Theor. Comput. Fluid Dyn.
DOI 10.1007/s00162-012-0281-y

ORIGINAL ARTICLE

Jesse Dorrestijn · Daan T. Crommelin · A. Pier Siebesma · Harm J. J. Jonker

Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data

Received: 26 January 2012 / Accepted: 22 August 2012

Abstract In this paper, we report on the development of a methodology for stochastic parameterization of convective transport by shallow cumulus convection in weather and climate models. We construct a parameterization based on Large-Eddy Simulation (LES) data. These simulations resolve the turbulent fluxes of heat and moisture and are based on a typical case of non-precipitating shallow cumulus convection above sea in the trade-wind region. Using clustering, we determine a finite number of turbulent flux pairs for heat and moisture that are representative for the pairs of flux profiles observed in these simulations. In the stochastic parameterization scheme proposed here, the convection scheme jumps randomly between these pre-computed pairs of turbulent flux profiles. The transition probabilities are estimated from the LES data, and they are conditioned on the resolved-scale state in the model column. Hence, the stochastic parameterization is formulated as a data-inferred conditional Markov chain (CMC), where each state of the Markov chain corresponds to a pair of turbulent heat and moisture fluxes. The CMC parameterization is designed to emulate, in a statistical sense, the convective behaviour observed in the LES data. The CMC is tested in single-column model (SCM) experiments. The SCM is able to reproduce the ensemble spread of the temperature and humidity that was observed in the LES data. Furthermore, there is a good similarity between time series of the fractions of the discretized fluxes produced by SCM and observed in LES.

Keywords Stochastic parameterization · Atmospheric convection · Large-Eddy Simulation · Markov chain · Clustering · Grey zone

1 Introduction

The effect of clouds and convection on the large-scale atmospheric state is one of the major sources of uncertainty in weather and climate models. To resolve the convective dynamics realistically, a numerical model resolution of at least 100 m is required. Current operational numerical weather prediction (NWP) models are still far too coarse to resolve convection: global NWP models are approaching $O(10$ km$)$ resolutions while high-resolution limited-area models operate at $O(1$ km$)$ resolution. The atmospheric components of coupled climate models currently use resolutions of $O(100$ km$)$ or more because of the long simulation time spans

Communicated by W. Dewar.

J. Dorrestijn (✉) · D. T. Crommelin
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
E-mail: J.Dorrestijn@cwi.nl

A. P. Siebesma · H. J. J. Jonker
Delft University of Technology, PO Box 5046, 2600 GA Delft, The Netherlands

A. P. Siebesma
Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

Published online: 11 September 2012

for which climate models are used. In all of these models, the effects of clouds and convection in individual vertical model columns must therefore be represented through a parameterization, that is, the effect of these processes have to be taken into account statistically in terms of the resolved mean state of the model column.

As pointed out in the seminal paper of Arakawa and Schubert [1], there are 2 fundamental assumptions underlying all traditional convection parameterizations: (i) the horizontal model grid size is large enough for each model column to contain a representative statistical ensemble of convective clouds, (ii) the cloud ensemble is in quasi-equilibrium with the resolved large-scale variables [13]. These assumptions justify a deterministic convective parameterization: the resolved-scale state determines a unique ensemble of convective clouds that is well sampled and that produces unique convective transport and cloud properties.

With increasing model resolution, the above assumptions become problematic. With decreasing grid size, the size of the ensemble of convective clouds in a model column decreases, so that the ensemble is more likely to deviate significantly from the theoretical distribution (see Plant and Craig [24]), and as a result, it is expected that the cloud ensemble will give a fluctuating response to the same mean state. Furthermore, the life cycles of individual convective events become more prominent, so that quasi-equilibrium is less likely to hold. Clearly, the one-to-one correspondence between the resolved mean state and the convective response breaks down and a traditional deterministic convection parameterization will not be able to incorporate these fluctuations.

A promising strategy to tackle parameterization under conditions, where traditional approaches break down, is the use of stochastic methods [4,14,16–18,21,23,24,30]. Rather than fixing the subgrid-scale response to a given resolved-scale state (as in a deterministic parameterization), the response is randomly sampled from a suitable probability distribution. This allows to account for the randomness of underresolved convection in a small model column.

In this paper, we report on the development of a methodology for stochastic parameterization of atmospheric moist convection. Our approach is based on the stochastic method introduced by Crommelin and Vanden-Eijnden [5] and has several key features. First of all, the stochastic process that represents the convective response of the subgrid scales in a model column is made *conditional* on the resolved-scale state in the same model column. Thus, the statistical properties of the stochastic subgrid-scale response change if the resolved-scale state changes. Secondly, the set of possible subgrid-scale responses is made finite (discrete), by using *finite Markov chains* as a stochastic process. This gives the advantage of an easy and straightforward computation and estimation. Thirdly, the properties of the stochastic process (i.e., the Markov chain) are *estimated from data*, where the data comes from high-resolution Large-Eddy Simulations (LES).

The Large-Eddy Simulations of moist convection are run at resolutions high enough to resolve convection explicitly. The LES data and thus the Markov chains are *precomputed*, that is, they are determined before the stochastic parameterization is put to use. The conditional Markov chain (CMC) parameterization is designed to reproduce, in a statistical sense, the convective behaviour observed in the LES data. Thus, it can be seen as a *statistical emulator* of the high-resolution LES model. Because of its high computational cost, the LES model can only cover the horizontal domain of a few model columns of an operational NWP or climate model. Using a statistical emulator type parameterization, trained on LES data, allows one to use realistic, LES-emulating convection at low computational cost.

Atmospheric moist convection can be distinguished in two categories. One category is *shallow convection* characterized by fair weather cumulus that have a limited vertical extent of no more than 3 km. As a result, precipitation does play a minor role, and for these clouds, its feedback on the dynamics can be neglected. Shallow cumulus convection plays an important role in the determination of the vertical temperature and humidity profiles. Locally, it determines the vertical transport; non-locally, it has strong influence on the planetary-scale circulation, especially over the sub-tropical oceans where it enhances the moisture transport towards the inter-tropical convergence zone (ITCZ), thereby intensifying the Hadley circulation. Despite their limited size, they are the most abundant cloud type in our climate system and their response to global warming forms one of the largest sources of uncertainty in climate modelling. For a comprehensive introduction to shallow convection, see Siebesma [29].

The second category is that of *deep convection* by cumulus towers that reach heights up to 15 km. Deep convection occurs especially in the tropics in the ITCZ where they provide extra kinetic energy to the Hadley circulation through the net latent heat release as a result of the precipitation. The dynamics of these deep convective clouds is, mainly through the interaction between the precipitation and the cloud dynamics, a far more complex phenomenon than shallow convection.

In this paper, we will concentrate on shallow cumulus convection, for several reasons. As already mentioned, its dynamics is conceptually simpler than that of deep cumulus convection, because precipitation feedback can be neglected. Furthermore, due to its smaller spatial extent, Large-Eddy Simulations are able to

resolve the dynamics of shallow convection numerically on domains large enough to contain a representative ensemble of convective clouds. As a result, we can create a numerical data set that can be coarse-grained from resolutions that fully resolve the dynamics, through resolutions that partly resolve dynamics and that will require a stochastic parameterization, all the way to coarse resolutions for which deterministic statistical parameterizations are sufficient. The focus will be on coarse-grained resolutions of a few kilometres, the so-called *grey zone* or *terra incognita*, see [7,11,32,33] at which individual shallow clouds cannot be resolved but on the other hand, at which a statistical approach is also not possible. We will explore how to use the stochastic approach from [5] to parameterize the vertical convective transport of heat and moisture in a realistic way, taking into account the variability of the transport.

Designing a CMC type parameterization for shallow convection poses several challenges that were not encountered in [5] because of the relative simplicity of the test model used there. In [5], the Lorenz 96 (L96) model [19] was used for testing and demonstrating the CMC parameterization approach. In the L96 model, both the resolved-scale state and the subgrid-scale response at each grid point are scalar quantities. For shallow convection, the situation is much more complicated:

1. The resolved-scale state consists of 5 *functions* (vertical profiles) in each model column (wind velocities, temperature and humidity). Conditioning on the resolved-scale state, a key element of the CMC approach, is therefore highly nontrivial.
2. The subgrid-scale variables consist of 2 vertical profiles, the heat and moisture turbulent fluxes. These fluxes are strongly correlated and must be treated as such in the CMC parameterization.

In [5], discretizing the subgrid-scale response was rather easy because, in the L96 model, the response is a single scalar. Here, we are facing the challenge of summarizing the infinite variety of possible heat and moisture fluxes in a handful (finite) number of states; in other words, we have to discretize an infinite-dimensional function space. To achieve this, we use a *clustering* method, where each cluster centroid represents a heat and moisture flux pair (thereby taking care of the observed correlations between the heat and moisture fluxes).

This paper is organized as follows. In Sect. 2, we introduce the variables and equations that are used in weather and climate models. We describe our approach of parameterizing convection by conditional Markov chains. In Sect. 3, we describe the high-resolution data obtained from LES. We divide the LES domain into subdomains of smaller size to obtain highly intermittent turbulent fluxes for which the use of stochastic parameterization is necessary. In Sect. 4, we describe in detail how to construct a CMC, and in Sect. 5, results are given and the CMC is tested in a *single-column model* (SCM) setting. Finally, in Sect. 6, we summarize and discuss our findings and make some suggestions concerning future work.

