A data-driven multicloud model for stochastic parameterization of deep convection

J. Dorrestijn[1], D.T. Crommelin[1], J.A. Biello[2], S.J. Böing[3]

[1]*CWI, Amsterdam, Netherlands.* [2]*Department of Mathematics, University of California, Davis, CA, USA.* [3]*Delft University of Technology, Delft, Netherlands.*

Stochastic subgrid models have been proposed to capture the missing variability and correct systematic medium term errors in general circulation models. In particular, the poor representation of subgrid scale deep convection is a persistent problem which stochastic parameterizations are attempting to correct. In this paper we construct such a subgrid model using data derived from large-eddy simulations (LESs) of deep convection. We use a data driven stochastic parametrization methodology to construct a stochastic model describing a finite number of cloud states. Our model emulates, in a computationally inexpensive manner, the deep convection resolving LES. Transitions between the cloud states are modelled with Markov chains. By conditioning the Markov chains on large-scale variables we obtain a conditional Markov chain which reproduces the time-evolution of the cloud fractions. Furthermore, we show that the variability and spatial distribution of cloud types produced by the Markov chains becomes more faithful to the LES data when local spatial coupling is introduced in the subgrid Markov chains. Such spatially coupled Markov chains are equivalent to stochastic cellular automata.

Key words: conditional Markov chains, stochastic cellular automata, large-eddy simulation, climate, variability.

:

1. Introduction

General circulation models (GCMs) are unable to capture the medium term variability in the tropical atmosphere. Lin et al. [1] made a comprehensive study of the tropical wave spectra determined from the IPCC GCMs and showed that none were able to reproduce the observed power spectrum [2] of convectively coupled Kelvin waves, two day waves, westward inertio-gravity waves and, least of all, the Madden-Julian oscillation [3]. These are the waves that modulate weather on intraseasonal time scales in the tropics and are increasingly seen to affect two week weather forecasts in the middle latitudes [3].

One bias that [1] identify in these GCMs is "the persistence of equatorial precipitation", which occurs at the subgrid scales. In the parlance of dynamical systems, the subgrid dynamical models quickly attain their equilibrium values and remain there too long. Palmer [4] used simple arguments from dynamical systems to show how the reduction of a chaotic dynamical system to a smaller number of degrees of freedom can suppress the chaos. While this has the obvious effect of suppressing the variability, he argued that it can have the, even more insidious,

Phil. Trans. Roy. Soc. A 1–23; doi: 10.1098/rspa.00000000
September 11, 2012, revised

effect of driving systematic errors in the mean state. A stochastic parameterization of the unresolved convection introduces variability in the GCM description of these processes and these parameterizations are increasingly being seen as the next generation of subgrid models [4, 5, 6, 7, 8, 9, 10].

Khouider et al. [7] created a stochastic multicloud model based on the deterministic multicloud model of [11]. The deterministic multicloud model was derived to correspond to the observed behaviour of tropical waves [12], where a focus on three cloud types is needed to capture the observed structure of convectively coupled waves. Furthermore, the deterministic model was calibrated so that the dynamics of the waves matched those of the tropical wave spectrum [2]. When implemented in a GCM, it has been shown to capture much of the convectively coupled equatorial wave [13] activity.

In the stochastic model [7], convection is modelled on a 2-dimensional micro-lattice by letting the local convective state at each lattice site switch randomly between four possible states (three cloud types, and clear sky) with a given probability. At the macroscopic level, the area fractions of these four states evolve randomly over time. The fractions effectively determine the feedback from the micro-scale to the macro-scale. Even in the setting of a single column [7] showed that the stochastic multicloud model has a large degree of variability. When coupled to a one dimensional dynamical core [8] it produces a large degree of gravity wave variability.

Crommelin & Vanden-Eijnden [14] proposed a data-driven stochastic parameterization methodology, where the stochastic processes driving the parameterization are systematically inferred from data (e.g. from high-resolution models). This method was used by [15] on data from a large-eddy simulation (LES) of shallow convection. This approach leads to a model with random jumps between a finite number of possible subgrid-scale states where, both the discrete states as well as the switching probabilities, are estimated from data. Furthermore, the switching probabilities are dependent (conditional) on the macroscopic, resolved-scale state of the atmosphere.

For the shallow convection parameterization in [15], vertical turbulent fluxes of heat and moisture were collected from the LES data and discretized using a clustering method. By contrast, the discrete states used in [7] are cloud types (congestus clouds, deep convective clouds, stratiform clouds, and clear sky) rather than flux states. The states and switching probabilities used in [7] are based on physical intuition and observations; they are not inferred from data.

The objective of the current study is to determine a stochastic multicloud parameterization approaches from [7] using a data driven approach [14, 15]. Much as in [7], we use pre-specified cloud types as a basis for discretizing the subgrid-scale states, and study their (time-evolving) fractions on macroscopic domains. The precise discretization, as well as the switching probabilities and the conditioning on the resolved-scale state, are all inferred from LES data, as in [14, 15].

Specifically, we use eight hours of simulation of the development of tropical convection based on an idealization of observed conditions in North-West Brazil [16]. Simulated cloud top and rain water path are stored to classify states on the LES (horizontal) grid nodes. We use five states: *clear sky* (1) and the four cloud types *shallow cumulus* (2), *congestus* (3), *deep* (4) and *stratiform* (5). Strictly speaking clear sky is not a cloud type, but from now on we will refer to five cloud

types. At the beginning of the simulation, only clear sky is present. Gradually, shallow cumulus develops, followed by (raining) congestus clouds. After about five hours, deep convective towers with heavy precipitation develop. The deep convective towers turn into passive stratiform decks that spread and dissolve.

The paper is organised as follows. In Section 2 we discuss how we model transitions between cloud types with Markov chains, and how these Markov chains can be made conditional on the environment, or on the cloud types at neighboring lattice sites. We describe the LES data and specify the cloud classification in Section 3. The stochastic multicloud model is described in Section 4. In Sections 5-7 we infer the transition probabilities of the Markov chains and assess their ability to reproduce (emulate) the cloud filling fractions from the LES data. In Section 5 we use a Markov chain without conditioning, in Section 6 a Markov chain conditioned on the environment, and in Section 7 a Markov chain conditioned on cloud types at neighboring lattice sites. Then, we discuss implementation of the multicloud model into a simple single column model (Section 8), again calculating cloud filling fractions. Finally, conclusions about our multicloud model, how stochastics can change dynamics and its implications for climate models are given in Section 9.

2. Modelling cloud type transitions with Markov chains

A central element in the stochastic parameterization approach used here and in [7, 14, 15] is discretization of the subgrid-scale (e.g., convective) states. Here, each grid point at the microscopic level can be in only one of five possible states. Let us denote by $Y_i(t) \in \{1, 2, 3, 4, 5\}$ the state at time t at grid point i. The time evolution of $Y_i(t)$ is modelled as a Markov chain (MC), so $Y_i(t)$ changes randomly in accordance with a set of transition probabilities. In the most basic form, these probabilities are simply

$$p(\alpha, \beta) = \text{Prob}\left(Y_i(t + \Delta t) = \beta \,|\, Y_i(t) = \alpha\right). \tag{2.1}$$

However, in this basic formulation, the probability of e.g. a congestus state at grid point i turning into a deep convective state is independent of the environment (macroscopic state) for i. To include such dependency, in [7, 14, 15] the transition probabilities are *conditioned* on the macroscopic state. If we denote by $X_i(t)$ a variable that is representative of the environment of i (e.g. convectively available potential energy (CAPE), convective inhibition (CIN), or mid-troposphere relative humidity), the transition probabilities of such a conditional Markov chain (CMC) are

$$p_\gamma(\alpha, \beta) = \text{Prob}\left(Y_i(t + \Delta t) = \beta \,|\, Y_i(t) = \alpha,\ X_i(t) = \gamma\right). \tag{2.2}$$

As can be seen, the transition probabilities in (2.1) and (2.2) are not explicitly dependent on the convective states of neighbouring grid points. If i and j are neighbouring grid points, Y_i and Y_j are completely uncoupled in case of (2.1). They are coupled indirectly via X_i and X_j in case of (2.2), because X_i and X_j are coupled at the macroscopic level. Since i and j are neighbouring grid points, X_i and X_j will be strongly correlated. In this paper we also explore explicit conditioning on the neighbourhood, as this is likely to improve the spatial correlation of the parameterized convection patterns. We do this by considering

the conditional transition probabilities

$$p_\delta(\alpha,\beta) = \mathrm{Prob}\left(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha,\ Y_{\{i\}}(t) = \delta\right), \tag{2.3}$$

and

$$p_{\gamma,\delta}(\alpha,\beta) = \mathrm{Prob}\left(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha,\ X_i(t) = \gamma,\ Y_{\{i\}}(t) = \delta\right), \tag{2.4}$$

where $\{i\}$ denotes the neighbourhood of i (e.g., the 8 direct neighbours on the lattice). We note that by conditioning the Markov chain on neighbouring states, as in (2.3), the Markov chain effectively becomes a stochastic cellular automaton (SCA). A schematic overview of the generalizations of the Markov chains is shown in Fig. 1.

