Investigating cloud–radiation error compensations stemming from climate model calibration

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:

- Ensembles of perturbed-parameter simulations of single column models are used to investigate cloud–radiation compensating errors
- We find that structural errors in radiative transfer models due to lack of 3D effects are compensated by overestimated cloud fractions
- Radiative transfer models with and without 3D effects should not be used with the same overlap and heterogeneity parameters

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 forced error compensation due to targeting top-of-atmosphere radiative fluxes in the tuning process while using an inaccurate radiative transfer parameterization. Here, we investigate parametric and structural errors in state-of-the-art radiative transfer models and ask if these errors might be compensated by errors in cloud properties, to what extent, and in which circumstances. Convection and cloud parameters of two versions of a Single-Column version of a climate Model (SCM), with and without a parameterization of 3D radiative effects of clouds, are tuned targeting reference solar fluxes calculated on Large-Eddy Simulation cloud fields. When 3D effects are neglected, reasonable fluxes are obtained only at the expense of overestimated cloud fractions, partly compensating for underestimated cloud reflectivity at low sun. Targeting mean solar-angle fluxes instead of detailed ones removes this mechanism entirely.

Plain Language Summary

Global climate models are used to understand the climate system and anticipate the consequences of global warming. They are based on a set of equations that represent the various aspects of the system. Among these, cloud physics and radiative transfer (light propagation) play a major role. The equations, called parameterizations, are based on our understanding of the physics, complexity–accuracy compromises, and include parameters that need to be calibrated against observations. Satellite-observed radiative fluxes are often targeted in this calibration process, with the risk of obtaining the right fluxes (close to the observed ones) for the wrong reasons, and in particular for the wrong cloud properties. In this study, we show that the radiative transfer models currently used in climate models present structural errors that could, through the calibration process, be balanced by introducing additional errors in cloud properties. This implies that persistent systematic biases in clouds simulated by global models could be partly due to structural errors in the radiative transfer model, rather than to inadequate modeling of cloud physics. We use a History Matching approach based on machine learning to shed light on these questions and study error compensations in detail, albeit in an idealized framework.

1 Introduction

General circulation models (GCMs) used for climate projections are, like any model, imperfect representations of the climate system. Their behavior 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. Reducing them is a motivation to develop more physically robust parameterizations. Finding ways to detect, characterize, and understand sources of compensating errors 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 doing so, 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.

An important step was taken in this direction in the recent years, with the seminal work of Couvreux et al. (2021) and Hourdin et al. (2021), who proposed a tuning strategy relying on a hierarchy of model configurations, from Single Column Model (SCM) to GCMs. In those studies, SCMs are tuned to ensure that boundary-layer clouds are well simulated, using a few representative cases for which Large Eddy Simulations (LES, i. e. 3D simulations with resolution of a few tens of meters on domains of a few tens of km) are used as a reference. Following the History Matching approach proposed by Williamson et al. (2013), this tuning of the SCM does not produce one optimal parameter set for boundary layer parameterizations, but rules out parameter values that lead to unsatisfyingly large errors in the simulated boundary layers. Then, the GCM is tuned in the remaining "Not Ruled-Out Yet" (NROY) parameter space. This limits error compensations between boundary layer physics and the other components of the model.

Here, we go one step further by investigating compensation errors that may occur (i) *inside* one parameterisation, and (ii) between the parameterisations of the SCM, in cumulus clouds situations, and focus on their solar radiative effect. In particular, we investigate (i) compensations that may arise between different aspects of cloud-geometry inside a radiative transfer scheme and (ii) errors in convection and cloud schemes that may arise from the need to compensate errors in the radiative transfer scheme when radiative transfer is inaccurate.

Indeed, until recently most radiative transfer schemes implemented in GCMs neglected or misrepresented important aspects of cloud geometry such as non-maximally overlapped cloud fractions inside low-cloud layers, internal inhomogeneity of cloud water, and 3D radiative effects associated with horizontal transport of light through cloud sides, all of which affect cloud reflectivity. As detailed geometrical properties of clouds vary in space and time, and might evolve in a warmer climat, it is important that their individual effects, as well as their non-linear interactions, be well understood and captured in radiative transfer schemes. However, Villefranque et al. (2021) has shown that the right reflectivity could be obtained for quite different sets of cloud-geometry parameters for cumulus clouds. Thus we ask: how to ensure that reflectivity is accurate for the right reasons, i.e., for the right cloud-geometry parameters? On the other hand, we acknowledge that cloud reflectivity errors will probably remain in the next generation of GCMs; yet errors on cloud reflectivity will lead to errors on radiative fluxes even if vertical profiles of cloud fraction and cloud water content are perfectly simulated by the boundary layer parameterisations. Thus we ask: to which extent might reflectivity errors be compensated by cloud property errors, and thus lead to the right fluxes for the wrong reasons?

To answer these questions, we design cheap, idealized, SCM tuning experiments, extending on previous work of Couvreux et al. (2021); Hourdin et al. (2021); Villefranque et al. (2021). The specific radiative scheme and SCM that will be used in our study are described in Section 2, as well as reference data used as target for the calibration, and details on the History Matching tuning protocol. Section 3 is dedicated to the question of error compensations inside the radiative scheme, and Section 4 to the question of radiation–cloud error compensations. In Section 5 we discuss our results and their implications, as well as the methodological choices on which they are based.

[[ci dessous une liste des qrev qui ont sauté vu les modifs sur l'intro]

- *[rev2 2)Some references are not cited correctly. Mais je ne vois pas l'erreur ?????]*
- *[rev2 3) sur les problèmes d'accronymes]*

- *[rev2 1) "Words such as "Spartacus," "Tripleclouds," and "History Matching" should preferably be introduced after their first occurrences]*
- *[Pour être plus explicite sur la première occurance de History Matching, même si c'est la deuxième ici. Remarque 1 rev2]*
- *[rev2 3) The full names and their abriavations are miused throughout the manuscript]*
- *[rev3 : This sentence is confusing. Would it be possible to add a few words more precise then "more appropriate ecRad parameters"? I was confused by the words "(instead of best Spartacus)". It seems trivial to me that using the parameters tuned with Spartacus would not work well with Tripleclouds.]* [MCD : D'ailleurs, c'est surtout le fait de viser l'angle moyen qui enlève la surestimation systématique de la clous cover nan ?] (la phrase était : Then, using more appropriate ecRad parameters for the 1D-RT scheme (instead of parameters values tuned for 3D-RT), and targeting 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 the 1D-RT scheme.)

2 Data, models and tools

To run 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 the protocol objective and automatic.

In this study, two different kinds of tuning experiments are explored. They both use the same tuning tool, which is 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, described in Section 2.1. All experiments are made using idealized cumulus cases, following Villefranque et al. (2019).

The difference between the two types of experiments is the model that is tuned: In Section 3, cloud-geometry parameters of the radiative transfer schemes of ecRad (Hogan & Bozzo, 2018) are tuned in a "perfect cloud" setup (ecRad is run on vertical profiles of cloud fraction and water content taken from horizontally-averaged 3D reference LES fields). The radiative transfer scheme and parameters are described in Section 2.2. In Section 4, cloud and boundary-layer parameterizations of the SCM version of LMDZ are tuned, targeting radiative fluxes. The SCM and its parameters are described in Section 2.3 and Supporting Information Table S1.

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 cooling and surface conditions are imposed throughout the simulation for each case. The effects of large-scale advection and radiative cooling are represented by prescribed source terms in the heat and moisture evolution equations, applied in each column of the domain. These sources/sinks of heat and moisture are functions of height and time and replace an explicit resolution of the large-scale dynamics and of radiation. In the ARMCu case, surface fluxes are prescribed as a function of time, following a typical diurnal cycle of turbulent fluxes over land, while in RICO, the sea surface temperature is set to a constant during the whole simulation and turbulent fluxes are calculated by the model. *[Demande de précision de*

