Reprise papier Maëlle

M. Coulon-Decorzens[1], F. Hourdin[1], N. Villefranque[2]

[1]Laboratoire de Meteorologie Dynamique, Sorbonne Université/IPSL/CNRS, Paris, France
[2]Centre National de Recherches Météorologiques, Météo-France, CNRS, Toulouse, France

Key Points:

-
-
- Structural errors in radiative transfer models are compensated by simulating wrong cumulus-cloud properties

Corresponding author: M. Coulon–Decorzens, `maelle.coulon-decorzens@lmd.ipsl.fr`

Abstract

Compensating errors are an obstacle to the development of climate models. We wonder if systematic errors in simulated cloud properties might result from error compensation when targetting top-of-atmosphere radiative fluxes in the tuning process while using a inaccurate radiative transfer parameterization. Here, we investigate structural errors in radiative transfer models and how they might be compensated by errors in cloud properties in an idealized tuning experiment. Convection and cloud parameters of two versions of a Single-Column version of a climate Model (SCM), with and without a parameterization of cloud 3D radiative effects, are tuned targetting reference radiative fluxes obtained from Large-Eddy Simulations. When 3D effects are neglected, accurate fluxes are obtained only at the expense of overestimated cloud fractions, compensating underestimated cloud reflectivity at low sun. Aiming at fluxes averaged over solar angles removes this mechanism.

Plain Language Summary

Blabliblou

1 Introduction

Contexte général de l'étude : tuning des GCMs et compensations d'erreurs

- GCM, tuning, compensating errors
- elles sont là, comment vivre avec
- focus rayonnement car central : importance majeure pour le tuning, biais systématiques genre too few too bright, questions du moment avec ecrad, est ce qu'il faut mettre les effets 3D...
- sachant que de toute façon à la fin il faudra que le modèle soit équilibré donc il sera tuné pour que les flux radiatifs globaux soient bons ; on voudrait que ce soit pour de bonnes raisons. Notamment, le bon rayonnement pour les bons nuages
- "avoir les bons nuages" c'est avoir la bonne répartition de l'eau dans le système, c'est à dire avoir les bonnes climatologies de profils verticaux de fractions et de contenu en eau. Ces propriétés sont le résultat d'interactions non linéaires entre les différents processus modélisés dans le GCM. En plus d'avoir le bon rayonnement pour les bons nuages, on voudrait aussi avoir les bons nuages pour de bonnes raisons, c'est à dire pour une représentation correcte des processus et de leurs interactions.
- Une façon standard de tenter d'assurer que la modélisation des processus physiques est correcte est de travailler dans un cadre 1D, qui permet de travailler sur la physique seule, sans qu'elle n'interagisse avec la dynamique.
- Des travaux précédents ont exploité ce cadre et permis des avancées... puis gain de puissance récemment avec les outils d'exploration paramétrique qui vont jusqu'à la calibration automatique... on arrive à assurer que les paramétrisations physiques de couche limite sont correctes, et on apprend à connaitre leurs erreurs quand ce n'est pas le cas

Qu'est ce qu'on fait là ?

- avec tout ça en tête...
- l'étude présentée ici est partie d'une question pratique : dans le cadre du développement du modèle de climat LMDZ, on se pose la question de quelle configuration du schéma radiatif il faut prendre : quels paramètres il faut choisir pour la géométrie nuageuse ? est ce qu'il faut tenir compte des effets 3D en mode climat ?
- ces questions pratiques nous ont amené à nous interroger sur le modèle et sur les potentiels effets compensatoires, internes au rayonnement ou entre nuages et rayonnement
- pour ce qui est des paramètres à choisir dans la paramétrisation de rayonnement, on avait une proposition pour les cumulus cf papier tuning 3. Comme on a voulu faire des tests avec et sans effets 3D, on s'est demandé dans quelle mesure les paramètres réglés pour un solver tenant compte des effets 3D (spartacus) étaient valides pour un autre solver (tripleclouds). Autrement dit, est ce que les flux simulés par spartacus sont bons pour de "bonnes raisons", c'est à dire pour de bons processus radiatifs "purement 1D", liés à l'overlap et à l'hétérogénéité horizontale de l'eau nuageuse ? Si on désactive les effets 3D, est ce qu'on retombe bien sur ce que devraient être des "flux 1D" ?
- pour ce qui est de prendre en compte ou non les effets 3D dans les modèles de climat, ce qu'on sait déjà c'est que ça coute plus cher et que ça a l'air d'avoir peu d'effet en moyenne sur les tests qui ont été effectués en mode global ; mais quid des compensations d'erreurs sur les nuages via la calibration ? si le calcul de rayonnement est biaisé en présence de nuages (car manque d'effets 3d) et qu'on vise des flux globaux observés pour équilibrer le modèle, est ce que ça peut mener à choisir des configurations du modèle dans lesquelles les propriétés nuageuses sont biaisées ?
- on a finalement abouti à une étude en deux parties où (1) on examine les compensations internes entre les différents paramètres de géométrie nuageuse au sein de la param de rayonnement ; et (2) on examine les compensations entre nuages et rayonnement. Ce sont les deux sections de résultats, sections 3 et 4. Dans la section 2, on décrit l'approche générale et les différents outils / modèles que l'on va utiliser. Dans la section 5, on discute.

General circulation models (GCM) used for climate projections, are, like any model, imperfect representations of the climate system. Their behaviour depends on free parameters that need to be adjusted, which is achieved through calibration. One of the challenges of calibrating complex models such as GCMs is to avoid overfitting through unwarranted compensating errors. Hourdin et al. (2017) report that a common practice in calibrating (tuning) climate models is to target observed top-of-atmosphere (TOA) radiative fluxes by adjusting parameters associated with the most uncertain processes controlling these fluxes: those related to clouds. In so doing, accurate TOA fluxes are often obtained at the expense of cloud-related compensating errors: between cloud properties and e.g. albedo or jet position (Hourdin et al., 2013), between low-, middle- and high-level clouds (Webb, Senior, Bony, & Morcrette, 2001; Nam, Bony, Dufresne, & Chepfer, 2012) or even between physical, optical and radiative properties of a given cloud regime (Konsta et al., 2022).

1 INTRO

1.1 Contexte général de l'étude : tuning des GCMs et compensations d'erreurs

General circulation models (GCM) used for climate projections, are, like any model, imperfect representations of the climate system. Their behaviour depends on free parameters that need to be adjusted, which is achieved through calibration.

When calibrating numerical models as complex as GCMs, it is very difficult, if not impossible, to avoid overfitting through unjustified error compensation. The issue of reducing these compensation errors and finding ways to better characterize and control them remains a major challenge in climate modeling, one that we hope to address more effectively thanks to increased computing power and machine learning algorithms. This issue is essential to the reliability of climate change projections.

"non voulues" ?

Hourdin et al. (2017) report that a common practice in calibrating (tuning) climate models is to target observed top-of-atmosphere (TOA) radiative fluxes by adjusting parameters associated with the most uncertain processes controlling these fluxes: those related to clouds. In so doing, accurate TOA fluxes are often obtained at the expense of cloud-related compensating errors: between cloud properties and e.g. albedo or jet position (Hourdin et al., 2013), between low-, middle- and high-level clouds (Webb, Senior, Bony, & Morcrette, 2001; Nam, Bony, Dufresne, & Chepfer, 2012) or even between physical, optical and radiative properties of a given cloud regime (Konsta et al., 2022).

To understand these errors, and in particular compensations between cloud physics and radiative transfer (RT), it is necessary to disentangle model errors stemming from each parameterization.

climatology ?
(je n'aime pas trop représentation ici, trop flou / imprécis)

Cloud parameterizations typically provide vertical profiles of cloud fraction and water condensate in each atmospheric column. Having the "right clouds" in a GCM means having a good representation of those (these) profiles, which results from complex non linear interactions between the various processes accounted for in the GCM. Additional (cloud) modelling assumptions must be made to compute RT from these profiles, such as vertical overlap, horizontal in-cloud heterogeneity or 3D radiative effects. They are usually made inside the RT scheme and their first-order effects on TOA fluxes are quite well known (see e.g., McKee and Cox (1974); Barker, Stephens, and Fu (1999); Várnai and Davies (1999); Shonk, Hogan, Edwards, and Mace (2010); Hogan, Fielding, Barker, Villefranque, and Schäfer (2019)): The widely used maximum-random overlap assumption tends to underestimate total (vertically integrated) cloud cover, and consequently, TOA fluxes. Neglecting in-cloud optical-depth heterogeneity systematically leads to overly reflective clouds. Neglecting 3D effects resulting from horizontal transport of photons produces either too-bright or too-dim clouds depending on solar zenith angle, surface and cloud properties. The resent development of the ECrad code at ECMWF opens the opportunity to investigate issues related to the cloud approximations since it allows to activated different cloud solver s in the same code, including for the first time a parameterization of 3D cloud radiative effects.

facilitates the investigation of cloud-geometry effects in radiative transfer models

A now standard way to improve the representation of cloud physics in GCMs is to work in the single column framework, which allows to work on physics parameterizations without interaction with large scale dynamics. In this approach, ~~Large Eddy Simulations (LES)~~, 3D simulations of the same cloud scene with resolution of a few tens of meters on domains of a few tens of km, (Large Eddy Simulations, LES) are used as a reference for model evaluations of simulations with (output from) the Single Column Model (SCM). This SCM/LES approach has been recently empowered by machine learning ~~approaches~~ and automatic tuning procedures by Couvreux et al. (2021), using the history matching approach proposed ~~by~~ by Williamson et al. (2013). This approach is based on global sensitivity experiments that enable the separation of parametric and structural errors in the model, thus providing new perspectives to the long-lasting issue of compensating errors. This framework has led to significant advances in the parameterizations of boundary layer convection and associated cumulus and stratocumulus clouds (see e.g. Hourdin et al. (2019)). It is at the heart of the hierarchical tuning process promoted by Couvreux et al. (2021); Hourdin et al. (2021); Villefranque et al. (2021), upon which the present work is built.

In Couvreux et al. (2021) and Hourdin et al. (2021), parameters of an SCM's boundary-layer parameterizations are tuned targetting LES cloud properties. In Villefranque et al. (2021), cloud-geometry parameters of an RT scheme are tuned by running offline radiation upon mean LES vertical profiles, targetting reference solar fluxes obtained from 3D Monte Carlo simulations.

This work combines these approaches by tuning the SCM parameters targetting radiative fluxes (ou autre phrase du genre ?)

on garde ce découpage ?

