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

[FH : Reprendre l'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

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 remove compensating errors. The issue of reducing compensating 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 crucial for the reliability of climate change projections.

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. surface 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 those involving cloud physics and radiative transfer (RT), it is necessary to disentangle model errors stemming from each parameterization.

Cloud parameterizations typically provide vertical profiles of cloud fraction and water condensate in each atmospheric column. They result from complex non linear interactions between the various processes taken into account in the GCM. Having the "right clouds" in a GCM means simulating these profiles accurately. Additional assumptions concerning clouds can or must be made to compute RT from these profiles, such as their vertical overlap, horizontal inhomogeneity or size. These assumptions 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 overlap assumption for low clouds tends to underestimate total (vertically integrated) cloud cover, and consequently, TOA fluxes. Neglecting in-cloud optical-depth inhomogeneity systematically leads to overly reflective clouds. Assuming infinite cloud size means neglecting 3D radiative effects, which produces either too-bright or too-dim clouds depending on solar zenith angle. To remove these errors, radiative transfer models that are able to treat "cloud geometry" (overlap, heterogeneity and size) in a more sophisticated way than maximally-overlapped homogeneous slabs have been developed in the last decades (e.g., Shonk and Hogan (2008); Hogan and Shonk (2013); Hogan et al. (2019); Pincus, Barker, and Morcrette (2003)). The development of the ecRad code at ECMWF (Hogan & Bozzo, 2018), which implements these new radiative transfer models (solvers) in the same software, has remarkably facilitated the investigation of cloud-geometry effects in radiative transfer models.

A now standard way to improve the representation of atmospheric physics in GCMs is to isolate one column, working on physics parameterizations without interaction with large scale dynamics. In this approach, Large Eddy Simulations (LES, i. e. 3D simulations with resolution of a few tens of meters on domains of a few tens of km) of a cloud scene are used as a reference for evaluation of simulations run with the Single Column Model (SCM) version of a GCM. Radiation is still often poorly represented in LES, but reference fluxes associated with the 3D cloud scenes can now be computed using offline 3D radiative codes based for instance on Monte Carlo methods (Villefranque et al., 2019). In recent years, the SCM/LES approach has also been empowered by machine learning and automatic tuning procedures (Couvreux et al., 2021) based on the history matching approach proposed by Williamson et al. (2013). This approach performs global sensitivity experiments to separate parametric from 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); Villefranqu~~ Couvreux et al. (2021), Hourdin et al. (2021) and Villefranque et al. (2021), upon which the present work builds.

In Couvreux et al. (2021) and Hourdin et al. (2021), parameters of an SCM boundary-layer parameterization 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 on the 3D LES clouds. Here, we begin with an extension of the Villefranque et al. (2021) study, and we then present a new type of tuning experiments, specifically designed to study cloud–radiation compensations.

This work was originally motivated by 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). One particular question was to activate or not the `Spartacus` solver of ecRad, which is numerically more expensive than its 1D counterpart `Tripleclouds`, but allows for the first time to account for 3D radiative effects of clouds in the atmosphere. [FH : Les deux phrases ci-dessous sont fragiles. A discuter. Que veut on dire là ? Est-ce qu'on a vraiment besoin de ça ?] Previous studies had revealed that 3D effects changed global mean flux at TOA by about 1 W·m^{-2}, (**REF???**), but this was done without retuning the model. We wonder how calibration interacts with cloud radiative effects, and if retuning the model with a new radiation scheme might change previously reported conclusions. With this question in mind, we designed cheap, idealized, column experiments, to mimic what may occur when tuning cloud parameterizations in a climate model while ~~targetting~~ targeting TOA radiative fluxes observed by satellites (as is the common practice reported by Hourdin et al. (2017)). In these experiments, we focus on two kinds of compensating errors: those that occur within the RT scheme itself, and those that occur between cloud parameterizations and the RT scheme. Under the pretext of addressing our own specific questions, we also mean to show how History Matching can be used to guide model development, and in particular to evidence and investigate compensating errors.

The first part of the study follows that of Villefranque et al. (2021). History Matching is applied to the calibration of cloud-geometry parameters of an RT scheme. PPEs are produced by running the RT scheme on horizontally-averaged cloud profiles output from LES ("perfect cloud" experiments). The reference targets in the calibration exercise are 3D solar fluxes computed using Monte Carlo methods on the 3D LES cloud fields. In this approach, ecRad and MC computations thus "see" the same average cloud scenes. Three ecRad parameters relative to cloud geometry are varied in these experiments: overlap decorrelation length, degree of in-cloud inhomogeneity, and cloud size. In Villefranque et al. (2021), ecRad was run using the `Spartacus` solver. Ranges of parameters were identified that give a very good agreement with MC computations, including a good representation of the dependency of reflected radiation to solar zenith angle, which requires a representation of 3D effects. A more systematic investigation is done here, comparing tuning experiments with the `Spartacus` and `Tripleclouds` solvers.

The second part of the study investigates error compensations between radiative transfer (including the representation of cloud geometry) and cloud physics (which ~~provide~~ provides vertical profiles of temperature, cloud fraction and liquid water content). To that end, cloud-physics parameters of the SCM are tuned~~(~~, as in Hourdin et al. (2021)~~) while targetting~~ , but targeting reference radiative metrics computed on the LES rather than targeting directly the cloud fraction and thermodynamic profiles of the LES. That is, clouds are adjusted to get the right radiative fluxes, for a given version of the RT model. In these experiments, ecRad uses either `Tripleclouds` or `Spartacus`, with their free parameters set to values identified as "best" values in the "perfect cloud" ecRad tuning experiments. We find that when radiative fluxes at various solar zenith angles are simultaneously targetted in the tuning process, experiments using `Tripleclouds` produce wrong cloud profiles to compensate for errors in the RT scheme. Then, using more appropriate ecRad parameters for `Tripleclouds` (instead of best `Spartacus`), and targetting one SZA-averaged flux (instead of three SZA-dependent fluxes), we show that both clouds and radiative fluxes can be reasonably well simulated in the experiments using `Tripleclouds`.

The rest of the paper is organized as follows: Section 2 describes the data, tools and models that are used in the study; Section 3 investigates compensations that are internal to the radiation scheme; Section 4 investigates compensations that arise between clouds and radiation schemes; Section 5 presents the conclusions of the study, and points out that, beyond these specific results, this work also shows how History Matching can be used to go deeper into the understanding of climate models.

2 Data, models and tools

To implement tuning experiments, three ingredients are needed: reference data (simulated or observed) that will serve as targets for the tuning; a model, which includes free parameters that will be adjusted in the tuning process; a tuning tool (methodology and software) that will make part of the protocol objective and automatic.

In this study, two different kinds of tuning experiments are implemented. Both use the same tuning tool, described in Section 2.4, and the same reference data: radiative fluxes computed using a 3D Monte Carlo code, run on 3D cloud fields output from LES. These are described in Section 2.1. The difference between the two types of experiments is the model that is tuned. In the first type of experiments, the models are the `Tripleclouds` and `Spartacus` solvers of the radiative transfer scheme ecRad, run on horizontally-averaged LES cloud profiles. In the second type of experiments, the model is the SCM version of LMDZ. They are respectively described in Section 2.2 and Section 2.3. All experiments are made using idealized cumulus cases, following Villefranque et al. (2019).

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 × 12.8 × 4 km^3 domain. Large scale dynamics, ra-