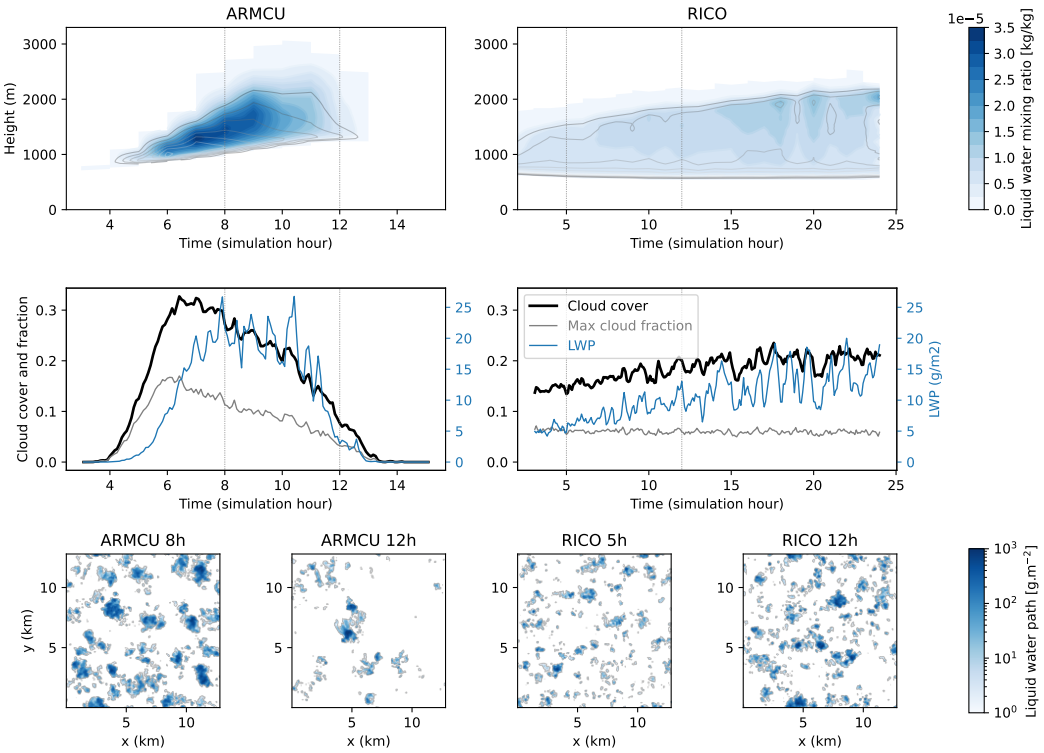

Figure 1. Reference simulations of the ARMCU (left) and RICO (right) cases used in this study. All variables are computed from LES 3D cloud fields. a-b: time evolution of vertical profiles of cloud fraction (contour lines, 0 to 0.16 every 0.02 in ARMCU, 0 to 0.07 every 0.01 in RICO) and cloud water mixing ratio (blue shading). c-d: time evolution of total (vertically integrated) cloud cover, maximum cloud fraction and domain-mean liquid water path. e-h: liquid water path maps at the two hours retained for tuning (ARMCU 8th and 12th hours, RICO 5th and 12th hours, identified by vertical lines in a-d plots).

la part de reviewer 1 sur cette phrase : q2.2 Please elaborate on this. Naj: j'ai rajouté des détails, dites moi si vous pensez que ça suffit ou pas ?]. 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). The reference simulations are presented in Figure 1. Four scenes from these two simulations will be used as constraints in the tuning experiments. They are documented in Table 1.

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. 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 Approximation (ICA) by solving radiative transfer independently in each column of the 3D LES cloud field and taking the average flux. Assuming independent columns is the same as removing 3D radiative effects from the

Table 1. Cloud properties in four scenes extracted from ARMCU and RICO LES.

Case, hour	ARMCU, 8h	ARMCU, 12h	RICO, 5h	RICO, 12h
Cloud cover (%)	26.07	6.99	13.43	20.12
Maximum cloud fraction (%)	11.11	3.09	5.45	5.96
Cloud layer depth (km)	1.625	1.20	1.25	1.675
Cloud cov. / max. cloud frac.	2.346	2.26	2.462	3.379

calculation; the resulting flux is hence called 1D MC flux. Taking the difference between 3D and 1D MC fluxes provides an estimate of "3D radiative effects".

[MCD : harmoniser ICA et 1D dans le texte]

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. 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 cloud droplets (not shown). *[rev3 2.3]*

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, 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). 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, Fielding, Barker, Villefranque, & Schäfer, 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). In the rest of this article, the `Tripleclouds` solver is referred to as 1D-RT and the `Spartacus` solver as 3D-RT, referring to the presence or absence of 3D radiative effects.

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.

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 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 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. 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 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 pre-

dictions 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

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

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. *[rev1 q2.3 : Have the samples that were generated during previous iterations (but still within the NROY space) been reused when training the emulator?]* Then a new ensemble is run, metrics are evaluated, emulators are built, etc. In practice, the parameter vectors used to run the first ensemble of simulation are sampled randomly in the hypercube using a latin hypercube sampling algorithm. The ensemble of next waves used the $10\times N$ first members that satisfies the implausibility condition, among another bigger latin hypercube sampling of typically 10^5 samples. *[Pour répondre à la question 2.4 de rev1]*.

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 σ are significantly smaller than tolerances to error T 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).

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) [MCD : J'ai déjà introduit la signification de PPE dans l'intro : je la remet ici ou pas ? Naj : il y est plus dans l'intro je crois !]. 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:

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

[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 $G_i^* \sim \mathcal{N}(\mu_i, \sigma_i)$.

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 The right cloud reflectivity for the right cloud-geometry parameters?

In this section, HighTune:Explorer is used to explore the behaviour of the 1D- and 3D-RT solvers in a "perfect-cloud" setup. Their cloud-geometry parameters are varied to produce PPEs that are then analysed. These parameters are the overlap decorrelation length ℓ, the fractional standard deviation of in-cloud water FSD, and the cloud effective scale C_s (only active in the 3D-RT model). As the focus is on internal RT compensating error related to cloud geometry, flux differences that might be related to physical cloud properties are removed by using mean LES cloud fraction and liquid water content profiles as inputs to the 1D- and 3D-RT solvers. These profiles were horizontally-averaged over the 3D fields that were used as inputs to the Monte Carlo reference RT simulations.

Effects of cloud geometry on cloud reflectivity have been studied for a long time (see e.g., McKee and Cox (1974); Barker, Stephens, and Fu (1999); Várnai and Davies (1999); Shonk, Hogan, Edwards, and Mace (2010); Hogan et al. (2019); Villefranque et al. (2021)).*[rev2 2)Some references are not cited correctly. Mais je ne vois pas l'erreur ????? Naj : je pense que c'est la liste de noms au lieu du "et al." mais c'est le format de JAMES on y peut rien]* In particular for solar reflectivity of cumulus clouds, it is well-known that:

- the widely used maximum overlap assumption tends to underestimate total (i.e. vertically integrated) cloud cover, and consequently, cloud reflectivity; [MCD : on a précisé qu'on parlait du SW ici ? Naj : maintenant oui !]
- neglecting in-cloud optical-depth inhomogeneity systematically leads to overly reflective clouds;
- assuming infinite cloud size means neglecting 3D radiative effects, which produces clouds that are too bright by around 10W·m^{-2} when the sun is at zenith, and too dim by around 10 W·m^{-2} when the sun is low on the horizon. Note that since 3D effects shift from positive to negative values depending on Solar Zenith Angle (SZA), *[rev2 3) sur les problèmes d'accronymes Naj : ok c'est bien la première occurence ici pour l'instant]* they are null at around 40-50°, and partially cancel out over a day. *[rev3 2.4 : Does it account for 10 W/m2 for high and low sun? J'ai pas compris ce qui allait pas dans la phrase Naj : la phrase a changé, est ce que c'est plus clair maintenant ? (J'avais pas compris non plus la question)]* [[je déplace ça de la sec 3.2 à ici car pas vraiment un nouveau résultat - mais j'ai pas encore repris / vérifié ta phrase] This can be explain by the competing effects between side leakage, that lead to less reflective clouds and dominate at low SZA, and side illumination that lead to more reflective clouds and dominate at high SZA. [MCD : Une ref ? Je veux bien que vous vérifier les sens.] *[Pour répondre à rev2 q5) Another interesting thing in section 3 is that the results in both spartacus and tripleclouds are close when the SZA is 44degree. What is the underlying mechanism?]*

Recent RT models such as those implemented in ecRad Shonk and Hogan (2008); Hogan and Shonk (2013); Hogan et al. (2019); Pincus, Barker, and Morcrette (2003) include terms that represent effects of cloud geometry on radiative fluxes, and are able to simulate fairly

accurate cloud reflectivity when compared to reference 3D MC simulations. Yet, different choices of cloud-geometry parameter values may result in the right cloud reflectivity. Here, we investigate these configurations, to determine if the right reflectivity might be achieved for the wrong reasons, that is, for the wrong cloud geometry.