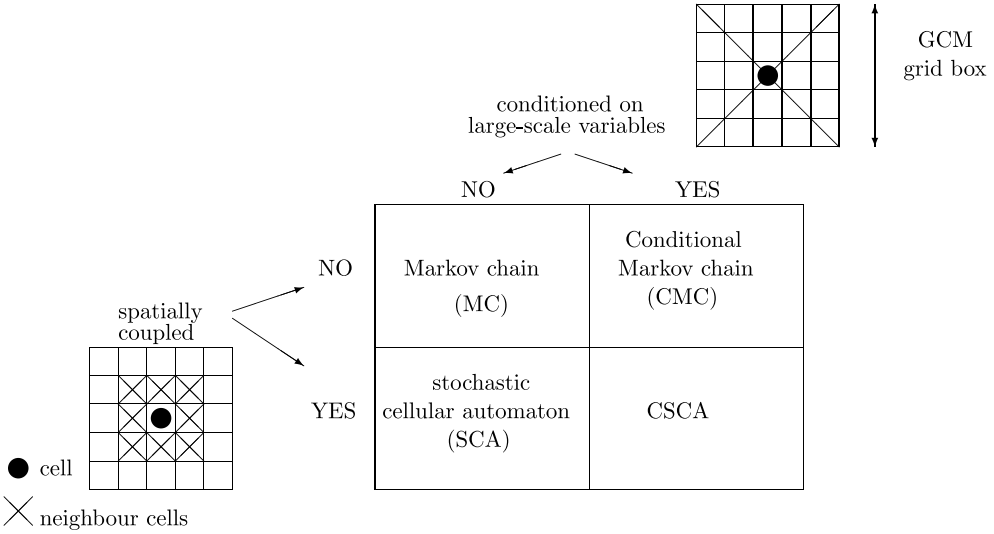

Figure 1. A Markov chain can be conditioned on the macroscopic state to obtain a CMC or on the state of the nearest neighbours to obtain a SCA.

Each gridpoint on the micro lattice has a state that evolves randomly according to the same set of transition probabilities, e.g. (2.2). At the macroscopic level, square blocks of micro lattice sites are grouped together, and we study the filling fractions (or area fractions) of the various convective states. For each block we have

$$\sigma_\alpha(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(Y_i(t) = \alpha), \tag{2.5}$$

where n is the number of micro lattice sites in the macroscopic block, and $\mathbf{1}(.)$ is the indicator function. The filling fractions are time-dependent and random, and must sum up to one for each macroscopic block: $\sum_\alpha \sigma_\alpha(t) = 1$ for all t. By matching the size of the macroscopic blocks to the (horizontal) size of GCM model grid boxes, the filling fractions can be used as input for parameterizing vertical transport due to convection.

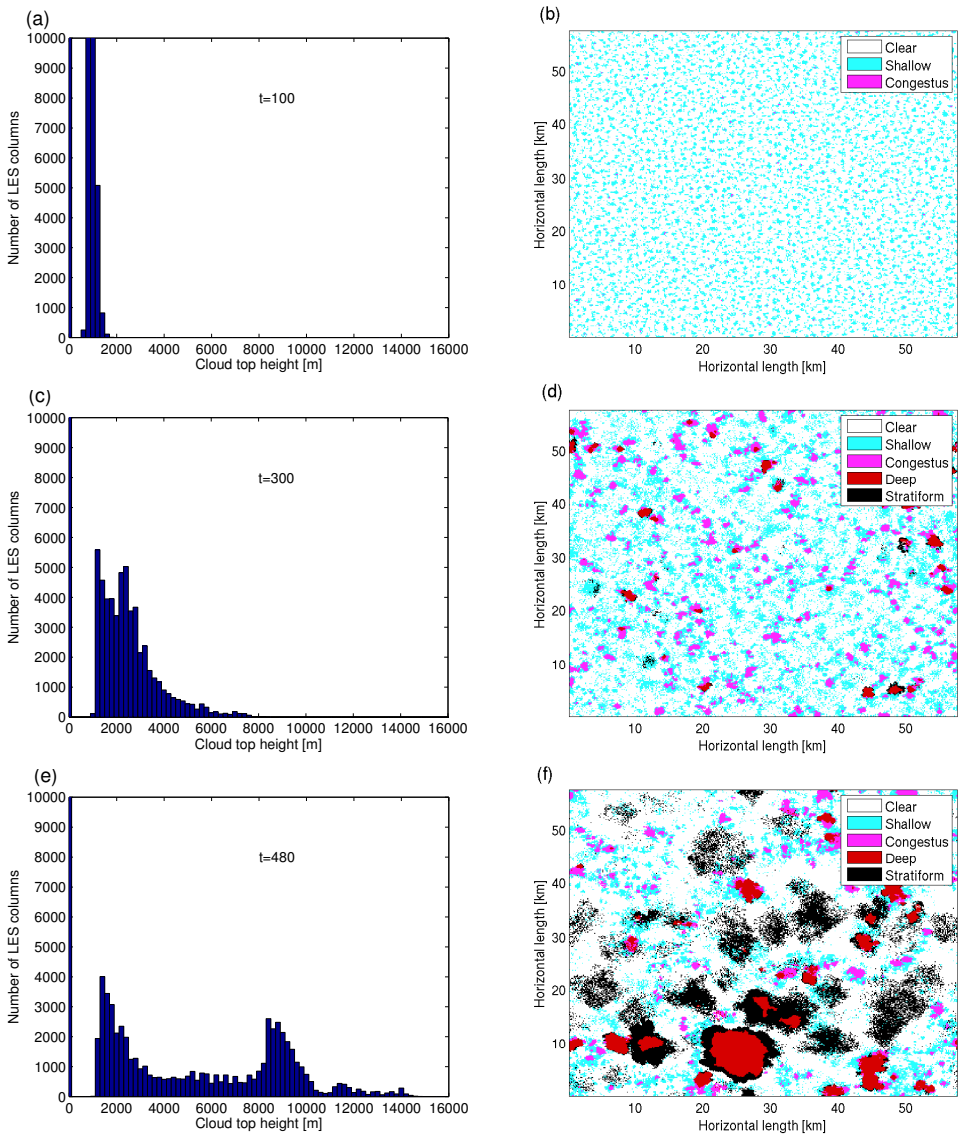

Figure 2. (a,c,e) Histograms of the cloud top at different time instances of the simulation. (b,d,f) Three snapshots of the LES field for which all columns are assigned to one of the five cloud types.

3. Large-eddy simulation

We use the Dutch Atmospheric LES (DALES) model to produce high-resolution data. DALES is a non-hydrostatic model that resolves atmospheric convection explicitly by solving the spatially filtered Navier-stokes equations under the anelastic approximation. The model has an ice microphysics scheme, but does not

account for latent heat release due to freezing. For further details about DALES we refer to [17]. The simulation is based on an idealization of observed conditions [16] during the tropical convection experiment TRMM-LBA carried out in North-West Brazil in 1998/1999. There is no horizontal shear, and surface heat and moisture fluxes are held constant throughout the simulation. At the start of the 8-hour simulation, the entire LES domain consists of clear sky. Convection develops gradually, first shallow convection, eventually (after about five hours) also deep convection. We emphasize that it is a non-stationary case of the development of deep convection. The simulation and the resulting data are described in more detail by [18].

The horizontal size of the LES domain is 57.6×57.6 km^2 and the vertical extent is 25 km. The horizontal grid spacing is 150 meter and the vertical spacing increases exponentially from 40 near the surface to 200 meter at the upper levels. For every column we store the simulated cloud top height, rain water path (the vertically integrated rain water content), CAPE and CIN. We also store liquid water potential temperature θ_l and total water specific humidity q_t at two model levels, one in the boundary (subcloud) layer at 413 meter, the other in the lower free troposphere at 2345 meter. These variables are defined by

$$\theta_l = \theta - \frac{L}{c_p \pi} q_l \qquad \text{and} \qquad q_t = q_v + q_l \,, \tag{3.1}$$

with θ the potential temperature, L the latent heat of vaporization, c_p the specific heat of dry air at constant pressure, q_l the non-raining liquid water content and q_v the water vapour specific humidity. Furthermore, π is the Exner function, the ratio of absolute and potential temperature. In the absence of precipitation θ_l and q_t are conserved for moist adiabatic processes. We store the data at time intervals of one minute during eight hours, resulting in 480 time slices of the variables mentioned above in each of the 384×384 LES model columns. Below we discuss how these variables are used for classification of each model column state into five cloud types.

