Process-based climate model development harnessing machine learning: III. The Representation of Cumulus Geometry and their 3D Radiative Effects

Najda Villefranque[1], Stéphane Blanco[2], Fleur Couvreux[1], Richard Fournier[2], Jacques Gautrais[3], Robin J. Hogan[4], Frédéric Hourdin[5], Victoria Volodina[6], Daniel Williamson[6,7]

[1]CNRM, Université de Toulouse, Meteo-France, CNRS, Toulouse, France
[2]LAPLACE, Université de Toulouse, CNRS, Toulouse, France
[3]Centre de Recherches sur la Cognition Animale (CRCA), Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, Université Paul Sabatier, France
[4]ECMWF, Reading, United Kingdom
[5]LMD-IPSL, Sorbonne University, CNRS, 4 pl Jussieu, Paris, France
[6]The Alan Turing Institute, 96 Euston Road, London, United Kingdom
[7]Exeter University, Exeter, United Kingdom

Key Points:

- A novel calibration approach is applied to an offline radiation scheme to disentangle sources of uncertainty in cloud radiative effects
- The SPARTACUS solver is run on cloud profiles derived from LES cumulus fields and compared to Monte Carlo 3D radiative transfer computations
- Adjusting SPARTACUS cloud geometry parameters provides effective values that improve surface and TOA fluxes compared to LES-derived values

Corresponding author: Najda Villefranque, `najda.villefranque@gmail.com`

Abstract

Process-scale development, evaluation and calibration of physically-based parameterizations of clouds and radiation are powerful levers for improving weather and climate models. In a series of papers, we propose a strategy for process-based calibration of climate models that uses machine learning techniques. It relies on systematic comparisons of single-column versions of climate models with explicit simulations of boundary-layer dynamics and clouds (LES). This paper focuses on the calibration of cloud geometry parameters (vertical overlap, horizontal heterogeneity and cloud size) that appear in the parameterization of radiation. The solar component of a radiative transfer scheme that includes a parameterization for 3D radiative effects of clouds (SPARTACUS) is run in offline single-column mode on an ensemble of input cloud profiles synthesized from LES outputs. The space of cloud geometry parameter values is efficiently explored by sampling a large number of parameter sets (configurations) from which radiative metrics are computed using fast surrogate models that emulate the SPARTACUS solver. The sampled configurations are evaluated by comparing these radiative metrics to reference values provided by a 3D radiative transfer Monte Carlo model. The best calibrated configurations yield better predictions of TOA and surface fluxes than the one that uses parameter values computed from the 3D cloud fields: the root-mean-square errors averaged over cumulus cloud fields and solar angles are reduced from ~ 10 Wm^{-2} with LES-derived parameters to ~ 5 Wm^{-2} with adjusted parameters. However, the calibration of cloud geometry fails to reduce the errors on absorption, which remain around 2 to 4 Wm^{-2}.

Plain Language Summary

A way to improve the accuracy of climate models is to improve the physical formulations that represent the effects of small-scale processes on the evolution of atmospheric state. Processes that involve clouds and radiation are particularly important due their key role on climate. Choosing values for the parameters inherent to these formulations is a difficult task. This series of papers presents a rigorous strategy for calibrating models. It is based on comparisons between high-resolution models that accurately represent clouds and single-column versions of a climate model, on the basis of process-oriented metrics such as cloud height. A set of acceptable parameters is efficiently found using machine learning techniques. In this third part, the parameters that control the radiative effects of cloud geometry are calibrated. A recent radiation model that includes realistic representation of the radiative effect of cloud heterogeneity, cloud vertical structure and cloud size is evaluated and calibrated using references that are provided by a ray-tracing algorithm that tracks virtual photons in virtual cloud fields produced by high-resolution models (LES). Calibration improves the model with respect to using parameters diagnosed in the LES. Good agreement is found only when interception of sunlight by cloud sides is represented.

1 Introduction

Cloud–radiation interactions, through their strong impact on the Earth's global energy balance (Ramanathan et al., 1989), are key processes in the evolution of the Earth's climate. The radiative effect of cumulus clouds is particularly important due to their permanent presence in large regions of the Earth's troposphere and their large optical thickness (Berg et al., 2011). They are responsible for a large proportion of the uncertainty around climate sensitivity (Dufresne & Bony, 2008; Bony et al., 2015). Cloud–radiation interactions are also key for climate model tuning. A common practice involves adjusting cloud parameters to match the observed cloud radiative effect (CRE) (Hourdin et al., 2017). This can lead to selecting model configurations in which errors in cloud properties and in the parameterization of radiative transfer (RT) compensate for each other. A famous example of that is the "too few too bright" syndrome found in numerous climate models (Karlsson et al., 2008; Nam et al., 2012), in which the underestimated cover and overestimated optical depth of low clouds yield an acceptable global CRE.

Accurately predicting the radiative effects of cumulus clouds is, therefore, of major importance, yet remains challenging, particularly when the detailed 3D structure of these geometrically complex clouds is unknown, as is the case in large-scale models (see e.g. Barker et al. (2003)). The effects of cloud geometry are most often separated into three aspects: the vertical overlap of cloudy regions occupying distinct model layers, which controls the total cloud cover; the horizontal variability of in-cloud water content, which controls the mean transmissivity of the cloudy region of the layer (inhomogeneous clouds are less opaque than their homogeneous counterpart; Newman et al. (1995)); and the cloud size which controls the intensity of radiative transfers between clouds and their "clear-sky" environment, called 3D effects. Examples of 3D effects include the interception of direct sunlight by cloud sides when the sun is not at zenith (McKee & Cox, 1974), which decreases transmission; or the channelling of downward flux and entrapment of upward flux towards the surface (Várnai & Davies, 1999; Hogan et al., 2019), which both increase transmission. The sign of resulting 3D effects depends on solar zenith angle. Gristey et al. (2020) found that 3D effects of sub-tropical land cumulus fields act to heat the surface when averaged over a diurnal cycle; neglecting these effects in climate models might introduce significant errors in the predicted evolution of the system.

Various propositions have emerged in recent years to take these effects into account. In the Monte Carlo Independent Column Approximation (McICA) of Pincus et al. (2003), 1D radiative transfer is solved in sub-columns that are sampled based on vertical overlap and horizontal heterogeneity assumptions. The representation of cloud geometry is hence separated from the resolution of radiative transfer. In SPARTACUS (Hogan & Shonk, 2013; Schäfer et al., 2016; Hogan et al., 2016, 2019), the 3D structure of clouds are intrinsically mixed with the two-stream equations that are used to solve RT within the cloudy column. SPARTACUS is the only parameterization that includes 3D effects in addition to overlap and heterogeneity effects. This paper is dedicated to the evaluation and calibration of SPARTACUS, with specific attention paid to its internal modelling of cloud 3D geometry.

This is the third part of a series of papers in which a novel approach for climate model tuning is defended. A first calibration step is advocated for, during which Single Column Models (SCM) and Large-Eddy Simulations (LES) are compared using process-scale metrics in order to eliminate regions of the parameter space where the SCM parameterizations produce unsatisfying results. During the final global model tuning, only the parameter values that were not rejected during the first step are explored. This ensures that only model configurations that reach the calibration target for the good reasons (for instance, produce the right CRE for the right clouds) can be selected, thereby limiting compensation errors. Part I (Couvreux et al., 2020) describes this approach and the associated numerical tools. Part II (Hourdin et al., 2020) applies them to the calibration of a 3D climate model after prior calibration of the parameterization of shallow convection in the SCM/LES framework.

Here, we go one step further in this effort to untangle the sources of uncertainties in climate models by calibrating SPARTACUS cloud geometry parameters assuming perfect cloud profiles.

In practice, 3D RT is solved by Monte Carlo (MC) in 3D cloud field outputs from LES of four idealized cumulus cases to provide reference radiative metrics. These same 3D cloud fields are summarized to a handful of vertical profiles (most importantly cloud fraction and liquid water content (LWC)) that are provided as inputs to SPARTACUS, whose outputs are compared to the MC references. SPARTACUS also requires the specification of parameters related to cloud geometry. As these parameters have a physical interpretation, values can be derived from the LES cloud fields. Alternatively, they can

Table 1. Configuration of ecRad in the following work.

Property	Option	Reference
Gas model	RRTMG-IFS	Iacono et al. (2008)
Aerosols	None	
Liquid cloud optics	SOCRATES	Manners et al. (2017)
Liquid water content distrib. shape	Gamma	
Cloud overlap scheme	Exp-Ran	Hogan and Illingworth (2000)
Solver	SPARTACUS	Schäfer et al. (2016); Hogan et al. (2016)
Entrapment	Explicit	Hogan et al. (2019)

be adjusted using the calibration tool described in Part I (Couvreux et al., 2020). This latter approach is arguably more appropriate given the structural errors in the model (see discussion in Section 4).

The paper is organised as follows: Section 2 describes the ecRad RT scheme, the MC model, the 3D LES and the resulting 1D profiles. In Section 3, the High-Tune:Explorer calibration tool is briefly described before being applied to SPARTACUS. Four calibrated configurations are then analysed. The main results are discussed in Section 4.

2 Radiative Transfer Models and Cloudy Atmosphere Data

This section presents the SPARTACUS solver of the ecRad radiation scheme we are calibrating (Hogan & Bozzo, 2018), the MC model (Villefranque et al., 2019) that serves as reference, the LES clouds and the methodology used to translate these 3D fields into the 1D profiles used as inputs to ecRad.

2.1 ecRad

The ecRad scheme (Hogan & Bozzo, 2018) has been operational in the Integrated Forecasting System (IFS) at the European Centre for Medium-Range Weather Forecasts (ECMWF) since 2017. Recent efforts have led to a notable increase in flexibility as well as in efficiency compared to previous schemes. Another important step was the development of SPARTACUS (Schäfer et al., 2016; Hogan et al., 2016, 2019), a two-stream based solver that explicitly represents the 3D effects of clouds. An offline version of ecRad is freely avalaible at `https://github.com/ecmwf/ecrad`. The configuration used in this paper is summarized in Table 1.

Three parameters need to be provided to SPARTACUS in addition to standard cloud profiles. They relate to the three main aspects of cloud geometry mentioned in the introduction: vertical overlap, horizontal heterogeneity and cloud size.