3.1 Experimental desgin

In all experiments, the three-dimensional parameter space that is explored is $\Omega_{\ell,FSD,C_s} = [] \times [] \times []$ [MCD : c'est laissé vide pour mettre les bornes de chaque paramètre ?]. All tuning experiments target solar reflected fluxes of the four scenes of ARMCU and RICO cases (see Figure 1) for three distinct SZA, 0°, 44° and 77°, following Villefranque et al. (2021). This choice of distinct SZA values is made to have metrics that are sensitive to 3D radiative effects. [MCD : J'enleve l'histoire des signes opposés car c'est dit plus haut]

Reference radiative fluxes used as targets are 3D-MC fluxes for the 3D-RT tuning experiments, and 1D-MC fluxes for the 1D-RT experiments. It is particularly significant that we require the 1D-RT model to perform well in 1D and not in 3D: we do not force it to compensate for a structural error that we know is present in the solver, namely the absence of 3D radiative effects. Using 1D-MC references is a common practice in 1D-RT model evaluation (see e.g. Barker et al. (1999), among many), though it is not necessarily the strategy that would effectively be employed in the tuning of a GCM, where "true" radiative fluxes might be targeted even at the expense of internal compensating errors.

In the tuning experiments, parameter values that result in errors larger than three times the tolerance to error $T = 3.5$ W·m^{-2} are rejected. This value of tolerance to error is an estimate of the `Spartacus` structural error on cumulus clouds, based upon the work of Villefranque et al. (2021) (see Supporting Information Text S1).

For each experiment, ten waves (iterations) of history matching are produced, each consisting in 101 simulations. *[q2.4 reviewer1 : Which sampling algorithm was applied? e.g., LHS? QMC? Naj : on l'a pas dit dans la section 2 ? Maelle Si ça a été rajouté]* 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, i.e. the bests ones, are selected. These ensembles constitue the PPEs under study in this section, named 1D-PPE for the `Tripleclouds` tuning experiments and 3D-PPE for the `Spartacus` one.

3.2 Independent calibration of 1D- and 3D-RT solvers

Figure 2 shows the location of the 1D-PPE and 3D-PPE members in parameter space. The first three rows and columns of the subplot matrix are relative to parameter values, while the last row and column is relative to the score $S(\boldsymbol{\lambda})$ of each configuration, as per Equation (2). Since the C_s parameter is not used in the 1D-RT solver, the red points corresponding to this configurations spread randomly along the C_s axis.

First, note that all configurations of the 1D-PPE have scores below 1, while all those of 3D-PPE have scores above 1. Of course it does not mean that 1D-RT is better than 3D-RT in an absolute sense, since the scores of 1D-PPE are computed against 1D-MC references, intentionally neglecting 3D effects. It just means that 1D-RT is closer to reference 1D fluxes than 3D-RT is close to reference 3D fluxes. It might not be too surprising for a 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 the two-stream framework. In fact, when runing the same tuning experiment with the 1D-RT targetting 3D-MC fluxes intead of the 1D-MC ones, tolerance to error had to be

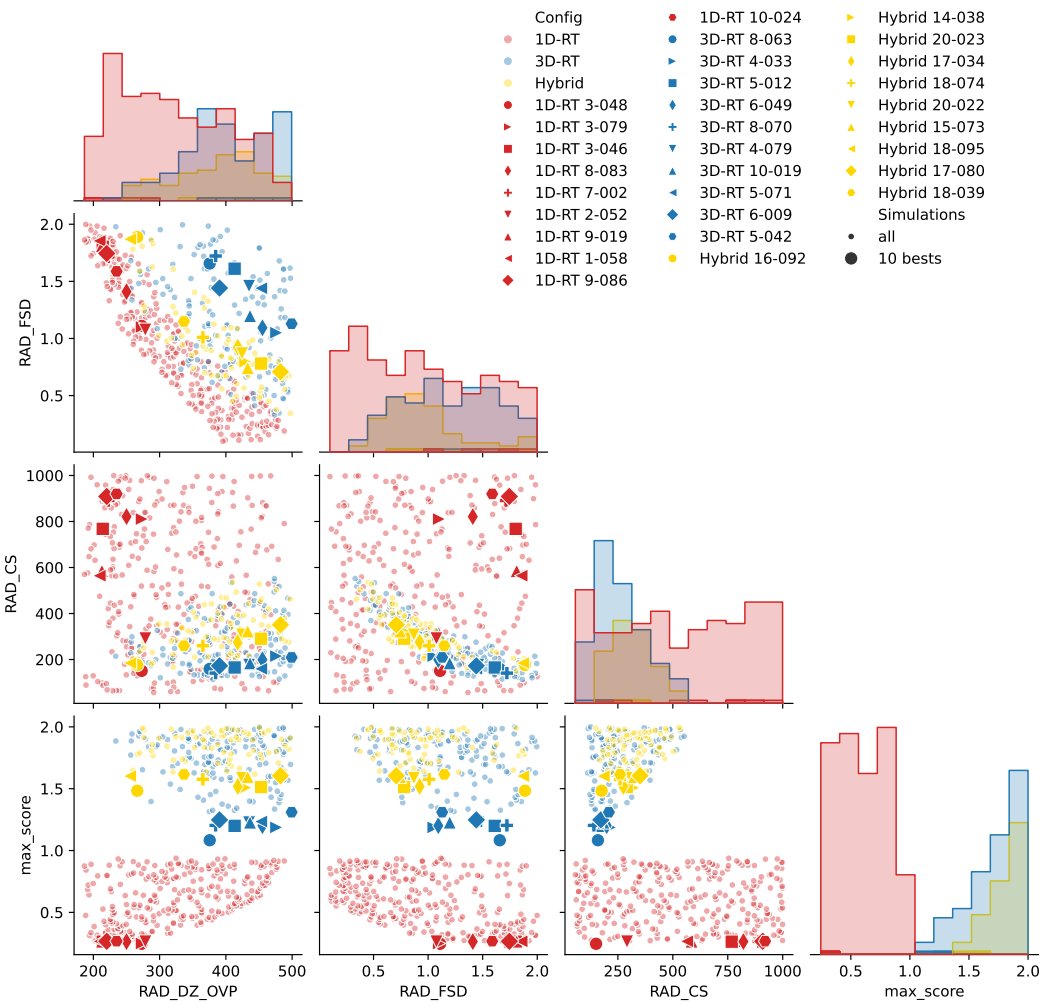

Figure 2. Parameters of 300 best simulations for three "perfect-cloud" ecRad tuning experiments: 1D-RT `Tripleclouds` (red), 3D-RT `Spartacus` (blue) and Hybrid (gold). [MCD : faire en sorte de ne pas avoir besoin de la phrase qui suit] RAD_DZ_OVP is the vertical overlap decorrelation length parameter ℓ, RAD_FSD is the relative in-cloud inhomogeneity parameter FSD, C_s is the cloud size parameter C_s. The max_score is $S(\boldsymbol{\lambda})$. Each column corresponds to one parameter. The top subplot of each column shows histograms of the parameter values for the 300 best configurations, for each experiment. Then, each row corresponds to a second parameter, and subplots show the location of the configurations in the 2D space of the column $\times$ row parameters.

increased from 3.5 $\mathrm{W{\cdot}m^{-2}}$ to 9 $\mathrm{W{\cdot}m^{-2}}$ to reach a non-empty NROY space, along with an increase of the best score simulation (see). Nevertheless, the NROY space of the two experiment are very similar and simulated very similar radiative fluxes (not shown). The difference of tolerance to error is in fact explain by the change of fluxes reference values, and not by the change in fluxes simulation. This result mean that 1D-RT cloud geometry parameter can't compensate for the lack of 3D radiative effects.

The best 1D-RT simulations, highlighted by large red symbols in Figure 2, exhibit a strong negative correlation between ℓ and FSD. In 1D radiation, this is easy enough to understand: with a smaller decorrelation length, clouds are more randomly (or less maximally) overlapped, leading to a larger cloud cover and smaller cloud optical thickness 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,

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

The optical thickness at a given location on the horizontal is the sum of the optical thicknesses of the cloudy layers above it. *[rev3 2.6 There is a part of this sentence missing. J'ai rajouté it mais pas sur que ça suffise]* Distributing the individual optical thicknesses of the cloud layers horizontally on a wider area (large cloud cover with clouds of moderate optical thickness) 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. This explains why the tuning algorithm selects smaller values of the FSD to compensate for larger values of ℓ in order to simulate the right fluxes.

In the 3D-PPE (blue symbols on Figure 2), the relationship between ℓ and FSD also seems to exist, but the 10 best configurations of this PPE show more dispersion than the 10 best 1D-PPE ones. Our first guess was that horizontal transport, driven by the third parameter C_s, modulates reflectivity in 3D-RT while it has no effect in 1D-RT, which means that, in 3D-RT, a given value of FSD can be associated with various values of ℓ and still produce the same reflectivity. For instance a smaller cloud cover could be compensated by more intense 3D effects to maintain reflectivity at large SZA. However, 3D effects tend to decrease reflectivity when the sun is close to zenith and increase it otherwise; and C_s does not vary much between these 10 best configurations so this explanation might not be sufficient.