(*a*) *Classification of cloud types*

In the vein of [19] and [7] we consider five cloud types: clear sky, shallow cumulus, congestus, deep convection and stratiform. Fig. 2 (left) shows histograms of the cloud top height. At $t = 480$ we see three categories (clear sky, low clouds and high clouds), which can be well distinguished with thresholds at 200 meter and 5000 meter. Furthermore, to distinguish the heavily raining deep convective towers from their passive, modestly raining stratiform remnants, we use the rain water path divided by the cloud top height. We call this the column rain fraction:

$$CRF := \frac{\text{rain water path}}{\text{cloud top}}. \tag{3.2}$$

By dividing by the cloud top height we obtain a measure of the rain intensity, from which the vertical extent of raining cloud has been factored out. The CRF makes it easier to identify stratiform clouds, which have high cloud top and low, but not always negligible rain water path. Furthermore, we can use the same threshold of the CRF, 10^{-5}, to distinguish deep from stratiform as well as non-raining shallow cumulus from raining congestus clouds. In Fig. 3 we plot the CRF against the

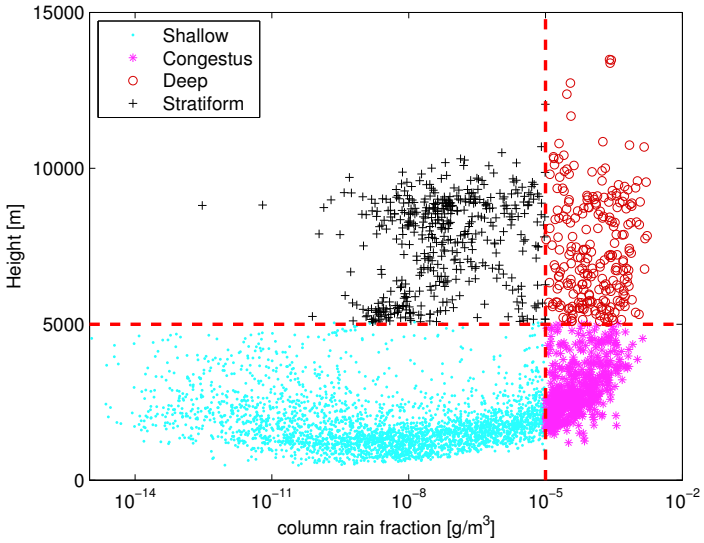

Figure 3. Classification of cloud types using cloud top and CRF.

Table 1. Classification of the clouds. CRF defined in (3.2).

Cloud type	cloud top	rain
Clear sky	n/a	n/a
Shallow cumulus	$200 \le h < 5000$ m	CRF $\le 10^{-5}$
Congestus	$200 \le h < 5000$ m	CRF $> 10^{-5}$
Deep	$h \ge 5000$ m	CRF $> 10^{-5}$
Stratiform	$h \ge 5000$ m	CRF $\le 10^{-5}$

cloud top height and indicate 4 cloud types with different symbols. The clear sky group is not shown because its CRF is not well-defined. In Table 1 we summarize the cloud classification.

We can now assign the state of each LES column, at every time step, to one of the five cloud types. Fig. 2 (right) shows snapshots of the LES domain with all columns assigned to one of the cloud types. At $t = 100$ we see clear sky sites combined with shallow cumulus clouds and some congestus clouds. At $t = 300$ the development of deep towers start. At $t = 480$ we see larger deep towers and dissolving stratiform decks.

4. The stochastic multicloud model

With the LES data discretized according to Table 1, we can choose the size of the macroscopic blocks and calculate the filling fractions $\sigma_\alpha(t)$ on each of these blocks using (2.5). In what follows, the LES blocks always consist of 32×32 microscopic lattice sites (so that $n = 32^2$), unless explicitly stated otherwise. The corresponding physical size of these blocks is 4.8 km by 4.8 km. The total LES domain is covered by 12^2 of such (non-overlapping) blocks. In Fig. 4a we show the time evolution of the means and standard deviations of the filling fractions, taken over the 12^2 different blocks. We emphasize that these are the filling fractions as computed directly from the LES data.

With the stochastic multicloud model, we aim to emulate the time evolution of the LES filling fractions. This is done by evolving the state (cloud type) of each micro lattice site as a Markov chain. The states on the micro lattice sites can be grouped again in macroscopic blocks (of any desired size), leading to emulated filling fractions. As already mentioned, the number of Markov chains grouped together in the multicloud model in one macroscopic block will be 1024, except for the creation of plots in Fig. 7b, Fig. 8b and Fig. 11b where we use blocks of 64 Markov chains.

The transition probabilities that characterize the Markov chain are of the form (2.1), (2.2), (2.3) or (2.4). Their numerical values are estimated from the LES data. We use a time step Δt of 1 minute, matching the saving time step of the LES data. We assess the performance of the various forms (2.1) - (2.4) in the following sections. The choice of the macroscopic environment variable $X_i(t)$, used in (2.2) and (2.4), are discussed there as well.

Eventually, the multicloud model has to provide not just filling fractions, but vertical profiles for heating and moistening that can be used for parametrization purposes in a GCM. In Section 8b we explain how we deal with heating and moistening in a single-column model experiment.

5. Markov chains

We start by using the simplest form (2.1), i.e. the form where the Markov chain is not conditioned on macroscopic environment variables or on neighbour states. The transition probabilities determine a single 5×5 stochastic matrix in which the entry at the k-th row and l-th column is the probability that a site that is in state k will switch to state l in the next minute. We count transitions in the LES data to estimate the transition probability matrix, resulting in

$$\hat{\mathbf{M}} = \begin{pmatrix} 0.95 & 0.04 & 0.00 & 0.00 & 0.00 \\ 0.14 & 0.84 & 0.02 & 0.00 & 0.00 \\ 0.02 & 0.06 & 0.90 & 0.02 & 0.00 \\ 0.01 & 0.00 & 0.03 & 0.94 & 0.03 \\ 0.10 & 0.03 & 0.00 & 0.01 & 0.86 \end{pmatrix}$$

We use all data of the entire simulation to estimate transition probabilities. In this case we do not take into account the strong dependence of the transitions on time. The reader is reminded that the case we consider is a non-stationary case of

the development of deep convection. Next, we will test the skills of this Markov chain.

(*a*) *Filling fractions of the MC*

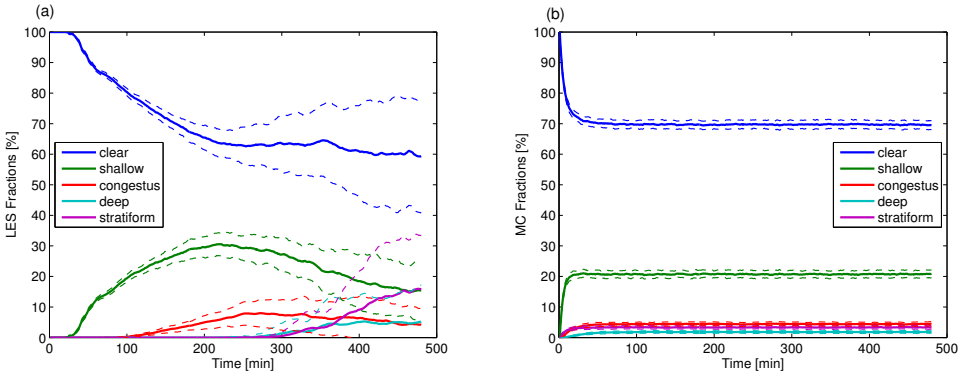

Figure 4. (a) Mean filling fractions observed in the LES data using $n = 32^2$ micro lattice sites per macroscopic block (solid) plus and minus the standard deviations over the 12^2 macroscopic blocks (dashed) and (b) reproduced mean filling fractions using 1024 MCs (solid) plus and minus the standard deviation of 144 realizations (dashed).

