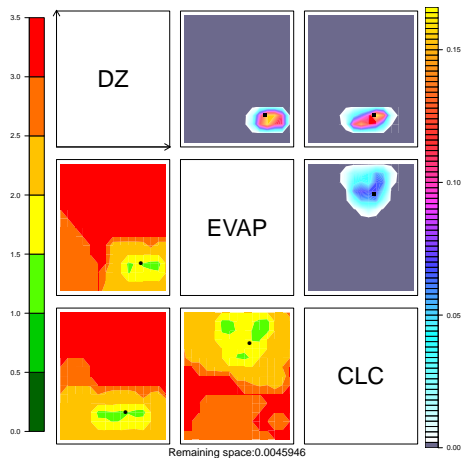


Figure 5.

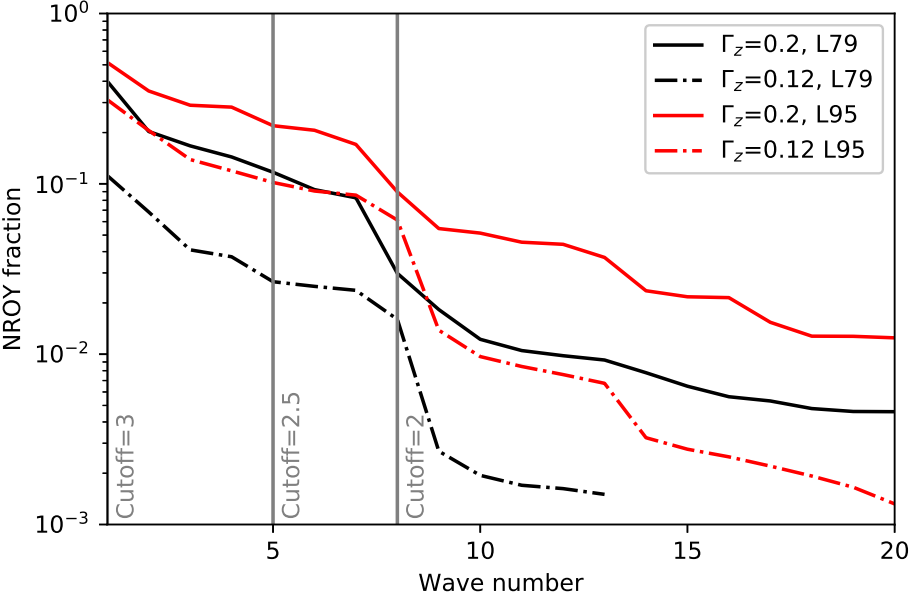
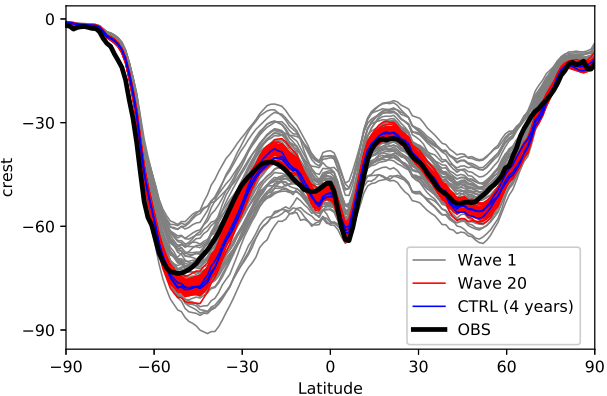


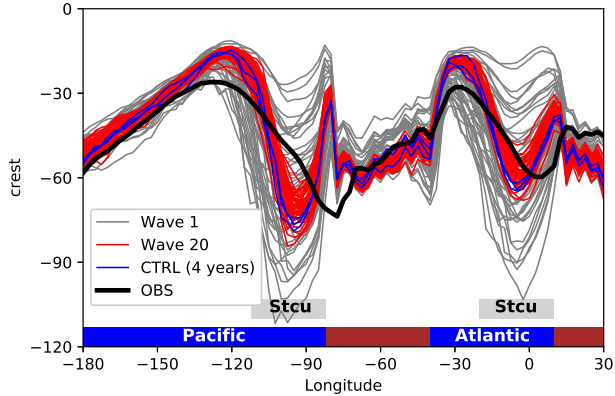


Figure 6.

Zonal mean

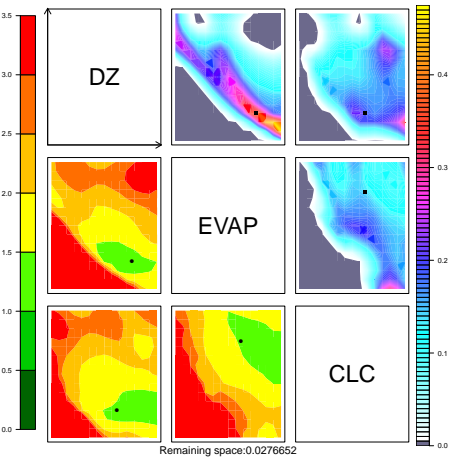


[20S:5S] longitudinal cross section

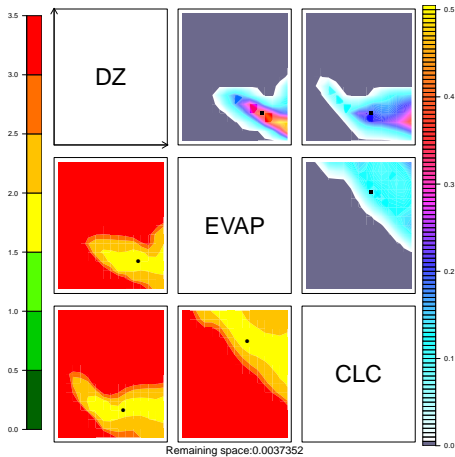


**Figure 7.**

3D alone, Wave 1



1D+3D, Wave 1



1D+3D, Wave 20

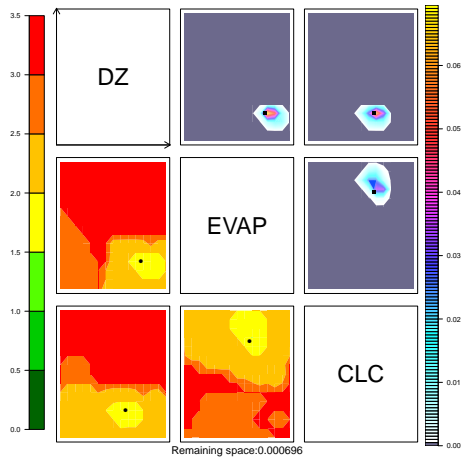


Figure 8.