1.2 Qu'est ce qu'on fait là ?

The study presented here started from the practical need to tune atmospheric radiation and clouds after the introduction of ECrad in the LMDZ GCM Hourdin et al. (2020), the atmospheric component of the IPSL coupled model IPSL-CM, used in particular for CMIP exercises (Boucher et al., 2020). Which configuration should be chosen, ? ~~including or not~~ 3D radiative effects, ? with which valuer of ECrad free parameters ?

Should it include

These questions led us to consider more in depth the question of error compensation, both inside the radiative transfer code, between for instance 3D effects and cloud overlap, and between clouds and radiation.

Villefranque et al. (2021) compared off line computations of RT with ECrad, ~~compared~~ with reference 3D ray-tracing Monte-Carlo simulations run on the same LES cloud scenes, to propose a tuning of the three ECrad free parameters relative to cloud geometry: the overlap decorrelation length, heterogeneity and cloud size. ~~The~~ ECrad was run with the Spartacus solver, the only one able to account for 3D radiative effects. Ranges of parameters were identify that give a very good agreement with MC computations, including a good representation of the dependency of reflected radiation to the solar zenith angle, which requires a representation of 3D effects. ~~The~~ Spartacus is however costly numerically and the effect on climate simulations not obvious, questioning the need to use it in climate simulations. Spartacus is built ~~by adding~~ on the TripleClouds solver, mostly by adding a parameterizations of horizontal transfer of radiation between clear sky and clouds or inside clouds. We first compare here three different tuning of ECrad: 1) ~~a~~ tuning ~~with the~~ Spartacus ~~solver~~ targeting 3D MC computations as in Villefranque et al. (2021); ~~a~~ tuning ~~with the~~ TripleClouds ~~solver~~ targeting 1D MC computations (consisting in computing 3D radiation independently in every column of the LES assumed independant, as in the MCica solver); and 3) ~~a~~ tuning ~~targeting~~ both, 3D radiation with Spartacus and 1D radiation with TripleClouds ~~when~~ using the same values for the parameters that are common ~~to both~~: the overlap decorrelation length and heterogeneity parameter. Comparisons of this three tuning exercises is used to enlighten the possible error compensations that can be at work within the radiative code itself, arrizing from the representation of ~~the~~ cloud geometry. The idea of tuning 3) is to avoid error compensations between 3D effects and other part of the radiative code, tuning Spartacus with parameters that give a correct simulation of 1D radiation when activating TripleClouds only.

(en vrai il y a aussi l'entrapment qui est un peu différent)

cloudy regions of the same layer

they have in

We then investigate ~~the~~ error compensations between radiative transfer computation (including the representation of ~~the~~ cloud geometry) and clouds physics (that provide the vertical profiles of temperature, cloud fraction and water content), by running SCM tuning experiments that target radiative metrics. In these simulations, we fix set the ECrad free parameters to the best values previously identified in the off line ECrad tuning, i. e. assuming ~~a~~ perfect ~~representation of~~ clouds ~~physics (by LES)~~. When targeting radiative fluxes at various solar zenith angle, TripleClouds produces errors to compensate for not accounting for 3D effects. However, by using the best parameters issued from the combined Spar-

on est surs que c'est ça le résultat premier ?

tacus/TripleClouds ~~Clouds~~ off-line tuning, and averaging metrics obtained at various zenith angles, we show that the mean diurnal fluxes can be reasonably well simulated by TripleClouds, with a correct representation of clouds physics.

We ~~this~~ finally present ~~arrived at~~ a two-part study in which (1) we examine the internal compensations between the different cloud geometry parameters within the radiation scheme parameter; and (2) ~~we examine~~ the compensations between clouds and radiation. These are the two sections ~~of~~ results, ~~sections~~ 3 and 4. In section 2, we describe the general approach and the different tools/models that we ~~will~~ use, before discussing the results and concluding in Section 5.

References

Barker, H. W., Stephens, G. L., & Fu, Q. (1999). The sensitivity of domain-averaged solar fluxes to assumptions about cloud geometry. *Quarterly Journal of the Royal Meteorological Society*, *125*(558), 2127-2152. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712555810 doi: 10.1002/qj.49712555810

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., . . . Vuichard, N. (2020, July). Presentation and Evaluation of the IPSL-CM6A-LR Climate Model. , *12*(7), e02010. doi: 10.1029/2019MS002010

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., . . . Xu, W. (2021). Process-Based Climate Model Development Harnessing Machine Learning: I. A Calibration Tool for Parameterization Improvement. *Journal of Advances in Modeling Earth Systems*, *13*(3), e2020MS002217. doi: 10.1029/2020MS002217

Hogan, R. J., Fielding, M. D., Barker, H. W., Villefranque, N., & Schäfer, S. A. K. (2019). Entrapment: An Important Mechanism to Explain the Shortwave 3D Radiative Effect of Clouds. *Journal of the Atmospheric Sciences*, *76*(7), 2123-2141. doi: 10.1175/JAS-D-18-0366.1

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. *Climate Dynamics*, *40*(9-10), 2193–2222. doi: 10.1007/s00382-012-1343-y

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. *Journal of Advances in Modeling Earth Systems*, *11*(9), 2910–2933. doi: 10.1029/2019MS001666

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. *Bulletin of the American Meteorological Society*, *98*(3), 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. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS001892. doi: 10.1029/2019MS001892

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., . . . Volodina, V. (2021). Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2020MS002225. doi: 10.1029/2020MS002225

Konsta, D., Dufresne, J.-L., Chepfer, H., Vial, J., Koshiro, T., Kawai, H., . . . Ogura, T. (2022). Low-Level Marine Tropical Clouds in Six CMIP6 Models Are Too Few, Too Bright but Also Too Compact and Too Homogeneous. *Geophysical*

maellecd
8:07 PM

partie 1: super
partie 2 : J'ai eu du mal à comprendre cette partie. Bon c'est sur que je suis pas dedans et que mon niveau d'anglais et pas fou ; mais je pense que ça pourrait être plus clair. J'ai cloner pour faire des modifs mais j'ai pas réussi à faire des phrases.

Le premier paragraphe : pas de soucis
Après j'ai eu du mal à saisir la ligne directrice et j'ai trouvé que ça aller trop dans le détail ; impression qui est surement du au fait que ça manque de hierarchie
J'ai pas directement compris que le paragraphe qui commence par « Villefranque et al. (2021) compared off line » aller expliquer le « error compensa-
tion, [both] inside the radiative transfer code »
Et en lisant « we finally present » (début du dernier paragraphe) j'ai cru qu'il y allait avoir un truc en plus (mais ça c'est surement du détail de tournure). C'est en lisant cette phrase que j'ai compris la structure de ce que je venais de lire.
=> Je dirais qu'il faut ennoncer plus clairement les deux phases de l'étude + faire le lien avec l'objectif ennoncé au paragraphe 1 (on veut si-possible éviter de choisir des vecteurs de paramètres de ecRad qui conduisent à des compensations d'erreurs)
=> J'ai du mal à comprendre ce que le paragraphe qui va de « Villefranque et al. (2021) compared off line … We first compare here three different tuning of ECrad » fait la. C'est peut être juste qu'il manque un lien avant le « we first compare ». Je suis pas sur que l'aspect cout numérque et importance de spartacus doit être dit à cet endroit ; en tout cas je pense que ça m'a perdu que ça apparaisse la.
Je me demande s'il faut pas « retourner » ce troisième paragraphe, en commençant par l'objectif, puis en disant le contexte de villefranque 2021, puis en détaillant les trois expériences.

Peut être que ce serait bien d'expliciter plus clairement / de mettre plus de poids sur le fait qu'on a aussi fait ces expériences pour apprendre à étudier les compensations d'erreurs grâce à htexplo, et qu'on y arrive, et que du coup c'est un peu une « preuve de concepte » qu'on sait le faire, sur des cas simple où les maitrises les choses etc. Soit dans la partie 2 de l'intro, soit au niveau de « This approach is based on global sensitivity experiments that enable the
separation of parametric and structural errors in the model, thus providing new perspectives to the long-lasting issue of compensating errors » voir même a la fin du paragraphe pour rajouter de l'emphase. Je reste convaincue que c'est un résultat super important et que c'est cool de bien l'assoire dans l'intro.

Bon désolée pour le pavé, et désolée de pas avoir réussi à écrire direct dans le document. Mais j'ai préféré écrire mes pensées quelque part plutôt que rien vous envoyer du tout faute de temps !
Bonne soirée:)

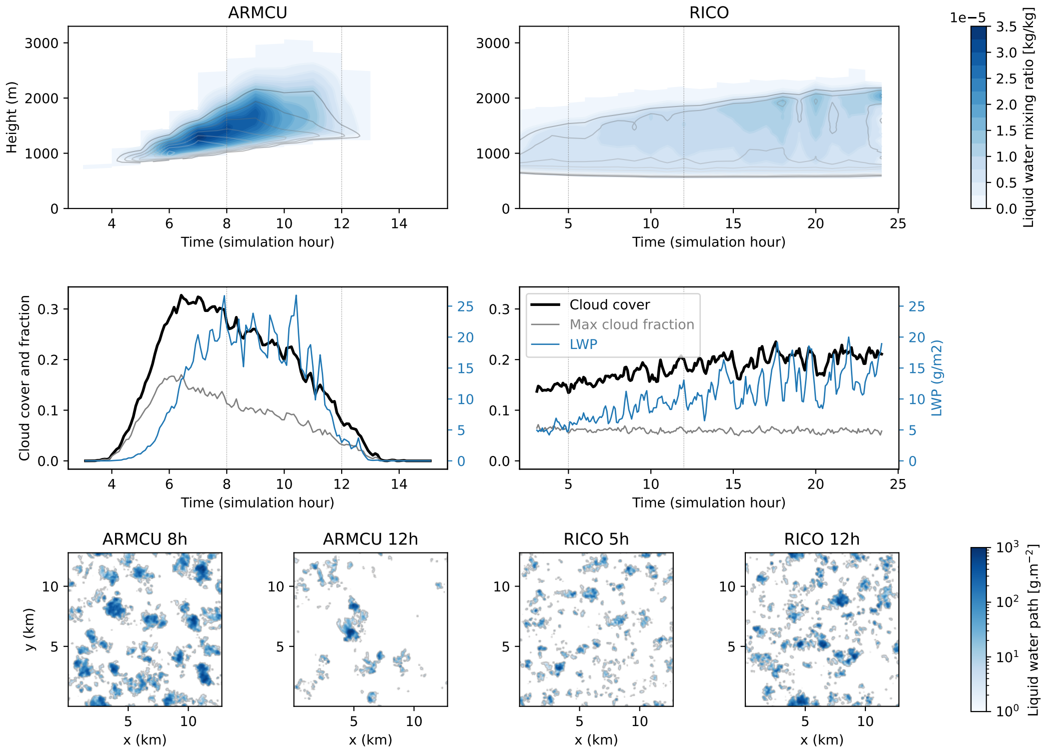

Figure 1. LES data for ARMCU (left) and RICO (right) cases. Contours of cloud fraction and cloud water mixing ratio vertical profiles as a function of time; evolution of total cloud cover, maximum cloud fraction and domain-mean liquid water path; liquid water path maps at the two hours of interest (ARMCU 8th and 12th hours, RICO 5th and 12th hours

2.1 Reference data

Two typical cumulus cloud cases are used from the set of idealized cases that are distributed in a standardized format by Dephy (Développement et Evaluation PHYsiques des modèles atmosphériques, Dephy (2020)). The ARMCU case (Brown et al., 2002) is typical of the development of boundary-layer clouds over continent during the day, while the RICO case (vanZanten et al., 2011) is typical of trade-wind cumulus developing over a stationnary ocean. LES of these two cases are run with the Meso-NH model (Lafore et al., 1998; Lac et al., 2018) at 25 metres horizontal and vertical resolutions on a 12.8 $\times$ 12.8 $\times$ 4 km^3 domain. Large scale dynamics, radiative heating and surface conditions are imposed. These simulations provide reference values for the thermodynamic and cloud variables, and their uncertainties are quantified running sensitivity experiments to numerical and physics options as described in Couvreux et al. (2021).