All things considered, the internal dispersion of the best 3D-RT configurations might be a detail compared to the fact that the 10 best configurations of each experiment are located in distinct regions of the $\ell \times FSD$ space: The best 3D-RT configurations are located in the "large ℓ, large FSD" corner of the parameter space, where none of the best 1D-RT simulations can be found. It implies that cloud geometry parameters yielding strongly overlapped, highly heterogeneous clouds are not acceptable from a 1D-RT perspective. Does this mean that reflectivity in the best 3D-RT configurations is right for the wrong reasons?

To investigate this point further, let us look at cloud covers calculated by the exponential-random overlap model, for the 10 decorrelation length values considered as "best" values in 1D-PPE compared to the 10 best values in the 3D-PPE. The upper panel of Figure 3 presents a comparison of reference 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 1D-RT or 3D-RT 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

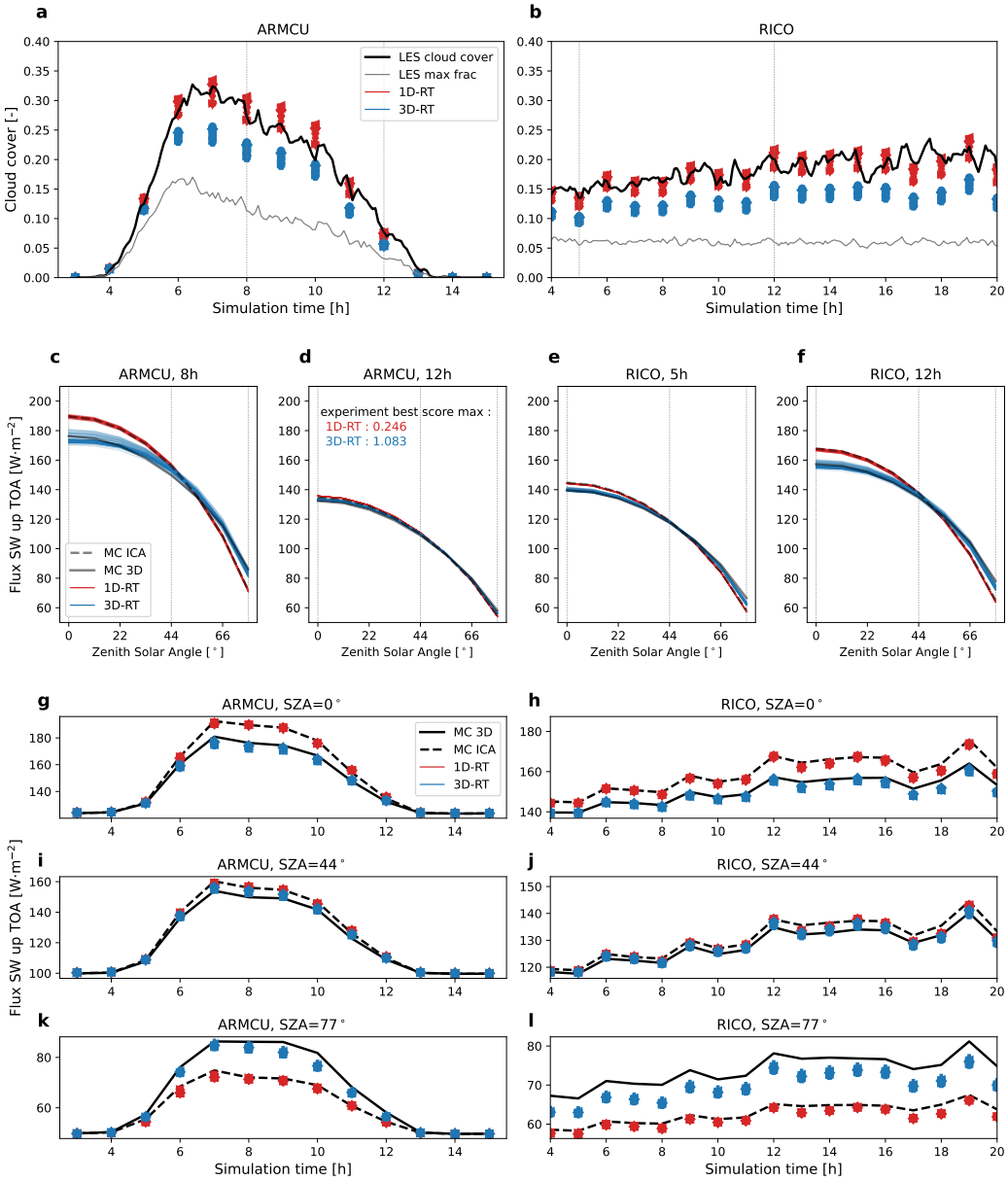

Figure 3. [MCD : harmoniser MC ICA-1D] Cloud and radiation variables for the 10 best simulations of two perfect-cloud experiments (1D-RT with `Tripleclouds` in red, 3D-RT with `Spartacus` in blue), compared to reference. Left column: ARMCU. Right colum: RICO. a-b: cloud cover (black) and maximum cloud fraction (gray) in LES fields, and cloud cover computed by ecRad (color points) using input cloud fraction profiles calculated from LES cloud fields, and the exponential–random overlap model parameterized by various decorrelation lengths (RAD_DZ_OVP values in Figure 2). c-l: Fluxes simulated by Monte Carlo 3D (black full lines) and 1D (black dashed lines), 1D-RT with `Tripleclouds` (red) and 3D-RT with `Spartacus` (blue). c-f: upward TOA fluxes as a function of solar zenith angle, for the 4 cloud scenes used as constraints in the tuning experiments: ARMCU 8th (c) and 12th (d) hours, RICO 5th (e) and 12th (f) hours. For SZA=0°, 44° and 77°, fluxes plotted are directly those used as metrics. For the other one (SZA=11°,22°,33°,55°,66°), they are further computed following the same framework, for diagnosis only. *[Pour répondre à rev1 q2.5 "it seems the experiments were run at more than three solar angels, but I didn't see any description about this configuration in the method section."]* g-l: upward TOA fluxes, calculated at every hour of the two LES simulations, each time for three solar angles. g-h: SZA=0°, i-j: SZA=44°, k-l: SZA=77°.

regions of a given layer, to cloudy and clear regions of adjacent layers. These plots show that cloud covers corresponding to the best 1D-RT simulations are very close to those of the LES. In comparison, cloud covers corresponding to the best 3D-RT simulations are systematically smaller than those of the LES (in agreement with the fact that ℓ values are larger). This suggests that in these best 3D-RT configurations, reflectivity is right for the wrong reasons, i.e., for clouds that are too maximally overlapped (we might say, too compact). In this view, we conclude that compensating errors are at work between overlap, inhomogeneity and 3D effects in the 3D-RT solver. This is an important result, as it means that finding the best 3D-RT configuration and removing 3D effects (e.g., using 1D-RT with the same (ℓ, FSD) values, or setting C_s to infinity in 3D-RT), will produce wrong fluxes compared to the 1D MC reference.

A last important result shown in Figure 3(c-f) is that only the 3D-RT solver is able to capture the angular dependency of cloud reflectivity. *[rev1 2.7 What is "the second line"? Naj : en fait c'est parce que le bon mot est "row"]* [MCD : pas sur que ce soit bien dit le "alone" dans la phrase, moi je compris qu'il a besoin de personne d'autres pour bien faire la dépendance angulaire, alors qu'on veut plutôt dire que c'est le seul à pouvoir la faire ? Ou bien j'ai mal compris ?] The 1D-RT always underestimates (resp. overestimates) cloud reflectivity at high SZA (resp. at low SZA) compared to 3D radiation. *[rev1 2.8 : Compared to zenith or compared to Spartacus? In addition, Tripleclouds was calibrated targeting 1D MC flux. Can we improve the angular dependency of Tripleclouds if targeting 3D MC flux? Maelle : remarque bien prise en compte dans ce paragraphe]* [[on l'a dit plus haut du coup j'enlèverai ça : Note that 1D-RT is equal to 3D-RT at middle SZA (around 44°) because 3D radiative effects generally cancel out around middle SZA, as can be deduced from the 1D- and 3D-MC lines of Figure 3 (c-f). Maelle : oui bien sur] This is not an artifact of the calibration experiment: tuning the 1D-RT scheme against 3D-MC fluxes leads to the same result (see Figure S2 in Supplementary Information).