1. **Overlap decorrelation length.** Following Hogan and Illingworth (2000), the cloud cover $C_{i,i+1}$ of two adjacent layers of cloud fractions c_i, c_{i+1} is expressed as

$$C_{i,i+1} = \alpha_{i,i+1} C_{\text{max}}(c_i, c_{i+1}) + (1 - \alpha_{i,i+1}) C_{\text{rand}}(c_i, c_{i+1}) \quad (1)$$

where C_{max} and C_{rand} are two cloud covers computed respectively from the "maximum" and "random" overlap of cloud fractions and α is the overlap parameter. It is modeled as an exponential function:

$$\alpha_{i,i+1} = \exp(-\frac{\Delta z(i, i+1)}{z_0}) \quad (2)$$

where $\Delta z(i, i{+}1)$ is the vertical distance that separates the center of the two layers and z_0 is the overlap decorrelation length.

2. **Fractional standard deviation of in-cloud liquid water.** Following the Tripleclouds model of Shonk and Hogan (2008), the effect of horizontal variations of LWC on radiation is accounted for by dividing each layer's cloudy region into two thin and thick sub-regions. To distribute the LWC into the two sub-regions of a given layer and infer their respective optical depths, a gamma-shaped distribution of the liquid water is assumed, characterized by a mean and a standard deviation σ. The fractional standard deviation (FSD) of the distribution (ratio of σ to the mean in-cloud LWC) is used to characterize the horizontal variability of LWC in each layer.

3. **Radiatively effective cloud scale.** Following Hogan and Shonk (2013); Hogan et al. (2016, 2019), terms are added in the two-stream equations of Tripleclouds to account for horizontal transport. These terms are proportional to the length of the interface between clear and cloudy regions: for a given cloud fraction, 3D effects will be larger for a large number of small clouds than for a single large cloud. The cloud perimeter density p (perimeter length per surface units) is modeled as:

$$p = \frac{4c(1-c)}{C_s} \tag{3}$$

where c is the cloud fraction and C_s is the radiatively effective cloud scale (or size).

2.2 Monte Carlo reference computations of solar 3D RT

A Monte Carlo (MC) method is used to compute solar 3D RT in 3D cloud fields, considered as the "truth" in comparisons to ecRad estimates.

MC methods are widely used to accurately compute 3D RT in complex media (see for example Marchuk et al. (1980), Mayer (2009) or Marshak and Davis (2005)). The model used here is based on the High-Tune library described in Villefranque et al. (2019), and is freely available online at `https://gitlab.com/najdavlf/scart_project`. The algorithm consists in tracking a large number of virtual photon paths throughout a virtual medium, explicitly simulating all radiative processes such as emission, absorption, scattering and surface reflection. Whenever a path hits the ground or the TOA, its weight is added to a virtual sensor. Paths are terminated upon absorption or escape in space.

In this work, each simulation consists of ten million paths so that the Monte Carlo errors in our metrics are around 0.1%. Fewer paths would have been necessary to estimate the boundary fluxes to the same accuracy. The relative error on absorption is larger because absolute absorption is small and because absorption is computed from TOA and surface fluxes, therefore the error on the absorption is the sum of errors for TOA and surface estimates.

The optical properties input to the MC model are the same as in ecRad, that is, RRTM-G data for gas (Iacono et al., 2008) and SOCRATES data for clouds (Manners et al., 2017). Spectral integration is performed in both models on the $0.2 - 12.2$ μm interval. This prevents compensating errors between cloud geometry effects and mismatched optical properties.

Three important differences between SPARTACUS and the MC model remain. First, SPARTACUS is a two-stream model that relies on the asymmetry parameter g instead of the detailed angular scattering phase function that is used in the MC model (see Supplemental Information for details). Second, SPARTACUS (as with many atmospheric two-stream RT solvers) is based on the δ-Eddington approximation of Joseph et al. (1976), which corresponds to scaling the optical properties of the clouds to account for large amounts of energy scattered in a very small solid angle around the forward direction. Third, SPARTACUS only sees vertical profiles that summarize the 3D structure of clouds while the MC model acts upon the fully detailed 3D cloud field. Using this MC model as a reference to adjust geometry parameters will mask compensating errors between geometry effects and pure radiative transfer. We argue that this is legitimate since these aspects are fundamentally entangled in SPARTACUS.

2.3 3D fields from LES

For this study, four idealized cumulus cases have been simulated using the French LES model Meso-NH (Lafore et al., 1997; Lac et al., 2018):

- ARM-Cumulus (ARMCu; Brown et al. (2002)), a case of continental cumulus developing over the Southern Great Plains, with a clear signature of the diurnal cycle of the boundary layer in the cloud characteristics. Cloud cover ranges from 0 to 30%;
- BOMEX (Siebesma et al., 2003), a case of marine shallow cumulus forced with constant surface fluxes through the simulation. Cloud cover ranges from 10 to 20%;
- RICO (vanZanten et al., 2011), a second case of marine cumulus, forced with constant sea surface temperature through the simulation. Cloud cover ranges from 15 to 25%;
- SCMS (Neggers et al., 2003), a case of continental cumulus developing in Florida, with strong moisture advection into the domain caused by the nearby ocean. Cloud cover ranges from 0 to 45%.

All simulations were performed on small domains (6.4 km $\times$ 6.4 km $\times$ 4 km) with high spatial resolution (25 m $\times$ 25 m $\times$ 25 m). The horizontal boundary conditions are periodic. The four cases are standards of the literature used in LES intercomparison exercises. Detailed descriptions of the setups, initial conditions and forcings can be found in the reference papers. From these four simulations, thirty-five 3D fields of temperature, pressure, mixing ratio of water vapor and liquid water are used in this study, among which eight will be used in the calibration process of Section 3 (the colored entries in Table 2).

Using an object-identification tool (freely available at `https://gitlab.com/tropics/objects`; Brient et al. (2019)), individual clouds are labelled in each field. A cloud is defined as an ensemble of contiguous cells where the liquid mixing ratio is greater than 10^{-6} kg/kg. Each scene is then described in terms of cloud characteristics, some of which are presented in Table 2. The cloud cover is the fraction of cloudy columns in the domain. To first order, cloud cover controls the transmitted and reflected solar fluxes. The number density is the total number of identified clouds in the scene divided by the horizontal surface of the domain. For a given cloud cover, a larger number density indicates a longer interface between clouds and clear sky, hence more 3D radiative effects. The maximum depth is the highest minus lowest altitudes at which clouds are present. When the sun is not at zenith, the "effective" cloud cover (that is, the cloud cover projected in the sun's direction) depends on the cloud layer depth. Surface CREs computed by MC are also provided at solar zenith angles (SZA) 0 and 77 degrees. CREs are computed as the difference between a full-sky simulation (including clouds) and a clear-sky simulation (where clouds are removed).

Some of the cloud fields might not be realistic because of the small domain size or other numerical constraints (see e.g. Gristey et al. (2020)). Calibration tests have been performed with a wider domain showing only weak sensitivity (not shown). Another limitation of the LES is that clouds were simulated using a one-moment microphysical scheme that did not predict droplet concentrations, hence no detailed information on droplet size was directly available in the 3D fields. In the radiation computations, the droplet size

Table 2. Cloud characteristics from the 35 scenes issued from four standard cumulus cases simulated by LES. Scenes selected for the calibration process are in bold and colors.

Case	Hour	Cover [%]	Number density [km^{-2}]	Max depth [km]	Surface CRE [Wm^{-2}] SZA 0°	Surface CRE [Wm^{-2}] SZA 77°
ARMCu	04	2.722	0.73	0.175	-1.10	-1.35
ARMCu	05	13.174	1.59	0.300	-7.79	-8.63
ARMCu	06	27.139	1.39	0.525	-53.36	-30.74
ARMCu	07	29.416	2.00	0.825	-74.24	-38.08
ARMCu	**08**	**26.343**	**1.64**	**1.225**	**-69.87**	**-38.32**
ARMCu	09	26.180	1.44	1.050	-63.63	-39.03
ARMCu	10	23.499	1.61	1.375	-61.15	-33.63
ARMCu	11	23.029	1.15	1.275	-71.51	-41.70
ARMCu	**12**	**12.663**	**0.81**	**1.450**	**-36.79**	**-19.32**
BOMEX	04	13.884	2.71	1.025	-20.03	-18.63
BOMEX	**05**	**16.301**	**2.17**	**1.200**	**-30.29**	**-22.15**
BOMEX	06	18.001	2.71	1.200	-28.67	-22.10
BOMEX	07	18.204	2.69	1.125	-35.71	-25.26
BOMEX	08	19.081	2.25	1.375	-37.20	-27.50
BOMEX	09	14.175	2.39	1.075	-23.52	-17.64
BOMEX	10	16.585	2.05	0.975	-34.17	-23.67
BOMEX	**11**	**10.318**	**2.00**	**0.775**	**-14.40**	**-11.16**
BOMEX	12	14.294	2.15	0.650	-20.23	-15.05
RICO	04	13.933	2.27	0.950	-18.30	-18.68
RICO	**05**	**13.802**	**2.15**	**0.850**	**-19.84**	**-17.41**
RICO	06	17.195	2.25	1.025	-27.90	-26.22
RICO	**07**	**18.054**	**2.34**	**1.175**	**-33.32**	**-27.50**
RICO	08	19.252	2.69	1.225	-40.16	-29.69
RICO	10	23.451	2.20	1.425	-59.46	-31.64
RICO	11	21.048	2.25	1.125	-41.24	-30.15
RICO	12	16.768	2.32	1.350	-34.01	-25.02
SCMS	04	44.035	4.86	1.050	-103.42	-56.07
SCMS	**05**	**37.947**	**3.71**	**1.450**	**-104.15**	**-55.75**
SCMS	**06**	**32.010**	**2.78**	**1.400**	**-90.75**	**-42.64**
SCMS	07	29.108	2.51	1.450	-78.74	-44.19
SCMS	08	20.961	2.05	1.725	-52.24	-34.19
SCMS	09	15.678	1.88	1.600	-33.70	-22.65
SCMS	10	18.272	1.81	1.200	-38.85	-26.65
SCMS	11	11.980	0.93	1.050	-28.47	-18.79
SCMS	12	1.502	0.51	0.325	-1.24	-1.20

distribution is therefore assumed to be the same everywhere within the clouds, with an effective radius of 10 μ m. In both cases, what matters the most for the calibration is that ecRad and MC see exactly the same clouds.

