Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement

Fleur Couvreux[1], Frédéric Hourdin[2], Danny Williamson[3], Romain Roehrig[4], Victoria Volodina[5], Najda Villefranque[6], Catherine Rio[7], Olivier Audouin[8], James Salter[3], eric bazile[9], Florent Brient[10], Florence Favot[1], Rachel Honnert[11], Marie-Pierre Lefebvre[1], Jean-Baptiste Madeleine[12], Quentin Rodier[1], and Wenzhe Xu[3]

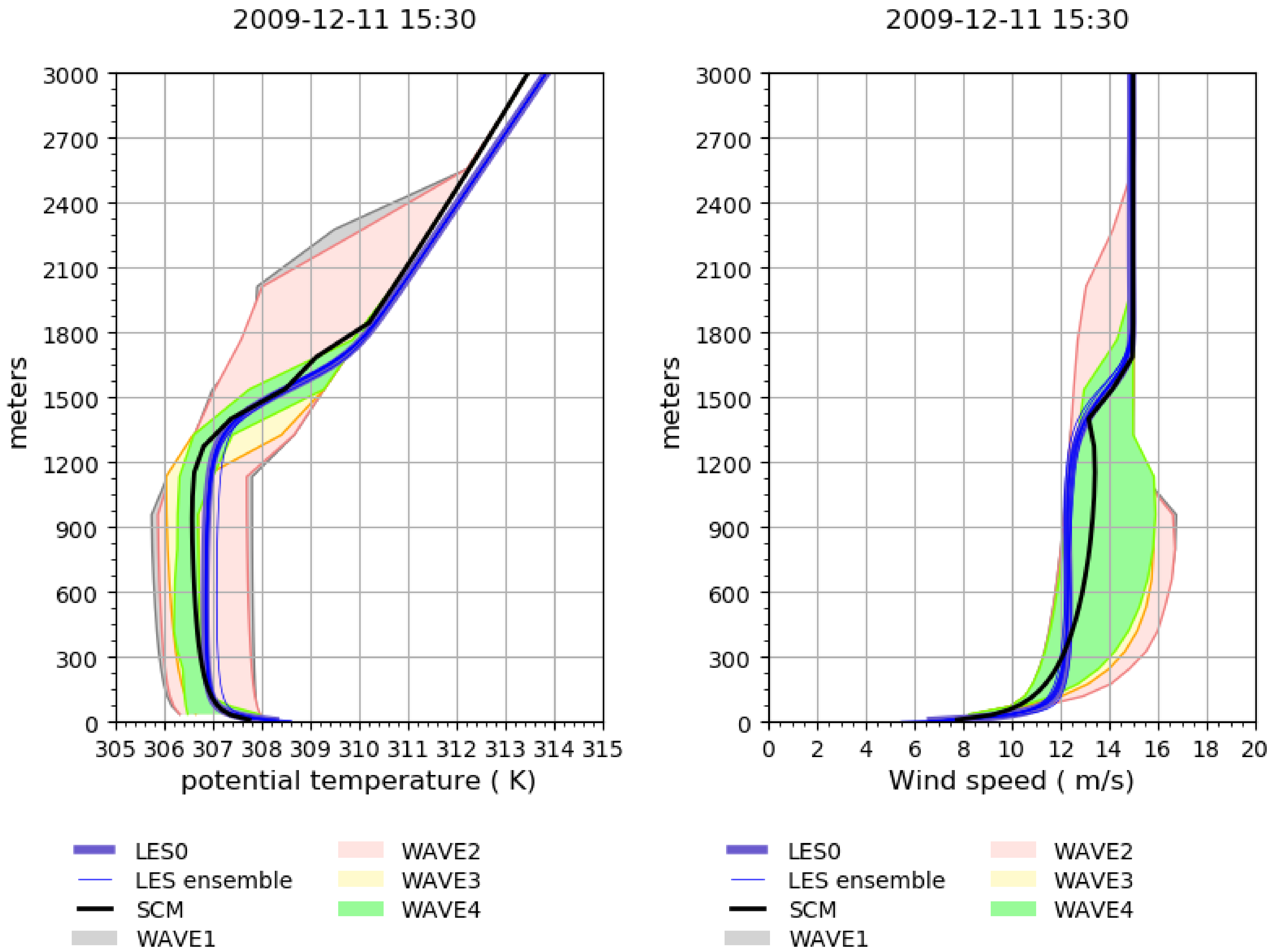
[1]Université Toulouse, CNRM, Meteo-France, CNRS
[2]LMD
[3]University of Exeter
[4]CNRM, Université de Toulouse, Météo-France, CNRS
[5]The Alan Turing Institute
[6]Centre National de Recherches Météorologiques
[7]Centre national des recherches météorologiques (CNRM), Université de Toulouse, Météo-France, CNRS
[8]CNRM, 9 University of Toulouse, Meteo-France, CNRS
[9]Meteo-France/CNRS
[10]CNRM/CNRS/Météo-France
[11]Météo-France, CNRM-CNRS UMR-3589
[12]Laboratoire de Météorologie Dynamique

November 22, 2022

Abstract

The development of parameterizations is a major task in the development of weather and climate models. Model improvement has been slow in the past decades, due to the difficulty of encompassing key physical processes into parameterizations, but also of calibrating or â\euro tuningâ\euro the many free parameters involved in their formulation. Machine learning techniques have been recently used for speeding up the development process. While some studies propose to replace parameterizations by data-driven neural networks, we rather advocate that keeping physical parameterizations is key for the reliability of climate projections. In this paper we propose to harness machine learning to improve physical parameterizations. In particular we use Gaussian process-based methods from uncertainty quantification to calibrate the model free parameters at a process level. To achieve this, we focus on the comparison of single-column simulations and reference large-eddy simulations over multiple boundary-layer cases. Our method returns all values of the free parameters consistent with the references and any structural uncertainties, allowing a reduced domain of acceptable values to be considered when tuning the 3D global model. This tool allows to disentangle deficiencies due to poor parameter calibration from intrinsic limits rooted in the parameterization formulations. This paper describes the tool and the philosophy of tuning in single-column mode. Part 2 shows how the results from our process-based tuning can help in the 3D global model tuning.

Figure 4.



2009-12-11 15:30
meters
potential temperature (K)
305 306 307 308 309 310 311 312 313 314 315
0 300 600 900 1200 1500 1800 2100 2400 2700 3000
2009-12-11 15:30
meters
Wind speed (m/s)
0 2 4 6 8 10 12 14 16 18 20
0 300 600 900 1200 1500 1800 2100 2400 2700 3000
LES0
LES ensemble
SCM
WAVE1
WAVE2
WAVE3
WAVE4

Figure 6.

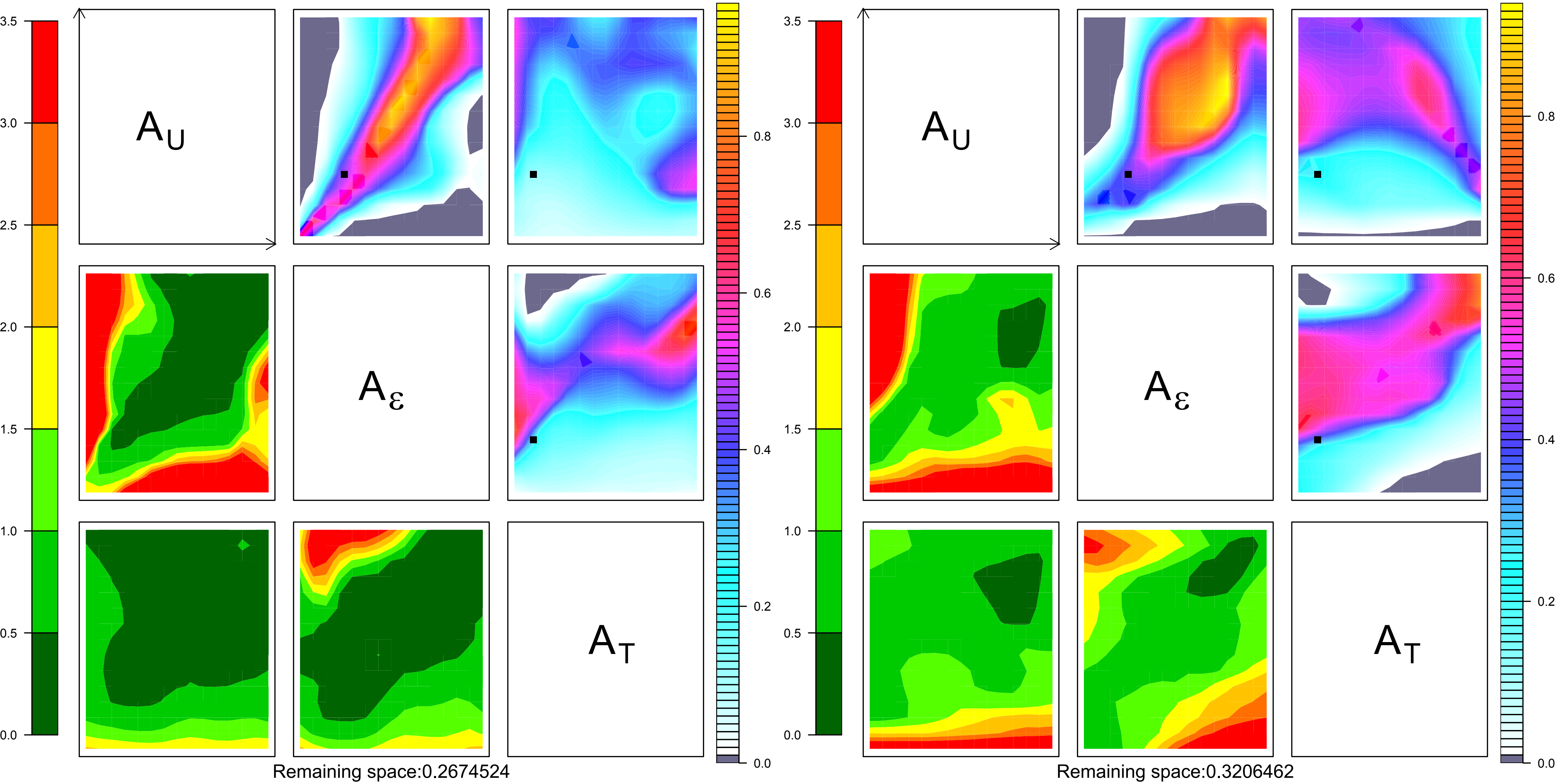
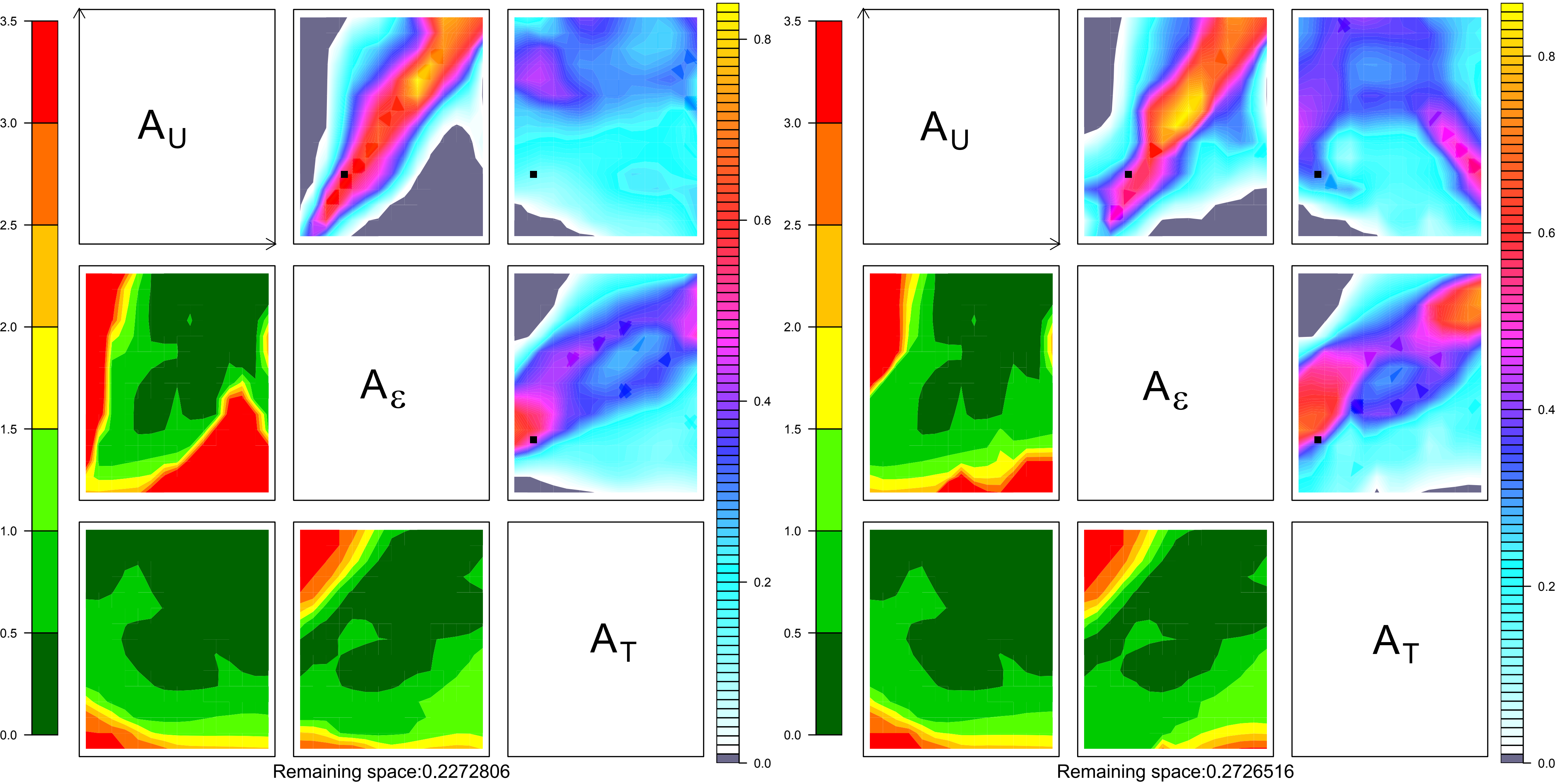

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
A_U
A_ε
A_T
0.8
0.6
0.4
0.2
0.0
Remaining space:0.2674524
Remaining space:0.3206462

Figure 5.



A_U
A_ε
A_T
Remaining space:0.2272806
Remaining space:0.2726516
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.8
0.6
0.4
0.2
0.0

Figure 1.