Table 1. Quelques propriétés des champs nuageux LES...

cas	armcu008	armcu012	rico005	rico012
cover	0.26	0.07	0.13	0.20
max frac	0.11	0.03	0.06	0.06
epaisseur	1.5	1.3	1.1	1.5
cover/mf	2.36	2.3	2.24	3.35

Reference solar fluxes are computed using a 3D Monte Carlo (MC) code run on 3D cloud fields extracted every hour from the LES, as described in Villefranque et al. (2019). 3D fields of liquid water content are taken from the LES and cloud-droplet effective radius is homogeneously set to 10 μm. Cloud optical properties are obtained from Mie theory. Gas optical properties are calculated using the k-distribution model RRTMG-IFS included in the ecRad radiation scheme (Hogan & Bozzo, 2018), for temperature, pressure and humidity profiles corresponding to the LES horizontal mean below 4 km, and to Standard Mid-Latitude Summer profile above. Solar constant is set to 1368 W.m^{-2} and surface albedo to 0.08. Additional MC calculations are made under the Independent Columns Assumption (ICA), which removes 3D radiative effects from the calculation. Differences between 3D and ICA MC fluxes yield estimates of 3D radiative effects.

2.2 LMDZ Single-Column Model

LMDZ-6A (Hourdin et al., 2020) is the atmospheric component of the IPSL-6A General Circulation Model, which participated in the sixth phase of the Coupled Model Intercomparison Project (CMIP6). Here, its single-column version is used with a refined 95-level grid as in Hourdin et al. (2019, 2021) to simulate ARMCU and RICO cases. The same large-scale dynamics, radiative trends and surface conditions are imposed as in the LES so that physical parameterizations are the only active part of the model. **Dire quelque chose du fait qu'on sait que RICO a tendance à exploser / déclencher de la cvp ?**

More specifically, the parameterizations that are active here are the boundary-layer transport and cloud schemes. The parameterization of vertical sub-grid transport is based on an Eddy-Diffusivity and Mass-Flux approach. The Eddy-Diffusivity model parameterizes the effects of small-scale turbulence on the mean state using the Turbulent Kinetic Energy prognostic equation formulated by Yamada (1983) with a 1.5-order closure. The Mass-Flux model parameterizes the effects of organized convective cells or rolls on the mean state using an effective thermal plume model. The plume transports air and state variables from the surface to the boundary-layer top. Exchanges with the environment are modeled through lateral entrainment and detrainment formulations (Hourdin et al., 2019). Water condensate and cloud fraction profiles are computed using a bi-Gaussian probability density function of the saturation deficit, with one mode accounting for saturation deficit in the thermal plume and one mode in the environment (Jam, Hourdin, Rio, & Couvreux, 2013). This combination of Eddy-Diffusivity-Mass-Flux scheme with a bi-Gaussian cloud scheme provides a unified framework that has been shown to accurately represent both dry and cloudy convective boundary layers with cloud regimes ranging from cumulus to stratocumulus (Hourdin et al., 2019). The conversion of cloud water into precipitation and the evaporation of precipitation are detailed in Madeleine et al. (2020).

2.3 Radiation parameterization

The radiative scheme under investigation in this study is ecRad, the radiative transfer model developed at European Centre for Medium-Range Weather Forecasts (Hogan & Bozzo, 2018). ecRad provides a flexible interface that allows users to configure various aspects of the radiation model. Cloud droplet effective radius, gas optics, clear-sky profiles (gas concentrations, temperature and pressure) and radiative boundary conditions are set as in the MC simulations so that they are excluded from causes of possible differences between parameterized and reference fluxes. In "perfect clouds" experiments, input liquid water content and cloud fraction profiles are taken from horizontally averaged LES 3D fields. They are hence also excluded from potential causes off differences between parameterized and reference fluxes. In SCM experiments however, liquid water content and cloud fraction profiles are taken from SCM outputs.

Cloud optics are those of the SOCRATES model (Manners, Edwards, Hill, & Thelen, 2017), which slightly differ from the reference Mie optical properties used in the Monte Carlo simulations. **non dans la dernière version d'ecrad ce n'est pas socrates mais une table de Mie, qui ressemble à celle utilisée par le Monte Carlo mais pas tout à fait les mêmes valeurs. Un mot sur l'ordre de grandeur qu'on attend comme différence ? Moins que de se tromper d'un microns sur les rayons ef-**

fectifs. En tout cas à priori cette "erreur" est la même dans les deux solvers. The RT model is a modified two-stream model (Meador & Weaver, 1980) that directly integrates the effects of cloud geometry on transport through assumptions on vertical overlap, horizontal heterogeneity and 3D effects.

Two configurations of the RT model are studied. In both, vertical overlap is represented using the exponential-random model parameterized by its decorrelation length ℓ (Hogan & Illingworth, 2000), and a two-region cloud representation based on the Triple-Clouds model (Shonk & Hogan, 2008) is used to account for in-cloud water sub-grid heterogeneity, whereby layer-wise optical depths in thin-cloud and thick-cloud regions are calculated according to the fractional standard deviation (FSD) parameter. In the `TripleClouds` configuration, no 3D effects are taken into account, whereas in the `Spartacus` configuration (Hogan, Schäfer, Klinger, Chiu, & Mayer, 2016; Schäfer, Hogan, Klinger, Chiu, & Mayer, 2016; Hogan et al., 2019), intensity of 3D effects is proportional to cloud-side perimeter length (Hogan & Shonk, 2013), itself a function of cloud fraction and cloud effective scale (C_s). It was shown previously that 3D effects in cumulus clouds account for around 10 W/m^2 of the reflected flux at high suns and removes 10 W/m^2 at low suns.

2.4 The High-Tune:Explorer tuning tool

ça ou une version abrégée ?? plus besoin de détailler aussi rigoureusement les convergences etc car on s'en servira pas vraiment, mais intéressant d'avoir ce texte en annexe ? Je le retravaille pas pour l'instant.

High-Tune:Explorer is a tuning tool based on History Matching with iterative refocusing (Vernon, Goldstein, & Bower, 2010; Williamson et al., 2013). It aims at finding the subspace of model free parameters that match a set of constraints. The parameter space hypercube, $[\lambda^1_{min}, \lambda^1_{max}] \times [\lambda^2_{min}, \lambda^2_{max}] \times ... \times [\lambda^N_{min}, \lambda^N_{max}]$, with $\lambda^1, ..., \lambda^N$ the N free parameters to tune, is iteratively reduced by ruling out parameter vectors for which the model's predictions, for a set of user-defined metrics, do not match reference values within the range of user-defined tolerance to error.

To accelerate the exploration of the hypercube, Gaussian Process based emulators are build for each metric. Emulators are trained on metrics computed from an ensemble of model runs (with typically $10 \times N$ members), to then provide rapid predictions of metric values for huge sets of free parameter vectors.

Gaussian Processes provide a prediction of the i-th metric at a given point of the parameter space as the expectation μ_i of a random variable together with its standard deviation σ_i. The prediction uncertainty is combined with tolerance to error to avoid ruling out parameter vectors that might in fact be acceptable configurations of the model. To this end, the implausibility is defined as a function of parameter vector $\boldsymbol{\lambda}$,

$$I(\boldsymbol{\lambda}) = \max\left\{\frac{\mid r_1 - \mu_1(\boldsymbol{\lambda})] \mid}{\sqrt{\sigma_1^2(\boldsymbol{\lambda}) + T_1^2}}; ...; \frac{\mid r_p - \mu_p(\boldsymbol{\lambda})] \mid}{\sqrt{\sigma_p^2(\boldsymbol{\lambda}) + T_p^2}}\right\}, \tag{1}$$

with r_i the reference (target) value and T_i its tolerance to error.

The parameter vector $\boldsymbol{\lambda}$ is ruled out if its implausibility $I(\boldsymbol{\lambda})$ is greater than an arbitrary value Γ, which represents the size of the confidence interval (reference $\pm\Gamma$ times uncertainty), typically between 2 and 3. At the end of each iteration, the new Not-Ruled-Out-Yet (NROY) space of parameters is determined using this implausibility condition. The next iteration starts by sampling a set of parameter vectors in the NROY space of the previous iteration. Then a new ensemble is run, metrics are evaluated, emulators are built, etc.

As the iterative process progresses, the NROY space narrows down, mostly because emulators uncertainty decreases, which is due to denser information being collected for training (same amount of points in a smaller NROY space). The tuning experiment is considered to have strictly converged when emulator uncertainties are significantly smaller than tolerances to error for every metric. In this case, the final NROY space is exactly the subspace of free parameters that matches the user-defined constraints and emulators can be considered perfect models for the metrics.

In practice, emulator uncertainties rarely fall one order of magnitude under tolerances to error for all metrics and hence the experiment rarely strictly converges. It is therefore useful to define another kind of convergence: The experiment is considered to have weakly converged when adding new iterations does no longer significantly reduce the NROY space. In that case, the final NROY space is larger than the sought parameter subspace. To investigate the quality of model configurations still in the NROY space, a score $S(\boldsymbol{\lambda})$ is defined as

$$S(\boldsymbol{\lambda}) = \max\left\{\frac{\mid r_1 - f_1(\boldsymbol{\lambda}) \mid}{T_1}; ...; \frac{\mid r_p - f_p(\boldsymbol{\lambda}) \mid}{T_p}\right\}, \tag{2}$$

where $f_i(\boldsymbol{\lambda})$ is the actual model output for metric i and parameter vector $\boldsymbol{\lambda}$ (instead of emulator prediction in Equation (A1)). This score is used to select a set of best simulations (those with smallest scores).