2.4 1D profiles from 3D fields

From each 3D cloud field output from LES, 1D profiles are derived to serve as inputs to ecRad. Temperature, pressure, vapor and liquid mixing ratios are horizontally averaged from the 3D fields on each vertical level and extended above the LES domain top using the I3RC (Cahalan et al., 2005) mid-latitude summer (MLS) cumulus profiles provided in the ecRad package. There is no cloud above the LES domain. Gas mixing ratios (other than water vapor) are set as in the I3RC MLS cumulus case. Cloud fraction is computed at each level as the fraction of cells where the liquid mixing ratio is positive in the 3D field. The three parameters needed to characterize cloud geometry for SPARTACUS can also be estimated directly from the LES fields.

The overlap parameter can be computed from a 3D cloud field between each pair of layers by inverting Equation (1). Vertical profiles of overlap diagnosed in the 35 LES scenes are illustrated in Figure 1a. Overlap is most often greater than 0.7, with an average value (over the scenes and the vertical levels) of 0.876. It shows relatively small variations on the vertical as well as between the different scenes. Inverting Equation (2) for the average α yields an average decorrelation length z_0 of around 189 meters, close to the values found by Neggers et al. (2011) in LES cumulus fields yet much smaller than the range reported by Hogan and Illingworth (2000), probably because of our smaller vertical resolution as hinted by the sensitivity analysis presented in their Table 1.

The FSD, that is, the ratio of in-cloud LWC horizontal standard deviation to mean in-cloud LWC is easily diagnosed in each layer of the LES 3D fields since the LWC horizontal distribution is directly accessible. Computed FSD profiles are illustrated in Figure 1b. Again, relatively small variations are observed as both height and scenes change. The FSD ranges from 0.3 to 1 with an average value of 0.7, in agreement with the literature (see e.g. Shonk et al. (2010)).

In 3D cloud fields, the true (resolution-dependent) cloud perimeter could be diagnosed in each layer. However, Schäfer et al. (2016) have shown that accounting for small-scale fluctuations of cloud edges leads to an overestimation of the radiatively effective perimeter and hence of 3D effects. They advocate the use of a cloud perimeter corresponding to the perimeter of an ellipse fitted to the cloud. Following this recommendation, the total cloud perimeter is computed in each layer as follows: for each labeled cloud in the layer, the length of the semi-major axis of the fitted ellipse is taken as the maximum distance between the cloud geometric barycenter and any cell that belongs to the cloud. The area of the ellipse is taken as the cloud area. The perimeter of the ellipse is then computed from its area and the length of its semi-major axis. The individual ellipse perimeters are summed to obtain the total radiatively effective cloud perimeter and to derive C_s by inverting Equation (3). Vertical profiles of diagnosed C_s are illustrated in Figure 1c. C_s ranges from 50 to 600 meters with some variability both in height and between the different cloud fields, with an average value of 249 m. They are slightly smaller than those found by Hogan et al. (2016) and Fielding et al. (2020) in the I3RC LES cumulus cloud field of Hinkelman et al. (2005). Their simulation is also based on the ARMCu case, with the same forcings and domain size, but their larger resolution of (67 m)$^2\times$ 40 m explains the differences.

3 Parametric exploration of SPARTACUS

This section presents a parametric exploration of the SPARTACUS parameterization of 3D radiation. Can we find a set of cloud geometry parameters for which SPAR-

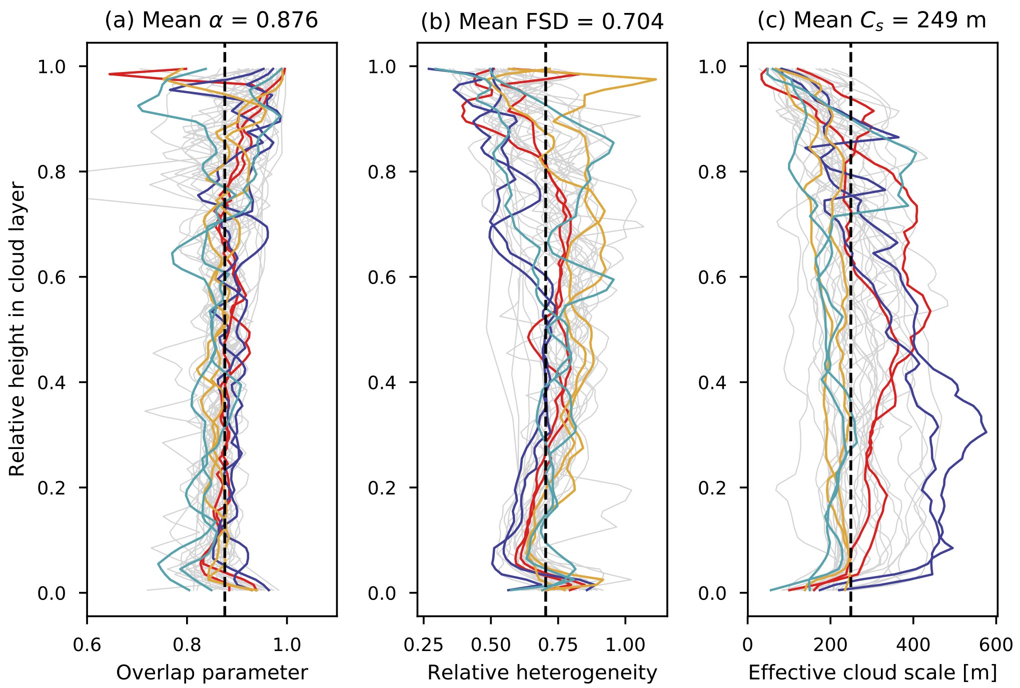

Figure 1. Vertical profiles of the three geometric parameters, scaled on the cloud layer depth (height 0 is the base of the cloud layer, height 1 is the top of the cloud layer). Gray and colored curves are for individual cloud scenes (colored curves are the fields used for calibration) and dashed black line is the average value over all cloud scenes and heights.

TACUS predictions of CREs lie within a reasonable distance from reference Monte Carlo estimates of the same quantities? How accurately can a unique configuration of SPARTACUS predict different radiative metrics computed in a large sample of cumulus fields under various illumination conditions? Does the best choice for cloud parameters match the LES-derived values of Section 2? The High-Tune:Explorer calibration tool (Couvreux et al., 2020; Hourdin et al., 2020) is used in the following to answer these questions. The tool is fully described in Part I. We give here only the information needed to understand the calibration procedure, before presenting the results.

3.1 Setup of High-Tune:Explorer

High-Tune:explorer (htexplo) is a statistical tool that automatically explores the behaviour of a model throughout an arbitrarily large parameter space. It is based on Gaussian process surrogates and implements history matching to reduce the parameter space to a sub-space of parameter vectors, or model configurations, that are "acceptable" in view of a given set of predetermined reference targets. The tool automatically performs most of the computations but the results crucially depend on the choices made for the calibration setup: the parameters to adjust, the metrics that will measure the model quality, the reference target and its associated uncertainty, and the uncertainty associated with the model structural error.

The SPARTACUS parameters that enter the calibration process are the three parameters described in Section 2.1: the overlap vertical decorrelation length z_0; the fractional standard deviation of the horizontal distribution of in-cloud liquid water FSD; and the cloud scale C_s. The parameter ranges that define the original parameter space $\mathcal{P}$ (a 3D space formed by the cartesian product of parameter ranges) were determined from numerical stability constraints in ecRad and other calibration experiments (not shown) in which larger ranges of parameter values were explored without adding value to the calibration exercise. Finally,

- z_0 ranges in [50, 500] (mean LES-derived value : 189 m)
- FSD ranges in [0.1, 2] (mean LES-derived value : 0.704)
- Cs ranges in [50, 1000] (mean LES-derived value : 249 m)

Three metric types were used in the calibration of SPARTACUS, all based on solar fluxes horizontally averaged over the LES domain: the reflected flux at the TOA $F_t^{\uparrow}$; the total absorbed flux in the atmosphere F_{abs} and the atmospheric radiative effect measured at the surface which is the difference between downward flux at TOA and downward flux at the surface, $F_t^{\downarrow} - F_s^{\downarrow}$. For each of these fluxes, three solar angles are used to explore the different mechanisms that drive the radiative effect of clouds under different illumination conditions. These angles were chosen arbitrarily: 0, 44 and 77 degrees from zenith. Each of these nine metrics (three fluxes × three solar angles) are computed in eight different cloud fields selected among the 35 available cumulus fields described in Table 2. These eight scenes, illustrated in Figure 2, were chosen for their contrasting characteristics to enable us to explore the distribution of available cumulus fields.

The reference values used as targets for these 72 metrics (nine metrics × eight cloud fields) are provided by the Monte Carlo model described in Section 2. The associated uncertainty is taken as the standard deviation of the MC estimate, typically smaller than 0.1%.

The structural error of SPARTACUS is unknown. In a sense, it is the error that would remain after the parameters are well calibrated. However, its characterisation is a prerequisite to the calibration process, as it prevents the tool from rejecting configurations that predict metric values within the structural error around the reference target. We hence rather use the term "tolerance to error": an acceptable distance between the parameterization estimate and the reference target, arbitrarily set by the modeler. Here, it is set as the third quartile of the distributions of relative errors between MC and SPARTACUS runs using the mean LES-derived parameter values, for each type of metric and solar angle:

- for the atmospheric radiative effect at the surface ($F_t^{\downarrow}$ - $F_s^{\downarrow}$), the relative tolerances to error are 3% for SZAs 0 and 44, and 4% for SZA 77
- for the absorbed flux in the atmosphere, the relative tolerances to error are 1%, 2% and 4% respectively for SZAs 0, 44 and 77
- for the reflected fluxes at SZAs 0, 44 and 77, the relative tolerances to error are set to 6%, 3% and 4% respectively.

Once this setup is fixed, the htexplo tool automatically computes the following:

1. An "experimental design" is built by sampling a small number of points (around ten times n, here 45 at the first iteration and 80 in the following ones) in the parameter space. A maximin Latin Hypercube sampling method is used that maximizes the minimum distance between samples (Williamson, 2015). ecRad is run for the sampled configurations on the eight selected cloud scenes with the three selected solar zenith angles.

2. The chosen metrics are computed from the model outputs and used as a training set in the construction of emulators (one per metric, each based on a Gaussian Process). These fast surrogate models are then used to compute estimates (the expectation of the process) for the metrics on a large sample of points in the parameter space (here, 10^5), along with the associated statistical uncertainties (the standard deviation of the process).

