Stochasticc convection parameterization

Dorrestijn, Jesse

DOI
10.4233/uuid:d80246c5-41dc-451d-9beb-c293c445a8f3

Publication date
2016

Document Version
Final published version

Citation (APA)
Dorrestijn, J. (2016). *Stochasticc convection parameterization*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:d80246c5-41dc-451d-9beb-c293c445a8f3

Stochastic Convection Parameterization - Jesse Dorrestijn

Stochastic Convection Parameterization

Jesse Dorrestijn

i n v i t a t i o n

You are cordially
invited
to the defence of the
dissertation

Stochastic Convection
Parameterization

by

Jesse Dorrestijn

Thursday
8 September 2016
at 14:30 hours

Senaatszaal
Auditorium
Mekelweg 5, Delft

Stochastic Convection Parameterization

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben;
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op donderdag 8 september 2016 om 15:00 uur

door

Jesse DORRESTIJN

Master of Science (MSc) in de Mathematische Wetenschappen,
Universiteit Utrecht,
geboren te 's-Hertogenbosch.

This dissertation has been approved by the
promotors: Prof. dr. A.P. Siebesma and Prof. dr. D.T. Crommelin.

Composition of the doctoral committee:

Rector Magnificus	chairman
Prof. dr. A.P. Siebesma	Delft University of Technology
Prof. dr. D.T. Crommelin	University of Amsterdam
Prof. dr. H.J.J. Jonker	Delft University of Technology.
Independent members:	
Prof. dr. F.H.J. Redig	Delft University of Technology
Prof. dr. C. Jakob	Monash University
Prof. dr. ir. J.E. Frank	Utrecht University
Prof. dr. A.A.M. Holtslag	Wageningen University
Prof. dr. ing. R. Klees	Delft University of Technology, reserve member.

Royal Netherlands
Meteorological Institute

This work is funded by the program *Feedbacks in the Climate System* of the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). The usage of supercomputer facilities is sponsored by the National Computing Facilities Foundation (NCF) with financial support of NWO.

Printed by: Ipskamp Printing, Enschede.
Cover design: Jesse Dorrestijn.
ISBN 978-94-028-0275-7

To my parents

Contents

Samenvatting (Summary in Dutch)

Wolken zijn chaotische, moeilijk te voorspellen, maar bovenal prachtige natuurverschijnselen. Er zijn verschillende soorten wolken: *stratus*, een dikke wolkenlaag waaruit het soms de hele dag miezert, *cirrus*, wolken die hoog in de atmosfeer te vinden zijn, en *cumulus*, stapelwolken die als bloemkolen de atmosfeer inschieten. De laatste variant duidt op convectie.

Een voorbeeld van convectie in de atmosfeer is thermiek, welbekend bij vogels en zweefvliegers die dankbaar gebruik maken van deze opwaartse luchtbeweging. Thermiek ontstaat als de zon het aardoppervlak verwarmt. Warme vochtige lucht stijgt in thermiekbellen naar boven. Warmte en vocht worden zo door convectie verticaal in de atmosfeer getransporteerd en verspreid. Convectie gaat vaak samen met wolkvorming en hevige regenval. Met name in de tropen zorgen cumuluswolken voor veel regen. Verder beïnvloeden convectie en wolkvorming de grootschalige windcirculatie op aarde. Ze hebben aldus een grote impact op de atmosfeer en daarmee op weer en klimaat op aarde.

Evenzo spelen deze processen een grote rol in simulaties van weer en klimaat. In globale circulatiemodellen worden grootschalige windstromingen en grootheden als temperatuur berekend op een driedimensionaal rooster dat gespannen is over de hele aarde. Kleinschalige processen, zoals convectie en wolkvorming, kunnen hiermee niet expliciet berekend worden. Deze moeten daarom worden geparametriseerd: er wordt een schatting gemaakt van het effect dat ze hebben op de grootschalige modelvariabelen. Voor een grofmazig rooster kan een dergelijke schatting statistisch worden gedaan, omdat het effect van een groot aantal realisaties van dezelfde kleinschalige processen goed uitmiddelt. Zo kan bijvoorbeeld het gezamenlijke effect van een groot aantal wolken in principe statistisch worden gerepresenteerd.

De zaak verandert doordat operationele weer- en klimaatmodellen met steeds fijnmazigere roosters werken. Met een fijnmaziger rooster kunnen stromingen in de atmosfeer nauwkeuriger berekend worden waardoor het voorspellend vermogen van deze modellen meestal verbetert. Er komt echter een moment waarop de modelroosters zo fijnmazig zijn dat er nog maar een paar wolken in een roostercel passen. Dan wordt het chaotische gedrag van wolkvorming een belangrijke factor en is het door de parametrisaties berekende effect niet meer representatief. De toename van variabiliteit en willekeur is een motivatie voor het introduceren van stochastiek in convectieparametrisaties voor modellen met een relatief fijnmazig rooster.

In dit proefschrift staat stochastische convectieparametrisatie centraal. Kansprocessen worden gebruikt in de parametrisaties van convectie en bijbehorende wolkvorming. Een meerwaarde ten opzichte van traditionele deterministische parametrisaties is dat stochastische parametrisaties fluctuaties rond het verwachte effect kunnen genereren. Stochastiek kan op meerdere manieren worden inge-

voerd. In dit proefschrift wordt gebruik gemaakt van Markovketens, kansprocessen die zijn vernoemd naar de bekende Russische wiskundige Andrei Markov (1856-1922). Deze kansprocessen hebben een eindig aantal toestanden, waarvan de overgangskansen geschat kunnen worden uit data. Door de overgangskansen te schatten met data van convectie, wordt het gedrag van convectie nagebootst.

Een Large-Eddy Simulation model is gebruikt om data te produceren, een model dat convectie en wolken zeer nauwkeurig simuleert. Met de data zijn Markovketens gemaakt die convectie en wolkvorming, zoals waargenomen in een meetcampagne nabij Barbados, nabootsen. Hetzelfde is gedaan voor wolkvorming in Brazilië. Voor een beperkt scala aan atmosferische omstandigheden werken deze Markovketens goed. Een ander Markovmodel is gemaakt met een grote dataset waarnemingen van een regenradar in Darwin in Australië. Deze Markovketens werken voor algemenere atmosferische omstandigheden. Ze zijn gebruikt voor het testen van stochastische convectieparametrisatie in een klimaatmodel. Dit heeft geleid tot verbeteringen in de variabiliteit van de gesimuleerde convectie en ook de verdeling van de gesimuleerde regen in de tropen is verbeterd. Helemaal perfect werkt het Markovmodel nog niet, maar er is wel een grote stap gezet in de ontwikkeling van deze stochastische methode voor convectieparametrisatie in weer- en klimaatmodellen.

Links: cumuluswolken in Amsterdam. Rechts: dezelfde soort wolken boven Duitsland gefotografeerd vanuit het vliegtuig. Foto's gemaakt door JD.

Summary

Clouds are chaotic, difficult to predict, but above all, magnificent natural phenomena. There are different types of clouds: *stratus*, a layer of clouds that may produce drizzle, *cirrus*, clouds in the higher parts of the atmosphere, and *cumulus*, clouds that arise in convective updrafts.

Thermals, rising air that is often used by birds and gliders to gain height, are an example of atmospheric convection. When the sun heats Earth's surface layer, warm and moist air rises in thermals to higher parts of the atmosphere. In this way, convection transports heat and moisture vertically in the atmosphere. This often leads to the formation of clouds and heavy rainfall. A major part of the rainfall on Earth, especially in the tropics, is produced by cumulus clouds. Furthermore, convection and cloud formation affect the large-scale planetary circulation. In the atmosphere, these processes are of major importance for Earth's weather and climate.

Convection and clouds also play a major role in numerical simulations of weather and climate. With general circulation models, the large-scale wind circulation and variables such as temperature and humidity are calculated on a three-dimensional global grid. The model grid resolution is low, and therefore, smaller-scale processes such as convection and cloud formation can not be calculated explicitly. The impact of these small-scale processes has to be determined in another way. They are represented by parameterizations that give an estimate of the effect of the small-scale processes on the large-scale model variables. For models with relatively large columns, the presence of a large number of realizations of the same small-scale process justifies the expression of their effect on the large-scale variables in terms of statistical properties. For example, the effect of a large number of clouds can be represented statistically.

A problem arises from the fact that the resolution of operational weather and climate models tends to increase. Generally speaking, with higher model resolutions the atmosphere can be simulated more accurately. However, if resolutions keep increasing, the expression of the small-scale effects in terms of statistical properties can no longer be justified. In a small model column, there is for example only space for a small number of clouds. The chaotic behavior of convective clouds becomes an important factor and deterministic parameterizations no longer give accurate estimates. The increase of fluctuations and randomness is a motivation for using stochastic convection parameterizations.

The central research theme in this dissertation is stochastic convection parameterization. Stochastic processes are used in the representation of convective clouds. Traditional deterministic parameterizations only give an estimate of the expected value of the effect of small-scale variables. Stochastic parameterizations can deviate from this expected value and can produce a range of convective responses. Especially in models with a relatively high resolution, it is important that parame-

terizations can represent fluctuations around the expected value. There are several ways of introducing stochastics. In this dissertation, Markov chains are examined, stochastic processes that are named after the famous Russian mathematician Andrei Markov (1856-1922). Markov chains have a finite number of states of which the transition probabilities can be estimated from data. By inferring transition probabilities from high-resolution data of convection, Markov chains mimic convective behavior.

A Large-Eddy Simulation model is used to construct a data set. Large-Eddy Simulation models are able to resolve clouds and convection in detail. After inference of the Markov chains, they are able to mimic clouds and convection as observed in a field-experiment near Barbados. The same method has also been applied for convective clouds in Brazil. These Markov chains only work for a very specific range of atmospheric circumstances. Therefore, another Markov chain model is constructed from a large observational data set from a rain radar in Darwin, Australia. A larger range of atmospheric circumstances is covered, and the Markov chains can be applied more generally. The Darwin Markov chains are implemented in a climate model to stochastically parameterize convection. This improves the variability related to convection as well as the distribution of the simulated tropical precipitation. The Markov-chain model is not perfect yet; however, a large step has been made in the development of this stochastic method for usage in state-of-the-art weather and climate models.

Left: Cumulus clouds in Amsterdam. Right: the same type of clouds in Italy photographed from an airplane. Photos by JD.

Chapter I
Introduction

1.1 Atmospheric convection

Everyone knows what a cloud is. We can see them floating in the sky with our own eyes. There are different types of clouds. We will focus on *convective clouds* that are related to the process of *atmospheric convection*; which is a less known natural phenomenon. To explain what convection is, it is easier to come down to Earth. You likely heard of *lava lamps* (Fig. 1.1). In these lamps, wax floats in a closed glass filled with a liquid. The glass is heated from below by a lamp. The temperature of the wax that is close to the bottom increases and as a result the wax starts rising. At the top of the glass it cools down and descends. Warm wax is lighter than cold wax: its density is lower which gives it positive buoyancy. This process is an example of convection which could be defined as 'buoyancy-driven turbulent flow'.

In the atmosphere, there are temperature differences as well, mainly because the sun warms the Earth's surface and the atmosphere loses heat by emitting in-

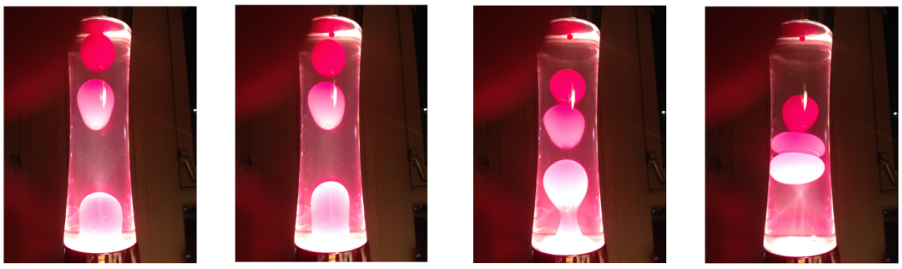

Figure 1.1: Convection in a lava lamp. If the wax is close to the bottom of the glass, a lamp heats it and as a result it starts rising. At the top of the glass it cools down and decends. Heat is transported from the bottom to the top of the glass.

frared radiation into space. Just compare the sun with the lamp that warms the bottom of the glass in the lava lamp. Sunlight penetrates the atmosphere quite easily and therefore the heating of the atmosphere by the sun is mainly done from the surface. Air *parcels* are heated and start rising in so-called *updrafts* or *thermals*, similar to the rising of the warm wax in the lava lamp. In this way, heat from the surface is transported vertically in the atmosphere. One of the main contributions of convection to Earth's atmospheric system is that it transports heat, moisture, momentum and various other physical quantities vertically in the atmosphere.

Convection is also visible in the atmosphere. Sometimes when the sun heats the Earth's surface, you can see that the air is trembling a bit. The warm air rises right into the colder air and the density differences cause refraction of the light. A far more easy way to discern convection in the atmosphere is by looking at … clouds! Convection can result in the formation of clouds. Rising air cools, because it is expanding. The air in the parcel contains water, but in the gas phase (water vapor), which can not be seen. If the temperature in the rising air parcel drops below the condensation temperature, the air becomes oversaturated, the moisture starts to condensate and a cloud appears.

These convective clouds are called *shallow convective cumulus clouds* if they are of limited vertical extent and *deep convective cumulus clouds* in case they are larger and produce rain. In the Dutch and English summaries of this thesis you can see pictures of shallow convective cumulus clouds and deep convective cumulus clouds, respectively. Let us summarize the types of convection in the atmosphere: a distinction is made between *dry* and *moist* convection, and the latter can further be divided into shallow and deep convection.

The role of convection in Earth's atmosphere and climate

Convection plays a major role in Earth's atmosphere and climate [3, 137]. Locally, it *stabilizes* the atmosphere by vertical *transport* of heat and moisture. The atmosphere is unstable when layers of relatively warm less-dense air are below layers of

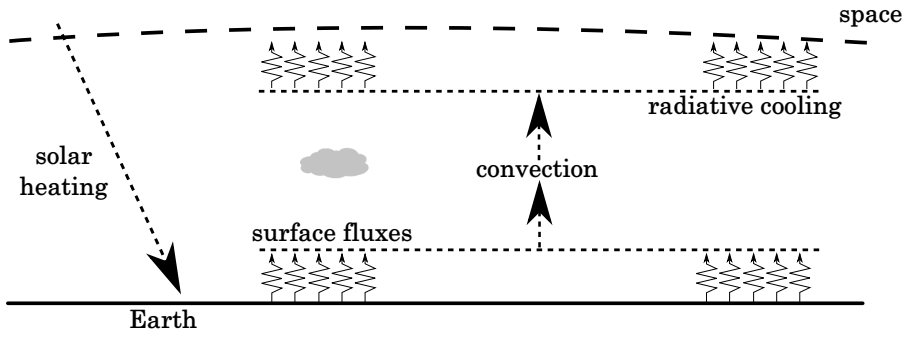

Figure 1.2: Schematic illustration: convection counteracts the destabilizing large-scale forcing. For example, convection transports the excess of heat at the surface, caused by solar heating, to higher levels in the atmosphere where it balances radiative cooling.

colder denser air, which can be the result of surface heating by the sun or radiative cooling (the emission of infrared radiation into space), which are two examples of *large-scale forcings*. The large-scale forcing is defined as 'the destabilizing effects of large-scale processes' [5] and the processes can be referred to as large-scale forcings. Convection counteracts the large-scale forcing (Fig. 1.2): convection tends to stabilize the atmosphere by redistributing heat and moisture, thereby removing *instabilities*. If the large-scale forcing continuously destabilizes the atmosphere, a balance is formed between forcing and convection. We can schematically express this as follows:

$$\frac{\partial\overline{\phi}}{\partial t} = \frac{\partial\overline{\phi}}{\partial t}_{\text{convection}} + \frac{\partial\overline{\phi}}{\partial t}_{\text{forcing}}, \qquad ||\frac{\partial\overline{\phi}}{\partial t}|| \ll ||\frac{\partial\overline{\phi}}{\partial t}_{\text{forcing}}||,$$

where ϕ can be temperature or moisture and $\overline{\phi}$ is the horizontal average of the variable over a large area of the order of 100^2km^2. This means that the atmospheric circumstances over a large area are changing at a much slower rate than that convection is counteracting the large-scale forcing. This possibly slowly changing balance is called *quasi-equilibrium*.

The appearance of clouds in convection makes convection a process of even more importance. Convective clouds affect the *large-scale planetary circulation* [19]. Shallow cumulus clouds are abundant in the trade wind region and the moisture that they transport to higher atmospheric levels is advected further by the trade winds towards the equator. There it works as an extra supply of moisture in the *Intertropical Convergence Zone* (ITCZ) (Fig. 1.3). Shallow cumulus clouds supply the tropical atmosphere with moisture, which facilitates the formation of deep convection. In the ITCZ, air rises as part of the Hadley circulation and deep convection intensifies this upward motion of air by latent heat release. We see that shallow and deep convection intensify the Hadley circulation [128].

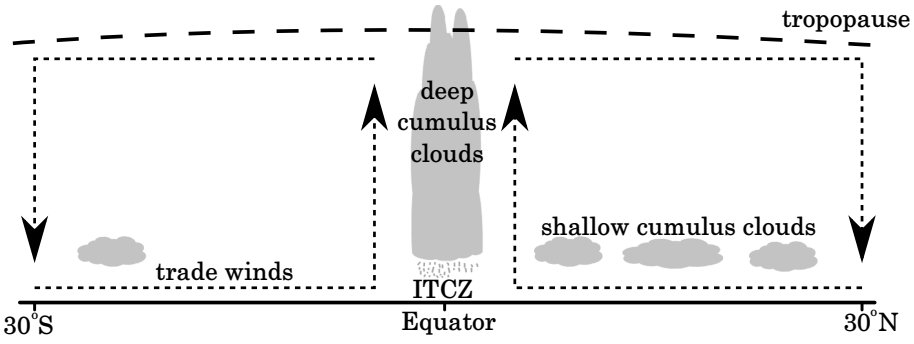

Figure 1.3: Schematic depiction of the north and south cell of the Hadley circulation. At the equator, in the Intertropical Convergence Zone (ITCZ), deep convective cumulus clouds intensify the strong upward motion. At a height of around 15 km in the atmosphere winds blow northward and southward and descend slowly at 30°N and 30°S. In these regions with subsiding air, shallow cumulus clouds form and are advected by the trade winds in the direction of the equator. There they act as an extra moisture supply for deep convection.

Besides stabilizing the atmosphere, transporting heat, moisture and momentum and affecting the large-scale circulation, deep convection also largely determines *precipitation*. Moreover, deep convection is related to spatially organized large structures of deep convective events that are called *convectively coupled equatorial waves* [76, 85, 143], that only occur around the equator. The structures are called waves, but don't look like waves as you know from the beach, because they are much larger with a wavelength of the order of $1,000 - 10,000$ km. The convectively coupled equatorial waves determine in part the variability of precipitation around the equator. Finally, convective clouds affect the *planetary energy budget*, as all clouds do. Clouds reflect sunlight back into space and they absorb and emit infrared radiation.

Now that we have some idea of what convection is and what its role is in Earth's atmospheric system, we will have a closer look at shallow and deep convection.

Shallow convection

Shallow cumulus convection is most common in the subtropics, as explained especially in the trade-wind region above the ocean, but it is also frequently observed in the tropics and in the mid-latitudes above land and sea. A classic field experiment with shallow cumulus convection over sea in the trade-wind region is the Barbados Oceanographic and Meteorological Experiment (BOMEX) carried out in 1969 [58]. To give an indication of the vertical extent of the clouds: a typical cloud base height was found around $500 - 600$ m and the cloud top around $1,500 - 1,600$ m.

Since large-scale forcings are typically homogeneous over large areas of the ocean in the trade-wind region and constant over long time periods, shallow cumulus clouds appear in large *cumulus ensembles*, i.e., a large number of cumulus clouds spread over the area. The ensemble continuously counteracts the large-scale forcings, forming a quasi-equilibrium. Note that individual cumulus clouds in the ensemble can be at different stages of their life cycle, e.g., some of the clouds may just have arisen and others may already be dissolving. On average, however, a nearly constant number of cumulus clouds is distributed randomly over the area.

Let us look in more detail at the vertical structure of the atmosphere in the case of shallow cumulus convection. The easiest way to do this, is by comparing the *virtual potential temperature* of an ascending parcel $\theta_{v,p}$ with the virtual potential temperature of the environmental air $\overline{\theta_v}$ through which it penetrates. The virtual potential temperature is defined as:

$$\theta_v := \theta(1 + 0.61 q_v - q_l),$$

with q_v the water vapor specific humidity [126], q_l the liquid water specific humidity [126], and θ the *potential temperature* [126, 135]. The difference between θ_v of the parcel and the environment is proportional to their density difference. Therefore it can be used as a measure of buoyancy. In Fig. 1.4, we see a schematic depiction of θ_v of a rising parcel and its environment. The height above Earth's surface is on the vertical axis and θ_v on the horizontal axis. The layer near the surface is unstable, which causes air parcels to rise upward. The virtual potential

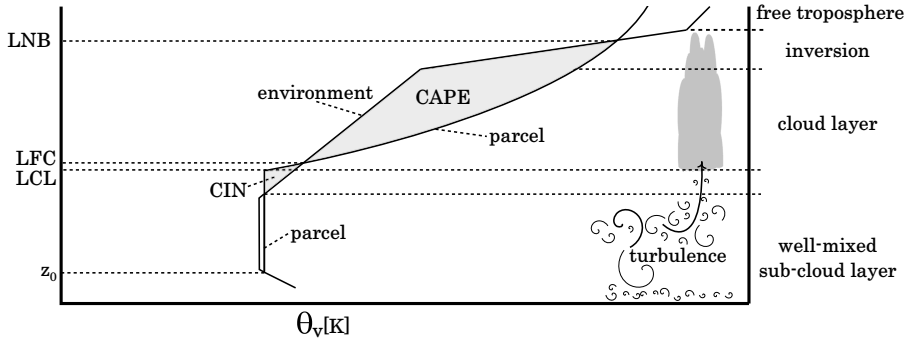

Figure 1.4: Schematic depiction of the virtual potential temperature θ_v of an entraining air parcel and the environment. Several important height levels are indicated on the left vertical axis and several layers on the right vertical axis. CIN is the area below the LFC that is enclosed by the two curves and CAPE the area between the LFC and the LNB (after Siebesma (1998) [126]).

temperature of these parcels is higher than of their environment, resulting in a positive buoyancy force:

$$B = g \cdot \frac{\theta_{v,p} - \overline{\theta_v}}{\overline{\theta_v}},$$

with $g = 9.81$ m s^{-2}, the gravitational acceleration constant. The parcels will rise and may reach the level where water vapor starts to condensate, depending on their strength and on the vertical extent of the *well-mixed boundary layer* (Fig. 1.4).

The level at which water vapor condensates is called the lifting condensation level (LCL). Moist convection is a more complex process than dry convection, mainly because phase changes of water are accompanied with energy release or energy costs that change the buoyancy of the rising air parcels. Latent heat release in a saturated updraft generates extra buoyancy, because the temperature increases. This enables saturated updrafts to reach much higher levels in the atmosphere than unsaturated updrafts. In effect, it works as an extra engine for reaching higher in the atmosphere. However, an updraft coming from the surface first has to penetrate through a layer in which it experiences negative buoyancy.

Only strong moist updrafts can penetrate through the layer of negative buoyancy and reach higher levels in which they experience positive buoyancy, solely due to latent heat release. The level above which a parcel experiences positive buoyancy is called the level of free convection (LFC). If we look in Fig. 1.4, we see that the level is located where $\theta_{v,p} = \overline{\theta_v}$. Strong updrafts that reach this level can potentially ascend up to much higher stable layers in the atmosphere.

A measure for the strength of the negative buoyancy below the LFC is the *convective inhibition* (CIN) and is defined by:

$$\mathrm{CIN} := -\int_{z_0}^{\mathrm{LFC}} B^{-} dz,$$

where $B^- = \min(B, 0)$. A large CIN value indicates that it is difficult to form convective clouds. In Fig. 1.4 it is the area between the two curves (θ_v of updraft and environment) just below the LFC. Updrafts that come above the LFC experience positive buoyancy and they can rise all the way to the level of neutral buoyancy (LNB) which is the level higher in the atmosphere where the buoyancy becomes zero again. A measure for the positive buoyancy above the LCL is the *convective available potential energy* (CAPE) and can be defined as:

$$\text{CAPE} := \int_{\text{LFC}}^{\text{LNB}} B dz.$$

Note that also other vertical levels can be chosen, for example a level close to the surface z_0 instead of the LFC. If the atmosphere is unstable, because it is too warm and moist in lower levels, CAPE will have a large value; while in case the atmosphere is stable, CAPE will have a small value or be equal to zero. This explains why CAPE is generally accepted as an *indicator* (i.e., predictor) of convection.

Since only strong moist updrafts can penetrate the layer of negative buoyancy, the result of this *selection* system is that moist convection is far more *intermittent* and *random* in character than dry convection in the boundary layer below the LCL. In the boundary layer, mixing by convection and turbulent eddies can be roughly seen as a very effective diffusion process with a corresponding eddy-diffusion coefficient that is orders of magnitude larger than the molecular diffusion coefficient [59]. The layer below the LCL is called the well-mixed subcloud layer (Fig. 1.4). In this layer, heat and moisture are horizontally and vertically well mixed: they are distributed such that the potential temperature is constant. Without a well-mixed subcloud layer reaching the LCL, it is difficult for updrafts to reach the LCL. Above the LCL, strong updrafts arrive in sudden bursts at intermittent rates and shallow cumulus clouds form at the LCL and rise up to the LNB. The layer in between these two levels is therefore called *the cloud layer* (Fig. 1.4). Generally, the clouds do not exceed this layer because it is capped by an inversion layer, in which the convective updrafts loose buoyancy, and therefore kinetic energy, quickly. Because shallow cumulus clouds are of limited vertical extent, precipitation effects are usually negligible for shallow convection. If updrafts are strong enough, they will reach even higher levels, and deep convective clouds will form. Deep convection is the topic of the next subsection.

A comprehensive introduction to shallow convection is found in Siebesma (1998) [126], in which all relevant terms and concepts, loosely mentioned in the present introduction (e.g., instability, several temperature definitions) are well defined and explained. Also, the important concepts of *entrainment* and *detrainment*, mixing of environmental air with cloud air, are discussed in detail. More information about detrainment in shallow cumulus can be found in [31, 129] and more about entrainment in deep convective clouds in [95, 122].

Deep convection

Precipitation can usually be neglected for shallow cumulus convection. This is not the case for *deep convection*, because its clouds are much larger, which enables the

formation of precipitation. Deep convection is characterized by heavily precipitating cumulus clouds, often accompanied with lightning, reaching very high levels in the atmosphere, sometimes up to 20 km. Deep convection is very common in the tropical belt, which is the reason that it is sometimes referred to as *tropical convection*. It is also common outside the tropics, for example at the end of a hot period in summer in the mid-latitudes. Deep convection is a more complex process than shallow convection. Deep convective updrafts are strong enough to penetrate the inversion layer that caps the shallow cumulus clouds (Fig. 1.5). As a result, their vertical extent is large enough to allow for the formation of precipitation. When precipitation takes place, liquid water falls from higher levels to lower levels of the atmosphere. It is possible that the liquid water falls through a warmer layer and evaporates before it reaches the surface, which cools the layer. When updrafts reach the *freezing level* at around 5 km (in the tropics), ice is formed and energy is released. This means that deep convective updrafts have a (modest) second buoyancy engine available, in addition to the latent heat release at the LCL. The amount of energy is, however, much less than the latent heat release at the LCL.

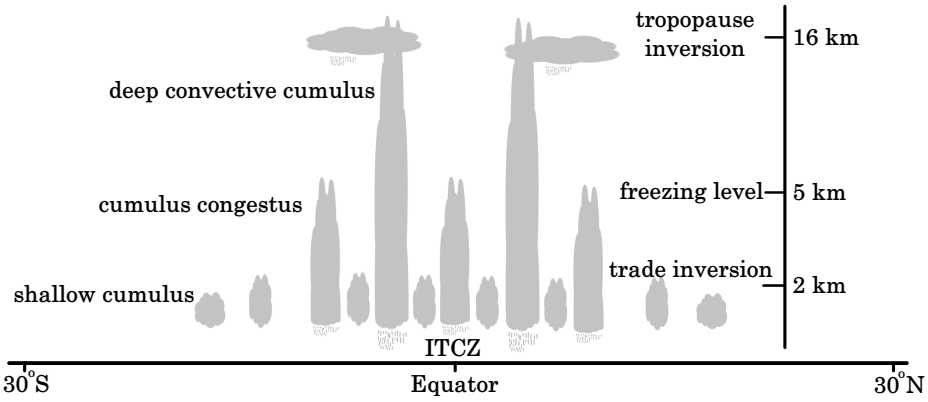

Figure 1.5: Schematic illustration of the typical location and vertical extent of shallow, congestus and deep convective cumulus clouds. Shallow cumulus clouds are capped by the trade inversion, congestus clouds do not reach higher than the freezing level and deep convective cumulus clouds reach the tropopause inversion (after Johnson et al. (1999)[65]).

The atmospheric conditions under which the initiation of deep convection is possible, are still debated. A necessary condition is a large CAPE value, because the atmosphere needs to be unstable in order for deep convection to occur. The role that shallow convection and cumulus congestus clouds (precipitating convective clouds that are not strong enough to go through the freezing level) play in the formation of deep convection is still not entirely understood. Johnson et al. (1999) [65] argue that by moistening the lower troposphere, thereby *preconditioning*, the shallow cumulus and congestus clouds enable the formation of larger deep convective cumulus towers. The importance of this preconditioning by the congestus clouds is debated by Hohenegger and Stevens (2013)[57]. They argue that large-

scale *moisture convergence* is more important for the formation of deep convection. Furthermore, organization of clouds also plays an important role: deep convective clouds help the formation of other deep convective clouds, because deep convective clouds are triggered at points where spreading *cold pools* originating from different convective clouds meet [17, 138]. In case deep convection occurs, the deep convective updrafts reach the strong tropopause inversion and spread horizontally and form a modestly raining *stratiform* anvil. Stratiform decks dissolve slowly due to precipitation and mixing with the environmental air.

As is the case for shallow cumulus clouds, deep convective cumulus clouds often appear in cumulus ensembles. Deep convective clouds are much larger than shallow cumulus clouds, both in vertical and horizontal extent. Updrafts are so strong that deep convection causes horizontal convergence of air. On the other hand, convergence or moisture convergence also supports the formation of deep convection [53, 57]. We conclude that convergence and deep convection form a *positive feedback system*.

We have seen that convection and clouds play several important roles in Earth's atmosphere and climate. Further, we have seen that convection and cloud formation are complicated processes that take place at several length scales and affect the atmosphere and climate in several ways over a large range of space and time scales. Because the prediction of intermittent and randomly occurring moist convection is complex, clouds and convection are a large source of uncertainty in weather and climate prediction models [117]. It is for example difficult to predict how clouds will respond to a warming climate and climate models do not show agreement [18, 19, 22, 69]. In order to make reliable weather and climate predictions, moist convection should be accurately represented in *general circulation models* (GCMs), used in numerical weather and climate models. The next section explains what a GCM is, how moist convection is currently represented in GCMs, and what the shortcomings of these representations are.

1.2 Parameterization

GCMs simulate Earth's entire atmosphere. Vilhelm Bjerknes [15] was the first to propose that the weather can be predicted by solving equations. Given the initial conditions, boundary conditions and external forcings it is, in theory, possible to calculate the time evolution of temperature T, pressure p, wind velocity in three directions (u,v,w), air density ρ and humidity q_t (defined below). The most important equations that apply to the movement of any incompressible fluid are the incompressible *Navier-Stokes equations*, which are the fluid-equivalents of Newton's second law (relating force, mass and acceleration) combined with the conservation of mass:

$$\frac{Du_i}{Dt} = -\frac{1}{\rho_0}\frac{\partial p}{\partial x_i} + \nu\Delta u_i + F_i, \qquad i \in \{1,2,3\}, \tag{1.1}$$

$$\mathrm{div}(\mathbf{u}) = 0,$$

with $u_i \in \{u,v,w\}$, the material or total derivative $D/Dt = \partial/\partial t + u\cdot\partial/\partial u + v\cdot\partial/\partial v + w\cdot\partial/\partial w$, the divergence div, the kinematic viscosity ν and $\Delta = \sum_{i=1}^{3}\frac{\partial^2}{\partial x_i^2}$ [21], and ρ_0

a reference density. Any force that is acting on a fluid parcel results in an acceleration if it is not counteracted by another force. These forces are for example the pressure force which is the first term on the right-hand side, Earth's gravitational force, the third term on the right-hand side for the vertical velocity equation with $F_3 = -g$ and, as a result of Earth's rotation, the Coriolis force, also the third term, but with $F_1 = cv, F_2 = -cu$, in which c is the Coriolis parameter.

In practice, it is impossible to resolve the full Navier-Stokes equations for a domain as large as Earth's entire atmosphere. Therefore in atmospheric GCMs, which are at the core of global numerical weather and climate prediction models, the simpler *primitive equations* are used instead [27]. These simplifications are obtained by using scale analysis of all terms in the equations for Earth's atmosphere [27]. In addition, instead of using the 'normal' temperature T and the humidity, the *liquid water potential temperature* θ_l and *the total water specific humidity* q_t can be used, because they are conserved for moist adiabatic processes in the absence of precipitation. They are defined as:

$$\theta_l = \theta - \frac{L}{c_p \pi} q_l, \tag{1.2}$$

$$q_t = q_v + q_l \tag{1.3}$$

where L is the latent heat of vaporization, c_p the specific heat of dry air at constant pressure, and π the Exner function: the ratio of absolute and potential temperature.

Furthermore, the variables in the equations are Reynolds averaged, i.e., decomposed in a mean part and a deviation part:

$$\phi = \overline{\phi} + \phi',$$

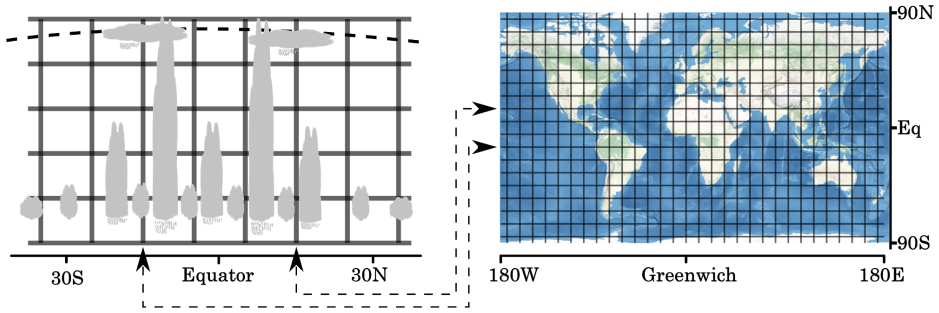

Figure 1.6: Schematic illustration of a global three-dimensional grid of a GCM. The left panel displays a part of a meridional cross-section of the grid. We see seven *vertical columns* and each column contains six *z-levels*. The right panel displays the grid on a map of Earth. The arrows indicate grid points located at latitudes 15°N and 15°S in both panels. Model variables are only resolved in grid points and, therefore, grid point values represent variables in the entire grid box. In each vertical column the most important *subgrid processes* have to be represented by *parameterizations*. Note that state-of-the-art GCM grids can be much finer than the grid in this illustration.

where $\overline{\phi}$ is again the horizontal average of a generic quantity ϕ over a large area of the order of 10^2km^2-100^2km^2, and ϕ' is a deviation from this average, see [12, 27] and [110] for similar treatments. To give an example, after Reynolds averaging of Eq. (1.1), the momentum equation for the zonal velocity, becomes:

$$\frac{\bar{D}\overline{u}}{Dt} = -\frac{1}{\rho_0}\frac{\partial \overline{p}}{\partial x} + \nu\Delta\overline{u} + \overline{F_i} - \Big(\frac{\partial \overline{u'u'}}{\partial x} + \frac{\partial \overline{v'u'}}{\partial y} + \frac{\partial \overline{w'u'}}{\partial z}\Big), \tag{1.4}$$

in which the last three terms are the Reynolds stresses, effects of small-scale processes that are smaller than the area size over which has been Reynolds averaged.