2 Problem formulation and strategy

The prognostic equations for heat and moisture in large-scale models are most conveniently written in terms of the liquid water potential temperature θ_l and the total water specific humidity q_t which can be written as

$$\theta_\mathrm{l} = \theta - \frac{L}{c_\mathrm{p}\pi} q_\mathrm{l}, \tag{1}$$

$$q_\mathrm{t} = q_\mathrm{v} + q_\mathrm{l} \tag{2}$$

where θ is the potential temperature, L is the latent heat of vaporization, c_p is the specific heat of dry air at constant pressure, q_l is the liquid water content and q_v is the water vapour specific humidity. We also introduced the Exner function π, the ratio of absolute and potential temperature. In the absence of precipitation θ_l and q_t are conserved for moist adiabatic processes and the grid box averaged prognostic equations for climate and numerical weather prediction models can be written, using the Boussinesq approximation, as

$$\frac{\partial \overline{\theta_\mathrm{l}}}{\partial t} = -\frac{\partial \overline{w'\theta_\mathrm{l}'}}{\partial z} - \overline{\mathbf{v}} \cdot \nabla \overline{\theta_\mathrm{l}} - \overline{w}\frac{\partial \overline{\theta_\mathrm{l}}}{\partial z} + \left.\frac{\partial \overline{\theta_\mathrm{l}}}{\partial t}\right|_\mathrm{rad} \tag{3}$$

$$\frac{\partial \overline{q_\mathrm{t}}}{\partial t} = -\frac{\partial \overline{w'q_\mathrm{t}'}}{\partial z} - \overline{\mathbf{v}} \cdot \nabla \overline{q_\mathrm{t}} - \overline{w}\frac{\partial \overline{q_\mathrm{t}}}{\partial z} \tag{4}$$

where $\mathbf{v}$ denotes the horizontal velocity vector, w the vertical velocity and the last term of the heat equation denotes the tendency due to radiation. Overbars denote a spatial average over the grid box, and primes denote deviations from this average. The first term on the right-hand side represents the turbulent flux divergence which needs to be parameterized. The second and the third terms denote horizontal and vertical advection which are resolved by the model. Since the horizontal turbulent flux divergences are much smaller than the vertical turbulent flux divergence at the resolution of large-scale models, they are omitted in (3) and (4). For shallow cumulus convection, the cloud fraction is usually small; therefore, the tendency due to radiation can be simply prescribed by a clear-sky cooling profile.

We can now schematically formulate our parameterization problem for $\phi \in \{\theta_\mathrm{l}, q_\mathrm{t}\}$ as

$$\frac{\partial \overline{\phi}}{\partial t} = \frac{\partial \overline{\phi}}{\partial t}_{\text{ Convection}} + \frac{\partial \overline{\phi}}{\partial t}_{\text{ Forcing}} \tag{5}$$

which states that the overall tendencies of heat and moisture can be broken down in a forcing term given by model-resolved advection and radiative cooling on the one hand and a turbulent flux divergence term as a result of convection that needs parameterization on the other hand. More precisely, we are searching for a parameterization of the turbulent flux in terms of the mean state and the forcing by means of a function f^ϕ such that

$$\overline{w'\phi'}(z) = f^\phi(z; \overline{\theta_\mathrm{l}}, \overline{q_\mathrm{t}}, F_\phi), \quad \phi \in \{\theta_\mathrm{l}, q_\mathrm{t}\}. \tag{6}$$

where F_ϕ is a short-hand notation for the forcing term of ϕ. This is in line with the definition of parameterization of Jakob [12].

Since the 1960s, researchers have proposed various ways to parameterize convective processes in a model column (see e.g. [2] for an overview). Arguably, the most widely used class of convection parameterization schemes at present is that of mass-flux parameterizations. In these schemes, the shapes of the turbulent fluxes are determined by an entraining plume model, a mass-flux closure at cloud base and several parameters depending on the resolved-scale variables. A straightforward way of designing a stochastic parameterization is to "stochasticize" one of the parameters of an existing, deterministic scheme, as in example [24]. The stochastic approach explored in this paper is different: we do not rely on physical concepts such as entraining plumes or mass-flux profiles, but instead, we infer the turbulent fluxes entirely from pre-computed LES data, thereby bypassing all existing ideas about convection parameterization. We compute the (time-dependent) vertical turbulent flux profiles $\overline{w'\theta_\mathrm{l}'}$ and $\overline{w'q_\mathrm{t}'}$ from the LES data and cluster these profiles in N_α different groups. We emphasize that each of the N_α cluster centroids represents a flux profile pair, that is, each centroid is associated with both a heat flux and a moisture flux. They are denoted by $(f_\alpha^{\theta_\mathrm{l}}(z), f_\alpha^{q_\mathrm{t}}(z))$, $\alpha = 1, \ldots, N_\alpha$ (thus, α is the cluster centroid index). Once the clusters and their centroids are determined, the time series of LES flux profiles $(\overline{w'\theta_\mathrm{l}'}, \overline{w'q_\mathrm{t}'})(z, t)$ can be mapped to a time series $\alpha(t)$ for the centroid index.

The key element of our parameterization approach is to infer a Markov chain stochastic process from the LES time series $\alpha(t)$ and to use this Markov chain to emulate the temporal behaviour of the LES turbulent fluxes. As time evolves, the Markov chain makes random transitions between different values of α, in accordance with transition probabilities that are estimated from the LES time series. The Markov chain that generated time series of α is mapped to a time series of turbulent fluxes by using the cluster centroids:

$$\left(\overline{w'\theta_\mathrm{l}'}(z,t), \overline{w'q_\mathrm{t}'}(z,t)\right)^{\mathrm{CMC}} = \left(f_{\alpha(t)}^{\theta_\mathrm{l}}(z), f_{\alpha(t)}^{q_\mathrm{t}}(z)\right). \tag{7}$$

The occurrence of convection depends in part on the resolved-scale state in the atmospheric model column. To account for this, the Markov chain transition probabilities are conditioned on the resolved-scale state. This conditioning is achieved by clustering the vertical profiles of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ into N_μ clusters. The time series of the LES resolved variable profiles can be mapped to a time series $\mu(t)$ for the resolved-scale state cluster index. Then, we let the transition probabilities for α depend on the cluster index μ in which the resolved-scale state is. Thus, the transition probabilities are encoded by N_μ different stochastic matrices, each of size $N_\alpha \times N_\alpha$.

3 Large-Eddy Simulations, turbulent fluxes and the grey zone

To produce high-resolution data, we use the Dutch Atmospheric LES (DALES), a non-hydrostatic atmospheric high-resolution model that is able to resolve clouds and convection, see Heus et al. [9]. The horizontal- and

vertical grid-point distance is on the order of tens of metres, while the horizontal size of the domain with doubly periodic boundaries is on the order of tens of kilometers and the vertical size is on the order of a few kilometres. The time step is on the order of a few seconds. The prognostic variables are u, v, w, θ_{l} and q_{t}. The equations of motions are based on the Navier-Stokes equations which are simplified using the Boussinesq approximation. The model calculates the liquid water content of all grid boxes to compute clouds. DALES has been used for numerous studies on clouds and convection, both shallow convection and deep convection, see [9].

As we focus on shallow cumulus convection, we run DALES based on a non-precipitating shallow cumulus case as observed during the undisturbed phase of the Barbados Oceanographic and Meteorological Experiment (BOMEX) [10]. During this phase, a typical steady state was observed for a period of 5 days where the large-scale drying and heating due to subsidence is balanced by radiative cooling and convective redistribution of the surface latent and sensible heat fluxes. This steady state can be well reproduced by LES and has been extensively described in the literature [26,27]. For the details of the initial profiles and the prescribed large-scale forcings, we strictly follow the case setup such as described in Siebesma [26].

As already discussed in the introduction, stochastic approaches to parameterization are particularly relevant for model resolutions in the grey zone. In this zone, model resolution is too low to resolve convection explicitly, but too high to rely on quasi-equilibrium to hold. Therefore, we consider three different length scales in the context of our LES model. The first is the horizontal size $L \times L$ of the entire LES domain, where we have chosen $L = 25.6$ km, see Table 1. For model resolutions of size L (or larger), deterministic parameterizations based on traditional equilibrium assumptions can be sufficiently adequate for shallow convection. The second length scale is Δx, the model resolution of the LES model itself. Convection is almost fully resolved at this resolution (which we put at $\Delta x = 50$ m). Finally, the grey zone length scale(s) lies in between L and Δx. To focus on this intermediate range, we divide the LES domain horizontally into subdomains, and we investigate the turbulent fluxes on these subdomains. This coarse-graining technique is similar to the one introduced by Shutts and Palmer [25].