3. For each sampled parameter vector $\check{\lambda}$, the distance between emulated and reference values is computed for each metric f (f_k is the kth metric). The samples where this distance is larger than a threshold for at least one of the N_{met} metrics are removed from the parameter space. The new parameter space is called the Not-Ruled-Out-Yet (NROY) space. In htexplo, this distance, called the *implausibility* $I_f(\check{\lambda})$, is defined as follows:

$$I_f(\check{\lambda}) = \frac{|r_f - \mathbf{E}[f(\check{\lambda})]|}{\sqrt{\sigma_{r,f}^2 + \sigma_{d,f}^2 + \sigma_f(\check{\lambda})^2}} \tag{4}$$

where $\mathbf{E}[f(\check{\lambda})]$ is the emulator estimate, r_f is the reference value, $\sigma_{r,f}$ is the uncertainty associated with the reference, $\sigma_f(\check{\lambda})$ is the statistical uncertainty associated with the emulator estimate and $\sigma_{d,f}$ is the model structural error. The implausibility threshold for rejecting points from the parameter space was set to three. This means that points were kept in the parameter space only if the distance between SPARTACUS and MC was closer than three standard deviations (according to all three uncertainties) for each of the 72 metrics.

4. A new experimental design is built from a sub-sample of the parameter vectors that were not rejected at the previous step, and the whole procedure is repeated until the NROY space converges. With each iteration, called "wave", the uncertainties associated with the emulators decrease until convergence, since the sampling of model configurations that serve to build the emulators is denser (the parameter space is smaller and the number of sampled points is unchanged).

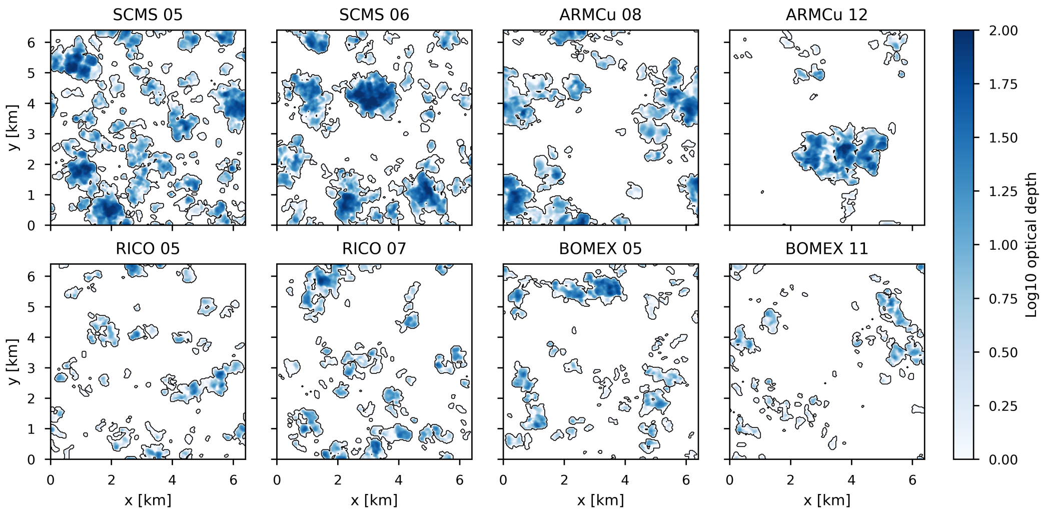

Figure 2. Maps of optical depths for the eight selected scenes. The shading uses a logarithmic scale and the black lines are the zero contours. The optical depth was estimated from the liquid water path field of the LES.

3.2 Reduction of the parameter space and global sensitivity analysis

Thirteen iterations were applied, reducing the NROY space from 11.7% of the original space after the first wave, to 8.40% after the twelfth wave and 8.39% after the thirteenth wave, where the process was assumed to have reached convergence. It would have been possible to further reduce the NROY space by decreasing the rejection threshold or adding new constraints (new metrics), as was done in Couvreux et al. (2020) and Hourdin et al. (2020). However, the aim of this study is not to determine a unique set of acceptable parameters but rather to analyse the structure of the parameter space and compare various configurations that are acceptable given the arbitrarily chosen tolerance.

Figure 3 illustrates the parametric dependency of the downward flux at the surface under the ARMCu 8th hour clouds at SZA 0° and 77° (two of the 72 metrics used in the calibration), obtained from the ∼1000 SPARTACUS configurations explored during the thirteen waves of history matching.

Large surface fluxes at high sun are only obtained when clouds are sufficiently heterogeneous (when FSD is large enough; Figure 3a), while the effect of heterogeneity in grazing sun conditions is less obvious (Figure 3d). The transmitted flux at 0° is strongly related to the decorrelation length (Figure 3b), but the transmitted flux at 77° does not seem driven by this parameter (Figure 3e). Indeed, when the decorrelation length increases, the overlap gets closer to maximum and the total cloud cover decreases. This leads to more energy reaching the surface, in particular for high sun. As the sun gets closer to the horizon, it is not the total cloud cover that matters but the effective cloud cover, projected in the direction of the sun, to which cloud sides contribute largely. At high sun, 3D effects (inversely proportional to cloud size C_s) lead to an increase in surface flux (Figure 3c), a signature of escape of light from cloud sides and entrapment. At low sun they lead to a decrease in surface flux (Figure 3f), explained by the interception of light by cloud sides. In multi-layered cloud scenes or with larger ground albedo, the entrapment effect would be stronger and the balance between positive and negative 3D effects as a function of SZA could be affected (entrapment leads to an increase of surface flux at all solar angles; Hogan et al. (2019)).

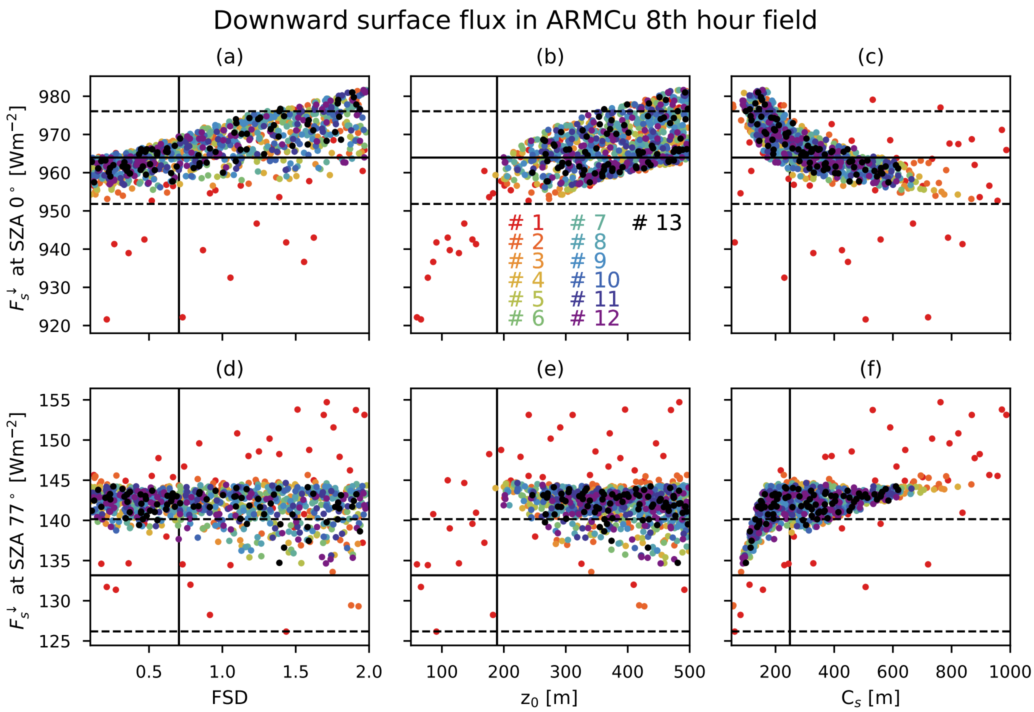

Figure 3. Downward flux at surface for various ecRad runs, as a function of three parameter values: (a,d) fractional standard deviation FSD, (b,e) overlap decorrelation length z_0, (c,f) cloud scale C_s, and of solar zenith angle: (a-c) 0° and (d-f) 77°. Full black horizontal lines represent the Monte Carlo reference value, dashed horizontal lines represent the tolerances to error. Full vertical lines represent the mean parameter value diagnosed in the LES. Different colors represent parameter sets sampled at different waves.

Metrics computed at different iterations in the calibration process are represented in different colors in Figure 3, showing that part of the parameter ranges are no longer sampled after a certain number of waves. For instance, after the first wave (red points), decorrelation length values smaller than ∼180 m have been excluded from the parameter space, independently of the values of the other two parameters. This is because for this subrange of decorrelation length values (in which the cloud cover is large) the 0° surface flux emulator predicts values that are too small compared to the MC estimate.

The implausibility matrix presented in Figure 4 reveals the structure of the NROY space obtained after the thirteenth wave. A large number of points was sampled in the original 3D space parameter, and the largest implausibility computed over the 72 metrics was associated with each sampled point, thereby building a unique implausibility cube. The NROY space corresponds to the regions of this cube filled with values smaller than 3 after wave 13 (note that points exceeding 3 at earlier waves have their implausibilities fixed at the value of their first excedence (when they were ruled out). To visualize the information contained in this cube, it is successively projected along each of the three dimensions to produce three 2D maps. The upper triangle of Figure 4 displays projections of the number density of points belonging to the NROY space. The lower triangle displays projections of the implausibility values, by taking the minimum value along the reduced dimension. The upper triangle gives the density of acceptable configurations, while the lower triangle informs on the quality of the "best" configurations.

The gray (red) zones in the upper (lower) triangle subplots represent the regions of the parameter space where no configuration is acceptable given the two parameter values that correspond to the pixel, whatever the value of the third parameter. For instance, the upper-left and lower-right subplots show that small values of the decorrelation length have been rejected, independently of the values of the other two parameters. This was already illustrated in Figure 3. Here, the plots additionally show that the set of parameter values derived from the 3D LES cloud fields do not belong to the NROY space of the thirteenth wave, in particular due to too small value of the FSD and/or of z_0.

On the upper-right subplot, we see that many (FSD, C_s) pairs have been rejected. The pairs that lead to acceptable configurations of the parameterization are cleary identified: small values of C_s are paired with large values of FSD and conversely (although very large values of C_s were all rejected). This means that an increase in heterogeneity can be compensated by a decrease in cloud size (more intense 3D effects), and that the uncertainties associated with the target metrics do not allow to determine which mode should be favored between small heterogenous or large homogeneous clouds.