In GCMs, equations are discretized for computation on a three-dimensional global grid (Fig. 1.6). The left panel of Fig. 1.6 displays a part of a meridional cross-section of a grid. The right panel of Fig. 1.6 displays a GCM grid from a 'top view'. State-of-the-art GCMs often work with finer grids than displayed in this illustration; for example, the EC-Earth model's spectral resolution corresponds to a horizontal grid spacing of 1.125° [55]. Model equations are truncated at the grid size and variables $\overline{\phi}$ are spatial representatives of the grid box. This truncation leads to subgrid processes of which the most important effects have to be represented by *parameterizations*. For example, for $\phi = \theta_l$, the subgrid term with the largest effect is the z-derivative of the turbulent vertical heat flux, the first term on the right-hand side of the equation for $\overline{\theta_l}$:

$$\frac{\bar{D}\overline{\theta_l}}{Dt} = -\frac{\partial \overline{w'\theta_l'}}{\partial z} + F_{LS},$$

with F_{LS} heating due to large-scale forcings. Representing subgrid processes simply and adequately in GCMs is a difficult topic and parameterizations (especially of clouds and convection) cause large uncertainties and errors in numerical weather and climate model predictions.

Parameterization of subgrid processes

The distance between the grid points determines the GCM's resolution. In case the model solves the dynamical equations with a spectral method, the number of spectral modes determines the resolution, but for the calculation of the parameterized physical effects, the variables are transformed to physical space on a grid, and therefore the distance between the grid points can still be seen as the model's resolution. A vertical grid column represents a geographical region in which it is located, a region of which the size depends on the model resolution.

Variables such as temperature can be seen as horizontal averages over the region. Processes that are of a scale much smaller than the grid size can not be explicitly resolved by the model. These subgrid processes, despite their small size, can have major effects on the resolved variables. For example, convection transports large quantities of heat and moisture vertically in the atmosphere, even in case the horizontally averaged vertical velocity is zero. Therefore, for correct calculation of the resolved variables, the most important subgrid processes need to be represented in some accurate yet simple way. In GCMs they are represented by parameterizations, which are functions of the resolved variables. For example,

the turbulent heat flux is expressed in terms of the resolved-scale variables by a function f:

$$\overline{w'\theta_l'} = f(\overline{u},\overline{v},\overline{w},\overline{\theta_l},\overline{q_t},F_{LS}).$$

Parameterizations should be simple, since they should not cost too much computationally. In GCMs, the variables that are resolved on the grid are called *prognostic variables* or *large-scale variables* (e.g., $\overline{u},\overline{\theta_l}$), and the variables that are representing processes of a scale much smaller than the grid size are called *subgrid variables* or, in case one refers to the corresponding terms in the governing equations, *subgrid terms*, e.g., the z-derivative of the turbulent heat and moisture fluxes and Reynolds stresses.

Multi-scale modeling

The problem of parameterization of subgrid processes in weather and climate prediction models can be placed in a more general mathematical context: the atmospheric flow consists of many scales of motion [89], and hence, global simulation of the atmospheric flow is a typical *multi-scale problem*. The main objective is resolving the large-scale flow, with for example an efficient macroscopic model that does not resolve processes at the microscopic scale. At the same time, however, the microscopic processes partly determine the large-scale flow. The macroscopic model is efficient, but in order to be accurate it should incorporate the effects of the microscopic-scale processes. This is for example solved by assuming a separation of the two scales (the macro and micro scale), such that the microscopic effects can be obtained by a micro-scale model assuming a fixed large-scale state. Multi-scale modeling focuses on linking micro- and macro-scale models. Multi-scale problems are not only common in atmospheric sciences, but occur in many fields, e.g., computational chemistry and physics, biological systems, mathematics, material science [38, 49, 67, 70, 111]. This means that the parameterization approach that is examined in this dissertation, could also be applied in various other fields. In this dissertation the focus is on parameterization of moist convection, which is the topic of the next subsection.

Parameterization of moist convection

In order to resolve moist convection, a horizontal grid resolution of $10-100$ m is needed, while GCMs operate with resolutions of the order of $10-100$ km. Therefore, moist convection is a subgrid process and has to be represented by parameterizations. For every grid column, the trigger function of the model's convection scheme determines if there is convection present; and if so, it also determines the type of convection: dry, shallow or deep. This is done with a simple cloud parcel model [64]. A virtual air parcel is released from the surface with small excesses in temperature and moisture, as illustrated in Fig. 1.4. If there is moist convection present, the cloud parcel model determines the cloud base, the cloud top and in-cloud variables such as temperature, moisture, and the liquid water content. Almost all convection schemes are mass flux based, they determine the vertical profile of the mass flux in the atmosphere. This profile is a function of height and can be used to calculate subgrid fluxes (of for example heat and moisture) that are

necessary to evolve the prognostic variables of the GCM, i.e., calculate the time-derivatives of the prognostic variables. The turbulent heat and moisture flux are calculated with the following expression:

$$\overline{w'\phi'} = M(\phi_u - \overline{\phi}), \qquad \phi \in \{\theta_l, q_t\},$$

in which the mass flux $M = \rho a_u w_u$, with a_u the convective area fraction multiplied by w_u, the vertical velocity in the updraft [126], and ϕ_u is ϕ in the updraft.

The usage of mass-flux based parameterizations of convection in GCMs relies on the assumption that in a grid column a cumulus ensemble and the large-scale forcing are in quasi-equilibrium [5]. This means that the cumulus ensemble responds quickly to large-scale cooling and drying. Since the large-scale forcing is typically not entirely constant in time, the equilibrium can also change slowly in time. The quasi-equilibrium ensures that the heat and moisture transport of the cumulus ensemble can be expressed in terms of the prognostic model variables. To do so, one switches from the convective properties of a single cloud to the statistical properties of the ensemble, for which for example the cloud area fraction at cloud base is more important than cloud life cycles of individual clouds.

As mentioned, in the dry convective boundary layer, parameterizations based on eddy-diffusivity are appropriate (although this is already a rough approximation), while for shallow cumulus convection parameterizations based on mass flux are more appropriate. In the eddy-diffusivity mass flux (EDMF) approach [112, 130, 132] these two types are combined:

$$\overline{w'\phi'} = -K\frac{\partial \overline{\phi}}{\partial z} + M(\phi_u - \overline{\phi}), \qquad \phi \in \{\theta_l, q_t\},$$

in which K is the eddy-diffusivity. This scheme has been improved with the dual mass-flux closure for shallow and dry convection by Neggers et al. (2009) [104]. The updraft θ_l and q_t values are found with a cloud parcel model. The mass flux vertical profile can be calculated only if the mass flux at cloud base is determined by a closure. There are several mass flux closures [105], several indicators (e.g., CAPE, CIN) can be used [29], and as we will see in Chapter 4, there is not yet a general consensus on the best closure.

Parameterizations of convection can have major impact on model results, e.g., the location of the ITCZ [100], since the process has a large effect on the atmosphere and climate. We will see in Chapter 5 that convection parameterizations can have major impacts on GCM climate values, e.g., precipitation. Errors made in convection parameterizations are reflected in biases and uncertainties in the simulation of the atmosphere, and therefore, also in weather and climate predictions. Shortcomings in convection parameterizations can arise from for example:

- inadequate closures of the mass flux at cloud base; Should the closure be based on dynamical (e.g., convergence) or thermodynamical variables (e.g., CAPE, CIN)? (Discussed further in Chapters 3, 4);
- classification of only three types of convection: dry, shallow, deep, while congestus clouds may deserve their own treatment [73], (Chapters 3, 4);

- the occurrence of multiple types of convection in the same model column at the same time is often not possible [39]. In reality, deep convective clouds can be surrounded by shallow convective clouds (Chapter 3);
- the trigger function: sometimes the trigger function tends to switch on and off too rapidly and destroys smoothly decaying convection (Chapter 5);
- no unified treatment of convection and clouds [4]. For example, cloud cover should be related to convective area fractions (Chaper 4);
- no scale-adaptive/aware convection parameterizations, but dependent on a fixed coarse resolution (Chapters 4, 5);
- not enough subgrid-scale variability associated with convection (Chapters 2, 3, 4, 5);
- the usage of deterministic parameterizations instead of stochastic parameterizations. By using deterministic functions to determine the effect of convection on the resolved model variables, random fluctuations around the expected values can not be captured. Stochastic parameterizations do allow correct representation of variability (topic of Section 1.3).
- no direct coupling to neighboring model columns (spatial dependencies) (Chapters 3, 4, 5);

In convection parameterizations there are many parameters of which the values have to be estimated, either from observations or from high-resolution computer simulations, before implementation in a GCM. One of these high-resolution computer simulations is *Large-Eddy Simulation* (LES), which is described in the next subsection and which we will use in Chapters 2 and 3 to construct convection parameterizations.

Large-Eddy Simulation

To examine clouds and convection one can make use of LES (Fig. 1.7). On a domain with a horizontal size of the order of $10-1{,}000$ km, depending on computer capacity, the evolution of the three-dimensional flow in the atmosphere is calculated on a grid with a resolution of $10-100$ m. This resolution is high enough to resolve convection. The spatially filtered Navier-Stokes equations are solved using the *Boussinesq* approximation [27, 133] for shallow convection or the *anelastic* approximation [131] for deep convection. In grid points in which the air is oversaturated, liquid water is present and this will be the case if there is a cloud. With an integration time step of a few seconds, the clouds are simulated from formation to the point when they dissolve. Researchers are able to test assumptions about clouds, convection and turbulence in this virtual laboratory [103, 129, 140, 148]. Depending on the predetermined initial state of the atmosphere, the boundary conditions and the large-scale forcings, different types of clouds (e.g., shallow or deep) can be examined. These predetermined conditions are typically obtained from field experiments, for example from the aforementioned BOMEX. The model can also be

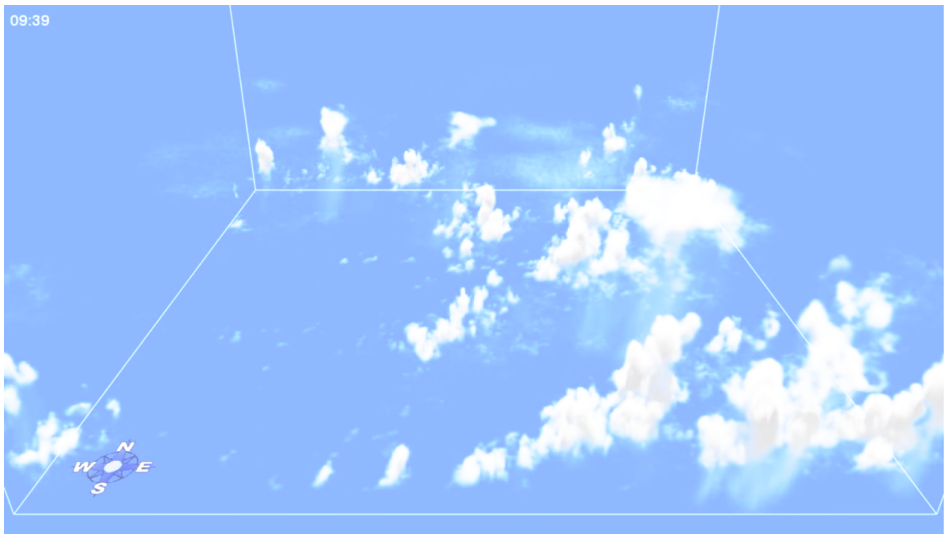

Figure 1.7: A snapshot from a three-dimensional simulation of deep convection using a Large-Eddy Simulation model (credits: J. Schalkwijk).

validated by comparison with data of such field experiments. Several LES models have been developed at different universities and research institutes, with comparable results [127]. The LES model that is used for the research described in this thesis is called DALES (Dutch Atmospheric LES), developed at KNMI, University Wageningen and Delft University of Technology, described in detail by [56]. A comprehensive introduction to the general aspects of LES models (e.g., equations, subgrid-scale filters, boundaries) is given by [12]. Running an LES model is computationally expensive and this limits the capacity of the model in terms of domain size and simulation time period.

We now have some basic notion of moist convection, its important role in Earth's atmosphere and climate, and its representation in GCMs. These representations have shortcomings which can lead to model errors and uncertainties. In the next section we will see that the usage of *stochastics* may improve representations of convection.

1.3 Stochastics

Traditionally, GCMs are deterministic models. In deterministic models, each tendency of each model variable ϕ is a deterministic function of the model's prognostic variables and large-scale forcings $\mathbf{x} = \{u, v, w, \theta_l, q_t, \ldots\}$:

$$\frac{\partial \phi}{\partial t} = f(\mathbf{x}). \tag{1.5}$$

The initial state $\mathbf{x}_0$ determines the future time state $\mathbf{x}_T$ in a deterministic way (Fig. 1.8). In stochastic models, random numbers affect the model variable tendencies:

$$\frac{\partial \phi}{\partial t} = f(\mathbf{x}, \alpha), \tag{1.6}$$

in which $\alpha(x, y, z, t)$ is a stochastic process [66, 145], producing *random numbers* that depend on time t and grid point location (x, y, z). In the schematic illustration in the right panel of Fig. 1.8, we see that model variable trajectories of ϕ are not uniquely determined by the initial state $\mathbf{x}_0$, instead several trajectories are possible and consequently several outcomes for the future time state $\mathbf{x}_T$. You may see similarities between the right panel of Fig. 1.8 and ensemble prediction model outcomes in which initial conditions are slightly perturbed in order to generate several possible predictions, reflecting the uncertainty in the initial conditions. Still, stochastic modeling is something different than perturbing initial conditions since random numbers affect the time derivatives every time step, thereby possibly changing model behavior.

Now we have some idea about differences between deterministic and stochastic modeling. However, we have not yet discussed what the distribution is of the random numbers, and where and how they affect model tendencies. Furthermore, we have to discuss why we introduce *stochastics*. Therefore, first of all, we will motivate why we use stochastics. After that, we will look in detail how random numbers can be incorporated in models. We are mainly interested in introducing stochastic elements in the convection parameterization scheme of GCMs, as reflected in the title of this thesis.

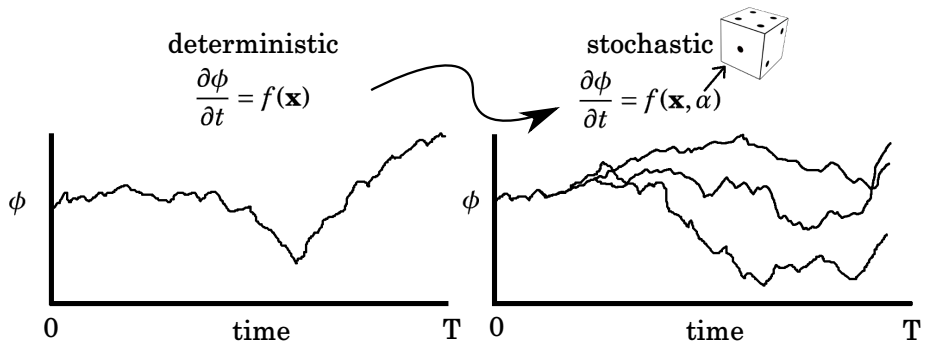

Figure 1.8: In deterministic models, each tendency of each model variable ϕ is a deterministic function of the model's prognostic variables and large-scale forcings $\mathbf{x} = \{u, v, w, \theta_l, q_t, \ldots\}$. The initial state $\mathbf{x}_0$ determines the future time state $\mathbf{x}_T$ in a deterministic way. In stochastic models, random numbers affect the model variable tendencies. In this illustration, the stochastic process is indicated by α. Model variable trajectories of ϕ are not uniquely determined by the initial state $\mathbf{x}_0$, instead several trajectories are possible and consequently several outcomes for the future time state $\mathbf{x}_T$.

The Grey Zone

GCMs are typically defined on a grid for which each vertical grid column represents a region on Earth so large that if moist convection is present, it is reasonable to assume that moist convection is in quasi-equilibrium with the large-scale forcings. The quasi-equilibrium assumption is valid if the resolution of the grid is coarse enough to be sure that a large number of convective updrafts is present in one column: for example a horizontal grid point distance of a few hundred kilometers. In this case, convective transport by cumulus clouds can be reasonably represented by parameterizations [5]. In case the resolution is so high that individual convective clouds are resolved by a model, which is the case for cloud resolving models or high-resolution models such as LES, no convection parameterization is needed.

With higher model resolutions, atmospheric flows can be simulated in more detail. Therefore, atmospheric models tend to get more accurate if the resolution increases. The availability of more computational resources enabled modelers to increase the grid resolution of GCMs for decades. This partly explains the major improvements of numerical weather and climate models. At the moment, however, complications are encountered when increasing model resolutions, because GCMs operate with resolutions that are getting close to, or are already in, the *Grey Zone* [48, 60, 149] or *terra incognita* [147]. For models with grid resolutions such that convection is partly resolved and partly unresolved, the so-called *Grey Zone resolutions*, grid resolutions in between the two extreme situations described above (Fig. 1.9), transport has to be represented in a different way. The quasi-equilibrium assumption is no longer valid since the ensemble of cumulus clouds is too small. Individual cloud life cycles are important in this case and traditional mass-flux parameterizations are not correct. Entirely omitting convection parameterizations, as is done for LES, is also not possible, because then convective transport would be underestimated. Note that by definition the Grey Zone is a range of grid resolutions that is dependent on the subgrid process that is considered. For example, the Grey Zone for deep convection differs from the Grey Zone for shallow convection, because the processes have different typical sizes. The range of Grey Zone

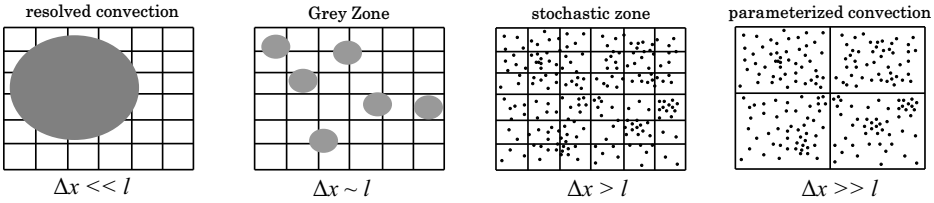

Figure 1.9: Top view of the atmosphere. Each dot represents a convective updraft with horizontal length scale l. The panels illustrate different situations: for models with high-resolutions ($\Delta x \ll l$), convection is explicitly resolved (left panel); for models with coarse resolutions ($\Delta x \gg l$), convection can be parameterized (right panel); for models with resolutions in the Grey Zone ($\Delta x \sim l$) in between the two extremes, convection is partly resolved (second panel from the left); and for models with resolutions Δx that are only slightly larger than l, convection can be parameterized, but since the number of updrafts that are present in a model column varies significantly, stochastics are needed (second panel from the right).

resolutions corresponding to shallow convection is shifted to smaller grid sizes as compared to the range of Grey Zone resolutions of deep convection.

As computational power increases, GCMs will get or already are in the Grey Zone for deep convection, followed by the Grey Zone for shallow convection. This is a problem that has to be addressed and can not be neglected.

Stochastic parameterization of convection

The intermittent and random character of moist convection vanishes when only statistical properties of an ensemble over large areas are important, and therefore, for coarse grid resolutions, *deterministic* parameterizations of moist convection are appropriate. For higher resolutions and, in particular, in the Grey Zone, the intermittent and random character of moist convection is reflected in the random fluctuations of turbulent heat and moisture fluxes around the expectation values. In the second panel from the right in Fig. 1.9, it can be seen that the number of updrafts that are present in a model column varies significantly, the number of updrafts ranges from only two to more than ten updrafts. Therefore, in or close to the Grey Zone, as is demonstrated in Chapter 2 of this thesis, *stochastic* parameterizations have more potential to adequately represent convection. One answer to the question how, for resolutions in or close to the Grey Zone, subgrid variability related to convective transport can be represented by parameterizations is 'by introducing stochastics in the parameterizations'. Stochastic processes can be used in parameterizations to represent unpredictable random effects of individual cumulus clouds and increase the variability in the output of parameterizations. This idea is one of the main topics of this thesis. We have the following questions:

- How can stochastics be introduced in convection parameterizations in an adequate way?
- What is the effect of stochastic parameterization of convection on GCM behavior?
- Is it necessary to introduce stochastics in the parameterizations, or can the same effect be obtained with deterministic parameterizations?

Assume that we have a deterministic convective parameterization scheme in a GCM. How can we make the scheme stochastic? There are several ways to do this, and indeed, different approaches have been explored by researchers:

- *multiply* the output values of the convection scheme with random numbers $r \in [1-x, 1+x]$ every time step. This is a very ad hoc method, but has been used in an even more general context by multiplying all subgrid terms with random numbers, for example in Buizza et al. (1999) [20] with $x = 0.5$. In Teixeira and Reynolds (2008) [136] only the convective tendencies are perturbed. It is important to take spatial and time correlations into account: there should be a large correlation between grid points that are close to each other, e.g., a vertical profile of the vertical heat flux is a smooth function and if the random numbers are independent, this smoothness could be destroyed.

The same is true for horizontal correlations and correlations in time. Convection gradually develops and gradually vanishes, and this could be destroyed by multiplying with random numbers without any correlation;

- *add* random numbers (with zero mean) or a stochastic process to the output of a parameterization. On top of the deterministic parameterization a random process can be used to make it stochastic. Lin and Neelin (2000) [86] showed that by adding red noise, they were able to improve the simulated total convective variance in a tropical atmospheric model of intermediate complexity and showed that the model results were sensitive to the autocorrelation time of the stochastic process.

- take *a part of the convection scheme* and make it stochastic, e.g., the trigger function [134], the mass flux at cloud base [88], or the entrainment [121]; or make one *suitable parameter* of the convection scheme stochastic instead of all outcomes of the scheme, e.g., Grant's constant for shallow convection [51] is a potential candidate. In this way, correlations are retained between fluxes inside each grid column;

- focus only on *shallow convection* [123];

- employ a *stochastic multiplume* model (introduced by Plant and Craig (2008) [114]); or a

- *stochastic multicloud* model (introduced by Khouider et al. (2010) [72]), see Section 1.3.8.

Immediately testing new stochastic parameterization approaches in *state-of-the-art* GCMs (e.g., EC-Earth [55], CAM [23]) is a large step, therefore, new parameterization approaches can be tested in less complex models, such as for example:

- mathematical multi-scale test models or '*toy-models*'; a well-known example of a toy-model is the Lorenz '96 set of equations [91]. It has two types of variables: large-scale variables and small-scale variables and has been used frequently as a testbed for new parameterization approaches, e.g., [6, 25, 81, 144]. To give an additional example of a two-layer idealized model with interacting small-scale and large-scale variables, we refer to Harlim and Majda (2013) [54]. Another test model that can be used to test parameterizations is the Kac-Zwanzig heat bath [142]. Furthermore, differential equations of the form $\frac{dq}{dt} = S$, where S is a source term with stochastic elements have been explored by [134];

- *single-column models*; in single-column models (SCMs), parameterizations can be tested in a simple environment: their interactions with the model variables can be tested in one column [10]. The model variables are not interacting with the large-scale variables in neighboring columns, because they are not present in a SCM. External forcings such as horizontal divergence and subsidence are not calculated as is done in multi-column models; instead, they are prescribed. This creates a clean test environment, without

effects due to large-scale advection etc. Furthermore, testing in SCMs is computationally inexpensive and errors can be found prior to implementation in a GCM. Preferably, testing is done with the SCM variant of the GCM - i.e., an SCM which is similar to the GCM, using for example identical codes - but without large-scale dynamics. Behavior of the parameterization in an SCM - e.g., its responds to the large-scale variables and forcings - gives a good indication of its behavior in a GCM;

- *tropical circulation models*; models in which the dynamics are confined to the tropics. These models can for example be used to examine convectively coupled equatorial waves. Often, they employ the β-plane approximation for the Coriolis force; in which case, the models solve the anelastic hydrostatic Euler equations on an equatorial β-plane [14];

- *aqua-planet models*: in these models, the entire surface of the planet is assumed to be covered by water; and they often use prescribed sea-surface temperatures (SSTs). An aqua-planet comparison study is described by [16] and a stochastic parameterization of convection has been implemented in an aqua-planet GCM by [116];

- GCMs of *intermediate complexity* usually use prescribed SSTs, have coarse resolutions, and use simplified parameterizations. They can be run with or without seasonal cycles or daily cycles, etc. Examples of intermediate complexity GCMs are: AMIP-type models [47] or the SPEEDY model [101], which is introduced in Chapter 5 of this thesis.

We have seen that stochastics can be implemented in GCMs in several ways. In this thesis, the focus is on stochastic parameterization of moist convection. The motivation is that moist convection is of major importance in Earth's atmosphere and climate and it has a major impact on model results. Furthermore, the process has a random character, and therefore, when model grid resolutions get finer, this randomness has to be represented somehow. Further confinements and choices have to be made: we have to clarify what kind of stochastic processes we use and how we will asses the stochastic parameterizations. The latter will become clear in the core chapters of this thesis (Chapters 2, 3, 4, 5). The former can be clarified as follows.

We build on the stochastic approach based on *data-driven conditional Markov chains* originally introduced by Crommelin & Vanden-Eijnden (2008) [25]. This Markov chain method for parameterization of subgrid processes has been shown to adequately represent the effects of subgrid processes in the Lorenz '96 model by [25]. Therefore, since it already has proven itself in a simplified model, a natural step is to extend it to the usage in parameterizations in more complicated models with GCMs as a final goal. We will explain the several aspects of this stochastic approach in the following sections.

Markov chains

If time correlation is desirable for the random numbers that are used in a stochastic convection scheme, this can be attained by making random numbers in param-

eterizations at time $t + \Delta t$ dependent on the random numbers at time t. If the probabilities only depend on time t and random numbers used before time t do not affect these probabilities the stochastic process is Markovian. Markov processes are computationally effective, because time correlation is present, without the need of storing long sequences of random numbers. However, care should be taken, because the effects of convection on the large-scale state are to some extent non-Markovian; and hence, using Markov models in the representations of convection could lead to errors [26, 41]. As we will see, the Markov models that are examined in this thesis are conditioned on the large-scale variables, such that memory effects due to the interaction of convection with the large-scale state of the atmosphere are included.

If only a finite number of states can be attained by the Markov process and only at equidistant discrete time points, the Markov process is called a *finite state Markov chain* [2, 52, 107] (Fig. 1.10). Discrete models with a finite number of states have been examined frequently in the context of weather and climate modeling, e.g., [25, 74, 93, 141]. The idea of using discrete stochastic models with only a few states stems from statistical mechanics [93]. In statistical mechanics these models have proven to be effective in modeling physical or chemical processes, e.g., the movement of molecules, that are too complicated to resolve explicitly [68, 83, 84]. This explains the choice of examining the use of discrete models with a finite number of states in the representations of convection and clouds, as is done in this thesis.

A finite state Markov chain is determined by its initial state distribution and a transition probability matrix of size $N \times N$, where N is the number of attainable states. The probability of a transition of a Markov chain $Y(t)$ from state m to state n is given by:

$$P(m,n) = \text{Prob}\big(Y(t+\Delta t) = n | Y(t) = m\big) \tag{1.7}$$

In the next section, we will explain, how these transition probabilities are estimated from data.

A Markov chain can be used to make a convection scheme stochastic in the

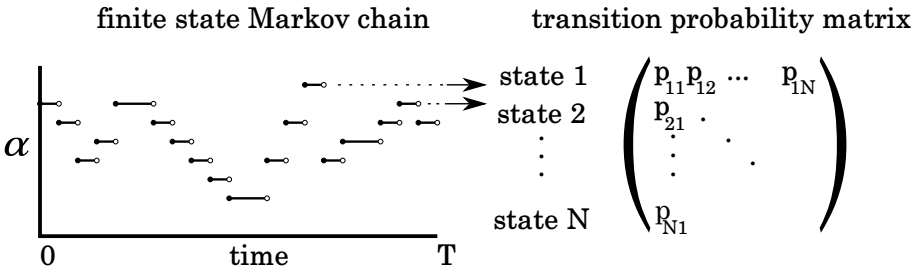

Figure 1.10: Schematic illustration of a finite state Markov chain. The Markov chain α switches between N states and only at discrete times. The probabilities of switching from state i to state j, p_{ij}, form a transition probability matrix.

several ways described before; for example by modeling a suitable parameter with a Markov chain, or by adding random numbers generated by a Markov chain to the output of the convection scheme. The states of the Markov chain can be chosen to be parameter values (Chapter 5), flux profiles (Chapter 2), or cloud types (Chapters 3, 4, 5), or some other model quantity.

Data-driven

In case a process is too complicated to derive laws and construct parameterizations from first principles, data-driven methods can be used to construct parameterizations by inferring from data. As mentioned before, to make realistic convection schemes, parameters should be estimated or at least be compared with observational data or with high-resolution LES data. Finite state Markov chains can be *estimated from data* directly. Data-driven models are getting more and more attention since computing power is increasing and more data is available. Think of deep learning algorithms, based on neural networks, that are *trained with* large data sets with a specific aiming task, such as recognizing images. Data-inferred Markov chains are also an example of a learning model, because the Markov chains are trained with data with the aim of *mimicking* observed behavior afterwards. Since convection is a complicated process for which it is unknown how to derive parameterizations from first principles, in particular for GCMs with resolutions close to or in the Grey Zone, we will use data-inferred Markov chains to mimic convective behavior as observed in data and construct convection parameterizations. The Markov chains are 'trained with' (i.e., *inferred from*) data of convection. A finite state Markov chain is inferred from data by estimating its transition probability matrix. For a matrix of size $N \times N$, it means that N^2 matrix entries have to be estimated. The transition probability matrix entry $P(m,n)$ in (1.7) is estimated as in [25]:

$$\hat{p}(m,n) = \frac{T(m,n)}{\sum_n T(m,n)}, \tag{1.8}$$

with $T(m,n)$ counting transitions from m to n observed in a given training data set:

$$T(m,n) = \sum_t \mathbf{1}\big(Y(t+\Delta t) = n\big)\mathbf{1}(Y(t) = m),$$

in which $\mathbf{1}$ is the indicator function: $\mathbf{1}(A) = 1$ if A is true and $\mathbf{1}(A) = 0$ if A is false, and t runs over time instances in the training data set. It can be shown that the estimator $\hat{p}(m,n)$ in (1.8) is the maximum likelihood estimator of $P(m,n)$ [2].

For accurate estimation, enough data should be available, which depends on the size of the matrix. When choosing the number of states of the Markov chain, one has to take into account how much data is available. The lower the number of states, the smaller the matrix, and the better its entries can be estimated. Furthermore, a Markov chain model with a smaller number of states is simpler and faster when it is implemented in a convection scheme of a GCM (e.g., smaller matrices are loaded faster). In this thesis, we will show how one can use data from high-resolution convection resolving models, in particular LES, as well as observational data to construct Markov chain models for stochastic parameterization of shallow convection and deep convection.

In the context of parameterization in weather and climate models, data-driven models have been employed by many authors, amongst others: [6, 25, 34–37, 50, 61, 81, 144].

Conditional Markov chains

The occurrence and strength of moist convection depends on the large-scale physical state of the atmosphere. Some states are favourable for moist convection and others prevent convection (e.g., an unstable versus stable atmosphere). This means that if a Markov chain is used to mimic the process of convection, it can be improved by taking the large-scale state of the atmosphere or a suitable indicator of convection into account. This can be done by making probabilities depend on the large-scale state X_t of the atmosphere. Probabilities take the following form:

$$P_\gamma(m,n) = \text{Prob}\big(Y(t+\Delta t) = n | Y(t) = m, X(t) = \gamma\big). \tag{1.9}$$

If a finite number of large-scale states is considered, then a conditional Markov chain (CMC) is constructed by estimating a transition probability matrix for each possible large-scale state [25]. In the training stage, the data is first classified into different large-scale states and then for each large-scale state, a matrix is estimated by counting transitions for each large-scale state (see again [25]:

$$P_\gamma(m,n) = \frac{T_\gamma(m,n)}{\sum_n T_\gamma(m,n)}, \tag{1.10}$$

in which:

$$T_\gamma(m,n) = \sum_t \mathbf{1}\big(Y(t+\Delta t) = n\big)\mathbf{1}(Y(t) = m)\mathbf{1}(X(t) = \gamma).$$

In this thesis, we will condition the Markov chains on several indicators of convection.

Clustering

To be able to work with finite state conditional Markov chains, the subgrid-scale state and the large-scale state need to be discretized into a finite number of states. Therefore, we need a way to classify continuous quantities (e.g., the subgrid variables and the large-scale variables) into a finite number of classes/states. To give a very simple example of classification, imagine a variable with values in the interval $[a,b]$ that we would like to classify into two classes. This can be done by choosing a threshold, say c, to obtain two classes $[a,c)$ and $[c,b]$. If more variables have to be classified into more than two classes, one can choose several thresholds. In case the training data is not uniformly distributed, choosing thresholds is difficult and could result in classes to which no data is assigned and classes to which almost all data is assigned. One way of classifying data less arbitrary and in some sense 'optimal way' is using clustering algorithms. Clustering, usually requires a distance that has to be defined beforehand. A cluster algorithm that is very useful, easy to use, and fast is *k-means* [92]. It clusters data into k classes, where k has to be chosen beforehand. After clustering, the data set is represented by k cluster centers (called *centroids*) that are positioned such that the total distance of the data points

to their nearest-by centroid is minimized. Note that k-means converges to a locally optimal solution, which is possibly not the global optimum. In this thesis we will see that with k-means we are able to discretize the subgrid-scale state as well as the large-scale state. Kwasniok (2012) [81] also works with clusters in the context of data-driven stochastic parameterizations.