3.3 Obtaining the right reflectivity for the right cloud geometry

The best 3D-RT configurations were found to yield the right reflectivity for the wrong cloud geometry, which we interpret as a sign of internal compensating errors. Our main hypothesis here is that a "good" 3D-RT model should be able to simulate 3D fluxes as well as 1D fluxes when 3D effects are removed from the simulation. In the following, we look for configurations of the 3D-RT scheme that satisfy this requirement. *[Ce paragraphe correspond a la remarque de rev3 : It is not clear to me why the authors try to find a paramete set that works well for the two different models simultaneously. Is it to reach the objectives of section 4? Please clarify how at the beginning of this section. Et aussi a sa remarque 2.8 Could you clarify here why you want to satisfy both solvers at the same time?]*

We thus design a third tuning experiment where (ℓ, FSD) values that do not result in both accurate 3D fluxes, and accurate 1D fluxes when 3D effects are removed, are ruled out from the acceptable 3D-RT parameter space. It is called the Hybrid experiment, since it looks for (ℓ, FSD, C_s) values that satisfy both solvers at the same time.

The 300 best Hybrid configurations are presented in Figure 2 (yellow marks). Their scores are generally worse than for previous experiments (mostly, above 1.5). This is expected since the best 3D-RT configurations of Section 3.2 are bad choices for 1D-RT, and conversely. However, a compromise was found within our tolerance to error.

The location of the best Hybrid configurations in the (ℓ, FSD) space illustrates this compromise: for a given choice of FSD, the 10 best simulations of the Hybrid experiment correspond to ℓ values between best 1D-RT and best 3D-RT 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.

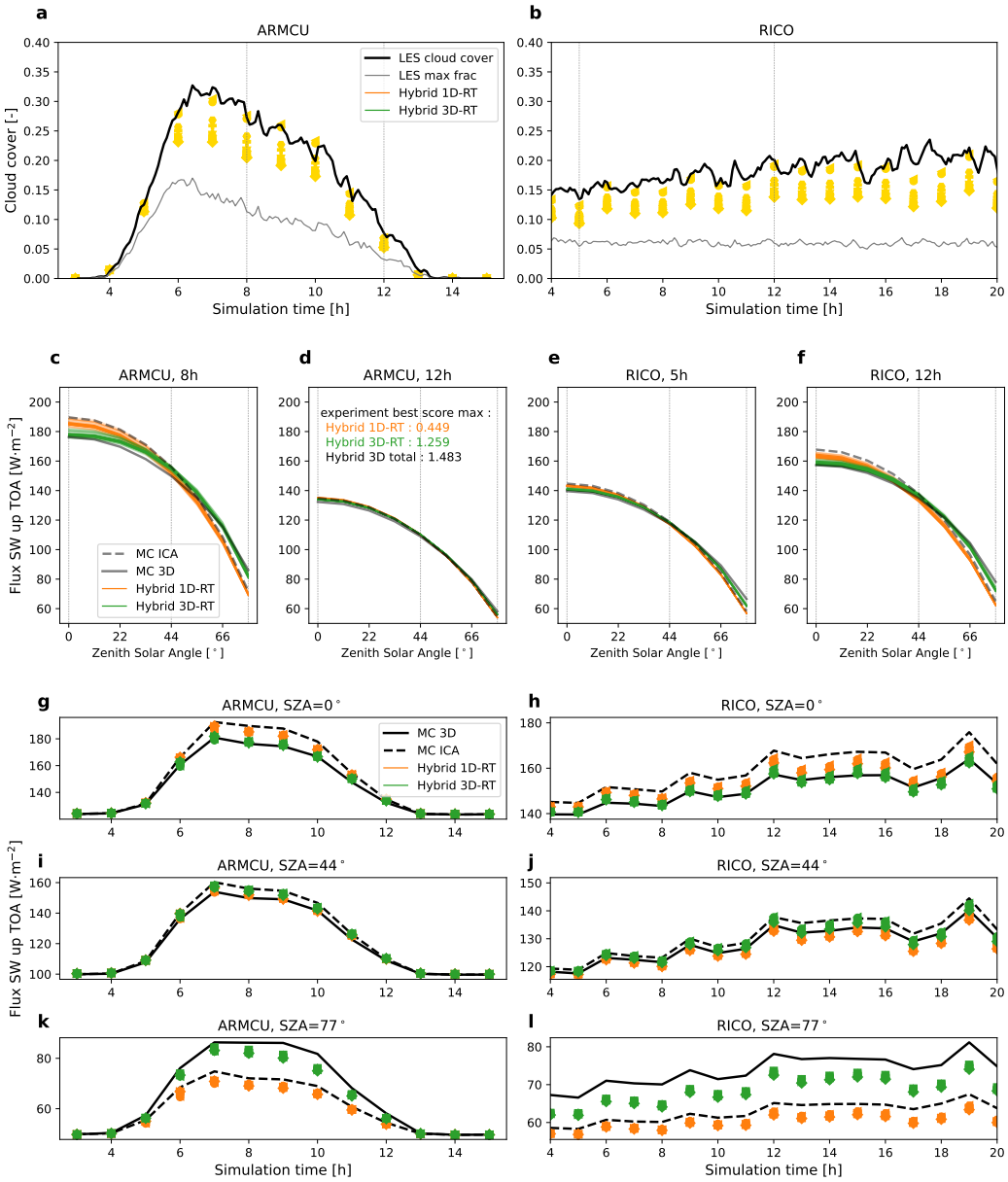

Figure 4. *[rev1 2.10. Figure 4, What are the yellow marks? J'ai fait des changements plus loin pour répondre à cette question + il faudrait ajuster dans la légende du jaune pour Hybrid]* Cloud and radiation variables for the 10 best configurations of the Hybrid experiment, compared to reference. As in Figure 3 except that: in a-b, decorrelation length parameters used to compute cloud cover (yellow marks) are those of the 10 best simulations from the Hybrid experiment; in c-l, 3D-RT (green) and 1D-RT (orange) are both configured with the same 10 best Hybrid configurations.

The very best configuration in the Hybrid experiment corresponds to large FSD and small ℓ values, close to the best 1D-RT simulations. Accordingly, in the upper panel of Figure 4, this best Hybrid configuration corresponds to the yellow marks that are closest to the 3D LES cloud cover. Comparing fluxes in Figure 4 with those of Figure 3, we can see that 1D-RT 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 1D-RT performances. As for 3D-RT, 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 3D-RT experiment (compare for instance, yellow square with blue diamond). This would lead to more reflective clouds than before, if nothing else changed. However, these configurations are also associated with larger C_s values, which means less intense 3D effects, and therefore less reflective clouds at large SZA. In the end, reflectivity at large SZA is largely unchanged with respect to the best 3D-RT of Section 3.2. 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, if our hypothesis that a good 3D-RT model should be able to simulate 1D fluxes once 3D effects are removed is valid, then the fact that scores are smaller in the best 3D-RT configurations of Section 3.2 is a sign of over-fitting. As a consequence, the scores of the best Hybrid configurations, and not those of the best 3D-RT, should be used to derive estimates of structural errors of the 3D-RT solver (errors that remain once parametric errors are removed). This assumption and its implications will be further discussed in Section 5. In the meantime, let us conclude that two 3D-RT configurations are potential candidate for "best reflectivity": the "best 3D-RT" configuration and the "best Hybrid" one, even if the first one might be there for the wrong reasons.

4 Compensations between radiative transfer and cloud models

In Section 3, we went into details of internal compensations of cloud-geometry aspects within the RT schemes, in a perfect-cloud setup. In the following, we aim at characterzing error compensations between RT schemes and boundary-layer cloud parameterizations within a SCM. We wonder if errors in convection and cloud schemes may arise from the need to compensate cloud-reflectivity errors when RT is inaccurate. Indeed, the column (or scene) reflectivity is a combination of physical cloud properties in the column (namely, cloud fraction and liquid water content profiles) and cloud reflectivity (itself a function of cloud geometry). Building on the detailed analysis of Section 3, we investigate how cloud-reflectivity errors might propagate through tuning to errors on cloud fractions simulated by the SCM to result in accurate scene reflectivity. We thus study cloud properties of SCM PPEs produced under radiative constraints, in experiments that use different RT configurations with varying levels of errors on cloud reflectivity. Since they yield relatively small cloud-reflectivity errors, experiments based on 3D-RT configurations are regarded as control experiments, and error compensations are examined in equivalent 1D-RT-based experiments, that is, using 1D-RT configurations that are the same as the control 3D-RT but without 3D effects.

4.1 Experimental design

The experiments of this section are designed to mimic the tuning protocol often followed in GCMs when targeting top-of-the-atmosphere (TOA) radiative fluxes from satellite observations. In our idealized SCM version of the protocol, and contrary to Section 3, experiments based on either 1D- or 3D-RT target the same 3D MC fluxes ("true"

radiative fluxes). Also contrary to Section 3, in which the metrics were instantaneous fluxes, here the metrics are hourly-mean fluxes. References are obtained by averaging MC fluxes computed every 5 min on 3D LES outputs. In the SCM, fluxes are calculated on hourly-averaged cloud profiles.