The variations of implausibility in the parameter space reveal more of the parameterization behaviour than the implausibility absolute values, which are highly dependent on the arbitrarily set tolerance to error (see Equation 4). However, the subplots of the lower triangle show that for any configuration, there is always at least one metric that is farther away from its target than 1.5 times the root square sum of its uncertainties, which is dominated by the tolerance to error at wave thirteen. They also show that the best configurations have small heterogeneous clouds rather than large homogeneous ones, associated with large decorrelation lengths.

3.3 Evaluation of flux estimates in calibrated configurations

The various configurations that were sampled to construct emulators from true ecRad runs are evaluated using scores associated with each metric and configuration. It is the error between ecRad and the reference MC divided by the tolerance to error. For each SPARTACUS simulation run during waves three to thirteen, the RMS scores are computed over all metrics ("global"), and over reflected fluxes ("TOA up"), absorbed fluxes ("absorption") and surface fluxes ("surface down") separately. Then, the configurations with smallest RMS scores of each category are selected as "best" configurations. They

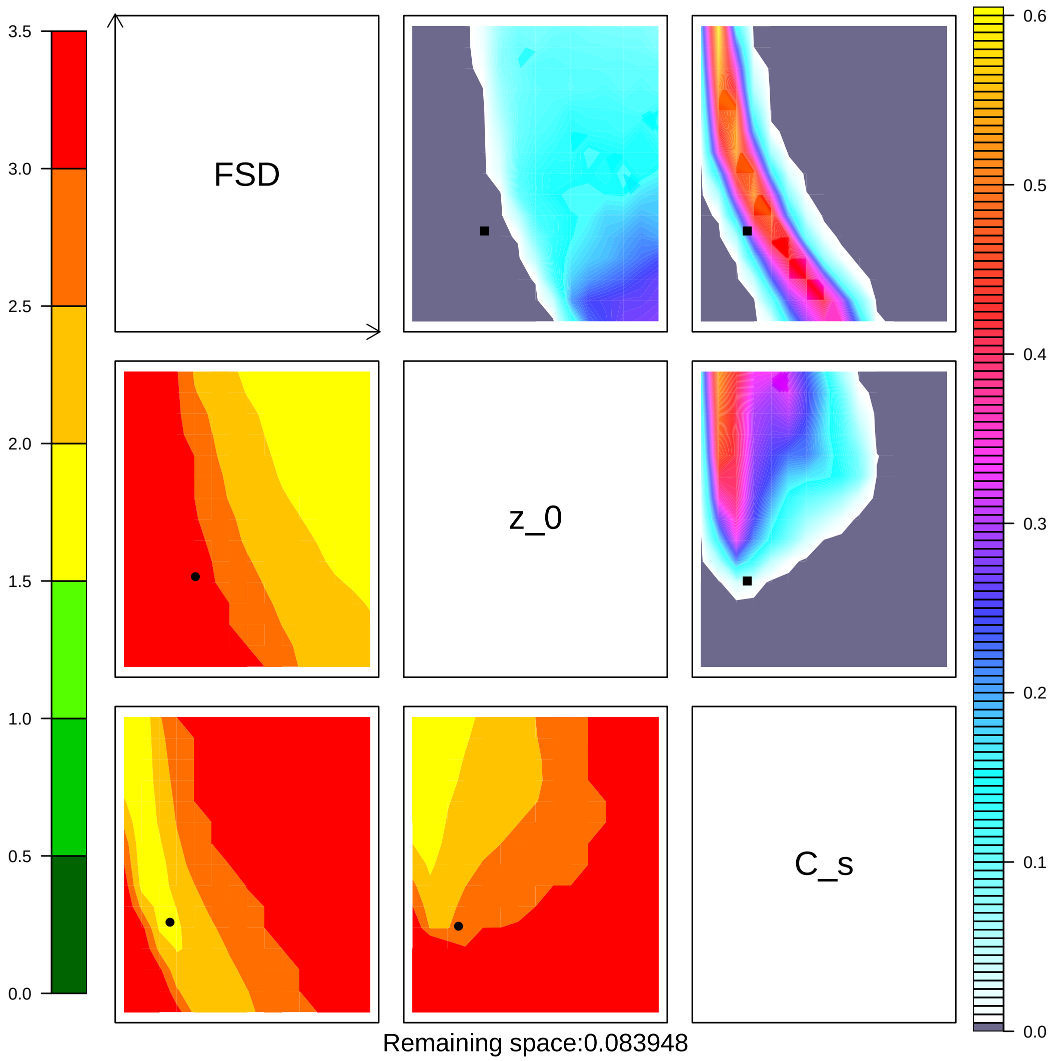

Figure 4. Visualization of the implausibility cube: NROY space density (upper triangle) and minimum implausibility (lower triangle) at wave 13. Implausibility is computed as the maximum over the metrics. Axes of the upper-triangle subplots are given by the parameter names on the diagonal. The (x,y) axes of the subplots are: (z_0,FSD) in row one, column two; (C_s,FSD) in row one, column three; (C_s,z_0) in row two, column three. The axes of the lower-triangle subplots are the same as the axes of their symmetric subplot in the upper triangle. Black dots correspond to the average parameter values derived from the LES cloud fields.

Table 3. Parameter values for the "best" configurations of ecRad

Parameters	FSD	z_0 [m]	C_s [m]
Mean LES-derived	0.705	187	247
Best global	1.079	436	155
Best TOA up	1.646	493	119
Best absorption	0.102	294	821
Best surface down	1.469	374	113

are presented in Table 3, along with the mean LES-derived parameters. Configurations that lead to best upward TOA and best downward surface fluxes are relatively similar, favoring small heterogeneous clouds. The configuration that leads to the better estimates of absorbed fluxes rather favors large homogeneous clouds. The configuration that leads to best global RMS is in between these two modes, but still selects smaller more heterogeneous clouds than found in the LES. The overlap decorrelation length parameter is always greater than the one diagnosed in the 3D cloud fields, yielding smaller cloud covers.

These four new configurations, obtained from a calibration process using only eight cloud fields and three solar angles, were tested on the 35 cloud fields of Section 2.3 and 11 solar zenith angles from 0 to 77 with step 11. The distributions of errors are represented in Figure 5. The RMSEs are given in the legends for each configuration. These numbers are different from the configuration scores as they are not divided by the tolerance to error. The configuration using the mean geometry parameters computed from the LES cloud fields is also represented.

The fluxes at TOA and surface are systematically improved compared to the configuration using the LES-derived parameter values, but all calibrated configurations are slightly worse for the absorption. The absorption bias associated with the "Best global", "Best TOA" and "Best surface down" configurations, which all have small heterogenous clouds, is always negative. It appears that most of the flux that should have been absorbed reaches the surface, inducing a positive mean bias in the transmitted fluxes.

To understand why small heterogeneous clouds lead to wrong estimates of the absorption, the CRE on absorption is analysed for various configurations as a function of SZA in one particular cloud field. Figure 6(a) shows results from the Monte Carlo, "Best surface down" and "Best absorption" simulations. Three sensitivity tests were performed, changing one parameter at a time, from the value corresponding to the "Best surface down" configuration to the value corresponding to "Best absorption" configuration, and keeping the two other parameters to the "Best surface down" values. Results are shown in Figure 6(b).

The "Best surface down" simulation with large 3D effects and important heterogeneity accurately reproduces the absorption dependency to solar angle but with a negative bias of 2 to 4 Wm^{-2}. Two-stream errors in plane-parallel homogeneous clouds absorption are around -4 Wm^{-2} on average (see Appendix A), which could explain the negative bias observed in the "Best surface down" simulation. The "Best absorption" simulation is closer to the Monte Carlo reference at small SZAs, which seems to result from the compensating effects of reduced inhomogeneity (smaller FSD increases absorption), reduced 3D effects (larger C_s decreases absorption) and structural errors independent from cloud geometry (e.g. from the two-stream approximation). At SZA 77°, reducing 3D effects increases the absorption which almost entirely cancels out the structural error. It thus appears that the "Best absorption" configuration yields correct absorption estimates for the wrong reasons, that is, the wrong radiative processes.

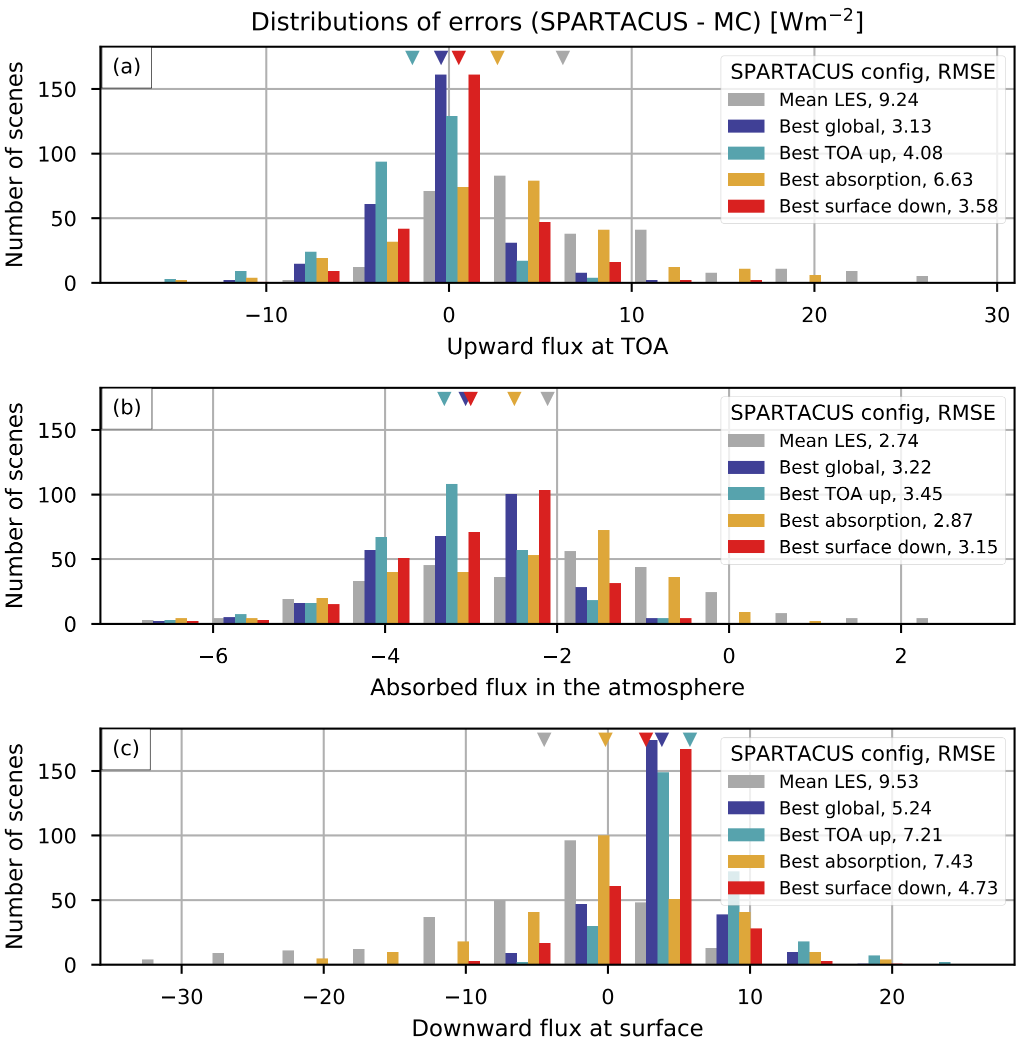

Figure 5. Histograms representing the distributions of differences between ecRad and Monte Carlo estimates for the three metrics: (a) upward flux at TOA, (b) absorbed flux in the atmosphere and (c) downward flux at the ground. Errors for all 35 cumulus scenes and 8 solar angles (from 0 to 77 with step 11 degrees) are distributed together. Each color corresponds to a different configuration of ecRad. The parameters values for each configuration are given in Table 3. Color triangles represent the mean error. The root mean square distances (RMSE) are given in the legends.