Fig. 4 shows cloud filling fractions observed in the LES data and reproduced by the Markov chain. The Markov chain filling fractions converge quickly to the filling fractions that correspond to the invariant distribution of the transition matrix. These fractions are therefore accurate in the sense that they are in agreement with the time averages of the fractions observed in the LES data. However, the standard deviations are too small and the overall time evolution of the LES cloud fractions is not captured at all.

From the results in Fig. 4 we can conclude that a Markov chain governed by (2.1) is not capable of emulating the LES cloud fractions satisfactorily. A longer time step (20 minutes) for the Markov chain did not improve any of these deficiencies (results not shown). Rather, the shortcomings are due to the insensitivity of the MC to both the macroscopic environment and the neighbour states. A natural way of to improve on this is to include dependency on environment or neighbours. Thus, in the next sections we generalize the Markov chain (2.1) by

1. conditioning on the macroscopic state (environment), leading to the conditional Markov chain (CMC) form (2.2), or

2. coupling to neighbouring cells, leading to the stochastic cellular automaton (SCA) form (2.3) .

In the most general form (2.4), both environment and neighbouring states are included. A schematic overview of these generalizations was shown in Fig. 1.

6. Conditional Markov chains

In this section we explore conditioning of the Markov chains on a function of some large-scale variables that could be resolved in a GCM. Large-scale variables such as CAPE, CIN, middle troposphere relative humidity, or (moist) convergence are considered to be potential indicators of convective behaviour. In Section 6a we discuss how mutual information can be used as an objective measure to quantify how good these indicators are.

For now, to explain our method we choose to condition on the CAPE and the CIN. These functions of large-scale variables have been used before in e.g. [20] and [7]. A reversibly lifted adiabatic parcel, using the mean thermodynamic properties at the 200-400 meter level is used to calculate the CAPE and the CIN in every LES model column. In the present context, CAPE and CIN mostly indicate the evolution of the surface properties, rather than the state of the free troposphere. CAPE and CIN are affected both by the gradual moistening and heating by surface fluxes and by the presence of cold pools (see e.g. [21]). The values depend on the choice of variable used to construct the adiabats, in our case θ_l. Although the CAPE values reported here, maximum values of around 4500 J/kg, are higher than what we had expected, seasonally averaged values as high as 7000 J/kg have been reported over tropical land masses by [22].

As before, we divide the whole LES domain in 12^2 macroscopic blocks (subdomains) and calculate spatial averages of CAPE and CIN on these subdomains. We thereby obtain 12^2 paths in the CAPE-CIN space, each 480 minutes long. An even larger part of the CAPE-CIN space could be sampled by combining data from several LES runs with different initial profiles for temperature and humidity; we will not explore this here.

After obtaining the paths in the CAPE-CIN space, we cluster the CAPE-CIN data points in K clusters using the K-means++ algorithm [24, 25, 26]. While clustering the CAPE-CIN space, we use the Euclidean distance with different rescaling factors for CAPE and CIN. The rescaling factors are such that the mean contribution to the distance to the centroids is equal for CAPE and CIN. The clustering algorithm also works for all other (combinations of) large-scale variables, with other scaling factors. The number of clusters K has to be chosen beforehand. It should be as small as possible, because for every cluster a 5×5 transition matrix has to be estimated. We refer to [15] and [23] where clustering has been used to construct conditional Markov chains.

In Fig. 5 we show the result of the clustering using $K = 20$. For $K = 20$ we will show that the CMCs are able to reproduce the correct filling fractions (see Section 6b). All 12^2 paths start at CIN ≈ 80 J/kg and CAPE ≈ 2400 J/kg. Then, CAPE increases and CIN decreases almost uniformly in the domain. When deep convection sets in, the domain starts to become very inhomogeneous, resulting in CAPE and CIN values that differ substantially over the subdomains. After the CAPE-CIN space is divided into K regions, the paths in the CAPE-CIN space can be mapped to paths in the space of cluster centroids.

To sum up: first we calculate the (time-evolving) subdomain averages of CAPE and CIN from the LES data, then we cluster these CAPE-CIN averages. To determine the environment state $X_i(t)$ for micro lattice site i we use the discretized (clustered) CAPE-CIN state of the subdomain to which site i belongs. Thus, $X_i(t)$

effectively takes values in the set of cluster indices: $X_i(t) \in \{1, 2, ..., K\}$. Using this $X_i(t)$ in the manner of (2.2) to condition the transition probabilities implies that we have a transition probability matrix associated with each CAPE-CIN cluster.

These transition probability matrices are estimated by counting transitions in the LES data (see also [14]). To estimate the probability $p_\gamma(\alpha, \beta)$ defined in (2.2) we use the estimator

$$\hat{p}_\gamma(\alpha, \beta) = \frac{T_\gamma(\alpha, \beta)}{\sum_\beta T_\gamma(\alpha, \beta)}\,, \tag{6.1}$$

where $T_\gamma(\alpha, \beta)$ is the number of cloud type transitions $\alpha \to \beta$ observed in the LES data with $X_i(t) = \gamma$. Thus,

$$T_\gamma(\alpha, \beta) = \sum_{t,i} \mathbf{1}(Y_i(t + \Delta t) = \beta)\, \mathbf{1}(Y_i(t) = \alpha)\, \mathbf{1}(X_i(t) = \gamma) \tag{6.2}$$

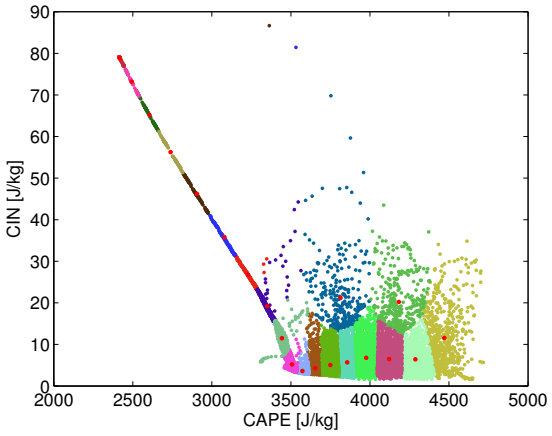

Figure 5. Clustered paths forming $K = 20$ regions in the CAPE-CIN space. The red dots are cluster centroids.

(*a*) *Mutual information between environment and cloud type*

Large-scale variables such as CAPE, CIN or middle troposphere relative humidity are considered to be potential indicators of convective behaviour. Below we discuss how mutual information can be used as an objective measure to quantify how good these indicators are.

Suppose we have two discrete random variables with a joint probability mass function $p^J(x, y)$ and marginal probability mass function $p(x)$ and $p(y)$. Then, the mutual information is the relative entropy or Kullback-Leibler distance between the joint distribution p^J and the product distribution $p^P(x, y) = p(x)p(y)$. It is

Large-scale variable(s)	Information
RH at 2345 meter & CIN	0.0992
RH at 2345 & w at 413	0.0948
CAPE & CIN	0.0946
CAPE & w at 413 meter	0.0897
CIN & w at 413 meter	0.0809
CIN	0.0757
RH at 2345 meter & CAPE	0.0710
w at 413 meter	0.0697
CAPE	0.0589
RH at 2345 meter	0.0590
u at 15843 meter	0.0290

Table 2. Mutual information between large-scale variables and cloud type at 4.8×4.8 km^2 subdomains.

given by

$$I(p^J, p^P) = \sum_{x,y} p^J(x,y) \log\left(\frac{p^J(x,y)}{p^P(x,y)}\right)$$

where the sum is over all values of x and y. $I(p^J, p^P)$ quantifies how much additional information p^J contains relative to p^P. For more details about mutual information and other information-theoretic concepts we refer to [27].

In our case x and y are the environment state $X_i(t)$ and the cloud type $Y_i(t)$ at the same location, respectively. The mutual information between their joint distribution and the product of their marginal distributions quantifies how good an indicator $X_i(t)$ is for $Y_i(t)$, and thus how useful it is to condition the Markov chain for Y_i on X_i. In [28], similar use is made of mutual information to select useful indicators for stochastic cellular automata. We note that in our case, the joint and marginal distributions are non-stationary, therefore we calculate the mutual information separately for every time t of the LES dataset.

In Fig. 6 we show three time series for mutual information between the large-scale variables and the cloud type. In the beginning of the simulation the mutual information is zero. The reason is that clouds have not evolved yet, and therefore the large-scale variables do not give information about the presence of a cloud. The mutual information is first calculated for every time instance and then the average is calculated over the last two hours (the phase in which deep convection is developed) to obtain a single value for the mutual information such that we can compare different choices of the large-scale variables. In Table 2 we list the time-averaged mutual information using various (clustered) quantities for X_i. To give an interpretation to the value of (mutual) information in nats we mention that the mutual information between the cloud type and the cloud type itself is 1.1486 (this would be the best possible score).