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 Supporting Information). Because we specifically want to investigate compensating errors between radiation and clouds, radiative parameters are not varied in these experiments. This is not necessarily the strategy that would be employed in the tuning of a GCM. *[rev3 2.9 Could you include the radiative transfer parameters in this experiment as well? Could it be a way to mitigate the compensating errors? Naj : j'ai précisé]*

Each tuning experiment consists in 30 iterations. At each iteration, 130 free-parameter vectors are sampled in the NROY space and ARMCU and RICO are simulated using these 130 configurations of LMDZ. One RT configuration is then run offline for the two chosen hours of the two cumulus cases and each of the 130 SCM 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 SCM parameter vectors are ruled out using a threshold that varies from 3 in the first 5 iterations to 2.5 in iterations 6 to 10, to 2 in iterations 11 to 30. The NROY space is thus efficiently narrowed down.

Tolerance to error is set to $T_{3D} = 3.5$ W·m^{-2} for all metrics of 3D-RT experiments, and $T_{1D} = 9$ W·m^{-2} for all metrics of 1D-RT experiments. T_{3D} was found by trial and error, seeking the smallest value still yielding non-empty NROY space without falling under 3.3 W·m^{-2}, which is the minimum estimated structural error of 3D-RT (see Supporting Information Text S1). T_{1D} was obtained following another, cheaper strategy: a preliminary experiment was run using an arbitrarily low tolerance to error, and the best simulations were selected amongst the 5 waves that were performed before the NROY space was emptied. Setting T_{1D} to the errors of these best simulations ensures that it exists configurations satisfying this requirement, but it might be a large overestimate of structural error if only bad configurations were explored during the preliminary experiment.

4.2 Cloud reflectivity vs. cloud fraction error compensations for too-compact clouds

[MCD : pas sur de ma traduction `Spartacus` et `Tripleclouds` ici. A reprendre a tete reposée] [[j'ai repris]

In these first experiments, the RT schemes are configured using the "best-3D" configuration of Section 3.2 (configuration 8-063 in Figure 2). The 3D-RT simulations use the full configuration while the 1D-RT ones only use overlap and inhomogeneity values. The version of 3D-RT that yields the best reflectivity is regarded as a control, and cloud-reflectivity errors are introduced by using 1D-RT instead of 3D.

For each experiment, maximum and average absolute errors (metrics-wise) of the best-score simulation are retained as a measure of the accuracy of the model. They are presented in Table 2. Note that 3D-RT fluxes are around as accurate as when 3D-RT was run on reference cloud profiles, since tolerance to error in Section 3 was of 3.5 W·m^{-2} and the best 3D-RT score was close to 1. On the other hand, flux errors in the 1D-RT experiment are about four times larger than those of 3D-RT. This means that even with the possibility of compensating radiation error with biased cloud properties, the 1D-RT scheme could not produce fluxes as accurate as 3D-RT in these experiments. Note that 1D-RT scores are much larger than in perfect cloud experiments but this is partly be-

Table 2. Fluxes and cloud fraction errors for the experiments of Section 4.

Configuration Best / Solver / Metrics	Shown in Fig.	Flux MxAE ($W{\cdot}m^{-2}$)	Flux MAE ($W{\cdot}m^{-2}$)	CF MxAE (-)	CF MAE (-)
3D-RT / 3D-RT / 3 SZA	5	3.6480	2.4931	0.0234	0.0156
3D-RT / 1D-RT / 3 SZA	5	14.2253	8.2945	0.1261	0.0941
Hybrid / 3D-RT / 3 SZA	6	3.7890	2.8047	0.0568	0.0444
Hybrid / 1D-RT / 3 SZA	6	14.2117	8.3226	0.1213	0.0841
Hybrid / 3D-RT / $\langle$ SZA $\rangle$	7	0.4360	0.2976	0.0462	0.0368
Hybrid / 1D-RT / $\langle$ SZA $\rangle$	7	1.2210	0.7972	0.0644	0.0482
Hybrid / 1D-RT / 2 SZA	S2	12.6082	8.7031	0.1230	0.0852
1D-RT / 1D-RT / 3 SZA	S2	11.2147	7.4093	0.1085	0.0717

MxAE: Maximum Absolute Error. MAE: Mean Absolute Error. CF: Maximum Cloud Fraction. Mean and max flux error is over the 4 scenes $\times$ number of SZA. Mean and max CF error is over the 4 scenes.

cause 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 ARMCU and RICO test cases. The bottom panel shows errors on the TOA upwelling fluxes as a function of solar zenith angles, for the four cloud scenes where radiative constraints are applied (8th and 12th hour of ARMCU, 5th and 12th hour of RICO). Errors due to neglecting 3D effects in the MC calculations, and errors made by 3D-RT and 1D-RT run on LES mean cloud profiles are also represented in the figures. It shows that although SCM simulations were constrained to be as close as possible to 3D MC fluxes, those based on 1D-RT they still closely follow the 1D MC fluxes. This is consistent with results shown in Figure S2 of Supporting Information [MCD : en dire un peu plus ? Pour Fredho : C'est la figure trpliclouds qui vise MC-3D en mode perfect clouds sur laquelle on voit que tripleclouds ne sait pas compenser les effets 3D et les meilleurs simus ont des paramètres ℓ et FSD dans la même région que tripleclouds qui vise du 1D. La seule différence est la tolérance à l'erreur de l'expérience qui doit être augmenté jusqu'à une dizaine de $W.m^{-2}$ soit l'ordre de grandeur des effets 3D] 3D-RT fluxes however match the 3D-MC reference for almost every angle, including those that were not constrained (not shown). In the upper plots, we see that, in addition to better simulating solar radiation, the 30 best SCM simulations of the 3D-RT experiment exhibit cloud fraction evolutions that are compatible with reference values for ARMCU, and close to the LES ones for RICO. Conversely, the 30 best 1D-RT simulations persistently overestimate maximum cloud fraction for both cases.

We interpret this as a sign of compensating errors due to wrong cloud reflectivity in the 1D-RT scheme. Indeed, plots in the lower row of Figure 5 show that the largest errors in 1D-RT simulations run on LES mean cloud profiles (red stars) are systematically the underestimation of reflected solar radiation at SZA=77°. The absolue error is the largest for the 8th hour of ARMCU where it reaches -20.5 $W{\cdot}m^{-2}$. The tuning protocol, which rejects configurations when the metric-wise maximum error is too large, systematically selects configurations where the error at SZA=77° is smaller than that of "perfect clouds", thanks to an increase of cloud fraction that compensates for a lack of cloud reflectivity. This increase of cloud fraction is limited by the fact that it also increases reflected radiation at SZA=0°, which cannot be too large either. The tuning finally selects configurations for which the bias shows the same absolute value for the maximum underestimation (about -15 $W{\cdot}m^{-2}$ at SZA=77° for the 8th hour of ARMCU) and max-

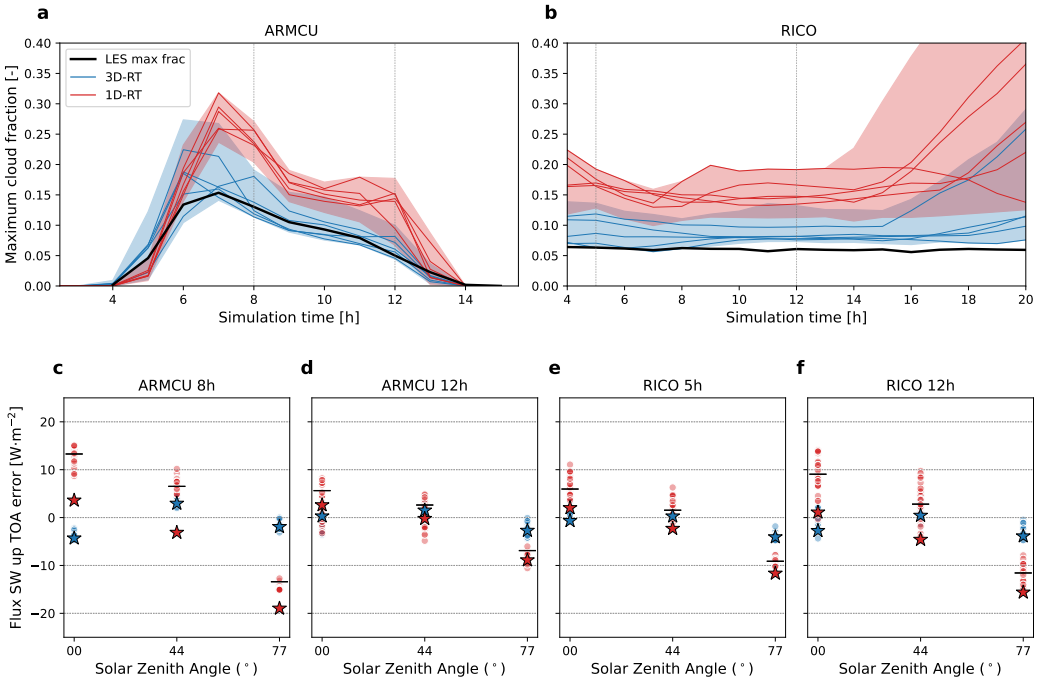

Figure 5. Clouds and radiation for the 30 best simulations for two experiments calibrating LMDZ parameters under radiative constraints using either `Tripleclouds` (red) or `Spartacus` (blue). Both solvers use parameters of the best `Spartacus` configuration from Section 3.2 (`Spartacus` 8-063). Left: ARMCU. Right:RICO. a-b: maximum cloud fraction (shadings represent the 30 best enveloppe, lines show the 5 best). c-f: error in upwelling TOA fluxes as a function of solar zenith angle, compared to MC 3D, for the four cloud scenes: ARMCU 8h (c), ARMCU 12h (d), RICO 5h (e), RICO 12h (f).

imum overestimation (about 15 W·m^{-2} at SZA=0° for the 8th hour of ARMCU and 12th hour of RICO).