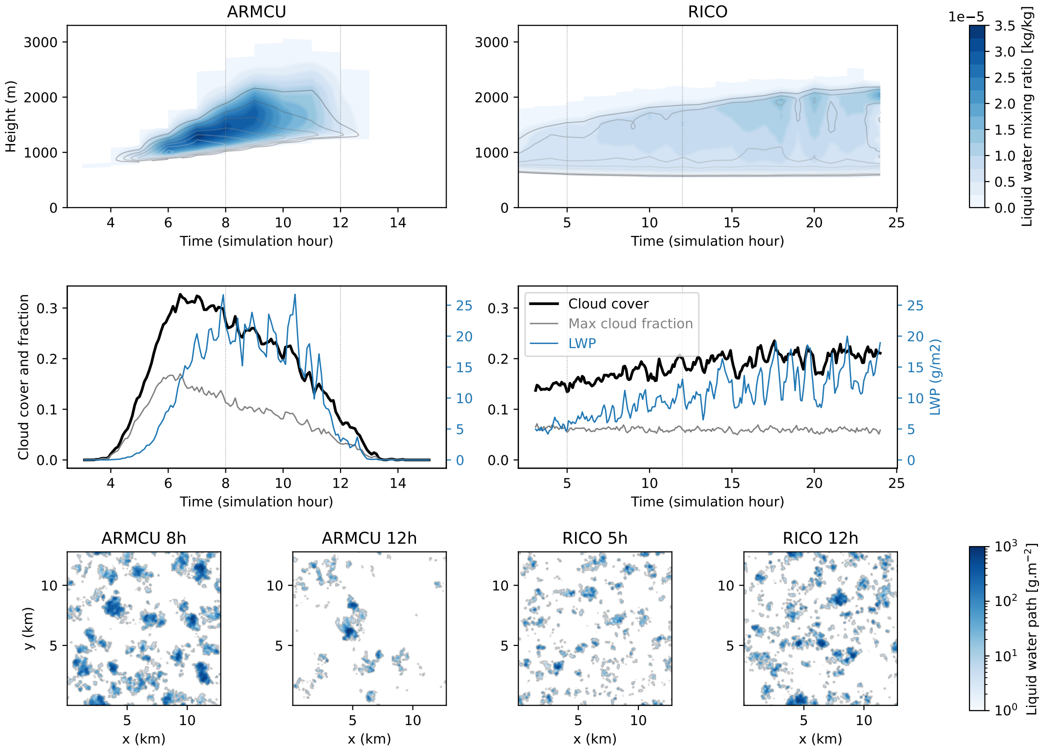


Figure 1. Illustrations from the LES ~~data for~~ of the ARMCU (left) and RICO (right) test cases used in this study. ~~Contours~~ The first row of figures shows the time evolution of the vertical profiles of the mean (averaged horizontally over the LES domain) cloud fraction (contour lines [FH : Il faut donner les valeurs]) and cloud water mixing ratio ~~vertical profiles as a function of~~ (blue shading). The second row shows the time ~~;~~ evolution of total cloud cover, maximum cloud fraction and domain-mean liquid water path~~;~~. The third row shows liquid water path maps at the two hours ~~of interest~~ retained for tuning (ARMCU 8th and 12th hours, RICO 5th and 12th hours).

diative 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).

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 radi-

Table 1. Quelques propriétés des champs nuageux LES... **A METTRE AU PROPRE**

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

ation 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~~ The solar constant is set to 1368 W.m^{-2} and the surface albedo to 0.08. For each 3D cloud field, additional MC calculations are made under the Independent Columns Assumption (ICA), by computing 3D radiation independently in distinct columns of the 3D LES cloud field and taking the average. This assumes independent columns, as in the MCICA model of Pincus et al. (2003), which removes 3D radiative effects from the calculation. Differences between 3D and ICA MC fluxes yield estimates of 3D radiative effects.

2.2 Radiation parameterization

The radiative models under investigation in this study are the `Tripleclouds` and `Spartacus` solvers implemented in 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 and are hence also excluded from potential causes of differences between parameterized and reference fluxes. *[Pour plus tard : Pas pour être discuté ici mais on oublie peut-être un peu vite un le critère retenu pour décider qu'on est dans le nuage ou pas dans la LES et le lien avec une fraction nuageuse effective pour le rayonnement ...]* In SCM experiments, liquid water content and cloud fraction profiles input to ecRad are taken from SCM simulation outputs.

Cloud optics are interpolated from a Mie look-up table provided with ecRad, similar to, but slightly different from, the one used in the Monte Carlo simulations. The difference between optical properties taken from the two tables are typically less than a change of 1 μm in the effective radius of clouds (not shown **ou en supplemental ? J'ai la figure qqpart**)[FH : Not shown suffit. Les reviewer verront. En tous cas pas un SI exprès pour ça].

The RT models at the heart of `Tripleclouds` and `Spartacus` are modified versions of the two-stream model (Meador & Weaver, 1980), which directly integrate the effects of cloud geometry on radiation transport through assumptions on vertical overlap, horizontal heterogeneity and~~cloud size (~~, in `Spartacus` only~~)~~, cloud size.

In our configurations of both `Tripleclouds` and `Spartacus`, vertical overlap is represented using the exponential-random model parameterized by its decorrelation length ℓ (Hogan & Illingworth, 2000), and a two-region cloud representation (the `Tripleclouds` model of Shonk and 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` solver, no 3D effects are taken into account, whereas in `Spartacus` (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 remove around 10 W·m^{-2} from TOA solar upward (reflected) flux when the sun is high (solar zenith angle close to zero), and account for about 10 W·m^{-2} at low suns, compared to 1D radiation (see e.g. Villefranque et al. (2019)).

2.3 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 ?**[FH : Pas là non. On le dira là on on en discute, si on en discute. Là on présente les outils.]

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).

As for radiation, ecRad was recently implemented in LMDZ and will be part of the ~~next version~~ forthcoming versions of the GCM. However, in this study, it is the offline version of ecRad that is run on the SCM output profiles. It calculates the radiative fluxes associated with the SCM cloud profiles and ecRad cloud-geometry parameters. The offline version was used to ~~be able~~ make more robust comparison, focused on cloud radiative effects. In particular it allows to use the same clear-sky profiles as in the MC simulation~~, and also because~~. This was done in practice by replacing the clear-sky profiles in the SCM outputs by the profiles of the LES averaged horizontally. Note that in the two cumulus cases that are used as constraints, surface fluxes and radiative cooling are prescribed as forcings: radiation is not interactive in the simulations ~~, thus the radiation scheme of LMDZ is never called during the SCM simulation itself~~and thus does not affect the clouds (consistently in the LES and SCM).

2.4 The High-Tune:Explorer tuning tool

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. To that end, 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, emulators based on Gaussian Processes are built for each metric. These 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 [1]. Prediction uncertainty is added to tolerance-to-error in order to avoid ruling out parameter vectors that might be far from the target because of bad emulators but actually acceptable configurations of the model. To this end, implausibility is defined for any parameter vector $\boldsymbol{\lambda}$ as

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

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 chosen 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 metrics. In that case, the final NROY space is exactly the subspace of free parameters that matches the user-defined constraints and emulators can be considered reliable models for the metrics (inside the final NROY space).

[1] Training an emulator means characterizing the law of the Gaussian Process $G \sim \mathcal{GP}(m, k(.,.))$, that is, specifying its expectation function m and variance kernel $k(.,.)$. This is done by fitting standard analytic functions to the simulated data for metric i, $\{f_i^j\}_{j\in 1,..10N}$ obtained from model runs at $10N$ points $\{\boldsymbol{\lambda}^j\}_{j\in 1,..10N}$ sampled in parameter space. Then, for any new $\boldsymbol{\lambda}^*$ in parameter space, the i-th emulator provides a prediction of the i-th metric as the expectation μ_i and standard deviation σ_i of the conditional random variable $G_i^* = G_i(\boldsymbol{\lambda}^*) | \left\{ G_i(\boldsymbol{\lambda}^j) = f_i^j, \forall j \in [1, 10N] \right\}$; and $G_i^* \sim \mathcal{N}(\mu_i, \sigma_i)$.

In practice, emulator uncertainties rarely fall one order of magnitude under tolerances to error for all metrics and hence the experiment rarely strictly converges. We therefore find it useful to define weak convergence: The experiment is considered to have weakly converged when adding new iterations does no longer significantly reduce NROY space. In that case, the final NROY space is larger than the sought parameter subspace.

Importantly, during the tuning experiment, many simulations will have been run, consituting a Perturbed-Parameter Ensemble (PPE). For each metric i and each simulation j, its distance to reference value can be calculated as $\frac{|r_i - f_i^j|}{T_i}$, where $f_i^j = f_i(\boldsymbol{\lambda}^j)$ is the actual model output for metric i and parameter vector $\boldsymbol{\lambda}^j$. The score of a simulation j is then defined as its worst-metric score:

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

where emulator predictions and uncertainties no longer appear, contrary to the implausibility defined in Equation (1). Keeping only simulations whose scores are under a given threshold provides, independently from the emulators convergence status, constrained PPEs that can be studied to learn about the quality of the model. Of course, if the emulators convergence is bad, the NROY space might remain too large and model configurations sample might not be dense enough, so low-probability, acceptable model behaviours might be absent from the PPEs.