The result in Table 2 shows that the combination of CAPE and CIN gives significantly more information about cloud type than either of them alone. We see that both the vertical velocity field and the CAPE/CIN fields contain information on the state of convection. Both of them may be used to reproduce some of the

time-dependent behaviour of convective organisation in low wind shear (e.g. cold pools). Here we choose for CAPE and CIN to obtain the best filling fractions. A more detailed study of the physical mechanisms behind the organisation of deep convection in the present case is given in [18].

As a final remark, we have included the mutual information of u at 15843 meter in Table 2 as a consistency check: u at 15843 meter is mainly determined by upward propagating gravity waves that can have a remote origin, and we do not expect it to be a good indicator of the state of convection and cloud type. The low value of the mutual information confirms this intuition.

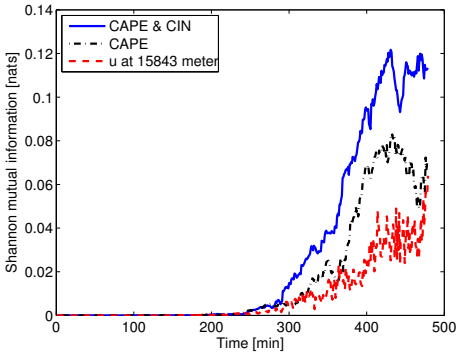

Figure 6. Time series of the mutual information between the large-scale variable at time t and cloud type at time t for different large-scale variables.

(*b*) *Filling fractions of the CMC*

Fig. 7 shows filling fractions produced by CMCs that are conditioned on CAPE and CIN with $K = 20$ clusters. The left panel shows the means and standard deviations of the fractions over 144 macroscopic blocks using 1024 CMCs per block. The time evolution of the means is in good agreement with the LES results, as can be seen by comparing with the left panel of Fig. 4. With a smaller number of clusters ($K = 10$) the agreement was unsatisfactory (results not shown). Further, the standard deviations are too small compared to the LES results. They can be increased by using a smaller number of CMCs (because fractions determined by a smaller number of Markov chains are more likely to deviate from the expected values). In the right panel of Fig. 7 we show the means and standard deviations using only 64 CMCs per macroscopic block. As expected, by using only 64 instead of 1024 CMCs, the standard deviations are larger and therefore in better agreement with the LES fractions. In Fig. 8 we show cloud filling fractions on a single macroscopic block. In the left panel the fractions of the LES data on a block of size $n = 1024$, in the right panel the fractions as produced by the multicloud model using 64 CMCs (conditioned on CAPE-CIN).

We have seen that by using CAPE and CIN to condition the CMCs, the time-evolution of the filling fractions is captured. This is not solely because CAPE and CIN are indicators of convection: in the first part of the simulation, CAPE

increases (and CIN decreases) steadily with time, so that conditioning on CAPE and CIN is similar to conditioning on time. However, this only holds true for the first part of the eight hours of simulation. In the last hours, CAPE no longer increases in all LES subdomains. Instead, we observe a decrease of CAPE in part of the subdomains.

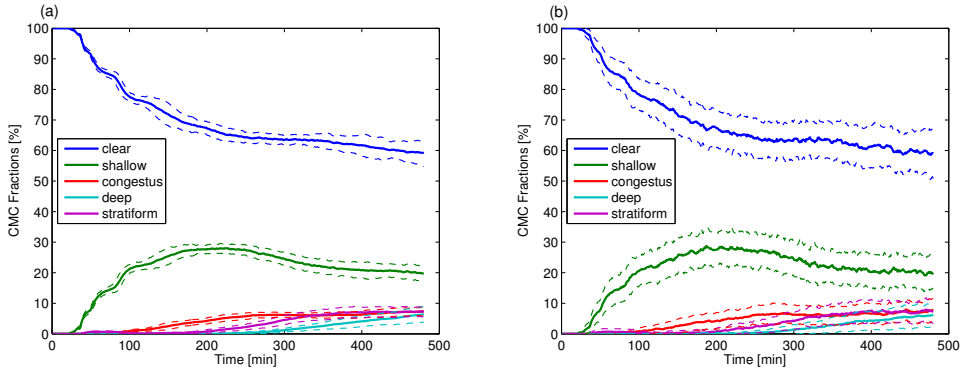

Figure 7. (a) Mean filling fractions produced by 1024 CMCs with $K = 20$ clusters of CAPE and CIN (solid) plus and minus the standard deviation (dashed). The CMC is driven by LES observed values of CAPE and CIN. (b) same as left but with 64 CMCs.

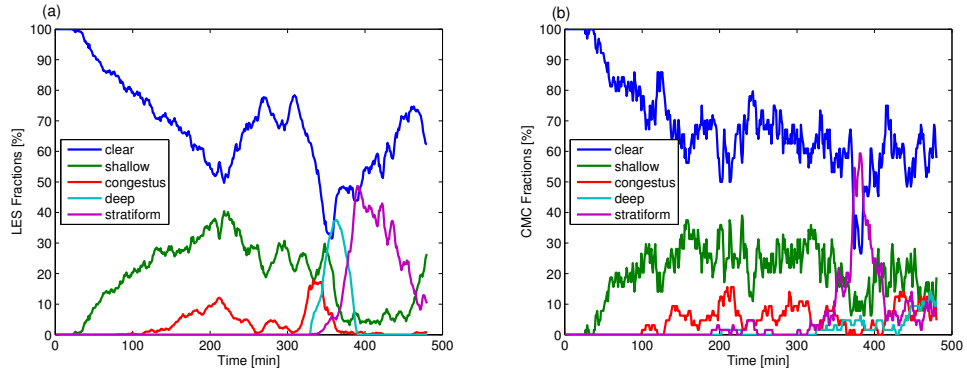

Figure 8. (a) Filling fractions observed in a single macroscopic block of $n = 32^2$ LES columns. (b) Filling fractions using 64 CMCs, where each CMC is conditioned on CAPE and CIN with $K = 20$.

7. Stochastic cellular automaton

In the previous section it was shown that conditioning of the Markov chain on the macroscopic environment strongly improves the behaviour of the filling fraction means, cf. Fig. 4 and Fig. 7. However, the variances of the CMC filling fractions are too small, and can only be brought in better agreement with the variances of the LES filling fractions by reducing the number of CMCs per macroscopic block. In this section we investigate whether coupling to neighbouring sites on the micro lattice can improve the emulated variances, without reducing the number of Markov chains. Thus, we study use of the forms (2.3) and (2.4) for the Markov chain. We expect that by coupling to neighbouring sites, the spatial correlations of the cloud type patterns will be better captured, thereby increasing the variance.

As mentioned earlier, by conditioning the Markov chain for lattice site i on the state of the neighbouring sites, as in (2.3), the Markov chain becomes a stochastic (or probabilistic) cellular automaton (SCA). Cellular automata (CA) have been used for parameterization purposes by [5, 6, 29]. In these studies, the CA have deterministic rules, not stochastic ones, and they are chosen by intuition rather than inferred from data. Also, in [5, 6, 29] the cells of the CA can take on two states, not five as is the case here.

First we estimate the SCA transition probabilities (2.3) from the LES data. As before, $Y_i(t)$ is the cloud type at site i at time t, $Y_i(t) \in \{1, 2, 3, 4, 5\}$. Use of (2.3) implies that in principle, for every state δ of the combined neighbouring sites $Y_{\{i\}}$, there is a different transition probability matrix. This reflects, for example, that the probability of a clear sky site turning into a shallow cumulus site may increase as the number of neighbouring shallow cumulus sites increases.

For the neighbourhood of site i, denoted $\{i\}$, we choose the 8 sites directly surrounding site i in the micro lattice (see also Fig. 1). As each site can take on 5 different values, there are 5^8 different configurations, i.e. 5^8 possible values of δ. This is too much to be practical, therefore we reduce the number of possibilities by conditioning not on $Y_{\{i\}}(t)$ directly, but on a simple reduction function f that depends on $Y_{\{i\}}(t)$. Thus, we use

$$p_\delta(\alpha, \beta) = \mathrm{Prob}\left(Y_i(t + \Delta t) = \beta \,|\, Y_i(t) = \alpha,\ f(Y_{\{i\}}(t)) = \delta\right) \tag{7.1}$$

rather than (2.3) itself.