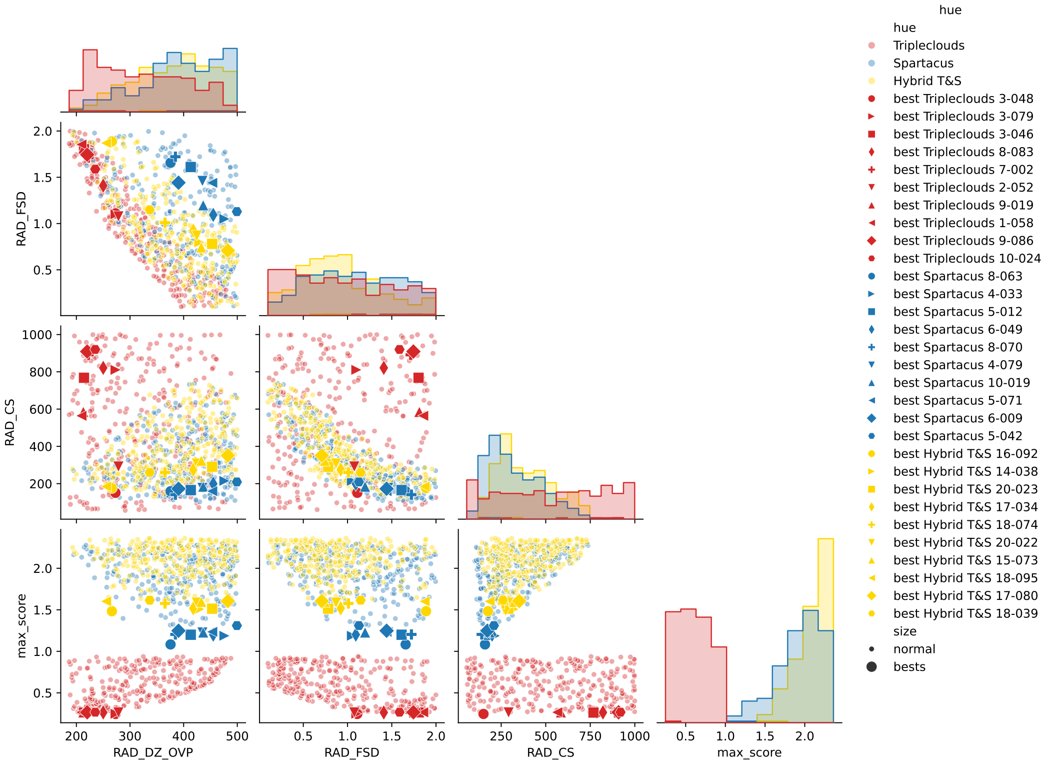

Figure 2. Parameters of 300 best simulations for the three ecRad offline tuning experiments

3 Compensations internes au schéma de rayonnement

First, we examine ecRad PPEs in a "perfect cloud" framework, and investigate compensating effects between overlap, heterogeneity and cloud size.

3.1 Les best tuning de triplecloulds visant du ICA et de spartacus visant du 3D sont incompatibles

First, we look at two experiments: one where triplecloulds is run on LES cloud profiles and the target is Monte Carlo 1D (ICA) fluxes at three solar zenith angles; and one where spartacus is run on LES cloud profiles and the target is Monte Carlo 3D fluxes at (the same) three solar zenith angles.

- Triplecloulds has better scores than spartacus, remember they do not compare to the same reference: it seems that it is easier to do 1D radiation than 3D... spartacus is not yet as good at 3D radiation as triplecloulds is at 1D radiation! We can still improve spartacus :)

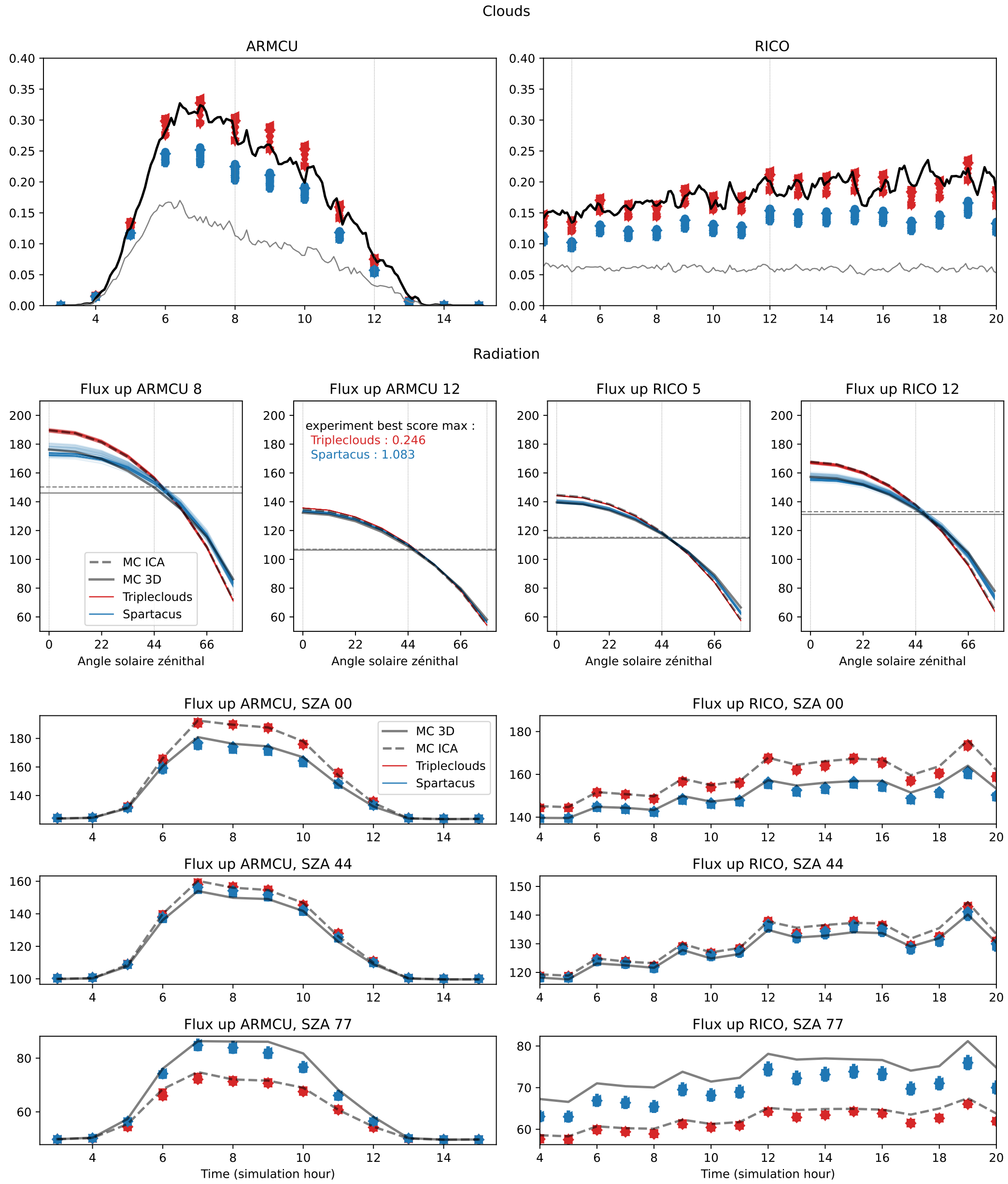

Figure 3. Cloud and radiation variables for 50 best simulations of two perfect-cloud experiments (Tripleclouds in red, Spartacus in blue), compared to reference. Left: cloud cover (black) and maximum cloud fraction (gray) in LES fields, and cloud cover computed by ecRad (color points) using input cloud fraction profiles horizontally averaged from LES at simulation hours 8 and 12 for ARMCU and 5 and 12 for RICO, and the exponential–random overlap model parameterized by various decorrelation lengths (RAD_DZ_OVP values in Figure 2). Right: upward TOA flux as a function of solar zenith angle, for the 4 cloud scenes used as constraints in the tuning experiment: ARMCu 8th and 12th hours, RICO 5th and 12th hours.

- Note that the RAD_CS parameter is not used in tripleclouds so the location of tripleclouds configurations in RAD_CS space is uniform random
- In Tripleclouds, larger decorrelation length leads to larger error, while in spartacus, it is the opposite.
- As a result, best tripleclouds have smaller decorrelation length, which means larger cloud covers, than best spartacus.
- There is a strong fsd / overlap relationship in Tripleclouds (less so in spartacus). Large fsd = more heterogeneous = less reflective, associated with small decorrelation length = more cover = more reflective leads to same flux as smaller fsd + larger decorrelation length.
- In Spartacus, large FSD appears to be preferentially associated with smaller clouds. Highly contrasted clouds and large transfer between regions yields the same fluxes as less contrasted clouds and less transfer between regions. Is this expected? Large FSD = less reflective, smaller clouds = more 3D effects = more reflective when the sun is low, but less reflective when the sun is high... Need to better understand interactions between FSD and 3D effects? Also, what is the perimeter length of heterogeneities?
- Configurations with large values of both fsd and overlap decorrelation length are not valid for tripleclouds but they are valid for spartacus.
- Best tripleclouds and best spartacus are not in the same parameter space. Spartacus does not just "add" 3D effects to tripleclouds. Effects of the three parameters are intertwined in the model. Changing cloud size from infinite to finite, to add 3D effects, leads to selecting new fsd × overlap parameters as best config. Parameters are *effective parameters*.

rajouter les y labels et units sur les figures et tout mettre en anglais

3.2 on trouve une configuration compatible avec les deux contraintes : compromis

Pour dépasser ces incompatibilités apparentes, on fait une expérience hybride qui permet de remplir les deux types de contraintes, quitte a dégrader un peu les scores

- on trouve des configurations compatibles, avec la tolérance qu'on s'était donné mais elles ne sont pas au même endroit que les bests des expériences précédentes.
- two of the ten bests = large fsd, small decorr len, small clouds ; vs fsd around or below 1, larger decorr len, slightly larger clouds.
- ils sont sur une droite overlap x cloud size ; c'est sous contraint ; quand on regarde les flux en fonction de l'angle solaire, tous les bests sont aussi nuls les uns que les autres...
- should we use cloud cover to constrain overlap parameter? Cloud cover from the LES 3D fields was computed and compared to cloud cover diagnosed by ecRad for the different configurations, that is, different overlap decorrelation lengths.
- the best hybrid is also the best cover compared to LES
- **look at individual metrics for hybrid only?**
- we can see in fig 4 that the best hybrid (simulation 16-092, blue disk), is also the "best cover" (with also the yellow diamond, simulation 18-095) ; these are the two violet and orange triangles in fig 1, with FSD approx 2 and small overlap decorrelation lengths.