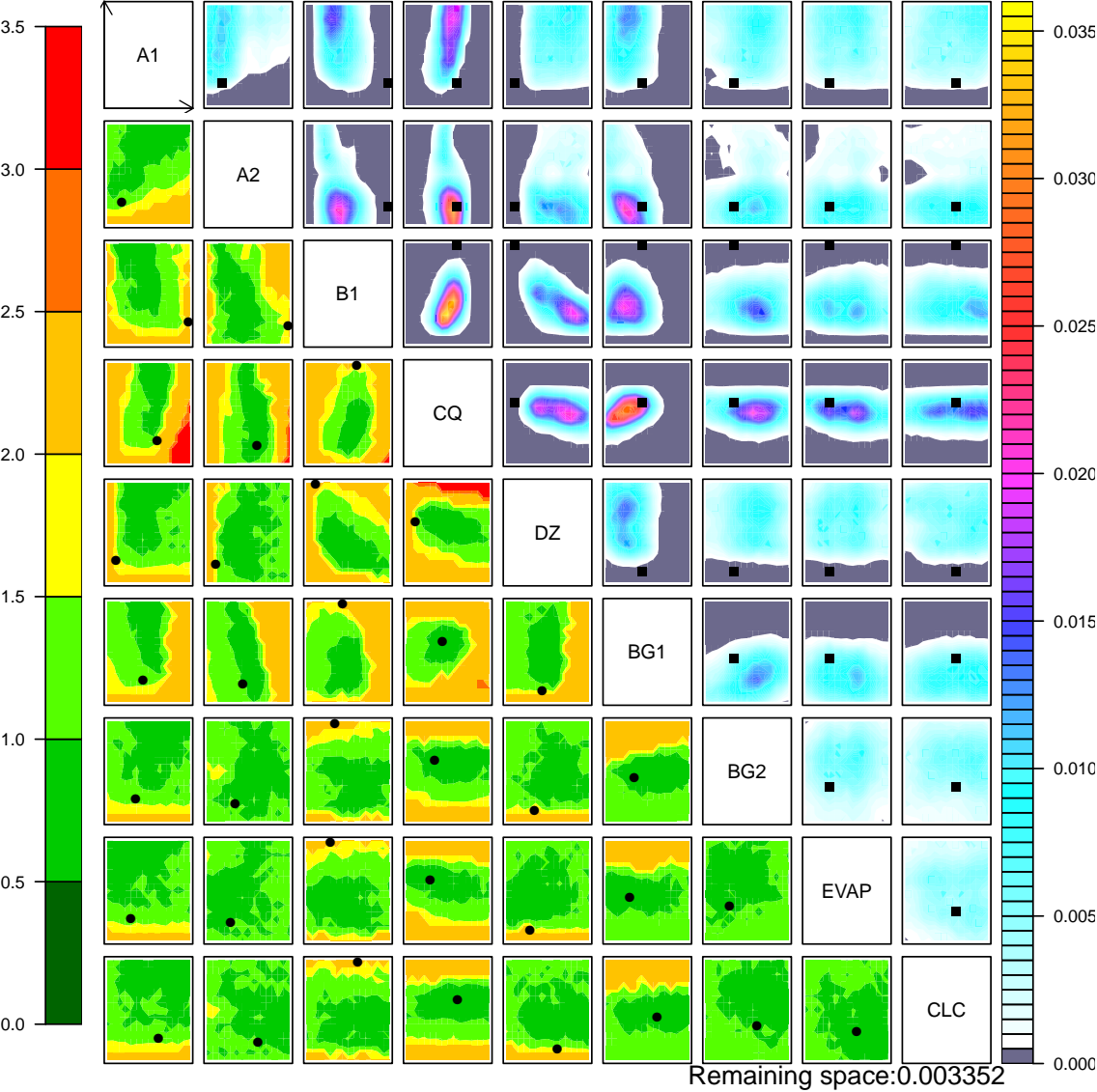


Figure 9.