We divide the whole LES domain of size $L \times L$ horizontally into K square subdomains of size $l \times l$, such that we can consider them as model columns of an atmospheric model with a resolution in or near the grey zone (Fig. 1). Each subdomain contains J grid-point values at every vertical level, which is determined by the spatial resolution of the LES. The values J and K and the length scales Δx, l and L are related as follows:

Table 1 A description of the LES data set

Domain size	*# grid points*	*Initialization time (hh:mm:ss)*
$25.6 \times 25.6 \times 3.2$ km^3	$512 \times 512 \times 80$, $J = 1{,}024$	04:00:00
Grid size	*Field experiment*	*# sampling time instances*
$50 \times 50 \times 40$ m^3	BOMEX	$N = 240$
Spatial averaging size	*Length scales*	*LES and sampling time step*
1.6×1.6 km^2, $K = 256$	$L = 25.6$ km, $l = 1.6$ km, $\Delta x = 50$ m	$\Delta t_{LES} \approx 6$ s and $\Delta t = 60$ s

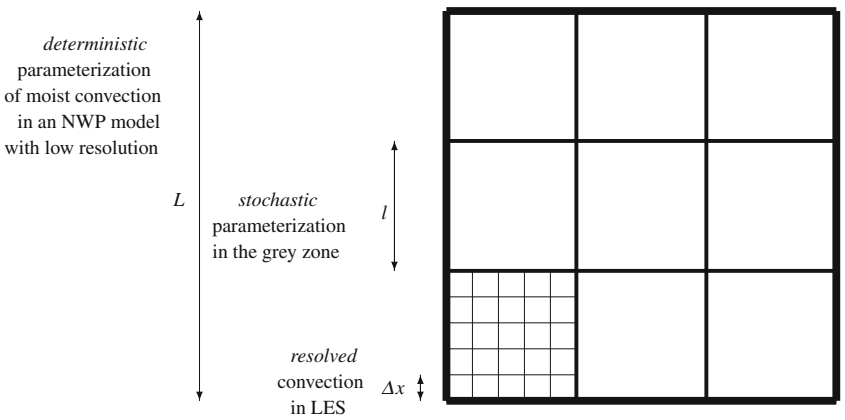

Fig. 1 A depiction of the three length scales discussed in Sect. 3. At the length scale L of the entire LES domain, *deterministic* parameterizations relying on equilibrium assumptions can still be adequate. At the length scale Δx of the LES model resolution, convection is explicitly *resolved*. In the *grey zone*, with model resolutions of size l, in between L and Δx, *stochastic* parameterizations are needed

$$J = \left(\frac{l}{\Delta x}\right)^2, \quad K = \left(\frac{L}{l}\right)^2. \tag{8}$$

We choose $l = 1.6$ km, so we have $K = 256$ subdomains that each contain $J = 1{,}024$ LES gridpoints.

The turbulent fluxes calculated over the subdomains do not simply add up to the turbulent flux calculated over the entire LES domain because the fluxes are determined using deviations from different averages. To clarify this, we define the following averages over the kth subdomain and over the entire domain:

$$\overline{\phi}^{l_k} := J^{-1} \sum_j \phi_{j,k}, \tag{9}$$

$$\overline{\phi}^{L} := (JK)^{-1} \sum_{j,k} \phi_{j,k} = K^{-1} \sum_k \overline{\phi}^{l_k}, \tag{10}$$

where $\phi \in \{w, \theta_l, q_t\}$. For the kth subdomain, one can calculate the turbulent flux relative to the subdomain average $\overline{\phi}^{l_k}$, or relative to the entire domain average $\overline{\phi}^{L}$. The first case gives

$$\overline{w'\phi'}^{l_k} = J^{-1} \sum_j \left(w_{j,k} - \overline{w}^{l_k}\right)\left(\phi_{j,k} - \overline{\phi}^{l_k}\right), \quad \phi \in \{\theta_l, q_t\}, \tag{11}$$

and is related to the second as follows:

$$J^{-1} \sum_j \left(w_{j,k} - \overline{w}^{L}\right)\left(\phi_{j,k} - \overline{\phi}^{L}\right) = \overline{w'\phi'}^{l_k} + \left(\overline{w}^{l_k} - \overline{w}^{L}\right)\left(\overline{\phi}^{l_k} - \overline{\phi}^{L}\right), \quad \phi \in \{\theta_l, q_t\}. \tag{12}$$

For the turbulent flux over the whole domain, we have

$$\overline{w'\phi'}^{L} = K^{-1} \sum_k \overline{w'\phi'}^{l_k} + K^{-1} \sum_k \left(\overline{w}^{l_k} - \overline{w}^{L}\right)\left(\overline{\phi}^{l_k} - \overline{\phi}^{L}\right), \quad \phi \in \{\theta_l, q_t\}. \tag{13}$$

As is clear, it is not equal to the sum of the subdomain fluxes obtained with (11). There is an additional term (the second term on the right-hand side), which is the contribution of the fluxes that are resolved at scale l but not at scale L. In the grey zone, the two contributions are of the same order, by definition of the grey zone. Remark that in this paper, we will calculate the turbulent fluxes on the subdomains with Eq. (11) and not with Eq. (12).

With Eq. (13), we can decompose for every length scale $\Delta x \leq l \leq L$, the turbulent flux on the whole LES domain of size L in a resolved part and an unresolved part. This decomposition is shown in Fig. 2. For this figure, we used two LES datasets for the BOMEX case: our standard dataset with $\Delta x = 50$ m resolution and $L = 25.6$ km domain length, and an additional dataset with $\Delta x = 12.5$ m and $L = 6.4$ km. Including the second dataset enables us to cover a wider range of length scales in Fig. 2 (without the large computational cost of simulating a 25.6×25.6 km^2 domain at 12.5 m resolution). The grey zone is clearly visible (see also Honnert [11]). The standard deviation of the unresolved flux gives an indication of the difficulty of constructing a parameterization for it. In the grey zone, this standard deviation is clearly large. Furthermore, we observe that for larger length scales, the standard deviation decreases as the subdomain size increases; however, it is still substantial until a horizontal domain size of around 10×10 km^2. This indicates that stochastic parameterizations are appropriate not only in the grey zone, but also for larger length scales up to about 10 km. Using the same argument, we could derive that also for length scales equal to or smaller than 50 m, stochastic parameterizations are appropriate; however, because for these length scales convection is almost resolved, the unresolved fluxes are small compared to the resolved fluxes, and therefore, the argument is not valid.

In Fig. 3, we display time series of the turbulent response to the prescribed large-scale cooling in the middle of the cloud layer ($z = 1{,}000$ m) for one of the subdomains of horizontal size 1.6×1.6 km^2 and for the whole domain of horizontal size 25.6×25.6 km^2. In the left panel, we plot the heating/cooling in Kelvin per day: on the whole domain, the turbulent heating is in equilibrium with the large-scale cooling, while in the subdomain, we see large fluctuations. In the right panel, we plot the corresponding heat fluxes for the whole domain and for the subdomain at the same height. It is not difficult to imagine that it is much easier to construct a parameterization for the flux on the whole domain than for the highly intermittent flux on the subdomain. Deterministic parameterizations can be used to calculate the flux in a model column if the resolution is low enough, see [28]. However, if we desire a parameterization that can produce fluxes such that besides the correct mean value of the flux, also the variability (in time) is captured for models with a resolution in the grey zone, we need a new kind of parameterization scheme. Below, we explore the characteristics of

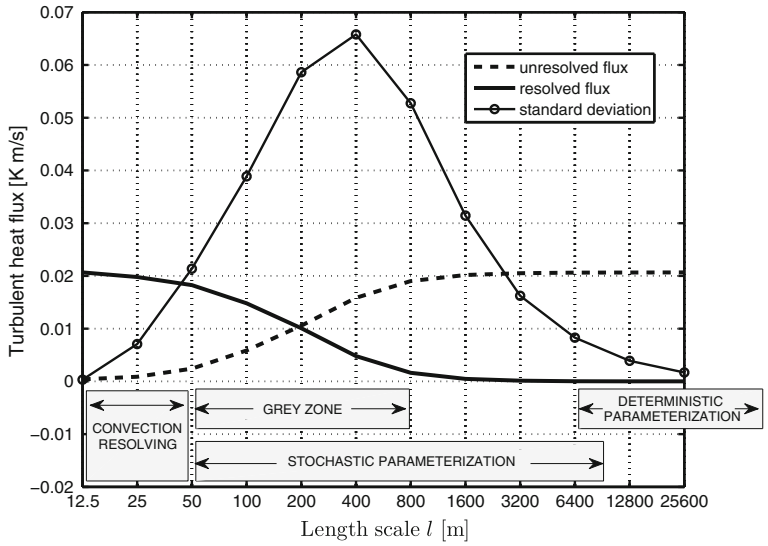

Fig. 2 A decomposition of the turbulent flux $-\overline{w'\theta_l'}^L$ of the whole LES domain of horizontal size 25.6×25.6 km^2 in a resolved part and an unresolved part according to Eq. (13) as a function of subdomain length l (at height 1,000 m). In the *grey zone* these parts are of the same order. The standard deviation of the unresolved fluxes is shown as a function of the subdomain length. The standard deviation is non-negligible up to $l = 10$ km. This indicates that for length scales larger than 10 km the column contains enough shallow convective clouds to use a deterministic parameterization for the unresolved turbulent fluxes. For smaller length scales, stochastic parameterizations are more appropriate