Cellular automata

In Markov chain models, time-correlation is present by construction. What about spatial correlations? Moist convection is spatially organized at different scales [95]. A convective updraft has a certain horizontal size which could be larger than the horizontal size of individual vertical columns of a high-resolution model or the resolution of the observational data could be so high that an updraft covers more than one observational pixel. As mentioned before, cumulus clouds are usually organized in ensembles: a large number of updrafts in a region of hundreds of kilometers. Cumulus clouds can also be organized in streets or bands [78] or organized in larger structures that span thousands of kilometers (e.g., the Madden-Julian oscillation (MJO) [150]). Therefore, if a Markov chain is used in a GCM to parameterize convection ideally it should be sensitive to the convective state of the neighboring GCM grid columns. By conditioning on the large-scale variables, as is done with CMCs, spatial correlation is indirectly present, because large-scale variables of neighboring grid columns are correlated. It is also possible to directly include spatial correlation. By conditioning transition probabilities spatially, one obtains *cellular automata* (CA). CA are well known thanks to Conway's Game of Life [46]. John Horton Conway introduced a mathematical system consisting of a two-dimensional grid with cells that are either alive or dead and interact with neighboring cells according to a small set of deterministic rules, showing totally unexpected chaotic behavior and organization. Here, we will examine spatial coupling of Markov chains by conditioning on the states of neighboring Markov chains. Our CA also live on a two-dimensional grid and interact with neighboring cells, but the cells will have more than two states and the rules that determine the evolution of the cells are probabilistic: CA that are known as *probabilistic cellular automata* (PCA) or *stochastic cellular automata* (SCA). Let a Markov chain Y_i be positioned on grid cell i with direct neighboring cells $\{i\}$, then the probability that it switches from state m to state n is now given by:

$$P_{\gamma,\delta}(m,n) = \text{Prob}\big(Y_i(t+\Delta t) = n | Y_i(t) = m, X_i(t) = \gamma, Y_{\{i\}} = \delta\big), \tag{1.11}$$

and inference is done similar to the conditional Markov chain estimation in Eq. (1.10), but additional matrices are estimated for each neighboring configuration $Y_{\{i\}}$. The number of neighboring configurations will often be impractically large in which case a 'reduction function' can be used as an ad hoc solution to reduce the number of neighboring states. A reduction function reduces the number of neighboring configurations by assigning the same value to configurations that are similar, e.g., if the configurations are the same after a rotation. In Chapter 3, we build in spatial correlations on a size smaller than a GCM grid column: CA will be used to capture spatial structures of convection *inside the GCM grid column*. The methodology could be extended to more general applications in weather

and climate models, for example by constructing CA with spatial correlations *in between neighboring GCM grid columns*. CA have been used before in weather and climate models to stochastically represent unresolved processes [11] and especially for stochastic convection parameterization [9]. Also, in other scientific fields, data-driven SCA have been explored earlier [99, 119]. In Chapter 3, we will use data-driven SCA (stochastic rules, estimated from data) for convection parameterization.

Multicloud models

In the atmosphere, several cloud types can be discerned: stratiform clouds, shallow convective clouds, deep convective clouds etc. Observations display a cycle between different regimes of cloud types [65]. Shallow convection can lead to convection with congestus clouds, and this in turn facilitates the formation of deep convection. At the top of deep convective towers, stratiform clouds are formed, spreading out horizontally. A type of Markovian model that implements this cycle is the stochastic multicloud model of [72] for convection parameterization, in which the states of the Markov chains are *cloud types*. The Markov chains are positioned on a two-dimensional micro grid with a grid cell size smaller than the GCM grid and switch between cloud types. In each GCM column, the *cloud type area fractions* are computed by counting the relative frequencies of the states of the Markov chains situated within the column. These area fractions are useful in parameterizations of clouds and convection in GCMs. For example, the deep convective area fraction can be used as a closure for the mass flux at cloud base, and the sum of the cloud type area fractions can be used to determine the cloud fraction. A schematic two-dimensional depiction of a multicloud model can be seen in Fig. 5.1. The stochastic multicloud models developed in this thesis are based on the framework of [72]. The main difference between the model of [72] and the models described in this thesis is that, here, the transition probabilities between the cloud types are directly estimated from data: in Chapter 3 estimation is done with LES data and with observational data in Chapter 4.

Recently, several papers present work that elaborates on the multicloud model of [73] using several cloud types (e.g., [1]) and the models have been made stochastic (e.g., [30, 35–37, 42, 43, 72, 113]). Finally, a *stochastic lattice gas model*, which is similar to a multicloud model, is presented by [116].

1.4 Research objectives and overview

The main objectives of this dissertation are to examine stochastic parameterization of shallow and deep convection with high-resolution simulations; and with observations as well. In particular, the use of data-driven (conditional) Markov chains to directly infer the probabilities in the stochastic schemes from high-resolution data will be explored.

Moist convection is a process of major importance in Earth's weather and climate system. Therefore, for reliable numerical weather and climate predictions, it must be represented accurately in GCMs. With increasing resolutions, resolutions that increase towards the Grey Zone, the large and intermittent fluctuations of the

process should be present in the subgrid fluxes of GCMs as well. Traditional deterministic parameterizations used in state-of-the-art GCMs, can no longer accurately represent the process [86]. Stochastic parameterizations have been proposed to capture this increased subgrid-scale variability. However, accurate stochastic parameterizations of convection are still missing, which means that new stochastic parameterizations are needed. It has been shown that with the data-driven conditional Markov chain approach, it is possible to accurately represent subgrid-scale processes in the Lorenz '96 model [25, 91]. Therefore, we examine the possible usage of this promising approach for stochastic parameterization of moist convection in GCMs. Particular challenges are:

- *Data*; large data sets of moist convection are needed. This can be data obtained from high-resolution convection-resolving models (e.g., LES or cloud resolving model data) or observational data (e.g., rain radar data or satellite data). Data sets should have a high spatial and temporal resolution (at least every 10 minutes), giving detailed information of the convective processes. The corresponding large-scale data sets should give an accurate description of the large-scale atmospheric circumstances (e.g., CAPE values), not necessarily at the same temporal resolution.

- *Statistical inference*; transition probabilities of the Markov chains have to be estimated. For accurate estimation, for each entry in the transition probability matrix enough transitions have to be observed in the 'training' data set. Therefore, the number of small-scale and large-scale states should be small, and the number of observations should be large. The limited size of the data set constrains the number of states of the Markov chains that can be chosen such that transition probabilities are accurately estimated.

- *Finite number of subgrid-scale states*; the subgrid-scale state is formed by vertical profiles of turbulent heat, moisture and momentum fluxes (Reynolds stresses), and is high-dimensional. We have to bring back this large number of combinations of vertical profiles to a finite number of representative vertical profiles, which are the states of the Markov chains.

- *Conditioning*; in order to condition the Markov chains on the large-scale resolved variables, the transition probabilities are made dependent on the large-scale resolved variables. As is the case for the subgrid-scale state, only a finite number of large-scale states can be considered. Choosing a suitable function of the large-scale variables (e.g., CAPE, CIN, vertical integral of a resolved variable), and considering a finite number of intervals or a finite number of clusters (in particular, in case the dimension of the indicator of convection is larger than one), is necessary to bring back the large number of degrees of freedom to a finite number of large-scale states.

- *Retaining correlations*; heat, moisture and momentum fluxes are correlated in space and time and correlate with each other. Parameterizations should retain these correlations, in particular when random numbers are introduced in the parameterizations.

- *Resolution dependency*; subgrid fluxes, in particular associated with convection, are model grid resolution dependent. Fluctuations in the parameterization outputs should have the correct amplitude and frequency, corresponding to the horizontal grid size. Especially, in the Grey Zone, fluctuations of subgrid fluxes are important. Parameterizations should be scale-aware, i.e., able to adapt to the horizontal resolution of the GCM.

- *From scratch or adaptation of existing parameterizations*; one can choose to design, construct and implement a new convection parameterization scheme starting from scratch. This approach has been explored in Chapter 2. However, in practice it can be advantageous to adapt an existing scheme, that has proven its usability in the GCM already, as we will see in Chapter 4 and 5. In the development of the GCM, the convection scheme has been adapted, tuned to the model, and vice versa, the model has been adapted to the convection scheme.

- *Implementation in a GCM*; after construction and testing (in for example SCMs) of a new parameterization, the implementation in a GCM is the next step to take. Coupling a new parameterization scheme to the resolved model variables of a GCM can be complicated, and all subgrid fluxes, needed by the particular GCM, have to be produced by the parameterization scheme. Further, the code has to be adapted to the code of the GCM. For example, the scheme has to be written in the programming language in which the GCM code is written. Usually numerical weather and climate models are written in FORTRAN. Translating code to FORTRAN code is of course not the most difficult challenge, but one has to keep in mind that it's a part of the process of developing.

- *Assessment of a new scheme*; and last but not least, when the new (or adapted) convection scheme is implemented in a GCM, the scheme and the GCM have to be assessed. Does the scheme better represent convection? And are the model results better? In Chapter 5 of this thesis, we will see ways to assess a new convection scheme that has been implemented in a GCM.

The following chapters present work from four sub-projects:

- in Chapter 2, LES data of shallow cumulus convection is used to construct a parameterization of shallow cumulus convection. Conditional Markov chains mimic shallow cumulus convection as observed in BOMEX. Its states are vertical profiles of heat and moisture fluxes. The parameterization is tested in an SCM with BOMEX forcings. Also the Grey Zone is examined by looking at subgrid fluxes at horizontal sub-domains of different sizes;

- in Chapter 3, LES data of deep convection is used to construct a data-driven multicloud model using conditional Markov chains and SCA that mimic the development of deep convection during one day. The focus lies on cloud area fractions for several cloud types. Relative entropy is used to find the best indicator of deep convection. CAPE and CIN values are clustered and the model is tested in an SCM;

- in Chapter 4, observational data from a rain radar is combined with large-scale re-analysis data to construct a conditional Markov chain multicloud model almost in the same way as is done in the Chapter 3. Cross-correlation analysis of the observational high-resolution data combined with large-scale re-analysis data is used to find the best indicator of deep convection;
- in Chapter 5, the conditional Markov chain multicloud model as described in Chapter 4, is implemented in a GCM of intermediate complexity (SPEEDY). Also a second conditional Markov chain model, similar to the model of [50], is implemented. Convective area fractions are used as a stochastic closure for the mass flux at cloud base.

Finally, Chapter 6 presents conclusions, a synthesis and an outlook for future research.

Chapter II
Stochastic parameterization of shallow convection

2.1 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 fluxes. 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 behavior 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.

2.2 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 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 pa-

This chapter has been published as Dorrestijn, J., Crommelin, D.T., Siebesma, A.P., Jonker, H.J.J., 2013: Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data, *Theor. Comput. Fluid Dyn.*, **23**, pp. 133–148. [36].

rameterization, 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 (1974) [5], there are two 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. 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 2008 [114]) 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 [20, 74, 86–88, 93, 109, 114, 136]. 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 (2008) [25], 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*, i.e., 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 behavior 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 three kilometers. 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 subtropical oceans where it enhances the moisture transport toward the Intertropical 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 modeling. For a comprehensive introduction to shallow convection, see Siebesma (1998) [126].

The second category is that of *deep convection* by cumulus towers that reach heights up to 15 kilometers. 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 kilometers, the so-called *Grey Zone* or *terra incognita*, see [48, 60, 147, 149] 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 [25] 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 [25], because of the relative simplicity of the test model used there. In [25], the Lorenz 96 (L96) model [91] 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 five *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 two vertical profiles, the heat and moisture turbulent fluxes. These fluxes are strongly correlated, and must be treated as such in the CMC parameterization.

In [25], 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 Section 2.3, 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 Section 2.4, 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 Section 2.5, we describe in detail how to construct a CMC, and in Section 2.6, results are given and the CMC is tested in a *single-column model* (SCM) setting. Finally, in Section 2.7, we summarize and discuss our findings and make some suggestions concerning future work.

2.3 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:

$$\theta_l = \theta - \frac{L}{c_p \pi} q_l, \tag{2.1}$$

$$q_t = q_v + q_l \tag{2.2}$$

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 the water vapor 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:

$$\frac{\partial \overline{\theta_l}}{\partial t} = -\frac{\partial \overline{w'\theta_l'}}{\partial z} - \overline{\mathbf{v}} \cdot \nabla \overline{\theta_l} - \overline{w} \frac{\partial \overline{\theta_l}}{\partial z} + \left.\frac{\partial \overline{\theta_l}}{\partial t}\right|_{\text{rad}} \tag{2.3}$$

$$\frac{\partial \overline{q_t}}{\partial t} = -\frac{\partial \overline{w'q_t'}}{\partial z} - \overline{\mathbf{v}} \cdot \nabla \overline{q_t} - \overline{w} \frac{\partial \overline{q_t}}{\partial z} \tag{2.4}$$

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 (2.3) and (2.4). For shallow 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 $\phi \in \{\theta_l, q_t\}$ as:

$$\frac{\partial \overline{\phi}}{\partial t} = \frac{\partial \overline{\phi}}{\partial t}_{\text{convection}} + \frac{\partial \overline{\phi}}{\partial t}_{\text{forcing}} \tag{2.5}$$

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:

$$\overline{w'\phi'}(z) = f^\phi(z; \overline{\theta_l}, \overline{q_t}, F_\phi), \qquad \phi \in \{\theta_l, q_t\}. \tag{2.6}$$

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 (2010) [63].

Since the 1960s, researchers have proposed various ways to parameterize convective processes in a model column (see e.g., [3] 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 (e.g., [114]). 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 $\overline{w'\theta_l'}$ and $\overline{w'q_t'}$ 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, i.e., each centroid is associated with both a heat flux and a moisture flux. They are denoted by $(f_\alpha^{\theta_l}(z), f_\alpha^{q_t}(z))$, $\alpha = 1, ..., N_\alpha$ (thus, α is the cluster centroid index). Once the clusters and their centroids are determined, the timeseries of LES flux profiles $(\overline{w'\theta_l'}, \overline{w'q_t'})(z,t)$ can be mapped to a timeseries $\alpha(t)$ for the centroid index.

The key element of our parameterization approach is to infer a Markov chain stochastic process from the LES timeseries $\alpha(t)$, and to use this Markov chain to

Table 2.1: 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

emulate the temporal behavior 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 timeseries. The Markov chain generated timeseries of α is mapped to a timeseries of turbulent fluxes by using the cluster centroids:

$$\left(\overline{w'\theta_l'}(z,t), \overline{w'q_t'}(z,t)\right)^{\mathrm{CMC}} = \left(f^{\theta_l}_{\alpha(t)}(z), f^{q_t}_{\alpha(t)}(z)\right). \tag{2.7}$$

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 $\overline{\theta_l}$ and $\overline{q_t}$ into N_μ clusters. The timeseries of the LES resolved variable profiles can be mapped to a timeseries $\mu(t)$ 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 $N_\alpha \times N_\alpha$.

2.4 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. (2010) [56]. The horizontal and vertical grid point distance is on the order of tens of meters, 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 kilometers. 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 [56].

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) [58]. Dur-

ing this phase, a typical steady state was observed for a period of five 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 literature [127] [128]. For the details of the initial profiles and the prescribed large-scale forcings, we strictly follow the case setup such as described in Siebesma [127].

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 2.1. 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. Obviously, convection is explicitly 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 [125].

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. 2.1). 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 $\Delta x, l$ and L are related as follows:

$$J = \left(\frac{l}{\Delta x}\right)^2, \qquad K = \left(\frac{L}{l}\right)^2. \tag{2.8}$$

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 k-th subdomain and over the entire domain:

$$\overline{\phi}^{l_k} := J^{-1} \sum_j \phi_{j,k}, \tag{2.9}$$

$$\overline{\phi}^{L} := (JK)^{-1} \sum_{j,k} \phi_{j,k} = K^{-1} \sum_k \overline{\phi}^{l_k}, \tag{2.10}$$

where $\phi \in \{w, \theta_l, q_t\}$. For the k-th subdomain, one can calculate the turbulent flux relative to the subdomain average $\overline{\phi}^{l_k}$, or relative to the entire domain average $\overline{\phi}^{L}$.

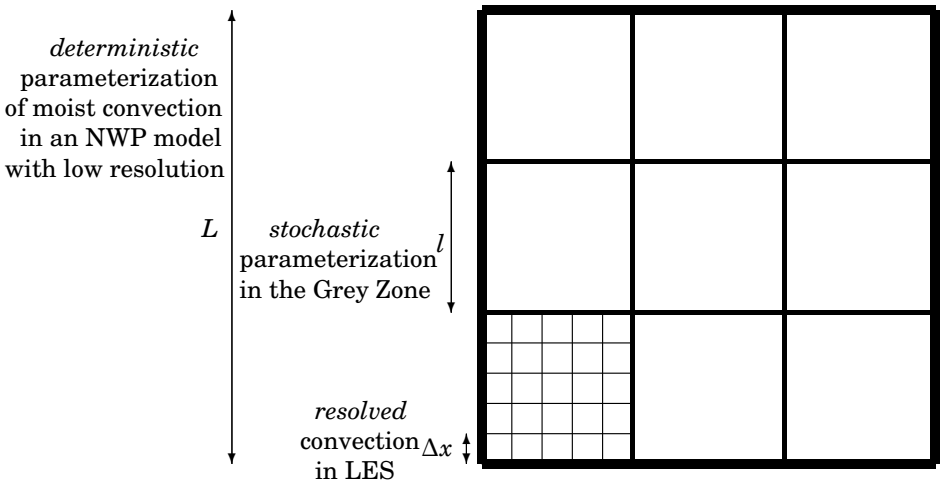

Figure 2.1: A depiction of the three length scales discussed in Section 2.4. At the length scale L of the entire LES domain, *deterministic* parameterizations relying on equilibrium assumptions can still be adequate. At the length scale Δx of the LES model resolution, convection is explicitly *resolved*. In the Grey Zone, with model resolutions of size l, in between L and Δx, *stochastic* parameterizations are needed.

The first case gives:

$$\overline{w'\phi'}^{l_k} = J^{-1}\sum_j (w_{j,k} - \overline{w}^{l_k})(\phi_{j,k} - \overline{\phi}^{l_k}), \qquad \phi \in \{\theta_l, q_t\}, \tag{2.11}$$

and is related to the second as follows:

$$J^{-1}\sum_j (w_{j,k} - \overline{w}^{L})(\phi_{j,k} - \overline{\phi}^{L}) = \overline{w'\phi'}^{l_k} + (\overline{w}^{l_k} - \overline{w}^{L})(\overline{\phi}^{l_k} - \overline{\phi}^{L}), \qquad \phi \in \{\theta_l, q_t\}. \tag{2.12}$$

For the turbulent flux over the whole domain we have:

$$\overline{w'\phi'}^{L} = K^{-1}\sum_k \overline{w'\phi'}^{l_k} + K^{-1}\sum_k (\overline{w}^{l_k} - \overline{w}^{L})(\overline{\phi}^{l_k} - \overline{\phi}^{L}), \qquad \phi \in \{\theta_l, q_t\}. \tag{2.13}$$

As is clear, it is not equal to the sum of the subdomain fluxes obtained with (2.11). 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. (2.11) and not with Eq. (2.12).

With Eq. (2.13), 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. 2.2. For this figure, we used two LES data sets for the BOMEX case: our standard data set with $\Delta x = 50$ m resolution and $L = 25.6$ km domain length, and an additional data set with $\Delta x = 12.5$ m and

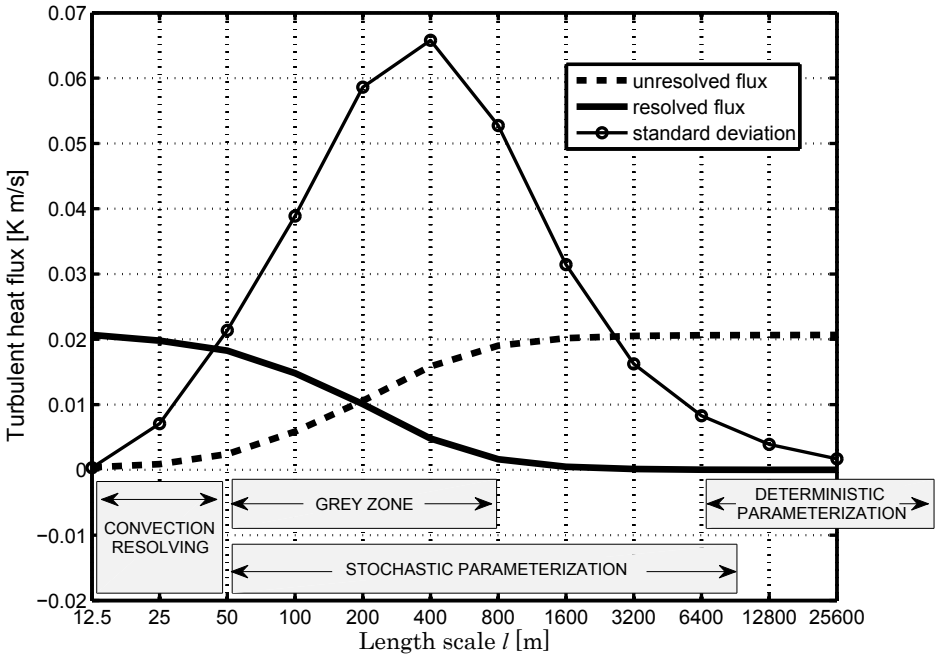

Figure 2.2: A decomposition of the turbulent flux $-\overline{w'\theta'_l}^L$ of the whole LES domain of horizontal size 25.6×25.6 km^2 in a resolved part and an unresolved part according to Eq. (2.13) as a function of subdomain length l (at height 1,000 m). In the *Grey Zone* these parts are of the 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 convective clouds to use a deterministic parameterization for the unresolved turbulent fluxes. For smaller length scales, stochastic parameterizations are more appropriate.

$L = 6.4$ km. Including the second data set enables us to cover a wider range of length scales in Fig. 2.2 (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 2011 [60]). 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. 2.3, 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 [130]. 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 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 $\overline{w'\phi'}^{l_k}$.

2.5 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.

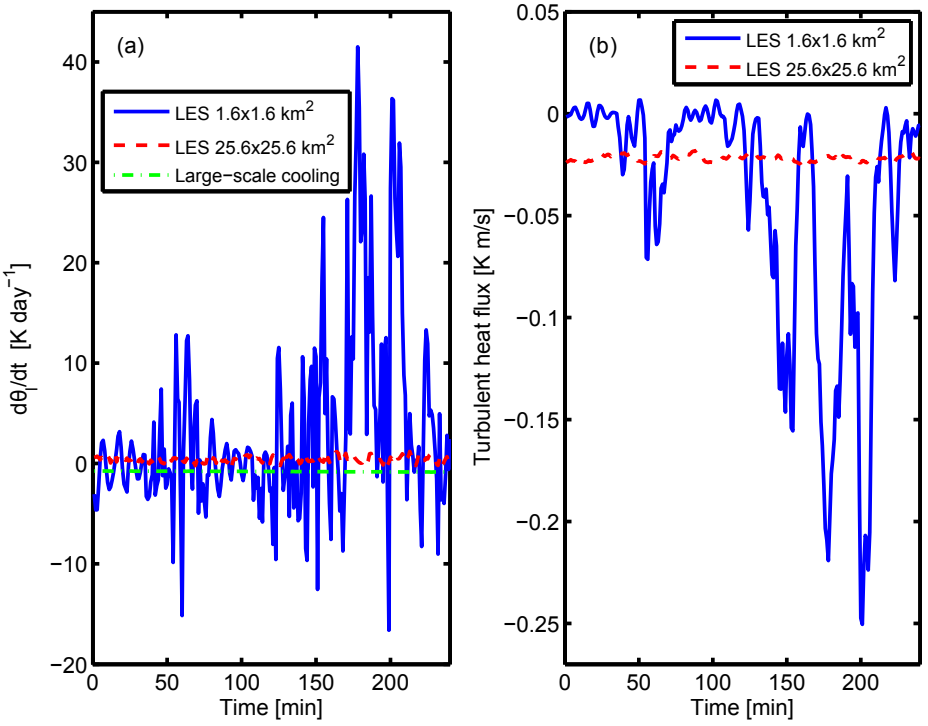

Figure 2.3: (**a**) At height 1,000 m the turbulent heating in the whole LES domain of horizontal size 25.6×25.6 km^2 (dashed line), i.e., $-\partial\overline{w'\theta_l'}^{L}/\partial z$, is in quasi-equilibrium with the large-scale cooling (dash-dotted line), while this is not the case for the turbulent heating in a subdomain of horizontal size 1.6×1.6 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).

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 [45]. 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 (2007) [7]. Apart from the initialization, the algorithm of k-means++ is the same as the k-means algorithm first described by Macqueen (1967) [92]. 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 Section 2.7, 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 Section 2.6, 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 $h = \big(h_1(z), h_2(z)\big)$:

$$d(g,h) = \sqrt{\sum_z c_1\big(g_1(z) - h_1(z)\big)^2 + c_2\big(g_2(z) - h_2(z)\big)^2}. \tag{2.14}$$

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. 2.4, 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 indiction 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 $8.0 \cdot 10^{-3}$ K m/s for the heat flux and $5.2 \cdot 10^{-5}$ m/s for the moisture flux. We have numbered the centroids such that $\alpha = 1$ corresponds to a

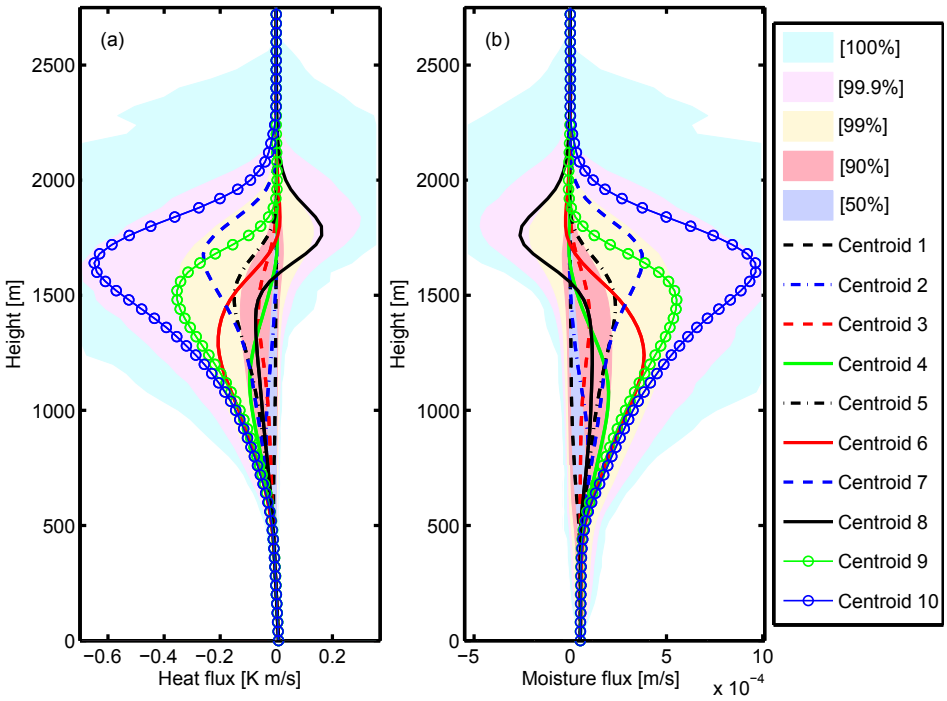

Figure 2.4: The ten centroids (i.e., pairs of turbulent (**a**) heat and (**b**) moisture flux profiles) calculated using the k-means++ clustering algorithm with $N_\alpha = 10$. The *shading* indicates for every height the percentage of (**a**) heat and (**b**) moisture flux profiles passing through 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.

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. 2.9a, 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 to 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 Section 2.6.3.

Conditioning on the resolved-scale state

We employ the same clustering method (k-means++) and the same distance function (2.14) 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 Section 2.6.1, 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.

Estimation of the transition probability matrices

Once the clustering of the turbulent fluxes and the resolved-scale states is completed, the LES data can be mapped to timeseries $(\alpha_k^{\mathrm{LES}}(t), \mu_k^{\mathrm{LES}}(t))$ 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 timeseries, 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:

$$\mathbf{P}_{nm}^{(i)} = \mathrm{Prob}[\alpha_k^{\mathrm{LES}}(t+\Delta t) = m \,|\, \alpha_k^{\mathrm{LES}}(t) = n, \mu_k^{\mathrm{LES}}(t) = i] \tag{2.15}$$

We do so using the following estimator (as in [25]):

$$\hat{P}_{nm}^{(i)} = \frac{T_{nm}^{(i)}}{\sum_m T_{nm}^{(i)}}, \tag{2.16}$$

where:

$$T_{nm}^{(i)} = \sum_k \sum_t \mathbf{1}[\alpha_k^{\mathrm{LES}}(t+\Delta t) = m]\mathbf{1}[\alpha_k^{\mathrm{LES}}(t) = n]\mathbf{1}[\mu_k^{\mathrm{LES}}(t) = i]. \tag{2.17}$$

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, $T_{nm}^{(i)}$ 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 [25], 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 [106].

Numerical integration with the CMC parameterization

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

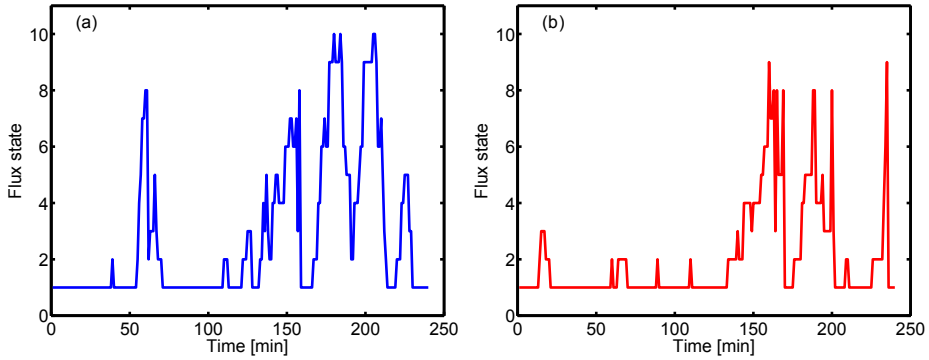

Figure 2.5: (**a**) Discretized turbulent fluxes (states) observed in one LES subdomain of horizontal size 1.6×1.6 km^2 with $N_\alpha = 10$. This discretization is part of the CMC construction algorithm, see Sect. 2.5.1 (**b**) Turbulent flux states produced by CMC (in Experiment 1) using the observed resolved-scale states of the same LES subdomain.

2. Update the resolved-scale state by integrating it, using (2.3) and (2.4), from t to $t+\Delta t$. During this step, the turbulent fluxes in column k are fixed at $(\overline{w'\theta_l'}(z), \overline{w'q_t'}(z)) = (f_n^{\theta_l}(z), f_n^{q_t}(z))$ with $n = \alpha_k^{\mathrm{CMC}}(t)$.

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 $\hat{P}_{nm}^{(i)}$ for m, with $n = \alpha_k^{\mathrm{CMC}}(t)$. Now $\alpha_k^{\mathrm{CMC}}(t+\Delta t) = m$. 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 (2.14) are needed. For step 2, the flux state centroids $(f_n^{\theta_l}(z), f_n^{q_t}(z))$ are required. The stochastic matrices $\hat{P}^{(i)}$ are used in the 3rd step.

2.6 Results

We construct and test the CMC parameterization using the LES data shown in Table 2.1. To construct the CMC we perform the three calculations mentioned at the start of Section 2.5: we determine $N_\alpha = 10$ turbulent flux centroids (Fig. 2.4), we consider $N_\mu = 10$ resolved-scale states determined by the vertical profiles of $\overline{\theta_l}$ and $\overline{q_t}$, and compute the ten 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 $\mu_k^{\mathrm{LES}}(t)$ as input. Thus, the CMC-produced flux profiles do not feed back onto 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.

Experiment 1: statistics of the CMC

In this experiment we use the resolved-scale state time series $\mu_k^{\mathrm{LES}}(t)$ obtained from the LES data to "drive" the CMC. The result is the CMC-generated time series

$\alpha_k^{\mathrm{CMC}}(t)$ with $k = 1,...,K = 256$ and $t = t_1,...,t_N$, $N = 240$. These can be compared to the LES time series $\alpha_k^{\mathrm{LES}}(t)$. For an example of a flux state sequence produced by LES and by CMC, see Fig. 2.5.

The CMC sequences $\alpha_k^{\mathrm{CMC}}(t)$ can be mapped to sequences for the turbulent fluxes by using the flux centroids $(f_\alpha^{\theta_l}(z), f_\alpha^{q_t}(z))$. In Fig. 2.6, we display the mean and the standard deviation of the vertical profiles of the 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 [139]. 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 (Section 2.6.2). 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 Eq. (2.3) and Eq. (2.4) 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.

Experiment 2: implementation in an SCM setting

We test the CMC, described in the first paragraph of Section 2.6, in an SCM setting. An SCM is a one-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 $\overline{\theta_l}$ and $\overline{q_t}$ using the CMC to generate turbulent fluxes. The governing equations for $\overline{\theta_l}$ and $\overline{q_t}$ are analogous to Eq. (2.3) and Eq. (2.4):

$$\frac{\partial \overline{\theta_l}}{\partial t} = -\frac{\partial \overline{w'\theta_l'}}{\partial z} - w_{\mathrm{LSS}} \frac{\partial \overline{\theta_l}}{\partial z} + \frac{\partial \overline{\theta_l}}{\partial t}_{\mathrm{rad}}, \tag{2.18}$$

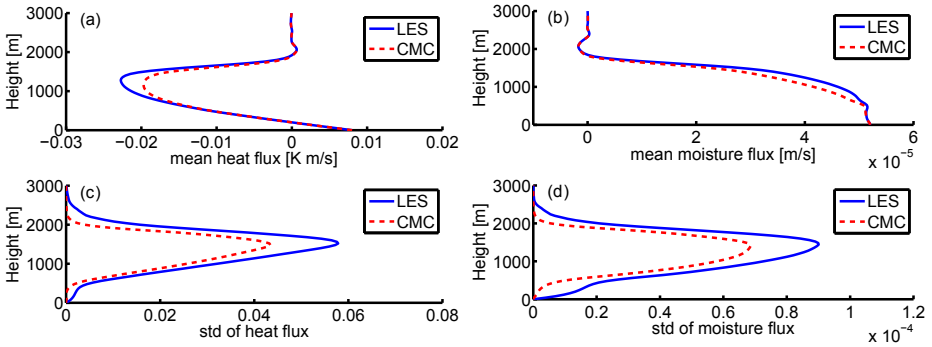

Figure 2.6: Mean vertical profile of the turbulent (**a**) heat and (**b**) moisture fluxes observed in LES subdomains of $1.6 \times 1.6 \times 3.2$ km^3 (*solid*) and produced by CMC (*dashed*) in the first experiment. (**c**, **d**) The corresponding standard deviations.

and:

$$\frac{\partial \overline{q_t}}{\partial t} = -\frac{\partial \overline{w' q_t'}}{\partial z} - F_{\mathrm{LSHA}} - w_{\mathrm{LSS}} \frac{\partial \overline{q_t}}{\partial z}, \tag{2.19}$$

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 $\overline{\theta_l}$ and $\overline{q_t}$ are radiative cooling ($\frac{\partial \overline{\theta_l}}{\partial t}_{\mathrm{rad}}$) and large-scale horizontal advection (F_{LSHA}), respectively.

We set the initial profiles of $\overline{\theta_l}$ and $\overline{q_t}$ 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. (2.18) and Eq. (2.19).