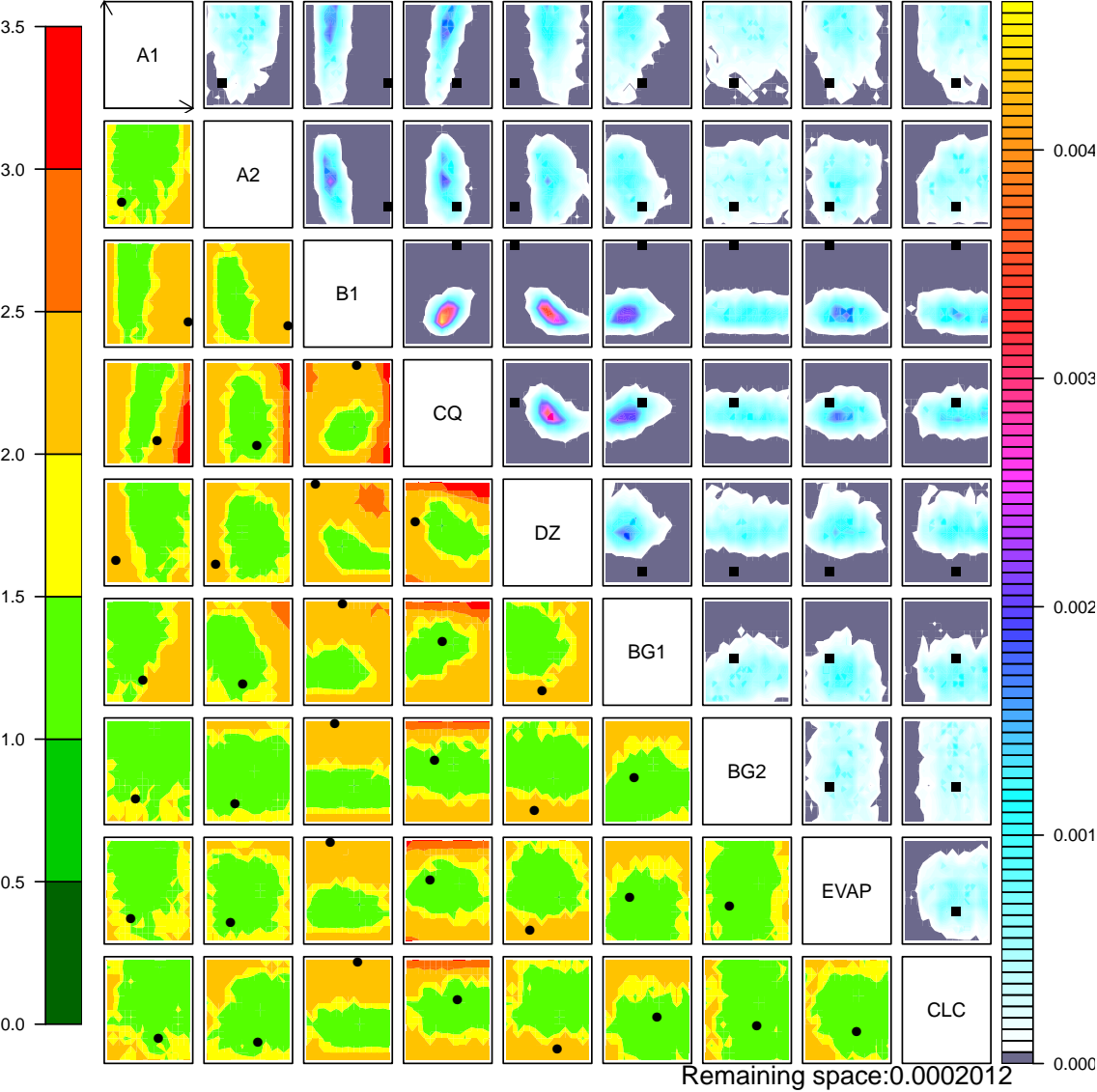


Figure 10.

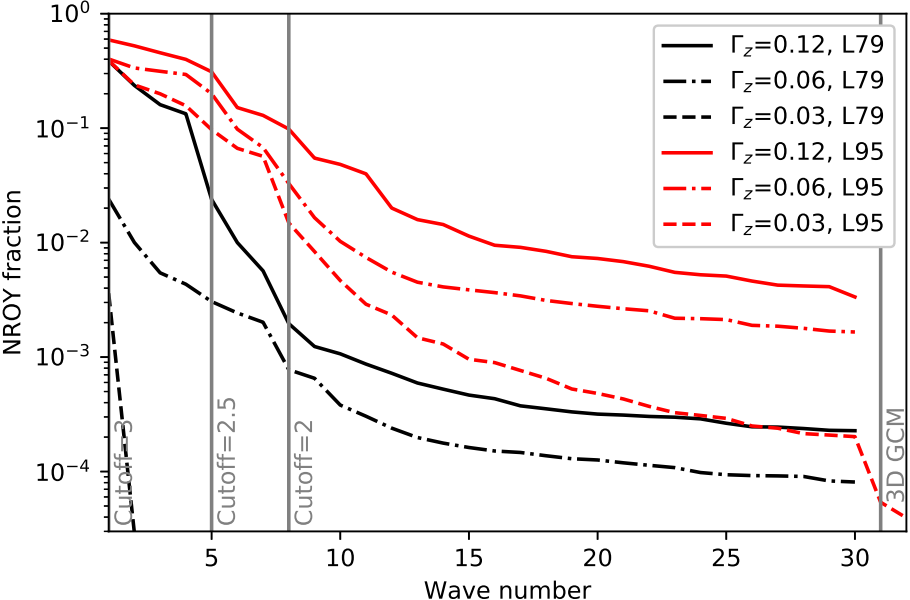
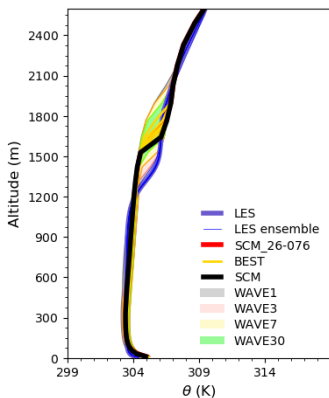
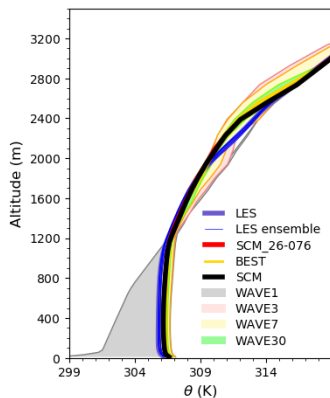


Figure 11.