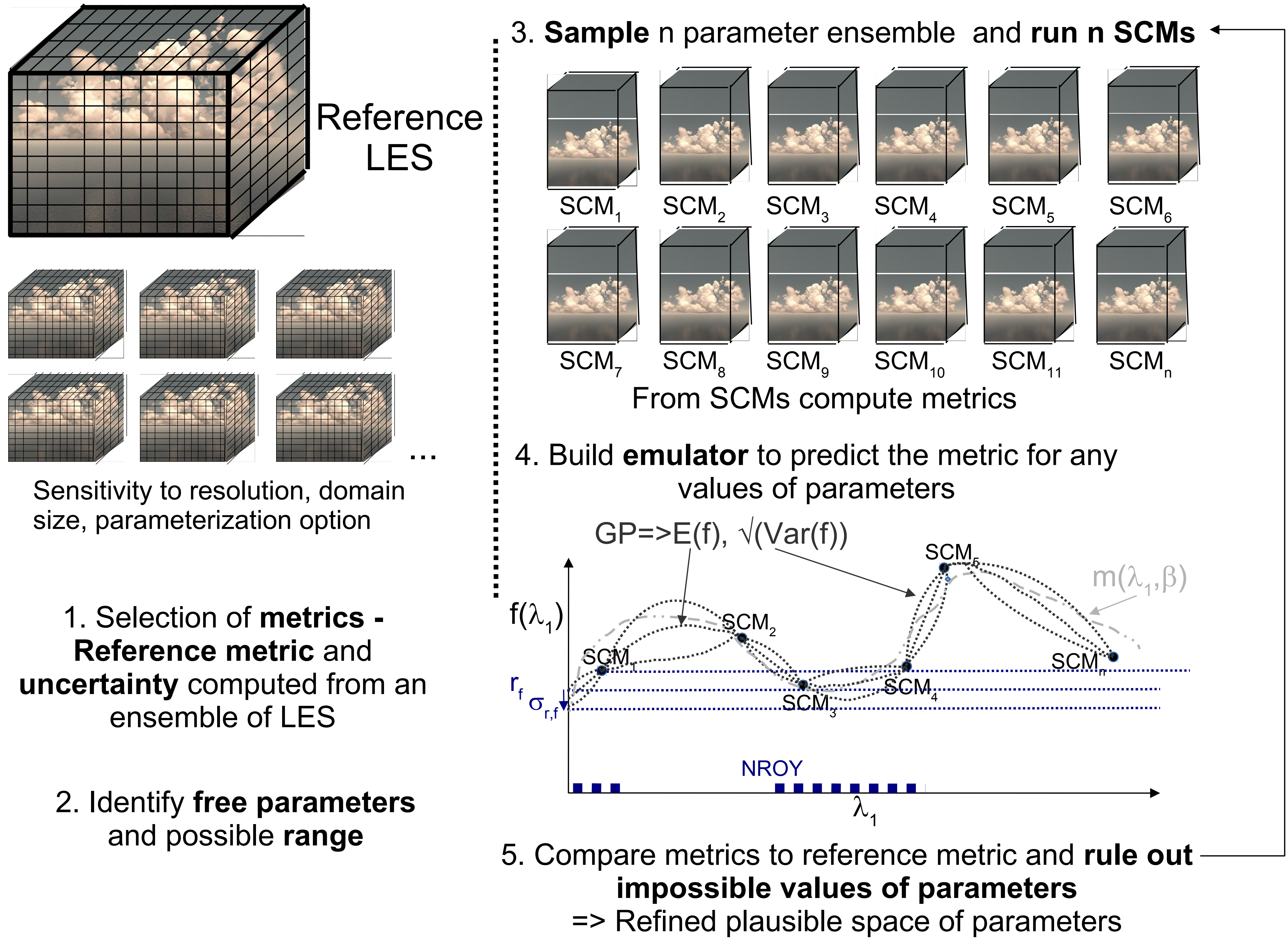
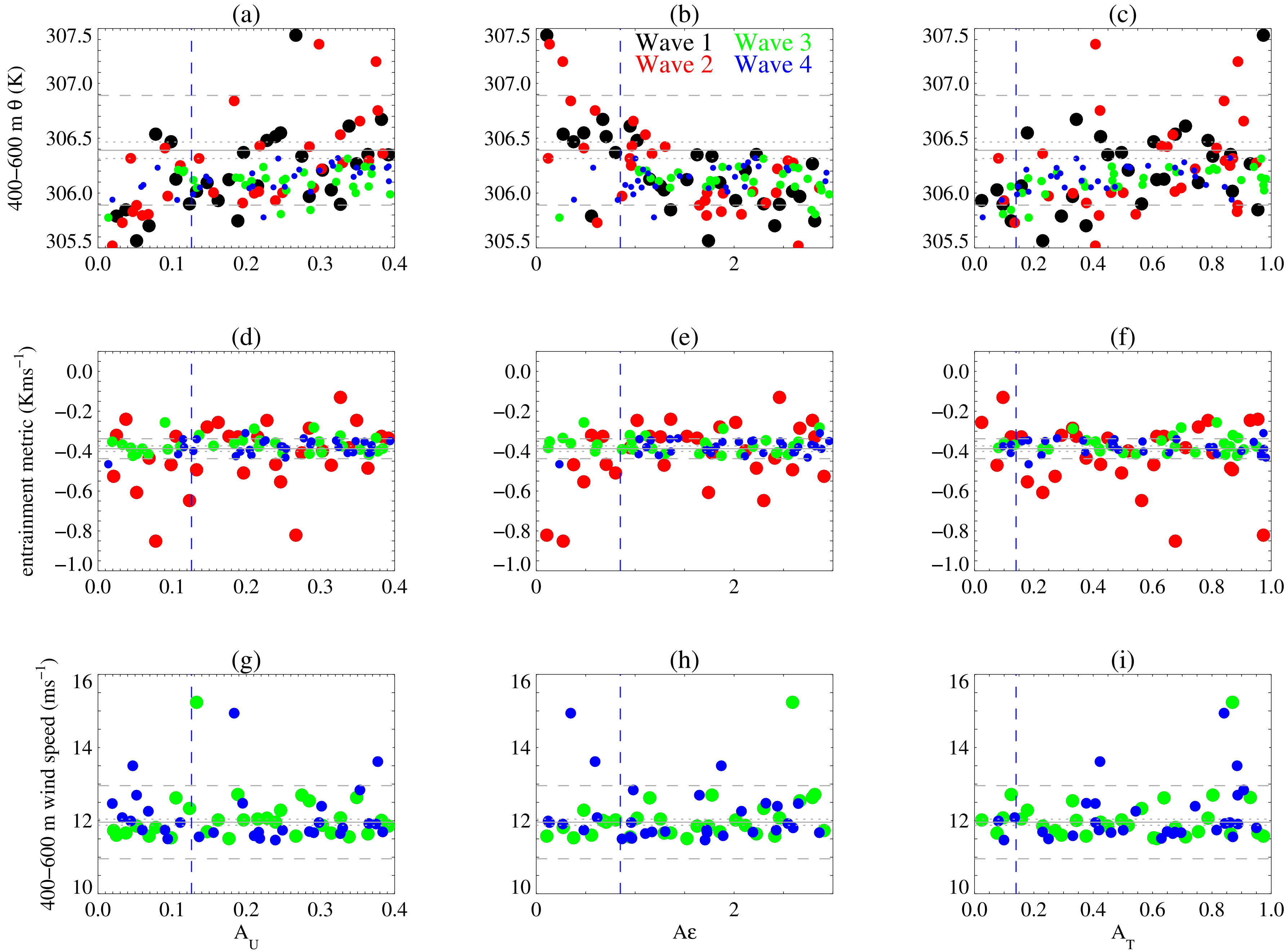
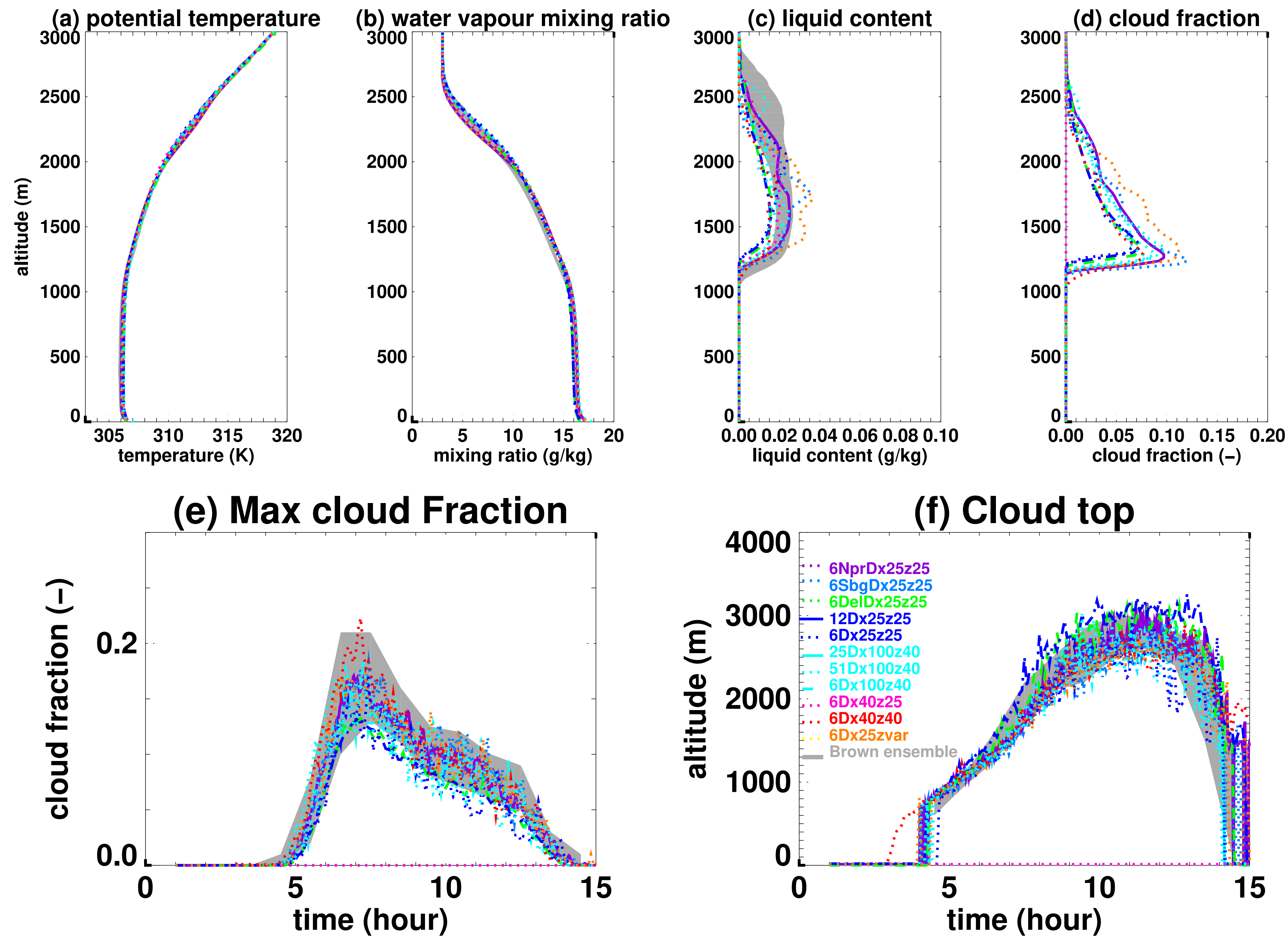

Reference LES
Sensitivity to resolution, domain size, parameterization option
...
1. Selection of metrics - Reference metric and uncertainty computed from an ensemble of LES
2. Identify free parameters and possible range
3. Sample n parameter ensemble and run n SCMs
SCM1
SCM2
SCM3
SCM4
SCM5
SCM6
SCM7
SCM8
SCM9
SCM10
SCM11
SCMn
From SCMs compute metrics
4. Build emulator to predict the metric for any values of parameters
GP=>E(f), √(Var(f))
m(λ1,β)
f(λ1)
rf
σr,f
NROY
λ1
5. Compare metrics to reference metric and rule out impossible values of parameters
=> Refined plausible space of parameters

Figure 3.



(a)
(b)
(c)
Wave 1
Wave 2
Wave 3
Wave 4
400–600 m θ (K)
307.5
307.0
306.5
306.0
305.5
(d)
(e)
(f)
entrainment metric (Kms^{-1})
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
(g)
(h)
(i)
400–600 m wind speed (ms^{-1})
16
14
12
10
0.0
0.1
0.2
0.3
0.4
0
2
0.6
0.8
1.0
A_U
Aε
A_T

Figure 2.



(a) potential temperature
(b) water vapour mixing ratio
(c) liquid content
(d) cloud fraction
altitude (m)
3000
2500
2000
1500
1000
500
0
305 310 315 320
temperature (K)
0 5 10 15 20
mixing ratio (g/kg)
0.00 0.02 0.04 0.06 0.08 0.10
liquid content (g/kg)
0.00 0.05 0.10 0.15 0.20
cloud fraction (–)
(e) Max cloud Fraction
cloud fraction (–)
0.2
0.0
0 5 10 15
time (hour)
(f) Cloud top
altitude (m)
4000
3000
2000
1000
0
6NprDx25z25
6SbgDx25z25
6DelDx25z25
12Dx25z25
6Dx25z25
25Dx100z40
51Dx100z40
6Dx100z40
6Dx40z25
6Dx40z40
6Dx25zvar
Brown ensemble
0 5 10 15
time (hour)

Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement

Fleur Couvreux[1], Frédéric Hourdin[2], Daniel Williamson[3,5], Romain Roehrig[1], Victoria Volodina[5], Najda Villefranque[1,4], Catherine Rio[1], Olivier Audouin[1], James Salter[3,5], Eric Bazile[1], Florent Brient[1], Florence Favot[1], Rachel Honnert[1], Marie-Pierre Lefebvre[1,2], Jean-Baptiste Madeleine[2], Quentin Rodier[1], Wenzhe Xu[3]

[1]CNRM, University of Toulouse, Meteo-France, CNRS, Toulouse, France
[2]LMD-IPSL, Sorbonne University, CNRS, 4 pl Jussieu, Paris, France
[3]Exeter University, Exeter, United Kingdom
[4]LAPLACE, University of Toulouse, CNRS, Toulouse, France
[5]The Alan Turing Institute, 96 Euston Road, London, United Kingdom

Key Points:

- We apply Uncertainty Quantification to Single-Column Model/LES comparison to calibrate free parameters
- We revisit model development strategy with an emphasis on processes for model calibration
- The proposed tuning tool allows to formalize the complementary use of multicases with various metrics

A major task in the development of atmospheric models is the development of parameterizations to account for processes not resolved by the dynamical core. The improvement of model is slow partly due to the difficulty of encompassing key processes into parameterizations and because parameterizations contain 'free' parameters that must be calibrated or 'tuned'. Considering the number of parameters in a model, their calibration is a complicated task, generally done manually. Recently, machine learning has been proposed as a replacement for these parameterizations. However, when models are

Corresponding author: Fleur Couvreux, `fleur.couvreux@meteo.fr`

to be used for long-term projections, exploring states far from the training data, sole use of machine learning might be dangerous. It also seems counter-intuitive to replace our strong physical understanding with unconstrained systems. Our proposition consists in retaining parameterizations but adjoining new tools relying on machine learning to accelerate model development. In particular we use Gaussian process-based methods from uncertainty quantification to calibrate the free parameters at a process level. To achieve this, we focus on the comparison of single-column simulations and reference large-eddy simulations over multiple boundary-layer cases. This paper describes the tools and the philosophy of tuning in single-column mode. Part 2 emphasizes how this framework can help accelerate model development.

Abstract

The development of parameterizations is a major task in the development of weather and climate models. Model improvement has been slow in the past decades, due to the difficulty of encompassing key physical processes into parameterizations, but also of calibrating or 'tuning' the many free parameters involved in their formulation. Machine learning techniques have been recently used for speeding up the development process. While some studies propose to replace parameterizations by data-driven neural networks, we rather advocate that keeping physical parameterizations is key for the reliability of climate projections. In this paper we propose to harness machine learning to improve physical parameterizations. In particular we use Gaussian process-based methods from uncertainty quantification to calibrate the model free parameters at a process level. To achieve this, we focus on the comparison of single-column simulations and reference large-eddy simulations over multiple boundary-layer cases. Our method returns all values of the free parameters consistent with the references and any structural uncertainties, allowing a reduced domain of acceptable values to be considered when tuning the 3D global model. This tool allows to disentangle deficiencies due to poor parameter calibration from intrinsic limits rooted in the parameterization formulations. This paper describes the tool and the philosophy of tuning in single-column mode. Part 2 shows how the results from our process-based tuning can help in the 3D global model tuning.

1 Introduction

Atmospheric global or regional circulation models used either for numerical weather prediction (NWP) or climate studies encompass a dynamical core and a physical component. The dynamical core computes the spatio-temporal evolution of atmospheric state variables by solving a discrete version of the fluid dynamic equations. The physical component quantifies the impact on the resolved variables of radiative, thermodynamical and chemical processes, as well as dynamical processes that occur at scales smaller than the computational grid. These processes are handled by a suite of sub-models, most often referred to as parameterizations, which provide source terms in the resolved-scale equations. Parameterizations (e.g., turbulence, convection, radiation, microphysics) are often based on a mixture of physical principles and heuristic description of the involved processes, of their interactions and of their impact on the larger resolved scales. Although it is difficult to trace back the origin of the term "parameterization" in climate modelling, it semantically points to the fact that the sub-models summarize the processes as functions of the model state vector $\boldsymbol{x}$ (typically the value of zonal and meridional wind, surface pressure, temperature and water phases at each point of the 3D model grid) that depends on some free parameters. These free parameters arise from the simplification of the complex nature of the subgrid processes (e.g., assuming a bulk thermal plume instead of a population of plumes, stationarity). The atmospheric model can be summarized as

$$\frac{\partial \boldsymbol{x}}{\partial t} = \mathcal{D}(\boldsymbol{x}) + \sum_p \mathcal{P}_p(\boldsymbol{x}, \boldsymbol{\lambda}_p) \tag{1}$$

where $\mathcal{D}$ stands for the discretized form of the fluid dynamic equations, $\mathcal{P}_p$ for the source term provided by the parameterization of the process p and $\boldsymbol{\lambda}_p$ for the associated free parameters. This equation may however be too simplistic, as, in reality, a given parameterization often depends on intermediate variables provided by other parameterizations (e.g., cloud fraction used in radiation, turbulence variance used in the cloud scheme) and computes additional prognostic variables (e.g., turbulence kinetic energy). Nevertheless, with this simplified framework, improving models through parameterization development means both to propose more appropriate functional forms $\mathcal{P}_p$ and to identify acceptable or better values of the free parameters $\boldsymbol{\lambda}_p$.

Among the different parameterizations, those involved in the representation of turbulence, convection and clouds still challenge state-of-the art NWP and climate models (Holtslag et al., 2013; Nam et al., 2012; Nuijens et al., 2015; Klein et al., 2017; Randall et al., 2003; Bony et al., 2015). Innovative and diverse concepts and ideas have been proposed over the past decade to improve this representation (Rio et al., 2019). A detailed understanding of the physical processes leading to the formation of low-level clouds can be obtained by Large-Eddy simulations (LES) (Guichard & Couvreux, 2017), which reproduce, with high fidelity, the turbulent dynamics within the clouds (e.g., Siebesma & Cuijpers, 1995; Neggers, Duynkerke, & Rodts, 2003; Wang & Feingold, 2009). LES are therefore increasingly used to derive and evaluate the conceptual models at the root of boundary-layer and shallow cloud parameterizations. The choice of the parameterization free parameters is also crucial for the simulation of clouds. Their calibration or "tuning" consists in searching for acceptable or optimal values of these parameters, such that the associated model configuration has a realistic behavior under various conditions and compared to a suite of observations (Mauritsen et al., 2012). Calibration is therefore a fundamental aspect of NWP or climate model development. However, it is often conducted without much control on the way it modifies the parameterization behavior at the process level as the calibration focuses more on regional or global constraints, such as the radiative balance of the Earth System for climate models, or performance metrics (e.g. root mean square error, skill scores) for NWP models. Hourdin et al. (2017) compile the tuning strategies of several climate groups and emphasize that most of the parameters used to tune climate models (droplet size, fall velocity, entrainment rate) are related to clouds (see also Golaz et al., 2013), i.e. the most uncertain processes that affect radiation, the primary engine of the atmospheric circulation.