3 Internal compensations between parts of the radiative transfer model

First, we examine ecRad PPEs in a “perfect cloud” framework. “Perfect clouds” means that ecRad is run on horizontally-averaged LES cloud profiles. The ~~same LES cloud fields~~ 3D cloud fields from the same LES were used as inputs for the reference Monte Carlo radiative simulations. Errors in ecRad fluxes can thus not be attributed to wrong cloud fractions or mean liquid contents, ~~which would be output~~ unlike in cases where cloud fraction and liquid content are taken from boundary-layer parameterizations ~~in a “modelcloud” experiment~~of a climate model, as discussed in the following section. We use these PPEs to investigate compensating effects between various aspects of cloud geometry. In `Tripleclouds` and `Spartacus`, the effect of cloud geometry on radiation ~~(i.e. the fact that clouds are not horizontally infinite homogeneous slabs)~~ [FH : On ne peut pas dire ”(i.e. the fact that clouds are not horizontally infinite homogeneous slabs)”, si ? Les nuages ne sont pas infiniement homogènes dans les codes précédents. On avait une partie ciel clair et une partie nuageuse, elle étant effectivement plus ou moins homogènes. Mais il me semble que cette phrase est trompeuse. Soit on vire soit on détaille, non ? A court terme je crois qu’on peut virer.] is taken into account by adding state variables and source terms to the two-stream model of Meador and Weaver (1980). Cloud geometry includes vertical overlap, heterogeneity of in-cloud water, and cloud size. Free parameters associated with these new terms are overlap decorrelation length ℓ, fractional standard deviation FSD, and cloud effective scale C_s. In the present "perfect cloud" experiment, they are the only parameters that are calibrated in order to obtain fluxes as close as possible as those computed using Monte Carlo methods on the 3D cloud fields. ~~In other words~~

Here, we look for the ~~parameters~~ parameter values that provide an accurate description of ARMCU and RICO cumulus cloud geometry *from a radiative transfer model perspective.* ~~Note that this~~ This approach is fundamentally different from trying to ~~use parameter values derived~~ extract estimates of these parameter values directly from the simulated 3D ~~fields, or from observations, to constrain ecRad~~cloud fields. Here, we accept that cloud geometry parameters ~~make sense~~ are *effective parameters* that make sense *in the context of ~~the~~ a particular model* (here the RT solver), rather than in the absolute. ~~Given that, the~~

[FH : Pas sûr qu'il faille garder la fin de cette sous section. Ou la garder comme ça.] Given that one question driving this study was: can we (should we) choose the same parameter values when using `Tripleclouds` or `Spartacus`? Which led us to asking: how do the various submodels, which describe radiative effects of specific aspects of cloud geometry, interfer with each other? Can we think of each effect separately, or do we need to consider the coupled problem? In the coupled problem, do cloud-geometry parameters have the same role as in separate submodels?

3.1 Where we find that `Tripleclouds` and `Spartacus` should not use the same cloud-geometry parameter values

First, we look at two experiments: one where `Tripleclouds`~~is run on LES cloud profiles and the target is~~ targets Monte Carlo 1D (ICA) fluxes ~~at three solar zenith angles;~~ and one where `Spartacus`~~is run on the same LES cloud profiles and the target is Monte Carlo~~ targets Monte Carlo 3D ~~fluxes at (the same)~~ fluxes. By targeting 1D MC fluxes for `Tripleclouds`, we avoid asking the model to compensate for a structural error that we know is present in the solver, namely the failure to take into account 3D radiative effects Targeting 1D reference computation is a classical way of evaluating two-stream RT models that don not take 3D effects into account (`https://doi.org/10.1175/JAS-D-15-0033.1` `https://doi.org/10.1002/qj.49712555810`

The fluxes targeted as tuning metrics are taken at three solar zenith angles. Simulation ensembles are produced varying the three radiative parameters (ℓ, FSD and C_s, only the first two being active in Tripleclouds) under the constraint that targets are reached within ~~3~~ three times the tolerance-to-error, which is set to 3.5 W·m^{-2}. ~~Here, we expect to be able to reproduce Monte Carlo 1D fluxes if values of overlap decorrelation length and fractional standard deviation of liquid water content distribution are well chosen. Similarly, we expect to be able to reproduce Monte Carlo 3D fluxes if its three parameters are well chosen.~~ [FH : Penser à mettre à jour le cuttof, le nombre de vagues ...]

For each of these two experiments, ten waves (iterations) of history matching are produced, each consisting in 101 simulations. In total, for each experiment, 1010 configurations of the solver (either Tripleclouds or Spartacus depending on the considered experiment) are sampled and run. Among these configurations, the 300 simulations with the lowest scores are selected. Figure 2 shows the location of these 300 best configurations per experiment in parameter space. The first three lines and columns of the subplot matrix are relative to parameter values, while the last line and column is relative to the score $S(\boldsymbol{\lambda})$ of each configuration, as per Equation (2). ~~Note that~~ Since the RAD_CS parameter is not used in Tripleclouds~~so the location of~~, the red points corresponding to Tripleclouds configurations ~~in~~ spread randomly along the RAD_CS ~~space is uniform: unlike for the two other parameters, simulations are not concentrated in any “good” region of this parameter space~~axis.

First, we note that all configurations in the Tripleclouds experiment have scores below 1, while all those of Spartacus have scores above 1. ~~However, this~~ Of course it does not mean that Tripleclouds is better than Spartacus~~, since their scores are computed relatively to different references. It thus~~ in an absolute sense, since the scores for Triplecloudsare computed against referenced intentionally biased by not accounting for 3D effects. It just means that Tripleclouds is closer to reference 1D fluxes than Spartacus is close to reference 3D fluxes. It might not be too surprising for a modified two-stream model to perform better at simulating 1D radiation than 3D, especially when considering the many years spent on the 1D problem (i.e., how to account for vertical cloud structure and horizontal heterogeneity in vertical light propagation) compared to the relatively short time dedicated to modelling horizontal transport within a the two-stream framework.

The next thing one can notice is that the 10 best configurations of each experiment are located in distinct regions of the $\ell \times FSD$ space. The best Tripleclouds simulations, high-

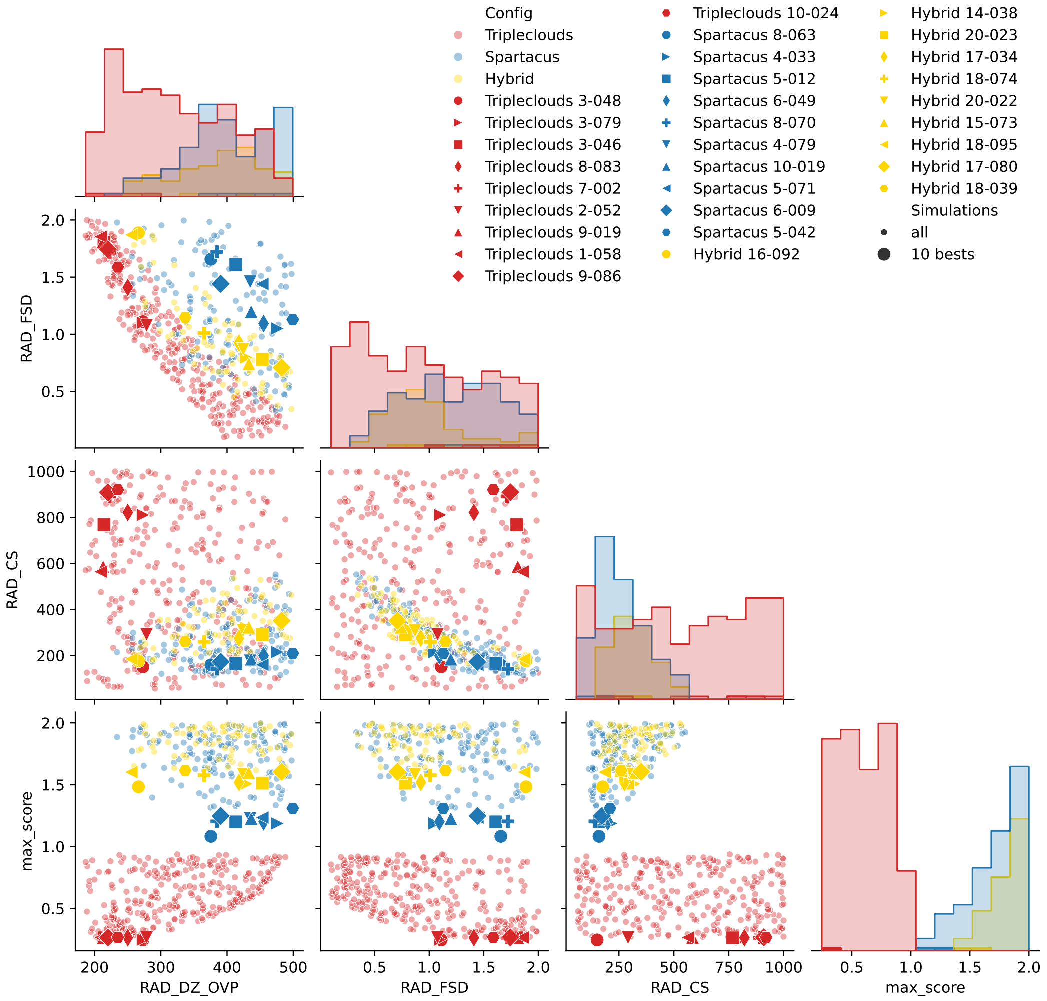


Figure 2. Parameters of 300 best simulations for three "perfect-cloud" ecRad tuning experiments. RAD_DZ_OVP is the vertical overlap decorrelation length ℓ, RAD_FSD is the relative in-cloud inhomogeneity FSD, RAD_CS is the cloud size C_s~~,~~. The max_score is $S(\boldsymbol{\lambda})$. Each column corresponds to a parameter. Top subplot of each column shows histograms of this parameter value for the 300 best configurations for each experiment. Then, each line corresponds to a second parameter, and subplots show the location of the configurations in the 2D space of the column × line parameters.

lighted by large red symbols in Figure 2, have small ℓ and large FSD, while the best `Spartacus` simulations (large blue symbols) correspond to larger ℓ values, for approximately the same range of FSD values. The best `Tripleclouds` simulations exhibit a strong negative correlation between ℓ and FSD.