In the control experiment based on 3D-RT, angular dependence of reflected radiation is much better represented, avoiding compromises between errors at high and low zenith angles. Overall, there is no sign of cloud–radiation compensating errors when the 3D-RT scheme is used: The best SCM configurations obtained with a 3D-RT-based tuning produce at the same time (i) fluxes that is as good on average as in the perfect-cloud experiments, and (ii) maximum cloud fractions that correspond reasonably well to those of the LES.

Here, the only difference between the two experiments is the removal of 3D effects in the 1D-RT-based experiment. The intuitive conclusion at this stage would be that, in the 1D-RT-based experiment, cloud fractions simulated by the SCM are overestimated to compensate for lack of 3D effects. However, we saw in Section 3 that removing 3D effects from the best 3D-RT configurations led to bad 1D-RT configurations, that is, to (ℓ, FSD) values which result in an underestimation of cloud reflectivity. This means that in this 1D-RT-based experiment, cloud reflectivity is not only underestimated at large SZA because 3D effects are neglected, but also because the other geometry parameters are not adequately chosen. Even if the difference between the control and 1D-RT experiments are indeed the strict removal of 3D effects, it would be inaccurate to conclude that it is the lack of 3D effects that leads to overestimating cloud fraction, when it might instead be due, to some extent, to parametric errors.

4.3 Cloud reflectivity vs. cloud fraction error compensations due to lack of 3D effects

In order to quantify the impact of neglecting 3D effects in a more robust way, additional tuning experiments are performed. First, the "best Hybrid" configuration is used instead of the "best 3D-RT" one.*[rev3 2.10 Reading this, it seems that the "best Tripleclouds" was not used in sections 4.2. Naj : ben oui c'est le cas. Ca veut dire qu'il a cru qu'on avait pris la best TC au coup d'avant... j'espère que c'est plus clair dans la nouvelle version]* The rationale is that in the perfect-cloud setup of Section 3, the overlap parameter of the "best Hybrid" configuration yields cloud cover estimates that are close to the 3D LES, which in turn leads to more accurate 1D-RT reflectivity. Therefore, using the "best Hybrid" configuration, we expect the remaining 1D-RT cloud-reflectivity errors to be smaller than with the "best 3D-RT" and hence, if our hypothesis is correct, cloud fraction overestimation to be less pronounced. *[rev3, remarques generales sur ce paragraphe : "I read the first part of this section multiple times without completely understanding the objective. The title of this section is not clear either. "]*

This is verified in Figure 6, which is the same as Figure 5 but with both solvers using the "best Hybrid" configuration instead of "best 3D-RT". The score of the 3D-RT experiment is a little bit larger than before, suggesting that the degradation introduced in changing the configuration is not compensated by clouds. The time evolution of maximum cloud fractions in the 3D-RT experiments resembles those of Figure 5, which confirms that changing cloud-geometry parameters to a slightly less good configuration does not significantly impact 3D-RT fluxes. In the 1D-RT experiment, the radiative score was not improved by much. Yet, the evolution of maximum cloud fraction is closer to the LES: To reach the same radiative accuracy with the 1D-RT scheme, that is, the same scene reflectivity, cloud fractions did not have to be as large as before. Our interpretation is that the error in cloud reflectivity that is compensated by increasing cloud fraction is smaller: perfect-cloud 1D-RT errors at SZA 77° are smaller with this best Hybrid configuration, around -18.8 W·m^{-2} (against -20.5) at the 8th hour of ARMCU.

Yet, maximum cloud fraction is still overestimated in the 1D-RT-based experiment, compared to 3D-RT. This might be for two reasons: in the best Hybrid configuration,

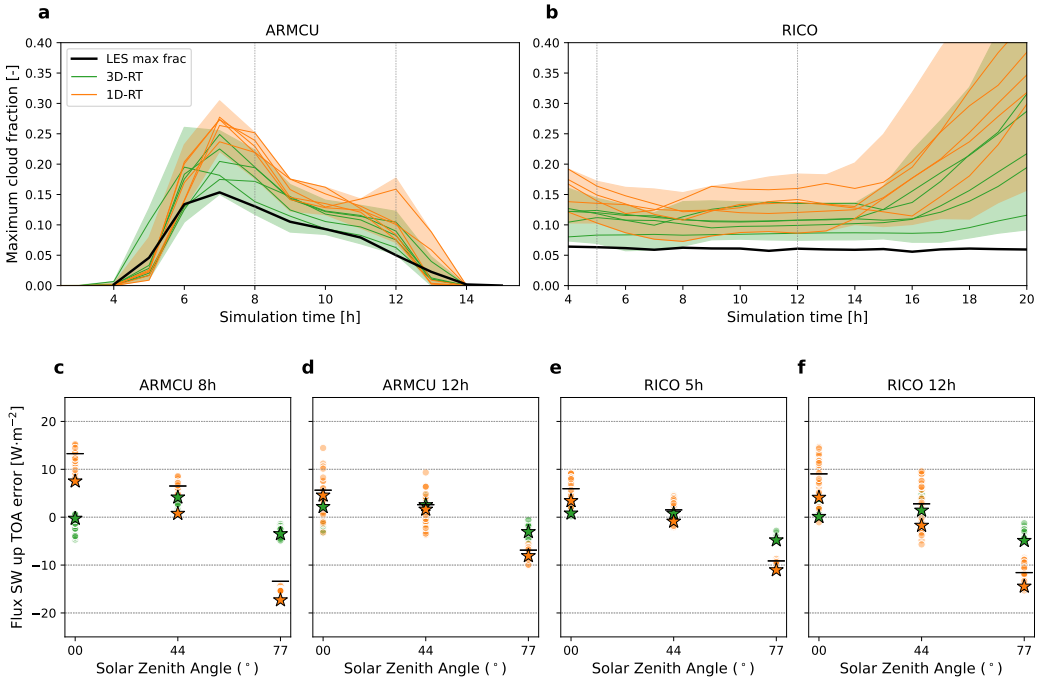

Figure 6. Clouds and radiation for the 30 best simulations for two experiments calibrating LMDZ parameters under radiative constraints using either `Tripleclouds` (orange) or `Spartacus` (green). Same as Figure 5 except that both solvers now use parameters of the best Hybrid configuration from Section 3.3 (Hybrid 16-092).

overlap parameter is similar to best 1D-RT values, but the associated FSD is larger (for the same ℓ) which means that clouds are less reflective; again, in 3D-RT, this overlap-heterogeneity combination is balanced by 3D effects, but in 1D-RT a large FSD leads to an underestimated cloud reflectivity. Hence, maximum cloud fraction might again be overestimated in compensation of this parametric error. The second reason is that, independently of heterogeneity, accurate 3D fluxes cannot be obtained simultaneously for all solar angles, for the right clouds, if 3D effects are not represented in the model.