IHOP/REF: 2002-06-14 13:30



ARMCU/REF: 1997-06-21 20:30



RICO/REF: 2004-12-27 21:30

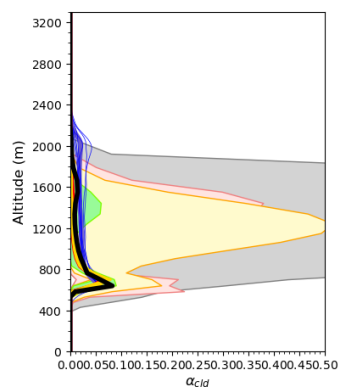
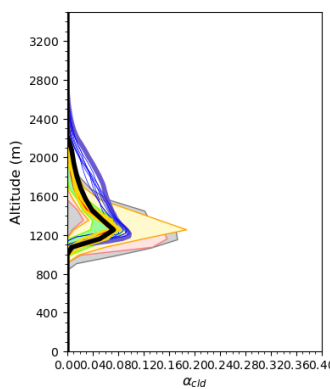
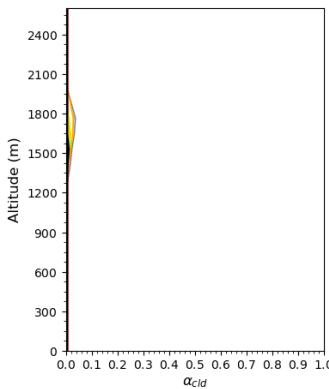
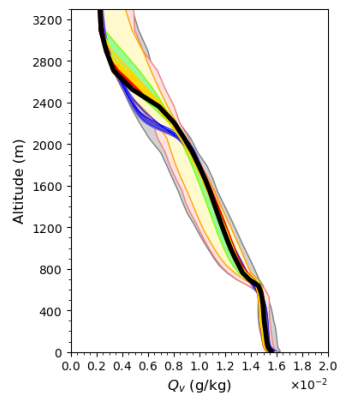
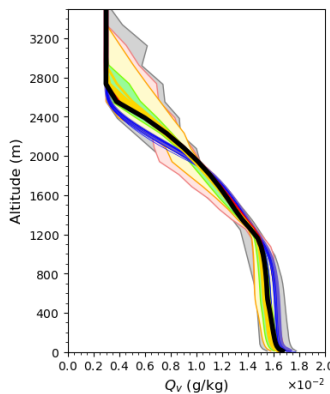
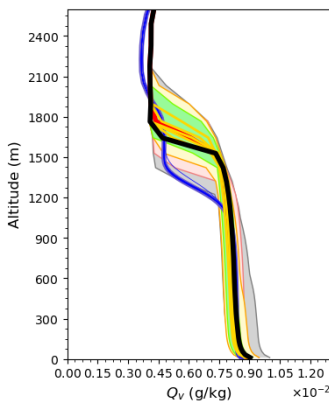
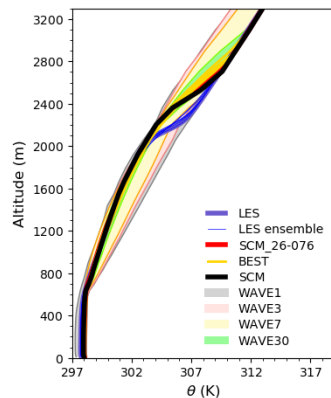
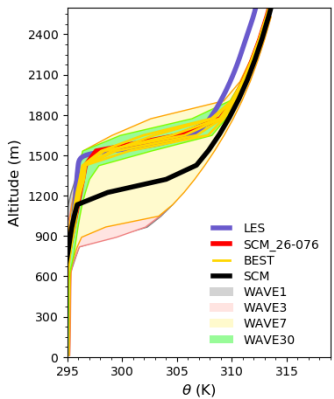
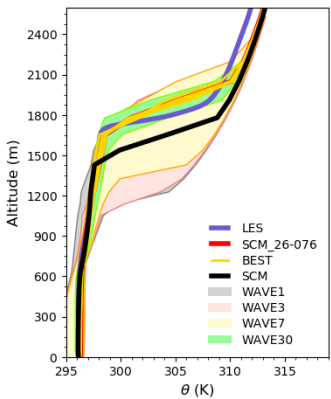


Figure 12.