In 1D radiation, this is easy enough to understand: with a smaller decorrelation length~~in the vertical overlap model means that~~, clouds are more randomly (~~and~~ or less maximally) overlapped, ~~which yields larger~~ leading to a larger cloud cover and smaller cloud optical thickness~~, given the same input cloud profile. With smaller ℓ, a larger part of the domain is occupied by clouds, but they are less optically thick. This, overall, leads to a more reflective scene: the~~ in regions covered by clouds. The radiative effect of increasing cloud cover dominates that of decreasing cloud optical thickness, because reflectivity is an exponential function of optical thickness τ,

$$R \sim 1 - \exp(-\tau).$$

~~Replacing part of the clear-sky region of the domain~~ The optical thickness at a given location on the horizontal is the sum of the optical thicknesses if the cloudy layers above. Distributing the individual optical thicknesses of the cloud layers horizontally on a wider area (large cloud cover with clouds of moderate optical thickness~~increases domain reflectivity more than the decrease due to smaller cloud optical thickness in the cloudy part of the domain.~~) leads to a larger reflectivity than when cloudy layers overlap, because of the saturation of the exponential term.

The same reasoning can be applied to in-cloud water horizontal heterogeneity: because of the exponential function, horizontally homogeneous clouds are more reflective than heterogeneous clouds that have the same horizontally-averaged optical depth. Thus, decreasing ℓ leads to more reflective clouds, and increasing FSD leads to less reflective clouds. ~~In , choosing a value for ℓ provides a strong constraint on~~ This explains why the tuning algorithm selects smaller values of the FSD ~~value that must be set to have~~ to compensate for larger values of ℓ in order to simulate the right fluxes.

In `Spartacus`, the relationship between ℓ and FSD also seems to exist, but the 10 best configurations of the `Spartacus` experiment show more dispersion than the 10 best `Tripleclouds` ones. Our first guess was that horizontal transport, driven by the third parameter C_s, modulates reflectivity in `Spartacus` while it has no effect in `Tripleclouds`, which means that, in `Spartacus`, a given value of ℓ can be associated with various values of FSD and still produce the same fluxes~~, for instance if~~ . For instance a larger FSD ~~is~~ could be compensated by more intense 3D effects.

However, C_s does not vary much between these 10 best configurations so this explanation might not be sufficient. Moreover, 3D effects tend to decrease reflectivity when the sun is close to zenith and increase it otherwise; so we need to consider our three metrics, which correspond to three different solar zenith angles, in order to understand if significantly larger FSD can indeed be compensated by slightly smaller clouds.

All things considered, the internal dispersion of the best Spartacus configurations might be a detail compared to the fact that best Spartacus configurations can be found in the "large ℓ, large FSD" corner of the parameter space, where none of the best Tripleclouds simulations can be found. This is an important result, as it means that finding the best Spartacus configuration and removing 3D effects (e.g., using Tripleclouds with the same (ℓ, FSD) values, or setting C_s to infinity in Spartacus), will produce wrong fluxes compared to the 1D Monte Carlo reference. Indeed, none of these "large ℓ, large FSD" configurations leads to 1D fluxes closer to the reference than $2{\times}3.5 = 7$ W·m^{-2} (~~not shown~~ ~~**ou en supplemental, la même figure mais avec toutes les simu**~~ ~~It is interesting to realize that the~~ figure not shown).

[FH : Le paragraphe du dessous est une proposition de paragraphe alternatif à "It is interesting to realize that the Spartacus solver does not just "add" 3D effects to Tripleclouds fluxes. Effects of the three parameters are intertwined in the model: Adding 3D effects leads to selecting new $\ell{\times}FSD$ values to get the right fluxes. One interpretation is that, as suggested earlier, models' free parameters are ever only *effective parameters*, and hence that there is no objective reason to set them to (often loosely equivalent) observed values. Alternatively, one might consider that a model such as Spartacus *should*, once set to its best configuration, simulate accurate 1D fluxes when 3D effects are removed; In which case, Figure 2 evidences a default in the parameterization of 3D effects in Spartacus."]

The Spartacus~~solver does not just "add"~~ solver was built upon the Tripleclouds sover by adding 3D effects~~to~~. So one could wish Spartacus to well simulate both 3D radiation and 1D radiation when 3D effects are removed. In this view, the results shwon here suggest that compensation errors are at work betwen the Tripleclouds~~fluxes. Effects of the three parameters are intertwined in the model: Adding~~ solver and 3D effects ~~leads to selecting new $\ell \times FSD$ values to get the right fluxes. One interpretation is that, as suggested earlier, models'~~ in Spartacus. However, when running Spartacus, there is no real distinction between non 3D and 3D effects. It is the full radiation which is computed and affected by the parameterization of 3D effects. An other, less ambitious but probably more relevant view, is to accept that the free parameters are ~~ever~~

only *effective parameters* ~~,~~ and hence that there is no objective reason to set them to (often loosely equivalent) observed values. ~~Alternatively, one might consider that a model such as *should*, once set to its best configuration, simulate accurate 1D fluxes when 3D effects are removed; In which case, Figure 2 evidences a default in the parameterization of 3D effects in .~~

To illustrate this point further, let us look at cloud covers calculated by the exponential-random overlap model, for the ten decorrelation length values considered as "best" values in `Tripleclouds` and `Spartacus` experiments. The upper panel of Figure 3 presents a comparison of true cloud covers diagnosed in the 3D fields of ARMCU and RICO LES simulations, with those calculated by the overlap scheme. Note that covers calculated in ecRad, when using `Tripleclouds` or `Spartacus` solvers, do not enter directly in the flux equations. They are only computed as a diagnosis. Rather, the overlap model is used locally, at each interface between layers, to distribute fluxes from cloudy and clear regions of a given layer, to cloudy and clear regions of the next layer. These plots show that covers corresponding to the best `Tripleclouds` simulations are very close to those of the LES. It is tempting to conclude that `Tripleclouds` fluxes are right *for the right reasons*, i.e., the right cloud geometry. In comparison, cloud covers corresponding to the best `Spartacus` simulations are systematically smaller than those of the LES (in agreement with the fact that ℓ values are larger). Yet, `Spartacus` fluxes are globally accurate. Is it for wrong reasons, i.e. the wrong cloud geometry? Does this fall under the "compensating error" category? Do we want to eliminate this behaviour or shall we accept, and even expect, that "cloud cover" (most often, a vertically integrated quantity), does not have the same meaning, and does not play the same role, in 1D vs. 3D radiative transfer models? ~~This is a modelers' debate which the present PPE experiments can help support, but in which we will not provide further insights here~~These are open question that the modelers should have in mind when using, evaluating and tuning their models.

What fundamentally shows the improvement of the `Spartacus` solver compared to `Tripleclouds` is its ability to well capture the angular dependency of the "true" reflected radiation (3D MC reference)as clearly seen in the second line of panels in Figure 3. The `Tripleclouds` is able to well simulate 1D radiative reference computations but it will always underestimate the reflected radiation at high solar angle compared to zenith.