Given the societal needs for reliable climate simulations and weather forecasts, the progress achieved by the global atmosphere modeling community has been found slow (Jakob, 2010). Several systematic errors in state-of-the-art models have been modestly reduced, such as those regarding the surface temperature over the eastern oceans (Richter, 2015), the rainfall distribution in the Tropics (Flato et al., 2013), the variability of the liquid water path (Jiang et al., 2012) and the low clouds (Nam et al., 2012). The deadlock of the cloud parameterization, highlighted by Randall et al. (2003), is still an issue today. This too slow improvement of models can be attributed to remaining deficiencies in the structure of the parameterization itself (the function $\mathcal{P}_p$) but also to the calibration of model parameters that can be considered as a bottleneck in model development. On the one hand, the calibration may not be done efficiently enough, and on the other hand, tuning may induce error compensations that contribute to slow model development. Indeed, a new model development usually starts with a model score degradation by breaking this compensation, as often experienced in the weather prediction centers where strong weight on well-established metrics slows down the implementation of new model development in the operational version (Sandu et al., 2013).

Various avenues have been proposed to get around these difficulties and accelerate climate model improvement. A first avenue seeks to exploit the high resolution, explicitly resolving convection, to reduce the number of involved parameterizations. With the recent increase of computer power, it is nowadays possible to run global kilometer-scale resolution simulations over a few months (Satoh et al., 2008, 2019; Stevens et al., 2019). However, the explicit simulation of the fluid dynamics associated with the life cycle of a cumulus requires grid resolution of the order of several tens of meters. Such resolution will not be accessible in the foreseeable future for climate change projections which require simulations of the global Earth System covering at least several hundreds of years (model spin-up plus transient simulations in response to anthropogenic forcing). The superparameterization approach (Randall et al., 2003) proposes an intermediate pathway by introducing a convection-permitting model in each column of a conventional general circulation model (GCM) to replace the deep convection parameterization (Khairoutdinov et al., 2005). The use of a large-eddy model instead of a convection-permitting model in such framework further removes the boundary-layer and shallow convection parameterizations (Grabowski, 2016; Parishani et al., 2017). A second avenue recently explored the potential of machine learning approaches, which ultimately envisions to replace some parameterizations by neural networks or similar algorithms, properly trained on convection-permitting model simulations or superparameterized GCM (Krasnopolsky et al., 2013; Brenowitz & Bretherton, 2018; Gentine et al., 2018).

A third proposition consists in retaining parameterizations in models but adjoining new tools relying on machine learning to accelerate model development. This choice is motivated by the fact that parameterizations summarize our current understanding of the dynamics and physics of atmospheric processes and offer the power of interpretation, crucial to build our confidence in the extrapolation beyond observed conditions realized by any climate projections. The ESM2.0, proposed by Schneider et al. (2017), belongs to this category. The authors defend that the major progress in Earth-System model development should come from a more systematic use of global observations and high-resolution simulations thanks to machine learning algorithms. They also underline the importance of climate model calibration. In particular, they stress that their new Earth System modeling framework comes with challenges such as developing innovative learning algorithms, identifying the best metrics, combining information from observations and high-resolution, innovating in the design of parameterizations such that they can more easily benefit from new observations or evolution of the models (e.g., refinement of resolution).

Along the same lines, we propose, in this paper, a new approach which allows the development of the parametrizations and their calibration to be tackled at the same time. We argue that a major slowdown of model improvement resides in the difficulty to clearly identify parameterization deficiencies and to properly disentangle them from the inherent calibration of their adjustable parameters at the process and global scales. It is likely that process-scale parameterization improvements are often hidden by the unavoidable full model retuning, required to maintain a reasonable radiative balance or acceptable scores. In the proposed approach, machine learning is harnessed in a principled way to calibrate parameterizations at process level. We promote a more systematic use of the multi-case comparison between Single-Column Model (SCM) and LES to evaluate and calibrate parameterizations, as we advocate that a lot still remains to be learnt from this comparison. Such a systematic use is not feasible however without more objective and automatic methods than the traditional trial/error approach used to fix parameter values during the parameterization development. Indeed, this trial/error approach is only applicable to one piece of a particular parameterization and one or two relevant cases at most. Here, we aim at assessing a set of parameterizations $\mathcal{P}_p$ for a series of test cases, which can be formalized as the question of the existence of a sub-space of the parameters λ_p that allows to match metrics between SCM and LES results for the series of cases, within a given tolerance to error.

Hourdin et al. (2017) reviewed the general practice for climate model calibration and proposed three different levels of calibration in a model development: a first calibration at the level of individual parameterizations, then a calibration of each component of the Earth System model and eventually a calibration of the full Earth System model. Distinguishing those three levels may avoid compensating errors that could arise if the calibration is only done at the last level. In this paper, we propose a methodology to address the first phase, *i.e.* the process-level calibration and defend that it can be part of the elaboration of a well-defined calibration strategy based on solid physical and statistical methodologies. By doing so, we tackle model development and parameter calibration together rather than independently as currently done for most climate model development.

Machine learning has already been proposed to calibrate free parameters (e.g., ensemble Kalman filters as in Schneider et al., 2017). The methodology retained here for model calibration uses history matching with Gaussian processes. History Matching is an efficient way to explore and reduce the domain of free parameters $\boldsymbol{\lambda}_p$ and document how a model physics, namely the suite of functions $\mathcal{P}_p$, behaves within this domain. Williamson et al. (2013) applied History Matching to tune the Hadley Climate Model and stressed its advantage: it accounts for the various sources of uncertainties in assessing the compatibility of the model with the reference: namely the reference uncertainty itself, the uncertainty introduced by the Gaussian process representation of the parametrization, and the intrinsic ability of the model to represent the reference (often referred to as structural error or model discrepancy). History matching inherently deals with the overconfidence issue, which emerges when model calibration is addressed as an optimization problem (Salter et al., 2019). It has been widely used to calibrate models in astrophysics (Vernon et al., 2010), epidemiology (Andrianakis et al., 2017) and hydrocarbon reservoirs (Craig et al., 1996). It has been applied to climate models (Williamson et al., 2015, 2017) and is starting to be used to find biases in models (McNeall et al., 2019).

Whilst history matching has been applied to calibrate 3D models, it has not been harnessed for process-level tuning, as we advocate here through application to SCM/LES comparison. The SCM approach provides confidence in the model's ability to represent some of the key processes whereas a direct calibration of the 3D global model targeting large-scale constrains may hide compensating errors (as discussed in Williamson et al., 2017). SCM calibration is able to reduce the domain of the free parameters for a parameterization, information that can be used for efficiently calibrating the full 3D global model (as we demonstrate in part II). The breakthrough proposed here was only possible thanks to a strong collaboration between the Uncertainty Quantification community and the atmospheric modelers.

The present paper focuses on parameterizations involved in the representation of boundary-layer clouds. Indeed, well-established case studies exist for such regimes and LES have been shown to realistically represent the main processes. However, this methodology can be easily expanded to other parameterizations and other objectives in the Earth System.

The paper is organized as follows: the next section describes the SCM/LES framework highlighting its advantages, recalls the different steps used in the development of a parameterization and details the new philosophy advocated here. Section 3 presents the statistical tool, with a focus on its philosophy and its main ingredients. Section 4 presents a guideline for its use based on a simple illustration. The paper ends with conclusions in Sect. 5. A companion paper (part II) illustrates the significant advances in model development offered by this tool. It exploits process-based calibration for model development and shows how this tool provides guidance for the tuning of a 3D global model.

2 A systematic use of the SCM/LES comparison

Observations only provide a sparse view, in time, space and variables, of the physical processes responsible for convection and clouds. In contrast, LES have the advantage of providing coherent 3D fields characterizing the dynamical and thermodynamical state of the atmosphere. Of course, LES models include turbulence and microphysics parameterizations and thus contain modeling uncertainties, but they have been shown to reproduce the turbulent dynamics of the clouds with high fidelity (e.g., Neggers, Duynkerke, & Rodts, 2003; Heus et al., 2009). As a result, LES have become a central tool in the development and evaluation of parameterizations of convection and clouds. Their analysis has helped in building the conceptual models behind several parameterizations (e.g., Neggers et al., 2002; Rio et al., 2010). LES are also used for the evaluation of the parameterizations in particular those involved in the representation of boundary layers and shallow clouds (e.g., Ayotte et al., 1996; Golaz et al., 2002; Hourdin et al., 2002; Neggers et al., 2004; Siebesma et al., 2007; Rio & Hourdin, 2008; Caldwell & Bretherton, 2009; Neggers, 2009; Pergaud et al., 2009; Rio et al., 2010; Suselj et al., 2013; Neggers et al., 2017; Tan et al., 2018; Suselj et al., 2019).

For their evaluation, parameterizations are often tested in a single-column framework, particularly relevant for global circulation model parameterizations, which are fundamentally 1D. SCM are built by extracting, from a 3D model, a single atmospheric column, which integrates the same set of subgrid parameterizations (boundary-layer, shallow convection, deep convection and microphysics schemes) and is run in a constrained large-scale environment (Zhang et al., 2016). The state vector of the SCM simulation is then a restriction to one column $\boldsymbol{x}_c$ of the full 3D state vector $\boldsymbol{x}$ and Eq. 1 reduces to Eq. 2. The dynamical term $\mathcal{D}(\boldsymbol{x})$ becomes a source term $\mathcal{F}_c$ specified as a function of time and altitude z; we however discard this dependency in the notation for simplicity. It can also depend on the column full state vector, $\mathcal{F}_c(\boldsymbol{x}_c)$, if for instance the large-scale advection is separated between a prescribed horizontal advection and a vertical advection computed as -$w\partial\boldsymbol{x}_c/\partial z$, where w is an imposed vertical velocity. During the SCM integration, some parameterizations can be deactivated in which case the corresponding source term is either neglected or included in the forcing $\mathcal{F}_c$. It is the case for instance when the radiative heating is imposed rather than being computed interactively by the model radiation scheme or when turbulent surface fluxes are imposed rather than computed by the model bulk paramerizations. What really matters in the SCM/LES approach is that both models use the exact same initial and boundary conditions and forcing terms. In a simplified formalism, the SCM thus corresponds to

$$\frac{\partial \boldsymbol{x}_c}{\partial t} = \sum_{p_{\text{activated}}} \mathcal{P}_p(\boldsymbol{x}_c, \boldsymbol{\lambda}_p) + \mathcal{F}_c(\boldsymbol{x}_c) \tag{2}$$

and the LES to

$$\frac{\partial \boldsymbol{y}}{\partial t} = \mathcal{L}(\boldsymbol{y}) + \mathcal{F}_c(\overline{\boldsymbol{y}}) \tag{3}$$

with

$$\boldsymbol{x}_c(t=0) = \overline{\boldsymbol{y}}(t=0) \tag{4}$$

where $\boldsymbol{y}$ stands for the full LES state vector, $\mathcal{L}(\boldsymbol{y})$ to the LES model equations (which include the LES parameterizations) and $\overline{\boldsymbol{y}}$ to the horizontal-domain average of the LES state vector. The SCM/LES framework thus provides a rigorous comparison between both simulations, as it removes the uncertainties, which may arise from different initial conditions or large-scale forcing when directly comparing SCM to observations. This constrained framework avoids the need to disentangle parameterization contributions from their coupling with the large-scale dynamics. Another important aspect of the method is that SCM simulations are computationally very cheap. The joint utilization of LES and SCM was first advocated by Randall et al. (1996); Ayotte et al. (1996) and has been, since then, widely used within the Global Energy and Water Exchanges (GEWEX) Cloud System Study (GCSS; Browning et al. (1993) community, now renamed the Global Atmospheric System Studies, GASS, community). One of the most important legacies of this group for the atmospheric modeling community is an ensemble of test cases that connect observations, LES and SCM, and which sample many typical situations over the globe, thought to be of importance for the climate system (e.g., Siebesma & Cuijpers, 1995; Brown et al., 2002; Duynkerke et al., 2004). As such, this framework has been increasingly used in model development (e.g., Hourdin et al., 2013; Gettelman et al., 2019; Hourdin et al., 2020; Roehrig et al., 2020), all the more so as SCM simulations have been shown to reproduce uniquely the behaviour of their GCM justifying the use of SCM simulations for improving weather and climate models (Hourdin et al., 2013; Neggers, 2015; Gettelman et al., 2019).

Traditionally, parameterizations are often tested over a few specific cases for which high-resolution simulations are available (e.g., Ayotte et al., 1996). Recently, the importance of using a wide benchmark of cases covering the different regimes encountered in reality instead of only a limited number of cases has been stressed (e.g., Neggers et al., 2012). We also highlight here the importance of using an extensive ensemble of cases. The use of multi-case is indeed essential for exploring the various degrees of freedom of the parameterization package. A stable boundary-layer case will constrain the turbulent diffusion; the combination of cloud free and cumulus topped convective boundary layers will ensure that cloud cover is obtained for a good representation of convection; transition cases from stratocumulus to cumulus will ensure the extension to stratocumulus regimes, etc. Combining multi cases and multi metrics is a much more robust assessment of model performance as also highlighted by (Neggers et al., 2017). To better use multi-cases, one important technical aspect is a common definition, in a predefined acknowledged format, for the description of the setup of reference cases, to be used both to perform SCM simulations or LES. This definition should include the description of the initial profiles and large-scale forcing but also contain information on the configuration to be used (e.g. the type of surface boundary conditions, the existence of any nudging towards reference vertical profiles, the way large-scale forcings are provided). An international initiative is ongoing to agree on the description of the format for this definition file. Sharing a standard to define cases will ease the realization of cases by any model and facilitate the share of new cases. The importance of creating libraries of high-resolution simulations representing different climate is another important aspect already identified as a goal by the GCSS community and stressed in Schneider et al. (2017). A common format and the libraries of LES are an important pre-requisite for the tool presented here. In addition, both will contribute to bringing the process-scale community and the community developing global models more closely together.

When comparing SCM and LES, the modeler has to decide which metrics to consider. Various types of metrics can be used. One can directly compare components of the SCM state vector $\boldsymbol{x}_c$ to their equivalent in LES, the horizontal domain-average state vector $\overline{\boldsymbol{y}}$ (e.g., vertical profiles of potential temperature, specific humidity and less often wind components). Assessing the ability of the parameterizations to reproduce the time evolution of $\boldsymbol{x}_c$ for a given forcing is indeed the ultimate goal. By doing so, one not only tests the behavior of one particular parmaterization but also its coupling with the other parameterizations activated in the SCM. However, it may make the determination of the behavior of the targeted parameterization more difficult and can hide compensating errors: for example, a given temperature turbulent flux can be obtained by different contributions from organized structures and small-scale turbulence when represented by two different parameterizations such as in the Eddy-Diffusivity Mass-Flux framework (Hourdin et al., 2002; Siebesma et al., 2007; Neggers, 2009; Pergaud et al., 2009). Another type of metrics targets parameterization-oriented variables, such as mass fluxes, heating source associated with one part of the motion only, subgrid-scale distribution of temperature or water, cloud vertical structure, updraft vertical velocity, area fraction or entrainment and detrainment rates. The metric, from the SCM point-of-view, is nolonger derived from the model state variables but corresponds to an internal variable to the parameterizations. However, additional uncertainty arises from the way such variables and associated metrics can be derived from LES. For example, clouds can be characterized in an LES as all the grid cells containing condensed water (e.g., Siebesma & Cuijpers, 1995). Combined with thresholds on the vertical velocity, cloudy updrafts can be separated from cloudy downdrafts. The analysis of the joint distribution of variables (Chinita et al., 2018) or the use of ad-hoc passive tracers can also be used in the LES to identify objects relevant with the conceptual model of the parameterization (e.g., Couvreux et al., 2010; Rio et al., 2010; Chinita et al., 2018; Brient et al., 2019). Such parameterization-oriented diagnostics have helped in the refinement of the conceptual model at the root of the parameterization (e.g., Rio et al., 2010; Jam et al., 2013; Rochetin et al., 2014). However, a question arises if such diagnostics should also be used as metrics in the calibration process. Answering this question on the relative importance to give to one type of metrics or another requires efficient algorithms, as the one proposed here, to explore the various options. Note also that using state vector-based metrics on a large set of cases that are more or less sensitive to one aspect of the parameterization may help avoid the error compensation issue. Neggers et al. (2017) propose to combine simple metrics on a unique score metric through the use of the bias and the root-mean square errors on each metric. As will be detailed later, we have decided to explicitly keep the individual information brought by each metric.

In line with Neggers et al. (2012), we advocate that, although not a new approach, the power of SCM/LES comparisons is largely underestimated and under-exploited. Applying history matching to this comparison is a way to fully take advantage of the SCM/LES on a large multi-case ensemble and explore whether there exists a sub-space of the parameter space for which the SCM is able to reproduce a series of LES simulations within a given uncertainty. Note that the metrics can be different from one case to the other.

This tool offers the possibility to revisit the different intercomparison exercises documented in the literature and to benefit from this rich database still underused.