SANDU/SLOW: 2006-07-18 00:00



SANDU/REF: 2006-07-18 00:00



SANDU/FAST: 2006-07-18 00:00

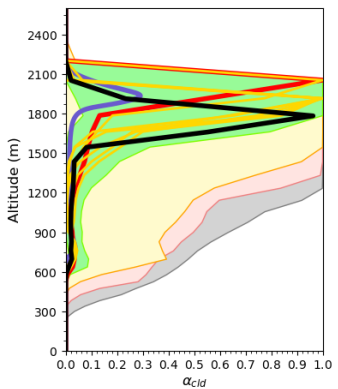
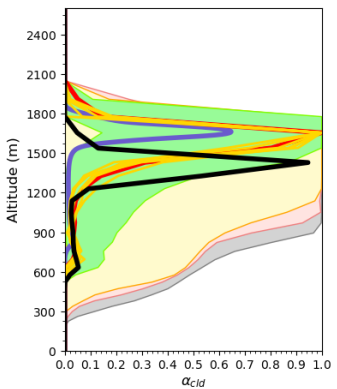
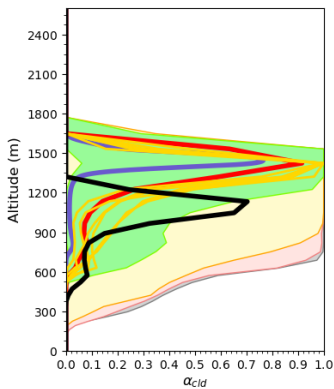
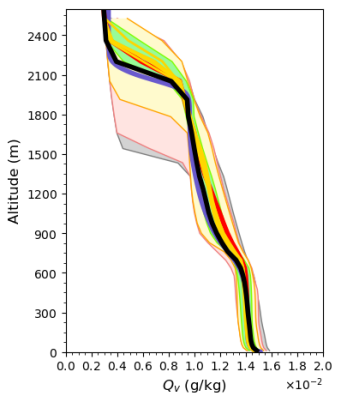
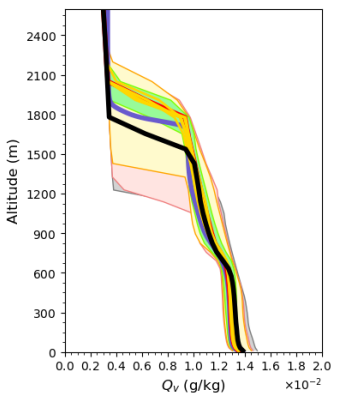
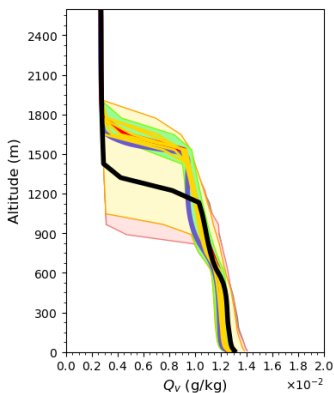
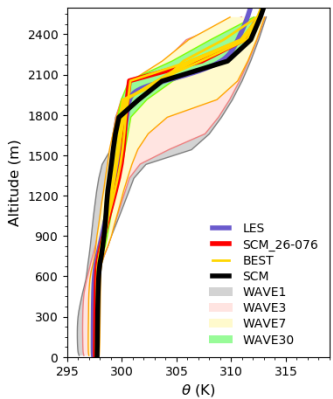


Figure 13.



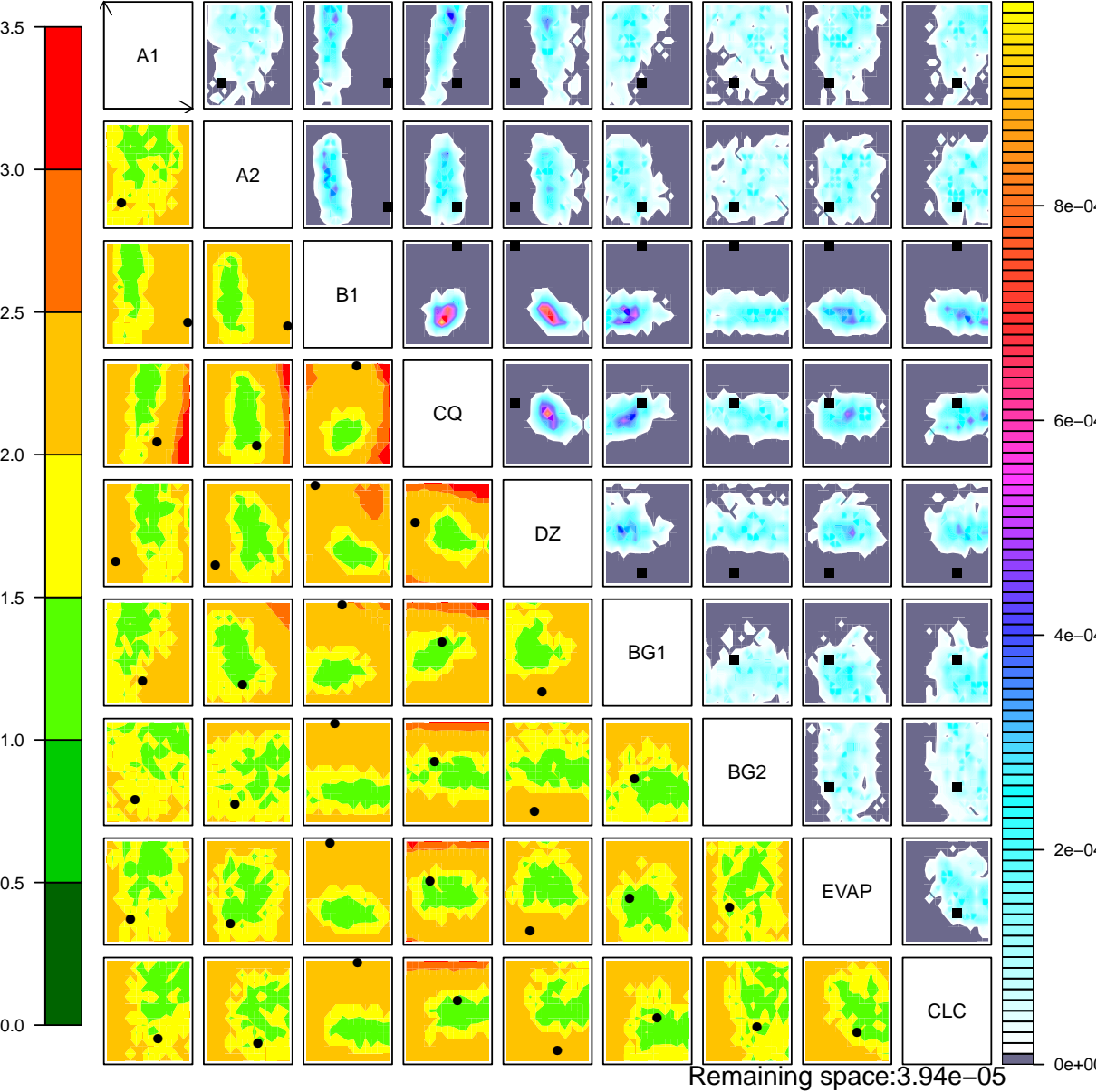
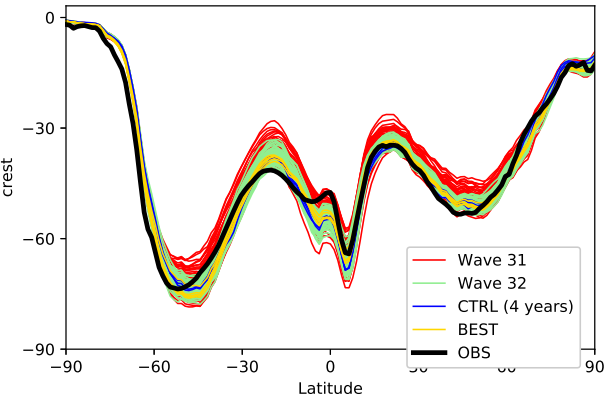


Figure 14.