We calculate the time evolution of $\overline{\theta_l}$ and $\overline{q_t}$ 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 four hours of integration. Then by calculating probability density functions (PDFs) of $\overline{\theta_l}$ and $\overline{q_t}$ at several heights. In Fig. 2.7 we see the vertical profiles of $\overline{\theta_l}$ and $\overline{q_t}$ of 256 LES subdomains and 256 independent SCM realizations after four hours 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 $\overline{\theta_l}$ and $\overline{q_t}$ of LES and SCM in Fig. 2.8. Here we also plot the results of an SCM experiment in which we use an *unconditioned* Markov chain (MC), i.e., $N_\mu = 1$: 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 be equal, but after four hours of integration, *the ensemble spread for $\overline{\theta_l}$ and $\overline{q_t}$ resembles the spread of the profiles produced by LES.*

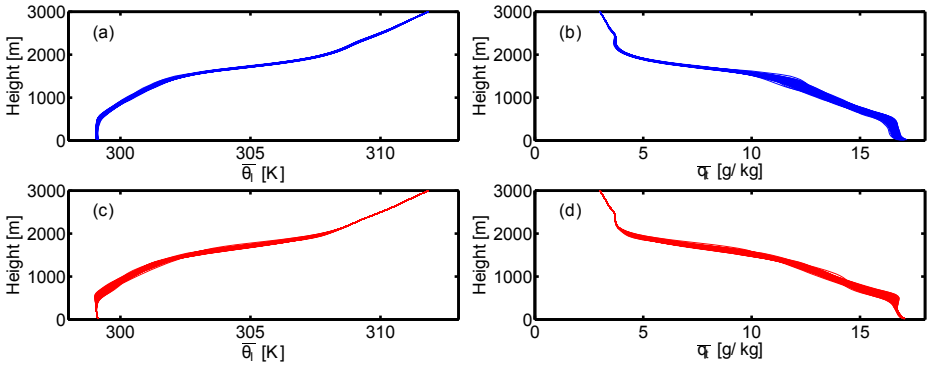

Figure 2.7: Superimposed vertical profiles of $\overline{\theta_l}$ and $\overline{q_t}$ of (**a, b**) 256 LES subdomains and (**c, d**) 256 independent SCM-CMC realizations after four hours of integration in the second experiment

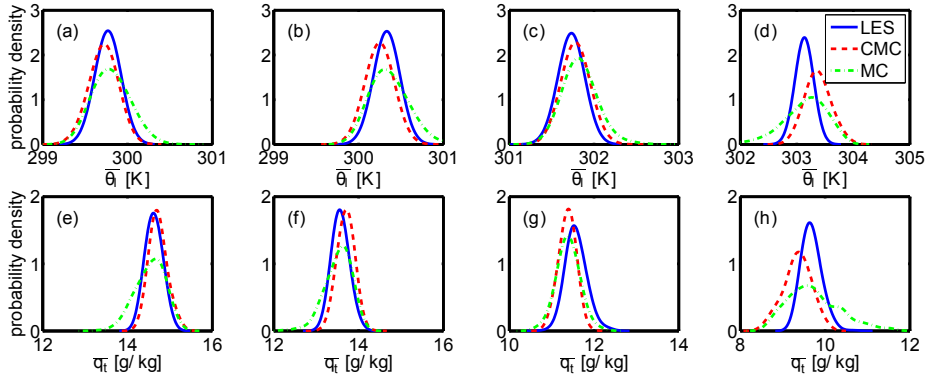

Figure 2.8: PDFs of $\overline{\theta_l}$ and $\overline{q_t}$ at heights (**a, e**) 800, (**b, f**) 1,000, (**c, g**) 1,400 and (**d, h**) 1,600 m of 256 LES subdomains (solid line) and 256 independent SCM realizations after four hours of integration using CMC (dashed line) and MC (dash-dotted line) in the second experiment

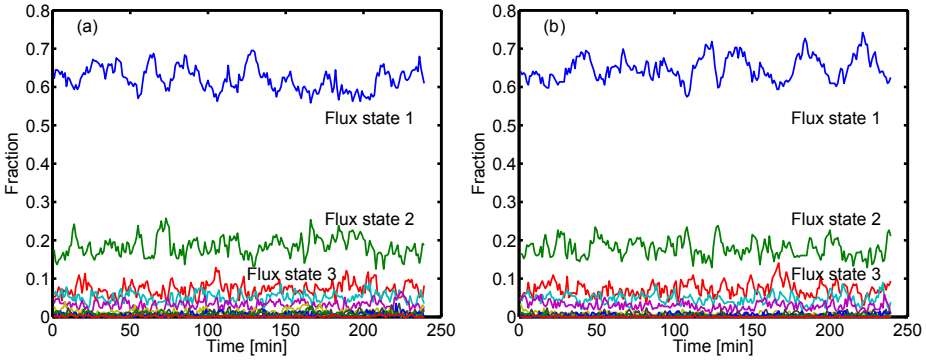

Figure 2.9: Time series of the fractions of the ten flux states (**a**) observed in the 256 LES subdomains and (**b**) produced in the 256 SCM-CMC realizations in the third experiment.

Experiment 3: implementation 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 k-th run set equal to the profiles of the k-th 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. 2.9. The figure is inspired by a similar figure in Khouider et al. (2010) [72]. 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 one minute. What is not well visible in Fig. 2.9 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 Section 2.6.1.

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

2.7 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 modeled 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 (2008) [25], in which the condi-

tional 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.

Modeling 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 behavior of LES by randomly jumping between the centroids, according to transition probabilities estimated from the LES data.

We demonstrated in Section 2.6 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.

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 are present 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 not only conditioning the transition probability on the state of the subdomain of interest but also on the state of the neighboring 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 data set 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 [25], 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. 2.2). 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. (2010) [72] 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.

2.8 Acknowledgment

The project is funded by the NWO-program "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:
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.

Chapter III
Stochastic parameterization of deep convection

3.1 Abstract

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 parameterization 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 modeled 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.

3.2 Introduction

General circulation models (GCMs) are unable to capture the medium-term variability in the tropical atmosphere. Lin et al. [85] made a comprehensive study of the tropical wave spectra determined from the Intergovernmental Panel on Climate Change (IPCC) GCMs and showed that none were able to reproduce the observed power spectrum [143] of convectively coupled Kelvin waves, two day waves, westward inertio-gravity waves and, least of all, the Madden-Julian oscillation [150]. 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 [150].

One bias that [85] 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 [109] used simple arguments from dynamical systems to show how the reduction of a chaotic dynamical system to a smaller number

This chapter has been published as Dorrestijn, J., Crommelin, D.T., Biello, J.A., Böing, S.J., 2013: A data-driven multi-cloud model for stochastic parametrization of deep convection, *Phil. Trans. R. Soc. A*, **371**, pp. 20120374[34]

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, 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 [11, 42, 72, 108, 109, 114, 124].

Khouider et al. [72] created a stochastic multicloud model based on the deterministic multicloud model of [73]. The deterministic multicloud model was derived to correspond to the observed behavior of tropical waves [97], 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 [143]. When implemented in a GCM, it has been shown to capture much of the convectively coupled equatorial wave [75] activity.

In the stochastic model [72], convection is modeled on a two-dimensional microlattice 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 [72], it was shown that the stochastic multicloud model has a large degree of variability. When coupled to a one-dimensional dynamical core [42], it produces a large degree of gravity wave variability.

Crommelin & Vanden-Eijnden [25] 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 [36] 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 [36], 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 [72] are cloud types (congestus clouds, deep convective clouds, stratiform clouds, and clear sky) rather than flux states. The states and switching probabilities used in [72] 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 as in [72] using a data-driven approach from [25, 36]. Much as in [72], 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 [25] and in [36].

Specifically, we use eight hours of simulation of the development of tropical con-

vection based on an idealization of observed conditions in northwest Brazil [146]. Simulated cloud top and rain water path are stored to classify states on the LES (horizontal) grid nodes. We use five states: (i) *clear sky* and the four cloud types (ii) *shallow cumulus*, (iii) *congestus*, (iv) *deep* and (v) *stratiform*. 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 5 h, deep convective towers with heavy precipitation develop. The deep convective towers turn into passive stratiform decks that spread and dissolve.

The paper is organized as follows. In Section 3.3, we discuss how we model transitions between cloud types with Markov chains (MCs), and how these MCs 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 3.4. The stochastic multicloud model is described in Section 3.5. In Sections 3.6-3.8, we infer the transition probabilities of the MCs and assess their ability to reproduce (emulate) the cloud filling fractions from the LES data. In Section 3.6, we use a MC without conditioning, in Section 3.7 a MC conditioned on the environment, and in Section 3.8 a MC conditioned on cloud types at neighboring lattice sites. Then, we discuss implementation of the multicloud model into a simple single-column model (SCM) (Section 3.9), 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 3.10.

3.3 Modeling cloud type transitions with Markov chains

A central element in the stochastic parameterization approach used here and in recent studies [25, 36, 72] 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 modeled as a MC, so $Y_i(t)$ changes randomly in accordance with a set of transition probabilities. In the most basic form, these probabilities are simply:

$$p(\alpha,\beta) = \text{Prob}\left(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha\right). \tag{3.1}$$

However, in this basic formulation, the probability of, for example, 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 recent studies [25, 36, 72], 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 (RH)), the transition probabilities of such a conditional MC (CMC) are:

$$p_\gamma(\alpha,\beta) = \text{Prob}\left(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha,\, X_i(t) = \gamma\right). \tag{3.2}$$

As can be seen, the transition probabilities in (3.1) and (3.2) are not explicitly dependent on the convective states of neighboring grid points. If i and j are neigh-

boring grid points, Y_i and Y_j are completely uncoupled in case of (3.1). They are coupled indirectly via X_i and X_j in case of (3.2), because X_i and X_j are coupled at the macroscopic level. Since i and j are neighboring grid points, X_i and X_j will be strongly correlated. In this paper, we also explore explicit conditioning on the neighborhood, as this is likely to improve the spatial correlation of the parameterized convection patterns. We do this by considering the conditional transition probabilities:

$$p_\delta(\alpha,\beta) = \text{Prob}(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha,\, Y_{\{i\}}(t) = \delta), \tag{3.3}$$

and:

$$p_{\gamma,\delta}(\alpha,\beta) = \text{Prob}(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha,\, X_i(t) = \gamma,\, Y_{\{i\}}(t) = \delta), \tag{3.4}$$

where $\{i\}$ denotes the neighborhood of i (e.g., the eight direct neighbors on the lattice). We note that by conditioning the MC on neighboring states, as in (3.3), the MC effectively becomes a stochastic cellular automaton (SCA). A schematic overview of the generalizations of the Markov chains is shown in Fig. 3.1.

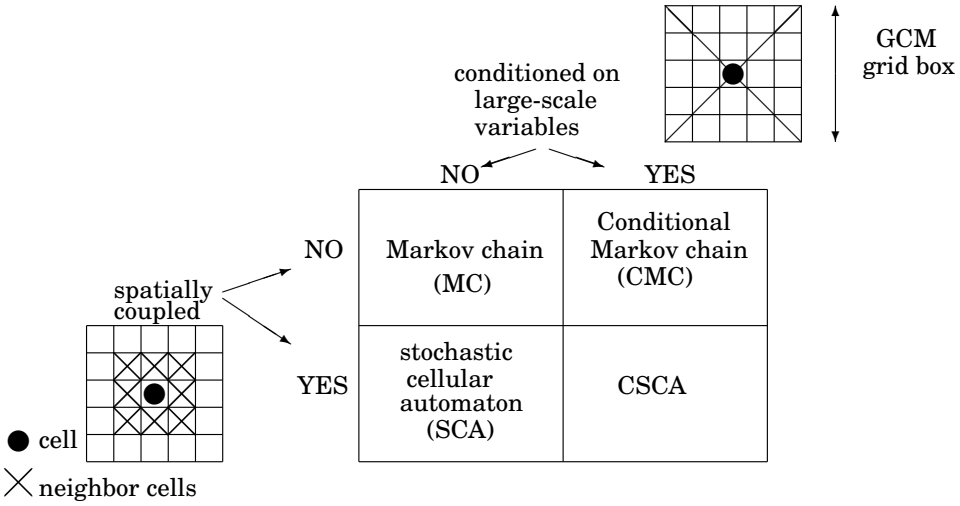

Figure 3.1: A Markov chain (MC) can be conditioned on the macroscopic state to obtain a CMC (eq. (3.2)) or on the state of the nearest neighbors to obtain a SCA (eq. (3.3)).

Each grid point on the microlattice has a state that evolves randomly according to the same set of transition probabilities, e.g., (3.2). At the macroscopic level, square blocks of microlattice sites are grouped together, and we study the filling fractions (or area fractions) of the various convective states. For each block, we have:

$$\sigma_\alpha(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}(Y_i(t) = \alpha), \tag{3.5}$$

where n is the number of microlattice 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.

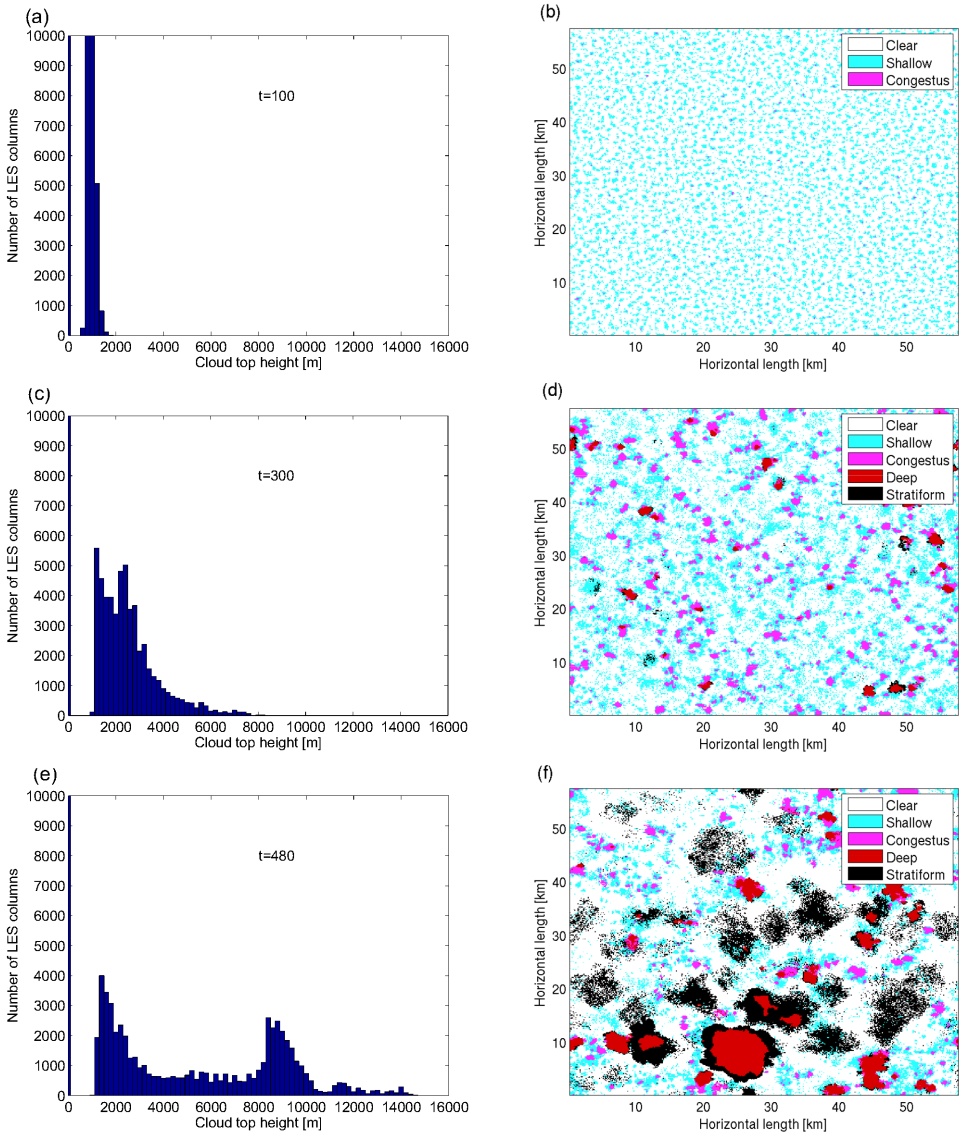

Figure 3.2: (a,c,e) Histograms of the cloud top at different time instances of the simulation. (b,d,f) Three snapshots of the LES field for which all columns are assigned to one of the five cloud types.

3.4 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 [56]. The simulation is based on an idealization of observed conditions [146] during the tropical convection experiment TRMM-LBA carried out in northwest 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 eight-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 [17].

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 m and the vertical spacing increases exponentially from 40 m near the surface to 200 m 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 m, the other in the lower free troposphere at 2,345 m. These variables are defined by:

$$\theta_l = \theta - \frac{L}{c_p \pi} q_l \qquad \text{and} \qquad q_t = q_v + q_l \,, \tag{3.6}$$

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 vapor 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.

Classification of cloud types

In the vein of [96] and [72], we consider five cloud types: clear sky, shallow cumulus, congestus, deep convection and stratiform. Fig. 3.2 (a,c,e) 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 and 5,000 m. 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:

$$CRF := \frac{\text{rain water path}}{\text{cloud top}}. \tag{3.7}$$

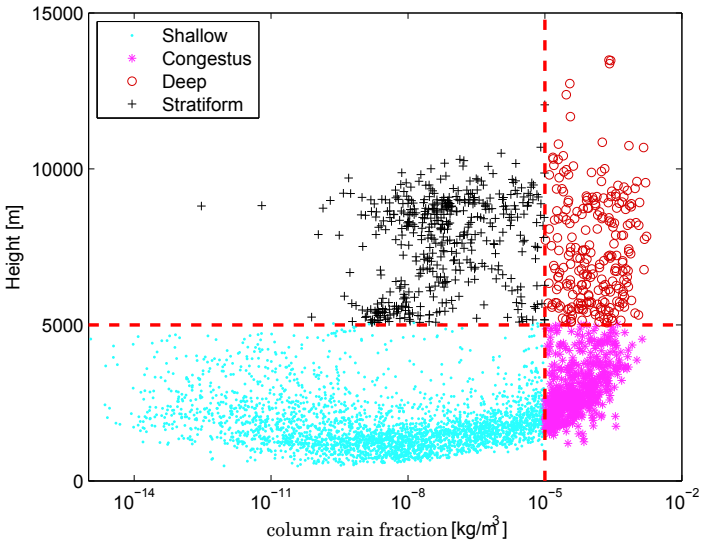

Figure 3.3: Classification of cloud types using cloud top and CRF.

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} kg m^{-3}, to distinguish deep from stratiform as well as non-raining shallow cumulus from raining congestus clouds. In Fig. 3.3, we plot the CRF against the cloud top height and indicate four cloud types with different symbols. The clear sky group is not shown because its CRF is not well defined. In Table 3.1, 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. 3.2 (b,d,f) 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 $t = 300$, deep towers start to develop. At $t = 480$, we see larger deep towers and dissolving stratiform decks.

3.5 The stochastic multicloud model

With the LES data discretized according to Table 3.1, we can choose the size of the macroscopic blocks and calculate the filling fractions $\sigma_\alpha(t)$ on each of these blocks using (3.5). 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 entire LES domain is covered by 12^2 of such (non-overlapping) blocks. In Fig. 3.4a we show the time evolution

Table 3.1: Classification of the clouds. CRF defined in (3.7).

Cloud type	cloud top [m]	rain [kg m^{-3}]
Clear sky	n/a	n/a
Shallow cumulus	$200 \le h < 5,000$	CRF $\le 10^{-5}$
Congestus	$200 \le h < 5,000$	CRF $> 10^{-5}$
Deep	$h \ge 5,000$	CRF $> 10^{-5}$
Stratiform	$h \ge 5,000$	CRF $\le 10^{-5}$

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 microlattice site as a MC. The states on the microlattice sites can be grouped again in macroscopic blocks (of any desired size), leading to emulated filling fractions. As already mentioned, the number of MCs grouped together in the multicloud model in one macroscopic block will be $1,024$; except for the creation of plots in Fig. 3.7b, 3.8b and 3.11b where we use blocks of 64 MCs.

The transition probabilities that characterize the MC are of the form (3.1), (3.2), (3.3) or (3.4). Their numerical values are estimated from the LES data. We use a time step Δt of one minute, matching the saving time step of the LES data. We assess the performance of the various forms (3.1) - (3.4) in the following sections. The choice of the macroscopic environment variable $X_i(t)$, used in (3.2) and (3.4), 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 parameterization purposes in a GCM. In Section 3.9.2, we explain how we deal with heating and moistening in a single-column model experiment.

3.6 Markov chains

We start by using the simplest form (3.1), i.e., the form where the Markov chain is not conditioned on macroscopic environment variables or on neighbor 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:

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

Note that in this matrix we display probabilities rounded up to two decimal places, while we keep calculating with higher precision. 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.

Filling fractions of the Markov chain

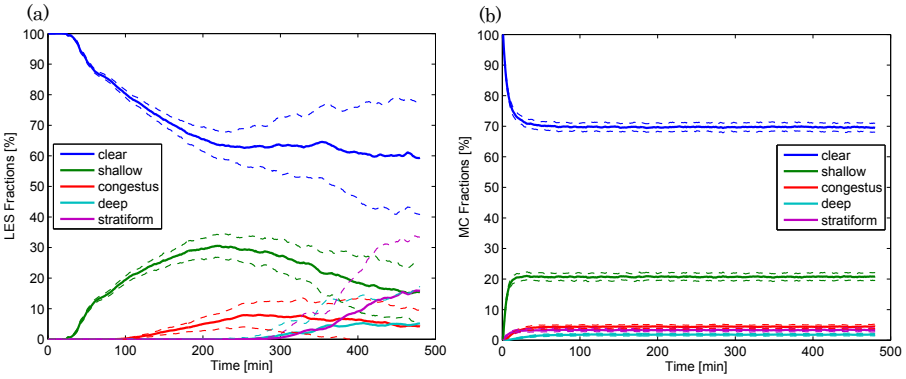

Figure 3.4: (a) Mean filling fractions observed in the LES data using $n = 32^2$ micro lattice sites per macroscopic block (solid) plus and minus the standard deviations over the 12^2 macroscopic blocks (dashed) and (b) reproduced mean filling fractions using 1,024 MCs (solid) plus and minus the standard deviation of 144 realizations (dashed).

Fig. 3.4 shows cloud filling fractions observed in the LES data and reproduced by the MC. The MC 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. 3.4, we can conclude that a MC governed by (3.1) is not capable of emulating the LES cloud fractions satisfactorily. A longer time step (20 minutes) for the MC 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 neighbor states. A natural way to improve on this is to include dependency on environment or neighbors. Thus, in the next sections we generalize the MC (3.1) by:

1. conditioning on the macroscopic state (environment), leading to the CMC form (3.2), or

2. coupling to neighboring cells, leading to the SCA form (3.3).

In the most general form (3.4), both environment and neighboring states are included. A schematic overview of these generalizations was shown in Fig. 3.1.

3.7 Conditional Markov chains

In this section, we explore conditioning of the MCs on a function of some large-scale variables that could be resolved in a GCM. Large-scale variables such as CAPE, CIN, middle troposphere RH, or (moist) convergence are considered to be potential indicators of convective behavior. In Section 3.7.1, 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 [74] and [72]. A reversibly lifted adiabatic parcel, using the mean thermodynamic properties at the 200-400 m 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 [138]. 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 $4,500$ J/kg, are higher than what we had expected, seasonally averaged values as high as $7,000$ J/kg have been reported over tropical land masses by [120].

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 [7, 45, 92]. 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 [36] and [81] where clustering has been used to construct CMCs.

In Fig. 3.5, 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 3.7.2). 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 (3.2) 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 [25]). To estimate the probability $p_\gamma(\alpha,\beta)$ defined in (3.2) we use the estimator:

$$\hat{p}_\gamma(\alpha,\beta) = \frac{T_\gamma(\alpha,\beta)}{\sum_\beta T_\gamma(\alpha,\beta)}, \tag{3.8}$$

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,

$$T_\gamma(\alpha,\beta) = \sum_{t,i} \mathbf{1}(Y_i(t+\Delta t) = \beta)\mathbf{1}(Y_i(t) = \alpha)\mathbf{1}(X_i(t) = \gamma) \tag{3.9}$$

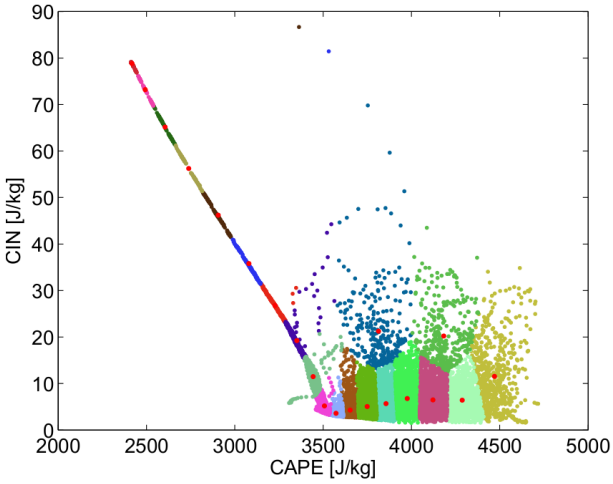

Figure 3.5: Clustered paths forming $K = 20$ regions in the CAPE-CIN space. The red dots are cluster centroids.

Mutual information between environment and cloud type

Large-scale variables such as CAPE, CIN or middle troposphere RH are considered to be potential indicators of convective behavior. 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

Large-scale variable(s)	Information
RH at 2,345 m & CIN	0.0992
RH at 2,345 m & w at 413 m	0.0948
CAPE & CIN	0.0946
CAPE & w at 413 m	0.0897
CIN & w at 413 m	0.0809
CIN	0.0757
RH at 2,345 m & CAPE	0.0710
w at 413 m	0.0697
CAPE	0.0589
RH at 2,345 m	0.0590
u at 15,843 m	0.0290

Table 3.2: Mutual information between large-scale variables and cloud type at 4.8×4.8 km^2 subdomains.

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 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 [24].

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 MC for Y_i on X_i. In [99], 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 data set.

In Fig. 3.6, 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 3.2, 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 3.2 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 (w) 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 behavior of convective organization 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 organization of deep convection in the present case is given in [17].

As a final remark, we have included the mutual information of u at 15,843 m in Table 3.2 as a consistency check: u at 15,843 m 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.

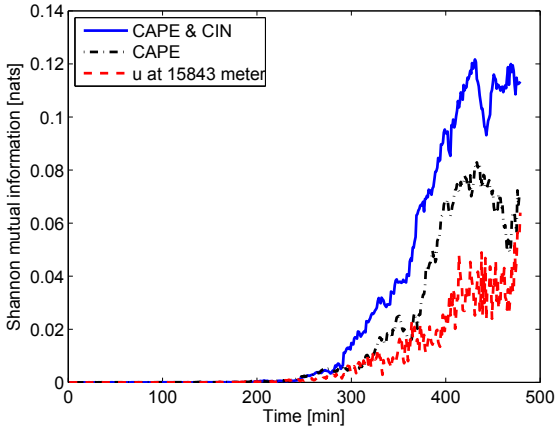

Figure 3.6: Time series of the mutual information between the large-scale variable at time t and cloud type at time t for different large-scale variables.

Filling fractions of the CMC

Fig. 3.7 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 1,024 CMCs per block. The time evolution of the means is in good agreement with the LES results, as can be seen by comparing with Fig. 3.4a. 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 Fig. 3.7b, we show the means and standard deviations using only 64 CMCs per macroscopic block. As expected, by using only 64 instead of 1,024 CMCs, the standard devia-

tions are larger and therefore in better agreement with the LES fractions. In Fig. 3.8, we show cloud filling fractions on a single macroscopic block; in Fig. 3.8a, the fractions of the LES data on a block of size $n = 1,024$; in Fig. 3.8b, the fractions as produced by the multicloud model using 64 CMCs (conditioned on CAPE-CIN). We see 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 8 h 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.

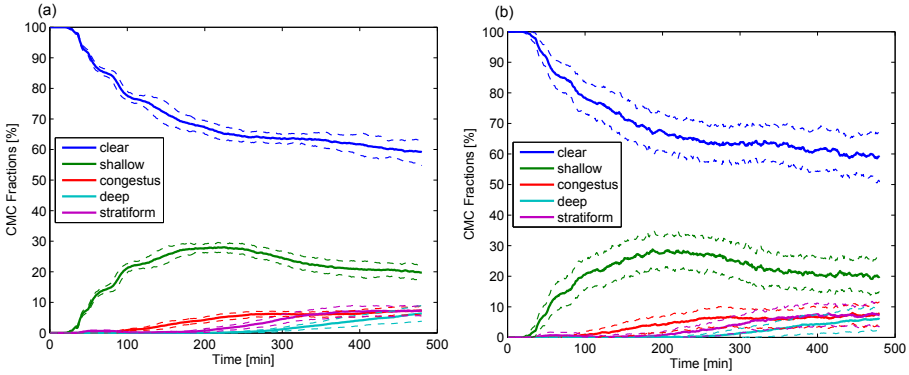

Figure 3.7: (a) Mean filling fractions produced by 1,024 CMCs with $K = 20$ clusters of CAPE and CIN (solid) plus and minus the standard deviation (dashed). The CMC is driven by LES observed values of CAPE and CIN. (b) Same as left but with 64 CMCs.

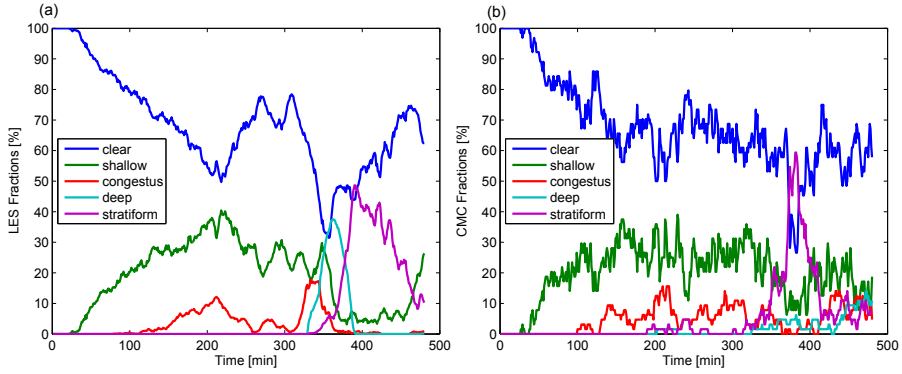

Figure 3.8: (a) Filling fractions observed in a single macroscopic block of $n = 32^2$ LES columns. (b) Filling fractions using 64 CMCs, where each CMC is conditioned on CAPE and CIN with $K = 20$.

3.8 Stochastic cellular automaton

In Section 3.7.2, it was shown that conditioning of the MC on the macroscopic environment strongly improves the behavior of the filling fraction means, cf. Fig. 3.4 and Fig. 3.7. 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 neighboring sites on the micro lattice can improve the emulated variances, without reducing the number of Markov chains. Thus, we study use of the forms (3.3) and (3.4) for the MC. We expect that, by coupling to neighboring sites, the spatial correlations of the cloud type patterns will be better captured, thereby increasing the variance.

As mentioned earlier, by conditioning the MC for lattice site i on the state of the neighboring sites, as in (3.3), the MC becomes a SCA. Cellular automata (CA) have been used for parameterization purposes by [8, 11, 124]. In these studies, the CA have deterministic rules, not stochastic ones, and they are chosen by intuition rather than inferred from data. Also, in those studies [8, 11, 124], the cells of the CA can take on two states, not five as is the case here.

First, we estimate the SCA transition probabilities (3.3) 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 (3.3) implies that in principle, for every state δ of the combined neighboring 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 neighboring shallow cumulus sites increases.

For the neighborhood of site i, denoted $\{i\}$, we choose the eight sites directly surrounding site i in the micro lattice (see Fig. 3.1). As each site can take on five 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:

$$p_\delta(\alpha,\beta) = \text{Prob}(Y_i(t+\Delta t) = \beta \,|\, Y_i(t) = \alpha,\ f(Y_{\{i\}}(t)) = \delta) \tag{3.10}$$

rather than (3.3) 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:

$$f(Y_{\{i\}}(t)) = 1 * |SH|_i + 2 * |CO|_i + 3 * |DE|_i + 4 * |ST|_i\,. \tag{3.11}$$

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 neighboring sites in a state of convection the larger the value of f. Furthermore, a neighboring site with cloud type congestus increases f more than a neighboring site with cloud type shallow. The function increases even more if there is a neighboring 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 [99] for some ideas on this), however we will not pursue this here. Estimating the probabilities (3.10) is straightforward, using an estimator analogous to (3.8)-(3.9).

We obtain 33 different transition matrices of size 5×5, because $0 \leq f \leq 32$. For each site, the state of the neighborhood is determined by counting the numbers of different cloud types surrounding it, and computing the corresponding value of $f_i(t)$ as in (3.11). This value determines which transition matrix is used at lattice site i at time t.

We initialize 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. 3.9a.

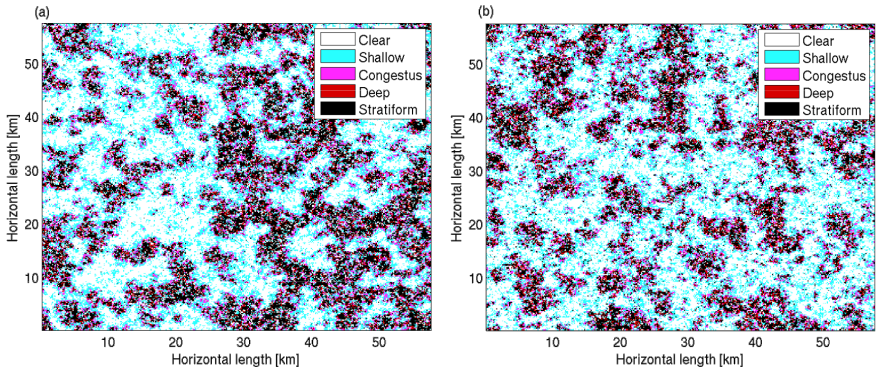

Figure 3.9: Patterns formed (a) by SCA at $t = 480$ and (b) by CSCA additionally conditioned on CAPE using $K = 5$ clusters.

Fig. 3.10a shows filling fractions (mean and standard deviation) for the SCA, using (3.10)-(3.11). The standard deviation is taken over macroscopic blocks of size $n = 1,024$. Both the time evolution and the magnitude of the standard deviations are in much better agreement with the LES data (Fig. 3.4a) than those produced by the CMC (Fig. 3.7). 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 MC both on the macroscopic state $X_i(t)$ and on

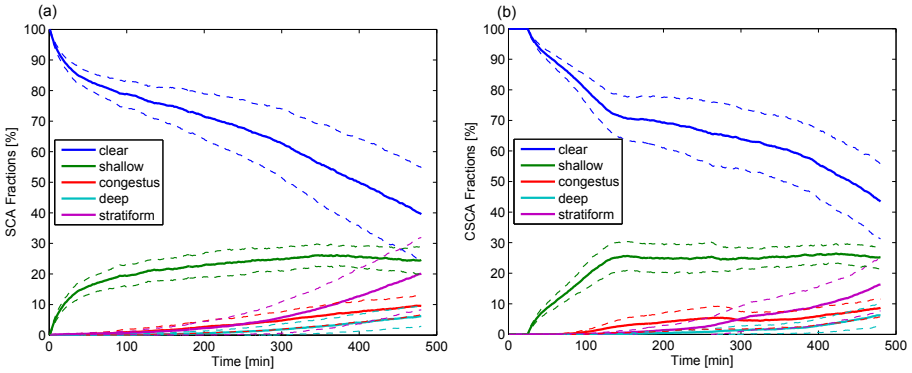

Figure 3.10: (a) Mean filling fractions of the SCA (solid) plus and minus the standard deviation calculated over blocks of 1,024 cells (dashed) and (b) the same for a CSCA conditioned on CAPE using $K = 5$ clusters.

the neighboring 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.

The filling fractions of the CSCA are shown in Fig. 3.10b. As before, we used the function (3.11) rather than $Y_{\{i\}}(t)$ to condition the CSCA on the neighboring sites. Thus, the transition probabilities are as in (3.4), but with $Y_{\{i\}}(t)$ replaced by the function (3.11). For conditioning on the macroscopic state $X_i(t)$ we used CAPE, clustered with five centroids. The patterns are similar to the patterns of the SCA; compare the panels of Fig. 3.9. 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 [99], and with methods to reduce the parameter space as in [81]. We leave this for future study.