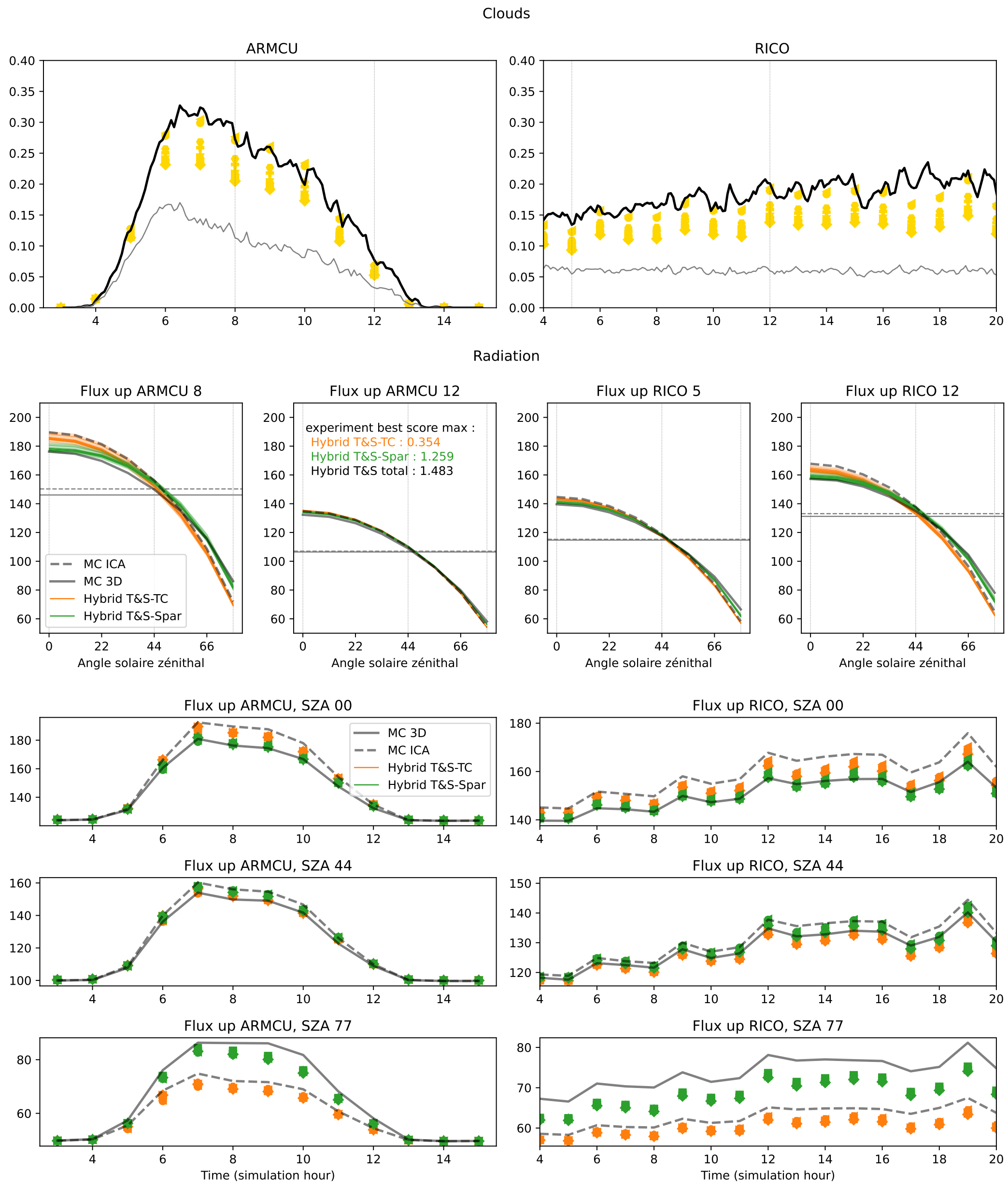

Figure 4. Cloud and radiation variables for 50 best simulations of perfect-cloud "Hybrid" experiment, compared to reference. Left: cloud cover (black) and maximum cloud fraction (gray) in LES fields, and cloud cover computed by ecRad (color points) using input cloud fraction profiles horizontally averaged from LES at simulation hours 8 and 12 for ARMCU and 5 and 12 for RICO, and the exponential–random overlap model parameterized by various decorrelation lengths (RAD_DZ_OVP values in Figure 2). ecRad configurations are those of the hybrid experiment, run either with TripleClouds (orange) or Spartacus (green). Right: upward TOA flux as a function of solar zenith angle, for the 4 cloud scenes used as constraints in the tuning experiment: ARMCu 8th and 12th hours, RICO 5th and 12th hours.

4 Compensations rayonnement / nuage

Second, we examine PPEs of the LMDZ SCM model run for our two cumulus cases under radiative constraints, using the ecRad configuration chosen in previous section.

We investigate cloud–radiation compensation errors.
Pas ouf les scatter matrix pour ces expériences. C'était mieux les figures de Maelle avec max frac vs flux ?

4.1 Les best tuning de tripleclouds visant trois angles font des max de fraction trop grandes par rapport à la LES et à spartacus

- Various ecRad configurations from the previous experiments were selected and SCM tuning was performed with these versions of ecRad, using either tripleclouds or spartacus
- Contrary to the previous section, we target MC 3D even for tripleclouds. We want to see if errors in the radiation scheme are compensated by selecting wrong clouds. Remember that in a true GCM tuning process, the target would be fluxes infered from satellite measurements, hence would include 3D effects.
- tolerance = 9 W/m2 for tripleclouds, 3.5 for spartacus, target = flux at three angles, Monte Carlo 3D
- for configuration best spartacus (8-063 on Fig2), the max of cloud frac is systematically overestimated by tripleclouds, but not by spartacus Figure 5
- for configuration best hybrid (16-092 on Fig2), it is still the case but less strong Figure 7
- when ecRad is optimally tuned for spartacus, tuning clouds using spartacus constraints leads to the right clouds; however, tuning clouds using tripleclouds constraints leads to selecting LMDZ configurations in which cloud fractions are overestimated. A possible explanation is that we saw in Fig 3 that for this ecrad configuration, the overlap parameter leads to underestimating cloud cover (blue points). When spartacus is used, this "lack" of cover is compensated by 3D effects and the fluxes remain correct. However, when tripleclouds is used, this lack of cover cannot be internally compensated; hence to still have the right fluxes (because this is the constraint of the tuning experiment), cloud fractions are overestimated.
- the best hybrid ecRad configuration, however, does not underestimate cloud cover as much, as we saw in Fig4; this configuration is supposed to work reasonably well for both tripleclouds and spartacus (even if it is a little less accurate for spartacus compared to the "best spartacus" configuration)
- as a result, cloud fractions selected using tripleclouds constraints are closer to those of the LES than before, and the spartacus ones are a bit further.

4.2 How do RT models fare at computing radiative fluxes?

Here, final tolerances to error are 3.5 W.m^{-2} for `Spartacus`, 6 W.m^{-2} for `TripleClouds` and 7 W.m^{-2} for `Homogeneous`. Yet, experiments have not strictly converged per the definition of Section A1, as for some metrics, tolerances do not dominate emulator uncertainties. Therefore, they do not provide direct information on the structural error.

For a sounder analysis, LMDZ+ecRad simulations of each RT configuration are ranked according to their score as defined by Equation (A2). The product of a simulation's score

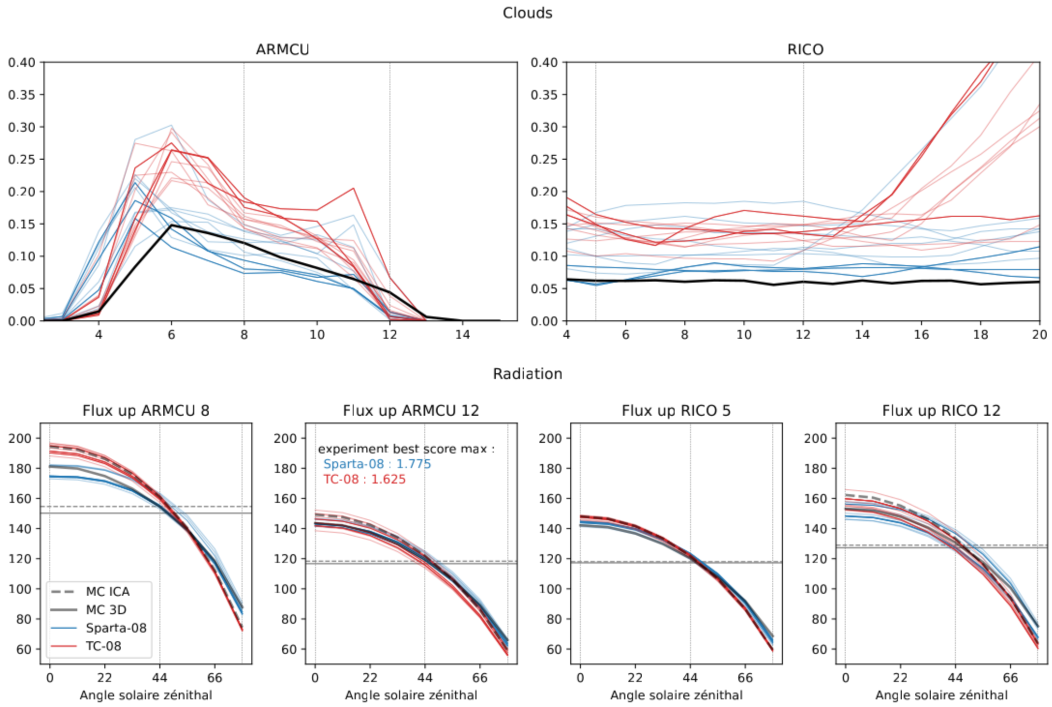

Figure 5. Clouds and radiation for the 10 best for two experiments calibrating LMDZ parameters using radiative constraints: tripleclouds with config best spartacus targetting 3 angle fluxes in red, spartacus with config best hybrid targetting 3 angle fluxes in blue. Left: ARMCU; Right:RICO. First row: maximum cloud fractions; Second row: upwelling TOA fluxes as a function of solar zenith angle.

by experiment's tolerance to error is the error associated with the worst metric of the simulation. The worst-metric error of the best simulation is retained as a measure of the accuracy of the RT configuration: 4.1 W.m^{-2}, 11.3 W.m^{-2} and 12.8 W.m^{-2} for `Spartacus`, `TripleClouds`, `Homogeneous` configurations respectively.