Zonal mean



[20S:5S] longitudinal cross section

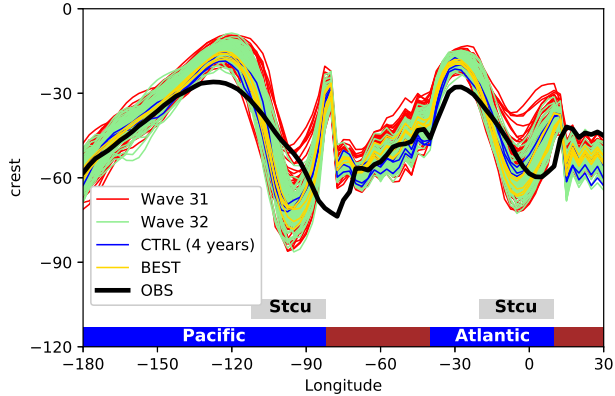


Figure 15.

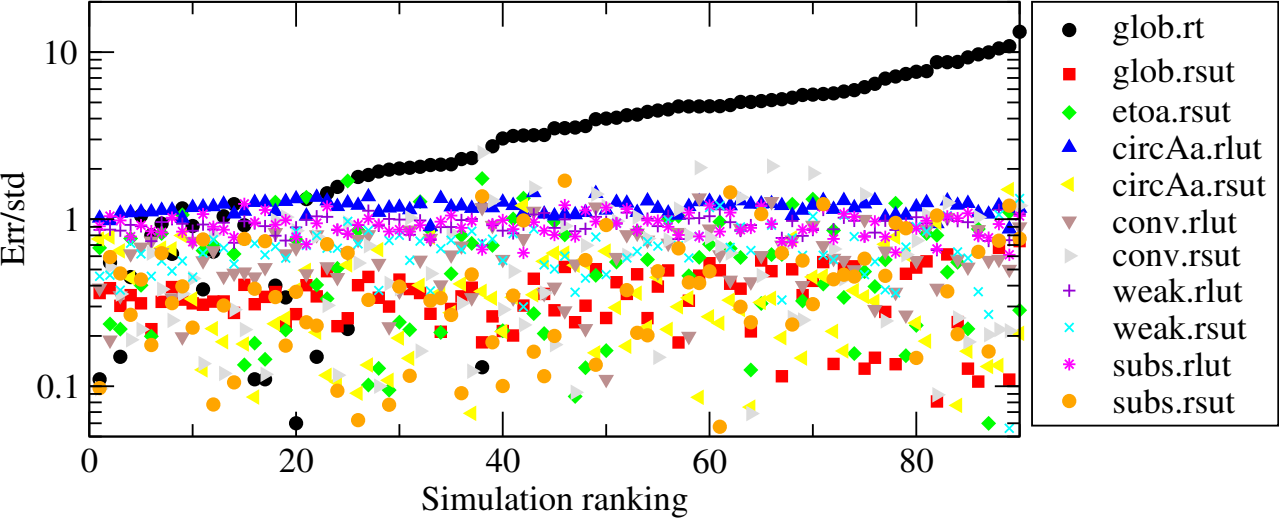


Figure 16.