Eventually, a point that becomes crucial when using LES for parameterization evaluation and tuning is the assessment of LES reliability and its uncertainties. Although it has been shown, through the comparison to observations, that LES is able to correctly reproduce boundary-layer processes and shallow clouds (Couvreux et al., 2005; Neggers, J, & Siebesma, 2003; Heus & Jonker, 2008), LES, as in many models, come with uncertainties associated to the advection scheme and the parameterizations still active in such simulations concerning small-scale turbulence, microphysics, radiation and surface fluxes. Sullivan and Patton (2011) have shown that a horizontal resolution of a few tens of meters for convective boundary layers is enough to get convergence for the mean, fluxes and variances but 10m resolution is needed in order to get convergence on skewness. The sensitivity of LES of shallow convection to resolution, size of the domain, subgrid model and advection scheme has been widely investigated (Brown, 1999; Matheou et al., 2011; Pressel et al., 2017; Zhang et al., 2017; Wurps et al., 2020). In particular, it has been shown that most of the ensemble-averaged turbulence statistics are reasonably insensitive, allowing one to use LES results to develop and evaluate convection parameterizations. However, some characteristics of the cloud fields (e.g. size distribution of individual clouds) are more sensitive to resolution or advection scheme (Brown, 1999; vanZanten et al., 2011). Uncertainty around this reference should be documented so that history matching can explicitly take it into account.

3 ***High-Tune Explorer*** (htexplo), a statistical tool to calibrate model parameters and more

3.1 Overview

The present section describes the tool proposed to perform process-based calibration. Its objective is twofold: (i) characterize the domain of the model parameter values that allows the model to appropriately capture process-level metrics and which can be used for subsequent calibration of the global model, and (ii) identify the model parameters that limit model performance and thus highlight the need for model parameterization revision. The tool relies on history matching approach developed by Vernon et al. (2010) and first used for climate studies by Williamson et al. (2013). This method aims at removing “unphysical” regions of parameter space iteratively, refocusing the search for "acceptably tuned" models at each step. The tool finds the subspace of the model parameter space containing simulations consistent with the reference metrics, acknowledging the various sources of uncertainty. This tool has already been successfully applied to identify the acceptable range of model parameter values in the 3D configuration of the Hadley Centre climate model (Williamson et al., 2013, 2015) or in the NEMO oceanic model (Williamson et al., 2017). It is here used for the first time in the context of the SCM/LES comparison for a given set of cases.

As already stated in the previous section, we focus here on the parameterizations involved in the representation of boundary-layer clouds (turbulence, convection, cloud micro and macrophysics, radiation). However, this methodology can be easily expanded to other parameterizations and other objects of the Earth system as soon as reliable references are available.

Figure 1 sketches the main steps of the *High-Tune Explorer* (htexplo in the following for an explorer to use High-resolution simulation to improve and Tune parameterizations) tool:

- 1. **Metric selection and references** First, the cases and associated target metrics are selected. The relevant reference for each metric is then identified and the associated uncertainty is estimated. In the present case, the reference is an LES and the associated uncertainty is based on an LES ensemble. Observations could also be used with an associated error when an LES is not available. This phase is not model-specific and could be shared between different models.
- 2. **Selection of model parameters** The model parameters to be calibrated are identified and their possible range of values are determined.
- 3. **Experimental design and SCM runs** The experimental design consists of defining the ensemble of experiments (or SCM) to be run. The goal is to optimally sample the parameter space and provide a small set of parameter values for which the single-column model will be run. Metrics are computed from each of the SCM simulations and form the training data-set on which emulators are built.
- 4. **Building emulators**, i.e. construction of surrogate models, also called "emulators", one for each metric. Each emulator is based on a Gaussian Process (GP) and predicts the corresponding metric value at any point of the full parameter space, without running the SCM. The GP statistical model also provides a probability

distribution of its prediction, thus quantifying the prediction uncertainty for use in calibration.

- 5. **History matching** The comparison between the reference metrics and those inferred with the emulators is based on a distance that accounts for reference uncertainty, modeler tolerance to error or model discrepancy (induced by e.g., misrepresentation of specific processes, inaccuracy of numerical solvers, model resolution) and emulator uncertainty. History matching rejects parameter values that lead to unacceptable model behavior (too large distance from the reference) and thus defines a not-ruled out yet (NROY) space, the model parameter space that cannot be further reduced given the sources of uncertainty.
- 6. **Iterative refocusing** To reduce the emulator uncertainty, but only where needed, new iterations (or waves) following steps 3 to 5 are performed, sampling the NROY space obtained at the end of the previous wave for the design and only constructing emulators over the NROY domain.

Details on the different steps are given below. For simplicity, we first describe them for the first iteration and only one metric. Subsequent iterations and the addition of other metrics are discussed in Sect. 3.7. This section ends with a discussion about the relationship between the present tool and more common tools used for calibration and sensitivity analysis.

3.2 Step 1: Metric selection and references

The metrics used to evaluate the SCM behavior depend on the physical situation considered and the parameterization hypothesis. Scalar metrics based on a dynamical or thermodynamical variable (e.g., potential temperature, water vapor mixing ratio, wind speed, cloud fraction) sampled at a given time can be used, such as the value at a given vertical level, the average or the maximum over a given layer (e.g., boundary layer, cloud layer), or the maximum over the whole atmospheric column. Radiation-oriented metrics are particularly relevant to enhance the link between the present process-oriented model calibration and the calibration of the corresponding 3D configuration. Ideally, the chosen metric should be as insensitive as possible to the model vertical resolution. In that regard, integrals (or averages) are good candidates for scalar metrics, as will be illustrated in Part II. Root-mean square errors are not encouraged for two reasons, i/ there are usu-

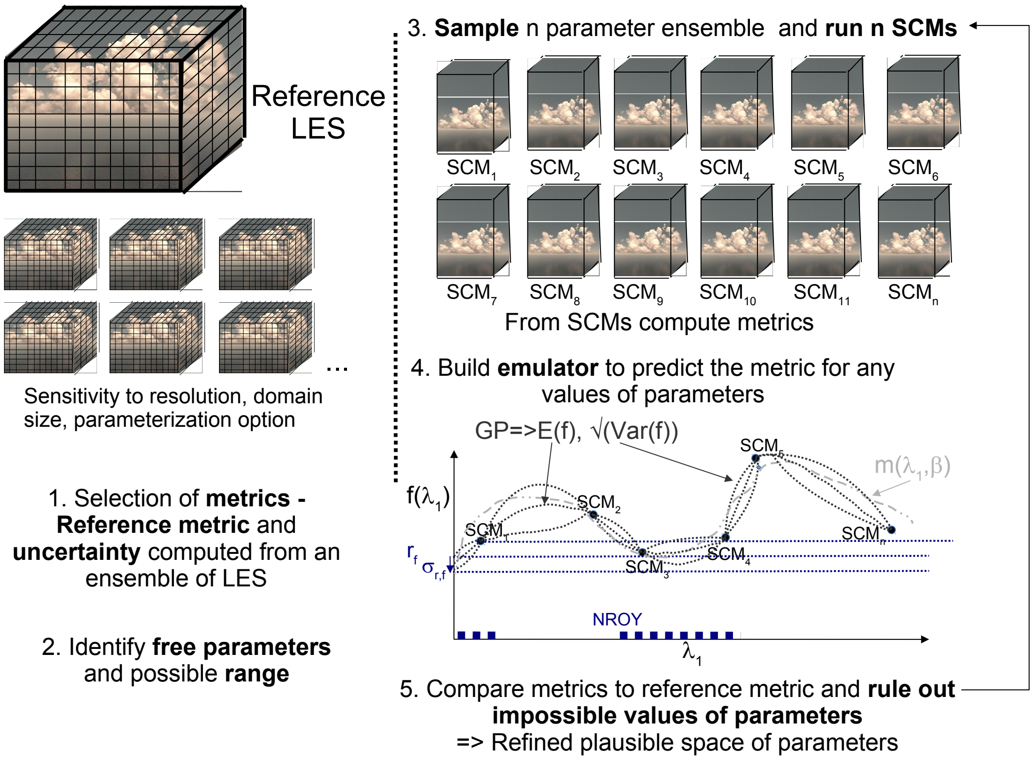


Figure 1. Schematic of the different steps of the htexplo tool

ally associated to a smaller signal to noise ratio and ii/ the Implausibility (see Sect. 3.6) is already a kind of root-mean square error. The number of metrics to be used is generally of the order of ten, but it can be many more.

More complex metrics such as vertical profiles, time series or spatial fields, can also be considered. In that case, methods are used to reduce the dimensions of the outputs and principal component decomposition is one option (e.g., Salter et al., 2019). However, scalar metrics, taken at a given time, or averaged over a short period of time, seem often sufficient to robustly constrain most of the SCM simulations (Personal communication O Audouin). Therefore, in the present paper and in Part II, only scalar metrics will be used. They include boundary-layer average potential temperature and maximum cloud fraction.

References and their associated uncertainty are estimated from an LES ensemble. There are a priori two possibilities to build such an ensemble, which can be combined. The first consists in building the ensemble from simulations performed by different large-eddy models, as has been done in several GCSS intercomparison exercises (Brown et al., 2002; Siebesma et al., 2003; vanZanten et al., 2011; Pressel et al., 2017). The reference thus corresponds to the LES ensemble mean, while the uncertainty is quantified by the LES ensemble variance. The second option, used in this paper, relies on only one large-eddy model and estimates the uncertainty around the reference model configuration by performing sensitivity experiments to horizontal and vertical resolution, domain size, and parameterization options (e.g., turbulence, microphysics, surface fluxes, radiation). In this study, we have chosen to use the simulation realized with the higher resolution over the largest domain and with the most relevant parameterization options as the reference, but the ensemble mean could also be used. The large-eddy model used in this study is the LES-configuration of Meso-NH (Lac et al., 2018). It makes use of a fourth-order centered discretization associated with an explicit fourth-order Runge-Kutta time integration. Figure 2 illustrates the spread obtained from a Meso-NH LES ensemble exploring the sensitivity to horizontal, vertical resolution, domain size and options in the turbulence and cloud schemes for one given case, namely the ARM Cumulus case, which is a golden case for the study of continental cumulus (Brown et al., 2002). Table A2 in the Appendix describes in detail the different simulations used to estimate the uncertainty. Consistently with the literature (Brown et al., 2002; Matheou et al., 2011; vanZanten et al., 2011; Zhang et al., 2017), domain-average conserved thermodynamical quantities are weakly sensitive to changes in resolution, domain size and parameterization choices while the domain-average liquid water content and cloud fraction exhibit more spread. Metrics derived from those latter quantities will therefore be associated to a larger uncertainty. Figure 2 also indicates in grey shading the spread obtained from the LES intercomparison of Brown et al. (2002) highlighting a similar uncertainty estimate between the two methods mentioned above. Similar results are obtained for LES ensembles of other intercomparison exercises (not shown). For a given metric f, r_f is the reference metric value, estimated from the reference LES simulation or the average of the LES ensemble and $\sigma^2_{r,f}$ is the associated square error estimated from the LES ensemble. Note that, in the absence of available LES, observations can also be used as a reference to be compared to the SCM runs as illustrated in Ahmat Younous et al. (2018) but the observation error needs to be quantified.

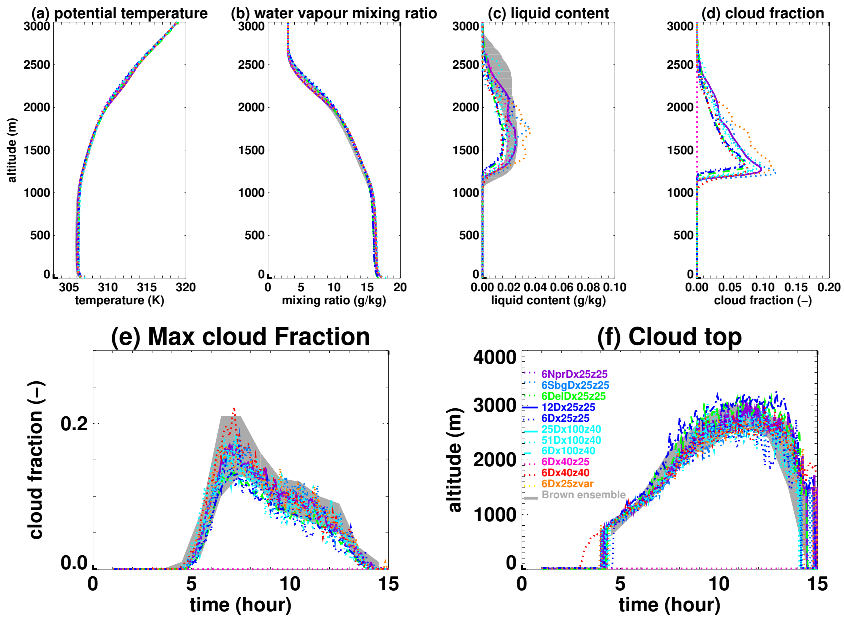


Figure 2. Vertical profile of (a) potential temperature, (b) water vapour mixing ratio, (c) liquid water content and (d) cloud fraction averaged over the horizontal domain at the 10^{th} hour of the simulation (1530 LT) and time series of (f) the cloud top and (e) the maximum cloud fraction over the atmospheric column. The grey shading corresponds to the results of the Brown et al. (2002) intercomparison. The different color lines correspond to different sensitivity tests realized with Meso-NH changing either, one by one, the size of the domain, the vertical or horizontal resolution and some option in the cloud scheme, microphysics scheme or turbulence scheme (detailed in Table A2).

3.3 Step 2: Selection of model parameters

The number of model parameters can be large (generally on the order of 10 for each parameterization). Estimating the prior range of values that needs to be explored for each of them requires the modeler's expertise and experience about the model and parameterizations. The definition of this range is an important step as the results are only valid in this predefined parameter space (Williamson et al., 2013). So, we advise to choose a range as wide as possible in the absence of physical reasons or numerical concerns for constraining it. Nevertheless, the user might consider some tradeoff as the smaller the ranges, the smaller the space to explore.

As the tool samples any parameter independently from the others (see Step 3), the method remains efficient even though a parameter with no influence on the results was included. A sensitivity analysis (Oakley & O'Hagan, 2004) could be used as a preliminary step in order to reduce the number of parameters selected but may not be a good idea in general (see Sect. 3.8). Depending on the predefined range of parameter values, the user can consider either linear or logarithmic variations of the parameter values.

In the following, we consider a set of parameters $\boldsymbol{\lambda} = (\lambda_k)_k$, which is a subset of the model parameters $(\boldsymbol{\lambda}_p)_p$ (see Sect. 1).

3.4 Step 3: Experimental design and SCM runs

Once the model parameters are selected and their range of values defined, an experimental design is built. It corresponds to the selection of a relatively small set of values for the model parameters $(\boldsymbol{\lambda}_i)_{i=1,\ldots,n}$, usually on the order of ten times the number of parameters. It explores the initial (or input) space of the parameter values in the range given for each parameter. An SCM simulation is performed for each of them and provides the state vector $\boldsymbol{x}_c(\boldsymbol{\lambda}_i)$. The objective is to "fill" the parameter space as uniformly as possible maximizing the minimum distance between points. Here, as classically used for the design of computer experiments, a Latin Hypercube (LHC) (Williamson et al., 2015) is used to efficiently sample the input parameter space. Classically, a LHC for a n-member ensemble uniformly divides each dimension of the input space into n bins that are sampled once each and only once. All the parameters are thus varied simultaneously in contrast to other sensitivity analysis approaches such as in the Morris sensitivity analysis (Saltelli, 2002), where parameters are varied one by one. The LHC sampling used here maximizes the minimum distance between the selected points of the input space.

More precisely, here we use k-extended latin hypercubes as proposed by Williamson (2015). It consists in producing several LHCs, added sequentially, which ensure that each additional LHC samples an area of the space that has not been sampled yet by the previous LHCs. Such a design provides the advantage of being able to robustly check the GP performance on well-designed sub-LHCs.

3.5 Step 4: Building emulators

The selected metric (see Step 1) is computed for each SCM simulation, noted $f(\boldsymbol{\lambda}_i)$ for $i = 1, \ldots, n$. These numbers serve as a training dataset for the building of an emulator. The emulator is then used to predict the metric values $f(\boldsymbol{\lambda})$ for any vector of parameter values $\boldsymbol{\lambda}$ in the input space.

Specifically, we use a Gaussian process (GP), a well known statistical model which has the advantage of interpolating observed model runs and provides a probabilistic prediction. The emulator gives a probability distribution for f written as:

$$f(\boldsymbol{\lambda}) \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim \mathrm{GP}\left(m(\boldsymbol{\lambda}, \boldsymbol{\beta}), k(\cdot, \cdot, \sigma^2, \boldsymbol{\delta})\right),$$

where $m(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is a prior mean function with parameters $\boldsymbol{\beta} = (\beta_i)_i$ and k a specified kernel (a covariance function describing the covariance between any 2 points). The kernel has a parameter that normally controls variance, σ^2, and parameters δ_k for each dimension of the input parameter λ_k that control the correlation attributed to each input. To start with, we assume a stationary kernel, i.e., the covariance only depends on the distance between points and not the absolute position. The GP is such that any finite collection $f(\boldsymbol{\lambda}_1), \ldots, f(\boldsymbol{\lambda}_n)$ has a multivariate normal distribution with mean vector $m(\boldsymbol{\lambda}_1, \boldsymbol{\beta}), \ldots, m(\boldsymbol{\lambda}_n, \boldsymbol{\beta})$, and variance matrix $\boldsymbol{\Sigma}$ with $\Sigma_{ij} = k(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j, \sigma^2, \boldsymbol{\delta})$. Let the training data be $\boldsymbol{F} = (f(\boldsymbol{\lambda}_i))_{i=1,\ldots,n}$, then

$$f(\boldsymbol{\lambda}) \mid \boldsymbol{F}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta} \sim \mathrm{GP}\left(m^*(\boldsymbol{\lambda}, \boldsymbol{\beta}), k^*(\cdot, \cdot, \sigma^2, \boldsymbol{\delta})\right),$$

where there are well-known closed form expressions for m^* and k^* (Williamson et al., 2017). Note that m^* and k^* are the updated mean and covariance representing what the emulator has 'learned' from the data, $\boldsymbol{F}$.