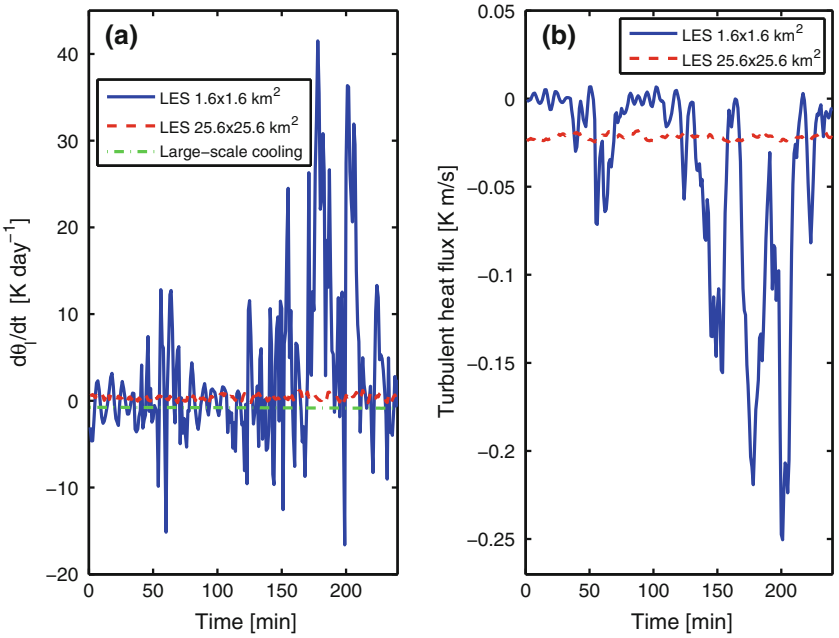

Fig. 3 **a** At height 1,000 m the turbulent heating in the whole LES domain of horizontal size 25.6×25.6 km^2 (*dashed line*), i.e., $-\partial\overline{w'\theta_l'}^L/\partial z$, is in quasi-equilibrium with the large-scale cooling (*dash-dotted line*), while this is not the case for the turbulent heating in a subdomain of horizontal size $1.6 \times 1.6\ km^2$ (*solid line*), i.e., $-\partial\overline{w'\theta_l'}^{l_k}/\partial z$. **b** The fluctuations of the corresponding turbulent heat flux, $\overline{w'\theta_l'}^{l_k}$, in the subdomain (*solid line*) are much larger than the turbulent heat flux, $\overline{w'\theta_l'}^L$, in the whole domain (*dashed line*)

a new stochastic method based on conditional Markov chains. From now on, we will focus on turbulent flux profiles and resolved-scale variable profiles on the subdomains of size $1.6 \times 1.6 \times 3.2$ km^3. The resolution of LES will be $\Delta x = 50$ m. Further, we will omit the l_k upperscript in the $\overline{w'\phi'}^{l_k}$.

4 Construction of the CMC

To construct a CMC, we perform three calculations:

1. Cluster the pairs of turbulent heat and moisture flux profiles to obtain N_α different *flux centroids* (i.e., pairs of representative heat and moisture flux profiles) that determine the *flux states*, indexed by $\alpha \in \{1, \ldots, N_\alpha\}$;
2. Cluster the vertical profiles of the resolved-scale variables to form the *resolved-scale states*, indexed by $\mu \in \{1, \ldots, N_\mu\}$;
3. Count transitions between different flux states to obtain a transition probability matrix for every μ.

Below, we describe these steps in more detail.

4.1 Clustering the turbulent flux profiles

We need to find a finite number of functions that can represent the variability of the turbulent heat and moisture flux profiles observed in LES. We use *clustering* of the observed profiles to obtain such functions [6]. To take into account correlations between the heat and moisture fluxes, both fluxes are clustered *simultaneously*. The resulting cluster *centroids* are the representative pairs of heat and moisture flux profiles that we seek.

For clustering, one needs to choose a clustering method and one has to define a distance function that has to be minimized. We use the *k-means++* algorithm, a partitional center-based clustering method introduced by Arthur [3]. Apart from the initialization, the algorithm of k-means++ is the same as the k-means algorithm first described by Macqueen [20]. The k-means++ algorithm is summarized in the Appendix. It minimizes the cost function defined as the sum over all distances d between data points and their closest centroids. In the present context, a data point of the algorithm is an equal-time pair of heat and moisture flux vertical profiles as observed in the LES data set. The number of clusters N_α has to be chosen a priori. In Sect. 6, we will briefly discuss how to make this choice.

The method is computationally inexpensive; it conserves the mean of the data; and it produces smooth (pairs of) functions as centroids. We observe convergence to a local minimum after a finite number $O(20)$ of iterations. This local minimum does not have to be a global minimum because the optimization problem is non-convex. For the present study, this is not a problem, as long as the centroids can represent the variability of observed LES fluxes. A drawback of k-means++ is that the standard deviation of the clustered data is smaller than the standard deviation of the original data. In Sect. 5, we will say more about this.

As *distance* function d, we choose the following Euclidean distance between two pairs of vertical profiles $g = \big(g_1(z), g_2(z)\big)$ and $h = \big(h_1(z), h_2(z)\big)$:

$$d(g,h) = \sqrt{\sum_z c_1\big(g_1(z) - h_1(z)\big)^2 + c_2\big(g_2(z) - h_2(z)\big)^2}. \tag{14}$$

The summation over z is the summation over all 80 vertical levels. The weight factors c_i are included to non-dimensionalize the contributions from the two different fluxes (heat and moisture). We choose them to be $c_i = \langle\sqrt{\sum_z \big(g_i(z) - h_i(z)\big)^2}\rangle$, $i \in \{1, 2\}$, that is, the average distance between the vertical profiles and their closest centroids. Remark that these averages may change every iteration step in the cluster algorithm.

In Fig. 4, we display the centroids calculated using the k-means++ cluster algorithm with $N_\alpha = 10$. The shaded areas show, for every height, percentile intervals of the observed LES flux profiles, giving an indication of the distribution of the LES fluxes. The centroids cover the range (variability) of the LES flux profiles quite well. The percentile intervals show that the turbulent fluxes are mostly close to 0, with infrequent, large fluctuations. Remark that the surface fluxes for the BOMEX case are fixed at 8.0×10^{-3} Km/s for the heat flux and 5.2×10^{-5} m/s for the moisture flux. We have numbered the centroids such that $\alpha = 1$ corresponds to a clear atmosphere, a higher centroid and flux-state number corresponds to a more convectively active atmosphere and $\alpha = 10$ corresponds to the most convectively active atmosphere.

Jumping briefly forward to Fig. 9a, one can see for every α, the time series of the observed fraction of LES subdomains that are in this flux state. We see that 60–70 % of the subdomains are in flux-state number 1, around 20 % in flux-state number 2, and lower percentages for higher flux-state numbers. We will discuss this in Sect. 5.3.

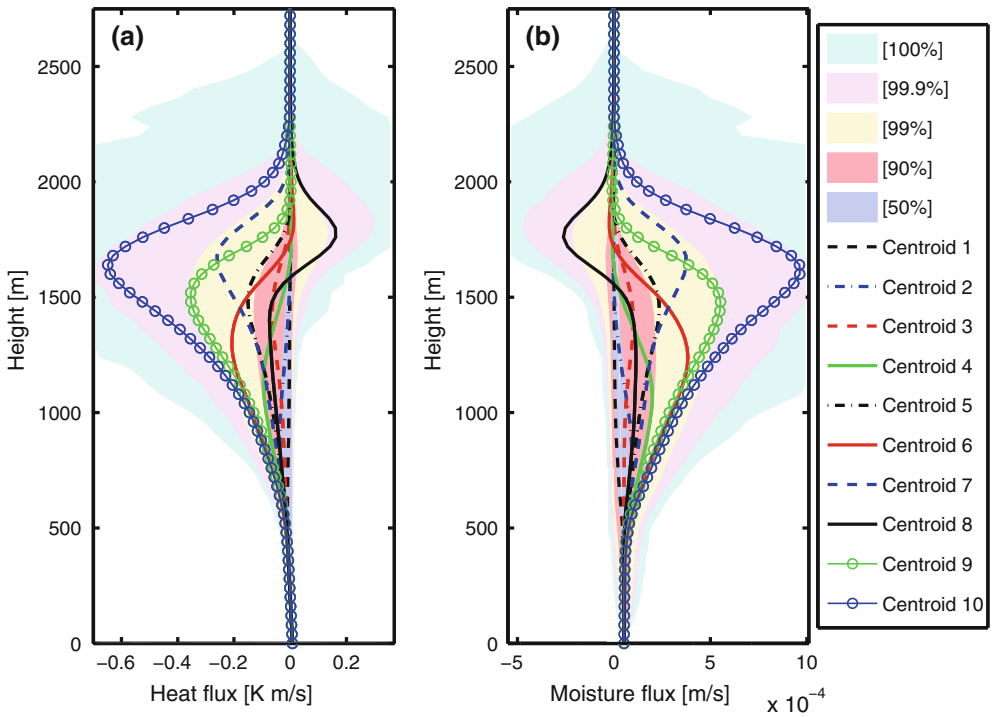

Fig. 4 The 10 centroids (i.e., pairs of turbulent **a** heat and **b** moisture flux profiles) calculated using the k-means++ clustering algorithm with $N_\alpha = 10$. The *shading* indicates for every height the percentage of **a** heat and **b** moisture flux profiles passing through that interval. The centroids cover the range of possible heat and moisture flux profiles that are produced in LES on $1.6 \times 1.6 \times 3.2$ km^3 subdomains