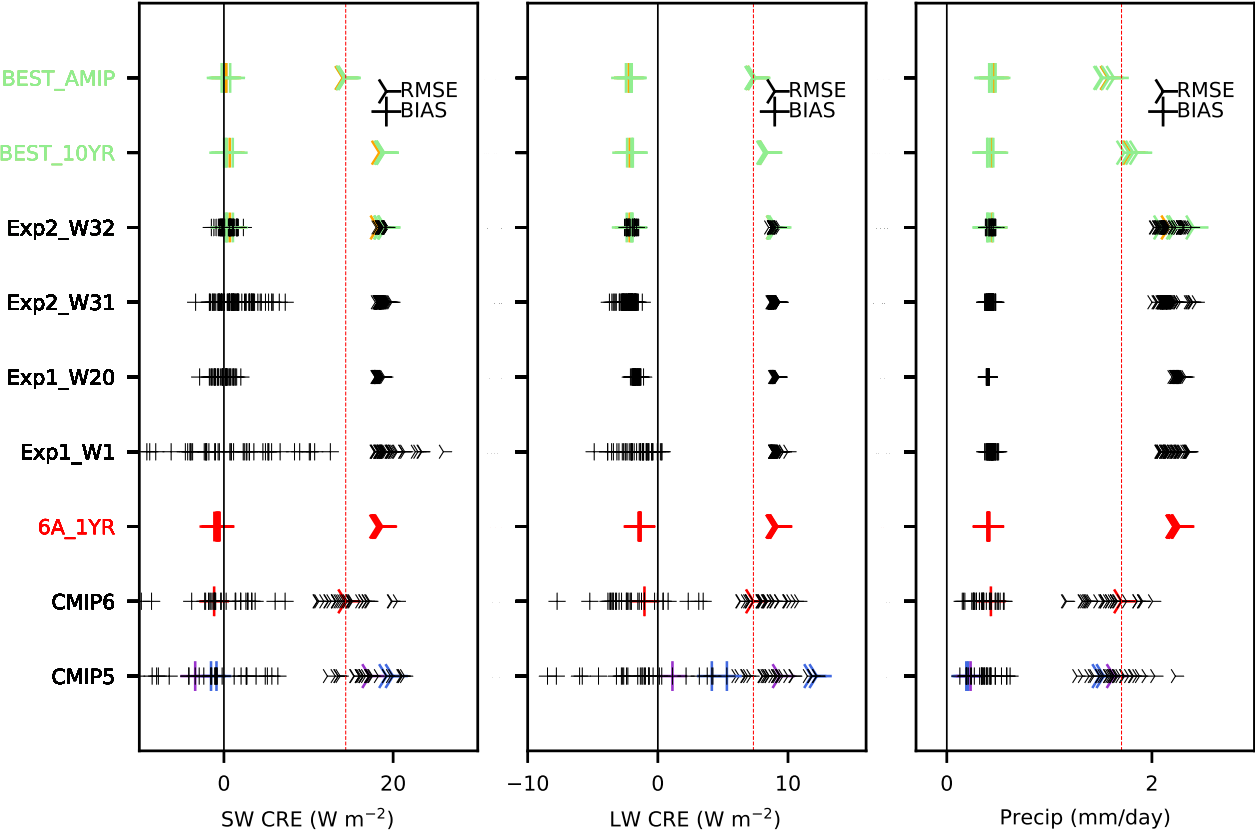
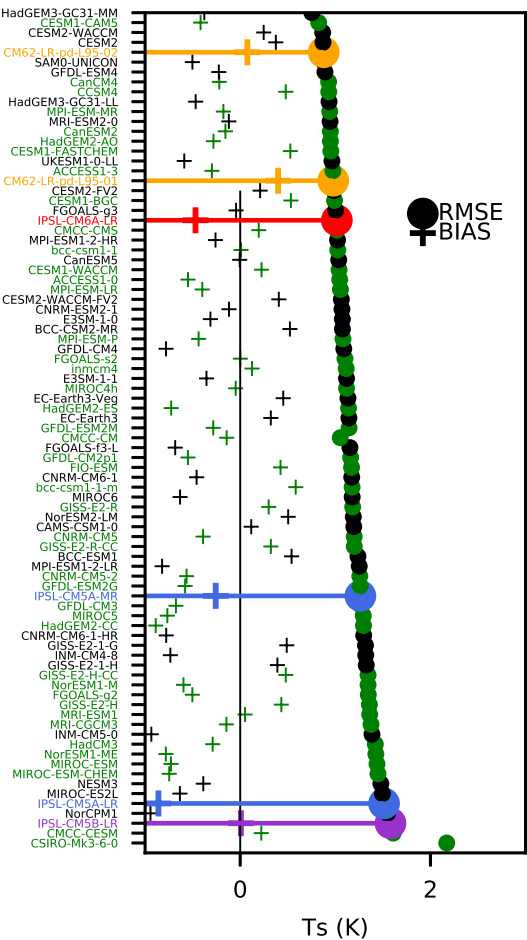


Figure 17.



35S-35N



65S-65N

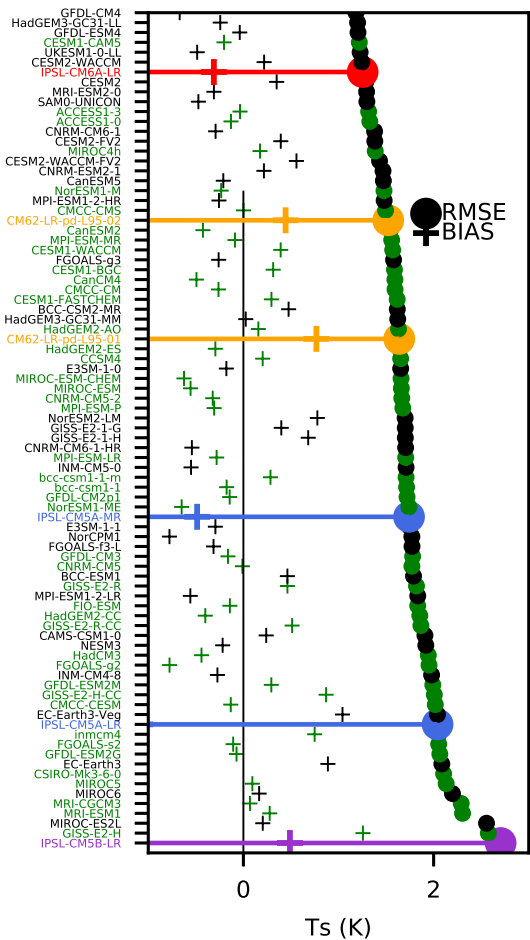
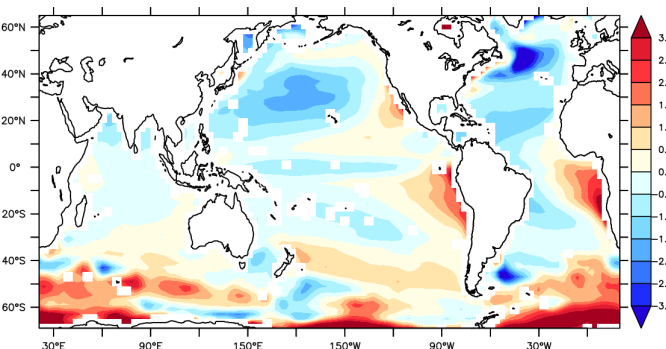
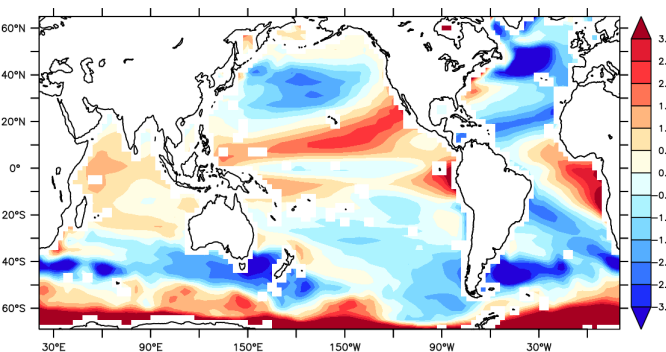


Figure 18.