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

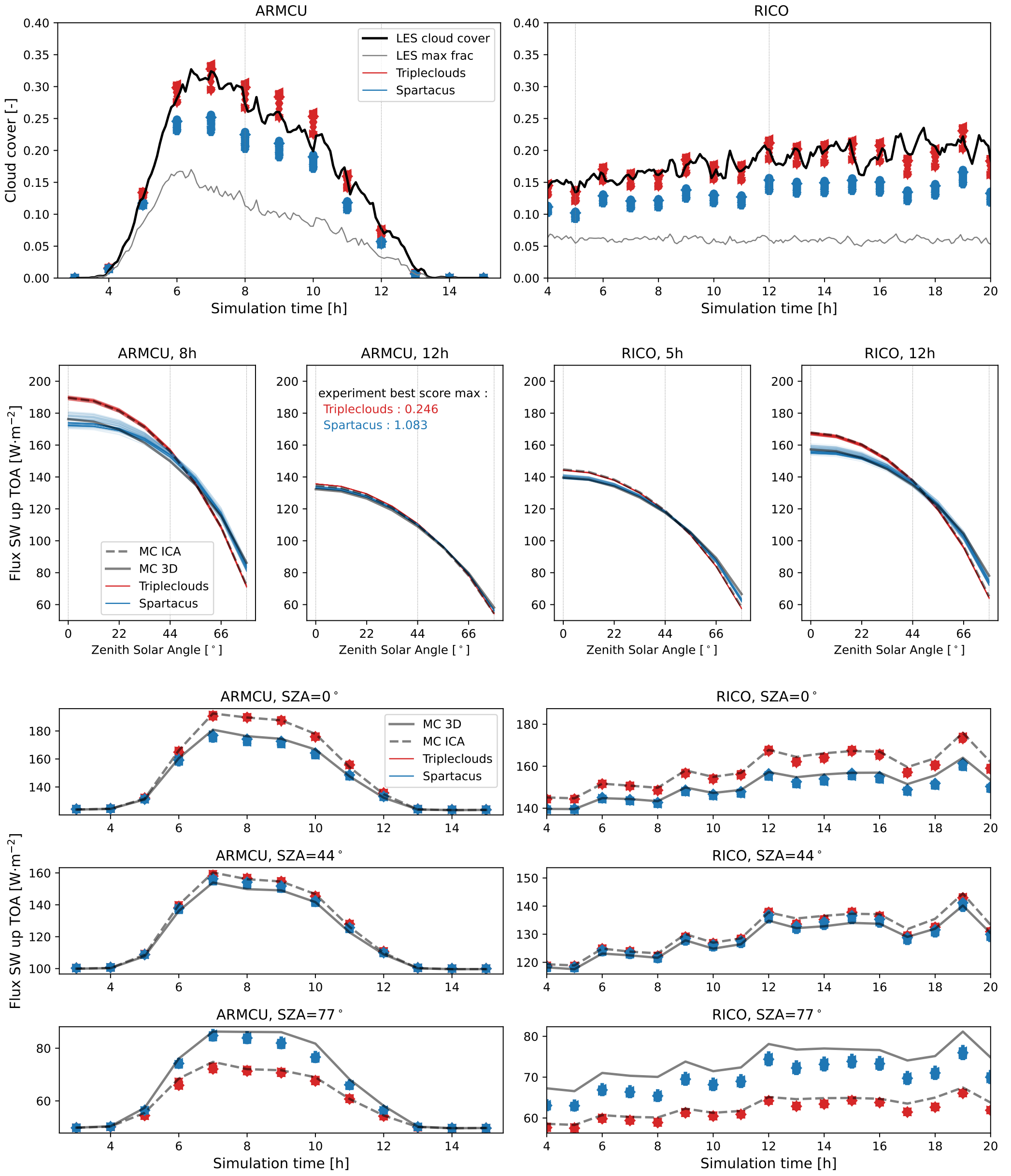


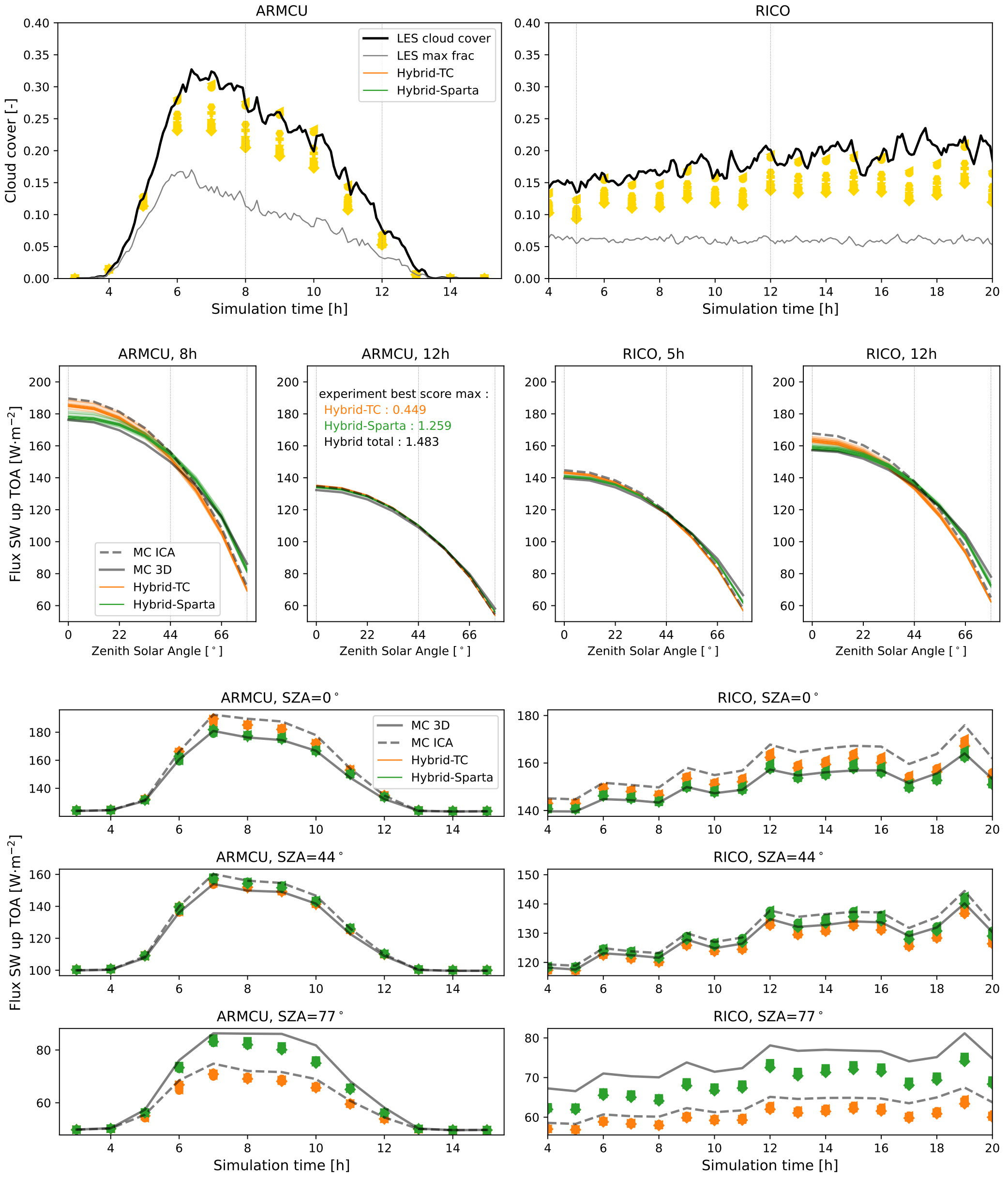
Figure 3. Cloud and radiation variables for 10 best simulations of perfect-cloud experiments (`Tripleclouds` in red, `Spartacus` in blue), compared to reference. Left column: ARMCU. Right colum: RICO. Top line: cloud cover (black) and maximum cloud fraction (gray) in LES fields, and cloud cover computed 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). Second line: upward TOA fluxes as a function of solar zenith angle, for the 4 cloud scenes used as constraints in the tuning experiments: ARMCu 8th and 12th hours, RICO 5th and 12th hours. Monte Carlo references are in black (full lines for 3D fluxes, dashed lines for 1D fluxes), `Tripleclouds` fluxes are in red and `Spartacus` fluxes are in blue. Lines 3-5: upward TOA fluxes, calculated at every hour of the LES simulations, each time for three solar angles (line 3 for SZA=0°, line 4 for SZA=44°, line 5 for SZA=77°). Monte Carlo fluxes are in gray (full line for 3D fluxes, dashed line for 1D fluxes) and ecRad fluxes correspond either to `Tripleclouds` fluxes (~~orange~~ red lines) or `Spartacus` fluxes (~~green~~ blue lines) ~~, set with the same parameters (those of the ten best~~ configurations~~from the Hybrid experiment)~~.