Whilst there are many possible prior choices of m and k, htexplo uses a 2-phase approach. First, we impose a structured mean surface $m(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{g}(\boldsymbol{\lambda})$ as a linear combination of simple functions of the input parameters contained in the vector $\boldsymbol{g}(\boldsymbol{\lambda})$ (e.g. monomials, Fourier functions and interaction terms are chosen through the forwards selection and backwards elimination method described in Williamson et al. (2013)). In the second stage, we use the squared exponential kernel function and Hamiltonian Monte Carlo (HMC, implemented in Stan – Carpenter & Coauthors, 2017) to sample from the posterior distribution of the parameters $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\delta}$ given $\boldsymbol{F}$ (note that the mean surface $m(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is not directly fitted in phase 1, but its structure is chosen, with Bayesian inference ultimately used in fitting for phase 2).

The choice of HMC implemented in Stan was motivated by requiring robust automation of emulator building across many metrics and cases, without needing the constant statistical expertise to diagnose MCMC convergence issues and to fix them by hand each time. Stan affords us with the ability to specify flexible and intuitive priors, and we use weakly informative priors as advocated by Gelman (2006). With the exception of the intercept term (which is uniform), our prior for each $\boldsymbol{\beta}$ is $\mathrm{N}(0, 10)$ and we use the OLS fitted values as starting values for the HMC. We set $\delta_k \sim \mathrm{Gamma}(4, 4)$ for all k to allow a wide range of potential correlation structures (this is a weakly informative prior) whilst penalising very small values that typically have high likelihoods, but lead to emulators with no predictive power (for discussion, see Volodina, 2020). Our prior for σ^2 is a truncated Normal (at 0), with mean at the residual from our OLS fits, and variance set using the variability of the ensemble (full details for these choices in Volodina, 2020).

The emulator is then tested using standardized Leave One Out diagnostics on the training data. These tests remove one point at a a time from the training set and use the emulator fitted on the remaining data to predict the removed point. Repeated over the training set, we then check whether the majority of left out points lie within 95% prediction intervals (we would expect 5% to miss). We also remove sub-designs from the training set and attempt to predict the whole sub-design, again checking to see if we have good posterior coverage of the ensemble. If the emulator fails these checks we revisit the computation of the emulator. For example, the procedure described in Volodina and Williamson (2020) (and available in htexplo) can be used to derive an appropriate non-stationary kernel k before refitting the emulator by HMC. Once fitted, the GP expectation $\mathrm{E}\left[f(\boldsymbol{\lambda})\right]$

provides an estimation of the metric for any given $\boldsymbol{\lambda}$, and its variance $\mathrm{Var}\left[f(\boldsymbol{\lambda})\right]$ provides an uncertainty around this estimation.

SCM runs are computationally cheap, but the fitted emulators are even cheaper and thus allow the computation of millions of predictions, with associated uncertainties, in a short time (a few minutes). This enables us to numerically define the space containing acceptable sets of parameters with respect to the chosen metrics and in particular, to visualize it (Step 5). The choice of Stan has proven effective for this project, though it does not scale well to larger ensembles. Going forward, a new version of the tools defaulting to MAP estimation and using efficient parallel implementation has just been released enabling millions of predictions in just a few seconds (Williamson & Volodina, 2020).

3.6 Step 5: History matching

The htexplo tool relies on the history matching technique, which seeks to rule out parameter values from the input space that are "implausible", given the SCM behavior for these parameter values and the sources of uncertainty. These sources include the reference (observation) error, treated as a random quantity with mean 0 and variance $\sigma^2_{r,f}$, and the SCM discrepancy, which has mean 0 (unless the user knows the direction in which the model is biased) and variance $\sigma^2_{d,f}$ (Sexton et al., 2011). The emulator is used to estimate the model behavior on a much larger sample of the input space than possible with the SCM. To history match the SCM behavior, we introduce the "Implausibility" measure for the metric f (Williamson et al., 2013), $I_f(\boldsymbol{\lambda})$, which is a distance between the metric prediction $f(\boldsymbol{\lambda})$ by the emulator at $\boldsymbol{\lambda}$, and the reference metric value, r_f, with respect to the norm induced by our second-order uncertainty specification, noted $|| \, ||_H$ below. The Implausibility reads

$$\begin{aligned} I_f(\boldsymbol{\lambda}) = ||r_f - f(\boldsymbol{\lambda})||_H &= \frac{|r_f - \mathrm{E}\left[f(\boldsymbol{\lambda})\right]|}{\sqrt{\mathrm{Var}\left[r_f - \mathrm{E}\left[f(\boldsymbol{\lambda})\right]\right]}} \\ &= \frac{|r_f - \mathrm{E}\left[f(\boldsymbol{\lambda})\right]|}{\sqrt{\sigma^2_{r,f} + \sigma^2_{d,f} + \mathrm{Var}\left[f(\boldsymbol{\lambda})\right]}}. \end{aligned} \tag{5}$$

The model discrepancy for the metric f, $\sigma_{d,f}$, accounts for the model structural error due to the inherent inability of the SCM to reproduce the LES exactly (due to unresolved physics or missing processes, for example). It could be defined as the minimum error possible when exploring the full set of parameters, however, this could permit the SCM to be close to the reference for the wrong reasons and does not account for multiple metrics and cases, so we avoid this definition. Instead it is typically defined to be the uncertainty left in the difference between the SCM metric when the parameters are fixed at their best values (fixed the same for all metrics) and the references. This quantity is perhaps the target of model development in the first place and, as such, is unknown. For example, suppose we want to test the ability of a new parameterization to capture the behaviour of the reference. With the standard definition of discrepancy, the uncertainty needed so that the new parameterization captures the behaviour of the reference, it is not clear how to proceed with testing. Our approach instead is to treat model discrepancy as a “tolerance to error” as detailed in Williamson et al. (2017). The tolerance to error is the distance between model results and the reference that the modeler would be satisfied with, enabling modellers to place confidence in certain metrics/parts of their parameterization, and relax restrictions on others as needed. As illustrated in Sect. 4 and Part II, defining this tolerance to error can be a difficult a-priori task; however experimenting with this value provides important insights into the behavior and its inherent limitations. The most attractive feature of this approach to discrepancy is that, for a given tolerance to error, if the induced NROY space is empty it means that the parameterization is not able to reproduce the reference under the given tolerance. Either the tolerance can be relaxed, accepting the limitations of the current set of parameterizations, or the parameterization can be revisited.

The implausibility defines a membership rule for NROY space after the first iteration:

$$\mathrm{NROY}_f^1 = \{\boldsymbol{\lambda} \mid I_f(\boldsymbol{\lambda}) < T\}.$$

where T is a chosen threshold (or cutoff). For scalar metrics, it is standard to use $T =$ 3 justified using Pukelsheim's rule that states 95% of the probability density for any unimodal distribution is within 3 standard deviations of the mean (Pukelsheim, 1994). Using this threshold makes it unlikely that good parameter values are ruled out by chance. To measure and visualize NROY space the Implausibility $I_f(\boldsymbol{\lambda})$ is calculated on a random LHC sampling of a large number (on the order of hundreds of thousands or millions) of vectors $\boldsymbol{\lambda}$.

Note that $I_f(\boldsymbol{\lambda})$ can be smaller than the chosen threshold T either because $\mathrm{E}\left[f(\boldsymbol{\lambda})\right]$ is close to the reference or because the sum of the different errors is large. When the uncertainty of the emulator is larger than the tolerance to error and observation error, points that should be ruled out are kept in the NROY. In this case, further iterations are desirable in order to increase the density of the sampling of NROY and hence improve the emulator quality and reduce the associated uncertainty.

3.7 Iterative refocusing and multi-metrics

One advantage of this method is to progressively optimize the design of simulations to be run. New simulations are iteratively added only where it is useful to increase the emulator accuracy. This is performed by iterating the same process previously described several times in "waves", (this is termed "iterative refocusing" and is a fundamental part of the history matching approach). Each new iteration n starts from the remaining space NROY_f^{n-1} estimated at the end of the previous wave. Because of its complex geometry, a LHC sampling, as in the first wave, cannot be applied, and therefore the remaining space is re-sampled uniformly. A new SCM simulation ensemble is performed with this design and is used to proceed with steps 4 and 5. The new emulator is only valid in the new parameter space, namely NROY_f^{n-1}. Outside this space, we rely on the emulators from the previous waves. As in Step 5, to measure and visualize NROY_f^n, the implausibility is computed over a large number of points in the input space. The threshold T may be varied between waves, but we advise to keep it to 3 as long as the process has not converged (i.e. the emulator variance within the current NROY space remains large – see also Sect. 4 and Part II). The iterative refocusing stops when the convergence of the sequence $(\mathrm{NROY}_f^n)_n$ has been qualitatively achieved.

So far, we have considered only one metric, but several metrics $(f_k)_k$ can be combined at the same time. An Implausibility is then computed for each metric and the total NROY^n space is the intersection of the $\mathrm{NROY}_{f_k}^n$ associated with each metric:

$$\mathrm{NROY}^n = \bigcap_k \mathrm{NROY}_{f_k}^n = \left\{ \boldsymbol{\lambda} \mid \#\{k \mid I_{f_k}^n(\boldsymbol{\lambda}) > T\} \leq \tau \right\},$$

$\#$ represents the number of metrics fulfilling the condition indicated into brackets (where the implausibility is greater than the threshold) and τ, the number of metrics for which the model is allowed to be far from the reference while still kept in the NROY space. If $\tau = 0$, all metrics must satisfy our implausibility cutoff. If there are a large number of metrics then τ should be increased ($\tau \geq 1$) to avoid multiple testing problems meaning that too many good parameter values are ruled out by chance. If a modeller seeks to prioritize certain metrics, they can either be introduced in early waves, ensuring that the NROY space satisfies priority metrics first before introducing new ones, or the tolerance to error, which is defined for each metric, can be used to impose priorities (a larger tolerance to error induces a less constraining metric).

3.8 Sensitivity analysis provided by the tool

The htexplo tool provides its own sensitivity analysis, which, due to the use of multi-wave history matching, is rather different from traditional methods applied to models throughout the literature. Traditional methods, either derivative-based (Saltelli, 2002), or variation-based (Oakley & O'Hagan, 2004), essentially seek to identify which parameters modify model output. This can help focus further study, model development or even observation collection to help understand these parameters. Note that the htexplo tool provides at the first iteration a sensitivity analysis over the entire space where correlation among parameters is included as the parameters are not varied one at a time.

However, for calibration purposes, once history matching is considered as a valid approach for a given model, the sensitivity analysis should not be done on the full model input space. By using history matching, we acknowledge that there is a large part of the model parameter space that is not useful for understanding reality. The Gaussian processes remove this uninformative space in order to target the space where the model becomes useful. Once we have this useful subspace, the usual and important questions that are posed by sensitivity analysis should be considered. For example, how is the model output changing as we move through parameter space and which parameters are responsible for these changes? As all models within the NROY space are consistent with our metrics, sensitivity analysis as described here is now really focused on the relevant subspace. Note that sensitivity analysis on the original input space does not answer these questions. Seen through the history matching lens, on the full space, sensitivity analysis is showing us which parameters are responsible for the variability in the space we are about to cut. Whilst informative for helping us cut the space efficiently, sensitivity analysis is not necessary at this stage. Our methods are already efficiently able to do this.

Performing variance-based sensitivity analysis in NROY space is not trivial and we are not aware of any methods that are currently able to do this. Variance-based sensitivity analysis requires independent input spaces (which is what we always start with in wave 1). But after cutting space, we have complex relationships between the parameters. NROY space may not even be simply connected, and can be highly non-linear. Efficient methods for calculating sensitivity in these unusual spaces would be interesting to apply for history matching as an avenue for further research.

As a practical tool, the density plots such as those given in Fig. 5, provide their own type of second-order sensitivity analysis. They allow us to see, as we move in two dimensions of a parameter space, how the shape is changing and, moreover, which combinations of parameters it is important to get right and, not usually included in a sensitivity analysis, how they need to be set in order to get sensible answers. As well as all of the benefits we have for tuning, we would argue that history matching is achieving many of the same things that a sensitivity analysis achieves in terms of informing the modelling, but concentrated only on the model input space that is consistent with the observations.

3.9 On the use of history matching and the avoidance of optimization

Whilst History matching is well established and is being used in a growing number of climate studies, other methods of calibration are more popular and we believe should be avoided for process-based model development. Whilst many methods based on optimizing a cost function exist (Hourdin et al., 2017), the most popular in the UQ community is Bayesian calibration (Kennedy & O'Hagan, 2001). Bayesian calibration requires a similar set up to history matching (emulators, observation errors and model discrepancy) and then jointly finds the posterior probability distribution of the "best" value of the input parameters and the model discrepancy (strong prior information on the discrepancy is required to make this sensible, Brynjarsdóttir & O'Hagan, 2014). Optimization methods like these do not afford us with the chance to falsify a parameterization (they always find the best value), nor do they give all parameter values that are consistent with the observations (in our case reference LES) that can then be used when tuning the 3D model (see Part II).

4 Illustration of htexplo on a simple case

In this section, the use of htexplo is illustrated for the ARPEGE-Climat 6.3 atmospheric model based on a single 1D case. More comprehensive exploitation of the tool will be given in Part II.

4.1 Model, parameters and case-study

ARPEGE-Climat 6.3 is the atmospheric component of the CNRM-CM6-1 climate model (Voldoire et al., 2019; Roehrig et al., 2020). It has 91 vertical levels, 15 of them below 1500 m. The model time step is 15 minutes. Here, we use its SCM version and focus on its representation of a clear convective boundary layer. To simulate the processes involved in the boundary layer, the model combines a turbulence scheme with a mass-flux scheme, thus following the Eddy-Diffusivity Mass-Flux framework (e.g. Hourdin et al., 2002; Soares et al., 2004; Siebesma et al., 2007; Pergaud et al., 2009). The mass-flux scheme represents convection in a unified way from the clear convective boundary layer regime to the shallow cumulus and deep convection regimes (Piriou et al., 2007; Gueremy, 2011). In this section, we aim at analyzing the importance of the values of free parameters of the turbulence scheme on the simulation of an idealized clear boundary layer. A boundary-layer-top vertical entrainment is activated in the default version of ARPEGE-Climat 6.3 (see (Roehrig et al., 2020)). For the sake of simplicity of the present illustration, and also because this parameterization is weakly active in the analyzed case, it is fully deactivated in the following section. Similar results are obtained when it is activated.

The turbulence scheme is based on Cuxart et al. (2000) which aims at providing the vertical turbulent fluxes from which the turbulent source term is derived for the prognostic variables (see more details in Roehrig et al., 2020). The scheme relies on a prognostic equation of the grid-scale turbulence kinetic energy, $\overline{e}$:

$$\frac{\partial e}{\partial t} = \frac{-1}{\rho}\frac{\partial(\rho\overline{w'e'})}{\partial z} - (\overline{w'u'}\frac{\partial\overline{u}}{\partial z} + \overline{w'v'}\frac{\partial\overline{v}}{\partial z}) + \beta\overline{w'\theta'_{vl}} - \frac{\overline{e}^{3/2}}{L_\epsilon} \tag{6}$$

where the advection terms, the pressure fluctuations and the diffusion transport have been neglected. ρ is the air density, w the vertical velocity, u and v the zonal and meridional wind components, β is the buoyancy parameter (equal to $\frac{g}{\theta}$ with g the gravitational constant, θ being the potential temperature), θ_{vl} is the liquid virtual potential temperature and L_ϵ the dissipation length. Primes indicate fluctuations with respect to the grid-scale values indicated with overbars. The different turbulent vertical fluxes are diagnosed using $\overline{e}$ following, for any variable φ:

$$\overline{w'\varphi'}(z) = -K_\varphi\frac{\partial\overline{\varphi}(z)}{\partial z} \tag{7}$$

with

$$K_\varphi = \sqrt{e}L_m A_\varphi \Phi_\varphi \tag{8}$$

with Φ_φ a stability function also computed at each altitude (for more details see Cuxart et al. (2000)) and A_φ a free parameter. The mixing length, L_m, is computed following Bougeault and Lacarrere (1989); it consists in computing the vertical displacement an air parcel can travel upwards and downwards with its available turbulence kinetic energy according to the thermal stratification. Also, L_ϵ in Eq. 6 is defined by $L_\epsilon = A_\epsilon \times L_m$ with A_ϵ another free parameter. Finally, we have selected three parameters for this analysis namely, A_ϵ controlling the expression of the dissipation length-scale as a function of the mixing length-scale and A_U and A_T that respectively enter into the expression of the exchange coefficient in Eq. 8 for the wind and the temperature (the same coefficient, A_U, is used for both the zonal and meridional component of the wind). The range of variation explored for each parameter is indicated in Table 1 and the parameters are varied linearly in those ranges (when parameter ranges span many orders of magnitude, we typically vary them on a log scale and htexplo is set up to do this). The turbulence parameterization includes other free parameters but to keep the example simple, the three most influencial parameters for this case have been selected and no free parameters of the mass-flux scheme are considered.

Table 1. List of the free parameters of the turbulence scheme that are varied in this example with default values and range of variation

Names	A_U	A_ϵ	A_T
Default	0.126	0.85	0.14
Minimum	0.01	0.1	0.01
Maximum	0.4	3.	1.