4.2 Conditioning on the resolved-scale state

We employ the same clustering method (k-means++) and the same distance function (14) to construct N_μ different clusters of the resolved-scale variables. The resolved-scale variables we choose to condition on are the entire vertical profiles of $\overline{\theta}$ and $\overline{q_t}$, and to retain correlation, we cluster *pairs* of heat and moisture profiles. Other choices are possible: one can choose any combination of the resolved-scale variables $\overline{u}$, $\overline{v}$, $\overline{w}$, $\overline{\theta_l}$ and $\overline{q_t}$, at any number of vertical levels. We found that conditioning the Markov chain on the combination of the entire vertical profiles of $\overline{\theta_l}$ and $\overline{q_t}$ gives the best results. In Sect. 5.1, we discuss how to choose the number of clusters N_μ.

The whole idea behind conditioning the Markov chain on the resolved-scale state is that the probability of switching between flux states depends on the resolved-scale state. For example, a small difference in temperature can influence the probability that a thermal becomes a cloud or not. Rather than choosing these probabilities ad hoc, we estimate them systematically from the LES data. In the next section, we describe this in more detail.

4.3 Estimation of the transition probability matrices

Once the clustering of the turbulent fluxes (Sect. 4.1) and the resolved-scale states (Sect. 4.2) is completed, the LES data can be mapped to time series $(\alpha_k^{\mathrm{LES}}(t), \mu_k^{\mathrm{LES}}(t))$ for the cluster indices. Thus, $\alpha_k^{\mathrm{LES}}(t) = m$ means that the LES fluxes in the kth subdomain at time t belong to cluster m, and similarly for the resolved-scale state index $\mu_k^{\mathrm{LES}}(t)$. From these time series, we can estimate the transition probabilities for α, conditioned on μ. This is done in a straightforward way, by counting transitions and normalizing in an appropriate way afterwards.

More specifically, we need to estimate the probabilities

$$\mathbf{P}_{nm}^{(i)} = \mathrm{Prob}\left[\alpha_k^{\mathrm{LES}}(t+\Delta t) = m \,|\, \alpha_k^{\mathrm{LES}}(t) = n, \mu_k^{\mathrm{LES}}(t) = i\right] \tag{15}$$

We do so using the following estimator:

$$\hat{P}^{(i)}_{nm} = \frac{T^{(i)}_{nm}}{\sum_m T^{(i)}_{nm}}, \tag{16}$$

where

$$T^{(i)}_{nm} = \sum_k \sum_{\mathrm{t}} \mathbf{1}\left[\alpha^{\mathrm{LES}}_k(t+\Delta t) = m\right] \mathbf{1}\left[\alpha^{\mathrm{LES}}_k(t) = n\right] \mathbf{1}\left[\mu^{\mathrm{LES}}_k(t) = i\right]. \tag{17}$$

The time t runs over the time points t_1 to t_{N-1}, and k runs from 1 to K so that all subdomains contribute to the estimation of the probabilities. The function $\mathbf{1}[.]$ is the indicator function, satisfying $\mathbf{1}[\alpha = m] = 1$ if $\alpha = m$ and $\mathbf{1}[\alpha = m] = 0$ if $\alpha \neq m$. Thus, $T^{(i)}_{nm}$ counts the number of transitions from (n, i) to $(m, \cdot)$.

In total, we obtain N_μ matrices $\hat{P}^{(i)}$ of size $N_\alpha \times N_\alpha$, one stochastic matrix for every μ. This set of matrices can be used to emulate the time evolution of the turbulent fluxes of the LES model. Comparing with the CMC described in [5], we have omitted the conditioning on μ at the next time point $t + \Delta t$. In this way, we reduce the number of used matrices without huge loss of accuracy. See also [22].

4.4 Numerical integration with the CMC parameterization

Using the CMC for parameterization during the numerical time integration of an atmosphere model proceeds as follows. Let $(\overline{u}, \overline{v}, \overline{w}, \overline{\theta_{\mathrm{l}}}, \overline{q_{\mathrm{t}}})_k(z, t)$ be the resolved-scale state in model column k at time t, and let $\alpha^{\mathrm{CMC}}_k(t)$ be the flux cluster index for the same model column at time t.

1. Determine to which cluster μ_k the resolved-scale state in column k belongs.
2. Update the resolved-scale state by integrating it, using (3) and (4), from t to $t + \Delta t$. During this step, the turbulent fluxes in column k are fixed at $(\overline{w'\theta'_{\mathrm{l}}}(z), \overline{w'q'_{\mathrm{t}}}(z)) = (f^{\theta_{\mathrm{l}}}_n(z), f^{q_{\mathrm{t}}}_n(z))$ with $n = \alpha^{\mathrm{CMC}}_k(t)$.
3. Update the fluxes in column k using the stochastic matrix $\hat{P}^{(i)}$ with $i = \mu_k$, i.e. sample m randomly from the probability distribution $\hat{P}^{(i)}_{nm}$ for m, with $n = \alpha^{\mathrm{CMC}}_k(t)$. Now, $\alpha^{\mathrm{CMC}}_k(t + \Delta t) = m$. Repeat this step for all k, using independent sampling for different k.

In the first step, the resolved-scale state centroids and the distance function d (14) are needed. For step 2, the flux-state centroids $(f^{\theta_{\mathrm{l}}}_n(z), f^{q_{\mathrm{t}}}_n(z))$ are required. The stochastic matrices $\hat{P}^{(i)}$ are used in the 3rd step.

5 Results

We construct and test the CMC parameterization using the LES data shown in Table 1. To construct the CMC, we perform the three calculations mentioned at the start of Sect. 4: we determine $N_\alpha = 10$ turbulent flux centroids (Fig. 4), consider $N_\mu = 10$ resolved-scale states determined by the vertical profiles of $\overline{\theta_{\mathrm{l}}}$ and $\overline{q_{\mathrm{t}}}$, and compute the 10 transition probability matrices $\hat{P}^{(i)}$. We test the CMC in three different experiments. In the first experiment, we let the CMC produce the turbulent fluxes while using the LES time series $\mu^{\mathrm{LES}}_k(t)$ as input. Thus, the CMC-produced flux profiles do not feed back onto the resolved-scale state. In the second experiment, this feedback is present, by performing integrations in a single-column model (SCM) setting. The third experiment is similar to the second experiment: only the initial profiles are chosen in a different way.

5.1 Experiment 1: statistics of the CMC

In this experiment, we use the resolved-scale state time series $\mu^{\mathrm{LES}}_k(t)$ obtained from the LES data to "drive" the CMC. The result is the CMC-generated time series $\alpha^{\mathrm{CMC}}_k(t)$ with $k = 1, \ldots, K = 256$ and $t = t_1, \ldots, t_N$, $N = 240$. These can be compared to the LES time series $\alpha^{\mathrm{LES}}_k(t)$. For an example of a flux-state sequence produced by LES and by CMC, see Fig. 5.

The CMC sequences $\alpha^{\mathrm{CMC}}_k(t)$ can be mapped to sequences for the turbulent fluxes by using the flux centroids $(f^{\theta_{\mathrm{l}}}_\alpha(z), f^{q_{\mathrm{t}}}_\alpha(z))$. In Fig. 6, we display the mean and the standard deviation of the vertical profiles of the

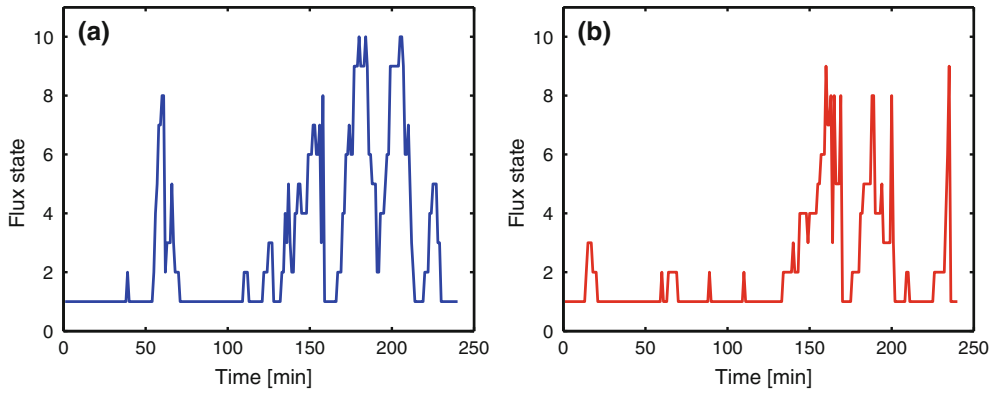

Fig. 5 **a** Discretized turbulent fluxes (states) observed in one LES subdomain of horizontal size $1.6 \times 1.6\,\mathrm{km}^2$ with $N_\alpha = 10$. This discretization is part of the CMC construction algorithm, see Sect. 4.1. **b** Turbulent flux states produced by CMC (in Experiment 1) using the observed resolved-scale states of the same LES subdomain