3.2 Where we look for a compromise between Tripleclouds and Spartacus preferred regions of cloud-geometry parameter space

~~Rather, we focus~~ We rather focus here on finding a configuration that is acceptable for both Tripleclouds and Spartacus models. Because best Spartacus configurations are bad choices for Tripleclouds, it implies that Spartacus simulations, set with a compromise configuration, will be less accurate than the previously examined best Spartacus simulations. ~~We thus~~ To that end, we design a Hybrid ~~experiment, by setting constraints on both fluxes (still aiming at~~ tuning experiment, in which we target both 1D ~~fluxes) and fluxes (still aiming at~~ MC fluxes with the Triplecloudssolver and 3D ~~fluxes),~~ MC fluxes with Spartacus at the same time. This means that for every sampled ~~set of (ℓ, FSD, C_s) parameter~~ parameter vector (ℓ, FSD, C_s), we run both Tripleclouds and Spartacus on the four clouds scenes (two hours from ARMCU et two hours from RICO) and three solar zenith angles (0, 44 and 77 degrees). The score associated with each configuration is the worst (metric-wise) score over all the Tripleclouds and Spartacus metrics.

The 300 best configurations are presented in Figure 2 (yellow points). ~~Looking at the scores , we see that they~~ The scores obtained are generally worse than for previous experiments (mostly, above 1.5). The location of the best configurations in the (ℓ, FSD) space best illustrates the compromise that was found: for a given choice of FSD, the 10 best simulations of the Hybrid experiment correspond to ℓ values between best Tripleclouds and best Spartacus values. Note that relatively small values of FSD of around 0.75, associated with relatively large C_S of around 400 m, yield scores that are among the ten smallest ones, which was not the case in any of the previous experiments.

The very best configuration in the Hybrid experiment corresponds to large FSD and small ℓ values, close to the best Tripleclouds simulations. Accordingly, in the upper panel of Figure 4, this best Hybrid configuration corresponds to the points that are closest to the 3D LES cloud cover. Comparing fluxes in Figure 4 with those of Figure 3, we can see that Tripleclouds fluxes, which were previously extremely accurate, are now less so: most of the ten best configurations from the Hybrid experiment lead to smaller cloud cover than in the 3D LES field, which negatively impacts Tripleclouds performances. As for Spartacus, it seems to perform better or equally at SZA=0, and to be slightly worse than before at large solar zenith angles. If we look at the "large ℓ, small FSD" configurations of the Hybrid experiment, we note that, for a given ℓ value, FSD is smaller than in the pure Spartacus experiment, which should lead to more



[FH : Il me semble qu'il faut remplacer la légende par une légende plus courte disant que c'est comme la figure précédente sauf que ...]

Figure 4. Cloud and radiation variables for 10 best simulations of perfect-cloud "Hybrid" experiment, compared to reference. Left column: ARMCU. Right colum: RICO. Top line: cloud cover (black) and maximum cloud fraction (gray) in LES fields, and cloud cover computed by ecRad (yellow 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). Second line: upward TOA fluxes 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. ecRad configurations are the ten best of the Hybrid experiment, run either with `Tripleclouds` (orange lines) or `Spartacus` (green lines). Lines 3-5: upward TOA fluxes, calculated at every hour of the LES simulations, each time for three solar angles (line 3 for SZA=0°, line 4 for SZA=44°, line 5 for SZA=77°). Monte Carlo fluxes are in gray (full line for 3D fluxes, dashed line for 1D fluxes) and ecRad fluxes correspond either to `Tripleclouds` fluxes (orange lines) or

reflective clouds than before, if nothing else changed. However, these configurations are also associated with larger C_s values than the ten best of the pure `Spartacus` experiment, which means less intense 3D effects, which means more reflective at small SZA and less reflective at large SZA. The fact that smaller C_s values would not be chosen for these "large ℓ, small FSD" configurations, despite the fact that it would systematically improve fluxes at SZA=77, seems to indicate that fluxes at SZA=0 would be too large (more intense 3D effects would decrease reflectivity at SZA=0) and would become the limiting metrics.

Finally, we find that the very best configuration of the Hybrid experiment, that is, simulation number 16-092, is a good candidate for being our "unique" ecRad configuration in LMDZ, as far as cumulus cloud geometry is concerned.

4 Compensations between radiative transfer and cloud models

~~In this second part~~Here, we examine PPEs run with the single column version of ~~the LMDZ model~~ LMDZ for the same two cumulus cases, under radiative constraints. Using both `Tripleclouds` and `Spartacus` solvers, with configurations from the previous section, and testing two sets of radiative metrics, we investigate the conditions that lead to cloud–radiation compensating errors.

4.1 ~~Design of experiements~~Experimental design

We design experiments to mimic the tuning protocol often followed in GCMs when targeting top-of-the-atmosphere (TOA) radiative metrics from satellite observations. In our idealized 1D version of the protocol, simulations with both `Spartacus` and `Tripleclouds` target the same MC 3D radiative computations ("true" radiative fluxes). Tuning is performed using either `Tripleclouds` or `Spartacus`, in order to investigate in particular how convection and cloud physics could compensate for the structural error consisting in not accounting for 3D radiative transfer in `Tripleclouds`.

dire que les target sont plus les mêmes qu'avant car on vise des flux moyennés sur 1h !!

Thirteen parameters that control boundary-layer and cloud parameterizations in LMDZ are varied as in Hourdin et al. (2021) and Hourdin et al. (2023) (see details in Table S1 of Supplemental Information).

[FH : C'est pas préférable de commncer par dire comment on a ajusté pour des PPEs puis dire qu'on a gardé la même valeur pour les autres ? Là ca semble vraiment dire : on aurait du faire autrement mais on a eu la fleme.]

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}. In theory, it could be a way to quantify the structural error of the model. In practice, it is a heavy process and the resulting numbers still remains difficult to interpret. An alternative explored here was to do this only for one experiment, and once the final tolerance to error was found, to use the same value for other experiments. If these experiments go through their 40 iterations with this tolerance to error and yield approximately the same final size of NROY space, we change nothing and analyse the PPE as it is. If an experiment leads to an empty NROY after a few iterations, it is because the tolerance to error was too small ; looking at the best scores found in these first iterations can help resize the tolerance to error to a larger value, with which the experiment is run again from the beginning. Similarly, an experiment with a final NROY space of the same size as the initial hypercube means that the tolerance to error was too large. Again, looking at the best scores can help resize it to a smaller value. The tolerance to error is hence less well adjusted to each experiment than if the trial and error process had been repeated systematically, but we found it was sufficient to produce PPEs worth analysing.

4.2 Where we suspect that overestimated cloud fraction compensates lack of 3D effects

We start by comparing experiments using either `Spartacus` or `Tripleclouds`, with ecRad cloud-geometry parameters set to the values of the best simulation in the tuning of `Spartacus`, i.e. configuration 8-063 in Figure 2. The idea in this experiment is to use the best possible ver-

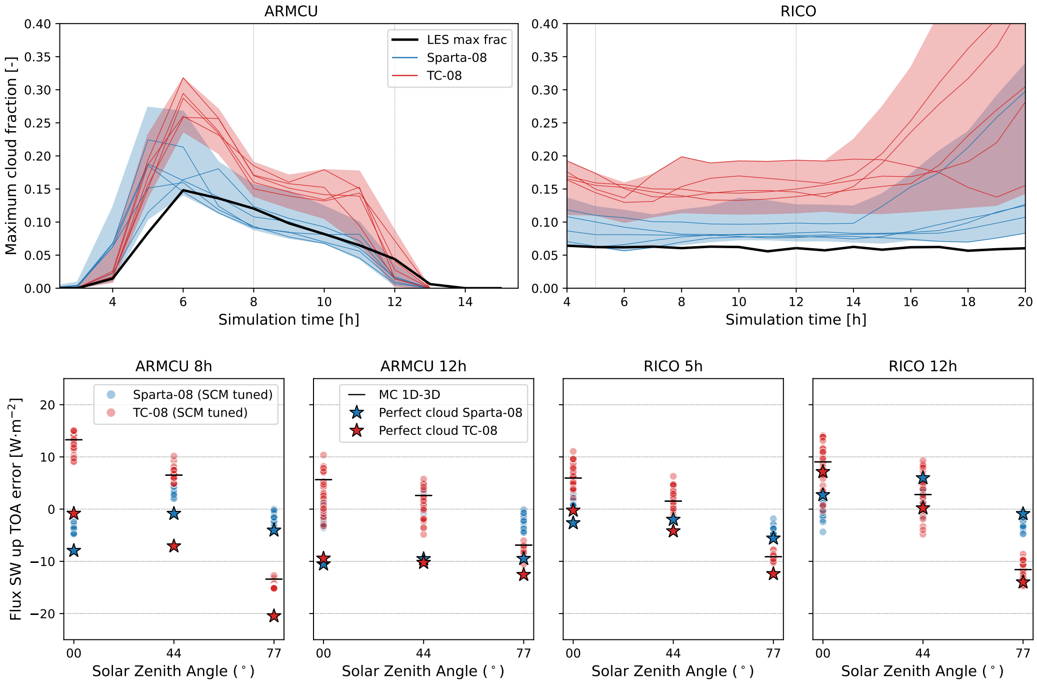


Figure 5. [FH : Verifier les légendes:] Clouds and radiation for the 30 best simulations for two experiments calibrating LMDZ parameters using radiative constraints: `Tripleclouds` with config best spartacus targetting 3 angle fluxes in red, spartacus with config best hybrid targetting 3 angle fluxes in blue. Left: ARMCU; Right:RICO. First row: maximum cloud fractions (shadings represent the 30 best enveloppe, lines show the 5 best); Second row: upwelling TOA fluxes as a function of solar zenith angle (thin lines are the 30 best simulations, thick lines are the 5 best simulations).

sion of `Spartacus` as a perfect model, and to introduce a structural error by removing 3D effects, using `Tripleclouds` instead of `Spartacus`. Fluxes at three angles are targeted as before.