To keep the example simple, only one case is used here. This case is a dry idealized case of a convective boundary layer with a constant-in-time large surface sensible heat flux of 0.24 Kms^{-1} with a strongly capped boundary layer documented in Ayotte et al. (1996), called 24SC in the following. The importance of combining different cases will be illustrated in part II.

We first document a sequence of three waves where additional metrics are added at each iteration (Experiment 1). We will then discuss the results obtained when adding all the metrics directly at wave 1 (Experiment 2), varying the threshold used to determine the NROY (Experiment 3 see also Sect. 3.5), using more SCM runs (Experiment 4), and varying the tolerance to error (Experiments 5 and 6).

4.2 Three consecutive waves adding metrics progressively

For the first iteration (or wave in the following) of Experiment 1, 30 SCM simulations of the 24SC case were realized by varying values for the three parameters exploring at best (using a LHC sampling, see Sect. 3.4) the range of each parameters (Table 1). Figure 3 illustrates that the parameters are randomly sampled as indicated by the distribution of the black dots along the different x-axes. Three different metrics are used to characterize the turbulent mixing in the boundary layer and are progressively introduced through the successive waves. The first chosen metric is the potential temperature averaged over the layer 400-600 m. It is a good proxy for the boundary-layer potential temperature, which is well mixed between the surface and the boundary-layer top, located around 1300 m. This metric is computed for the 30 SCM runs; these computations serve as training data for the construction of the emulator. The prior mean function (see Sect. 3.5), m for this emulator is a sum of linear and quadratic functions of the parameters. The stationary squared-exponential kernel provides a sufficient fit to the data according to the leave-one-out methodology explained in Sect. 3.5. Figure 3 presents the variation of the metric as a function of the parameters: some first-order relationships appear with the boundary-layer potential temperature increasing with A_U and A_T to a lesser extent A_T (due to an increased mixing associated to a larger diffusivity and larger fluxes) and decreasing with A_ϵ (due to a reduced mixing because of the increased dissipation). For this metric, we have chosen a tolerance to error of 0.5 K, a difference between SCM results and LES we are satisfied with. This may be a bit large for this very idealized case (with no moisture, an already convective initial state) but this is an error we will be satisfied with generally for boundary-layer potential temperature. Given this tolerance to error (indicated by the dashed horizontal grey line), the metric does not provide much constraint on the model behavior and the entire initial parameter space is kept (c.f. Table 2). Note that this tolerance to error is much larger than the uncertainty around the LES ($\sigma_{r,f} = 0.075$ K) and the emulator ($\mathrm{Var}\left[f(\boldsymbol{\lambda})\right] = 0.042$ K). Sect. 4.3 details the effect of a reduced tolerance to error.

A second wave is realized, with 30 runs sampling the NROY space of the first wave (the previous 30 SCM runs could also have been used for efficiency), which is in fact the

Table 2. Description of the model discrepancy (Disc.) of the given metric (indicated in the 2^{nd}, 3^{rd} and 4^{th} columns), the Cutoff, threshold used for Implausibility (5^{th} column), the Not-Ruled-out-Yet Space (fraction in % of initial space of parameters, 6^{th} column) and the emulator uncertainty quantified as the emulator standard deviation for each metric (7^{th} column) for each Experiment and wave.

N^o Expt N^o Wave	Disc. θ_{BL} [K]	Disc. Ay_θ [Kms^{-1}]	Disc. ws_{BL} [m s^{-1}]	Cutoff	NROY= % of initial space	Emulator Error for θ_{BL} / Ay_θ / ws_{BL}
Exp1-1	0.5	-	-	3	100	0.042/-/-
Exp1-2	0.5	0.05	-	3	30	0.022/0.014/-
Exp1-3	0.5	0.05	1	3	23	0.069/0.023/0.049
Exp2-1	0.5	0.05	1	3	40	0.042/0.019/0.22
Exp2-2	0.5	0.05	1	3	38	0.033/0.017/0.06
Exp2-3	0.5	0.05	1	3	27	0.13/0.036/0.14
Exp3-1	0.5	0.05	1	3	72	0.022/0.063/0.019
Exp3-2	0.5	0.05	1	3	32	0.060/0.021/0.15
Exp3-3	0.5	0.05	1	2.5	22	0.092/0.026/0.054
Exp3-4	0.5	0.05	1	2.	15	0.076/0.019/0.061
Exp4-1	0.5	0.05	1	3	27	0.038/0.013/0.033
Exp5-1	0.25	0.025	0.5	3	32	0.043/0.020/0.21
Exp6-1	0.1	0.01	0.25	3	31	0.041/0.020/0.21

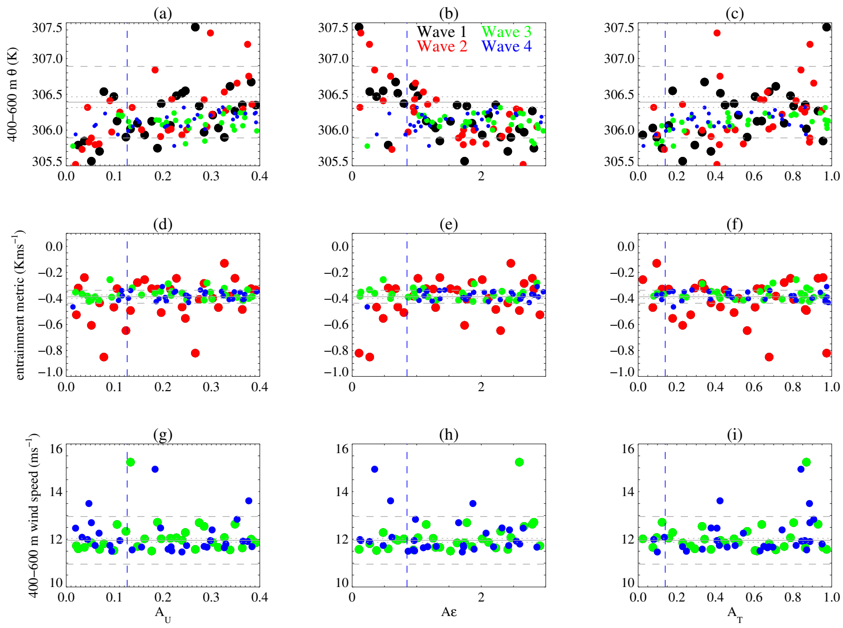


Figure 3. The three metrics, boundary-layer potential temperature (a–c), entrainment metric (d–f) and boundary-layer windspeed (g–i) are plotted as a function of the value of each parameter, A_U (a, d, g), A_ϵ (b, e, h) and A_T (c, f, i). A different color is used for the different waves of Experiment 1 (black for Wave 1, red for Wave 2, green for Wave 3 and blue for Wave 4). The vertical dashed blue line corresponds to the default value of the parameter used in the model, the horizontal thin full grey line correspond to the reference metric and the dotted lines indicates the uncertainty around this reference from the different LES simulations while the dashed lines indicate the tolerance to error around the reference.

entire initial parameter space as the first metric did not constrain the parameter space. Two metrics are computed from those 30 runs: the potential temperature averaged between 400 m and 600 m as in the first wave and the entrainment metric, A, quantifying the overshoot of the boundary layer relative to the initial profile as defined in Ayotte et al. (1996). A is computed as:

$$A = \frac{\int_{zi(t0)}^{H}(\theta(z,t_f) - \theta(z,t_0))dz}{t_f - t_0} = \frac{\int_{0}^{H}(max(\theta(z,t_f) - \theta(z,t_0), 0))dz}{t_f - t_0}$$

t_0 being the initial time, t_f the time at which the metric is computed and H the top of the model or a level largely above the boundary-layer top. This metric is less commonly used for evaluating models and it was more difficult to specify a tolerance to error which was taken as 0.05 K.m s^{-1}. An emulator is built for each metric. The second metric is more restrictive and the NROY space is now reduced to 30% of the initial parameter space (Table 2). The obtained NROY (not shown) is not very different from the one obtained for the third wave. It excludes values of the parameters that lead to simulations with too large or too small entrainment metric as indicated by the differences between the red dots and the green ones in Fig. 3.

A third wave is realized, with 30 new SCM runs sampling the new NROY. Three metrics are computed from those 30 runs: the two previous ones plus the wind speed averaged between 400 m and 600 m. For this last metric, we fixed the tolerance to error to 1 m s^{-1}. After this third iteration, the NROY is 23% of the initial space. As shown in Fig. 4, the spread of the different simulations that sampled the parameter values reduces progressively throughout the different waves and this tool allows to discard values of parameters that induce a too deep boundary layer. The wind-speed profiles did not completely converge and this is associated to the observation uncertainty which has been fixed to 1 ms^{-1}.

The final NROY space after the third wave is shown in Fig. 5. The metrics tend to reject preferentially low values of A_ϵ with high values of A_U or high values of A_ϵ with low values of A_U underlying some correlation between these two parameters. Note the default values of the parameters are within the NROY space confirming that they correspond to an acceptable calibration of the turbulence scheme, given the chosen tolerance to error and the LES uncertainty. This is also confirmed by the simulations of the last wave having a behavior similar to the default simulation as shown in Fig. 4.

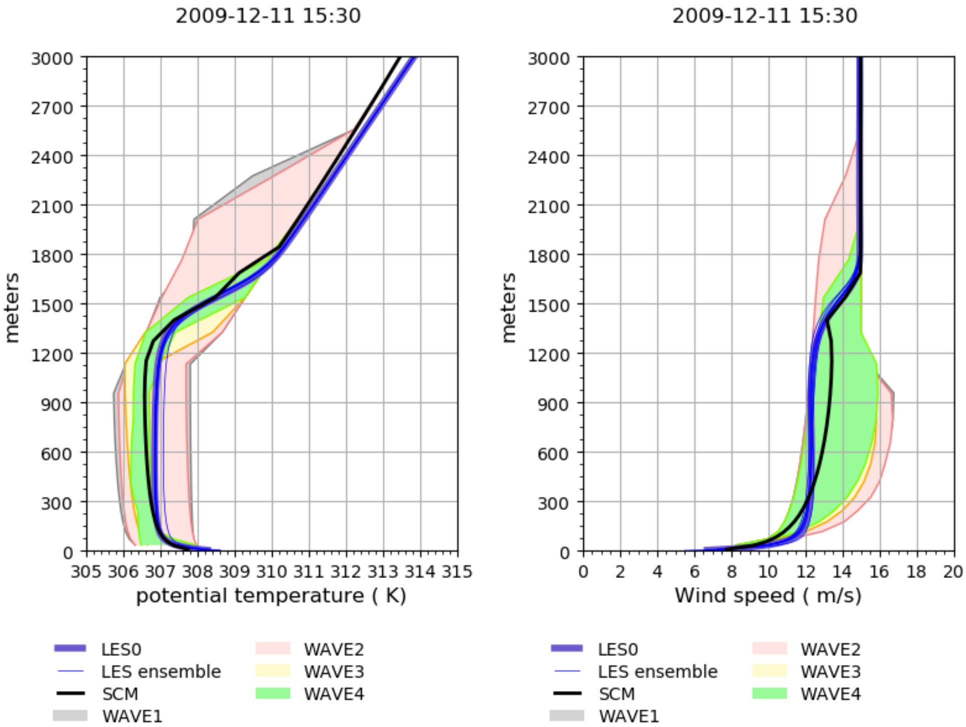


Figure 4. Vertical profile of (a) potential temperature and (b) wind speed for the last hour of the simulation with the spread of the ensemble of simulations used for the different waves indicated in different color shadings for Exp 1, the default simulation is in black, the reference LES in thick dark blue and the different elements of the LES ensemble in thin blue lines.

The uncertainty around the LES obtained from eight different LES runs with slightly different configurations, detailed in the appendix, is 0.075 K for θ_{BL}, 0.014 K m s^{-1} for A_θ and 0.083 m s^{-1} for ws_{BL}, on the same order of magnitude of the emulator uncertainty. For the first metric and third metric, the tolerance to error is much larger than the uncertainties of the reference and the emulator while for the second metric the three uncertainties are of the same order of magnitude. Concerning the tolerance errors, we can conclude that for this case and the selected metrics,the SCM is good enough for a sub-domain of the initial parameter space.

4.3 Robustness

In this subsection, we analyze the sensitivity of the results to i) the sequence of introduction of metrics (Experiment 2 uses the three metrics directly at wave 1), ii) the threshold used to determine the NROY space (Experiment 3), iii) the number of SCM runs used to form the training dataset (Experiment 4), and, iv) the tolerance to error (Experiments 5 and 6).

If the three metrics are introduced directly in the first wave (Experiment 2), the NROY space is similar to the one obtained after three waves (see Table 2 and Fig. 5) although the NROY space is larger (40% against 23%). Repeating more waves with the same metrics allows to progressively converge to the same NROY space. Note that a test with only one metric but the most constraining one, namely the entrainment metric, leads to very similar result ($NROY = 43\%$) for the first wave (not shown). Although not illustrated for this case, introducing one by one the metric, is sometimes important: i/ it can allow to give some priority among the metrics, finding first a space consistent with the first metric in which the second metric is then used as a constraint and ii/ if one metric has a strong non-linear behaviour reducing the initial parameter spaces with other metrics may ease the capacity of the emulator to reproduce the metric behaviour. These results also indicate that adding a new metric in the core of the process does not alter the selection, allowing to add supplementary metrics if one realizes that some behavior of the SCM is not constrained enough, a fundamental aspect of history matching.

In Experiment 3, we first realize two waves as in Experiment 2 and then progressively reduce the threshold used to determine the NROY space from 3 to 2.5 in Wave 3 and from 2.5 to 2 in Wave 4 (see Table 2) to explore the impact of less conservative threshold (a threshold of 3 corresponds to ruling out what exceeds three times the uncertainties and keeps 95% of the probability for any unimodal probability distribution). The differences in the NROY space of the first wave with Exp2-1 indicates that 30 SCM runs are probably not enough to robustly constrain the first iteration and more iterations are needed. Then, reducing the cutoff induces a smaller NROY space but the change is not radical. This was expected from the lower left figures of Fig. 5 that show the minimum value of the Implausibility for any variations of the other parameters (here, the third parameter). Indeed, the area with minimum value of $I_f(\boldsymbol{\lambda}) > 3$ (i.e. the points that are excluded from the NROY space whatever the value of the third parameter) is very similar to the area with minimum value of $I_f(\boldsymbol{\lambda}) > 2$.

All of the previous experiments have been realized using a rather small training dataset of 30 SCM runs (ten times the number of parameters). Experiment 4 has tested the impact of using 90 SCM runs instead of 30 for wave 1. This experiment produces directly a smaller NROY space (see Fig. 6) at the first wave than obtained from 30 SCM runs (see Exp3-1 or Exp2-1 in Table 2). Also, the emulator uncertainty is smaller for the first wave of Experiment 4 than the ones of the first wave of Experiment 2 or 3. A compromise must be found between a larger ensemble of simulations that increases robustness but is more costly.

The sensitivity to the tolerance to error is illustrated in Table 2 and Fig. 6 with Experiments 5 and 6. When reducing the tolerance to error by a factor of two the NROY space is 32% of the initial space in Exp5-1 (using the three metrics at once, so to be compared to 40%). The NROY space (31% of the initial space) is not much reduced further when reducing the tolerance to error twice more (Exp6-1), because the tolerance to error is not anymore the limiting uncertainty. It is interesting to note that even when strongly reducing the tolerance to error, the default values for the three selected parameters are still in the NROY space validating the choice of parameter values used in the control simulation. The lower left panel of the subfigures in Fig. 5 and Fig. 6 indicates the minimum Implausibility along the other dimensions of the space and as illustrated in Fig. 6, reducing the tolerance error (when larger than the other errors) induces a reduction of the denominator in the Implausibility and therefore an increase of Implausibility.

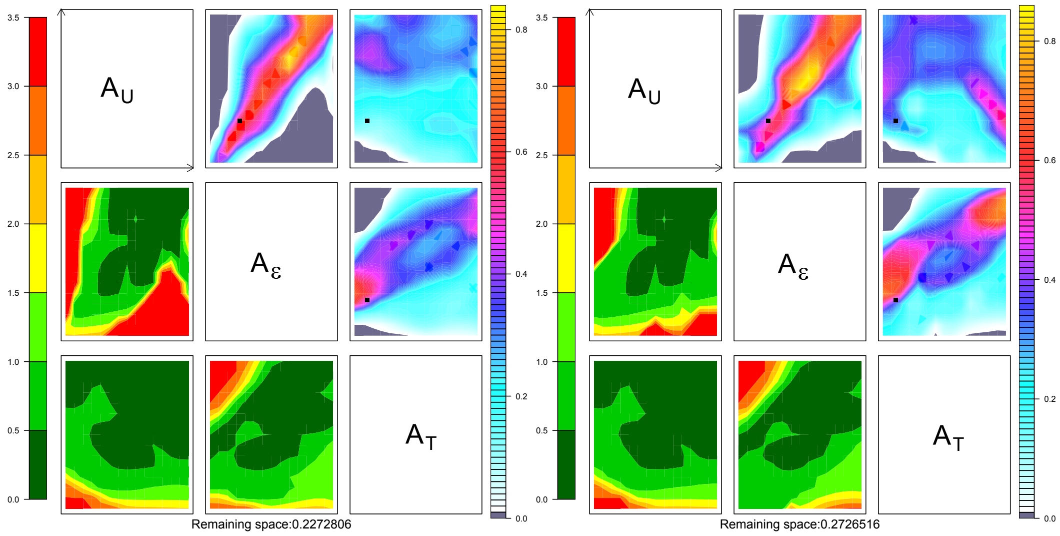


Figure 5. The left panel corresponds to the result of Exp1-3 and the right panel to Exp2-2. The upper right triangle contains 3 subfigures showing 2D sub-matrix. Each sub-matrix is a restriction to 2 parameters, the name of which are given in the diagonal of the main figure, and presents in colors the fraction of points with implausibility smaller than the threshold (here a value of 3). This fraction is obtained by fixing the two parameters at values of the x-axis and y-axis of the plotted location and searching the other dimensions (here the third dimension as we have only three parameters) of the parameter space. This allows to visualize in 2-D the full NROY which is 3-D here but can be n-D if n parameters are selected. The lower left triangle (with also 3 subfigures) presents the minimum value of Implausibility. These plots are orientated the same way as those on the upper triangle, for easier visual comparison. The black dots correspond to the default values used in the model.