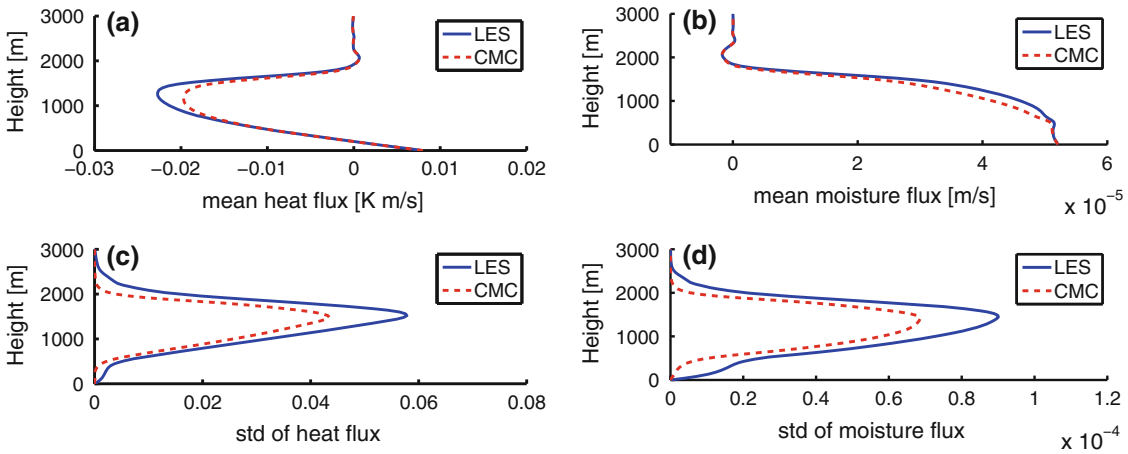

Fig. 6 Mean vertical profile of the turbulent **a** heat and **b** moisture fluxes observed in LES subdomains of $1.6 \times 1.6 \times 3.2\,\mathrm{km}^3$ (*solid*) and produced by CMC (*dashed*) in the first experiment. **c**, **d** The corresponding standard deviations

heat and moisture fluxes observed in the LES data and produced by CMC. There is a small discrepancy for both the mean value and the standard deviation of the heat and moisture flux. The reason for the discrepancy in the mean is that the turbulent flux states with a low probability are less frequently visited in the CMC sequence than in the LES sequence. The reason for this is not entirely clear and may be a subtle effect of the switching between different transition matrices in the CMC. The decrease in standard deviation is easier to understand: by replacing data with their corresponding cluster centroids, it can be proven using the Cauchy-Schwarz inequality that the standard deviation decreases. This problem could be solved by using a moment-preserving clustering method, see [31]. We will not pursue this here.

The choice of the *number of flux centroids* N_α and the *number of resolved-scale state clusters* N_μ influences the performance of the CMC. The smaller N_α the more reduction of the standard deviation of the fluxes. The larger N_α the larger the $N_\alpha \times N_\alpha$ transition matrices of the Markov chain, requiring more data for their estimation. The number N_μ is equal to the number of matrices one has to estimate, so the higher N_μ the more matrices one has to estimate. $N_\mu = 1$ produces the most accurate mean fluxes and standard deviations; however, for $N_\mu = 1$, the Markov chain is not conditioned on the resolved-scale state, giving poor results in the SCM test (Sect. 5.2). Better results in the SCM test are obtained with $N_\mu > 4$. We find the values $N_\alpha = 10$ and $N_\mu = 10$ to be a reasonable compromise between these different considerations.

With this test using resolved-scale states that we observed in LES, we showed that the CMC is able to produce flux profiles with approximately the right statistics. However, in an NWP or climate model, the turbulent fluxes *interact* with the resolved-scale state as in Eqs. (3) and (4) which is not the case in this test. Therefore,

to make a step forward towards this interactive model, we will test the CMC by implementing it in an SCM setting.

5.2 Experiment 2: implementation of CMC in an SCM setting

We test the CMC, described in the first paragraph of Sect. 5, in an SCM setting. An SCM is a 1-dimensional model in which the tendencies of the prognostic variables are only calculated for one column, considered as a column of an NWP or climate model. We will calculate the tendencies of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ using the CMC to generate turbulent fluxes. The governing equations for $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ are analogous to Eq. (3) and (4)

$$\frac{\partial \overline{\theta_\mathrm{l}}}{\partial t} = -\frac{\partial \overline{w'\theta_\mathrm{l}'}}{\partial z} - w_\mathrm{LSS}\frac{\partial \overline{\theta_\mathrm{l}}}{\partial z} + \frac{\partial \overline{\theta_\mathrm{l}}}{\partial t}_{\mathrm{rad}}, \tag{18}$$

and

$$\frac{\partial \overline{q_\mathrm{t}}}{\partial t} = -\frac{\partial \overline{w'q_\mathrm{t}'}}{\partial z} - F_\mathrm{LSHA} - w_\mathrm{LSS}\frac{\partial \overline{q_\mathrm{t}}}{\partial z}, \tag{19}$$

in which the large-scale subsidence, $\overline{w} = w_\mathrm{LSS}$, is a negative vertical wind velocity over the whole domain that was determined for BOMEX. The large-scale forcing for $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ is radiative cooling ($\frac{\partial \overline{\theta_\mathrm{l}}}{\partial t}_{\mathrm{rad}}$) and large-scale horizontal advection (F_LSHA), respectively.

We set the initial profiles of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ equal to the average profiles observed in the $K = 256$ LES subdomains at time t_1. The CMC does not provide $\overline{w'\phi'}(t_1)$ because to determine the turbulent flux profiles, it uses the turbulent flux profiles at the time instance before. Therefore, we choose one of the $N_\alpha = 10$ flux profiles at random with a probability given by the invariant distribution of the fluxes for the given resolved-scale state. For other time instances, the CMC can produce flux profiles $\overline{w'\phi'}$, which are used to determine the time evolution with Eq. (18) and (19).

We calculate the time evolution of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ for 256 runs of the SCM. We compare these time evolutions to the original time evolution of the LES variables: first, by looking at the entire vertical profiles observed in LES at time t_{240} and produced by the SCM (with implemented CMC) after 4 h of integration and then by calculating probability density functions (PDFs) of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ at several heights. In Fig. 7, we see the vertical profiles of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ of 256 LES subdomains and 256 independent SCM realizations after 4 h of integration.

At heights 800, 1,000, 1,400 and 1,600 m, we take a closer look by plotting the PDFs of the 256 values of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ of LES and SCM in Fig. 8. Here, we also plot the results of an SCM experiment in which we use an *unconditioned* Markov chain (MC), that is, $N_\mu = 1$: we clearly see that the *conditional* Markov chain performs better. At t_1, the profiles of the SCM are equal for all the 256 realizations, because we chose them to

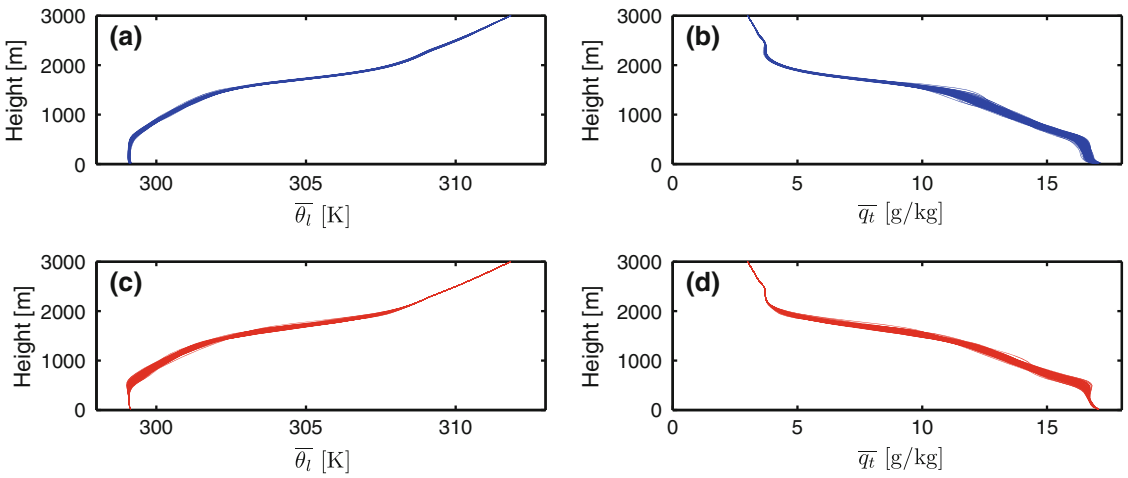

Fig. 7 Superimposed vertical profiles of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ of **a, b** 256 LES subdomains and **c, d** 256 independent SCM-CMC realizations after 4 h of integration in the second experiment