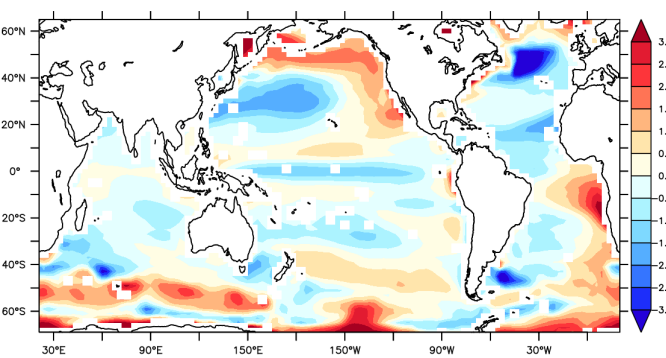
CMIP5



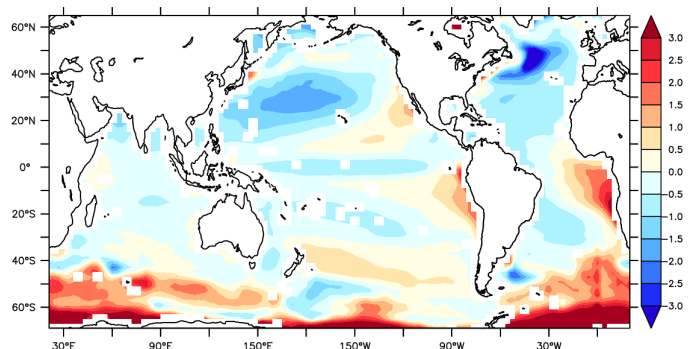
5A-MR



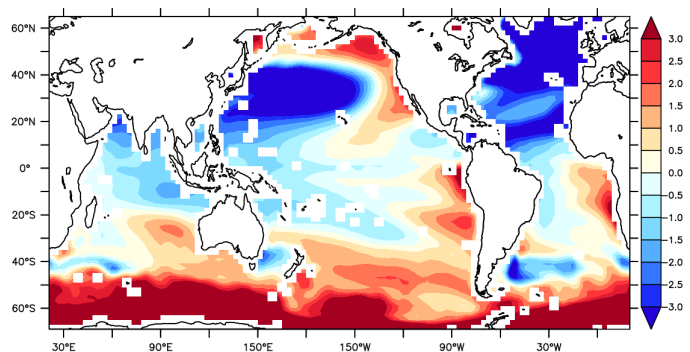
6A-LR



CMIP6



5B-LR



BEST

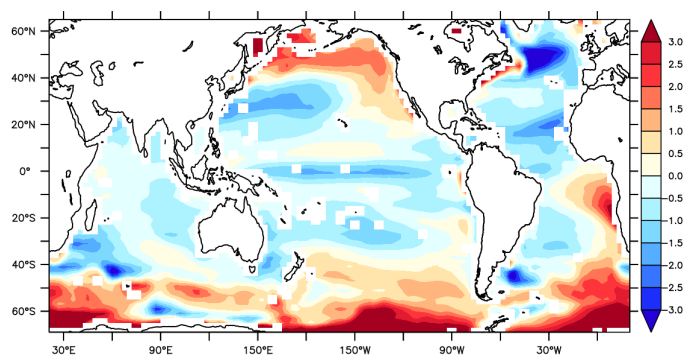
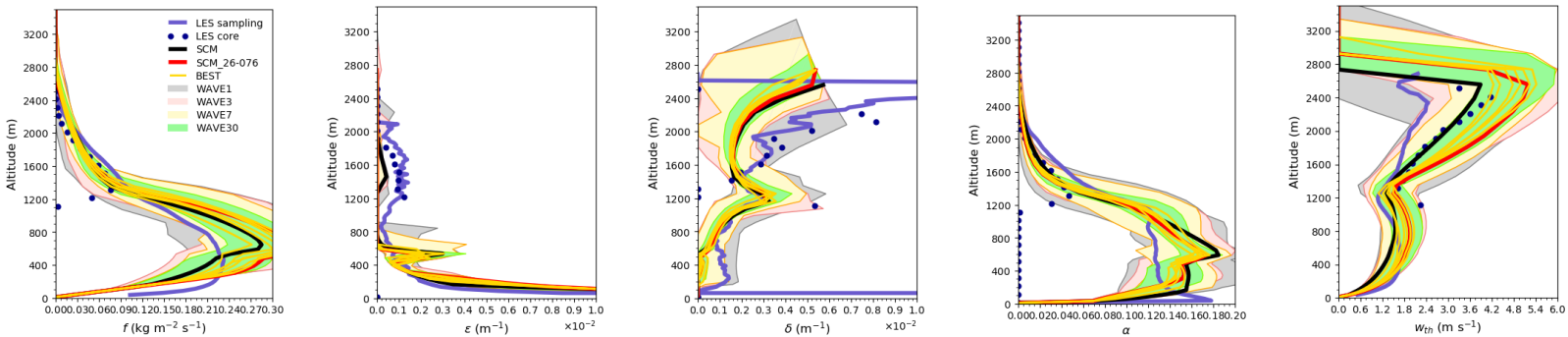


Figure 19.

AMRCU/REF



SANDU/REF