Let us denote by $|CL|_i$ the number of clear sky sites directly surrounding i, and similarly by $|SH|_i$, $|CO|_i$, $|DE|_i$ and $|ST|_i$ the number of surrounding shallow, congestus, deep and stratiform sites. These numbers are time dependent. Clearly, $|CL|_i + |SH|_i + |CO|_i + |DE|_i + |ST|_i = 8$ for all i and at all times. As the function f we now choose

$$f(Y_{\{i\}}(t)) = 1 * |SH|_i + 2 * |CO|_i + 3 * |DE|_i + 4 * |ST|_i \,. \tag{7.2}$$

The reason for choosing this particular reduction function is that it is a measure of the degree to which the direct environment is convectively active: the more neighbouring sites in a state of convection the larger the value of f. Furthermore, a neighbouring site with cloud type congestus increases f more than a neighbouring site with cloud type shallow. The function increases even more if there is a neighbouring deep site. The choice of the factor 4 for stratiform is somewhat debatable, but the coefficient has to be larger than 3 to indicate the presence

of stratiform instead of some other cloud type. Further the value has to be as small as possible to reduce the number of states (and therefore matrices) as much as possible. One can use information theory to perform a systematic search for functions that give the most information about the transitions (see [28] for some ideas on this), however we will not pursue this here. Estimating the probabilities (7.1) is straightforward, using an estimator analogous to (6.1)-(6.2).

We obtain 33 different transition matrices of size 5×5, because $0 \leq f \leq 32$. For each site, the state of the neighbourhood is determined by counting the numbers of different cloud types surrounding it, and computing the corresponding value of $f_i(t)$ as in (7.2). This value determines which transition matrix is used at lattice site i at time t.

We initialise the SCA-multicloud model using 384×384 cells all in a clear sky state, corresponding to the initial condition observed in the LES data. As time evolves, some cells switch to shallow cumulus and clusters of shallow cumulus cells appear. Later on, the SCA correctly produces congestus sites in the shallow cumulus clusters. At about 250 minutes after initialization, similar to LES, deep convective sites appear. These turn into stratiform decks. Eventually, the patterns of the SCA are clear sky areas with some shallow cumulus and areas of a mixture congestus, deep and stratiform. This mixture is not observed in the LES data, but the fractions turn out to be correct. First we show the patterns produced by the SCA in Fig. 9a.

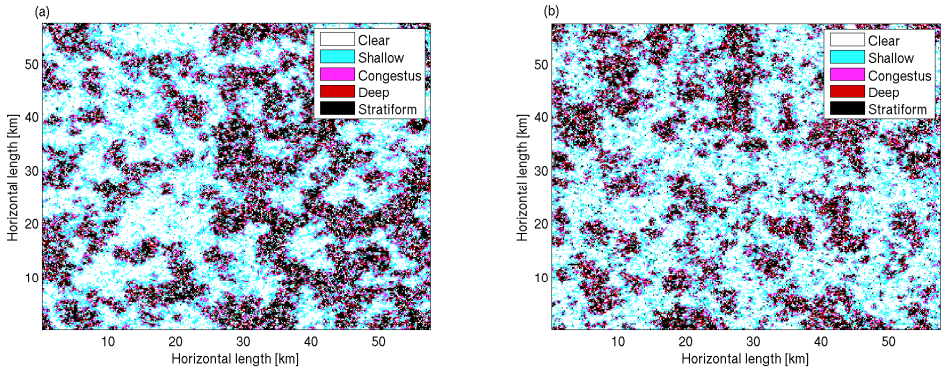

Figure 9. Patterns formed (a) by SCA at $t = 480$ and (b) by CSCA additionally conditioned on CAPE using $K = 5$ clusters.

Fig. 10a shows filling fractions (mean and standard deviation) for the SCA, using (7.1)-(7.2). The standard deviation is taken over macroscopic blocks of size $n = 1024$. Both the time evolution and the magnitude of the standard deviations are in much better agreement with the LES data (Fig. 4a) than those produced by the CMC (Fig. 7). The time evolution of the means are reasonable, but not as good as those of the CMC. Therefore, as a final step of refinement, we combine CMC and SCA by conditioning the Markov chain both on the macroscopic state $X_i(t)$ and on the neighbouring states $Y_{\{i\}}(t)$. We refer to this combination as CSCA (conditional SCA). To our best knowledge, a (stochastic) cellular automaton conditioned on an "external", time-evolving field (X, in our case) has not been studied before.

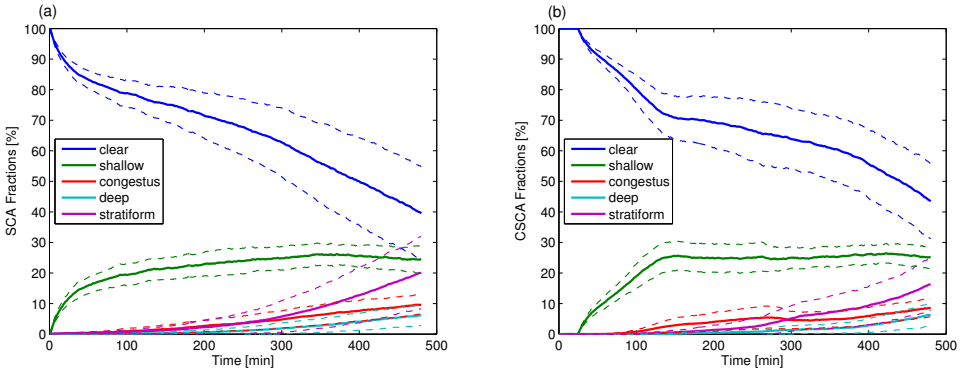

Figure 10. (a) Mean filling fractions of the SCA (solid) plus and minus the standard deviation calculated over blocks of 1024 cells (dashed) and (b) the same for a CSCA conditioned on CAPE using $K = 5$ clusters.

The filling fractions of the CSCA are shown in Fig. 10b. As before, we used the function (7.2) rather than $Y_{\{i\}}(t)$ to condition the CSCA on the neighbouring sites. Thus, the transition probabilities are as in (2.4), but with $Y_{\{i\}}(t)$ replaced by the function (7.2). For conditioning on the macroscopic state $X_i(t)$ we used CAPE, clustered with 5 centroids. Patterns are similar to the patterns of the SCA, compare the panels of Fig. 9. The time evolution of the filling fraction means is in better agreement with the LES data than was the case with the SCA. We anticipate that further improvement is possible, e.g. with search techniques as in [28], and with methods to reduce the parameter space as in [23]. We leave this for future study.

8. Single-column model

In the tests performed in the previous sections there was no interaction between the large-scale variables and the CMC or CSCA. Therefore, to take a step forward towards implementation in a GCM, we test the multicloud model in an SCM experiment. The SCM can be thought of as representative for the behaviour of a single GCM vertical model column. We use one macroscopic block, containing 1024 CMCs, to represent the GCM model column. These CMCs are conditioned on CAPE and CIN, as in Section 6. We choose suitable large-scale variables and use LES data to precalculate their tendencies. The tendencies are assumed to depend linearly on the filling fractions determined by the multicloud model. Thus, the large-scale variables and the cloud filling fractions are coupled to each other, and both evolve over time. Inspired by [7] we take four prognostic variables: $X_1 = q_t^{\text{low}}$, $X_2 = q_t^{\text{high}}$, $X_3 = \theta_l^{\text{low}}$ and $X_4 = \theta_l^{\text{high}}$, with q_t and θ_l as defined in (3.1). The low level is at 413 meter and the higher level is at 2345 meter in the atmosphere. These are the variables that we are going to resolve in our SCM.

We use the CMCs, conditioned on CAPE and CIN, to calculate the filling fractions of each cloud type. Therefore we have to express CAPE and CIN in terms of the prognostic variables $X = (X_1, \ldots, X_4)^T$.

(a) *CAPE* and CIN**

We assume that CAPE is a linear combination of X. We compute the coefficients by doing a least square fit with the CAPE values from the LES data and the values of X, also from the LES data. We write

$$\text{CAPE}^* = \lambda X, \tag{8.1}$$

where $\lambda = (\lambda_0, \ldots, \lambda_4)$ are the coefficients and where we add the constant term $X_0 = 1$. We solve

$$\min_{\lambda} \left((\text{CAPE} - \lambda X)^2 \right)$$

and find that the linear CAPE* is almost completely determined by q_t and θ_l at the low atmosphere level. The correlation coefficient of CAPE and CAPE* is 0.97, so we can use CAPE* as a proxy for CAPE. In general this is not the case, but free tropospheric properties change relatively slowly in the LES data. For CIN we do a linear fit of the logarithm of CIN. We write

$$\text{CIN}^* = e^{\mu X}. \tag{8.2}$$

Here $\mu = (\mu_0, \ldots \mu_4)$ are the coefficients for CIN*. For CIN and CIN* we find a correlation coefficient of 0.77, so we can use CIN* instead of CIN.