A final pair of experiments is done in which we change the target metrics to an average over eight solar zenith angles, from 0° to 77° with step 11°. As 3D effects change sign when sun zenith angle increases, they partially cancel out on average. If lack of 3D effects is indeed the error that is compensated by cloud fractions in the 1D-RT experiment, then it should be less pronounced when targeting an angular average of fluxes, 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, although 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 is the same in the two experiments, using 3D-RT or 1D-RT. It means that 1D-RT 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 1D-RT flux at 77° are 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

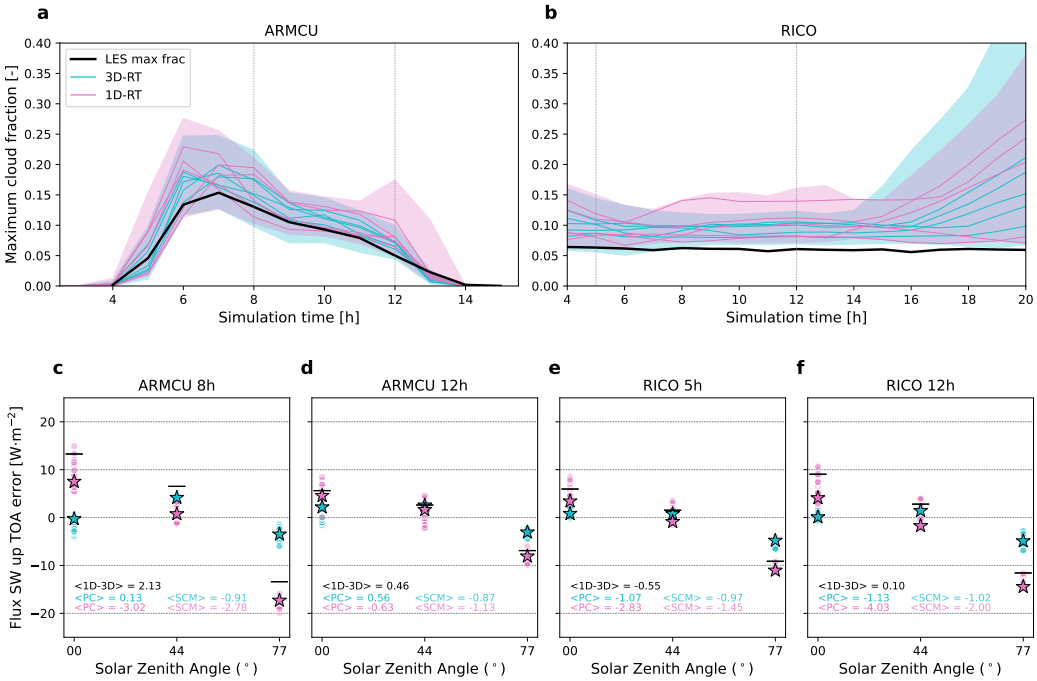

Figure 7. Clouds and radiation for the 30 best simulations for two experiments calibrating LMDZ parameters under radiative constraints using either `Tripleclouds` (pink) or `Spartacus` (cyan). Same as Figure 6 except that the radiative constraint is now the flux averaged over solar zenith angles, instead of flux values taken at three different SZA. c-f: In addition to flux errors as a function of SZA (markers), SZA-mean flux errors are written: ⟨1D-3D⟩ is the average 3D effects, ⟨PC⟩ is for Perfect Cloud experiments (associated with star markers), ⟨SCM⟩ is for SCM-tuned experiments (associated with bullet markers).

solar zenith angle; because increasing cloud fraction also increases reflected flux at all other angles, in 1D-RT experiments, 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 metrics tolerance to error. To support this statement, a complementary experiment where the SCM was tuned based on 1D-RT but using only constraints at SZA=44° and SZA=77° is presented in Supporting Information (Figure S2).

This last experiment, targeting average flux, is an important result for climate modeling: it confirms that 3D radiative effects could, for a large part, self compensate when averaged over diurnal cycle, seasons or latitudes. It also proposes a simple compromise to tune a model with a 1D RT code, against LES/MC reference simulations. However, these results also suggest the possibility of compensating errors at high latitudes, where cloud fraction might be overestimated to compensate for lack of 3D radiative effects, if the model was tuned targeting high-latitude metrics.

5 Conclusion and discussion

By tuning parameterizations of a climate model in SCM mode targeting TOA upward solar fluxes, we evidenced mechanisms of error ~~compensations~~ compensation both inside the RT model itself and in interaction with the cloud physics simulations.

Considering the RT computation itself, two kinds of ~~compensations~~ compensation were identified. ~~First we show~~ We showed first how the cloud vertical overlap and in-cloud heterogeneities can compensate each other inside the RT solver to simulate fairly accurate 1D fluxes, a compensation already known and studied in 1D-RT as well as 3D-RT (Villefranque et al., 2021). Secondly, when tuning 3D-RT computations targeting 3D-MC simulations, we showed how the best performing simulations selected favour parameters which underestimate the 1D reflectivity, in the sense that 1D-RT simulations run with the same parameter values underestimate the reflected solar radiation at TOA. This underestimation is consistent with the fact that the total cloud cover diagnosed in Ecrad by the exponential-random scheme is underestimated compared to the LES cloud cover. One way to interpret these results would be to say that errors in the computation of 3D effects in the 3D-RT `Spartacus` solver are compensated by an underestimation of cloud cover or 1D reflectivity.

This way of presenting the second mechanism for error compensation relies however on the hypothesis that 3D effects would be added on the results of a 1D-RT computation. In this case, it would be legitimate to privilege values that are optimal both in 1D-RT and 3D-RT mode, in order to fully avoid compensating errors between 3D effects and cloud cover or reflectivity. Actually, when running the 3D-RT solver, 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. It may ~~bot~~ so not be a surprise ~~so~~ that the tuning done in 1D-RT targeting 1D-MC, considering the number of numerous approximations and unavoidable error compensation inside the 1D-RT scheme itself, privilege parameter values which are not optimal when targeting 3D-MC fluxes using a more physical radiative transfer model. From this perspective, we consider free parameters as *effective parameters* that are meaningful only within their respective models and this constraint relaxes. Related to this, the cloud cover that is directly determined by the decorrelation length free parameter has no reason to be exactly the same in 1D and 3D-RT solver, nor between RT SCM computations and MC references. ~~Here also, this diagnostics only documents approximately how an effective cloud cover is related to the computation of RT, accounting for the effect of cloud overlap.~~ Thereby, constraining the 3D-RT solver to well simulate 1D fluxes when 3D radiative effects are removed is not a straightforward choice. ~~On the one hand, there~~ There is no clear or definite reason to privilege parameter values that would be the best both in 1D-RT and 3D-RT computations. ~~On the other~~

~~hand,~~ Note however that the fact that we were able to find parameter values for which RT computations match reasonably well, even if not optimally, both the 1D-MC and 3D-MC fluxes in 1D-RT computation and 3D-RT computation respectively, together with a reasonable diagnostic of the cloud cover compared to LES, is a strong indication that the "reasonably good simulations" were obtained for the "reasonably good physics". Beyond these specific results, this study illustrates the step forward allowed by being able to start address error compensation as a scientific question, ~~but opens as well~~ opening in the mean time a number of ~~open~~ questions that modelers should have in mind when using, evaluating and tuning their models.

We show in a second part that structural errors in the 1D-RT model can be compensated by errors in cloud properties when TOA radiative fluxes are targeted in a tuning process. ~~Maximum~~ The maximum cloud fractions obtained after tuning of the free parameters of the cloud and boundary layer model targeting the 3D-MC flux ~~computations,~~ are systematically overestimated to compensate two possible errors in the 1D-RT. The first one concerns the choice of cloud geometry parameter values, that lead to too compact and heterogeneous cloud from 1D-RT perspective when selected to perform well in the 3D-RT solver. The second one is a compensation of the structural error ~~of~~ due to not accounting for 3D radiative effects, that lead to a strong underestimation of cloud reflectivity at high SZA. 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 risk of degrading the clouds representations to obtain ~~,~~the right fluxes, in particular at high latitudes were the occurrence of high zenith angles is larger.

This demonstration is made in an idealized configuration, and our results should not be directly extrapolated to 3D coupled climate models. 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 constraint, but Couvreux et al. (2021); Hourdin et al. (2021)~~ Couvreux et al. (2021) and Hourdin et al. (2021) claim 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.

The main structural error under study here is known and introduced intentionally, using the 1D-RT instead of 3D-RT configuration, that is generally not the case when evaluating and tuning a climate model. This allowed us to demonstrate the potential of the History Matching tuning method for the quantitative investigation of compensating errors at work in climate models. Through tuning we explore model configurations and resulting simulated climates in a parameter space reduced 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 detect where the development effort have to be made to improve the simulation of the key climate characteristics by the model. On the other side, 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 uncertainties on various aspects of climate.

Finally, 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 tools to better understand climate. In the same spirit, we believe that research that aims at better understanding climate models and the act of modeling itself is a crucial aspect of climate sciences.

Open Research Section

High-Tune Explorer (htexplo) and LMDZ 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 568`. The LMDZ release used in the study can be configured and installed directly on Linux machines with an installation bash script `https://lmdz.lmd.jussieu.fr/pub/install_lmdz.sh` run with as `bash install_lmdz.sh -SCM -v 20250337.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 fa642e, 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

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). Combin-

ing 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