3.9 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 behavior of a single GCM vertical model column. We use one macroscopic block, containing 1,024 CMCs, to represent the GCM model column. These CMCs are conditioned on CAPE and CIN, as in Section 3.7. 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 [72] 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 (3.6). The low level is at 413 m and the higher level is at 2,345 m 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 $X = (X_1, \ldots, X_4)^T$.

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

$$\text{CAPE}^* = \lambda X, \tag{3.12}$$

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:

$$\text{CIN}^* = e^{\mu X}. \tag{3.13}$$

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.

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 [36] 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:

$$\dot{X}_m = \sigma \nu_m X, \tag{3.14}$$

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 $1 \leq m \leq 4$:

$$\min_{\nu_m} \sum_t (\dot{X}_m - \sigma \nu_m X)^2, \qquad 1 \leq t < 480. \tag{3.15}$$

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 (3.15) 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.

Integration of the single-column model

We integrate Eq. 3.14, 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 3.9.2. We recall that the CMCs are conditioned on CAPE* and CIN*.

Filling fractions of the SCM

We test the stochastic multicloud model in the SCM. In Fig. 3.11, we show filling fractions for SCM using 1,024 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 1,024 LES columns). Comparing Fig. 3.4a to Fig. 3.11, 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. Recall that the SCM has been constructed from LES data prior to integration.

Using a smaller number of MCs (64 instead of 1,024) increases the variance of the filling fractions in the SCM test, as can be seen in Fig. 3.11b. 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.

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.

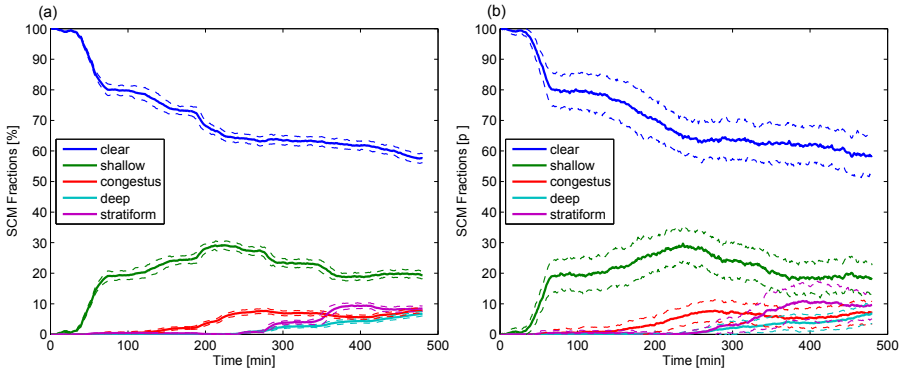

Figure 3.11: (a) Mean filling fractions produced in the SCM using 1,024 CMCs conditioned on CAPE* and CIN* (solid) plus and minus the standard deviations (dashed) and (b) the same using 64 CMCs.

As in [72], 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 behavior 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. 3.12, we plot time series for the prognostic variable X_4 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 2,500 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 MCs used. With a large number of MCs, the system behaves very regularly. For smaller n, the multicloud model is more stochastic, and the SCM-CMC model displays more variability.

3.10 Discussion and conclusion

In this paper, we combined, for the first time, the data-driven approach to stochastic parameterization from [25] and [36] with the stochastic multicloud model approach proposed in [72]. We used data from a convection-resolving LES model to infer a multicloud model similar to the one studied in [72]. The aim was to formulate a stochastic model that was able to emulate the coarse-grained convective behavior 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 cumulus, congestus, deep and stratiform. The coarse-grained convective behavior 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 (MC) makes random transitions between cloud types at each gridpoint, in accordance with transitions probabilities that are estimated from the LES data. A straightforward MC was not able to reproduce the correct evolu-

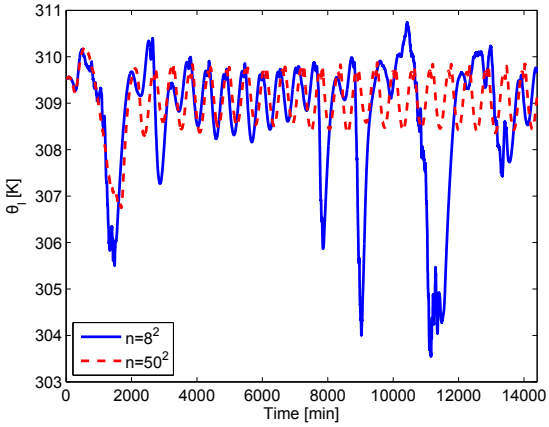

Figure 3.12: Time series for $X_4 = \theta_l^{\text{high}}$ in the single-column model integrated over ten days.

tion of the filling fractions corresponding to the five cloud types. Therefore, we explored two ways of improving the skills of the MC. First, by conditioning the Markov chain on large-scale variables, obtaining a CMC; second, by conditioning on the neighboring 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 neighboring 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 data set. Dividing the LES domain into subdomains, as was done here (as well as in [36]), 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 3.7, one can increase the sample variance even more by using data from multiple LES runs with different initial conditions.

We focused on a setting in which shear in the horizontal plane and spatially

varying terrain type have not been considered. In the case of a unidirectional shear with varying strength, the transition probabilities of the SCA may have to depend on the neighboring cells in an anisotropic way. The question of 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.

3.11 Acknowledgment

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.

Chapter IV
A multicloud model inferred from observational data

4.1 Abstract

Observational data of rainfall from a rain radar in Darwin, Australia, are combined with data defining the large-scale dynamic and thermodynamic state of the atmosphere around Darwin to develop a multicloud model based on a stochastic method using conditional Markov chains. The authors assign the radar data to clear sky, moderate congestus, strong congestus, deep convective, or stratiform clouds and estimate transition probabilities used by Markov chains that switch between the cloud types and yield cloud type area fractions. Cross-correlation analysis shows that the mean vertical velocity is an important indicator of deep convection. Further, it is shown that, if conditioned on the mean vertical velocity, the Markov chains produce fractions comparable to the observations. The stochastic nature of the approach turns out to be essential for the correct production of area fractions. The stochastic multicloud model can easily be coupled to existing moist convection parameterization schemes used in general circulation models.

4.2 The cumulus parameterization problem

The representation of clouds and convection is of major importance for numerical weather and climate prediction. Moist convection, also called cumulus convection, transports heat, moisture and momentum vertically in the atmosphere, it influences dynamical, thermodynamical and radiative processes and it has an impact on the large-scale global circulation. In general circulation models (GCMs), moist convection can not be explicitly resolved since the scale of the involved processes is too small, therefore the subgrid processes have to be represented by parameterizations, which are formulations of the statistical effects of the unresolved variables on the resolved variables. We refer to [3] for an overview of the the cumulus parameterization problem. Formulating moist convection parameterizations is a difficult problem: it introduces uncertainties in model predictions (e.g., [117]) and although models do agree that the cloud feedback is positive or neutral, they do not agree on the strength of the cloud feedback, e.g., [40]. It has been shown by [85] that the intraseasonal variability of precipitation is generally too small in models and that convectively coupled tropical waves are not well simulated.

An important issue considering cumulus parameterizations is that it is still not

This chapter has been published as Jesse Dorrestijn, Daan T. Crommelin, A. Pier Siebesma, Harmen J. J. Jonker, and Christian Jakob, 2015: Stochastic Parameterization of Convective Area Fractions with a Multicloud Model Inferred from Observational Data. *J. Atmos. Sci.*, **72**, 854–869.[35]

known which large-scale resolved variables are most strongly related to moist convection, and on which variables the closures of the parameterizations should be based. In general we have the choice between dynamical (e.g., vertical velocity) or thermodynamical (e.g., the convective available potential energy (CAPE), relative humidity (RH)) variables, which have been studied in a recent paper by [29]. Another important issue is that if parameterizations are chosen to be *deterministic* functions of the resolved variables, the subgrid response of moist convection to large-scale variations can not cover the variety of responses that is possible in reality, as deterministic parameterizations can only provide the expected value of the response of moist convection in a grid box. In view that GCMs resolutions are getting finer and finer, this issue becomes more important, because with smaller grid boxes the fluctuations around expected subgrid responses become larger. [109] pointed out that neglecting subgrid variability can result in model errors and that this can be corrected by using *stochastic* parameterizations to represent subgrid processes. This has for example been shown by [20] who improved the skill of numerical weather prediction (NWP) with the European Centre for Medium-Range Weather Forecasts's system by introducing stochastic elements in the physical parameterization tendency. Their pioneering work gave impulse to develop more sophisticated stochastic schemes.

Instead of perturbing all subgrid processes at once, it is possible to improve GCMs by introducing stochastic elements only in the deep convection parameterization, e.g., [9, 86, 88, 114, 136], or in the shallow convection parameterization, e.g., [123].

Rather than relying on physical intuition or deriving parameterizations from first principles, stochastic parameterizations can be inferred directly from data. [25] showed that Markov chains, with only a few states, for which the transition probabilities had been estimated from data, could represent the subgrid terms in the Lorenz '96 [91] model quite well, better than the determinstic parameterizations and the stochastic parameterizations, based on autoregressive processes, of [144]. The data-driven Markov chain model inspired [81] to develop a similar model based on cluster-weighted Markov chains. In [36] the Markov chain model of [25] was used to study stochastic parameterization of shallow convection and in [34] for deep convection.

A promising class of moist convection parameterizations based on the idea of evolving an ensemble of several (convective) cloud types, inspired by [96] and [65], is formed by *multicloud models*, e.g., [43, 72, 73, 94, 113]. The clouds follow a life cycle starting from clear sky to *congestus clouds*, to *deep cumulus* towers with *stratiform* anvil clouds as a remnant of the towers spreading over large areas, finally dissolving and come full circle at clear sky. In the multicloud model of [34] also *shallow cumulus* clouds are included.

In the present paper we use high-resolution ($\sim 2.5 \times 2.5$ km^2) observational data of rainfall in combination with data defining the large-scale ($\sim 150 \times 150$ km^2) dynamical and thermodynamical state of the atmosphere to infer such a stochastic multicloud model. The large-scale data are NWP analysis variable estimates improved with observations. The model is similar to the multicloud model of [34] in which Large-Eddy Simulation (LES) data was used to infer the model, as opposed

to the observational data of this study. The multicloud model produces area fractions for several cloud types which can be used as stochastic parameterizations in the deep convection and cloud schemes of GCMs. We also determine which large-scale variables are strongly related to deep convection.

In a late stage of the present study we became aware of work on stochastic parameterization of deep convection that is similar to our work [50]. Their stochastic models inferred from large-scale observational data also yield convective area fractions.

Our paper is organized as follows. In Section 4.3 we explain how we use Markov chains as a foundation for our multicloud model. Then, in Section 4.4 we give a description of the observational data, explain how we classified the data into cloud categories and how we dealt with advection while estimating transition probabilities between cloud states. In Section 4.5 we assess the skill of large-scale variables as indicators for deep convection. In Section 4.6 we construct our model, give expected area fractions and standard deviations and we discuss scale adaptivity, i.e., the ability to adapt to the size of a GCM grid box. We give results in Section 4.7 by comparing area fractions from the model with the observations and looking at their autocorrelation functions. In Section 4.8 we discuss the possibilities of implementation of the stochastic model in a convection parameterization of a GCM and make some concluding remarks.

4.3 Markov chains

The multicloud model we use in this study consists of Markov chains positioned on the nodes of a two-dimensional micro-grid. This model set-up has been used before in [34, 72, 113]. The state of each Markov chain at time t is denoted $Y_n(t)$, where n is the micro-grid index. Each Y_n can take on five different values, corresponding to the following categories: clear sky, moderate congestus, strong congestus, deep convective and stratiform. The choice of these specific categories will be discussed in Section 4.4. We will refer to these categories as *cloud types*. As time evolves, the Markov chains can switch, or "make a transition", between states every $\Delta t = 10$ minutes. All the Markov chains on the micro-grid together determine the area fractions σ_m for the various cloud types:

$$\sigma_m(t) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}[Y_n(t) = m], \tag{4.1}$$

in which $\mathbf{1}$ is the indicator function ($\mathbf{1}[A] = 1$ if A is true, 0 otherwise), N is the number of micro-grid nodes, and $m \in \{1, \ldots, 5\}$ the cloud type. We use radar data to estimate the transition probabilities, needed in the Markov chain model.

When used in a GCM, each GCM column contains N Markov chains that can switch to a different state every ten minutes, resulting in time-evolving area fractions σ_m for each cloud type and for each GCM column. These area fractions can be used in the convection and cloud schemes of a GCM. For example, the deep convective area fractions, σ_4, can serve as a mass flux closure at cloud base for a deep convection parameterization scheme:

$$M_b = \rho \sigma_4 w_{cb}, \tag{4.2}$$

in which ρ is the density and w_{cb} is the vertical velocity in a deep convective updraft at cloud base [4, 100]. More examples of possible applications in GCMs are given in Section 4.8.

As mentioned before, we use Markov chains with five possible states, so that the transition probabilities form a 5×5 transition matrix. Since these transition probabilities depend strongly on the large-scale state of the atmosphere, we make these probabilities conditional on functions of large-scale variables (i.e., the variables that are normally resolved by GCMs). These functions are called *indicators* of deep convection. In Section 4.5 we discuss appropriate indicators. The framework of conditional Markov chains (CMCs) for parameterization was introduced by [25].

For now, we consider a discretized indicator X, such that the possible states of X correspond to a finite number Γ of large-scale states. So, for each $\gamma \in \{1, \ldots, \Gamma\}$ we estimate a 5×5 transition probability matrix. The probability of CMCs switching from state α to state β given the large-scale state γ can be estimated as follows (see also [25]):

$$\text{Prob}(Y_n(t+\Delta t) = \beta | Y_n(t) = \alpha, X(t) = \gamma) = \frac{T_\gamma(\alpha,\beta)}{\sum_\beta T_\gamma(\alpha,\beta)} \tag{4.3}$$

where:

$$T_\gamma(\alpha,\beta) = \sum_{t,n} \mathbf{1}[Y_n(t+\Delta t) = \beta]\mathbf{1}[Y_n(t) = \alpha]\mathbf{1}[X_n(t) = \gamma]$$

counts the number of transitions observed in the data from cloud type α to β given that the large-scale state is γ. The indices n and t run over space and time covered in the *training data set* which is used to estimate the transition probabilities. We remark that we do not condition the Markov chains on $X(t+\Delta t)$, which reduces the number of matrices to estimate significantly. For the estimation of the transition matrices we use data sets corresponding to two different scales: data sets that are formed by high-resolution observations of rainfall at a scale that is equal to or smaller than the micro-grid scale of the CMCs and data sets that represent the large-scale atmospheric state at the grid scale of a GCM. In the next section we introduce the high-resolution observation data sets.

4.4 The radar data

The microscale data consists of observational data of precipitation obtained from the Darwin C-Band Polarimetric (CPOL) Radar in Darwin, North-Australia. This data is described in detail in [79]. In the same article it is explained how the radar data can be used to calculate cloud top height (CTH) and rain rates. For two time periods, 10 November 2005-15 April 2006 and 20 January 2007-18 April 2007, we have integer valued CTH and rain rate observations at ten-minute timesteps, for a circular area with radius 150 km and resolution of 2.5×2.5 km^2. In Fig. 4.1 we show a snapshot of the CTH and the rain rates at one time instance. The fields are rather noisy at the outer ring of the radar domain and the radar does not give observations in the center of the radar domain, which is known as the “cone of

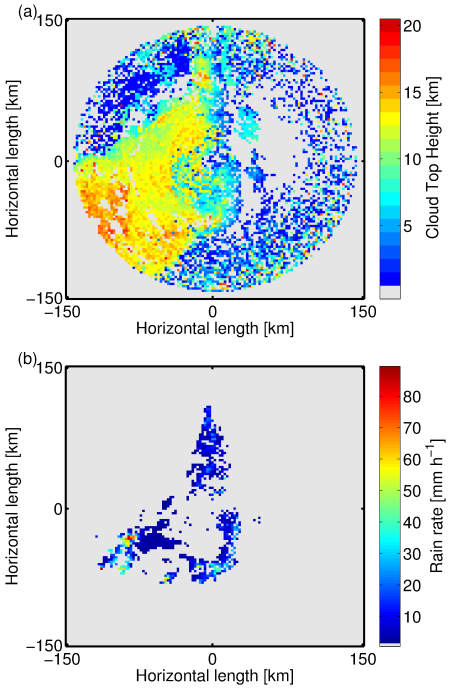

Figure 4.1: (a) A snapshot of the cloud top height derived from Darwin radar observations and (b) the corresponding rain rate.

silence" and is due to the 42° maximum elevation angle [98]. Therefore, we only use pixels in between 25 km and 97.5 km from the center of the domain. This forms an annular shaped subdomain consisting of 4,720 pixels of 2.5 × 2.5 km^2 corresponding to an area size of approximately 172 × 172 km^2. Fig. 4.2 contains histograms of the CTH and the rain rates, showing the distribution of these quantities. We consider CTH below 1.5 km as clear sky to avoid the influence of radar ground clutter. There is a bi-modal distribution of CTH, with a minimum at around 4 km, which is close to the freezing level at 5 km. To classify our cloud types, we use thresholds for CTH to distinguish high clouds, low clouds and clear sky. The bi-modal distribution in the cloud top histogram suggests a CTH threshold to distinguish low and high clouds (e.g., congestus and deep convective clouds) of around 4 or 5 km. Congestus clouds have been observed up to 9.5 km in the atmosphere [65]. We adopt the approach of [79], who developed a more objective identification of congestus and deep convective clouds, taking the value 6.5 km as a threshold. Further, we employ a rain rate threshold to make a distinction between clouds with intense precipitation and those with little or no precipitation. This enables us to make a distinction between deep convective clouds and stratiform clouds as well as a distinction between strong and moderate congestus. The rain rate histogram in Fig. 4.2b, shows an approximately exponential distribution, so it is impossible to

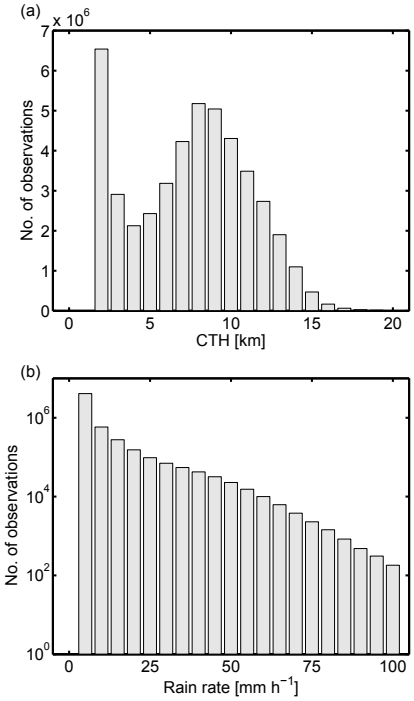

Figure 4.2: Histograms of (a) the cloud top height and (b) the rain rate observed with the Darwin radar in the periods November 2005 - April 2006 and January-April 2007.

argue for an obvious rain rate threshold. In the literature thresholds for partitioning convective and stratiform precipitation vary between 10 and 25 mm h^{-1}, and there are several methods for partitioning which are described in [82]. We choose a threshold of 12 mm h^{-1} to distinguish between deep convective and stratiform clouds and a threshold of 3 mm h^{-1} to distinguish between moderate and strong congestus. Combining these thresholds results in the following five cloud types: (1) clear sky, (2) moderate congestus, (3) strong congestus, (4) deep convective and (5) stratiform. In Table 4.1 we summarize the classification into cloud types. Note that, although desired, shallow cumulus clouds are not included in the model, for the obvious reason that the rain radar does not observe non-precipitating clouds.

After classification, we have two-dimensional fields with discrete values (integers from 1 to 5). In Fig. 4.3 we give an example of a classified field, which is the classified field corresponding to the CTH and rain rate fields shown in Fig. 4.1. After the classification *the observed area fractions*, σ_m, can be calculated according to (4.1), with Y_n the observed cloud type and $N = 4{,}720$ the number of radar pixels in the annular domain. The observed area fractions are strongly time-dependent, with σ_1 (clear sky) varying between 0% and 100%, σ_2 (moderate congestus) between 0% and 55%, σ_3 (strong congestus) between 0% and 2.5%, σ_4 (deep convective) ranging from 0 to about 10% and σ_5 (stratiform) ranging from 0 to about 99%.

Table 4.1: Cloud type classification using thresholds for the cloud top height and the rain rate.

CTH [km]	rain rate [mm h^{-1}]	
	≤ 12	> 12
≥ 6.5	stratiform ($m = 5$)	deep convective ($m = 4$)
	≤ 3	> 3
$\in [1.5, 6.5)$	moderate congestus ($m = 2$)	strong congestus ($m = 3$)
< 1.5	clear ($m = 1$)	

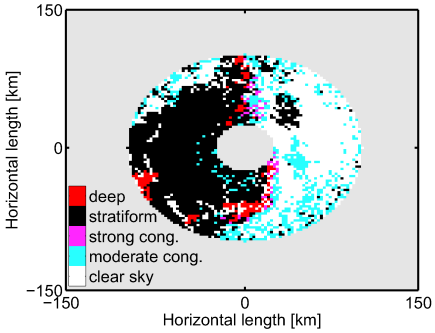

Figure 4.3: Example of radar data assigned to the categories clear sky, moderate congestus, strong congestus, deep convective and stratiform, corresponding to the CTH and rain rate snapshots of Fig. 4.1.

The observed fractions are depicted in Fig. 4.10 (discussed in Section 4.7) for a time period of five days for all cloud types, and the deep convective area fraction also in Fig. 4.7a (discussed in Section 4.7) for a longer period of three months.

Besides calculating observed area fractions for the different cloud types, the classified data are used to estimate transition probabilities between the cloud types for the CMCs, using (4.3). This is a key step in creating the multicloud model. To give an idea of the observed transition probabilities, not yet conditioned on the large-scale variables, we give the estimated transition matrix:

$$\hat{\mathbf{M}} = \begin{pmatrix} 0.8987 & 0.0668 & 0.0006 & 0.0011 & 0.0329 \\ 0.4147 & 0.4707 & 0.0033 & 0.0026 & 0.1086 \\ 0.2563 & 0.2686 & 0.2177 & 0.0545 & 0.2029 \\ 0.1757 & 0.0284 & 0.0124 & 0.4295 & \mathbf{0.3540} \\ 0.1185 & 0.0779 & 0.0010 & 0.0091 & 0.7935 \end{pmatrix}$$

The probability of a transition from cloud type m to cloud type n can be found in the nth column of row m. For example, the probability that a deep convective pixel will be assigned to stratiform ten minutes later, is **0.3540**. The probability that a deep site is again a deep site ten minutes later, is 0.4295, much larger than the expected deep convective area fraction (at most 0.03 as can be seen Fig. 4.6,

discussed later in this paper). This is comparable to the deep to deep transition probability of 0.5602 estimated from the LES data set of [34]. Most remarkable is that the stratiform decks in the LES data tend to dissolve faster than observed in the radar data. The transition probability for stratiform to stratiform is estimated 0.2266 in LES , as opposed to 0.7935 observed in the radar data. Some evidence for the life cycle can be seen in the transition matrix, a deep convective cloud likely turns into stratiform, which turns into clear sky. Some entries are artefacts of the estimation method, for example the probability of clear sky turning into stratiform is 0.0329, but in reality the stratiform cloud spreads out from the top of a deep cumulus cloud.

For correct estimation of cloud type transition probabilities, we have to take into account that clouds are advecting horizontally through the domain. To do this, we translate the advected clouds in a radar image back to their position in the previous image. In this way, we minimize transitions that are only a result of advection. The advection, with zonal wind u and the meridional wind v, is assumed to be a function of height and time only. We calculate this translation separately for every cloud type (as they are located at different heights in the atmosphere). Let $Z_m(x_i, y_j, t) = \mathbf{1}[Y(x_i, y_j, t) = m]$, with $Y(x_i, y_j, t)$ the discretized radar pixel at location (x_i, y_j) at time t and (x_i, y_j) running over all $N_{ij} = 4,720$ pixels in the annular shaped subdomain. We calculate for every cloud type m and for every time interval $[t, t+\Delta t]$ the optimal horizontal displacements $u_m \Delta t$ and $v_m \Delta t$ which maximize the correlation

$$\frac{1}{N_{ij}} \sum_{ij} Z_m(x_i, y_i, t) Z_m(x_i + u_m \Delta t, y_j + v_m \Delta t, t + \Delta t).$$

By applying the Correlation Theorem (e.g., [115]), fast Fourier transforms can be used to reduce the calculation time for finding the displacements. At the boundaries at the outer edge and in the center of the radar domain, clouds flow into and out of the domain. We also have to account for this during the estimation of cloud type transition probabilities. More specifically, we do not count transitions of "clouds" (including clear sky) that are inside the radar domain at time t, but which are outside the domain at the previous time step $t - \Delta t$ or at the next time step $t + \Delta t$, due to advection. Without corrections, the estimated probability transition matrix is significantly different: for example the probability that a pixel assigned to the deep convective cloud type is deep convective ten minutes later would be estimated at 0.29 instead of 0.43.

The focus in this paper will primarily be on the deep convective area fractions, when we determine the large-scale variable on which to condition the CMC (Section 4.5) and when we test the CMC (Section 4.8). Although the other fractions can have applications in GCMs, the deep convective area fractions are the most important. Describing the convective transport by deep convection accurately is crucial for a GCM to work properly. Conditioning each individual cloud type on different large-scale variables could improve the model, in particular for the strong congestus clouds, that precede deep convection.

4.5 The large-scale data

We have data available that defines the large-scale dynamic and thermodynamic state of the atmosphere around Darwin for the time periods November 2005-April 2006 and January 2007-April 2007 for which we also have the radar data. The large-scale fields are averages over six-hour intervals and have a vertical resolution of 40 pressure levels, from ground level to about 20 km altitude. The data has been prepared by [29] who used a variational analysis method to improve NWP analysis large-scale variable estimates by constraining the moisture budgets with observational rain data from the CPOL radar. The large-scale data is also used in [28]; [113] and [50]. Here, we use the data to investigate which large-scale variables are suitable indicators for the convective state of the atmosphere and compare our findings with the results of [29]. Then, we will use the large-scale data accordingly for conditioning the multicloud CMC model. As in [29], we consider thermodynamical and dynamical variables. In particular, we will consider the following well-known indicators: CAPE, the mean vertical velocity $\langle\omega\rangle$, and RH. CAPE is a measure for the stability of the atmosphere and is formally defined as follows:

$$\text{CAPE} := R_d \int_{p_{\text{NB}}}^{p_{\text{LFC}}} (T_{v,p} - \overline{T}_v) d\ln p\,,$$

in which $T_{v,p}$ is the virtual temperature of an undiluted parcel, $\overline{T}_v$ is the virtual temperature of the environment, R_d is the gas constant of dry air, p_{NB} the level of neutral buoyancy and p_{LFC} the level of free convection (e.g., [126]). The mean vertical velocity we define as:

$$\langle\omega\rangle := \frac{1}{p_0 - p^*} \int_{p^*}^{p_0} \overline{\omega}(p) dp,$$

in which $\overline{\omega}$ is the large-scale vertical velocity in hPa h^{-1}, p_0 the pressure at the surface, and p^* is pressure level 340 hPa, chosen because the resulting $\langle\omega\rangle$ gives the highest correlation with deep convective area fractions (as calculated with (4.4) that is given below). We find that the vertical integral over $\overline{\omega}$ gives higher correlations than $\overline{\omega}$ at a single pressure level. Further, the relative humidity is chosen at pressure level 640 hPa, also because it gives the highest correlation with deep convective area fractions. To assess how well an indicator correlates with deep convection, we calculate the time-lagged cross-correlation function (CCF) of the indicator and the deep convective area fraction.

Given the timeseries of the deep convective area fraction $\sigma_4(t)$ and the timeseries of the indicator $X(t)$, the normalized CCF of $X(t)$ and $\sigma_4(t)$ is:

$$\text{CCF}(\tau) = \int_{-\infty}^{\infty} \tilde{X}(t+\tau)\tilde{\sigma}_4(t) dt \tag{4.4}$$

with $\tilde{X}(t) = \frac{X(t)-\mu_X}{\sigma_X}$ (i.e., the indicator normalized by subtracting its mean μ_X and dividing by its standard deviation σ_X), $\tilde{\sigma}_4$ defined analogously, and τ the time lag of X w.r.t. σ_4. As such, the CCF lies in between -1 and 1. If the maximum value of the CCF is attained at positive time lag τ, the indicator $X(t)$ tends to *follow* rather than *precede* deep convection.

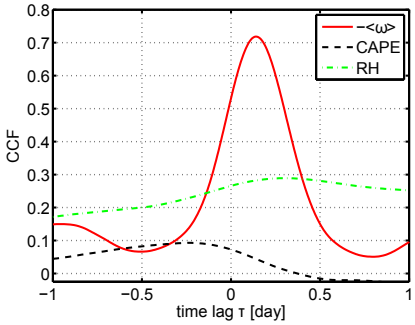

Figure 4.4: Cross-correlation functions (CCFs) of the deep convective area fraction with -$\langle\omega\rangle$, CAPE and RH at 640 hPa for the 2005/2006 data set.

In Fig. 4.4 we plot the CCFs of the indicators $-\langle\omega\rangle$, CAPE and RH with the observed deep convective area fraction for the 2005/2006 period. The figure for the 2007 period is similar (not included). Before calculating the CCF, we linearly interpolate X to get its values every ten minutes instead of every six hours, because the sequences X and $\tilde{\sigma}_4$ must have the same length. We see that $\langle\omega\rangle$ has a larger correlation at zero time lag than CAPE and RH. Moreover, also for negative time lags of a few hours this correlation is higher. In this respect $\langle\omega\rangle$ is the best indicator of deep convection. We note that the maximum correlation of $\langle\omega\rangle$ with σ_4 is attained at a positive time lag. This may seem to indicate that $\langle\omega\rangle$ is an effect rather than a cause of deep convection. However, this is a subtle issue, as $\langle\omega\rangle$ may also both be a trigger (i.e., cause) of deep convection and be reinforced by it, so that separating cause and effect becomes difficult. In [113] a related discussion can be found. For large-scale moisture and temperature advection we found correlations comparable to the correlation for $\langle\omega\rangle$ (not included in Fig. 4).

In order to use an indicator for constructing the CMC according to (4.3), it must be discretized into a finite number of states. If only one indicator is used, which is the case in this paper, a finite number (Γ) of intervals can be chosen, defined by thresholds. If a combination of several indicators is used, one can choose thresholds for each indicator separately, or use a clustering method as in [34, 36] and [81]. To give an example, in Fig. 4.5 we show a histogram of $\langle\omega\rangle$ discretized using 25 intervals. These intervals have been found by using a cluster method, k-means [45, 92], which minimizes the distance between the $\langle\omega\rangle$-values and the centers of the intervals. Using equidistant intervals is also an option, however, since the $\langle\omega\rangle$-values are not distributed uniformly, we prefer the non-equidistant intervals found by k-means. Interval number 25, corresponds to negative $\langle\omega\rangle$ or strongly positive large-scale vertical velocity (illustrated by the arrow), which is favorable for deep convection, and we will later see in Fig. 4.6 that the averaged observed deep convective and stratiform area fractions are large (around 3% and 90%, respectively) for interval number 25.

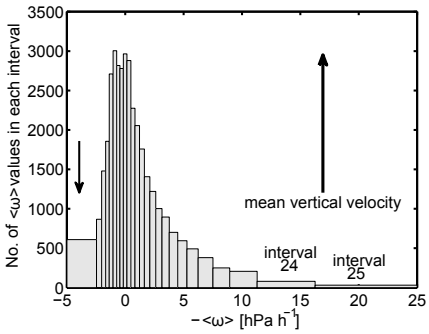

Figure 4.5: Histogram of the 25 intervals of -$\langle\omega\rangle$, found by clustering the linearly interpolated $\langle\omega\rangle$ values. The first and last (25th) intervals are open on one side. Because ω is a velocity in terms of pressure, positive $\langle\omega\rangle$ corresponds to downward mean large-scale motion and negative $\langle\omega\rangle$ to upward mean motion (as illustrated by the arrows).

4.6 A description of the multicloud model

Having classified the radar data into cloud types, and having identified (and discretized) a suitable large-scale indicator, $\langle\omega\rangle$, we estimate the transition probability matrices of the CMC using (4.3). We take the period from 10 November 2005 until 15 April 2006 as the training data set, and we set $\Gamma = 25$. So, we have to estimate 25 matrices each of size 5×5, giving 625 parameters in total. This may seem a large number, however the training data set is very large, containing $O(10^8)$ observations of transitions (radar images at ten-minute intervals during 157 days, with 4,720 pixels in each image).

In Section 4.7 we will validate the CMCs with the *test data set*, but since we have estimated transition matrices, we can already get some insight into the statistical properties of the cloud type area fractions generated by the CMC as compared to the observed area fractions in the *training data set*.

In Fig. 4.6, we plot the expected fractions and the standard deviation for both the observations and the CMC as a function of the $\langle\omega\rangle$-intervals seen before in Fig. 4.5. The expected values of the CMC correspond to the invariant distribution of the transition matrix for each $\langle\omega\rangle$-interval. The CMC expected values are almost equal to the observational expectations for all cloud types, the small differences can be ascribed to the way we corrected for horizontal advection (as described before in Section 4.4).

We see in Fig. 4.6a that the expected deep convective area fractions increase with increasing $\langle\omega\rangle$-interval (corresponding to increasing upward mean vertical velocities) and has its maximum of around 0.03 for interval number 24. Further, the strong congestus fractions in Fig. 4.6b, increase with increasing $\langle\omega\rangle$-interval, however, for interval number larger than 22, the fraction decreases rapidly, while expected deep and stratiform cloud fractions keep increasing. The expected stratiform fractions increase with increasing $\langle\omega\rangle$-interval up to very high expected values of 90%. The expected value of moderate congestus is around 15% for downward

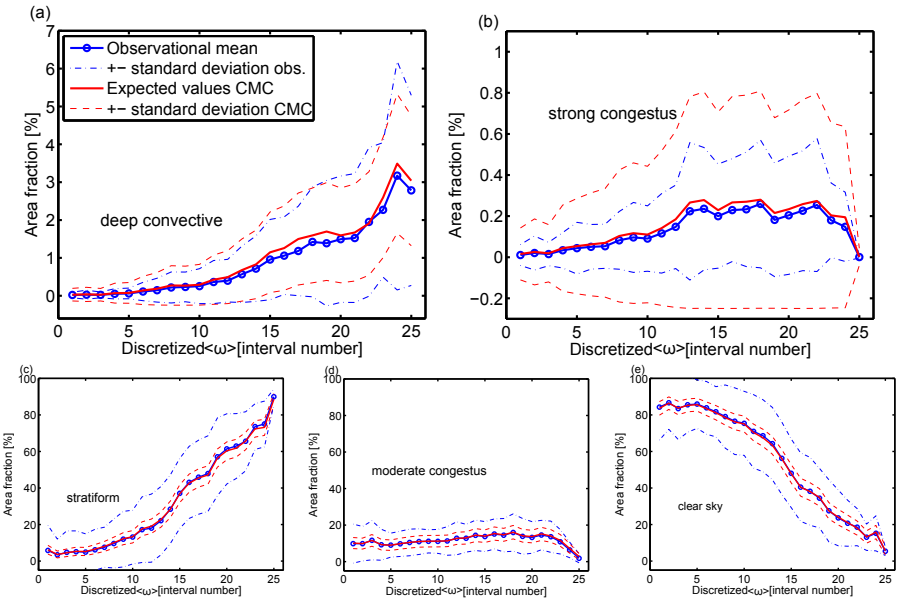

Figure 4.6: Observational mean cloud type area fractions as a function of the $\langle\omega\rangle$-intervals for the 2005/2006 training period (solid line with circles) plus and minus the standard deviation (dash-dotted line) and the CMC expected cloud type area fractions (solid line) plus and minus the standard deviation while using $N = 100$ CMCs (dashed line). Note the different scaling on the y-axis.

mean motion, increases slightly with increasing $\langle\omega\rangle$-interval number. For $\langle\omega\rangle$-interval numbers above 22, the expected value of moderate congestus decreases which is caused by the stratiform decks that are dominating the radar domain (for this $\langle\omega\rangle$-interval numbers). Expected clear sky fractions decrease rapidly as a function of the $\langle\omega\rangle$-interval.