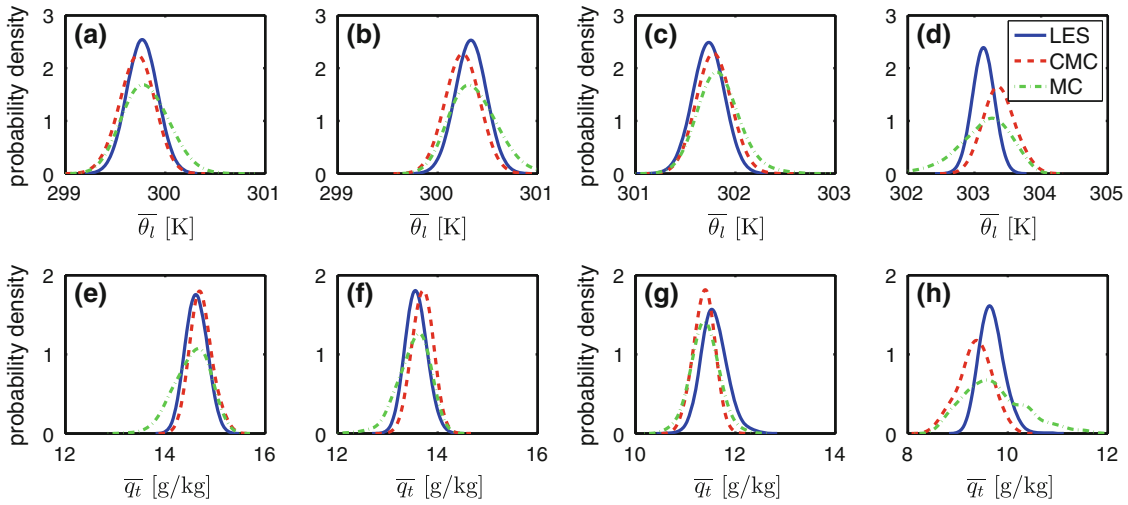

Fig. 8 PDFs of $\overline{\theta_l}$ and $\overline{q_t}$ at heights **a, e** 800, **b, f** 1,000, **c, g** 1,400 and **d, h** 1,600 m of 256 LES subdomains (*solid line*) and 256 independent SCM realizations after 4 h of integration using CMC (*dashed line*) and MC (*dash-dotted line*) in the second experiment

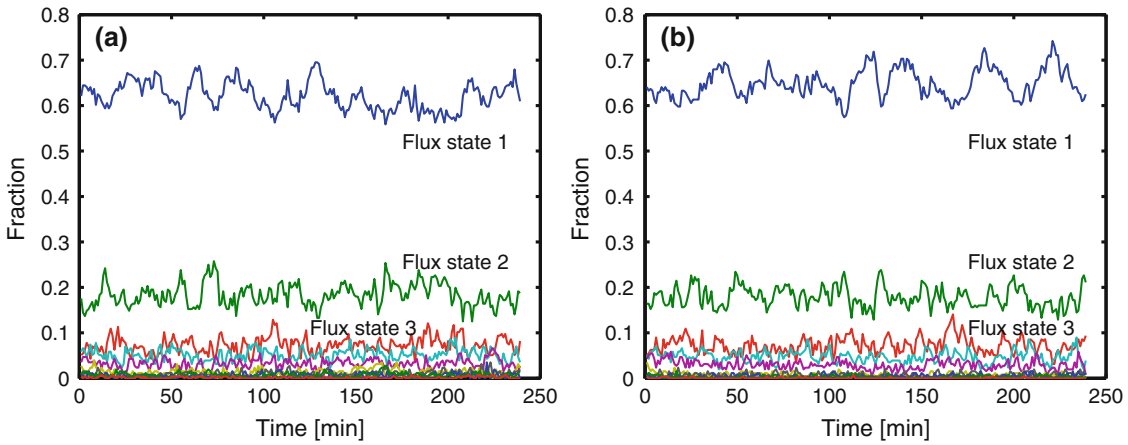

Fig. 9 Time series of the fractions of the 10 flux states **a** observed in the 256 LES subdomains and **b** produced in the 256 SCM-CMC realizations in the third experiment

be equal. Therefore, at t_1, the density function is a Dirac delta function. After 4 h of integration, *the ensemble spread for* $\overline{\theta_l}$ *and* $\overline{q_t}$ *resembles the spread of the profiles produced by LES.* If we continue integrating, the standard deviation of the SCM ensemble keeps growing. This is caused by ensemble members with $\overline{\theta_l}$ and $\overline{q_t}$ profiles too far outside the LES training dataset, so that the CMC parameterization is not trained to drive these ensemble members back to equilibrium. There is a way to solve this problem, by enlarging the training dataset to include such out-of-equilibrium profiles (e.g. by imposing these profiles in LES through nudging). For the unconditioned Markov chain (MC) parameterization, such an enlarged training set is unlikely to solve the problem, because the MC is not sensitive to the $\overline{\theta_l}$ and $\overline{q_t}$ profiles.

5.3 Experiment 3: implementation of CMC in an SCM setting with different initial conditions

We perform another experiment with the SCM. Now, we run the SCM-CMC again 256 times, but with initial profiles of $\overline{\theta_l}$ and $\overline{q_t}$ of the kth run set equal to the profiles of the kth subdomain observed in the LES data at time t_1. For both LES and the SCM, we count the fraction of realizations that are in flux state 1 to 10 as a function of time and plot the time series in Fig. 9. The figure is inspired by a similar figure in Khouider et al. [15]. We see a good similarity between the fractions produced by the SCM and observed in the LES. The equilibrium value of the fractions and the random fluctuations around it are well reproduced by the SCM. Remark that it

takes a few hours of calculation on a supercomputer to produce the LES time series, while the time series of the SCM with the implemented CMC can be calculated on a laptop within 1 min. What is not well visible in Fig. 9 is that the fractions of the least probable flux states (e.g. $\alpha = 10$) are not very well reproduced by the SCM. In the SCM-CMC simulation, these fractions are too low compared to the fractions observed in LES, as was already mentioned in Sect. 5.1.

As a final remark, we recall that we use the entire vertical profiles of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ to condition the Markov chain on. When conditioning on the values of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ at only a few vertical levels, then after 4 h of integrating SMC-CMC, the profiles of $\overline{\theta_\mathrm{l}}$ and $\overline{q_\mathrm{t}}$ were correct at these levels but (highly) inaccurate at other levels (results not shown).

6 Discussion and outlook

In this study, we considered the parameterization of shallow cumulus convection by data-inferred stochastic processes. The vertical turbulent fluxes of heat and moisture in an atmospheric model column were modelled with a stochastic process that is conditioned on the resolved-scale state in the same column. We adopted the approach from Crommelin and Vanden-Eijnden [5], in which the conditional stochastic processes, representing the feedback from unresolved scales, are chosen to be conditional Markov chains whose properties are estimated from data of high-resolution simulations. This approach has not been applied to convection parameterization before. We used LES at convection-resolving resolutions to simulate shallow convection in a realistic manner. The data from these simulations were used to estimate ("train") the CMC.

Modelling convective turbulent fluxes with a finite-state Markov chain requires discretization of the space of possible fluxes. This was achieved by using a clustering method, in which the LES-generated heat and moisture fluxes were clustered simultaneously in order to capture the correlation between the two fluxes. The resulting cluster centroids each represent both a heat and a moisture flux profile. The CMC emulates the convective behaviour of LES by randomly jumping between the centroids, according to transition probabilities estimated from the LES data.

We demonstrated in Sect. 5 that the CMC was able to reproduce the mean vertical profile of the LES-generated fluxes and the vertical profile of their standard deviations. Tests in an SCM setting showed that the CMC was able to produce realistic fluxes, as well as an ensemble spread comparable to the spread observed in the LES data. Also, the time series of the fractions of different flux states were very similar in SCM-CMC and LES. Altogether, the CMC was well able to mimic the turbulent heat and moisture processes corresponding to shallow cumulus convection in the LES model. The CMC can be regarded as a statistical emulator of the high-resolution LES model.

We mentioned the strong anti-correlation between the turbulent heat and moisture fluxes. One could consider taking only one of the fluxes into consideration and calculating the other from it. In the subcloud layer, however, this correlation switches to positive because both surface fluxes are positive. To stay as close to LES as possible, we therefore take both fluxes into account. More about correlation between θ_l and q_t can be found in Heus [8].

The added value of this present stochastic parameterization is not so much that it is capable of reproducing the observed mean state, but more so that it is able to reproduce the fluctuations at scales in the grey zone of the relevant process, in this case shallow cumulus convection. A crucial ingredient is that the constructed Markov chain is *conditional* on the resolved-scale state. This way it is possible to have the correct temporal evolution of the states of the subgrid domains, albeit in a stochastic way, reflecting the life cycle of the clouds that live in such a subdomain. The relevance of these fluctuations for the larger scales depends on whether they will cascade up to larger scales. These effects have not been investigated within the present study.