`Spartacus` accuracy is only slightly larger (+ 1 W.m^{-2}) than the accuracy of SPARTACUS run on reference cloud profiles as determined by Villefranque et al. (2021). This means that replacing LES cloud profiles by tuned LMDZ outputs has only a small effect on radiation, which implies that the best LMDZ configurations produced by radiative-based tuning yield similar cloud profiles as found in the reference LES, from a radiative perspective at least. On the other hand, `TripleClouds` and `Homogeneous` configuration errors are about three times larger than those of `Spartacus`. Lack of 3D radiative effects, and, to a lesser extent, of in-cloud horizontal heterogeneity and non-maximum vertical overlap, leads to a significant increase in solar radiation errors. Barker et al. (1999) show that overly simplistic overlap and heterogeneity assumptions, when considered separately and for 2 km-resolution cloud fields, are responsible for larger errors than 3D effects. Yet, as they act in opposite directions, the effect of combined exponential overlap and heterogeneity compared to maximum overlap and homogeneous clouds (`TripleClouds` vs `Homogeneous`) becomes much smaller than 3D effects (`Spartacus` vs `TripleClouds`).

Next, features of the 30 best simulations of each tuning experiment are analyzed. Figure **??**(a) shows that even though solar radiation is constrained at only two instants of each case, SCM simulations still capture ARMCU solar radiative fluxes diurnal cycle well enough (albeit with some difficulties in the first hours). This illustrates the power of machine learning when it is used at the service of physics: physically-based models contain enough structural constraints to extrapolate accurately the temporal evolution of the atmosphere from a small amount of learning data.

Figure **??**(b) shows that solar radiative fluxes of the first hours of RICO are also rather accurate, but a drift towards overly-bright cloud scenes occurs in every selected simulations after the time corresponding to the second metric (11:30). In the remaining of the analysis, only the first 12 hours of RICO are commented.

Figure **??**(a-b) also show that fluxes in the 30 best `Spartacus` simulations are almost always closer to reference values than the other two ecRad experiments. This again demonstrates the importance of including 3D radiative effects in solar RT parameter-

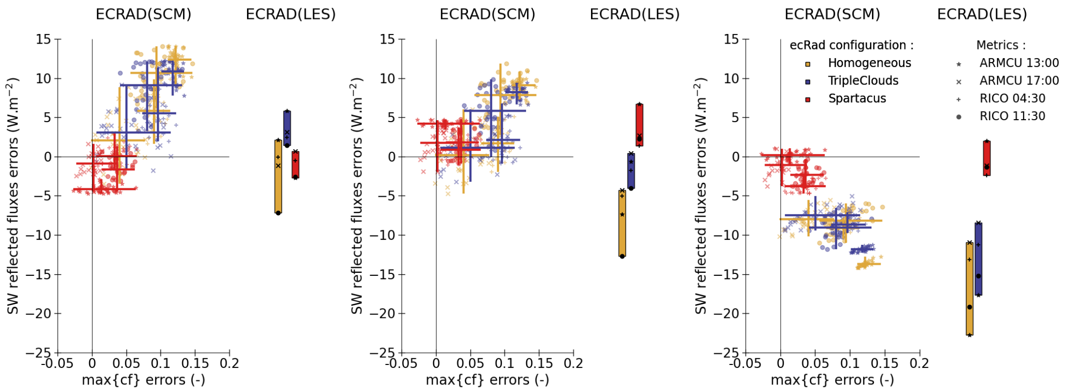

Figure 6. Reflected solar flux error (W.m^{-2}) vs maximum cloud fraction error for the 30 best simulations of each tuning experiment and for solar zenith angles of 0° (left), 44° (middle) and 77° (right). Different colors are associated with different ecRad configurations. Different symbols are associated with different metrics (case and hour). Bold crosses indicate the range of values covered by the 30 best simulations for each metric. Bar plots on the right represent errors when ecRad configurations are run on reference cloud profiles taken from LES instead of SCM outputs.

izations. Interestingly, `TripleClouds` and `Homogeneous` best simulations yield fluxes that are close to the ICA (no 3D effects) MC estimates.

4.3 Does tuning improve radiation at the expense of clouds?

Figure **??**(c-d) shows that in addition to better simulating solar radiation, the 30 best `Spartacus` simulations exhibit layer-wise maximum cloud fraction evolutions that closely match reference values. Conversely, the 30 best `TripleClouds` and `Homogeneous` simulations persistently overestimate maximum cloud fractions. Figure **??**(e-f) shows that liquid water path of selected simulations are close to reference values regardless of the radiative configuration used in the tuning experiment.

To evaluate the part of the error that comes from the radiative transfer model, ecRad is run on the reference cloud profiles averaged from the reference LES outputs. Figure 6 shows that, at 77°, `TripleClouds` and `Homogeneous` errors on LES profiles are the largest (up to -23 W.m^{-2}). This is due to lack of 3D effects, which leads to strong underestimation of reflectivity when the sun is low on the horizon. In the corresponding tuning experiments however, flux errors at 77° are systematically smaller than LES+ecRad ones, by up to 9 W.m^{-2}. Concurrently, maximum cloud fractions of the SCM clouds are significantly overestimated. This is the sign of compensating errors: The tuning process selects SCM configurations that overestimate cloud fractions because they partly compensate structural errors in the incomplete radiative transfer model. However, increasing cloud fractions increases reflected fluxes for all metrics, and therefore deteriorates fluxes at 0° where lack of 3D effects already led to an overestimation of reflected fluxes. As implausibilty is constrained by the worst metric, a compromise must be reached between errors at 0° and 77°. This leads to reflectivity overestimation at 0° and underestimation at 77° that are of equal magnitude. None of the examined `Spartacus` simulations exhibit such compensating error mechanisms.

To verify that this result is robust, the experiments are repeated with additional metrics: at 11:00 LT in ARMCU to better capture the beginning of the diurnal cycle, and at 21:30 in RICO to try preventing the drift in the second half of the simulation. While adding metrics slightly deteriorates fluxes for all configurations, we observe the same compensating error mechanisms as before in `TripleClouds` and `Homogeneous` (see text and figures in Supplementary Materials).

4.4 On trouve une configuration en visant un angle moyen : relache la contrainte

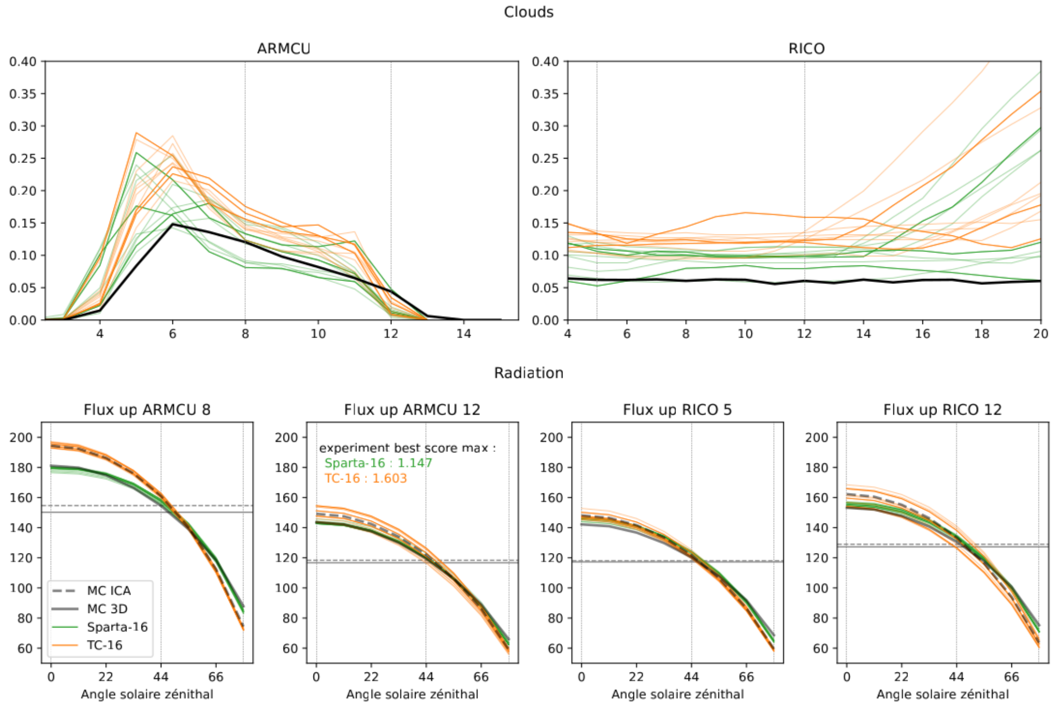

Figure 7. Clouds and radiation for the 10 best for two experiments calibrating LMDZ parameters using radiative constraints: tripleclouds with config best hybrid targetting 3 angle fluxes in orange, spartacus with config best hybrid targetting 3 angle fluxes in green. Left: ARMCU; Right:RICO. First row: maximum cloud fractions; Second row: upwelling TOA fluxes as a function of solar zenith angle.