The standard deviation of the observational deep convective area fractions tends to increase with increasing $\langle\omega\rangle$-interval number, so it tends to increase if the expected value increases and for high values of the $\langle\omega\rangle$-interval number the standard deviation is almost equal to the expected value. The standard deviation of the observational strong congestus area fractions depends on the expected values as well. The standard deviation of the observational stratiform area fractions tends to increase as a function of the $\langle\omega\rangle$-interval, but decreases if the expected values become very large because of the upper bound of 100%. For moderate congestus, the standard deviation ranges between 0.5 and 1 times the expected values. The standard deviation of the observed clear sky area fraction is around $10-20$%, independently of the $\langle\omega\rangle$-interval number, with an exception of interval number 25 for which the standard deviation is only 2.4%.

The standard deviation of a cloud type area fraction σ_m that is produced by N CMCs is defined as:

$$\sqrt{\mathrm{E}\left[(\sigma_m - \mathrm{E}[\sigma_m])^2\right]},$$

in which E is the expectation. One can derive that this is equal to $\sqrt{N^{-1}p(1-p)}$, in which $p = \mathrm{E}[\sigma_m]$ is the expected value of the fraction. Note that $\mathrm{E}[\sigma_m]$ is dependent on $\langle\omega\rangle$. So, the theoretical standard deviation depends only on the expected value of the fraction and the number of CMCs used to calculate the cloud type area fractions. We choose a value of $N = 100$ such that the standard deviation of the deep convective area fractions is comparable to the standard deviation of the observed deep convective area fractions in the training data set. This implies that the standard deviation of the fractions produced by the CMCs is too small for cloud types with larger observed standard deviations (clear sky, moderate congestus and stratiform) and too large for the strong congestus cloud type (which has a small observed standard deviation).

For the observational deep convective area fractions the *normalized* standard deviation, the standard deviation divided by the mean, is decreasing with increasing mean, with values decreasing from 5 down to about 1. So, we agree with the conclusion of [29] that noise (or stochastic behavior) decreases as a function of increasing forcing. This is also the case for the observational strong congestus area fractions, with a normalized standard deviation ranging from 1 (for relatively high fractions) up to 3 (for relatively low fractions).

Scale adaptivity

Ideally a parameterization of deep convection should be adaptive to the size of the GCM grid box, see [4]. By construction of the multicloud model, our parameterization of deep convection is indeed scale adaptive. The value N of the number of CMCs can be adapted to the horizontal grid spacing of the GCM. For a large size of the GCM grid box, a large number of clouds fit into the model column and therefore a large number of CMCs should be taken to calculate the cloud type area fractions. For very large GCM grids, the number of CMCs becomes very large and hence the σ_m tend to a deterministic limit (equal to the expected values associated with the large-scale interval number). For smaller grid box sizes, the number of CMCs is smaller and as a result, the area fractions generated by the multicloud model will be "more stochastic", fluctuating significantly around their expected values. It is difficult to say to which horizontal size a CMC corresponds exactly. The size corresponding to a CMC is equal to the typical horizontal size of the cloud type under consideration. Therefore, the horizontal size is larger than the area of a radar data pixel (2.5×2.5 km^2), which explains that producing area fractions with CMCs while using a number smaller than the number of radar pixels in the radar domain gives better results in Section 4.7, $N = 100$ versus $N = 4,720$. We emphasize that the value of $N = 100$ is found during the training phase and not during the the testing phase of the model.

4.7 Results

To assess how well the multicloud model reproduces the convective behavior observed in the radar data set, we first consider the cloud type area fractions. Then, we will look at autocorrelation functions (ACFs) of the fractions and $\langle\omega\rangle$.

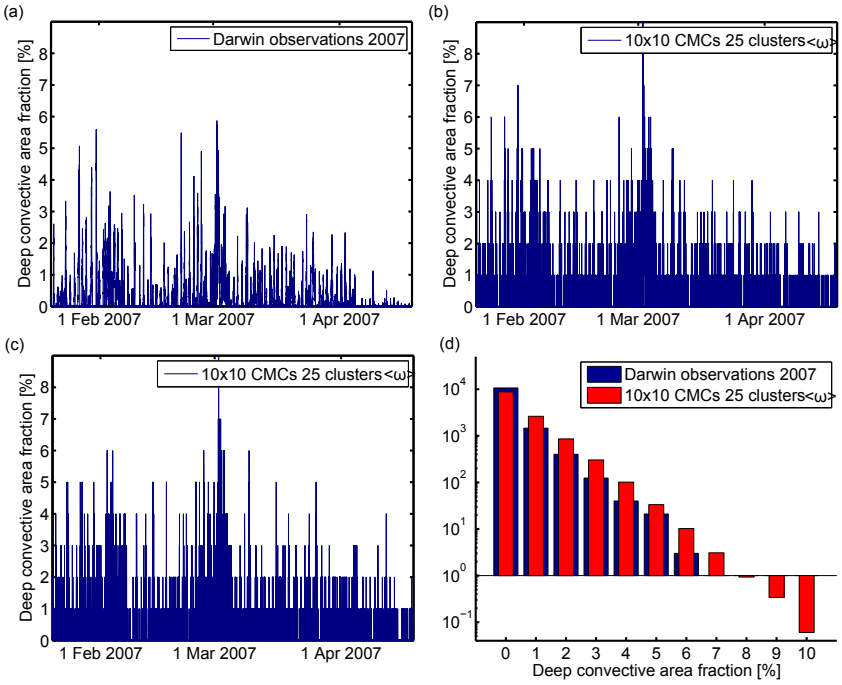

Figure 4.7: (a) Deep convective area fractions observed in Darwin (b,c) two realizations of deep convective area fractions produced by $N = 100$ CMCs conditioned on $\langle\omega\rangle$ and (d) the corresponding histograms comparing the CMC fractions (averaged over 100 realizations) with the observed fractions (binned into intervals) on a logarithmic y-axis.

Area fractions

As mentioned, the radar data can be used to calculate observed area fractions of each cloud type. We use $\langle\omega\rangle$ as indicator and take $N = 100$ CMCs. Then, we train the CMCs as explained in Section 4.6 using the training data set 2005/2006. We assess the model by driving the CMCs with $\langle\omega\rangle$ as observed in the other data set (from 2007). Thus, different data sets are used for training and evaluation.

In Fig. 4.7a we show the deep convective area fractions as observed in the Darwin radar test data set (2007). It can be seen that the deep convective events are very intermittent in the radar data, with periods of enhanced deep convection, periods with less wide-spread convective events and the deep convective area fraction is exactly zero in 52 % of the ten minute intervals. In Fig. 4.7b and 4.7c we give two realizations of the deep convective area fractions as reproduced by the CMCs. The CMC fractions display similar intermittent behavior, with maximum values that are slightly too high compared to the observations. The CMC fractions have discrete values, namely $\sigma_4 \in \{0, 0.01, 0.02, 0.03, \ldots\}$, because $N = 100$ CMCs are used. To further assess the quality of the deep convective fractions, we calculate histograms of the deep convective area fractions (Fig. 4.7d). Since the CMC fractions are integer multiples of 0.01, we bin the Darwin observed fractions into intervals

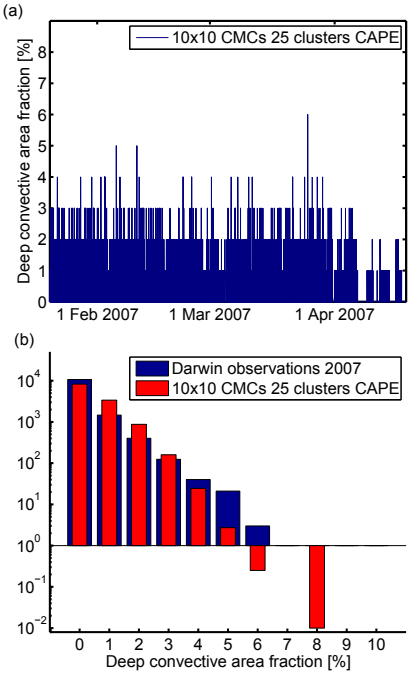

Figure 4.8: Deep convective area fractions produced by $N = 100$ CMCs conditioned on CAPE and (b) the corresponding histograms in which the CMC fractions (averaged over 100 realizations) are compared to the observed fractions (binned into intervals) on a logarithmic y-axis.

of length 0.01, apart from the first interval which is $[0, 0.005)$. Because high values of the deep convective fractions are rare, we plot the histograms on a logarithmic y-axis. We observe that the observational fractions decrease exponentially, as is expected since rain rates tend to decrease exponentially (see Fig. 4.2). The CMC fractions follow the exponential decrease well and the values are only slightly off.

We repeat the computations with CAPE as indicator instead of $\langle\omega\rangle$. In Fig. 4.8a we show the resulting CMC deep convective area fractions (compare to Fig. 4.7a). We observe that the fractions are also intermittent, but high fraction values are too rare. Further, although periods of enhanced convection and of less convective events are visible, they are not comparable with the observations. In the histograms with a logarithmic y-axis (Fig. 4.8b) it is indeed visible that fractions larger than 0.04 are too rare, although a fraction of 8% is reached in one of the 100 realizations. We conclude that in the present setting CAPE is less suitable as indicator for deep convection than $\langle\omega\rangle$.

As our third experiment, we use $\langle\omega\rangle$ again as indicator and keep everything as in the first experiment except for taking $N = 69^2 = 4{,}761$ which is (close to) the number of radar pixels used to train the CMCs. We observe (Fig. 4.9) that high values of the deep convective area fractions are not reached anymore, values are not higher than 0.04. Because N is much larger than before, the fractions are

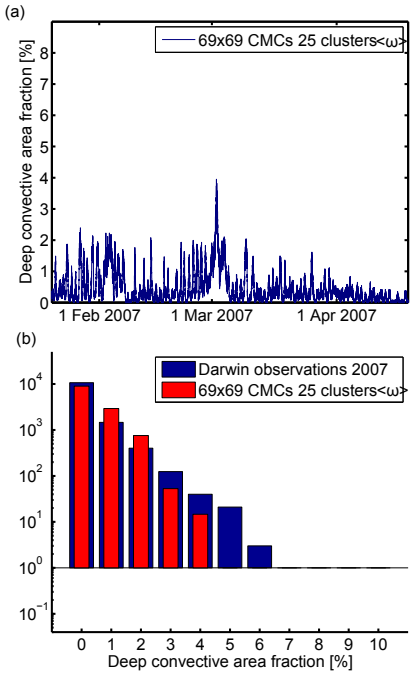

Figure 4.9: Deep convective area fractions produced by $N = 69^2$ CMCs conditioned on $\langle\omega\rangle$ and (b) the corresponding histograms of the binned CMC fractions averaged over 100 realizations compared to the binned observed fractions on a logarithmic y-axis.

rather close to the (deterministic) expectation values. This means that, although the number of CMCs is equal to the number of radar lattice sites, the CMC fractions show lower maxima. We note that in our current set-up the CMCs on the 2D micro lattice sites are independent of their lattice neighbors, which is not the case for the sites in the radar data. This is the underlying cause of the lower CMC maxima. Introducing local interactions between neighboring CMCs can improve this, but it makes the estimation of the CMCs much more complicated, see [34] and [71].

As a final experiment we take again $N = 100$ CMCs and $\langle\omega\rangle$ as indicator, but we interchange the roles of training data set and test data set. Thus, we train the CMCs with the 2007 data set and validate using fractions for the 2005/2006 period. The deep convective area fractions in the 2005/2006 radar data reach higher maxima than in the 2007 data set, with an overall maximum of about 10 percent (not shown). The fractions of the CMCs are less likely to attain these highest peak values. Notwithstanding this issue, the distribution of the CMC fractions is still comparable to that of the observed fractions.

For a more detailed look at the fractions, in Fig. 4.10 we show the area fractions of all five cloud types corresponding to the first experiment (with $N = 100$ and $\langle\omega\rangle$ as indicator) for a much shorter period of five days. The timing of the deep convective events produced by the CMCs is almost correct, although there is a small time

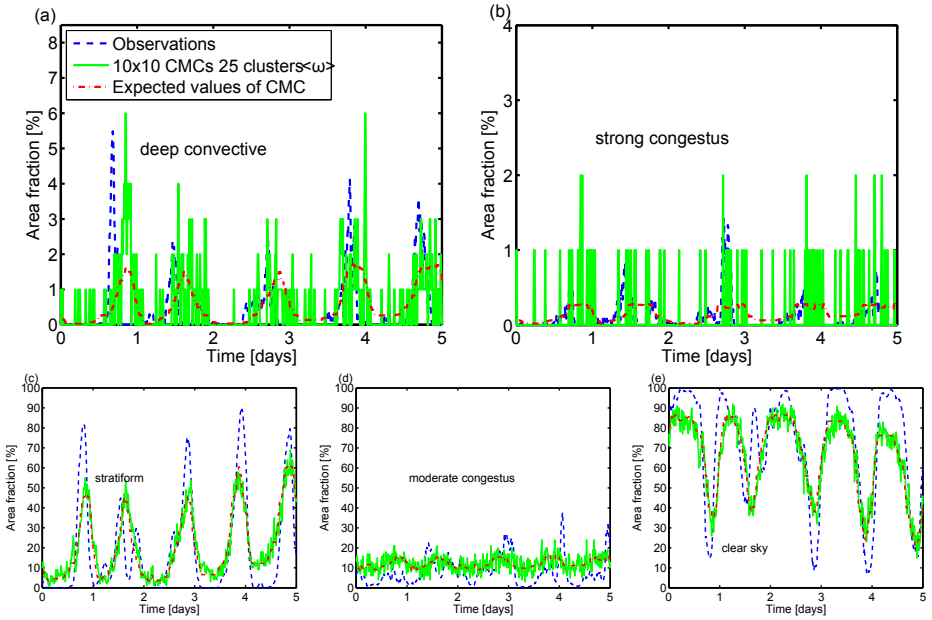

Figure 4.10: Area fractions of (a) deep convective, (b) strong congestus, (c) stratiform, (d) moderate congestus and (e) clear sky observed in Darwin (dashed line), produced by 100 CMCs (solid line) conditioned on $\langle\omega\rangle$ and the corresponding expected area fractions of the CMCs (dash-dotted line) for a period of five days. Note the different scaling on the y-axis.

lag visible in Fig. 4.10a. Furthermore, it is clear that the deep convective fractions of the CMC show maximum values of the peaks in agreement with the observations, which is not the case for the expected values of the CMC. The conclusion is that the stochastic fluctuations of the multicloud model fractions are needed in order to produce the correct maximum values of the deep convection area fraction peaks. The stochastic nature of the approach is essential for production of the correct area fractions. A day-night cycle can be seen in the deep convective fractions, owing to the presence of land in the radar domain. This cycle is also present in the CMC fractions.

The strong congestus fractions in Fig. 4.10b are small, so the CMC fractions, being integer multiples of 0.01, have difficulties attaining the observational fractions. So, $N = 100$ seems to be too small for the strong congestus area fractions. In Fig. 4.10c, we see stratiform area fractions. The CMC fractions follow the observations correctly (in a time sense), but the local maxima tend to be too low. The stochastic part of the fractions is not as prominent as for the deep convective area fractions. The observational moderate congestus fractions in Fig. 4.10d are difficult to follow for the CMCs: the value zero is never attained for the CMC fractions. A conclusion is that $\langle\omega\rangle$ is not such a good indicator of moderate congestus clouds. These depend probably more on boundary layer processes. The clear sky fractions (Fig. 4.10e) of the CMC follow the observations quite well, but the minimum values

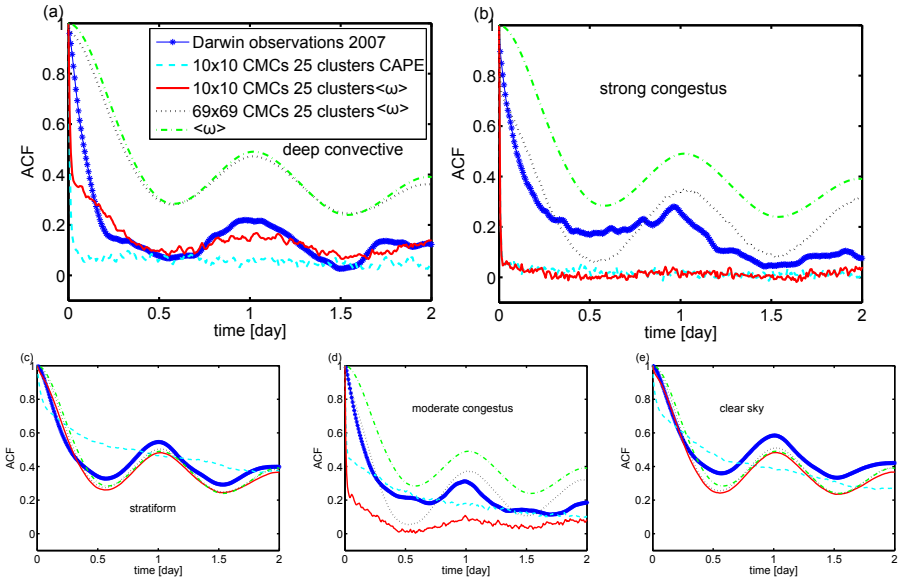

Figure 4.11: Normalized ACFs of the observational area fractions (solid lines with stars), the CMC area fractions with $N = 100$ conditioned on $\langle\omega\rangle$ (solid lines) and on CAPE (dashed lines), the ACF corresponding to 69^2 CMCs conditioned on $\langle\omega\rangle$ (dotted lines) for the cloud types (a) deep convective (b) strong congestus (c) stratiform (d) moderate congestus and (e) clear sky. Also the ACF of $\langle\omega\rangle$ is shown (dash-dotted lines).

are not small enough. The clear sky fractions are important, as $1-\sigma_1$ is the cloud cover observed by the radar, which is a usable quantity in GCMs, however, keep in mind that the radar is not able to detect all clouds.

Autocorrelation functions

As a final assessment in this paper, we inspect ACFs of the cloud type area fractions and $\langle\omega\rangle$. The ACF of the cloud type area fraction σ_m is:

$$\mathrm{ACF}(\tau) = \int_{-\infty}^{\infty} \tilde{\sigma}_m(t+\tau)\tilde{\sigma}_m(t)dt, \tag{4.5}$$

which is the CCF of $\tilde{\sigma}_m$ with itself, cf. (4). Recall that $\tilde{\sigma}_m$ is the normalized σ_m. The ACF of $\langle\omega\rangle$ is defined analogously. A main advantage of using Markov chains instead of drawing samples that are uncorrelated in time from the observed distribution of cloud types is that a Markov process should be better capable of capturing the observed ACF. In Fig. 4.11 we show normalized ACFs of the observed area fractions (solid line with stars), the CMC area fractions with $N = 100$ conditioned on $\langle\omega\rangle$ (solid line) and on CAPE (dashed line) and the ACF corresponding to 69^2 CMCs conditioned on $\langle\omega\rangle$ (dotted line), for (a) deep convective (b) strong congestus (c) stratiform (d) moderate congestus and (e) clear sky. Also the ACF of $\langle\omega\rangle$ is shown (dash-dotted line). In (a) we see that apparently, the ACF of the deep convective area fractions produced by $N = 100$ CMCs decreases too rapidly initially.

Without the correction for advection as explained in Section 4.4 the ACF decreases even more rapidly (not shown). The rapid initial decrease indicates that the probability of a transition from deep to deep is estimated too low. We see that the daily cycle is well captured in the case that we conditioned on $\langle\omega\rangle$. When CAPE is used as indicator the ACF decreases more rapidly than when conditioned on $\langle\omega\rangle$ and it can be seen that the daily cycle is not captured. The ACF for the observational data set of 2005/2006 is similar to the ACF for the 2007 data set (not shown). We note that for a large number of CMCs, close to the deterministic limit, the ACF follows the ACF of $\langle\omega\rangle$ almost perfectly. In (b), we see that in order for the CMCs to follow the observational strong congestus ACFs, the $N = 69 \times 69$ performs better than the $N = 10^2$. In (c) and (e) we see ACFs of the CMC, that are comparable to the observational ACF, only if conditioned on $\langle\omega\rangle$, not if conditioned on CAPE. The presence of a daily cycle in the fractions is clearly visible if conditioned on $\langle\omega\rangle$ except for strong congestus fractions produced with $N = 100$ CMCs. Considering all ACFs, we conclude that the ACFs for CMCs conditioned on $\langle\omega\rangle$ are better than if conditioned on CAPE (except for moderate congestus). For $N = 100$, the ACF of deep convection is better than for $N = 69^2$, while this is not the case for strong congestus and moderate congestus. For stratiform and clear sky, the number of CMCs does not strongly influence the ACFs. The deep convective, strong congestus and moderate congestus fractions are small and intermittent for the CMC with $N = 100$, which results in non-smooth ACFs.

4.8 Discussion and conclusion

In this study we constructed a stochastic multicloud model from observational radar data in Darwin, Australia, combined with large-scale data representing the atmosphere around Darwin. The multicloud model consists of CMCs switching between different cloud types (moderate congestus , strong congestus, deep convective and stratiform clouds and clear sky), a model set-up similar to [72] and [34]. The model is able to reproduce cloud type area fractions comparable to the observational fractions (especially for the deep convective area fractions, on which we focussed primary). The vertically averaged large-scale vertical velocity $\langle\omega\rangle$ was found to be a good indicator, whereas CAPE or RH were found to be less suitable indicators. This is in agreement with the findings of [29].

The number N of CMCs used to form cloud type area fractions was shown to be an important parameter of the model: for moderate values of N the model shows significant stochastic fluctuations and the model is able to produce area fractions comparable with the observational fractions. For large values of N the model is more deterministic and unable to reproduce fractions well. The stochastic nature of the model is essential for making the fractions comparable to the observations. Further, by changing N the multicloud model can be adapted to the horizontal scale if implemented in a GCM, providing a way to make the parameterization scale-adaptive. This makes the model suitable for GCMs using non-uniform grids. Further, the model can be used as a start for GCMs reaching grid sizes that fall in the Grey Zone, i.e., for grid sizes so small that subgrid convective flux terms are of the same order as the resolved flux terms (e.g., [36, 149]).

In the Grey Zone, besides the problem that the fluxes are partly resolved partly unresolved, the unresolved fluxes have a large standard deviation [36]. The stochastic multicloud model can produce stochastic fluctuations, resulting in a large standard deviation for the unresolved fluxes, that are difficult (or impossible) to produce with a deterministic model. Another advantage of using a multicloud model with a life cycle is that it will produce cloud type area fractions that are compatible with each other in case of large fluctuations. If the horizontal grid size is large the life cycle is not very important and a large number of Markov chains N can be used such that the model becomes effectively a deterministic model (expected area fractions can be used instead). Still, even then the multicloud model can be useful since the expected area fractions (that depend on the large-scale state) are directly inferred from observational data and can be used in the cumulus parameterizations. With this deterministic version of the multicloud model, we have a tool to examine directly the influence of the stochastic aspect of the model in a GCM. Obviously, for grid resolutions for which moist convection can explicitly be resolved our model is not useful. However, it will take a long time before global climate models can do runs with such fine resolutions.

The horizontal size to which a CMC corresponds is not clearly determined. In principle it corresponds to the horizontal size of the cloud type under consideration, which is different for all cloud types. Using a different number of CMCs for each cloud type is an option, but it is complicated and lies out of the scope of this research. During the training process, we arrived at a value of $N = 100$. This value was chosen because of the comparable standard deviations between model and observations. If local interaction is introduced for the CMCs, then a larger number of CMCs can be chosen while keeping a sufficiently large standard deviation (see [34]).

The fractions produced by the multicloud model depend on the thresholds of Table 4.1 that are used for the classification of the clouds in the radar data. If for example the threshold for rain rate is put from 12 mm h^{-1} to 25 mm h^{-1}, the observed cloud type area fractions change. The fractions produced by the CMCs constructed using the higher threshold also change. The CMC expected area fractions are then close to the new observational means and the same holds for the standard deviations. We conclude that the multicloud model is sensitive to the thresholds in the same way as the classification is sensitive to it.

The interaction of deep convection and the mean vertical velocity is a two-way interaction. If deep convection is triggered, it initiates a feedback system. It causes convergence of air, which in turn changes the mean vertical velocity. This convergence of air will cause more deep convection. In Fig. 4.4, we see that $\langle\omega\rangle$ and the deep convective area fraction attain maximum cross-correlation for positive time lag, suggesting that $\langle\omega\rangle$ can be seen more as an effect than a cause of deep convection. However, this correlation is already high for negative time lag and at time lag zero the deep convective area fraction correlates well with $\langle\omega\rangle$, better than with CAPE or RH. Therefore, we argue that $\langle\omega\rangle$ can be used to condition the Markov chains. In a GCM the deep convective area fractions are only used as a closure of the mass flux at cloud base as described in (4.2) in Section 4.3. In addition to the closure, every parameterization of deep convection further consists of a trigger

function, usually based on instability and/or humidity criteria, as well as a cloud model, which performs the parcel ascent in the vertical. Consequently, convection will only be initiated when the trigger function permits it and its vertical extent will be determined by the cloud model. The deep convective area fractions constructed by our multicloud model determine the strength of the deep convection only if the other conditions are met. By conditioning on $\langle\omega\rangle$ the observed feed-back system will be present in the GCM, but through the trigger function and cloud model deep convection will stop when relative humidity is too low, or when instability is not longer present in the atmosphere.

As the multicloud model was able to reproduce the cloud type area fractions quite well, a natural step is to test this model in a GCM. We are currently testing the multicloud model in a GCM of intermediate complexity (e.g., with prescribed sea-surface temperatures) and we will report on this in a separate paper. We use the deep convective area fractions σ_4 as a closure for the mass flux at cloud base. The strong congestus area fractions σ_3, which also represents convection, can be added with a different updraft velocity, and the same can be done with the moderate congestus fractions σ_2. As an alternative to using a parcel ascend cloud model it is possible to define vertical heat and moisture tendency profiles corresponding to each cloud type (e.g., [72]) or explicitly inferring vertical heat and moisture tendency profiles from data as in [36]. Another possible application of the model in a GCM is that $\sum_{m>1}\sigma_m$, or $1-\sigma_1$, can be used in the parameterization of cloud cover.

The main weakness of our model is that there is no spatial dependence between the CMCs other than through the large-scale state. In the atmosphere clouds are often organized into spatial structures, but with our model it is not possible to produce such spatial organization inside a grid box of a GCM. As mentioned, if spatial organization inside a grid box is desired, then introducing local spatial dependencies between the CMCs is a possibility. This is however, a difficult task and increases the complexity of the model (see [34]). The absence of local dependencies results in too small standard deviations for the CMC fractions when N is chosen to be equal to the number of radar sites. The area fraction of N CMCs converges fast to the expected value for increasing N, much faster than the fractions formed by radar pixels in the domain for which there is large dependence between neighboring pixels. Further, the peak values of the observational fractions of stratiform, moderate congestus and clear sky are difficult to produce while keeping N such that the peak values of the deep convective area fractions are good. The standard deviation for stratiform, moderate congestus and clear sky are too small and we noticed that the ACFs of the area fractions produced with $N = 100$ CMCs decrease too much initially (except for stratiform and clear sky).

How representative is our model? We showed that by training the CMCs with observational data from a five-month period in Darwin, the multicloud model was able to adequately produce fractions for a different three-month period at the same location. This indicates that the model works for a large range of large-scale atmospheric conditions and that a time series of five months is long enough to train the model for Darwin. In the experiment where we interchanged training and test data set we found that even training on a three-month period is enough to produce adequate fractions for the five-month period. We conclude that the time series

is long enough to make a representative parameterization of deep convective and cloud area fractions for Darwin itself.

The main advantage of using observational radar data over LES data is that a longer time period can be covered. The LES data set of the study of [34] was six hours as opposed to the ~eight-months period of the radar data. A simulation of eight months for a domain of the size of the radar domain is not yet computationally possible. Darwin is located in a tropical region where deep convection occurs frequently in the monsoon period, therefore it is representative for deep convection in the tropics. [50] show that only a small adaptation has to be performed to use their stochastic parameterizations of deep convection, also conditioned on ω, at a different location than where they have been trained. This supports that also our multicloud model could be used more globally. However, since convection is (in part) location dependent, e.g., the presence of land or sea, our model could be improved by using observations from multiple locations. Note that even in state-of-the-art GCMs, mass flux at cloud base closures are functions of large-scale variables only and are not specifically adapted to the location on the globe.

To summarize the strengths of our approach: realistic observational data is used to estimate the model; the CMC cloud type area fractions were shown to be comparable to the observations, which is notable, because we used different data sets for training and validation. Furthermore, we saw that the model can be adapted to the scale of the GCM, giving larger fluctuations when a smaller number of Markov chains is used to produce area fractions. Due to the conditioning, memory effects are build in that are often absent in conventional stochastic convection schemes. Implementation in a GCM for assessing the model in a dynamical environment is possible and it can be improved by using additional data from different locations.

4.9 Acknowledgment

We are grateful to Karsten Peters, Keith Myerscough and three anonymous reviewers for useful comments on the paper. This research was supported by the Division for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO).

Chapter V
Stochastic convection parameterization in a GCM

5.1 Abstract

Conditional Markov Chain (CMC) models have proven to be promising building blocks for stochastic convection parameterizations. In this paper, it is demonstrated how two different CMC models can be used as mass flux closures in convection parameterizations. More specifically, the CMC models provide a stochastic estimate of the convective area fraction that is directly proportional to the cloud base mass flux. Since, in one of the models, the number of CMCs decreases with increasing resolution, this approach makes convection parameterizations scale-aware and introduces stochastic fluctuations that increase with resolution in a realistic way. Both CMC models are implemented in a GCM of intermediate complexity. It is shown that with the CMC models, trained with observational data, it is possible to improve both the subgrid-scale variability and the autocorrelation function of the cloud base mass flux as well as the distribution of the daily accumulated precipitation in the tropics. Hovmöller diagrams and wave-number frequency diagrams of the equatorial precipitation indicate that, in this specific GCM, convectively coupled equatorial waves are more sensitive to the mean cloud base mass flux than to stochastic fluctuations. A smaller mean mass flux tends to increase the power of the simulated MJO and to diminish equatorial Kelvin waves.

5.2 Introduction

Deep convection is an atmospheric process of major importance in Earth's weather and climate system. Locally, it transports heat, moisture, and momentum vertically in the atmosphere [3]. Globally, it affects the large-scale circulation [118]. Further, deep convection largely determines precipitation in the tropics. Of specific interest is its coupling to equatorial waves (e.g., equatorial Kelvin waves, Rossby waves, and the MJO) that largely determine the variability of precipitation [76, 143]. Most GCMs do not resolve deep convection. Instead, this process is represented by *parameterizations*, assuming for example a cumulus ensemble that is in quasi-equilibrium with the large-scale forcing [5].

Availability of larger computational resources allows GCMs to be run at finer resolutions. At horizontal grid resolutions below ~ 100 km, and especially in the *Grey Zone* (1-10 km), where convection becomes partially resolved, the quasi-equilibrium assumption breaks down. As a result, the assumption that there is a unique

This chapter has been published as: Jesse Dorrestijn, Daan T. Crommelin, A. Pier Siebesma, Harmen J.J. Jonker, and Frank Selten, 2016: Stochastic Convection Parameterization with Markov Chains in an Intermediate-Complexity GCM. *J. Atmos. Sci.*, **73**, 1367–1382 [37]

relation between the cumulus ensemble and the large-scale conditions is not reasonable anymore, because the ensemble in the GCM grid column is too small, and life cycles of individual cumulus clouds cause large fluctuations in the convective response and associated subgrid fluxes. Therefore, convection parameterizations should become *scale-aware* [4] and stochastic ingredients are required in absence of quasi-equilibrium at increasing resolutions (e.g., [109, 114]). Stochastic physics have been introduced in GCMs for various reasons: to more realistically represent the subgrid-scale variability [86], but also to enlarge the model spread in ensemble prediction systems (e.g., [9, 20, 136]).

The stochastic subgrid-process parameterization approach used in this paper has been introduced by [25]. The main idea behind this approach is to represent subgrid processes of an atmosphere or ocean model by stochastic processes of which the properties are *inferred from high-resolution data* prior to implementation. More specifically, the processes are represented by finite state Markov chains with transition probability matrices that are estimated from data and are *conditioned on the resolved model variables*. In [25], the conditional Markov chains (CMCs) were shown to adequately represent subgrid-scale variables in the Lorenz '96 model [91]. Using the same CMC approach in a GCM to parameterize convection, is a challenging task.

In a GCM, both the large-scale and the subgrid-scale state are not single scalars as is the case in the Lorenz '96 model, but instead are formed by various vertical profiles of resolved and subgrid variables respectively. Another difficulty is the availability of high-resolution data of convection. As explained by [25], Markov chains can be inferred from high-resolution convection resolving model data as well as observational data. Inferring CMCs from high-resolution model data has been explored by [34, 36].

Inspired by the stochastic *multicloud model* of [72], [35] constructed a stochastic multicloud model on a two-dimensional square lattice, using CMCs inferred from observational data. The model was inferred from an extensive data set, consisting of a combination of high-resolution data of deep convection [79] and large-scale re-analysis data improved with observational data [29]. The high-resolution (2.5×2.5 km^2) data originated from a rain radar located in the tropics (Darwin, Australia) and was available every ten minutes for several months in a region of size $\sim 1.5^\circ \times 1.5^\circ$. Thresholds for the cloud top height and the rain rate were used for classification into a finite number of convective or stratiform cloud types [34, 72]. Observations of cloud type transitions were used to estimate the transition probabilities of the CMCs. When conditioned on the large-scale vertical velocity and choosing 100 CMCs, the cloud type area fractions of the scheme were comparable to the observational fractions in the radar domain. By varying the number of CMCs, the multicloud model could be adapted to the size of a GCM column, thereby making the parameterization scale-aware.

In [50], a similar data-driven stochastic scheme has been developed. Observational data sets from Darwin and Kwajalein were used to construct parameterizations of the *convective area fraction* σ_c, also conditioned on the large-scale vertical velocity. The convective area fraction was obtained by sampling directly from the area fraction distribution that was estimated from the data before, conditioned on

the large-scale state. Introducing time-correlation was explored as well by using CMCs. The scheme was able to adequately reproduce observational time series of σ_c.

Testing the schemes in a dynamical environment, in which the CMCs are interacting with the resolved model variables in a GCM, is a necessary step in the development of the CMC-based schemes for the usage in state-of-the-art GCMs. Therefore, in the present paper, we show results of the implementation of the stochastic multicloud model of [35], referred to as *Dor15*, and a scheme similar to the CMC scheme of [50], referred to as *Gott15*, in a GCM of intermediate complexity; the climate model SPEEDY (Simplified Parametrizations, primitivE-Equation DYnamics) [77, 101].