In order to do so, one may need to take into account spatial correlations through conditioning the transition probability not only on the state of the subdomain of interest but also on the state of the neighbouring subdomains. This way one could construct a data-driven cellular automaton that would be able to create spatial mesoscale structures, assuming that such structures are present in the dataset on which the system is trained. However, this is beyond the scope of the present study.

The main purpose of this paper is to simply demonstrate that the CMC that has recently been introduced and applied to the L96 model [5], which is a low-dimensional toy model, can actually successfully be applied to complex realistic high-dimensional atmospheric processes such as shallow cumulus convection.

We also demonstrated that the range of scales where stochastic parameterizations are required goes beyond the grey zone (see Fig. 2). For the present case of rather unorganized shallow cumulus convection, the grey

zone ranges from 50 to 800 m. The range where stochastic parameterizations are required on the other hand extends to scales up to 10 km, at which there are still significant fluctuations of the turbulent fluxes amongst the various subdomains that are subjected to the same large-scale forcing.

Finally, one might ask how one can make the present CMC more general applicable. After all in the present study, the CMC has been trained to reproduce a specific realization of shallow cumulus convection (BOMEX) and will hence only be able to reproduce this realization with all its variability. Of course, the aim is to develop a stochastic parameterization that will be able to reproduce moist convection more generally under a range of different conditions. We see various possibilities of using the present CMC to "stochasticize" existing moist convection parameterizations that operate on a wide scale of conditions. One possibility is to apply the present CMC technique on a multicloud model such as put forward by Khouider et al. [15] to infer the transition probabilities from data, rather than base them on physical intuition. Alternatively, one can apply this technique to more conventional moist convection mass-flux parameterizations. One can use LES data (or real observations if available) to find parameters in the parameterizations that will strongly fluctuate when diagnosed on smaller subdomains and train the CMC in order to stochasticize the fluctuating parameters. One obvious candidate is the cloud base mass flux which is a rather constant parameter at coarse resolution but that will start to fluctuate wildly if the subdomains reach scales on the order of the size of the clouds that constitute the moist convection.

Acknowledgments The project is funded by the NWO-programme "Feedbacks in the Climate System". In addition, we acknowledge sponsoring by the National Computing Facilities Foundation (NCF) for the use of supercomputer facilities, with financial support of NWO. The authors wish to thank Frank Selten and Jerome Schalkwijk for their help and fruitful discussions.

Appendix

The *k*-means++ algorithm: (see also [3,6,20]) Given data consisting of *data points* that have to be clustered into a finite number of clusters each represented by a cluster *centroid*. Let a distance between a data point and its nearest centroid be defined.

1. Choose a data point uniformly at random from the set of data points, this will be the first centroid.
2. Select a new data point at random from the set of data points with probability proportional to the squared distance to its nearest centroid, this will be the next centroid.
3. Repeat step 2 until the number of desired centroids has been reached.
4. Assign every data point to its closest centroid to form clusters.
5. In every cluster take the mean of its data points to form new centroids.
6. Repeat step 4 and step 5 till the centroids do not change anymore.

References

1. Arakawa, A., Schubert, W.H.: Interaction of a cumulus cloud ensemble with the large-scale environment, part I. J. Atmos. Sci. **31**, 674–701 (1974)
2. Arakawa, A.: The cumulus parameterization problem: past, present, and future. J. Clim. **17**, 2493–2525 (2004)
3. Arthurm, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. Proceedings of the 18th Annual ACMSIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
4. Buizza, R., Miller, M., Palmer, T.N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Q. J. R. Meteorol. Soc. **125**, 2887–2908 (1999)
5. Crommelin, D., Vanden Eijnden, E.: Subgrid-scale parametrization with conditional Markov chains. J. Atmos. Sci. **65**, 2661–2675 (2008)
6. Gan, G., Ma, C., Wu, J.: Data clustering: theory, algorithms, and applications. SIAM, Alex, VA (2007)
7. Gerard, L.: An integrated package for subgrid convection, clouds and precipitation compatible with meso-gamma scales. Q. J. R. Meteorol. Soc. **00**, 1–19 (2007)
8. Heus, T., van Dijk, G., Jonker, H.J.J., van den Akker, H.E.A.: Mixing in shallow cumulus clouds studied by Lagrangian particle tracking. J. Atmos. Sci. **65**, 2581–2597 (2008)
9. Heus, T., van Heerwaarden, C.C., Jonker, H.J.J., Siebesma, A.P., Axelsen, S., van den Dries, K., Geoffroy, O., Moene, A.F., Pino, D., de Roode, S.R., Vil-Gueraude Arellano, J.: Formulation of and numerical studies with the Dutch Atmospheric Large-Eddy Simulation (DALES). Geosci. Model Dev. **3**, 415–444 (2010)
10. Holland, J.Z., Rasmusson, E.M.: Measurement of atmospheric mass, energy, and momentum budgets over a 500-km square of tropical ocean. Mon. Weather Rev. **101**, 44–55 (1973)

11. Honnert, R., Masson, V., Couvreux, F.: A diagnostic for evaluating the representation of turbulence in atmospheric models at the kilometric scale. J. Atmos. Sci. **68**, 3112–3131 (2011)
12. Jakob, C.: Accelerating progress in global atmospheric model development through improved parametrizations. Bull. Am. Met. Soc. **91**, 869–875 (2010)
13. Jones, T.R., Randall, D.A.: Quantifying the limits of convective parameterizations. J. Geophys. Res. **116**, 1–19 (2011)
14. Khouider, B., Majda, A.J., Katsoulakis, A.: Coarse grained stochastic models for tropical convection and climate. Proc. Natl. Acad. Sci. **100**, 11941–11946 (2003)
15. Khouider, B., Biello, J., Majda, A.J.: A stochastic multicloud model for tropical convection. Commun. Math. Sci. **8**, 187–216 (2010)
16. Lin, J.W.-B., Neelin, J.D.: Influence of a stochastic moist convective parameterization on tropical climate variability. Geophys. Res. Lett. **27**, 3691–3694 (2000)
17. Lin, J.W.-B., Neelin, J.D.: Considerations for stochastic convective parameterization. J. Atmos. Sci. **59**, 959–975 (2002)
18. Lin, J.W.-B., Neelin, J.D.: Toward stochastic moist convective parameterization in general circulation models. Geophys. Res. Lett. **30**, 1162 (2003)
19. Lorenz, E.N.: Predictability a problem partly solved. Proc. 1995 ECMWF Seminar on Predictability, Read, UK, ECMWF, pp. 1–18 (1995)
20. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symp. Math. Stat. Probab. **1**, 281–297 (1967)
21. Majda, A.J., Khouider, B.: Stochastic and mesoscopic models for tropical convection. Proc. Natl. Acad. Sci. **99**, 1123–1128 (2002)
22. Nimsaila, K., Timofeyev, I.: Markov chain stochastic parameterizations of essential variables. Multiscale Model. Simul. **8**, 2079–2096 (2010)
23. Palmer, T.N.: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. Q. J. R. Meteorol. Soc. **127**, 279–304 (2001)
24. Plant, R.S., Craig, G.C.: A stochastic parameterization for deep convection based on equilibrium statistics. J. Atmos. Sci. **65**, 10–87 (2008)
25. Shutts, G.J., Palmer, T.N.: Convective forcing fluctuations in a cloud-resolving model: relevance to the stochastic parameterization problem. J. Clim. **20**, 187–202 (2007)
26. Siebesma, A.P., Bretherton, C.S., Brown, A., Chlond, A., Cuxart, J., Duynkerke, P.G., Jiang, H., Khairoutdinov, M., Lewellen, D., Moeng, C.-H., Sanchez, E., Stevens, B., Stevens, D.E.: A Large-Eddy Simulation intercomparison study of shallow cumulus convection. J. Atmos. Sci. **60**, 1201–1219 (2003)
27. Siebesma, A.P., Cuijpers, J.W.M.: Evaluation of parametric assumptions for shallow cumulus convection. J. Atmos. Sci. **52**, 650–666 (1995)
28. Siebesma, A.P., Soares, P.M.M., Teixeira, J.: A combined Eddy-Diffusivity mass-flux approach for the convective boundary layer. J. Atmos. Sci. **64**, 1230–1248 (2007)
29. Siebesma, A.P.: Shallow cumulus convection. In: Siebesma, A.P., Plate, E.J., Fedorovich, E.E., Viegas, X.V., Wyngaard, J.C. (eds.) Buoyant Convection in Geophysical Flows, pp. 441–486. Kluwer, Pforzheim (1998)
30. Teixeira, J., Reynolds, C.A.: Stochastic nature of physical parameterizations in ensemble prediction: a stochastic convection approach. Mon. Weather Rev. **136**, 483–496 (2008)
31. Tsai, W.H.: Moment-preserving thresholding: a new approach. Comput. Vis. Graph. Image Process. **29**, 377–393 (1985)
32. Wyngaard, J.C.: Toward numerical modeling in the "Terra Incognita". J. Atmos. Sci. **61**, 1816–1826 (2004)
33. Yu, X., Lee, T.-Y.: Role of convective parameterization in simulations of a convection band at grey-zone resolutions. Tellus A **62**, 617–632 (2010)