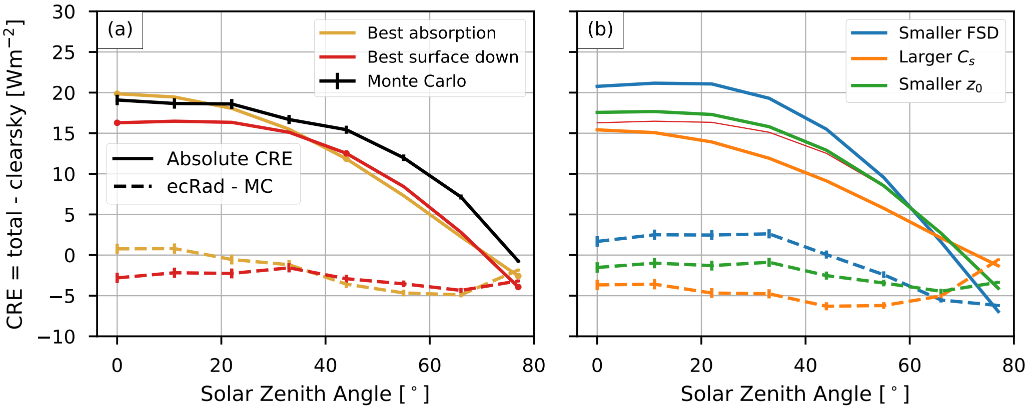

Figure 6. Absorption CRE in the SCMS 5th hour cloud field. Solid lines represent absolute values of the CRE and dashed lines represent errors compared to the MC reference. (a) shows the MC reference and two calibrated configurations ("Best surface down": FSD=1.469, z_0=374 m, C_s=113 m, and "Best absorption": FSD=0.102, z_0=294 m, C_s=821 m). (b) shows sensitivity tests, with two parameter values as "Best surface down" and one as "Best absorption": "Smaller FSD": FSD=0.102, z_0=374 m, C_s=113 m; "Larger C_s: FSD=1.469, z_0=374 m, C_s=821 m, and "Smaller z_0": FSD=1.469, z_0=294 m, C_s=113 m.

4 Discussion and outlook

The htexplo software enables efficient semi-automatic tuning for any aspect of a climate model. The "automatic" aspects of the tuning involve implementation of well developed techniques from the uncertainty quantification and machine learning communities: using Gaussian processes to quickly locate those regions of parameter space that are compatible with reference data sets. Yet htexplo is not a black box tuning software, and is designed as a tool to be harnessed by an expert physicist to assist with tuning. The physicist must define the parameters, metrics, references and their tolerances to error (if the structural error of the model is unknown). They must interpret the results and then perhaps adapt the tuning (introducing new metrics, adjusting tolerances, recognise compensating errors etc): htexplo cannot, alone, measure the quality of a model. In this section, we first discuss the choices that conditioned the calibration procedure of Section 3, and then some implications of the main results of our work.

A fundamental aspect of the tuning strategy advocated in this series of papers is that sources of errors related to different aspects of the model can be disentangled, while extending the modeler's capacity of analysis and level of comprehension. In Part II, this is achieved by performing a first calibration step using well-understood study cases in the LES/SCM framework to constrain the parameter space to values compatible with process-based metrics, before tuning the 3D global model. It is a way to ensure that the cloud radiative effect targeted in the 3D calibration is obtained for the right clouds. The focus of this paper (Part III) is on getting the right cloud radiative effect for the right radiative transfer. This is achieved by offline calibration of the radiative transfer parameterization, in which the cloud fraction and LWC profiles input to ecRad are computed directly from the 3D cloud fields that are acted upon by the reference MC model, instead of being parameterized.

We went one step further in our effort to disentangle potential sources of errors, by separating internal aspects of the radiative parameterization. Our choice of reference has determined the aspects of the parameterization that were allowed to compensate each other. On the one hand we chose to exclude the question of the representation of optical properties of clouds in order to focus on the representation of transport and cloud geometry, by computing the reference MC estimates using the same optical properties as ecRad. On the other hand, we chose to allow internal compensating errors between cloud geometry and pure radiative transfer by targeting MC simulations that use detailed Mie phase function instead of a delta-scaled two-stream version of the MC model. We also chose to calibrate all the geometry parameters together, although 1D geometrical effects could have been treated independently from the effects of horizontal transport. These choices were primarily driven by the inextricable aspect of the light transport formulation and the treatment of cloud geometry effects in the SPARTACUS radiative transfer model.

The choice of metrics is also a crucial aspect of the calibration setup. Here, we have used three metrics that are not independent from each other: the (known) incoming flux at the TOA is entirely distributed into reflected, absorbed by the atmosphere and absorbed by the surface fluxes. However, adding a metric that is a combination of the other two further constrains the parameters when each metric tolerance to error is smaller than the sum of the tolerances associated with the other two metrics.

The value of the tolerances to error for the different metrics were chosen here so as to reject the SPARTACUS configurations that are much less accurate than using the LES-derived parameter values. Other choices could have been made such as using a bulk value corresponding for instance to the tolerance of a climate model to local radiation errors, or to the radiation error that would result from a perturbation of the cloud fraction profiles typical of the errors found in cloud parameterizations. The results of the calibration are sensitive to the tolerance to error, therefore it should be set carefully, in concordance with the objectives of the tuning exercise. We also note that error tolerance can (and should) be adapted throughout a tuning exercise. We may find that our tolerances were too small, the model could not get close enough to the reference metrics, and the whole parameter space is ruled out. We should then increase our tolerance to error. We may also find our tolerances are too large (if we were being conservative at the beginning of the exercise), and that many of the models compatible with those tolerances are, in fact, poor relative to others in our later waves. It could be argued that adapting tolerance to error by observing the results of each wave will lead this tolerance to converge towards the true structural error. But the "true" structural error is not trivial to define; it is a modeler's judgement and likely has complex dependencies across metrics. It could be thought of as the error that remains once the parameters have been adjusted to remove parametric errors, but we have seen here that this "best" adjustment depends on the chosen metrics. Even if a model could be reduced to only one metric, say the absorption, the definition of the structural error would still depend on the modeler's appreciation: is it preferable to produce the best possible absorption estimates even if the representation of internal processes seems wrong as in the "best absorption" configuration? Or would we rather have a model that behaves slightly worse but for a more physical representation of the processes? In this example, the structural error of the former model would be smaller than that of the latter. With htexplo, we provide a framework within which modelers become able to continuously question, define, learn and explore the structural error of their model.

Beyond its implication for the calibration of SPARTACUS, the fact that the "best" parameters selected by htexplo do not match the LES-derived parameters questions the conceptual constraints that surround climate model development and tuning. The main goal of parameterization development is to derive functional forms that can be trusted to provide accurate source terms for the explicitly resolved variables of the model over a wide range of atmospheric regimes (including regimes that have not been observed yet but might appear in different climates). To achieve this, it is essential to base our developments on our understanding of physical processes. However, we argue that some flexibility should be allowed in the choice of parameter values. Results reported by Bastidas et al. (2006) and Hogue et al. (2006) also support this idea. They show that free parameters should be set to different values through different land surface models even though their physical interpretation is the same. Their conclusions were limited to so-called "functional" parameters that cannot be associated with direct measurements. We argue that observational constraints on "physical" parameters should also be alleviated. Indeed, it is an *effective* value of the parameters that is needed in the models. The calibration strategy advocated here is fundamentally a way to determine possible values for these effective parameters. An effective value laying too far from observations (e.g. outside the distribution of observed values) could however indicate that the physical images that supported the parameterization development are wrong or that important processes are missing.

Eventually, the improvement of SPARTACUS was obtained by calibrating a mean parameter, thereby neglecting parameter variations with height and between cloud scenes. This was probably only possible because all cloud fields used here represent cumulus clouds, with relative resemblance between the cases, although both marine and continental clouds were represented. An interesting follow-up would be to repeat this exercise with other cloud types, starting with other boundary-layer clouds such as stratocumulus and transition scenes involving both cloud types. A possible diagnosis of htexplo might then be that a single parameter is not able to represent different clouds. This would mean that a sub-parameterization should be developed to make this parameter depend on atmospheric conditions. Such parameterizations exist for example to predict cloud perimeter length in Fielding et al. (2020), or the degree of overlap in e.g. Sulak et al. (2020). Other parameters appear in these formulations, which can in turn be calibrated using the same procedure as described in this work.

Appendix A Estimation of various sources of errors in ecRad

Various aspects of the radiation scheme were identified in this study as potential sources of errors in ecRad flux estimates. The first category (a) groups the approximate optical properties and approximate radiative transfer model, that is, the two-stream equations and the delta-scaling approximation. A second category (b) relates to the degree of complexity in the representation of cloud geometry and horizontal transport. A third category concerns the errors due to neglecting inter-level and inter-scene variability of the parameters that describe cloud geometry. The last category (d) is the choice of the absolute values for these parameters.