The stochastic schemes produce σ_c which serves as a *closure for the cloud base mass flux* M_b in the convection parameterization scheme. So, SPEEDY's traditional deterministic convection scheme, a simplified Tiedtke mass flux scheme [137], is made stochastic by using σ_c as stochastic input for the determination of M_b. This is a crucial step in the coupling of the stochastic schemes to the convection scheme of SPEEDY. The coupling of a stochastic scheme to the convection scheme of a NWP model, via σ_c and M_b, has been successfully applied earlier by [9].

Our paper is organized as follows. In Section 5.3, we describe the Dor15 scheme, followed by a description of the Gott15 scheme in Section 5.4. Then, we explain how we implement the schemes in SPEEDY in Section 5.5. We specify the observational data sets in Section 5.6, and we present model results in Section 5.7. A discussion follows in Section 5.8.

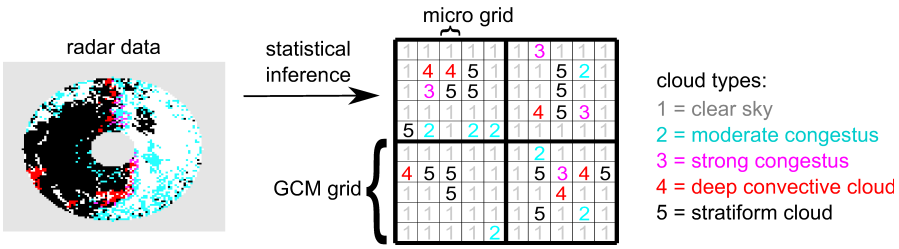

Figure 5.1: Illustration of the stochastic multicloud model (the Dor15 scheme). The thick black lines indicate the GCM grid of which we see four columns from a top view. Inside the four columns, the thin black lines form the two-dimensional micro grid of the multicloud model. Here, each GCM grid column contains $N = 25$ nodes, with a CMC on each node, switching between the five cloud types. A snapshot from the discretized radar data from Darwin is included to point out that the transition probabilities of the CMCs are estimated from observational data.

5.3 The Dor15 scheme

The stochastic multicloud model consists of a two-dimensional square lattice with N nodes, with at each node a CMC, denoted Y_n ($1 \le n \le N$), that switches, every ten min, between the following states: clear sky (1), moderate congestus (2), strong congestus (3), deep convective cloud (4) and stratiform cloud (5). We refer to these

states as *cloud types*. In Fig. 5.1, we illustrate how the multicloud square lattice, or "micro grid", can be embedded in a GCM grid. In the figure, we see four GCM columns with $N = 25$ CMCs for each column. The value $N = 25$ has only been chosen for the sake of illustration and results will be presented for $N = 100$ and $N = 500$.

The transition probabilities of the CMCs depend on the large-scale state of the atmosphere: they are conditioned on the vertical velocity averaged over the lower part of the troposphere defined by:

$$\langle\omega\rangle := \frac{1}{p_0 - p^*}\int_{p^*}^{p_0} \overline{\omega}(p)dp,$$

in which $\overline{\omega}$ is the large-scale vertical velocity in hPa h^{-1}, p_0 is the pressure at the surface, and p^* is the pressure level at 340 hPa. We condition the CMCs on $\langle\omega\rangle$, because in [35] this variable was shown to have the largest correlation with deep convection, see [29, 113] for similar findings. Since the CMCs have five states, the transition probability matrices are of size 5×5 and since we bin all possible values of $\langle\omega\rangle$ into 25 intervals, we obtain 25 matrices; for each interval there is a different 5×5 matrix.

In a GCM grid column, the N CMCs yield area fractions σ_m for each cloud type $1 \le m \le 5$ which are defined by:

$$\sigma_m = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}[Y_n = m], \tag{5.1}$$

in which $\mathbf{1}[\cdot]$ is the indicator function ($\mathbf{1}[Y_n = m] = 1$ if $Y_n = m$ and $\mathbf{1}[Y_n = m] = 0$ if $Y_n \neq m$). Previous studies based on observational data [35, 50] show that the expectation value of σ_4 is an increasing function of $\langle\omega\rangle$, with a maximum of around 0.03 for $\langle\omega\rangle \approx 15$ hPa h^{-1}.

Ideally, one would like to choose N such that the size of the micro-grid cells corresponds to the typical size of a convective updraft area L^2_{conv}. This implies that N should be the ratio between the GCM horizontal grid size area ΔX^2 and L^2_{conv}, i.e., $N \approx \Delta X^2/L^2_{\text{conv}}$. The parameter N is a scaling parameter enabling the Dor15 scheme to adapt to the GCM grid resolution and determines the magnitude of the stochastic fluctuations of the area fractions σ_m. The larger N, the smaller the deviations from the expectation values, to which the fractions converge if $N \to \infty$. This gives a deterministic version of the model. Previous off-line studies [35] showed that for $N = 100$ the temporal fluctuations of the deep convective fractions resemble the observational fluctuations on an area of size 170×170 km^2, and therefore, $L^2_{\text{conv}} \approx 17^2$ km^2. The value $N = 100$ is ideal for usage in a GCM with grid size $\Delta X^2 = 170^2$ km^2. We test the multicloud model in SPEEDY for the relatively small value $N = 100$, referred to as Dor15-100, to be able to assess the impact of stochastic fluctuations. As an extra sensitivity test, we do an additional experiment with $N = 500$, referred to as Dor15-500, which is a more appropriate value for SPEEDY.

For the implementation in SPEEDY we use the sum of the strong congestus and deep convective area fractions to estimate the convective area fraction:

$$\sigma_c = \sigma_4 + \sigma_3, \tag{5.2}$$

that is used in the closure for M_b in the convection scheme, explained in detail in Section 5.5.

More information about multicloud models can be found in e.g., [1, 32, 34, 35, 43, 72, 73, 94, 113].

5.4 The Gott15 scheme

In the Gott15 scheme, the CMCs switch between σ_c values instead of cloud types, and *only one* CMC is used for each GCM column (by contrast, the multicloud model has N CMCs in each GCM column). Thus, the scheme is less complex than the multicloud model, but it is not scale-aware. The fluctuations of σ_c can not be adapted to the GCM resolution. We will now describe in detail how we construct the Gott15 scheme.

Again, we use the discretized Darwin radar data set. The deep convective area fractions σ_4 are added to the strong congestus area fractions σ_3 forming σ_c. We cluster the fractions with k-means [45, 92], using $K = 10$ cluster centroids. This results in ten possible σ_c values, which are the states of the CMCs. We use the observational σ_c to estimate transition probability matrices of size 10×10. As in Dor15, the CMCs are conditioned on the 25 intervals of $\langle\omega\rangle$, so we estimate 25 matrices; for each interval of $\langle\omega\rangle$ there is a different 10×10 matrix. The transition probabilities of the CMC correspond to a time step of ten min, since observational fractions are available every ten min, and to an area size of $\sim 1.5° \times 1.5°$, which is the size of the radar domain.

The Gott15 scheme is implemented in SPEEDY in the same way as the multicloud model: σ_c is used as a closure for M_b. We stress that the main difference between the Gott15 scheme and the Dor15 scheme is that the Gott15 scheme does not make use of a multicloud model, instead its CMCs make transitions between σ_c values.

5.5 Implementation in SPEEDY

SPEEDY is a GCM of *intermediate complexity*: only the most important processes are incorporated in the model, they are represented in a simplified way, and the GCM's resolution is coarse [77]. It is a hydrostatic spectral model that solves the primitive equations on the entire globe. The prognostic variables are vorticity, horizontal divergence, absolute temperature, surface pressure, and specific humidity. The time integration is performed by a leapfrog scheme and the time step in the standard version of SPEEDY is 40 minutes. In our version, the horizontal resolution is T30, referring to a triangular truncation at total wavenumber 30. The prognostic model fields are expanded into series of spherical harmonical functions of total wavenumber 30 and smaller. Along latitude circles these functions correspond to cosine and sine functions with maximum zonal wavenumber 30. This corresponds to a size of $\sim 3.75° \times 3.75°$ for each of the $96 \times 48 = 4608$ vertical columns. In the vertical, the model has eight pressure levels. SSTs are prescribed by using observational climatological fields, while land skin temperatures are prognosed using a soil model. SPEEDY has a seasonal cycle, but no daily cycle. Simplified

parameterizations are used to represent short-wave and long-wave radiation, deep convection, clouds, surface heat and moisture fluxes, large-scale condensation and vertical diffusion (representing, e.g., shallow convection). Precipitation is the sum of the large-scale and convective precipitation. The large-scale precipitation is derived from a large-scale condensation scheme and the convective precipitation is derived from the deep convection scheme.

The reason why we choose such a simplified GCM is that it provides a perfect playground to explore new stochastic concepts in convection parameterizations and the impact on the representation of intraseasonal variability caused by equatorial waves. In that respect this explorative study should be considered as a natural intermediate step from recent off-line studies [35, 50] toward an implementation into the state-of-the-art GCMs.

The relaxation closure (CTRL)

The deep convection scheme is a simplified Tiedtke mass flux scheme [137]. Convection in a grid column is triggered if the atmosphere is conditionally unstable with respect to the lowest model level and if the relative humidity in the two lowest model levels exceeds a critical value (see the online SPEEDY manual by [102]). In the standard version of SPEEDY, the cloud base mass flux M_b is estimated by a relaxation closure. This closure determines a value of M_b such that the convection scheme relaxes back to a prescribed relative humidity threshold in six hours. The control experiments are done using this relaxation closure and are referred to as CTRL. In the vertical, the mass, heat and moisture fluxes are modified by a prescribed entrainment profile, while detrainment is assumed only to occur at the highest level where the convection scheme is active by depositing the convective updraft mass, heat and moisture into the environment.

Implementation of the stochastic schemes

The stochastic schemes are implemented by replacing the standard relaxation closure for M_b, which can instead be estimated by using the definition:

$$M_b = \rho w_c \sigma_c, \tag{5.3}$$

with a typical prescribed value of the *updraft momentum at cloud base*, $\rho w_c = 1$ kg m^{-2} s^{-1} [100]. For the multicloud model, we will also test the influence of this particular choice by varying this updraft momentum. In one experiment, we set ρw_c at 0.5 kg m^{-2} s^{-1} while using $N = 100$, referred to as Dor15-100w0.5, and later we choose other values of ρw_c.

When the multicloud model is used, we evolve $N = 100$ or $N = 500$ CMCs in every vertical column of SPEEDY, yielding cloud type area fractions σ_m for each cloud type at every model time step. The convective area fraction σ_c is calculated with (5.2) and used in (5.3). Note that we also evolve the CMCs for columns without deep convection (in case the trigger function did not activate convection), to be sure that the Markov chains do not have to spin up when convection is activated. Since the transition probabilities of the CMCs correspond to a time step of ten min, we set the time step of SPEEDY at ten minutes for all runs.

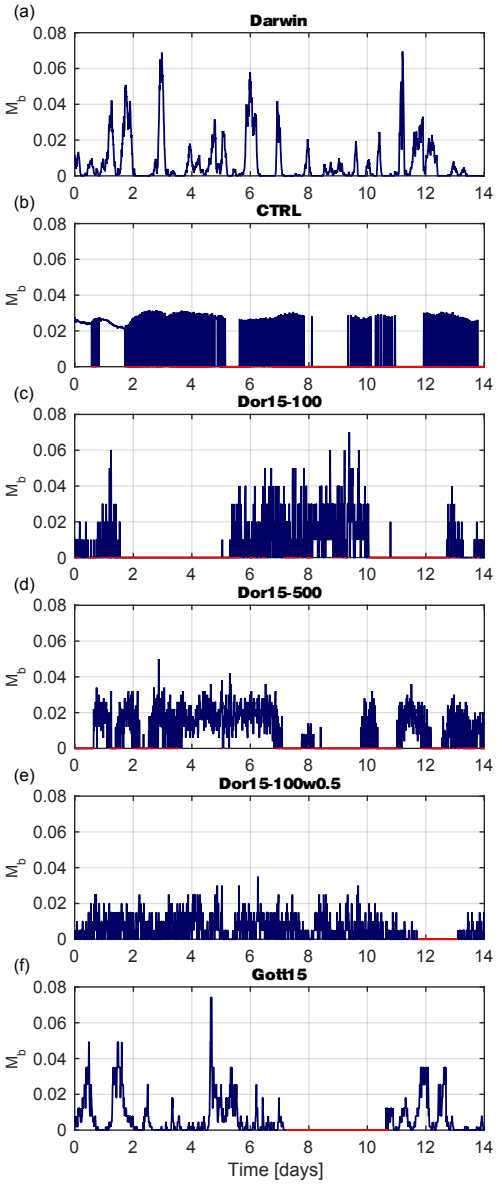

Figure 5.2: Typical time series of M_b in kg m^{-2} s^{-1} (a) observed in Darwin (σ_c observations used as a proxy for M_b, assuming ρw_c = 1 kg m^{-2} s^{-1}), and produced by SPEEDY at 130°E-13°S for (b) CTRL, (c) Dor15-100, (d) Dor15-500 (e) Dor15-100w0.5 and (f) Gott15. An inactive trigger function is indicated by a red dot at the horizontal axis.

In each vertical column, the input of the CMCs is the large-scale vertical velocity $\langle\omega\rangle$. The value $\langle\omega\rangle$ is assigned to one of the 25 interval numbers and the CMCs

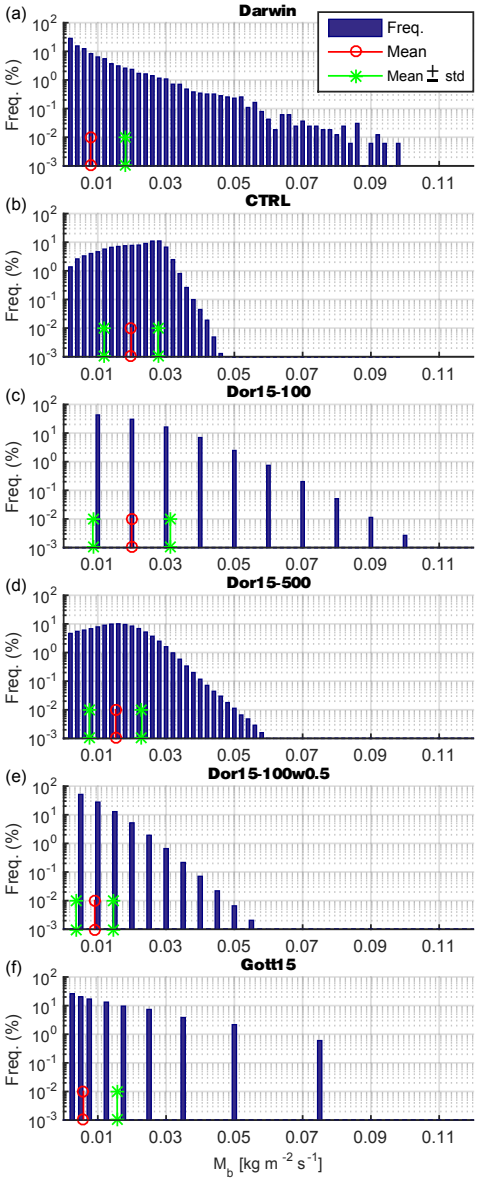

Figure 5.3: Histograms showing the relative frequency of occurrence of the non-zero M_b, the mean M_b and the standard deviation for (a) the Darwin observations (using σ_c as a proxy), and for SPEEDY (one year of model data between 15°N-15°S) using (b) CTRL (c) Dor15-100 (d) Dor15-500 (e) Dor15-100w0.5 and (f) Gott15.

will all use the same transition probability matrix that corresponds to this interval number. Given its cloud type, each CMC will switch to another cloud type (or does

not change cloud type), and after that, the new area fractions σ_m are calculated using (5.1). In the present study, we only use σ_4 and σ_3. In [35], the possible usage of the other cloud type area fractions is described.

When the Gott15 scheme is used, we evolve only 1 CMC in every column of SPEEDY, which directly yields σ_c. The value of $\langle\omega\rangle$ in a model column determines which transition probability matrix is used by the CMC.

5.6 Observations

We will compare the model behavior of SPEEDY with observations. We will use two observational data sets. The first data set is the Darwin radar data set. We will compare M_b at time step level (ten minutes) of the two stochastic schemes and CTRL with M_b observed in Darwin. We emphasize that we do not have observations of M_b, however, since we use σ_c of the stochastic schemes directly as M_b in (5.3), we will also use the observational σ_c as a proxy for the observational M_b by assuming again that ρw_c is equal to 1 kg m^{-2} s^{-1}.

The second observational data set is the daily accumulated precipitation GPI data set ($1^\circ \times 1^\circ$) [62]. Since SPEEDY has a resolution of $\sim 3.75^\circ \times 3.75^\circ$, we average the observational precipitation values over blocks of this size.

5.7 Results

We run SPEEDY several times for 11 years with different closures for M_b. In order to avoid spin-up effects, data from the first year are excluded. We store variables (e.g., M_b and precipitation values) at every time step and for all vertical columns around the equator between 15°N and 15°S, which are eight vertical columns for each longitude.

M_b at time-step level

To get a first impression of the convective behavior of SPEEDY with the several closures, we show M_b at time-step level for two weeks for a vertical column located at $\sim$ 130°E-13°S in Fig. 5.2. We choose this particular grid column, because it is closest to Darwin, Australia, for which we can show the time series of M_b using σ_c as a proxy in Fig. 5.2a. The time series should be compared in a statistical sense. The goal is not to give an identical reproduction of the time series observed in Darwin, instead we show "typical" time series of the several closures during the rain season.

In Fig. 5.2b, we see that the mass flux of CTRL is non-zero for specific time intervals, only when the trigger function is active (an inactive trigger function is indicated by a red dot at the horizontal axis). If the trigger function allows for convection, the mass flux is always close to 0.03 kg m^{-2} s^{-1}; CTRL has small variability. Further, there are periods when the trigger function switches convection on and off too rapidly, for example from day 2 till day 5. The too intermittent behavior of CTRL is due to the trigger function.

In Fig. 5.2c, we clearly see the discrete character of Dor15-100: only values that are integer multiples of 1/100 = 0.01 are attained, because $N = 100$ CMCs are used to calculate σ_c. If only one CMC is in a convective state (state 3 or 4),

σ_c = 1/100 = 0.01, if two CMCs are in a convective state, then σ_c = 2/100 = 0.02, etc. The mass flux fluctuates between 0 and 0.07 kg m^{-2} s^{-1}, in this period of this particular realization, which suggests that the variability has improved compared to CTRL. Note that a zero mass flux can be the result of an inactive trigger function or a convective area fraction σ_c equal to zero. For example, from day 2 till day 5 the zero M_b is a result of an inactive trigger function. The character of the time series is too intermittent compared to the series observed in Darwin, which can not be exclusively attributed to the trigger function.

In Fig. 5.2d, we see the mass flux produced by Dor15-500. Mass flux values higher than 0.04 kg m^{-2} s^{-1} are rare. Deviations from the expectation values are expected to be smaller compared to the N = 100 experiment. By increasing N even more, the time series start to resemble the series of CTRL. However, note that for the deterministic limit $N \to \infty$, the closure still differs from the standard relaxation closure, and therefore, convergence of the stochastic closure to CTRL should not be expected.

In Fig. 5.2e, we see that Dor15-100w0.5 produces lower M_b values than Dor-15-100 and that the values are multiples of 0.005. Lower mass fluxes imply that convective instabilities are less quickly removed, leading to prolonged periods of convective activity. As opposed to Dor-15-100 the trigger function is active for almost the entire period: it is only inactive around day 12.

Finally, the Gott15 scheme (Fig. 5.2f) produces M_b time series that are similar to the series as observed in Darwin. The highest value of M_b lies between 0.07 and 0.08 kg m^{-2} s^{-1}. The general shape of the convective peaks looks quite realistic for this scheme. It less intermittent than the multicloud and CTRL time series.

Clearly, compared to CTRL, the two stochastic schemes (Dor15 and Gott15) are better reproducing the fluctuations as observed in Darwin.

The distribution of M_b

In Fig. 5.3, the distributions of M_b are visualized by showing histograms of the relative frequency of occurrence of the non-zero M_b, and the corresponding mean and standard deviation observed in Darwin (Fig. 5.3a) and for model data between 15°N-15°S based on the different closures (Fig. 5.3b-f). The y-axes are scaled logarithmically, to make the tails of the distributions better visible. The Darwin histogram corresponds to a distribution that is approximately exponential with a maximum M_b of around 0.10 kg m^{-2} s^{-1}.

In Fig. 5.3b, we see that the mass flux of CTRL has a peak value at 0.03 kg m^{-2} s^{-1} and that the relative frequencies are rapidly decreasing to zero for larger mass fluxes. The maximum value lies below 0.05 kg m^{-2} s^{-1}. The mean mass flux of CTRL is larger than the mean mass flux observed in Darwin and the standard deviation is smaller. This is also the case if we evaluate the model data near Darwin instead of the entire tropical belt.

The mass flux of Dor15-100 (Fig. 5.3c) can attain values up to 0.10 kg m^{-2} s^{-1}. The discrete character of the scheme is visible, with only integer multiples of 0.01 kg m^{-2} s^{-1}. The mean flux is close to the mean flux of CTRL and its standard deviation is slightly larger. Dor15-500 (Fig. 5.3d) displays a histogram that resembles the histogram of CTRL, except that a higher maximum mass flux is possible. The

histogram looks smoother than the histogram of Dor15-100, since integer multiples of 0.002 kg m^{-2} s^{-1} can be attained. The mean mass flux is lower than the mean mass flux of Dor15-100 and it has a smaller standard deviation. Dor15-100w0.5 produces lower M_b than Dor15-100 and the histogram suggests that M_b is approximately exponentially distributed.

Gott15 (Fig. 5.3f) attains ten different mass flux values, which are exactly the values of the ten cluster centroids. Its maximum mass flux is around 0.07 kg m^{-2} s^{-1}; higher maximum values can be obtained, for example, by using a larger number of cluster centroids. This last option would, however, need reconstruction of the Gott15 scheme through a revised estimation of the transition matrices. The relative frequency of the bins of Gott15 seems to decrease approximately exponentially. The mean and standard deviation are close to the observational values.

We conclude that, compared to CTRL, the stochastic schemes (Dor15 and Gott15) produce mass flux distributions that are more similar to the Darwin distribution. However, the discrete character of the stochastic schemes is not very realistic.

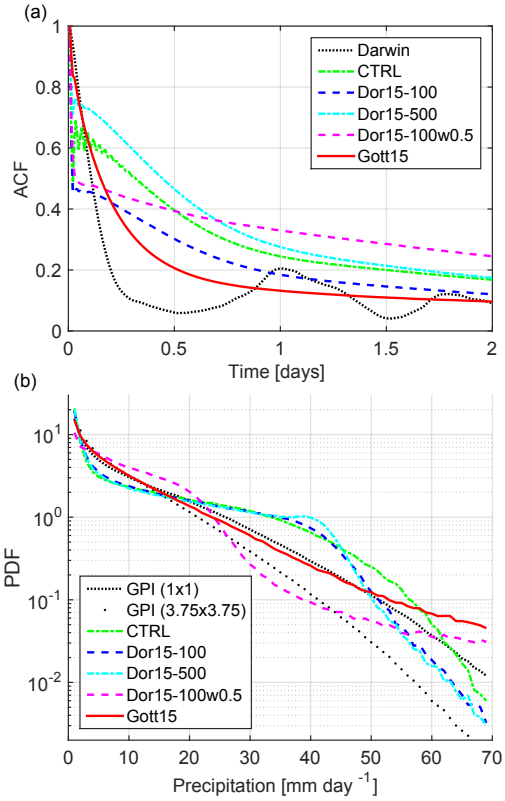

Figure 5.4: (a) The autocorrelation function of M_b for Darwin observations and for SPEEDY (15°N-15°S) with CTRL, Dor15-100, Dor15-500, Dor15-100w0.5 and the Gott15 scheme (b) the PDFs of the non-zero daily accumulated precipitation for GPI and for SPEEDY (15°N-15°S) with the same closures.

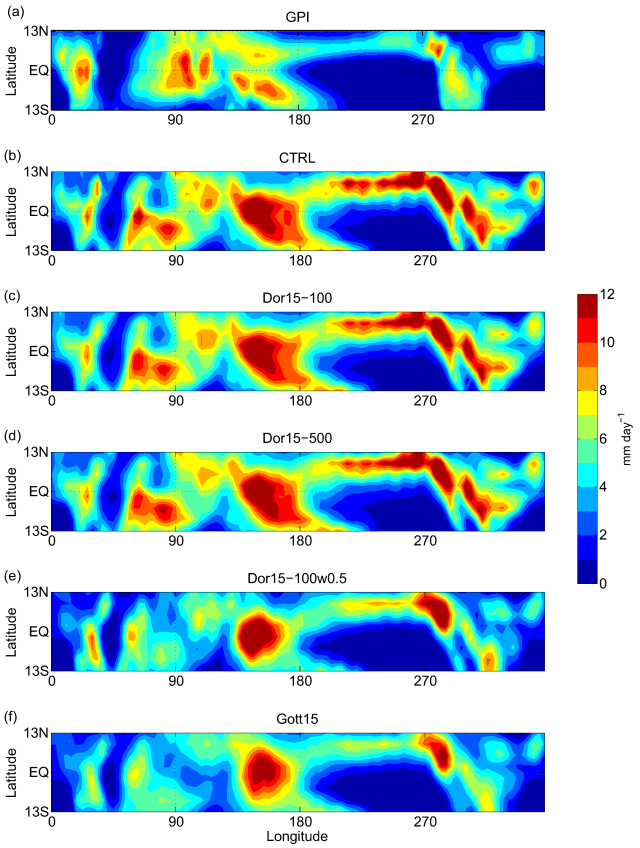

Figure 5.5: Mean equatorial precipitation (ten-year averaged) for (a) the GPI observations (3.75° ×3.75°) and SPEEDY (3.75° × 3.75°) with (b) CTRL (c) Dor15-100 (d) Dor15-500 (e) Dor15-100w0.5, (f) Gott15.

Autocorrelation functions

Deep convection is correlated in time and probabilities of the occurrence and strength of convection depend strongly on earlier time instances. This is one of the reasons why we choose to parameterize convection with Markov chain models; to be able to capture this correlation. How well the several closures reproduce observational correlations, can be assessed by calculating autocorrelation functions (ACFs) [35].

In Fig. 5.4a, we plot ACFs of M_b averaged over 15°N-15°S for one year of model data with the Gott15 scheme, the multicloud model (N = 100 and N = 500) and CTRL and compare them to the ACF of M_b observed in Darwin. Compared to the observations, the ACFs of all models except Gott15 decrease too rapidly initially due to the intermittent character and too slow thereafter. In contrast, the ACF of Gott15 is close to the observational ACF and the discrepancies can be partly contributed to the absence of a daily cycle in SPEEDY. The absence of a daily cycle

in SPEEDY contributes to a slower decay of the ACFs and the absence of a peak after one day.

Precipitation

The daily accumulated precipitation is an important output of GCMs. We will assess the different mass-flux closures by comparing the model's precipitation output with observations. In Fig. 5.4b, we show the PDFs of the non-zero daily accumulated precipitation for ten years of data between 15°N-15°S. Note the logarithmic scale of the y-axis. We see that the PDF produced while using Gott15 is very close to the PDF of the GPI observations ($1^\circ \times 1^\circ$) for precipitation values less than 50 mm day^{-1} and that higher values are too frequent. Its PDF is not so close to the GPI observations that are averaged over blocks of size $3.75^\circ \times 3.75^\circ$, only for precipitation values below 20 mm day^{-1} there is a good fit. Gott15 has been trained with data corresponding to an area of $\sim 1.5^\circ \times 1.5^\circ$ which may explain why it is closer to GPI $1^\circ \times 1^\circ$ than to GPI $3.75^\circ \times 3.75^\circ$.

The PDFs of Dor15-100, Dor15-500 and CTRL are similar, but not close to the observational PDFs. Above 45 mm day^{-1} the PDFs decrease with the correct slope compared to GPI ($3.75^\circ \times 3.75^\circ$). The PDF of Dor15-100w0.5 differs from the PDF of Dor15-100, but it is still not close to the observational PDFs. For values higher than 50 mm day^{-1}, the PDF is close to the PDF of Gott15. In Section 5.7.6, we will further examine the impact of ρw_c.

In Fig. 5.5, we show ten year averaged equatorial precipitation. The general patterns produced by SPEEDY (Fig. 5.5b-f) are somewhat similar to the GPI observations (Fig. 5.5a): a narrow ITCZ in the North East Pacific Ocean and a wide one over the Maritime Continent. However, there are some major errors: for example, the precipitation in CTRL, Dor15-100 and Dor15-500 in the North East Pacific Ocean is double as high as in GPI. Also SPEEDY's spatial patterns in the Indian Ocean differ significantly from the patterns in GPI.

Dor15-100 (Fig. 5.5c) and Dor15-500 (Fig. 5.5d) do hardly change the precipitation patterns of CTRL (Fig. 5.5b). So, the schemes, based on different closures, produce similar ten-year average precipitation. This can be explained by realizing that precipitation scales with mass flux at cloud base. Inspection of Fig. 5.3b-d shows that the different closures give similar mean mass flux values of 0.02 kg m^{-2} s^{-1}. Dor15-100w0.5 and Gott15 produce significantly lower mean mass flux values, which explains the reduction of the intensity of the precipitation patterns (Fig. 5.5e-f). These schemes do not improve the patterns in general. Only the ITCZ in the North East Pacific Ocean seems to improve. Precipitation in the warm pool (140°E) is still too intense and too localized compared to the observations. We conclude that, the intensity of M_b, rather than the variability of M_b, seems to have a major impact on mean precipitation in SPEEDY.

Equatorial waves

At the equator, the Coriolis force vanishes, and it increases north and south of the equator. This results in dynamics that are typical for the tropics. The governing equations of atmosphere and ocean admit solutions that describe waves traveling

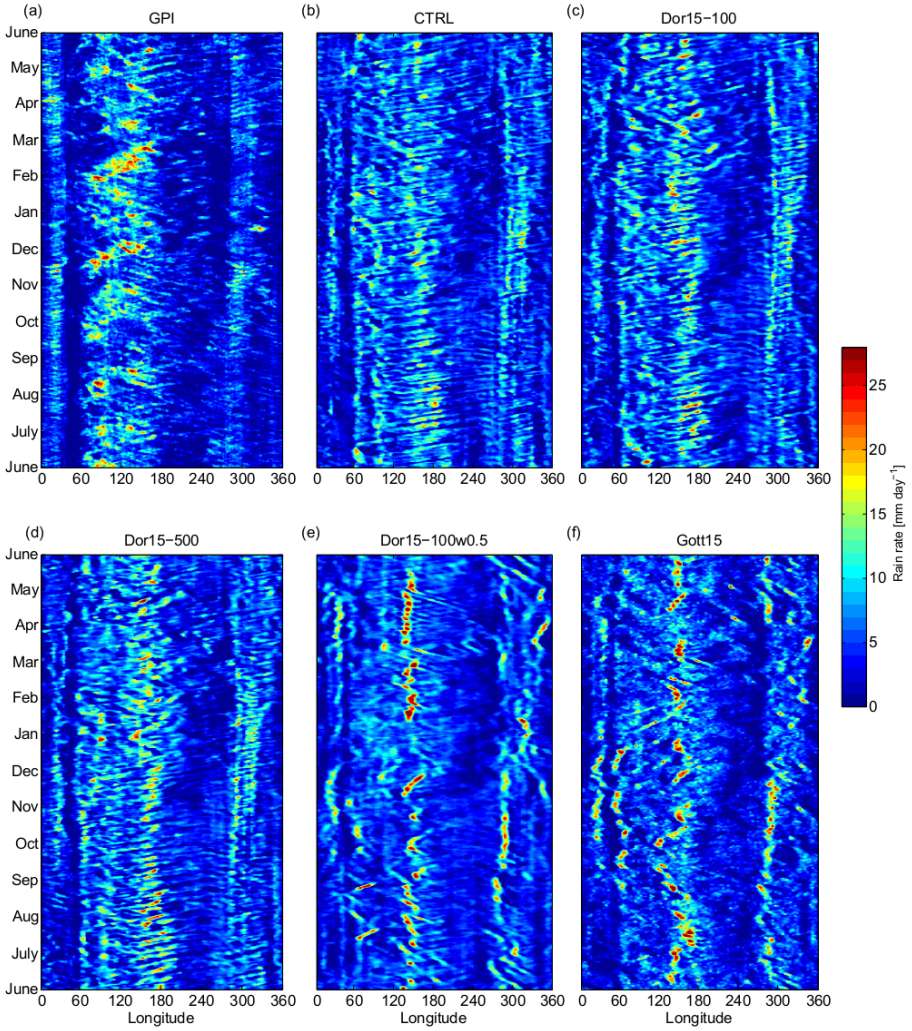

Figure 5.6: Hovmöller diagrams [150] of the daily precipitation (mm day^{-1}) averaged over 15°N-15°S for (a) the GPI observations ($1^\circ \times 1^\circ$) from June 2000 to May 2001, and a typical year of SPEEDY with (b) CTRL, (c) Dor15-100, (d) Dor15-500 (e) Dor15-100w0.5 and (f) Gott15. Note that the diagrams should be compared in terms of general patterns, e.g., equatorial Kelvin waves are better visible in (b) than in (e).

along the equator. It is possible to discern atmospheric waves in satellite observations of precipitation, because of their tendency to couple to deep convection.

A distinction can be made between waves that are mainly symmetrical with respect to the equator, e.g., equatorial Kelvin waves traveling eastward with 15 m s^{-1} (or $\sim 360^\circ$ month^{-1}), equatorial Rossby waves (ER) traveling westward, westward inertio-gravity (WIG) waves, eastward inertio-gravity (EIG) waves, and the

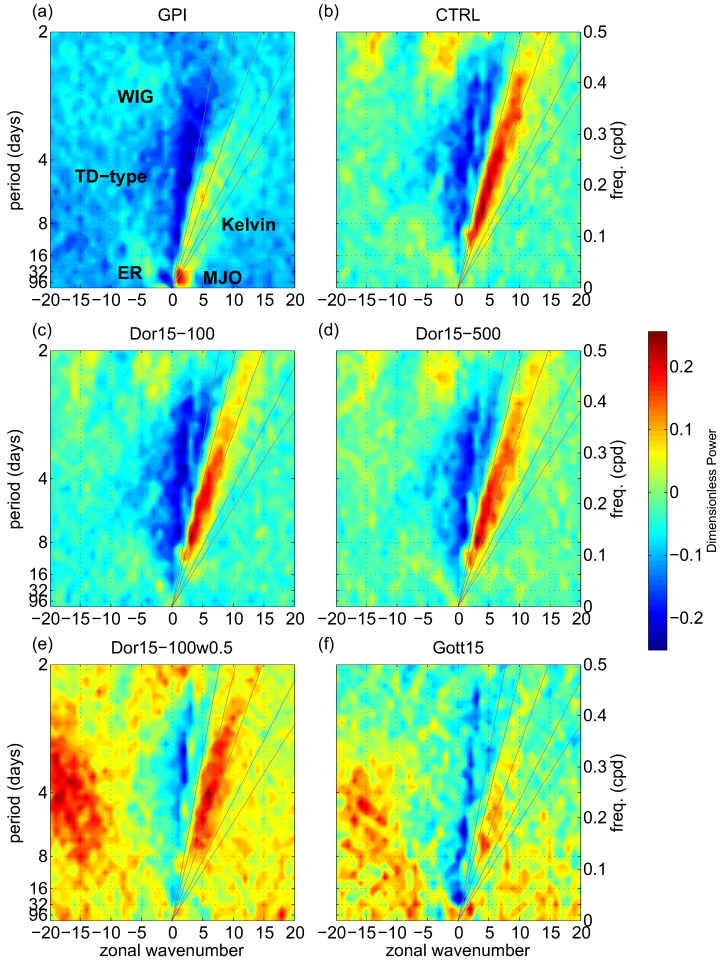

Figure 5.7: Zonal-wavenumber frequency diagrams [85, 143] of the symmetric part of the equatorial precipitation (15°N-15°S) divided by the background spectrum for (a) the GPI observations, (b) CTRL (c) Dor15-100, (d) Dor15-500, (e) Dor15-100w0.5 and (f) Gott15.