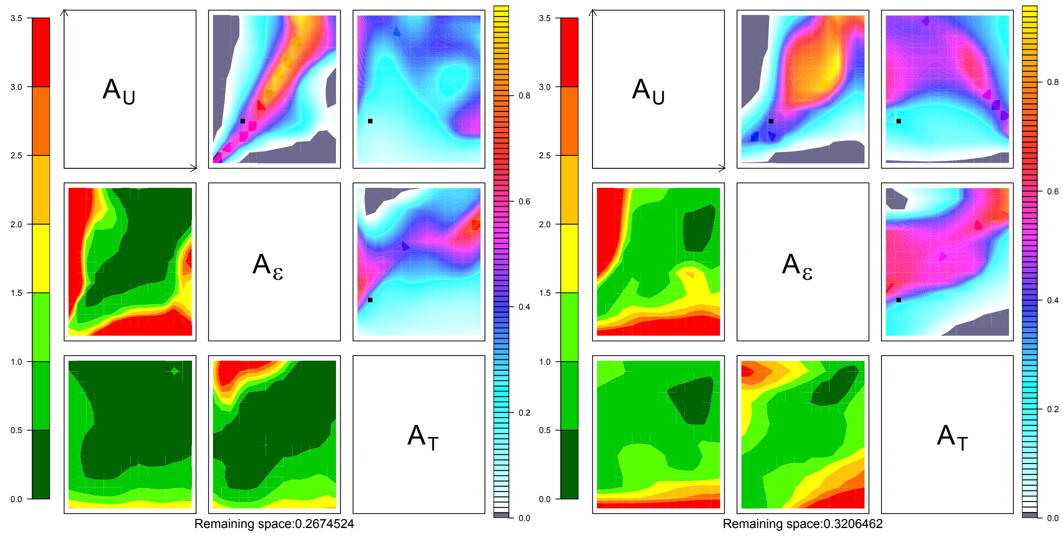


Figure 6. Same as Fig. 5 but for the sensitivity to the number of SCM runs (Experiment 4, left panel) and to the tolerance error (Experiment 5, right panel).

5 Conclusion

In this paper, we make a proposal to accelerate weather and climate model development. Our proposal tackles model development and calibration jointly. For that purpose, we have developed a tool that formalizes a process-based calibration, the *High-Tune Explorer* made available to the other modeling groups. It extensively exploits the SCM/LES comparison on a multicases, multi-metrics basis and benefits from machine learning techniques. In contrast with other recent proposals to use machine learning techniques in climate modeling, we keep parameterizations as key ingredients of these models because they summarize our current understanding of the main physical processes This choice is motivated in particular by the confidence needed when extrapolating the model results to a future climate.

The tool allows us to define the sub-domain of the parameter values for which SCM matches LES on selected metrics for a series of cases within a given uncertainty. The exploration of the free-parameter space is facilitated using Gaussian process emulators. These emulators, once trained on a limited number of real simulations, predict the SCM with uncertainty in a much shorter time than required to run the SCM. History matching using the emulator is performed iteratively to progressively shrink the space of acceptable parameter values. This iterative approach contrasts with the more traditional tuning strategy based on optimization, which i) seeks an individual "best" value where the SCM minimizes a cost function computed for given metrics, ii) is strongly dependent on the weights given to the metrics and iii) is highly sensitive to the choice of metrics. By pursuing a strategy for discarding parameter values, we are left with a free parameter domain that is (i) consistent with the metrics we have chosen, (ii) can be further reduced by introducing new metrics or altering our tolerance to model error, and (iii) does not claim a single best simulation which may be over-fitted to one or more metrics, needlessly biasing the simulation and potentially leading to less physical behavior, as the model is projected into different regimes, than other choices in our not-ruled-out-yet space. Our tool formalizes the consideration of the different sources of uncertainties associated to the reference, the statistical tool and the model. For the latter, we take a "tolerance to error" approach, allowing the question of whether a parameterization can match our reference as well as we think it ought to (based on any physical limitations we believe should be there), and enabling us to revisit those expectations and to understand the model's limitations throughout the process.

In the present study, we present applications of the *High-Tune Explorer* to the SCM/LES framework, focused on the repesention of the atmospheric boundary layer. We have illustrated how this tool allows us to objectively verify choices that have been made by model developers for the free-parameter values. Experimenting with the combination of the metrics with this tool allows us to clarify the importance of a given metric, the number or combination of metrics that should be used, and the possible redundancy between metrics all in an efficient way that was not possible without it. The tool also enables us to include new metrics at a new iteration so that we can pursue the calibration exercise, even though one realizes an important deficiency of the model is not addressed by the previously selected metrics. Our framework allows a progressive addition of metrics, cases or a gradual reduction of the tolerance to error and is therefore very flexible.

Although this new framework is tested here for the improvement of boundary-layer processes (turbulent transport in Part I and cloud representation in Part II) by running the full atmospheric physics on one model column considering well established test cases for which LES are particularly relevant, it has much broader application. It can be used for instance to calibrate elementary pieces of parameterization (e.g., entrainment formulation) without time integration. This methodology can be easily expanded to other parameterizations as well. The key ingredient for doing this is a reliable reference with documented uncertainty. This reference could come either from a detailed modeling of the process, as done here with LES, or from observations as long as the other sources of discrepancy as the uncertainty coming form the case definition are documented. Proposing new relevant metrics and estimation of associated uncertainties will become valuable now that we know how to include them in the model improvement process. An effort is currently done in that direction in parallel to the work presented here, consisting in providing reference radiative transfer computations on the classical cloud test cases currently used for parameterization development or (here) tuning. The development of the parameterization of boundary layer and clouds based on SCM/LES comparisons was indeed focused so far on the prepresentation of atmospheric transport and macrophyics of clouds, but the radiative transfer computations run in LES models were often not more reliable than those used in GCM. By developing fast and accurate radiative tools that accounts for the full 3D radiative transfer in LES cloud schene, as proposed by Villefranque et al. (2019), we can compute many types of radiative metrics, from monochromatic, local, and directional observable to integrated energetic quantities. The use of such radiative metrics will allow us to tackle calibration of radiative parameterizations but also to better link the calibration realized at the level of the parameterizations itself with the one realized for the final full 3D model calibration, which mainly targets the radiative forcing of the atmospheric general circulation.

To sum-up, the appication of the *High-Tune Explorer* on SCM/LES comparisons allows us: (i) to quantify the parametric uncertainty at process level, (ii) to identify parameters which limit model performance, whatever their value, and should be replaced by a more physical parameterization, and (ii) to reduce the domain of acceptable values of free parameters used in the final tuning of the global model.

We show indeed in Part II how the tool applied first to SCM/LES comparisons, on a multicase basis, can be used to reduce the range of acceptable values for the calibration of the complete 3D model configuration and considerably accelerate the resource and time consumption for this step of model development. The final 3D tuning becomes a part of the history matching process, by adding new metrics or constraints using the exact same codes.

We believe that this tool is a breakthrough for model development as it allows us to place the importance of the physical understanding of the processes at the heart of model development, based on an extensive use of the SCM/LES comparison, whilst harnessing important techniques in machine learning and uncertainty quantification. We advocate that the approach presented here leads to a well-defined strategy for calibration of the full model that may change the way we do climate modeling and result in a significant acceleration in model improvement.

Appendix A The different Large-Eddy Simulations

In total, eight different simulations have been run with Meso-NH (Lac et al., 2018), varying the resolution, domain size, turbulence formulation, intensity of the white noise introduced at the first level and initial time to trigger turbulence, activation of subgrid condensation and changes in the microphysics scheme for the cloudy cases. The Table A1 lists the different simulations of the Ayotte case used in Sect. 4 to estimate the uncertainty associated to the reference LES and the Table A2 lists the different simulations of the ARMCU case used in Sect. 3 to estimate the uncertainty associated to the reference LES. The reference LES is highlighted in bold.

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.B eyond the presentation of a new approach that we think could constiute a break through in climate model improvement, we intend to provide a tool for the climate community. All the programs, scripts and reference LES used will be made available together with the paper.

References

Ahmat Younous, A.-L., Roehrig, R., Beau, I., & Douville, H. (2018). Single-column modeling of convection during the cindy2011/dynamo field campaign with the cnrm climate model version 6. *Journal of Adavnces in Modeling Earth Systems*, *10*, 578–602.

Table A1. List of the different LES runs of the Ayotte case used to determine the uncertainty around the reference

Name	Resolution	White noise	Turbulence	Diffusion
Name	Dx, Dz	Standard deviation (K)	length-scale	Timescale
Reference	50 m,nested <25 m	0.01 K	Deardorff length scale	1800 s
WhiteNoise	”	0.1 K	”	”
WhiteNoiseLL	”	0.5 K	”	”
Turb	”	”	size of the grid	”
Difshort	”	”	”	300 s
Diflong	”	”	”	7200 s
Dx	25 m, ”	”	”	”
Dz	”, nested <12.5 m	”	”	”

Andrianakis, I., McCreesh, N., Vernon, I., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., … White, R. G. (2017). Efficient history matching of a high dimensional individual-based hiv transmission model. *SIAM/ASA Journal on Uncertainty Quantification*, *5*(1), 694–719.

Ayotte, K. W., Sullivan, P. P., Andren, A., Doney, S. C., Holtslag, A. A., Large, W. G., … Wyngaard, J. C. (1996). An evaluation of neutral and convective planetary boundary-layer parameterizations relative to large eddy simulations. *Boundary-layer Meteorol.*, *79*, 131–175.

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., … Webb, M. J. (2015, April). Clouds, ciruclation and climate sensitivity. *Nature Geoscience*, *8*(20), L20806. (WOS:000233104900005) doi: 10.1038/NGEO2398

Bougeault, P., & Lacarrere, P. (1989). Parameterization of orography induced turbulence in a mesobeta-scale model. *Mon. Wea. Rev.*, *117*, 1872–1890.

Brenowitz, N. D., & Bretherton, C. S. (2018, June). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. (WOS:000438499100052) doi: 10.1029/2018GL078510

Table A2. List of the different LES runs of the ARMCU case used to determine the uncertainty around the reference; the names indicated in the left column are those used in the legend of Figure 2

Name	Horizontal Resolution	Vertical Resolution	Domain side	Subgrid Condensation	Microphysics	Turbulence mixing length
12Dx25z25	25 m	25 m	12.8 km	No	Warm (ICE3)	Deardorff
6Dx25z25	”	”	6.4 km	”	”	”
6Dx40z25	40 m	25 m	6.4 km	”	”	”
6Dx40z40	40 m	40 m	6.4 km	”	”	”
6Dx25zvar	25 m	stretched grid	6.4 km	”	”	”
6Dx100z40	100 m	40 m	6.4 km	”	”	”
25Dx100z40	100 m	40 m	25.6 km	”	”	”
51Dx100z40	100 m	40 m	51.2 km	”	”	”
6DelDx25z25	25 m	25 m	6.4 km	”	”	$(Dx * Dy * Dz)^{1/3}$
6SbgDx25z25	25 m	25 m	6.4 km	Yes	”	Deardorff
6NprDx25z25	25 m	25 m	6.4 km	No	Only saturation adjustment	”

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*, 2854–2864.

Brown, A. R. (1999, January). The sensitivity of large-eddy simulations of shallow cumulus convection to resolution and subgrid model. *Quarterly Journal of the Royal Meteorological Society*, *125*(554), 469–482. (WOS:000079350100004) doi: 10.1002/qj.49712555405

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

Browning, K., Betts, A., Jonas, P., Kershaw, R., Manton, M., Mason, P., ... Simpson, J. (1993, March). The GEWEX Cloud System Study (GCSS). *Bulletin of the American Meteorological Society*, *74*(3), 387–399. (WOS:A1993KU53500004)

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *30*(11), 114007.

Caldwell, P., & Bretherton, C. S. (2009, January). Response of a Subtropical Stratocumulus-Capped Mixed Layer to Climate and Aerosol Changes. *Journal of Climate*, *22*(1), 20–38. (WOS:000262329100002) doi: 10.1175/2008JCLI1967.1

Carpenter, B., & Coauthors. (2017, May). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 00–00. doi: 10.18637/jss.v076.i01

Chinita, M. J., Matheou, G., & Teixeira, J. (2018). A joint probability density based decomposition of turbulence in the atmospheric boundary layer. *Monthly Weather Review*, *146*, 503–523.

Couvreux, F., Guichard, F., Redelsperger, J. L., Kiemle, C., Masson, V., Lafore, J. P., & Flamant, C. (2005, October). Water-vapour variability within a convective boundary-layer assessed by large-eddy simulations and IHOP_2002 observations. *Quarterly Journal of the Royal Meteorological Society*, *131*(611), 2665–2693. (WOS:000233475900005) doi: 10.1256/qj.04.167

Couvreux, F., Hourdin, F., & Rio, C. (2010, March). Resolved Versus Parametrized Boundary-Layer Plumes. Part I: A Parametrization-Oriented Conditional Sampling in Large-Eddy Simulations. *Boundary-Layer Meteorology*, *134* (3), 441–458. (WOS:000274013600004) doi: 10.1007/s10546-009-9456-5

Craig, P. S., Goldstein, M., Seheult, A., & Smith, J. (1996). Bayes linear strategies for matching hydrocarbon reservoir history. *Bayesian statistics*, *5*, 69–95.

Cuxart, J., Bougeault, P., & Redelsperger, J.-L. (2000). A turbulence scheme allowing for mesoscale and large-eddy simulations. *Q. J. R. Meteorol. Soc.*, *126*, 1–30.

Duynkerke, P. G., de Roode, S. R., van Zanten, M. C., Calvo, J., Cuxart, J., & Cheinet, S. (2004). Observations and numerical simulations of the diurnal cycle of the eurocs stratocumulus case. *Quarterly Journal of the Royal Meteorological Society*, *130*, 3269–3296.

Flato, J., G.and Marotzke, Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., ... Rummukaine, M. (2013, August). Evaluation of climate models. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assess-ment Report of the Intergovernmental Panel on Climate Change*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515–534.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018, May). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*, 5742–5751. doi: 10.1029/2018GL078202

Gettelman, A., Truesdale, J., Bacmeister, J., Caldwell, P., Neale, R., & Bogenschutz, P. (2019, May). The single column atmosphere model version 6 (scam6): Not a scam but a tool for model evaluation and development. *Journal of Advances in modeling earth systems*, *11*, 1381–1401. doi: 10.1029/2018MS001578

Golaz, J.-C., Horowitz, L. W., & Levy, H. (2013, May). Cloud tuning in a coupled climate model: Impact on 20th century warming. *Geophysical Research Letters*, *40* (10), 2246–2251. (WOS:000328840200064) doi: 10.1002/grl.50232

Golaz, J. C., Larson, V. E., & Cotton, W. R. (2002, December). A PDF-based model for boundary layer clouds. Part II: Model results. *Journal of the Atmospheric Sciences*, *59* (24), 3552–3571. (WOS:000179629800007) doi: 10.1175/

1520-0469(2002)059⟨3552:APBMFB⟩2.0.CO;2

Grabowski, W. W. (2016, June). Towards global large-eddy simulation: super parameterization revisited. *Journal of the Meteorological Society of Japan*, *94*(4), L20806. doi: 10.2151/jmsj.2016-017

Gueremy, J. F. (2011, August). A continuous buoyancy based convection scheme: one- and three-dimensional validation. *Tellus Series a-Dynamic Meteorology and Oceanography*, *63*(4), 687–706. (WOS:000292864500004) doi: 10.1111/j.1600-0870.2011.00521.x

Guichard, F., & Couvreux, F. (2017). A short review of numerical cloud-resolving models. *Tellus Dyn. Meteorol. Oceanogr.*, *69*, 1945–1960. doi: 10.1080/16000870.2017.1373578

Heus, T., & Jonker, H. J. J. (2008, March). Subsiding shells around shallow cumulus clouds. *Journal of the Atmospheric Sciences*, *65*(3), 1003–1018. (WOS:000254356600016) doi: 10.1175/2007JAS2322.1

Heus, T., Pols, C. F. J., Jonker, H. J. J., Van den Akker, H. E. A., & Lenschow, D. H. (2009, January). Observational validation of the compensating mass flux through the shell around cumulus clouds. *Quarterly Journal of the Royal Meteorological Society*, *135*(638), 101–112. (WOS:000265374900008) doi: 10.1002/qj.358

Holtslag, A. A. M., Svensson, G., Baas, P., Basu, S., Beare, B., Beljaars, A. C. M., ... Van de Wiel, B. J. H. (2013, November). STABLE ATMOSPHERIC BOUNDARY LAYERS AND DIURNAL CYCLES Challenges for Weather and Climate Models. *Bulletin of the American Meteorological Society*, *94*(11), 1691–1706. (WOS:000327926700007) doi: 10.1175/BAMS-D-11-00187.1