Errors related to these different aspects have been documented throughout the literature (see e.g. Barker et al. (2003, 2015) for categories (a) and (b)), although not always from the same metrics or clouds, which makes quantitative comparisons difficult. In our study, these different errors have been computed in a uniform way from various numerical experiments. Plane-parallel homogeneous clouds of several liquid water contents (yielding optical depths of 0.1, 0.25, 0.5, 1, 2.5, 5, 25, 50 and 100 at 800 nm) were used to estimate errors of category (a). The cumulus fields of Section 2.3 were used for the three other categories. Each plane-parallel and cumulus cloud field was combined into eight illumination conditions (sun at 0 to 77 degrees from zenith with step 11°). The results are displayed in Table A1.

In category (a), errors were diagnosed from different configurations of the Monte Carlo model: using detailed Mie data or the approximate SOCRATES model for cloud optical properties, and using detailed Mie phase function or the approximate Henyey-Greenstein (HG) phase function that only depends on the asymmetry parameter g, combined with the δ-Eddington approximation that is also used in ecRad. The difference between the Monte Carlo "as ecRad" (approximate optical properties and approximate phase function) and ecRad is interpreted as the error related to the two-stream model.

In category (b), errors were diagnosed from different configurations of ecRad using the 1D solver Tripleclouds (Shonk & Hogan, 2008) in plane-parallel homogeneous maximum overlap mode (PPH max ovp; by setting the overlap parameter to one and the heterogeneity parameter to zero) and in heterogeneous exponentional random mode (Tripleclouds (1D), by setting both the overlap and the heterogeneity parameters to LES-derived values), as well as the the 3D solver SPARTACUS with LES-derived parameters. These different ecRad estimates were compared to the reference Monte Carlo model using approximate optical properties and detailed phase functions (also used as the reference in Section 3). Figure A1 shows the error distributions for the various solvers.

In category (c), errors were diagnosed from the SPARTACUS solver parameterized with LES-derived values averaged along the vertical dimension (z-averaged) and on both the vertical dimension and the different cloud fields (case-z-averaged), compared to the SPARTACUS solver parameterized with scene-dependent profiles of parameters as derived from the LES.

In category (d), errors were diagnosed from various configurations of the SPARTACUS solver, with parameter values output from the calibration process.

Acknowledgments

This work received funding from grant HIGH-TUNE ANR-16-CE01-0010. It was supported by the DEPHY2 project, funded by the French national program LEFE/INSU and the GDR-DEPHY. Daniel Williamson was funded by NERC grant: NE/N018486/1 and by the Alan Turing Institute project Uncertainty Quantification of multi-scale and multiphysics computer models: applications to hazard and climate models as part of the grant EP/N510129/1 made to the Alan Turing Institute by EPSRC. The version of the code used in Part I and Part II is provided under doi:10.14768/70efa07b-afe3-43a4-8334-050354f9deac. The same code was used here but with additional scripts and data specific to ecRad. The complete version of htexplo that was used for this paper is publicly available via a Subversion through "svn checkout http://svn.lmd.jussieu.fr/HighTune". Note, however, that this tool is a new research tool, and, as such, is still evolving. Version 1.4.0 of the ecRad package is freely available under the terms of the Apache Licence Version 2.0. This paper uses a prior version (1.3.0) available upon request under the terms of the OpenIFS license. No major modification was made to the ecRad radiation code between versions 1.3.0 and 1.4.0. The source code of the Monte Carlo model that is used as a reference in this paper is available under the terms of the GNU General Public License. The scripts and data that were used to produce the figures and tables are available under doi:10.14768/medd-xa16. The authors would like to thank Robert Pincus and two anonymous reviewers for their very constructive feedbacks, as well as Igor Roffiac for extremely rich exchanges and sharing.

References

Barker, H. W., Cole, J. N. S., Li, J., Yi, B., & Yang, P. (2015). Estimation of errors in two-stream approximations of the solar radiative transfer equation for cloudy-sky conditions. *Journal of the Atmospheric Sciences*, *72*(11), 4053-4074. doi: 10.1175/JAS-D-15-0033.1

Barker, H. W., Stephens, G. L., Partain, P. T., Bergman, J. W., Bonnel, B., Campana, K., ... others (2003). Assessing 1d atmospheric solar radiative transfer models: Interpretation and handling of unresolved clouds. *Journal of Climate*, *16*(16), 2676–2699.

Bastidas, L. A., Hogue, T. S., Sorooshian, S., Gupta, H. V., & Shuttleworth, W. J. (2006). Parameter sensitivity analysis for different complexity land surface models using multicriteria methods. *Journal of Geophysical Research: Atmospheres*, *111*(D20). doi: https://doi.org/10.1029/2005JD006377

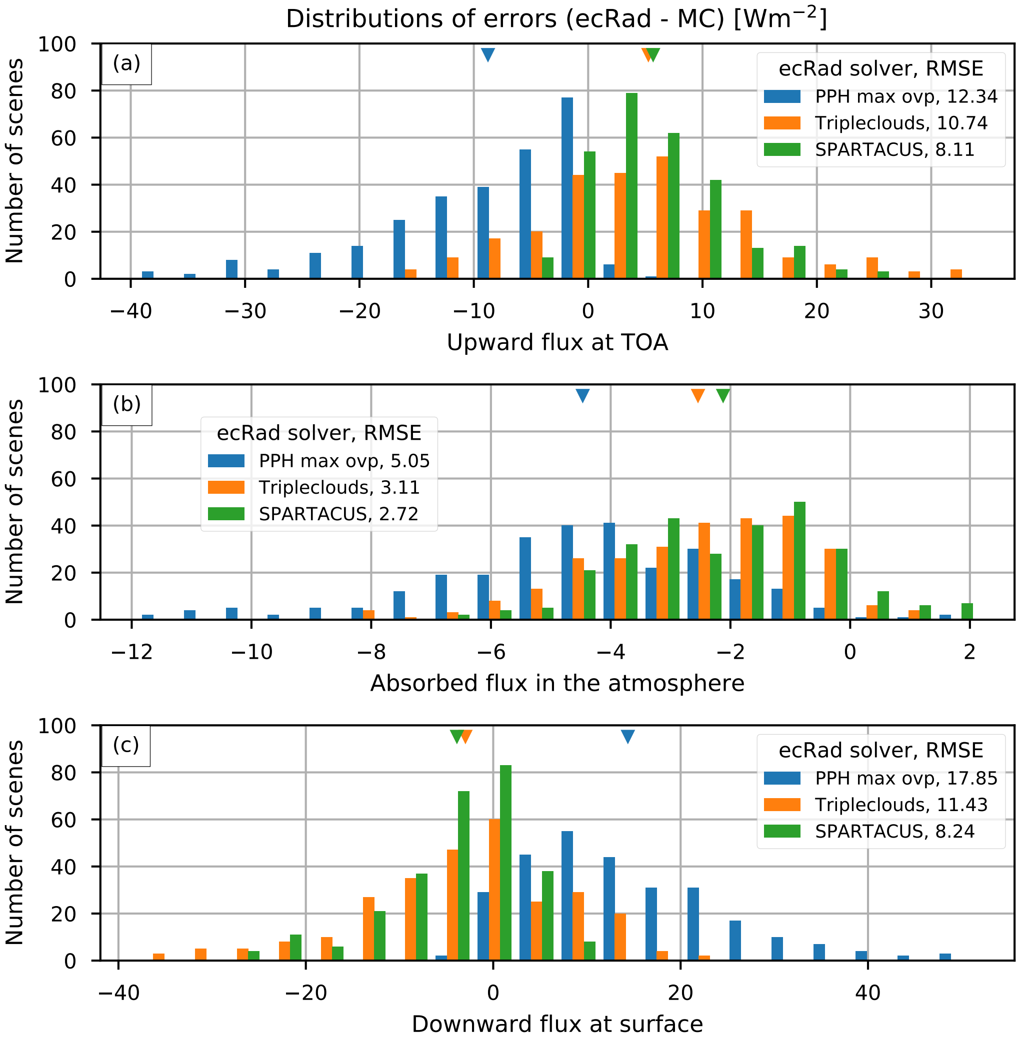

Figure A1. Histograms representing the distributions of differences between ecRad and MC estimates for the three metrics: (a) upward flux at TOA, (b) absorbed flux in the atmosphere and (c) downward flux at the ground. Each histogram represents the distribution of 280 data points: 35 scenes $\times$ 8 solar zenith angles (from 0 to 77 with step 11 degrees). Each color corresponds to a different configuration of ecRad. PPH max ovp corresponds to homogeneous clouds with maximum overlap and no 3D effects. Tripleclouds corresponds to heterogeneous clouds with FSD and α vertical profiles as diagnosed in the 3D LES field, without 3D effects. SPARTACUS is as Tripleclouds but with 3D effects, with C_s vertical profiles as diagnosed in the 3D LES fields. The mean error is represented by colored triangles. The RMSEs are given in the legends.

Table A1. Relative errors [%] for different aspects of the ecRad radiative transfer scheme.

(i) Experiment	(ii) Reference	(iii) Model	(iv) TOA up RMS	(iv) TOA up bias	(v) Absorbed RMS	(v) Absorbed bias	(vi) Surf. down RMS	(vi) Surf. down bias
(a) Experiments in plane parallel homogeneous clouds								
(1) SOCRATES	MC exact	MC SOCRATES	1.1	0.8	10.8	-10.5	1.9	-1.5
(2) δ-Eddington	MC exact	MC δ-Eddington	11.1	4.6	13.1	4.5	11.8	4.6
(3) two-stream	MC as ecRad	ecRad two-stream	5.4	0.3	15.0	-9.1	4.3	-2.2
Transport (2+3)	MC SOCRATES	ecRad two-stream	8.3	3.2	15.2	-4.4	7.0	1.0
Total (1+2+3)	MC exact	ecRad two-stream	8.6	4.1	17.8	-14.5	5.8	-0.7
(b) Experiments in cumulus, MC vs ecRad 1D and 3D solvers, parameters $\lambda = (\alpha, \mathrm{FSD}, C_s)$								
PPH max ovp	MC SOCRATES	1D, $\lambda = (1, 0, \infty)$	23.4	-20.9	54.2	-53.5	28.6	-27.0
Tripleclouds (1D)	MC SOCRATES	1D, $\lambda(z, case)$ LES	29.3	23.0	23.8	-18.9	23.7	15.1
SPARTACUS (3D)	MC SOCRATES	3D, $\lambda(z, case)$ LES	22.7	20.0	20.0	-10.4	18.3	14.4
(c) Experiments in cumulus, ecRad SPARTACUS, with LES-derived profiles vs averaged parameters								
z-averaged	$\lambda(z, case)$ LES	$\overline{\lambda}(case)$ LES	1.4	-0.1	1.6	-0.4	1.4	-0.1
case-z-averaged	$\lambda(z, case)$ LES	$\overline{\overline{\lambda}}$ LES	3.7	0.6	3.5	-0.4	3.6	0.4
(d) Experiments in cumulus, MC vs ecRad SPARTACUS with calibrated parameters (see Section 3)								
Best global	MC SOCRATES	$\overline{\overline{\lambda}}$ from htexplo (see Table 3)	8.3	-2.7	29.1	-28.1	10.2	-7.2
Best TOA up	MC SOCRATES	$\overline{\overline{\lambda}}$ from htexplo (see Table 3)	11.3	-8.5	33.3	-32.6	14.4	-12.8
Best absorption	MC SOCRATES	$\overline{\overline{\lambda}}$ from htexplo (see Table 3)	17.9	12.0	22.1	-18.8	14.9	6.3
Best surface down	MC SOCRATES	$\overline{\overline{\lambda}}$ from htexplo (see Table 3)	9.2	-0.4	28.0	-26.8	9.6	-5.1