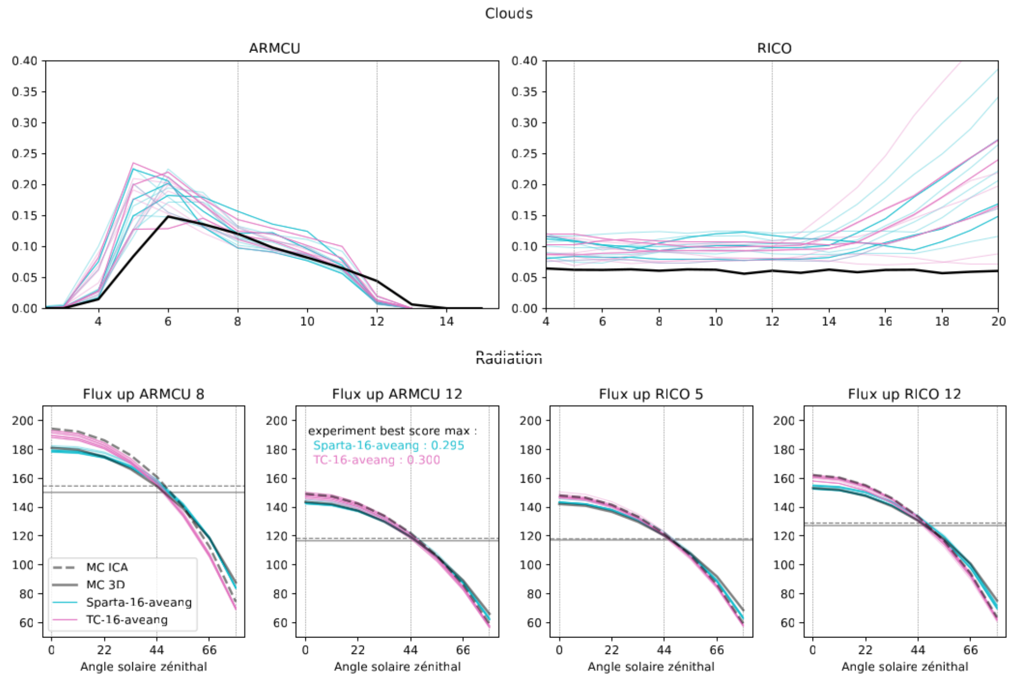

Figure 8. Clouds and radiation for the 10 best for two experiments calibrating LMDZ parameters using radiative constraints: tripleclouds with config best hybrid targetting mean-angle flux in pink, spartacus with config best hybrid targetting mean-angle flux in cyan. Left: ARMCU; Right:RICO. First row: maximum cloud fractions; Second row: upwelling TOA fluxes as a function of solar zenith angle.

5 Conclusion

We have shown that structural errors in radiative transfer models can indeed be compensated by errors in cloud properties when TOA radiative fluxes are targeted in a tuning process. Here, maximum cloud fractions are overestimated to compensate for underestimated cloud reflectivity at large zenith angles, stemming from the lack of 3D effects in the radiative model. This result provides a novel argument in favor of modelling 3D radiative effects in climate models: even if they were small on average and had a weak feedback on circulations and climate, we have shown that systematic errors in radiative transfer can generate systematic errors in other components of the model through tuning. A better radiative transfer model might remove the need for compensating errors and result in better clouds.

Here the demonstration was made in an idealized configuration, and our results cannot be directly extrapolated to climate simulations. Indeed, in the SCM setup considered here, only convection and cloud parameterizations can compensate structural radiative errors, whereas much more processes are at work in a 3D GCM, which can result in other compensating errors. Also, radiation was calculated as an off-line diagnosis whereas in a GCM radiative fluxes do feedback on clouds and dynamics. In addition, fluxes were targeted independently at various solar angles, which is particularly difficult to achieve without 3D effects; whereas GCM calibration targets fluxes that are integrated over time and space, where positive and negative 3D effects partly cancel out, therefore the error to compensate tends to be smaller. Finally, radiative fluxes were the only constraints, but Couvreux et al. (2021); Hourdin et al. (2021) suggest that compensating errors can be prevented or at least limited by process-based tuning in SCM mode before tuning the full GCM. With this strategy, constraints can be set directly on cloud properties to rule out model configurations that yield wrong cloud fractions. Note however that tuning towards radiative targets while preventing clouds from compensating radiation errors might generate compensating errors elsewhere in the system.

Our work goes beyond the question of radiative transfer and clouds: We propose to use tuning as a tool to investigate compensating errors and guide model development. Through tuning we explore parameter space, that is, model configurations and resulting climates, under a set of chosen constraints. This allows us to disentangle parametric from structural errors. Notably, when no set of parameters can be found for which all simulated metrics comply with user requirements, it indicates that structural errors are larger than tolerated errors, and hence that the model is incomplete. This is a powerful way to guide its development and accelerate its improvement. When simulated metrics do comply with prescribed requirements, resulting perturbed parameter ensembles of simulations (PPE) can be used to investigate compensating errors, better understand the model and its physics through global sensitivity studies, and quantify parametric uncertainty on various aspects of climate.

The tuning tool used here, High-Tune:Explorer, is based on machine learning techniques: predictive Gaussian Processes are trained on a small amount of simulated data and are then able to emulate the model's response much faster than the actual model. Thanks to this approach, the model's high-dimensional parameter space can be explored and shrunken efficiently. Machine learning is here at the service of physics; it helps saving computing time but not at the expense of the physical consistency of the model. This consistency is crucial for our confidence in climate projections and to keep using models as a tool to better understand climate.

Open Research Section

The High-Tune Explorer (htexplo) and LMDZ model are available through the open source version control system "subversion" (svn). htexplo is distributed under the GPL-v3 license, and LMDZ is distributed under the CeCILL version 2 license. The htexplo release used in the study can be downloaded through `svn checkout http://svn.lmd.jussieu.fr/HighTune -r 437`. The LMDZ release used in the study can be downloaded through `svn -r 4586 checkout http://svn.lmd.jussieu.fr/LMDZ/LMDZ6/trunk`. It can be configured and installed directly on Linux machines with an installation bash script `https://lmdz.lmd.jussieu.fr/pub/instal` running as `bash install_lmdz.sh -SCM -v 20230626.trunk` The ecRad offline package is freely available under the terms of the Apache License Version 2.0. The release used in this study corresponds to commit 210d791, which is based on version v1.6-beta. A tar file of the htexplo, LMDZ and ecRad codes used as well as the data that supports this research, the results of the SCM simulations, as well as the scripts for visualization WILL BE MADE AVAILABLE ON A DOI IF THE PAPER IS ACCEPTED FOR PUBLICATION. The corresponding DOIs will be provided during galley proofs by placeholder "IPSL data catalog."

Acknowledgments

The PhD scholarship of the first author was funded by Institut Pierre-Simon Laplace. This work was supported by the CNRS and the GDR DEPHY. The authors wish to thank the LMDZ team and the htexplo team for their investment in the development and maintainance of community tools.

References

Barker, H. W., Stephens, G. L., & Fu, Q. (1999). The sensitivity of domain-averaged solar fluxes to assumptions about cloud geometry. *Quarterly Journal of the Royal Meteorological Society*, *125*(558), 2127-2152. Retrieved from `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712555810` doi: 10.1002/qj.49712555810

Brown, A. R., Cederwall, R. T., Chlond, A., Duynkerke, P. G., Golaz, J.-C., Khairoutdinov, M., ... Stevens, B. (2002). Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Quarterly Journal of the Royal Meteorological Society*, *128*(582), 1075–1093. doi: 10.1256/003590002320373210

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., ... Xu, W. (2021). Process-Based Climate Model Development Harnessing Machine Learning: I. A Calibration Tool for Parameterization Improvement. *Journal of Advances in Modeling Earth Systems*, *13*(3), e2020MS002217. doi: 10.1029/2020MS002217

Dephy. (2020). *Dephy-scm: Single-column model standards and case drivers.* Retrieved from `https://github.com/gdr-dephy/dephy-scm`

Hogan, R. J., & Bozzo, A. (2018). A Flexible and Efficient Radiation Scheme for the ECMWF Model. *Journal of Advances in Modeling Earth Systems*, *10*(8), 1990–2008. doi: 10.1029/2018MS001364

Hogan, R. J., Fielding, M. D., Barker, H. W., Villefranque, N., & Schäfer, S. A. K. (2019). Entrapment: An Important Mechanism to Explain the Shortwave 3D Radiative Effect of Clouds. *Journal of the Atmospheric Sciences*, *76*(7), 2123-2141. doi: 10.1175/JAS-D-18-0366.1

Hogan, R. J., & Illingworth, A. J. (2000, October). Deriving cloud overlap statistics from radar. *Quarterly Journal of the Royal Meteorological Society*, *126*(569), 2903–2909. doi: 10.1002/qj.49712656914

Hogan, R. J., Schäfer, S. A. K., Klinger, C., Chiu, J. C., & Mayer, B. (2016). Representing 3-D cloud radiation effects in two-stream schemes: 2. Matrix formulation and broadband evaluation. *Journal of Geophysical Research: Atmospheres*, *121*(14), 8583–8599. doi: 10.1002/2016JD024875

Hogan, R. J., & Shonk, J. K. P. (2013, February). Incorporating the Effects of 3D Radiative Transfer in the Presence of Clouds into Two-Stream Multilayer Radiation Schemes. *Journal of the Atmospheric Sciences*, *70*(2), 708–724. doi: 10.1175/JAS-D-12-041.1

Hourdin, F., Ferster, B., Deshayes, J., Mignot, J., Musat, I., & Williamson, D. (2023, July). Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections. *Science Advances*, *9*(29), eadf2758. doi: 10.1126/sciadv.adf2758

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. *Climate Dynamics*, *40*(9-10), 2193–2222. doi: 10.1007/s00382-012-1343-y

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. *Journal of Advances in Modeling Earth Systems*, *11*(9), 2910–2933. doi: 10.1029/2019MS001666

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. *Bulletin of the American Meteorological Society*, *98*(3), 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. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS001892. doi: 10.1029/2019MS001892

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., … Volodina, V. (2021). Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global.