Hourdin, F., Couvreux, F., & Menut, L. (2002, March). Parameterization of the dry convective boundary layer based on a mass flux representation of thermals. *Journal of the Atmospheric Sciences*, *59*(6), 1105–1123. (WOS:000174019900006) doi: 10.1175/1520-0469(2002)059⟨1105:POTDCB⟩2.0.CO;2

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. (WOS:000318278700005) doi:

10.1007/s00382-012-1343-y

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., ... Williamson, D. (2017, March). The Art and Science of Climate Model Tuning. *Bull. Am. Meteorol. Soc.*, *98*, 589-602. doi: 10.1175/BAMS-D-15-00135.1

Hourdin, F., Rio, C., Grandpeix, J.-Y., Madeleine, J.-B., Cheruy, F., Rochetin, N., ... Ghattas, J. (2020, June). LMDZ6A: the atmospheric component of the ipsl climate model with improved and better tuned physics. *Journal of Advances in modeling earth systems*, *accepted for publication*, ??–?? doi: 10.1029/2019MS001892

Jakob, C. (2010, July). ACCELERATING PROGRESS IN GLOBAL ATMOSPHERIC MODEL DEVELOPMENT THROUGH IMPROVED PARAMETERIZATIONS Challenges, Opportunities, and Strategies. *Bulletin of the American Meteorological Society*, *91*(7), 869–+. (WOS:000280758700003) doi: 10.1175/2009BAMS2898.1

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. (WOS:000319475000004) doi: 10.1007/s10546-012-9789-3

Jiang, J. H., Su, H., Zhai, C., Perun, V., Del Genio, A., Nazarenko, L. S., ... L, S. G. (2012, July). Evaluation of cloud and water vapor simulations in CMIP5 climate models using nasa“a-train”satellite observations. *Journal of Geophysical Research*, *117*, 1–24. doi: 10.1029/2011JD017237

Kennedy, M. C., & O'Hagan, A. (2001, aug). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(3), 425–464. Retrieved from `http://doi.wiley.com/10.1111/1467-9868.00294` doi: 10.1111/1467-9868.00294

Khairoutdinov, M., Randall, D., & DeMott, C. (2005, July). Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes. *Journal of the Atmospheric Sciences*, *62*(7), 2136–2154. (WOS:000230962800006) doi: 10.1175/JAS3453.1

Klein, S. A., Hall, A., R., N. J., & Robert, P. (2017, October). Low-cloud feedbacks from cloud-controlling factors: a review. *Survey of Geophysics*, *38*(10), 1307–1329. doi: 10.1007/s10712-017-9433-3

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013, March). Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artifical Neural Systems*, *203*(3), 13. doi: 10.1155/2013/485913

Lac, C., Chaboureau, J.-P., Masson, V., Pinty, J.-P., Tulet, P., Escobar, J., ... Wautelet, P. (2018, January). Overview of the Meso-NH model version 5.4 and its applications. *Geosci. Model Dev. Discuss.*, *2018*, 1–66. Retrieved 2018-03-26, from https://www.geosci-model-dev-discuss.net/gmd-2017-297/ doi: 10.5194/gmd-2017-297

Matheou, G., Chung, D., Nuijens, L., Stevens, B., & Teixeira, J. (2011, September). On the Fidelity of Large-Eddy Simulation of Shallow Precipitating Cumulus Convection. *Monthly Weather Review*, *139*(9), 2918–2939. (WOS:000294932100014) doi: 10.1175/2011MWR3599.1

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., ... Tomassini, L. (2012, August). Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, *4*, M00A01. (WOS:000307467200001) doi: 10.1029/2012MS000154

McNeall, D., Williams, J., Betts, R., Booth, B., Challenor, P., Good, P., & Wiltshire, A. (2019). Correcting a bias in a climate model with an augmented emulator. *Geoscientific Model Development Discussions*, *2019*, 1–37. Retrieved from https://www.geosci-model-dev-discuss.net/gmd-2019-171/ doi: 10.5194/gmd-2019-171

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. (2009, June). A Dual Mass Flux Framework for Boundary Layer Convection. Part II: Clouds. *Journal of the Atmospheric Sciences*, *66*(6), 1489–1506. (WOS:000267263300002) doi: 10.1175/2008JAS2636.1

Neggers, R. A. J. (2015, December). Attributing the behavior of low-level clouds in large-scale models to subgrid-scale parameterizations. *Journal of Advances in modeling earth systems*, *7*(4), 2029–2043. doi: 10.1002/2015MS000503

Neggers, R. A. J., Ackerman, A. S., Angevine, W. M., Bazile, E., Beau, I., Blossey,

P. N., ... Heus, T. (2017, October). Single-column model simulations of subtropical marine boundary-layer cloud transitions under weakening inversions. *Journal of Advances in modeling earth systems*, *9*(6), 2385–2412. doi: 10.1002/2017MS001064

Neggers, R. A. J., Duynkerke, P. G., & Rodts, S. M. A. (2003, July). Shallow cumulus convection: A validation of large-eddy simulation against aircraft and Landsat observations. *Quarterly Journal of the Royal Meteorological Society*, *129*(593), 2671–2696. (WOS:000185187600011) doi: 10.1256/qj.02.93

Neggers, R. A. J., J, J. H. J., & Siebesma, P. (2003). Statistics of cumulus cloud populations in large-eddy simulations. *J. Atmos. Sci.*, *60*, 1060–1074.

Neggers, R. A. J., Siebesma, A. P., & Heus, T. (2012, September). Continuous single-column model evaluation at a permanent meteorological supersite. *Bulletin of the American Meteorological Society*, *93*(9), 1389–1400. (WOS:000309056400006) doi: 10.1175/BAMS-D-11-00162.1

Neggers, R. A. J., Siebesma, A. P., Lenderink, G., & Holtslag, A. A. M. (2004, November). An evaluation of mass flux closures for diurnal cycles of shallow cumulus. *Monthly Weather Review*, *132*(11), 2525–2538. (WOS:000225098900001) doi: 10.1175/MWR2776.1

Neggers, R. A. J., Siebesma, P., & J, J. H. J. (2002). A multiparcel model for shallow cumulus convection. *J. Atmos. Sci.*, *59*, 1655–1668.

Nuijens, L., Medeiros, B., Sandu, I., & Ahlgrimm, M. (2015, December). Observed and modeled patterns of covariability between low-level cloudiness and the structure of the trade-wind layer. *Journal of Advances in Modeling Earth Systems*, *7*(4), 1741–1764. (WOS:000368739800013) doi: 10.1002/2015MS000483

Oakley, J. E., & O'Hagan, A. (2004, December). Probabilistic sensitivity analysis of complex models: a bayesian approach. *Royal Statistical Society*, *66*(3), 751–769.

Parishani, H., Pritchard, M., Bretherton, C., Wyant, M., & Khairoutdinov, M. (2017, May). Toward low-cloud-permitting cloud superparameterization with explicit boundary layer turbulence. *Journal of Advances in modeling earth systems*, *9*, 1542–1571. doi: 10.1002/2018MS001409

Pergaud, J., Masson, V., Malardel, S., & Couvreux, F. (2009, July). A Parameterization of Dry Thermals and Shallow Cumuli for Mesoscale Numerical Weather Prediction. *Boundary-Layer Meteorology*, *132*(1), 83–106. (WOS:000267029600006) doi: 10.1007/s10546-009-9388-0

Piriou, J.-M., Redelsperger, J.-L., Geleyn, J.-F., Lafore, J.-P., & Guichard, F. (2007, November). An approach for convective parameterization with memory: Separating microphysics and transport in grid-scale equations. *Journal of the Atmospheric Sciences*, *64*(11), 4127–4139. (WOS:000251283000025) doi: 10.1175/2007JAS2144.1

Pressel, K. G., Mishra, S., Schneider, T., Kaul, C. M., & Tan, Z. (2017, June). Numerics and subgrid-scale modeling in large eddy simulations of stratocumulus clouds. *Journal of Advances in Modeling Earth Systems*, *9*(2), 1342–1365. doi: 10.1002/2016MS000778

Pukelsheim, F. (1994, February). The three sigma rule. *The American Statistician*, *48*, 88–91.

Randall, D., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003, November). Breaking the cloud parameterization deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547–1564. (WOS:000187163900019) doi: 10.1175/BAMS-84-11-1547

Randall, D., Xu, K., Somerville, R., & Iacobellis, S. (1996, August). Single-column models and cloud ensemble models as links between observations and climate models. *Journal of Climate*, *9*(8), 1683–1697. (WOS:A1996VG92100002) doi: 10.1175/1520-0442(1996)009⟨1683:SCMACE⟩2.0.CO;2

Richter, I. (2015, February). Climate model biases in the eastern tropical oceans: Causes, impacts and ways forward. *Wiley Interdisciplinary Reviews:Climate Change*, *6*(3), 345–358.

Rio, C., Del Genio, A. D., & F, H. (2019). Ongoing breakthroughs in convective parameterization. *Current Climate Change Reports*, *5*, 95–111.

Rio, C., & Hourdin, F. (2008, February). A thermal plume model for the convective boundary layer: Representation of cumulus clouds. *Journal of the Atmospheric Sciences*, *65*(2), 407–425. (WOS:000253406600007) doi: 10.1175/2007JAS2256.1

Rio, C., Hourdin, F., Couvreux, F., & Jam, A. (2010, June). Resolved Versus Parametrized Boundary-Layer Plumes. Part II: Continuous Formulations of Mixing Rates for Mass-Flux Schemes. *Boundary-Layer Meteorology*, *135*(3), 469–483. (WOS:000277635800007) doi: 10.1007/s10546-010-9478-z

Rochetin, N., Couvreux, F., Grandpeix, J.-Y., & Rio, C. (2014, February). Deep Convection Triggering by Boundary Layer Thermals. Part I: LES Analysis and Stochastic Triggering Formulation. *Journal of the Atmospheric Sciences*, *71*(2), 496–514. (WOS:000335491400003) doi: 10.1175/JAS-D-12-0336.1

Roehrig, R., Beau, I., Saint-Martin, D., Alias, A., Colin, J., Decharme, B., … Sénési, S. (2020, June). The CNRM global atmosphere model ARPEGE-Climat 6.3: description and evaluation. *Journal of Advances in modeling earth systems*, *accepted for publication*, ??–?? doi: 10.1029/2020MS002075

Saltelli, A. J. (2002, December). Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, *145*(2), 280–297.

Salter, J. M., Williamson, D. B., Scinocca, J., & Kharin, V. (2019, October). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. *Journal of the American statistical association*, *114*(528), 1800–1814. doi: 10.1080/01621459.2018.1514306

Sandu, I., Beljaars, A., Bechtold, P., Mauritsen, T., & Balsamo, G. (2013). Why is it so difficult to represent stably stratified conditions in numerical weather prediction (NWP) models? *Journal of Advances in Modeling Earth Systems*, *5*(2), 117–133. Retrieved from `http://dx.doi.org/10.1002/jame.20013` doi: 10.1002/jame.20013

Satoh, M., Matsuno, T., Tomita, H., Miura, H., Nasuno, T., & Iga, S. (2008, March). Nonhydrostatic icosahedral atmospheric model (NICAM) for global cloud resolving simulations. *Journal of Computational Physics*, *227*(7), 3486–3514. (WOS:000255005900005) doi: 10.1016/j.jcp.2007.02.006

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W., & P., D. (2019, September). Global cloud-resolving models. *Current Climate Change Reports*, *5*(3), 172–184. doi: 10.1007/s40641-019-00131-0

Schneider, T., Lan, T., Stuart, A., & Teixeira, J. (2017, December). Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters*, *44*, 12396–12417. doi: 10.1002/2017GL076101

Sexton, D. M., Murphy, J. M., Collins, M., & Webb, M. J. (2011, October). Multivariate probabilistic projections using imperfect climate models part i: outline of methodology. *Climate Dynamics*, *38*(1), 2513–2542.

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. *J. Atmos. Sci.*, *60*, 1201–1219.

Siebesma, A. P., & Cuijpers, J. W. M. (1995). Evaluation of parametric assumptions for shallow cumulus convection. *J. Atmos. Sci.*, *52*, 650–666.

Siebesma, A. P., Soares, P. M. M., & Teixeira, J. (2007, April). A combined eddy-diffusivity mass-flux approach for the convective boundary layer. *Journal of the Atmospheric Sciences*, *64*(4), 1230–1248. (WOS:000245742600011) doi: 10.1175/JAS3888.1

Soares, P. M. M., Miranda, P. M. A., Siebesma, A. P., & Teixeira, J. (2004). An eddy-diffusivity/mass-flux parameterization for dry and shallow cumulus convection. *Q. J. R. Meteorol. Soc.*, *130*(604), 3365–3383.

Stevens, B., Satoh, M., Auger, L., Bierchamp, J., Bretherton, C. S., Chen, X., … Chou, L. (2019). Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*, 1–17. doi: 10.1186/s40645-019-0304-z

Sullivan, P. P., & Patton, E. G. (2011, October). The Effect of Mesh Resolution on Convective Boundary Layer Statistics and Structures Generated by Large-Eddy Simulation. *Journal of the Atmospheric Sciences*, *68*(10), 2395–2415. (WOS:000296034700014) doi: 10.1175/JAS-D-10-05010.1

Suselj, K., Kurowski, M. J., & Teixeira, J. (2019, August). A unified eddy-diffusivity/mass-flux approach for modeling atmospheric convection. *Journal of the Atmospheric Sciences*, 2505–2537.

Suselj, K., Teixeira, J., & Chung, D. (2013, July). A Unified Model for Moist Convective Boundary Layers Based on a Stochastic Eddy-Diffusivity/Mass-Flux Parameterization. *Journal of the Atmospheric Sciences*, *70*(7), 1929–1953. (WOS:000322125600005) doi: 10.1175/JAS-D-12-0106.1

Tan, Z., Kaul, C. M., Pressel, G., K, Cohen, Y., Schneider, T., & Teixeira, J. (2018, March). An extended eddy-diffusivity mass-flux scheme for unified representation of subgrid scale turbulence and convection. *Journal of Advances in Modeling Earth Systems*, 770–800.

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*, M06001. (WOS:000303198400003) doi: 10.1029/2011MS000056

Vernon, I., Goldstein, M., & Bower, R. (2010). Galaxy formation: a bayesian uncertainty analysis. *Bayesian Analytics*, *5*, 619–846.

Villefranque, N., Fournier, R., Couvreux, F., Blanco, S., Eymet, V., Forest, V., & Tregan, J. M. (2019). A path-tracing monte carlo library for 3-d radiative transfer in highly resolved cloudy atmospheres. *Journal of Adavnces in Modeling Earth Systems*, *11*, 2449–2473.

Voldoire, A., Saint-Martin, D., Senesi, S., Decharme, B., Alias, A., Chevallier, M., … Waldman, W., R (2019, August). Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213. doi: 10.1029/2019MS001683

Volodina, V. (2020). *Uncertainty quantification for complex computer models with nonstationary output. bayesian optimal design for iterative refocussing* (Unpublished doctoral dissertation). University of Exeter.

Volodina, V., & Williamson, D. (2020, January). Diagnostics-driven nonstationary emulators using kernel mixtures. *Journal of Uncertainty Quantification*, *8*(1), 1–26.

Wang, H., & Feingold, G. (2009, November). Modeling Mesoscale Cellular Structures and Drizzle in Marine Stratocumulus. Part I: Impact of Drizzle on the Formation and Evolution of Open Cells. *Journal of the Atmospheric Sciences*, *66*(11), 3237–3256. (WOS:000271689700001) doi: 10.1175/2009JAS3022.1

Williamson, D. (2015, June). Exploratory ensemble designs for environmental models using k-extended Latin Hypercubes. *Environmetrics*, *26*(4), 268–283. (WOS:000353380200003) doi: 10.1002/env.2335

Williamson, D., Blaker, A. T., Hampton, C., & Salter, J. (2015, September). Identifying and removing structural biases in climate models with history matching. *Climate Dynamics*, *45*(5-6), 1299–1324. (WOS:000360507700010) doi: 10.1007/s00382-014-2378-z

Williamson, D., Blaker, A. T., & Sinha, B. (2017, April). Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model.

Geoscientific Model Development, *10*(4), 1789–1816. (WOS:000400181200002) doi: 10.5194/gmd-10-1789-2017

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013, October). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, *41*(7-8), 1703–1729. (WOS:000324812200002) doi: 10.1007/s00382-013-1896-4

Williamson, D., & Volodina, V. (2020). Exeteruq mogp an r interface to performing uq with mop emulator. *Documentation*. Retrieved from `https://bayesexeter.github.io/ExeterUQ_MOGP/`

Wurps, H., Steinfeld, G., & Heinz, S. (2020, March). Grid-Resolution Requirements for Large-Eddy Simulations of the Atmospheric Boundary Layer. *Boundary-Layer Meteorology*, *175*, 179–201. doi: 10.1007/s10546-020-00504-1

Zhang, M., Somerville, R. C. J., & Xie, S. (2016, February). The scm concept and creation of arm forcing datasets. *Meteorological Monographs*, *57*, 24.1–24.12. doi: 10.1175/AMSMONOGRAPHS-D-15-0040.1

Zhang, Y., Klein, S. A., Fan, J., Chandra, A. S., Kollias, P., Xie, S., & Tang, S. (2017, October). Large-Eddy Simulation of Shallow Cumulus over Land: A Composite Case Based on ARM Long-Term Observations at Its Southern Great Plains Site. *Journal of the Atmospheric Sciences*, *74*(10), 3229–3251. doi: 10.1175/JAS-D-16-0317.1