For each pair of reference computation (ii) / test approximation (iii), errors on the cloud radiative effects on TOA upward (iv), absorbed (v), and surface downward (vi) fluxes are quantified. For each column, the RMS and mean bias are first computed independently for each solar angle over the different cases, then RMS and mean bias are weighted by the cosine of the solar angle, and averaged over the 8 SZAs. Only data points where reference CRE > 2 Wm^{-2} are used to avoid division by zero. Only solar angles where at least 9 data points were available are used in the cosine-weighted average. The table subsections concern: (a) errors related to non-geometrical effects of clouds, (b) ecRad errors for different solvers, with increasing complexity in the representation of geometrical effects, (c) errors related to the neglect of parameters variations with height and cloud field, (d) ecRad errors for different choices of cloud-geometry parameters, output from the calibration exercise of Section 3.

CRE = total sky - clear sky. Relative error $r = 100\times$(model-ref)/ref. RMS = $\sqrt{\langle r^2 \rangle_{fields}}$. bias=$\langle r \rangle_{fields}$
MC exact: detailed Mie optical properties and phase function.
MC SOCRATES: parameterized optical properties and detailed Mie phase function.
MC δ-Eddington: detailed Mie optical properties and HG δ-Eddington phase function.
MC as ecRad: parameterized optical properties and HG δ-Eddington phase function.

Berg, L. K., Kassianov, E. I., Long, C. N., & Mills, D. L. (2011, January). Surface summertime radiative forcing by shallow cumuli at the Atmospheric Radiation Measurement Southern Great Plains site. *Journal of Geophysical Research-Atmospheres*, *116*, D01202. (WOS:000286058200002) doi: 10.1029/2010JD014593

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., ... Webb, M. J. (2015, April). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268. doi: 10.1038/ngeo2398

Brient, F., Couvreux, F., Villefranque, N., Rio, C., & Honnert, R. (2019). Object-oriented identification of coherent structures in large eddy simulations: Importance of downdrafts in stratocumulus. *Geophysical Research Letters*, *46*(5), 2854-2864. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL081499` doi: 10.1029/2018GL081499

Brown, A., Cederwall, R., Chlond, A., Duynkerke, P., Golaz, J.-C., Khairoutdinov, M., ... Stevens, B. (2002). Large-eddy simulation of the diurnal cycle of shallow cumulus convection over land. *Q. J. R. Meteorol. Soc.*, *128*, 1075–1093.

Cahalan, R. F., Oreopoulos, L., Marshak, A., Evans, K. F., Davis, A. B., Pincus, R., ... Zhuravleva, T. B. (2005, September). The I3RC: Bringing Together the Most Advanced Radiative Transfer Tools for Cloudy Atmospheres. *Bulletin of the American Meteorological Society*, *86*(9), 1275–1293. doi: 10.1175/BAMS-86-9-1275

Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N., ... Xu, W. (2020). Process-based climate model development harnessing machine learning: I. A calibration tool for parameterization improvement. *Journal of Advances in Modeling Earth Systems*. doi: https://doi.org/10.1029/2020MS002217

Dufresne, J.-L., & Bony, S. (2008). An assessment of the primary sources of spread of global warming estimates from coupled atmosphere–ocean models. *Journal of Climate*, *21*(19), 5135-5144. doi: 10.1175/2008JCLI2239.1

Fielding, M. D., Schäfer, S. A. K., Hogan, R. J., & Forbes, R. M. (2020). Parametrizing cloud geometry and its application in a subgrid cloud-edge erosion scheme. *Quarterly Journal of the Royal Meteorological Society*, *146*(729), 1651-1667. doi: 10.1002/qj.3758

Gristey, J. J., Feingold, G., Glenn, I. B., Schmidt, K. S., & Chen, H. (2020). Surface solar irradiance in continental shallow cumulus fields: Observations and large-eddy simulation. *Journal of the Atmospheric Sciences*, *77*(3), 1065 - 1080. doi: 10.1175/JAS-D-19-0261.1

Hinkelman, L. M., Stevens, B., & Evans, K. F. (2005). A large-eddy simulation study of anisotropy in fair-weather cumulus cloud fields. *Journal of the Atmospheric Sciences*, *62*(7), 2155-2171. doi: 10.1175/JAS3463.1

Hogan, R. J., & Bozzo, A. (2018). A flexible and efficient radiation scheme for the ECMWF model. *Journal of Advances in Modeling Earth Systems*. 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, July). Representing 3-D cloud radiation effects in two-stream schemes: 2. Matrix formulation and broadband evaluation. *Journal of Geophysical Research: Atmospheres*, *121*(14), 2016JD024875. 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

Hogue, T. S., Bastidas, L. A., Gupta, H. V., & Sorooshian, S. (2006). Evaluating model performance and parameter behavior for varying levels of land surface model complexity. *Water Resources Research*, *42*(8). doi: https://doi.org/10.1029/2005WR004440

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., … Williamson, D. (2017). 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., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., … Volodina, V. (2020). Process-based climate model development harnessing Machine Learning. Part II: Model Calibration from Single Column to Global. *Journal of Advances in Modeling Earth Systems*. doi: https://doi.org/10.1029/2020MS002225

Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., & Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Calculations with the aer radiative transfer models. *Journal of Geophysical Research: Atmospheres*, *113*(D13). doi: 10.1029/2008JD009944

Joseph, J. H., Wiscombe, W. J., & Weinman, J. A. (1976, December). The Delta-Eddington Approximation for Radiative Flux Transfer. *Journal of the Atmospheric Sciences*, *33*(12), 2452–2459. doi: 10.1175/1520-0469(1976)033⟨2452:TDEAFR⟩2.0.CO;2

Karlsson, J., Svensson, G., & Rodhe, H. (2008, June). Cloud radiative forcing of subtropical low level clouds in global models. *Climate Dynamics*, *30*(7-8), 779–788. (WOS:000255090500007) doi: 10.1007/s00382-007-0322-1

Lac, C., Chaboureau, J.-P., Masson, V., Pinty, J.-P., Tulet, P., Escobar, J., … Wautelet, P. (2018). 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., … de Arellano, J. V.-G. (1997, Dec 01). The meso-nh atmospheric simulation system. part i: adiabatic formulation and control simulations. *Annales Geophysicae*, *16*(1), 90–109. Retrieved from `https://doi.org/10.1007/s00585-997-0090-6` doi: 10.1007/s00585-997-0090-6

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

Marchuk, G. I., Mikhailov, G. A., Nazareliev, M. A., Darbinjan, R. A., Kargin, B. A., & Elepov, B. S. (1980). *The Monte Carlo Methods in Atmospheric Optics*. Berlin Heidelberg: Springer-Verlag.

Marshak, A., & Davis, A. (Eds.). (2005). *3d Radiative Transfer in Cloudy Atmospheres*. Berlin Heidelberg: Springer-Verlag. (DOI: 10.1007/3-540-28519-9)

Mayer, B. (2009). Radiative transfer in the cloudy atmosphere. *The European Physical Journal Conferences*, *1*, 75–99. doi: 10.1140/epjconf/e2009-00912-1

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

Nam, C., Bony, S., Dufresne, J.-L., & Chepfer, H. (2012, November). The 'too few, too bright' tropical low-cloud problem in CMIP5 models. *Geophysical Research Letters*, *39*, L21801. (WOS:000310690600003) doi: 10.1029/2012GL053421

Neggers, R. A. J., Duynkerke, P., & Rodts, S. (2003). Shallow cumulus convection

: A validation of large-eddy simulation against aircraft and Landsat observations. *Quarterly Journal of the Royal Meteorological Society*, *129*, 2671-2696.

Neggers, R. A. J., Heus, T., & Siebesma, A. P. (2011). Overlap statistics of cumuliform boundary-layer cloud fields in large-eddy simulations. *Journal of Geophysical Research: Atmospheres*, *116*(D21). doi: 10.1029/2011JD015650

Newman, W. I., Lew, J. K., Siscoe, G. L., & Fovell, R. G. (1995, 02). Systematic Effects of Randomness in Radiative Transfer. *Journal of the Atmospheric Sciences*, *52*(4), 427-435. doi: 10.1175/1520-0469(1995)052⟨0427:SEORIR⟩2.0.CO;2

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

Ramanathan, V., Cess, R. D., Harrison, E. F., Minnis, P., Barkstrom, B. R., Ahmad, E., & Hartmann, D. (1989). Cloud-radiative forcing and climate: Results from the earth radiation budget experiment. *Science*, *243*(4887), 57–63. doi: 10.1126/science.243.4887.57

Schäfer, S. A. K., Hogan, R. J., Klinger, C., Chiu, J. C., & Mayer, B. (2016, July). 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), 2016JD024876. 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

Siebesma, A. P., Bretherton, C. S., Brown, A., Chlond, A., Cuxart, J., Duynkerke, P. G., ... Stevens, D. E. (2003). A large eddy simulation intercomparison study of shallow cumulus convection. *Journal of the Atmospheric Sciences*, *60*(10), 1201-1219.

Sulak, A. M., Calabrase, W. J., Ryan, S. D., & Heus, T. (2020). The contributions of shear and turbulence to cloud overlap for cumulus clouds. *Journal of Geophysical Research: Atmospheres*, *125*(10), e2019JD032017. doi: https://doi.org/10.1029/2019JD032017

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).

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.

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*. doi: 10.1029/2018MS001602

Williamson, D. (2015). Exploratory ensemble designs for environmental models using k-extended latin hypercubes. *Environmetrics*, *26*(4), 268-283. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2335` doi: 10.1002/env.2335