(b) *Large-scale tendencies* $\dot{X}$

In a GCM a parameterization should deliver entire vertical heating and moistening profiles. In our SCM experiment we only have four prognostic variables and therefore we use LES data to determine the influence of the cloud filling fractions σ on these four prognostic variables X. Below, we propose a method of using data to calculate the heating and moistening (i.e. the tendencies $\dot{X}$); whether this method will work for a large number of variables remains to be explored.

In [15] this was done for shallow cumulus convection by clustering vertical heat and moisture fluxes observed in LES data. Here we will use a least-squares fitting method that we already used to calculate the CAPE* and CIN*. Every cloud type has influence on θ_l and q_t at the low and higher atmosphere level. This means that

$$\dot{X}_m = \sum_{\alpha=0}^{4} \sigma_\alpha F_m^\alpha,$$

where $\dot{X}_m$ is the tendency of X_m $(1 \leq m \leq 4)$ and F_m^α is the influence of cloud type α on prognostic variable X_m. We assume that F_m^α is a linear combination of

the prognostic variables X:

$$F_m^\alpha = \sum_n \nu_{mn}^\alpha X_n.$$

We now have:

$$\dot{X}_m = \sigma \nu_m X, \tag{8.3}$$

where σ is the 1×5 filling fraction vector, ν_m is a 5×5-matrix that has to be estimated separately for every prognostic variable X_m, and X is the 5×1 prognostic variables vector. For every prognostic variable X_m we estimate ν_m by least-square fitting. This is done as follows. Our aim is to calculate for every $1 \leq m \leq 4$:

$$\min_{\nu_m} \sum_t (\dot{X}_m - \sigma \nu_m X)^2, \qquad 1 \leq t < 480. \tag{8.4}$$

In every subdomain of LES we observe the prognostic variables X, tendencies $\dot{X}_m$ and the LES-filling fractions σ. This is the case for 479 time instances (at the last time instance $t = 480$ the tendencies are not estimated). We can write (8.4) in the form $y = Z\nu$. Then, the least square fit gives $\hat{\nu} = Z^T y (Z^T Z)^{-1}$. This gives the best least square estimate of the 25 entries in the 5×5 matrix ν_m.

(*c*) *Integration of the single-column model*

We integrate Eq. 8.3, to obtain the evolution of the prognostic variables $X_1, \ldots, X_4$. As initial condition we take $\sigma = (1, 0, 0, 0, 0)$. This means that each CMC starts in state 1 (corresponding to clear sky). The initial conditions for X are the average initial values observed in the LES data. The CMCs produce the filling fractions σ and the ν are pre-calculated in Section b. We recall that the CMCs are conditioned on CAPE* and CIN*.

(*d*) *Filling fractions of the SCM*

We test the stochastic multicloud model in the SCM. In Fig. 11 we show filling fractions for SCM using 1024 CMCs. To increase the standard deviation we do a second experiment using only 64 CMCs. To calculate the standard deviation in every experiment we use 12^2 independent runs of the SCM. In this way we can compare the standard deviation to the standard deviation that we observed in the 12^2 LES-blocks (each consisting of 1024 LES-columns). Comparing Fig. 4a to Fig. 11, we see that the SCM-CMC is capable of reproducing the time-evolution of the filling fractions from the LES data . This is a remarkable result because the SCM is not using any LES data during the integration. Remind that the SCM has been constructed from LES data prior to integration.

Using a smaller number of Markov chains (64 instead of 1024) increases the variance of the filling fractions in the SCM test, as can be seen in Fig. 11b. We expect that further improvement of the evolution of the standard deviations in the SCM is possible by using the SCA or the CSCA instead of CMC, but we did not perform these experiments here.

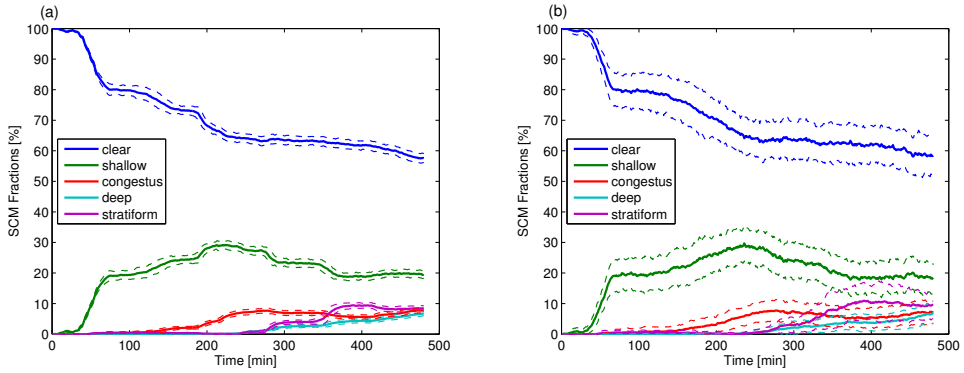

Figure 11. (a) Mean filling fractions produced in the SCM using 1024 CMCs conditioned on CAPE* and CIN* (solid) plus and minus the standard deviations (dashed) and (b) the same using 64 CMCs.

(e) *A ten day run of the SCM*

We have seen that the multicloud model produces correct filling fractions and that it can be used to enhance variability in the SCM. We integrate the SCM over a longer time period. Although the SCM-CMC has not been trained on a longer period, there are no practical restrictions on performing longer time integrations. As in [7] we integrate the SCM for ten days. Here, using the SCM, we do not aim to represent a realistic simulation of deep convection (as is the case for LES). Rather, we are interested in the long-term behaviour of the SCM as a dynamical system, seen as coarse extrapolation. We investigate whether or not the multicloud model can enhance variability in the SCM. In Fig. 12 we plot time series for the prognostic variable X_3 in the single column model integrated over ten days with a time step of one minute. The graphs for the other X_i are similar. For both runs, with 2500 CMCs and 64 CMCs, we see a cycle of around eight hours. This cycle is not caused by diurnal variations in the surface fluxes, because the CMCs have been trained on data from an LES run with fixed surface fluxes. We note that the trajectory depends strongly on the number of Markov chains used. With a large number of Markov chains, the system behaves very regularly. For smaller n, the multicloud model is more stochastic, and the SCM-CMC model displays more variability.

9. Discussion and conclusion

In this paper we combined, for the first time, the data-driven approach to stochastic parameterization from [14, 15] with the stochastic multicloud model approach proposed in [7]. We used data from a convection-resolving LES model to infer a multicloud model similar to the one studied in [7]. The aim was to formulate a stochastic model that was able to emulate the coarse-grained convective behaviour of the LES. Data for cloud top height and column rain fraction from the LES were used to determine five cloud types: clear sky, shallow

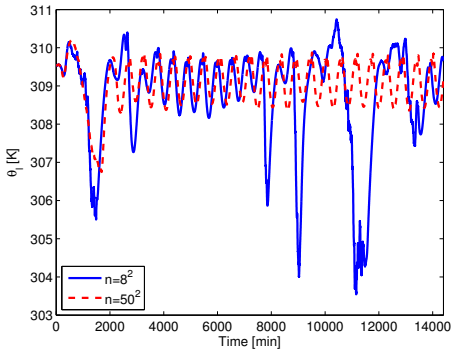

Figure 12. Time series for $X_4 = \theta_l^{\text{high}}$ in the single column model integrated over ten days.

cumulus, congestus, deep and stratiform. The coarse-grained convective behaviour of the LES was represented through the filling fractions, or area fractions, of the five cloud types on (horizontal) macroscopic blocks of 32^2 LES gridpoints.

The stochastic model (Markov chain) makes random transitions between cloud types at each gridpoint, in accordance with transitions probabilities that are estimated from the LES data. A straightforward Markov chain was not able to reproduce the correct evolution of the filling fractions corresponding to the five cloud types. Therefore, we explored two ways of improving the skills of the Markov chain. First, by conditioning the Markov chain on large-scale variables, obtaining a CMC. Second, by conditioning on the neighbouring cells, obtaining a SCA.

The CMC conditioned on a combination of CAPE and CIN was well capable of reproducing the time-evolution of the cloud fractions observed in the LES data. The standard deviations of the filling fractions were not very well reproduced by the CMCs. They were too small and not similar to the standard deviations observed in the LES data. The absence of direct spatial coupling between cloud types in neighbouring cells in the CMC made it difficult to capture the time-varying spatial patterns seen in the LES data. Therefore the enhanced variability due to these patterns could not be captured by the CMCs.