The final tolerance to error found by trial and error for `Spartacus` was 3.5 W·m^{-2}, and it was resized to 9 W·m^{-2} for `Tripleclouds`. For the analysis, LMDZ+ecRad simulations of each RT configuration are ranked according to their score as defined by Equation (2). The product of a simulation's score 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 model: 3.6 W.m^{-2} and 14.4 W.m^{-2} for `Spartacus` and `Tripleclouds` experiments respectively. That is, given these radiative models, the SCM cannot produce clouds that result in flux errors smaller than those numbers.

`Spartacus` fluxes are as accurate as when `Spartacus` was run on reference cloud profiles. ~~This means that replacing LES cloud profiles by tuned LMDZ outputs has only a small effect on radiation, which suggests 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.~~ [FH : J'enleverrais : "This means that replacing LES cloud profiles by tuned LMDZ outputs has only a small effect on radiation, which suggests 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." Ca ne veut pas dire grand chose si ?] On the other hand, flux errors in the `Tripleclouds` experiment are about four times larger than those of `Spartacus`. This means that even with the possibility of compensating radiation error with wrong clouds, `Tripleclouds` could not produce fluxes as accurate as `Spartacus` in these experiments. Note that `Tripleclouds` scores are much larger than in perfect cloud experiments but this is partly because in previous experiments it was compared to 1D MC fluxes, whereas it is now compared to 3D MC fluxes.

Features of the 30 best simulations of each tuning experiment are analyzed in Figure 5. The top panel presents temporal evolutions of layer-wise maximum cloud fraction for ~~the two cases~~ ARMCU and RICO test cases. The bottom panel shows, for cloud scenes when radiative constraints were applied (8th and 12th hour of ARMCU, 5th and 12th hour of RICO), TOA upwelling fluxes as a function of solar zenith angles (only three of these angles serve as constraints in the tuning exercise). In these bottom plots, we see that although `Tripleclouds` simulations were constrained to be as close as possible to 3D MC fluxes, they still closely follow the 1D MC curves. Fluxes computed in the `Spartacus` experiment however match the MC reference for almost every angles, including those that were not constrained. In the upper plots, we ~~can see that~~see that, in addition to better simulating solar radiation, the 30 best Spartacus simulations exhibit cloud fraction evolutions that are compatible with reference values for ARMCU, and close to the LES ones for RICO. Conversely, the 30 best Tripleclouds simulations persistently overestimate maximum cloud fractions for both cases.

We interpret this as a sign of compensating errors. ~~Indeed, we see in Figure ??**QUELLE FIG MONTRE CA** that~~ We see in the lower row of Figure 5 that the Triplecloudssimulation run on LES ~~cloud profiles led to larger errors in reflected fluxes than what we obtain with the tuning. Tuning clouds using constraints leads to selecting LMDZ configurations in which cloud fractions are overestimated. A possible explanation is that we saw in Figure 3 that for this configuration of ecRad,~~ fields with the parameters selected for this tuning experiment (red stars) systematically underestimate the reflected solar radiation at SZA=77°. The error is the largest (in absolute value) for the ~~overlap parameter leads to underestimating cloud cover (blue points). When is used, this ”lack ” of cover is internally compensated by~~ 8th hour of ARMCU where it reaches -21 W·m^{-2}. The tuning protocol, which targets to the maximum error normalized by the tolerance among the metrics chosen, systematically reduces the error at SZA=77° thanks to an increase of cloud fractions, compensating for the lack of 3D effects ~~and the resulting fluxes are accurate. However, when is used, this lack of cover cannot be internally compensated. To still have the right fluxes (because this is the constraint of the tuning experiment) , lack of cover is compensated by~~ (tilted shadows). This increase of cloud fractions is limited by the enhancement of the reflected radiation at SZA=0°. The tuning finally selects simulations for which the bias shows the same absolute value for the maximum underestimation (about -15W·m^{-2} for SZA=77○ at the 8th hour of ARMCU) and maximum overestimation (about 15W·m^{-2} at SZA=0W·m^{-2}, especially for the 8th hour of ARMCU and 12th hour of RICO).

The overestimated cloud fractions probably also compensate in part for the fact that the radiative parameters are not optimal for Tripleclouds, when using perfect clouds. We saw indeed in Figure 3 that for this configuration of ecRad, the overlap parameter leads to underestimating cloud cover (blue points) even with perfect vertical profiles of cloud fraction.

For Spartacus, the angular dependence of reflected radiation is much better represented, avoiding compromises between errors at high and low zenith angles. Furthermore, 3D effects can partially compensate for underestimated effective cloud cover. Overall, there is no clear sign of systematic compensation error with Spartacus. The best simulations obtained with a radiation-based setting give both radiation that is on average as good as that obtained with perfect clouds, and this for maximum cloud fractions that correspond reasonably well to those of the LES.

~~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.~~

~~**Naj : je m'occupe pas de ce paragraphe pour l'instant – je refais les exps avec des cutoff plus faibles c**~~ ~~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 **??** shows that , at 77°, 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 simulations exhibit such compensating error mechanisms.~~

4.3 Where we remove cloud fraction vs. 3D effects compensating errors

To verify our hypothesis that part of the radiative error that is compensated by clouds in TC-08 experiment is related to 3D effects, we do two more pairs of experiments. First, we change the ecRad configuration, to use the best Hybrid one instead of the best `Spartacus` one of Section 3. Indeed, the best Hybrid configuration used on LES cloud profiles led to cloud cover estimates closer to the LES than the best `Spartacus` one, which is more favourable to `Tripleclouds`. We expect remaining `Tripleclouds` flux errors to be smaller than before and hence, if our hypothesis is right, cloud fraction overestimation to be less important.

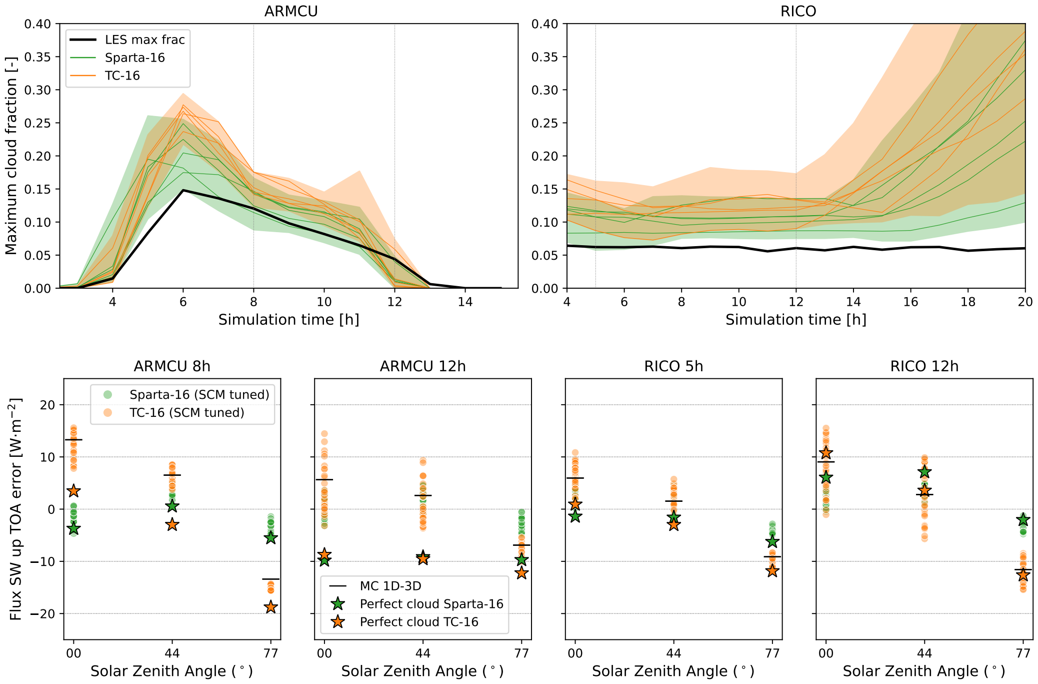


Figure 6. ~~Clouds and radiation~~ Same as Figure 5 but for the ~~10 best for two experiments calibrating LMDZ parameters using radiative constraints: with config best hybrid targetting 3 angle fluxes in orange~~Hybrid tuning of ecRad, ~~spartacus~~ with ~~config best hybrid targetting 3 angle fluxes in~~ green color corresponing to the tuning with `Spartacus`and orange to `Tripleclouds`solvers respectively. ~~Left: ARMCU; Right:RICO. First row: maximum cloud fractions; Second row: upwelling TOA fluxes as a function of solar zenith angle.~~[FH : Preciser si d'autres differences]