Journal of Advances in Modeling Earth Systems, *13*(6), e2020MS002225. doi: 10.1029/2020MS002225

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 Meteorology*, *147*(3), 421–441. doi: 10.1007/s10546-012-9789-3

Konsta, D., Dufresne, J.-L., Chepfer, H., Vial, J., Koshiro, T., Kawai, H., ... Ogura, T. (2022). Low-Level Marine Tropical Clouds in Six CMIP6 Models Are Too Few, Too Bright but Also Too Compact and Too Homogeneous. *Geophysical Research Letters*, *49*(11), e2021GL097593. doi: 10.1029/2021GL097593

Lac, C., Chaboureau, J.-P., Masson, V., Pinty, J.-P., Tulet, P., Escobar, J., ... Wautelet, P. (2018, May). Overview of the Meso-NH model version 5.4 and its applications. *Geoscientific Model Development*, *11*(5), 1929–1969. doi: 10.5194/gmd-11-1929-2018

Lafore, J. P., Stein, J., Asencio, N., Bougeault, P., Ducrocq, V., Duron, J., ... Vilà-Guerau de Arellano, J. (1998, January). The Meso-NH Atmospheric Simulation System. Part I: Adiabatic formulation and control simulations. *Annales Geophysicae*, *16*(1), 90–109. doi: 10.1007/s00585-997-0090-6

Madeleine, J.-B., Hourdin, F., Grandpeix, J.-Y., Rio, C., Dufresne, J.-L., Vignon, E., ... Bonazzola, M. (2020). Improved Representation of Clouds in the Atmospheric Component LMDZ6A of the IPSL-CM6A Earth System Model. *Journal of Advances in Modeling Earth Systems*, *12*(10), e2020MS002046. doi: 10.1029/2020MS002046

Manners, J., Edwards, J. M., Hill, P., & Thelen, J.-C. (2017). *SOCRATES Technical Guide Suite Of Community RAdiative Transfer codes based on Edwards and Slingo* (Tech. Rep.). FitzRoy Rd, Exeter EX1 3PB: Met Office. Retrieved from `http://homepages.see.leeds.ac.uk/~lecsjed/winscpuse/socrates_techguide.pdf`

McKee, T. B., & Cox, S. K. (1974). Scattering of Visible Radiation by Finite Clouds. *Journal of the Atmospheric Sciences*, *31*(7), 1885–1892. doi: 10.1175/1520-0469(1974)031⟨1885:SOVRBF⟩2.0.CO;2

Meador, W. E., & Weaver, W. R. (1980, March). Two-Stream Approximations to Radiative Transfer in Planetary Atmospheres: A Unified Description of Existing Methods and a New Improvement. , *37*(3), 630–643. doi: 10.1175/1520-0469(1980)037⟨0630:TSATRT⟩2.0.CO;2

Nam, C., Bony, S., Dufresne, J.-L., & Chepfer, H. (2012). The 'too few, too bright' tropical low-cloud problem in CMIP5 models. *Geophysical Research Letters*, *39*(21). doi: 10.1029/2012GL053421

Schäfer, S. A. K., Hogan, R. J., Klinger, C., Chiu, J. C., & Mayer, B. (2016). Representing 3-D cloud radiation effects in two-stream schemes: 1. Longwave considerations and effective cloud edge length. *Journal of Geophysical Research: Atmospheres*, *121*(14), 8567–8582. doi: 10.1002/2016JD024876

Shonk, J. K. P., & Hogan, R. J. (2008, June). Tripleclouds: An Efficient Method for Representing Horizontal Cloud Inhomogeneity in 1D Radiation Schemes by Using Three Regions at Each Height. *Journal of Climate*, *21*(11), 2352–2370. doi: 10.1175/2007JCLI1940.1

Shonk, J. K. P., Hogan, R. J., Edwards, J. M., & Mace, G. G. (2010, July). Effect of improving representation of horizontal and vertical cloud structure on the Earth's global radiation budget. Part I: review and parametrization. *Quarterly Journal of the Royal Meteorological Society*, n/a–n/a. doi: 10.1002/qj.647

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

Várnai, T., & Davies, R. (1999). Effects of cloud heterogeneities on shortwave radiation: Comparison of cloud-top variability and internal heterogeneity. *Journal of the Atmospheric Sciences*, *56*(24), 4206–4224.

Vernon, I., Goldstein, M., & Bower, R. G. (2010, December). Galaxy Formation: A Bayesian Uncertainty Analysis. *Bayesian Analysis*, *05*(04). doi: 10.1214/10-ba524

Villefranque, N., Blanco, S., Couvreux, F., Fournier, R., Gautrais, J., Hogan, R. J., ... Williamson, D. (2021). Process-Based Climate Model Development Harnessing Machine Learning: III. The Representation of Cumulus Geometry and Their 3D Radiative Effects. *Journal of Advances in Modeling Earth Systems*, *13*(4), e2020MS002423. doi: 10.1029/2020MS002423

Villefranque, N., Fournier, R., Couvreux, F., Blanco, S., Cornet, C., Eymet, V., ...

Tregan, J.-M. (2019). A Path-Tracing Monte Carlo Library for 3-D Radiative Transfer in Highly Resolved Cloudy Atmospheres. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2449–2473. doi: 10.1029/2018MS001602

Webb, M., Senior, C., Bony, S., & Morcrette, J. J. (2001, September). Combining ERBE and ISCCP data to assess clouds in the Hadley Centre, ECMWF and LMD atmospheric climate models. *Climate Dynamics*, *17*(12), 905–922. (WOS:000171263400001) doi: 10.1007/s003820100157

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. *Climate Dynamics*, *41*(7-8), 1703–1729. doi: 10.1007/s00382-013-1896-4

Yamada, T. (1983, January). Simulations of Nocturnal Drainage Flows by a q2l Turbulence Closure Model. *Journal of the Atmospheric Sciences*, *40*(1), 91–106. doi: 10.1175/1520-0469(1983)040⟨0091:SONDFB⟩2.0.CO;2

Appendix A Tools and Methods

A1 The High-Tune:Explorer tuning tool

High-Tune:Explorer is a tuning tool based on History Matching with iterative refocusing (Vernon et al., 2010; Williamson et al., 2013). It aims at finding the subspace of model free parameters that match a set of constraints. The parameter space hypercube, $[\lambda^1_{min}, \lambda^1_{max}] \times [\lambda^2_{min}, \lambda^2_{max}] \times ... \times [\lambda^N_{min}, \lambda^N_{max}]$, with $\lambda^1, ..., \lambda^N$ the N free parameters to tune, is iteratively reduced by ruling out parameter vectors for which the model's predictions, for a set of user-defined metrics, do not match reference values within the range of user-defined tolerance to error.

To accelerate the exploration of the hypercube, Gaussian Process based emulators are build for each metric. Emulators are trained on metrics computed from an ensemble of model runs (with typically $10 \times N$ members), to then provide rapid predictions of metric values for huge sets of free parameter vectors.

Gaussian Processes provide a prediction of the i-th metric at a given point of the parameter space as the expectation μ_i of a random variable together with its standard deviation σ_i. The prediction uncertainty is combined with tolerance to error to avoid ruling out parameter vectors that might in fact be acceptable configurations of the model. To this end, the implausibility is defined as a function of parameter vector $\boldsymbol{\lambda}$,

$$I(\boldsymbol{\lambda}) = \max \left\{ \frac{\mid r_1 - \mu_1(\boldsymbol{\lambda})] \mid}{\sqrt{\sigma_1^2(\boldsymbol{\lambda}) + T_1^2}}; ...; \frac{\mid r_p - \mu_p(\boldsymbol{\lambda})] \mid}{\sqrt{\sigma_p^2(\boldsymbol{\lambda}) + T_p^2}} \right\}, \tag{A1}$$

with r_i the reference (target) value and T_i its tolerance to error.

The parameter vector $\boldsymbol{\lambda}$ is ruled out if its implausibility $I(\boldsymbol{\lambda})$ is greater than an arbitrary value Γ, which represents the size of the confidence interval (reference $\pm\Gamma$ times uncertainty), typically between 2 and 3. At the end of each iteration, the new Not-Ruled-Out-Yet (NROY) space of parameters is determined using this implausibility condition. The next iteration starts by sampling a set of parameter vectors in the NROY space of the previous iteration. Then a new ensemble is run, metrics are evaluated, emulators are built, etc.

As the iterative process progresses, the NROY space narrows down, mostly because emulators uncertainty decreases, which is due to denser information being collected for training (same amount of points in a smaller NROY space). The tuning experiment is considered to have strictly converged when emulator uncertainties are significantly smaller than tolerances to error for every metric. In this case, the final NROY space is exactly the subspace of free parameters that matches the user-defined constraints and emulators can be considered perfect models for the metrics.

In practice, emulator uncertainties rarely fall one order of magnitude under tolerances to error for all metrics and hence the experiment rarely strictly converges. It is therefore useful to define another kind of convergence: The experiment is considered to have weakly converged when adding new iterations does no longer significantly reduce the NROY space. In that case, the final NROY space is larger than the sought parameter subspace. To investigate the quality of model configurations still in the NROY space, a score $S(\boldsymbol{\lambda})$ is defined as

$$S(\boldsymbol{\lambda}) = \max \left\{ \frac{\mid r_1 - f_1(\boldsymbol{\lambda}) \mid}{T_1}; ...; \frac{\mid r_p - f_p(\boldsymbol{\lambda}) \mid}{T_p} \right\}, \tag{A2}$$

where $f_i(\boldsymbol{\lambda})$ is the actual model output for metric i and parameter vector $\boldsymbol{\lambda}$ (instead of emulator prediction in Equation (A1)). This score is used to select a set of best simulations (those with smallest scores).

A2 Design of tuning experiments

In the framework of High-Tune:Explorer, designing a tuning experiment consists in choosing a set of parameters to explore, and a set of metrics to use as constraints. Here, thirteen parameters involved in boundary-layer parameterizations are varied as in Hourdin et al. (2021) and Hourdin et al. (2023) (see details in Table S1 of Supplemental Information). Twelve metrics are used as constraints: reflected solar fluxes for two hours of ARMCU (13:00 and 17:00 LT) and RICO (4:30 and 11:30) cases, each for three solar zenith angles (0°, 44°, and 77° following the choices of Villefranque et al. (2021)). Reference values are those from the 3D MC calculations. Associated tolerances to error are the same for all metrics.

To avoid overfitting in the tuning procedure, this tolerance to error must account for uncertainties involved in the comparison between modelled and reference metrics. These include the reference uncertainty due to MC noise and LES spread, biases introduced by the use of different droplet optical property models (1 W.m^{-2} according to Villefranque et al. (2021)) and structural errors of the model which are the intrinsic errors made by LMDZ and ecRad. The latter is never fully known, nor fully defined, and one of the main outcomes of a tuning experiment is to provide insights into these structural errors. Thanks to Villefranque et al. (2021) tuning experiment, SPARTACUS's structural errors for cumulus scenes can be derived from the error distribution between SPARTACUS run on LES mean profiles (reference 1D clouds) and MC run on the LES 3D fields. This error amounts to 3.1 W.m^{-2}. Finally, by taking the square root of uncertainties quadratic sum, the *minimum* tolerance to error is set to 3.3 W.m^{-2}.

Independent tuning experiments are performed for each of the three ecRad configurations described in Section 2.3. Each tuning experiment is made of 40 iterations. At each iteration, 130 free-parameter vectors are sampled in the NROY space, then 130 versions of ARMCU and RICO are simulated using these 130 configurations of LMDZ. ecRad is then run offline for the two chosen times of the two cases and the 130 LMDZ configurations, to compute solar reflected fluxes at three solar zenith angles each; in total, 1560 ecRad runs per iteration. Then one emulator is built for each of the 12 metrics, using their 130 evaluations as a learning database. Finally, implausible free-parameter vectors are ruled out and the NROY space is narrowed down.

Each 40-iteration experiment might be repeated following a trial and error process that seeks the smallest tolerance to error yielding non-empty NROY space without falling below 3.3 $W.m^{-2}$. This value is referred to as the “final tolerance to error” in the remaining of this letter.