The average filling fractions of the SCA were not as good as the CMC average filling fractions. Nevertheless, the SCA showed a much better evolution of the standard deviation of the filling fractions. By including spatial coupling, spatial and temporal patterns emerged, resulting in more realistic variability. We showed that further improvement can be achieved by additional conditioning on the large-scale variables, however this comes at the cost of a more complicated model.

A point of discussion is that the CMCs in the multicloud model have been trained on LES data of rather specific idealized (atmospheric) conditions. Clearly, not all possible large-scale states were sampled in this dataset. Dividing the LES domain into subdomains, as was done here (as well as in [15]), enlarges the sample of large-scale states. The large-scale states are defined as subdomain averages, so that the variability between the subdomains helps to increase the sample variance. As already mentioned in Section 6, one can increase the sample variance even more by using data from multiple LES runs with different initial conditions.

We focussed on a setting in which shear in the horizontal plane and spatially varying terrain type have not been considered. In case of a unidirectional shear with varying strength, the transition probabilities of the SCA may have to depend on the neighbouring cells in an anisotropic way. The question how strong this sensitivity is, has not been addressed here. With varying terrain, a possible solution is conditioning on several types of terrain.

We showed how the LES data can be used to produce heating and moistening rates. We tested the multicloud model in a simple SCM experiment. Using the CMCs, the LES filling fractions were faithfully reproduced by the SCM. The degree to which the multicloud model was stochastic had a large influence on the variability of the SCM.

10. Acknowledgement

The authors are grateful to Pier Siebesma, Harm Jonker and Christian Jakob for stimulating discussions. This research was supported by the Division for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO). The visit of J.A.B. to CWI was financially supported through an NWO visitor travel grant. In addition, we acknowledge sponsoring by the National Computing Facilities Foundation (NCF) for the use of supercomputer facilities, with financial support of NWO. J.A.B. is supported by a grant from the National Science Foundation, DMS-1009959.

References

[1] Lin, J.-L. et al. 2006 Tropical Intraseasonal Variability in 14 IPCC AR4 Climate Models. Part I: Convective Signals. *J. Climate* **19**, 2665–2690. (doi:10.1175/JCLI3735.1)

[2] Wheeler, M. & Kiladis, G.N., 1999 Convectively coupled equatorial waves: analysis of clouds and temperature in the wavenumber-frequency domain. *J. Atmos. Sci.* **56**, 374–399.

[3] Zhang, C. 2005 Madden-Julian Oscillation. *Rev. Geophys.*, **43**, RG2003 (doi:10.1029/2004RG000158)

[4] Palmer, T.N. 2001 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. doi: 10.1002/qj.49712757202

[5] Shutts, G. J. 2005 A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteor. Soc.* **131**, 3079–3102. (doi:10.1256/qj.04.106)

[6] Berner, J., Doblas-Reyes, F.J., Palmer, T.N., Shutts, G., Weisheimer, A. 2008 Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal predicition skill of a global climate model. *Phil. Trans. R. Soc. A* **366**, 2561–2579 (doi: 10.1098/rsta.2008.0031)

[7] Khouider, B., Biello, J., Majda, A.J. 2010 A Stochastic Multicloud Model for Tropical Convection. *Comm. Math. Sci.* **8**, 187–216.

[8] Frenkel, Y, Majda, A.J., Khouider, B. 2012: Using the Stochastic Multicloud Model to Improve Tropical Convective Parameterization: A Paradigm Example. *J. Atmos. Sci.* **69**, 1080–1105. (doi: 10.1175/JAS-D-11-0148.1)

[9] Palmer, T. & Williams, P. 2010 *Stochastic physics and climate modelling*, Cambridge University Press, Cambridge, UK.

[10] Plant, R.S. & Craig, G.C. 2008 A Stochastic Parameterization for Deep Convection Based on Equilibrium Statistics. *J.Atmos. Sci.* **65**, 87–105. (doi:10.1175/2007JAS2263.1)

[11] Khouider, B. & Majda, A.J. 2006 A simple multicloud parametrization for convectively coupled tropical waves. Part I: linear analysis, *J. Atmos. Sci.* **63**, 1308–1323.

[12] Mapes, B.E., 1998 The large-scale part of tropical mesoscale convective system circulations: A linear vertical spectral band model. *J. Meteor. Soc. Japan* **76**, 29–55.

[13] Khouider, B., St-Cyr, A., Majda, A.J., Tribbia, J. 2011 The MJO and Convectively Coupled Waves in a Coarse-Resolution GCM with a Simple Multicloud Parameterization. *J. Atmos. Sci.* **68**, 240–264 (doi: 10.1175/2010JAS3443.1)

[14] Crommelin, D. & Vanden Eijnden, E. 2008 Subgrid-Scale Parametrization with Conditional Markov Chains. *J. Atmos. Sci.* **65**, 2661–2675. (doi:10.1175/2008JAS2566.1)

[15] Dorrestijn, J., Crommelin, D.T., Siebesma, A.P., Jonker, H.J.J. 2012 Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theor. Comput. Fluid Dyn.* (doi: 10.1007/s00162-012-0281-y)

[16] Wu, C.-M., Stevens, B., Arakawa, A., 2009 What Controls the Transition from Shallow to Deep Convection? *J. Atmos. Sci.* **66**, 1793–1806. (doi: 10.1175/2008JAS2945.1)

[17] Heus, T. et al. 2010 Formulation of the Dutch Atmospheric Large-Eddy Simulation (DALES) and overview of its applications, *Geosci. Model Dev.* **3**, 415-444. (doi:10.5194/gmd-3-415-2010)

[18] Böing, S.J., Jonker, H.J.J., Siebesma, A.P., Grabowski, W. 2012 Influence of the subcloud layer on the development of a deep convective ensemble. *J. Atmos. Sci.* (doi: 10.1175/JAS-D-11-0317.1)

[19] Mapes, B., Tulich, S., Lin, J., Zuidema, P. 2006 The mesoscale convection life cycle: Building block or prototype for large-scale tropical waves? *Dyn. Atmos. Oceans* **42**, 3–29. (doi:10.1016/j.dynatmoce.2006.03.003)

[20] Khouider, B., Majda, A.J., Katsoulakis, M. 2003 Coarse grained stochastic models for tropical convection. *Proc. Nat. Acad. Sci. USA* **100**, 11941–1194.

[21] Tompkins, A. M. 2001 Organization of Tropical Convection in Low Vertical Wind Shears: The Role of Cold Pools. *J. Atmos. Sci.* **58**, 1650–1672. (doi:10.1175/1520-0469(2001)058<1650:OOTCIL>2.0.CO;2)

[22] Riemann-Campe, K., Fraedricha, K., Lunkeit, F. 2009 Global climatology of Convective Available Potential Energy (CAPE) and Convective Inhibition (CIN) in ERA-40 reanalysis. *Atmos. Res.* **93** , 534–545. (doi:10.1016/j.atmosres.2008.09.037)

[23] Kwasniok, F. 2012 Data-based stochastic subgrid-scale parameterization: an approach using cluster-weighted modelling. *Phil. Trans. R. Soc. A* **370**, 1061–1086 (doi: 10.1098/rsta.2011.0384)

[24] MacQueen, J.B. 1967 Some Methods for classification and Analysis of Multivariate Observations. *Proc. of 5-th Berkeley Symp. on Math. Stat. and Probab.* **1**, 281–297.

[25] Gan, G., Ma, C., Wu, J. 2007 *Data Clustering: Theory, Algorithms, and Applications.* SIAM, Alex., VA.

[26] Arthur, D. & Vassilvitskii, S. 2007 k-means++: the Advantages of Careful Seeding. *Proc. of the 18-th Annu. ACMSIAM Symp. on Discret. Algorithms*, 1027–1035.

[27] Cover, T.M. & Thomas, J.A. 1991 *Elements of information theory*, 2nd edn., John Wiley & Sons, Inc., Hoboken, New Jersey.

[28] Meyer, T.P., Richards, F.C., Packard, N.H. 1989 Learning Algorithm for Modeling Complex Spatial Dynamics. *Phys. Rev. Lett.* **63**, 1735–1738. (doi:10.1103/PhysRevLett.63.1735)

[29] Bengtsson, L., Körnich, H., Källén, E., Svensson, E. 2011 Large-Scale Dynamical Response to Subgrid-Scale Organization Provided by Cellular Automata. *J. Atmos. Sci.* **68**, 3132–3144 (doi: 10.1175/JAS-D-10-05028.1)