This is verified in Figure 6, which is the same as Figure 5 but with both solvers using the best Hybrid configuration of ecRad instead of best Spartacus. The score of the Spartacus experiment is a little bit larger than before, suggesting that the degradation introduced in changing ecRad configuration is not compensated by clouds. ~~Evolution~~ The time evolution of maximum cloud fractions in the Spartacus experiments ~~resemble~~ resembles those of Figure 5, which confirms that changing cloud-geometry parameters to a slightly less good configuration does not impact Spartacus too much. In the Tripleclouds experiment, the radiative score was not improved by much. Yet, the evolution of maximum cloud fraction is closer to LES and Spartacus ones than before: To reach the same radiative accuracy with Tripleclouds, in this experiment, cloud fractions did not have to be as large as before.

Maximum cloud fraction is still overestimated in the TC-16 experiment, compared to Sparta-16. It might be for two reasons: in the best Hybrid configuration, overlap parameter is similar to best Tripleclouds values, but the associated FSD is larger (for the same ℓ) which means that clouds are less reflective; again, in Spartacus, this overlap-heterogeneity combination is balanced with 3D effects, but in Tripleclouds a large FSD leads to an underestimation of cloud reflectivity. Hence, maximum cloud fraction might be overestimated in compensation. ~~Another possibility is that, independently~~ As for the best spartacus tuning and ndependently from heterogeneity, the right 3D fluxes cannot be obtained simultaneously for all solar angles, for the right clouds, if 3D effects are not represented in the model. ~~We test this latter assumption by changing the target metric: Instead of aiming three fluxes calculated at three distinct angles, we target a single flux value, averaged over seven solar~~ The cloud cloud fraction is not increased as much than in the previous "best spartacus" experiments because the errors with perfect cloud ware a little less negatively biased: around -19W·m^{-2}(against -21) for the 8th hour of ARMCU at SZA 77° and +11W·m^{-2}(instead of +8) for the 12th hour of RCIO at SZA 0°. [FH : Verifier les chiffres].

[FH : We show in SI additional experiments that confirm the above explanations : best tripleclouds... Sans ZSA=0 One not shwon by removing the SZA 0○ metrics form the tuning setup. In this case the maximum cloud fraction is still more increased (by a factor of 3 at least - Un paris à ce stade]

A last test is done in which we just change the target metrics to an average over seven angles, from 0° to 77° with step 11°. As 3D effects change sign when sun zenith angle increases, they partially cancel each other out when averaged. If lack of 3D effects is the error that is com-

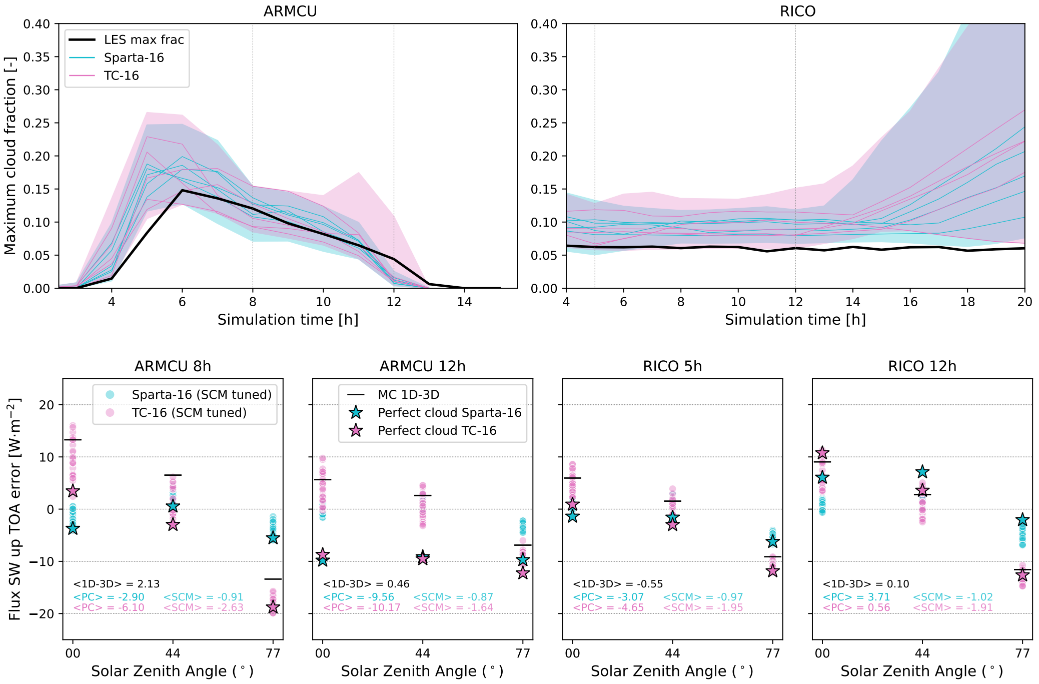


Figure 7. Clouds and radiation for the 10 best for two experiments calibrating LMDZ parameters using radiative constraints: `Tripleclouds` with config best hybrid targetting mean-angle flux in pink, spartacus with config best hybrid targetting mean-angle flux in cyan. Left: ARMCU; Right:RICO. First row: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

pensated by clouds in the `Tripleclouds` experiment, it should be less pregnant when targetting an average flux in which the expression of 3D effects is almost null. It is the unweighted arithmetic average that is considered here, as the simplest way to test our assumption; more complex averages that better represent solar angle distributions on Earth might be more relevant in a real GCM tuning exercise.

Results are presented in Figure 7. With the average flux metric, cloud fraction evolution are the same in the two experiments, using `Spartacus` or `Tripleclouds`. It means that `Tripleclouds` errors (*FSD* too large or lack of 3D effects) are no longer compensated by cloud fractions. Looking at the detailed flux errors, we see that `Tripleclouds` flux at 77° are ~~slightly worse~~ more negatively biased compared to previous experiments, in agreement with the fact that this particular value is no longer constrained. It confirms that maximum cloud fractions were previously overestimated to compensate for lack of 3D effects in particular at this large solar zenith angle; because increasing cloud fraction also increases reflected flux at all other angles, in `Tripleclouds`, flux errors could not be entirely compensated by clouds: increasing cloud fraction too much would have led to even larger flux overestimation at SZA=0°, which was not acceptable given the tolerance-to-error set in these experiments.

This last experiment is an important result for climate modeling: it confirms that 3D radiative effects could for a large part self compensate when averaging over the diurnal cycle, or seasons and latitudes. It also proposes a simple compromise for tuning with a RT code missing 3D effects against LES/MC reference simulations. However, results also clearly suggest the possibility of compensation error at high latitude with overestimated cloud fraction to compensate for the failure of 3D radiative effects in most parameterizations so far.

5 Conclusion

~~**C'est l'ancien texte ?**~~

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 should not be directly extrapolated to climate simulations. Indeed, in the SCM setup considered here, only shallow 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, 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. In the same spirit, we think that research that aims at better underestanding climate models and the act of modeling itself is a crucial aspect of climate sciences.

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/L` It can be configured and installed directly on Linux machines with an installation bash script `https://lmdz.lmd.jussieu.fr/pub/install_lmdz.sh` 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

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

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

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

Pincus, R., Barker, H. W., & Morcrette, J.-J. (2003, July). A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields. *Journal of Geophysical Research: Atmospheres*, *108*(D13), n/a–n/a. doi: 10.1029/2002JD003322

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 DATA

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.

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}.