MJO traveling eastward with 5 m s^{-1} (or ~ 120° $month^{-1}$), and waves with an anti-symmetric structure with respect to the equator, e.g., mixed Rossby-gravity (MRG). For a comprehensive treatise on equatorial waves, we refer to [143]. State-of-the-art GCMs should be able to reproduce these waves. Producing realistic equatorial waves (especially the MJO), is one of the major challenges for weather and climate modelers [13, 76].

Exactly as in [150], we show in Fig. 5.6 longitude-time plots, also known as Hovmöller diagrams, of the equatorial daily precipitation averaged over 15°N-15°S for one year of GPI observations and for the SPEEDY experiments. Hovmöller diagrams are useful to get a first insight in the model's ability to produce equatorial

waves.

The eastward moving Kelvin waves are clearly visible for CTRL, Dor15-100 and Dor15-500 (Fig. 5.6b-d). In the observations (Fig. 5.6a), these Kelvin waves are visible, but not as prominent. The Hovmöller diagrams of the multicloud model and CTRL are in general very similar. The multicloud model seems to produce slightly larger coherent structures of heavy rainfall, which are visible as tiny red blobs, for example in January at 90°E and 150°E in Fig. 5.6d. The MJO events in the GPI observations, for example in February (60°E-180°E), are prominent and are missing in CTRL, Dor15-100 and Dor15-500. In the Hovmöller diagram of the Gott15 scheme (Fig. 5.6f), large convective events are present (e.g., the red blobs between 120°E-180°E), considerably more than in CTRL. We even see, that MJO-like waves are present between 60°E-180°E in January. These MJO-like waves are, however, not as strong as in GPI, which indicates that the representation of spatial organization of convection is still inadequate.

The Hovmöller diagram of Dor15-100w0.5, with $\rho w_c = 0.5$ kg m^{-2} s^{-1}, differs from the Hovmöller diagram of CTRL: the Kelvin waves are less prominent and structures of heavy rainfall can be seen (mainly between 60°E-180°E) that are similar to the structures of Gott15. Also the MJO-like waves are present (60°E-180°E Jul.-Aug.), but are even weaker than for Gott15.

To further examine the model's ability to produce equatorial waves and investigate intraseasonal variability, we calculate *Wheeler-Kiladis diagrams* [143] of the equatorial precipitation. We focus on the symmetric part of the precipitation, since we are mostly interested in equatorial Kelvin waves and the MJO, the waves with the largest contributions to intraseasonal variance in precipitation. We calculate zonal-wavenumber frequency diagrams of the symmetric part of the equatorial precipitation (15°N-15°S) divided by the background spectrum, for which we apply smoothing with a 1-2-1 filter.

In Fig. 5.7, we plot the diagrams for the GPI observations [85], and the SPEEDY experiments. Note, first of all, that all the SPEEDY model results differ significantly from the GPI diagram. This is, besides the differences in the power of the waves, caused by the different background spectra by which the spectra are divided. In the observations (Fig. 5.7a), we clearly see the MJO peak (around zonal wavenumber 1-5 with a period between 32 and 96 days) and the Kelvin waves for positive wavenumbers. Further, we see the ER and the WIG less prominently. The diagrams of CTRL, Dor15-100 and Dor15-500 (Fig. 5.7b-d) are very similar to each other and show too prominent Kelvin waves while the MJO is essentially missing in these diagrams. These are typical model misrepresentations that occur in many state-of-the-art GCMs [85]. Our multicloud scheme is not able to improve the MJO. Successful MJO-like simulation with similar stochastic multicloud models is possible as demonstrated by [32].

In the diagram of the Gott15 (Fig. 5.7f), we see an MJO peak and the Kelvin waves are less prominent as in CTRL. The tropical depressions (TD-type) are too prominent. For Dor15-100w0.5 (Fig. 5.7e), the Kelvin waves slightly diminish in comparison to Dor15-100, the TD-type are even more prominent than in Gott15, and the MJO peak is missing.

The updraft momentum at cloud base

In the implementation of the stochastic schemes in SPEEDY, M_b was calculated by multiplying σ_c by $\rho w_c = 1$ kg m^{-2} s^{-1} in (5.3). We find that changing ρw_c has a major impact on the model behavior. If we lower ρw_c, then the equatorial Kelvin waves get less prominent and the MJO strength increases. Also the time-averaged equatorial precipitation changes (Fig. 5.5e) as compared to CTRL (Fig. 5.5b). To examine the influence of ρw_c, we do additional runs with Dor15-100 with ρw_c values ranging between $0 \leq \rho w_c \leq 1.5$ kg m^{-2} s^{-1} and calculate the power of the equatorial Kelvin waves and MJO as a function of ρw_c. We define the *power of the equatorial Kelvin waves and the MJO* as the average powers of the corresponding wave regions in the Wheeler-Kiladis diagram as defined in Fig. 6 of [143].

Fig. 5.8a displays the result of 12 independent 11 year runs of SPEEDY using the Dor15-100 scheme with different values of ρw_c. We see that ρw_c has indeed an impact on the power of the equatorial Kelvin waves and the MJO. The equatorial Kelvin waves tend to have less power for smaller ρw_c values. The GPI observational power is 0.08, so the figure suggests that the equatorial Kelvin wave power is only correct when $\rho w_c \approx 0.45$ kg m^{-2} s^{-1}. The MJO power tends to increase for smaller updraft momentum values, but never reaches the MJO observational power 0.14. Note that for $\rho w_c = 0$, the convection scheme is essentially switched off, and all precipitation is formed by large-scale precipitation. The relative contributions of the large-scale and convective precipitation as a function of ρw_c is plotted in Fig. 5.8b. The general idea we get, is that equatorial Kelvin waves are more prominent for schemes with a larger mean M_b and consequently a larger contribution of convective precipitation, and MJO-like features are more prominent for schemes with a smaller mean M_b and consequently a larger contribution of large-scale precipitation. The ratio between convective and large-scale precipitation seems to play a role in the type and the scale of organization of tropical convection in the model. [33] similarly find that the strength of stratiform heating affects the formation of MJO-like or equatorial Kelvin wave structures in an aqua-planet GCM.

With this novel method of calculating the power of the MJO and equatorial Kelvin waves, it is possible to express the model's ability to simulate these waves in a single scalar. This enables modelers, to directly tune parameters for optimal simulation of these waves. Note, however, that even if the powers are exactly equal to the observational powers, it is not yet sufficient to conclude that the model simulates the waves perfectly. Other requirements have to be fulfilled as well [150]. The power only gives an impression. For example, CTRL gives a too high equatorial Kelvin wave power, 0.12, and a too low MJO power, 0.02, which is consistent with the patterns found in the Hovmöller-diagrams (Fig. 5.6a-b).

5.8 Discussion

We have implemented two different stochastic parameterizations for the convective area fraction σ_c in the convection scheme of the intermediate complexity GCM SPEEDY and evaluated the impact in the tropics.

In both stochastic parameterizations σ_c is estimated with CMCs of which the

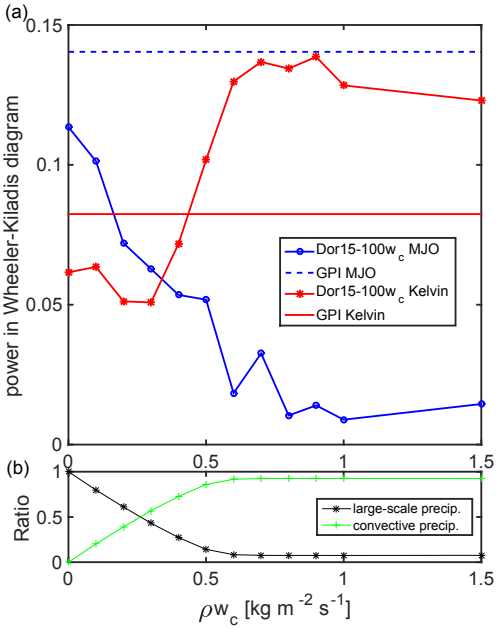

Figure 5.8: (a) The average power of the equatorial Kelvin waves (line with stars) and the MJO (line with circles) in the Wheeler-Kiladis diagram of SPEEDY, using Dor15-100, as a function of ρw_c. Compare with the GPI Kelvin (red solid line) and MJO (blue dashed line) average power (b) the relative contributions of the large-scale and convective precipitation as a function of ρw_c.

transition probabilities are conditioned on the large-scale vertical velocity $\langle\omega\rangle$, as this is the large-scale variable that displays the largest correlation with the occurrence of deep convection [35]. Note that a closure based on $\langle\omega\rangle$ effectively resembles a moist convergence closure, but due to the stochastic aspects our closures are not so hardwired as the more traditional deterministic moist convergence closures (e.g., [80, 137]). Although it is difficult to disentangle convergence and convection in terms of causality, there is no reason not to use the large-scale vertical velocity to condition the transition probabilities of the CMCs.

On a local grid point level both stochastic schemes produce mass-flux time series that are more realistic than the series produced by the standard CTRL version (Fig. 5.2). This is also reflected in a broader and more realistic frequency of occurrence distribution of the cloud base mass flux (Fig. 5.3). Gott15 and to a lesser extend Dor15 also improve the daily accumulated tropical precipitation compared to CTRL (Fig 5.4b). Substantial improvement of the temporal autocorrelation function for M_b is only observed for Gott15 (Fig. 5.4a).

Wheeler-Kiladis diagrams show that the equatorial Kelvin waves are too prominent in SPEEDY for CTRL and that the MJO is missing entirely. Gott15 significantly improves the representation of both the MJO and the equatorial Kelvin waves. Dor15 is only able to improve on this issue by strongly reducing ρw_c. By increasing the relaxation time-scale of the relaxation closure in CTRL, similar

changes are to be expected [44]. For Dor15 it seems that changing the mean M_b has a larger impact on the representation of the equatorial waves than changing the magnitude of stochastic fluctuations of M_b on a time step level.

How much of the model errors are due to the convection schemes and how much due to the large-scale forcings of SPEEDY? The range of $\langle\omega\rangle$ values produced by SPEEDY compares well with the range observed around Darwin. The time series of M_b in Fig. 5.2c and Fig. 5.2f compare well with time series produced by the same schemes using observed $\langle\omega\rangle$ values [35, 50]. In addition, the large range in different mass flux behavior displayed in Fig. 5.2b-f suggests that most of the discrepancies between the Darwin time series and the model time series are due to the convection parameterizations and not the large-scale forcings of SPEEDY. The too intermittent character of for example the Dor15-100 scheme is due to the scheme itself and not to SPEEDY.

An advantage of the Dor15 scheme over the Gott15 scheme is that it can be adapted to the scale of the GCM grid column, which makes it more universal in usage. We have, however, seen that the results for the Gott15 scheme are better than for the more involved Dor15 scheme. The main difference between the two methods is that the Gott15 scheme has been trained with the macroscopic data, i.e., averaged σ_c over the entire radar domain, while the Dor15 scheme has been trained on a finer scale: individual radar pixels. The Gott15 scheme works with only one CMC that directly yields σ_c corresponding to the size of the radar domain, while the Dor15 scheme works with N CMCs for which each CMC corresponds to the size of a convective updraft and σ_c is calculated later with (5.1) and (5.2). The main reason why Gott15 performs better, is that it implicitly inferred spatial interactions between neighboring radar pixels, which are not captured by the independently evolving CMCs of Dor15. This could also be the reason that the Gott15 scheme is less intermittent than the multicloud schemes (Fig. 5.2) and has a more realistic ACF (Fig. 5.4a).

Including local interactions between neighboring *cells* in the Dor15 model could improve its performance, but lies beyond the scope of this paper. Including spatial interactions makes the model more complicated, because for every configuration of the neighboring cells a different transition probability matrix is needed. For successful inclusion of spatial interaction we refer to [9] in which a cellular automata approach (deterministic and stochastic) is applied to make convection interact spatially between different grid boxes of a NWP model, leading to a more realistic representation of convective organization. Further, in [34] locally interacting CMCs have been inferred from LES data and in [71] the multicloud model of [72] is extended by including spatial dependencies.

The Dor15 multicloud model is inspired by the multicloud model of [72]. The models are similar because in both models CMCs are positioned on a micro grid and randomly switch cloud type with probabilities that depend on the large-scale forcing. The main difference between the models is that the transition probabilities of the Dor15 scheme are estimated from data while the transition probabilities used in [72] are derived by choosing typical time scales of formation of clouds, conversion between cloud types and decay of clouds which are based on physical intuition. Furthermore, in the multicloud model of [72] probabilities are conditioned

Table 5.1: Computational costs (seconds per model day) of SPEEDY with the several closures compared to CTRL. The third column shows the number of random numbers that has to be generated each model day. The last column shows the number of random numbers that has to be generated for each model column (4,608 columns) each time step (ten minutes). Calculations are performed on a PC with 7.7-GB memory and a 2.5-GHz processor.

Scheme	Computational costs (seconds per model day)	No. random numbers per day	No. random numbers per column per time step
CTRL	3.4	0	0
Dor15-100	5.0	$6.6 \cdot 10^7$	100
Dor15-500	11.5	$3.3 \cdot 10^8$	500
Gott15	3.4	$6.6 \cdot 10^5$	1

on CAPE and middle troposphere dryness instead of large-scale vertical velocity for Dor15. In [72], a stochastic coarse-grained birth-death system is derived for the multicloud model, such that each GCM column only uses one CMC, which makes the method very effective. Further, the model of [72] is scale-aware because the number of lattice sites in the the micro grid can be adapted to the GCM grid box size. We conclude that the beneficial properties of both methods could be combined to obtain an even better model. Especially the inclusion of spatial dependencies as in the extension in [71] is promising.

In some recent studies [1, 116], new convection parameterizations have been implemented in aqua-planet GCMs. SPEEDY can also run in aqua-planet mode, but for comparison to satellite observations, we have chosen to include land in the experiments.

A final remark on computational costs of the new stochastic schemes. The multicloud scheme, for which N CMCs have to be evolved for each grid column (including the generation of random numbers) increases the computational costs of the convective scheme substantially, while the computational burden of Gott15 is marginal. In Table 5.1, we list these computational costs. In GCMs with a large number of grid columns, using a large number of CMCs ($N > 100$) for each column, could become computationally problematic. [72] showed that the usage of birth-death-like processes, with the same characteristics, is a solution to this problem.

5.9 Acknowledgment

The GPI data set was provided by the NASA/Goddard Space Flight Center's Mesoscale Atmospheric Processes Laboratory. This research was supported by the Division for Earth and Life Sciences (ALW) with financial aid from the Netherlands Organization for Scientific Research (NWO). We are grateful to three anonymous reviewers for helpful comments.

Chapter VI
Epilogue

"Every bit of reality,
always starts with a dream."
Pete Philly

And … it was a dream of Edward Lorenz (1917-2008), mathematician and meteorologist, well-known as the inventor of chaos theory and the butterfly effect, to include random numbers in the equations of weather and climate prediction models [90, 110].

Random numbers have been used; however, stochastic weather and climate modeling is still in its infancy. This is reflected, for instance, in the fact that the use of stochastics has not yet been fine-tuned to individual parts of the prediction models. Scientists are now developing stochastic models and approaches for different parts of the prediction models (e.g., the dynamical core, the physical tendencies, the ocean, the atmosphere, the orography) such that random numbers are included in a realistic way.

The presented work in this dissertation sheds new light on how and why stochastic methods can and should be used for the representation of convection and clouds. In particular, the use of parameterization methods based on conditional Markov chains (CMCs) has been examined in detail. In each chapter, this has been done in a different setting. Stochastic representations of shallow and deep convection have been examined with a Large-Eddy Simulation (LES) model, and LES data has been used to infer Markov-chain models. Besides simulations of convection with LES, also observations from a rain radar have been used to examine deep convection and its associated clouds and Markov-chain based parameterizations have been inferred from the observations as well. Furthermore, it has been shown that Markov-chain based models can indeed be used to improve the representation of convection in a general circulation model (GCM) of intermediate complexity; the climate model SPEEDY.

This chapter presents detailed conclusions, a synthesis and an outlook for future studies.

6.1 Conclusions

It has been demonstrated that the stochastic subgrid parameterization approach of Crommelin & Vanden-Eijden (2008) [25], that had proven its usability in the Lorenz '96 model setting before, can also be used to construct parameterizations of shallow cumulus convection. CMC models have been inferred directly from LES data. In particular, the CMCs were able to stochastically produce vertical profiles of turbulent fluxes; as observed in an LES data set obtained from a simulation of shallow convection. The CMCs switched between a finite number of states and were conditioned on a finite number of resolved-scale states. In order to do so, discretizations of the subgrid-scale variables and the resolved-scale variables were needed. It has been shown that these discretizations could be obtained by using the clustering method k-means. Entire vertical profiles of the heat and moisture fluxes could be clustered simultaneously to obtain a finite number of representative pairs of heat and moisture profiles. Also, the resolved-scale variables, in particular: vertical profiles of heat and moisture, have been clustered in a similar way. Transition probabilities of the CMCs have been estimated from the LES data. The parameterization has been demonstrated to adequately produce fluxes in a single-column model (SCM) in which the turbulent fluxes, produced by the Markov chains, counterbalanced the large-scale forcings. The CMCs outperformed the Markov chains that were not conditioned on the large-scale variables; therefore, it could be concluded that conditioning on the large-scale variables improved the Markov-chain model. The random fluctuations around the expectation values of the heat and moisture fluxes could be captured quite well. However, a drawback of using a finite number of representative turbulent flux profiles was found to be that the standard deviation of the fluxes decreased compared to the standard deviation of the LES fluxes. A moment-preserving clustering method could solve this problem, but this was not further examined. Another drawback was that this parameterization only works for atmospheric circumstances that are similar to the atmosphere as observed in the field-experiment Barbados Oceanographic and Meteorological Experiment (BOMEX). In theory, this particular approach could be generalized to construct shallow convection parameterizations that can be used globally. This would, however, be a very complex task. In order to do so, in future studies, LES simulations could be done for a sufficiently large range of atmospheric circumstances with shallow convection. It is questionable whether with this particular approach a sufficiently large set of representative vertical turbulent flux profiles can be constructed that can be used to parameterize shallow convection in a GCM.

Furthermore, it has been clearly demonstrated that stochastic convection parameterizations can be useful for a range of resolutions that is larger than the Grey Zone. In the Grey Zone, the resolution is such that the resolved and unresolved tendencies of heat, moisture and momentum are of the same order, with as a result strongly fluctuating unresolved fluxes. Therefore, stochastic methods are useful in the Grey Zone. Fig. 2.2 showed that the standard deviation of the unresolved fluxes decreases slowly for coarser resolutions and that it is still significantly large for horizontal model resolutions outside the Grey Zone. Therefore, stochas-

tic methods could be useful for a much larger range of resolutions than the Grey Zone resolutions. These results were obtained by coarse-graining (i.e., averaging) LES model data over subdomains of increasing horizontal size and decomposing fluxes into a resolved and unresolved part. In future studies, this could be done for simulations of deep convection as well to better visualize the Grey Zone for deep convection.

The work presented in Chapters 2 and 3, showed examples of the approach of running a high-resolution or micro-scale model (e.g., LES) to infer Markov-chain models that can be used as representations of unresolved processes (e.g., convection) of a larger-scale or macro-scale model (e.g., GCM). The particular implementation of the approach - e.g., using clustering techniques to obtain the states of the Markov chains - could be applied to other multi-scale problems that are common in other fields as well. The approach in Chapters 2 and 3 can be fit into the Heterogeneous Multi-scale Methods (HMM) [38] framework by taking the GCM as the macroscopic model and LES as the microscopic model. The approach in this dissertation is an example of serial coupling: the microscopic model outputs were pre-computed and were used later in a macroscopic model. Concurrent coupling, for which the micro and macroscopic models are run at the same time, would be possible too and could be an interesting topic for future studies.

Although, the method of clustering entire vertical profiles of the turbulent heat and moisture fluxes, as performed in Chapter 2, was successful for shallow convection parameterization, this particular approach was not used in Chapters 3-5. A switch was made to the usage of multicloud models. In the multicloud models, the states of the Markov chains were cloud types, which turned out to be a major simplification compared to the vertical flux profiles. The multicloud models contained a number of CMCs per GCM column, the CMCs switched between different cloud types and produced cloud type area fractions. The transition probabilities could be inferred from high-resolution model data and from observations as well. An asset of the multicloud models compared to the flux profile method, was that the cloud type area fractions could be implemented in convection parameterizations in GCMs in a straightforward manner. For example, the convective area fraction could be used in the closure for the mass flux at cloud base (Chapter 5). This would be complicated to do with vertical flux profiles. Furthermore, the multicloud models could be inferred directly from observations of convective clouds (Chapter 4), which would be complicated to do for vertical flux profiles.

Another important asset of the multicloud model was that it could be adapted to the size of the GCM column; it was scale-adaptive. Each Markov chain corresponded to a certain size, and therefore, the size of the GCM model column determined the number of Markov chains. A smaller number of CMCs per model column resulted in stronger fluctuations around expectation values of the convective area fraction. A key issue turned out to be the lack of spatial coupling between the CMCs in the multicloud model. It caused the convective area fractions to be too intermittent, and the number of CMCs used in each model column had to be lowered artificially in order to obtain the desired standard deviation of the convective area fractions. Introducing spatial coupling between neighboring CMCs explicitly, was a solution to this particular problem. In Chapter 3, it has been demonstrated

that the CMCs in the multicloud model can be made dependent on the neighboring CMCs. This resulted in stochastic cellular automata that were able to produce spatial organization on the micro grid. In case the spatial structures as observed in LES were captured, also the standard deviation of the cloud type area fractions could be simulated in a more realistic way than without spatial coupling. A key issue was found to be the large number of neighboring configurations, which made the estimation of the transition probabilities complicated. This problem was solved by reducing the number of neighboring configurations by only counting the number of cloud types and not the exact neighboring configuration (Chapter 3). Introducing spatial coupling between data-inferred CMCs inside a GCM grid box, is an interesting field of almost unexplored research and it is a promising approach for improving convection and cloud parameterizations in weather and climate models.

The main motivation to switch from using LES simulations to observations was the limitation of the simulation time period. At the time that the deep convection simulation was performed (2013), doing such simulations was computationally demanding, and therefore simulations were relatively short (~8h). An LES simulation of months, in which a large range of atmospheric circumstances was covered, was not possible and therefore the step to the Darwin observations was made. A first asset of using observations was the large time period of the data set. This is expected to improve for LES simulations in the future, because computational resources increase. Furthermore, also the horizontal size of the domain was much larger for the radar observations (172×172 km^2) than for LES (58×58 km^2). This is also expected to improve for LES simulations in the future. Another important asset of observations of a rain radar was that it gave an accurate representation of the atmosphere in the region, however, observations are subjected to measurement errors. A drawback of using observations was that data was missing for some time intervals and that the data sets were unstructured, while at the other hand the LES data was well structured and no data points were missing. Furthermore, with LES simulations almost all variables that were desired could be obtained, while for the observations, only high-resolution data of cloud top and rain rate were available. It can be expected that in the future more variables are available. It can be concluded that both sources of data have their own assets and drawbacks and that for each particular situation, a choice can be made between the two, or a combination of the two sources is an option as well.

Using the Darwin radar data observations in combination with the large-scale analysis data (prepared by Davies et al. (2013) [28]), it has been demonstrated that the large-scale vertical velocity $\langle\omega\rangle$ correlates strongly with deep convection. The cross-correlation diagram (Fig. 4.4) showed that the correlation was the strongest for a time lag of around three hours for $\langle\omega\rangle$, which means that the time series of $\langle\omega\rangle$ had to be shifted back three hours to obtain the highest correlation with the time series of convection. This suggested that $\langle\omega\rangle$ was only an effect of convection. However, since the correlation was already strong for zero time lag, $\langle\omega\rangle$ could be used for conditioning the Markov chains; and consequently it could be incorporated in deep convection schemes of GCMs. The correlation of CAPE turned out to be much weaker and therefore, the choice of using $\langle\omega\rangle$ as an indicator instead of CAPE was justified. Concluding that CAPE is not a useful indicator of convection

would be incorrect, of course it is a measure for the potential and strength of deep convection. Using an instability measure for the initiation of convection and using $\langle\omega\rangle$ to stochastically determine the intensity of convection through the mass flux at cloud base, turned out to give realistic time series of the mass flux at cloud base in SPEEDY.

The simulation of convectively coupled equatorial waves has an effect on the medium range weather forecast global skill and is not confined to the skill of GCMs in the tropics. Therefore, the correct simulation of these waves is of major importance. With the assessment method described in Chapter 5 it was possible to express the model's skill to simulate a convectively coupled equatorial wave in a single scalar. The expression of the power of the Madden-Julian oscillation (MJO) and equatorial Kelvin waves in scalars, introduced in Chapter 5, was a novel idea. It was found to be a powerful indicator of the GCM's ability to simulate these large-scale phenomena. This method that has been used to assess equatorial waves in SPEEDY could also be used in state-of-the-art GCMs. In SPEEDY, the average mass flux at cloud base was found to have an effect on the strength of the MJO and the equatorial Kelvin waves. The MJO was weaker for larger mean values of the mass flux at cloud base in which case the Kelvin waves were stronger. In SPEEDY, the effect of stochastics in the convection scheme on the equatorial waves were found to be minor, despite the improvement of the time series of the cloud base mass flux at time step level. This could be a feature of SPEEDY, because it is a GCM of intermediate complexity, and could be tested in state-of-the-art GCMs in future studies.

6.2 Synthesis

In the presented chapters, convection parameterizations have been examined in different settings and with different approaches. First of all, stochastic schemes have been compared with deterministic schemes. The stochastic convection schemes were found to give more realistic estimates of the fluctuations of the convective area fraction, and consequently, of the cloud base mass flux around the expected values, without corrupting the mean cloud base mass flux.

The main stochastic schemes that have been examined are (i) schemes that stochastically determined the vertical profiles of the turbulent heat and moisture flux, (ii) the multicloud models that stochastically determined the cloud type area fractions of which for example the cloud base mass flux could be derived and (iii) a scheme that directly determined the cloud base mass flux. All schemes were Markov-chain based, the main difference lied in the states of the Markov chains. In (i), the states of the Markov chains were pairs of entire vertical profiles of heat and moisture fluxes, in (ii) the states were cloud types and in (iii) the states were cloud base mass fluxes. The choice to make a step from (i), in Chapter 2, to the multicloud model (ii), in Chapter 3-5, has been made because:

- (ii) was easier to construct from data than (i), because cloud types were easier to determine than vertical flux profiles;
- (ii) was more generally applicable than (i): the vertical flux profiles were con-

strained to the BOMEX case, while the convective area fractions generated by the multicloud scheme could yield a larger range of vertical flux profiles when used in a cloud or updraft parcel model; and

- (ii) was easier to implement in a GCM; cloud type area fractions could be used directly in GCMs. For example, the deep convective area fraction could be used in the closure for the mass flux at cloud base in GCMs with a mass flux scheme, and the shallow convective area fraction analogously. Other cloud type area fractions could be used in the determination of the cloud cover.

Furthermore, scheme (ii) had certain advantages over scheme (iii): it was scale-adaptive, i.e., it could be adapted to the size of the GCM column, it could in theory be used to produce 3 types of moist convective area fractions (shallow, congestus, deep) that are entirely compatible with each other, spatial correlation between the Markov chains could be introduced to capture spatial organization inside a grid column. An advantage of (iii) compared to (ii) was that it only used one Markov chain per GCM grid column and that it captured spatial organization inside a grid column by construction (since it had been trained with data averaged over a large area). In conclusion, (ii) was the most promising model compared to (i) and (iii), because of the important advantages it had over the other schemes.

Conditioning of the Markov chains by making the transition probabilities dependent on the large-scale variables has been examined in all chapters. A common approach was the use of the clustering method k-means. The main differences were the type of large-scale variables that were used to condition on, the number of large-scale variables to condition on, and the number of clusters. Another difference was the way how these large-scale variables were found. It is possible to condition on pairs of entire vertical profiles of heat and moisture (Chapter 2) or on indicators of convection (Chapters 3-5). The main advantage of using an indicator of convection is that it is a single scalar which is easier to cluster. Furthermore, generally accepted indicators of convection such as CAPE have proven their usability in models for decades, which is not the case for entire vertical profiles. The best indicator can be found by using the method of relative entropy or by correlation analysis. In Chapter 3, examination of the strongest indicator of deep convection has been done by using the method of calculating the relative entropy between deep convective area fractions and the large-scale indicator of convection, while on the other hand in Chapter 4, examination of the strongest indicator of convection has been done with correlation analysis. This choice has been made because, while examining the radar observations, it was found that both methods gave the same results in case only one indicator was assessed and in that case, correlation analysis was the easier approach. In case a combination of variables is assessed, the method of clustering and relative entropy may be the better option. This could be examined in more detail in future studies.

Clustering is an effective way of reducing the spaces of large-scale and small-scale variables in a finite number of classes. Pairs of entire vertical profiles can be clustered. Furthermore, combinations of large-scale variables can be clustered. An advantage of clustering over binning is that by clustering, the data is divided over classes in an optimal way (i.e., the clustering method aims to minimize the distance

to the nearest centroids). For example if two variables are binned, some of the bins could be empty, which would not be optimal. Furthermore, in case the indicator is distributed non-uniformly, the clustering method will automatically use clusters of unequal sizes. Of course, this can be done by hand, by choosing non-equidistant thresholds, but this may be a difficult task. In case the small-scale variables are cloud types, there are physical reasons to choose certain thresholds, in which case a clustering method would be less effective.

By using Markov chains for determination of the cloud base mass flux, the convection scheme turns into a prognostic scheme. This means that besides that the large-scale variables, also the convective state of the previous time step determines the outcomes of the convection scheme in a GCM column. This has to be kept in mind when introducing Markov chains in the (non-prognostic) convection scheme of a host GCM; if an improvement is found in the GCMs predictive skill, this may not be due to the fact that the convection scheme is stochastic, but due to the fact that the convection scheme has been turned into a prognostic scheme. With a prognostic scheme, the effects of convection on the resolved model variables can in theory be estimated more accurately, since it could capture the time-correlation that is present in convection in a large area (Fig. 4.11).

Assessment of the new parameterizations of convection has been done in SCMs and in a GCM of intermediate complexity. SCMs have been used in the past to assess new parameterizations. At the moment GCMs of intermediate complexity can be run many times without a large computational overhead, since computer speed has increased drastically. Testing in a GCM has some advantages over testing in an SCM. In a GCM, also if it is of intermediate complexity, a large range of large-scale atmospheric circumstances can be covered, while in an SCM this range is in general smaller. This means that the behavior of the new scheme can be tested for a large range of atmospheric circumstances, a range that is similar to the range of circumstances under which it has to work in a state-of-the-art GCM. Furthermore, interaction of the new scheme with the GCM resolved variables can be tested at scales that are larger than a single column. For example, the interaction of convection with convectively coupled equatorial waves can be tested with a GCM and not with an SCM. On the other hand, advantages of testing a new parameterization scheme in an SCM are that the scheme is tested in a constrained environment without interactions with the resolved-scale flows; that it is computationally inexpensive; that it can be tested in the SCM version of the host GCM; and that the range of atmospheric circumstances can be chosen to be limited, in the case that the new parameterization scheme is only designed for a limited range of atmospheric circumstances (e.g., the scheme in Chapter 2).

6.3 Outlook

Future studies should elaborate on the multicloud models presented in this dissertation, on multicloud models presented by other authors (e.g., [1, 30, 35, 37, 42, 43, 72, 73, 113]) and on multiplume models [104, 114]. Multicloud models have a large potential to improve convection and cloud parameterizations, because the models can be:

- inferred directly from high-resolution model simulation data and from observations as well;
- made scale-adaptive by adapting the number of cells that form cloud type area fractions;
- implemented in convection schemes of GCMs in a straightforward manner. Shallow and deep convective area fractions can be used in the closures of the cloud base mass flux. Moreover, the other cloud fractions could be used in the cloud schemes (e.g., to determine the cloud cover in a model column);
- conditioned on the large-scale variables of the GCM, such that interaction between the GCM and the multicloud model is possible;
- extended such that spatial interaction between cells in the micro-grid is present.

Almost the same list of assets could be made for the multiplume models, which are similar to multicloud models.

A next step would be the examination of spatial coupling of the cells in the multicloud model [36, 71]. These stochastic cellular automata could be used to capture spatial structures of convection and clouds inside GCM grid columns more realistically, resulting in more realistic cloud type area fractions and hence more realistic time series of the mass flux at cloud base. More realistic convection schemes will lead to more accurate weather and climate modeling and predictions.

Future studies with high-resolution simulations (e.g., LES, cloud resolving models) and observations (e.g., rain radar, satellite) can be used to improve existing multicloud models. Data from different locations at Earth can be used to improve its accuracy. It is also possible to combine studies with simulations and observations, for example: in the Darwin radar multicloud model, shallow cumulus clouds are not included, these can be added by doing additional simulations with LES.

An important feature of the presented schemes in this dissertation, and possibly the most important feature, is the usage of data (in particular: observations) to directly infer probabilities of the stochastic schemes. Using the presented data-driven techniques to construct future (stochastic) parameterizations is highly recommended.

References

References

[1] R. S. Ajayamohan, B. Khouider, and A.J. Majda. Simulation of monsoon intraseasonal oscillations in a coarse-resolution aquaplanet GCM. *Geophys. Res. Lett.*, 41:5662–5669, 2014.

[2] T.W. Anderson and L.A. Goodman. Statistical inference about markov chains. *Ann. Math. Statist.*, 28:89–110, 1957.

[3] A. Arakawa. The cumulus parameterization problem: Past, present, and future. *J. Climate*, 17:2493–2525, 2004.

[4] A. Arakawa, J.-H. Jung, and C.-M. Wu. Toward unification of the multiscale modeling of the atmosphere. *Atmos. Chem. Phys.*, 11:3731–3742, 2011.

[5] A. Arakawa and W.H. Schubert. Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *J. Atmos. Sci.*, 31:674–701, 1974.

[6] H.M. Arnold, I.M. Moroz, and T.N. Palmer. Stochastic parametrizations and model uncertainty in the Lorenz '96 system. *Phil. Trans. R. Soc. A*, 371, 2013.

[7] D. Arthur and S. Vassilvitskii. k-means++: The Advantages of Careful Seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[8] L. Bengtsson, H. Körnich, E. Källén, and E. Svensson. Large-scale dynamical response to subgrid-scale organization provided by cellular automata. *J. Atmos. Sci.*, 68:3132–3144, 2011.

[9] L. Bengtsson, M. Steinheimer, P. Bechtold, and J.-F. Geleyn. A stochastic parametrization for deep convection using cellular automata. *Q.J.R. Meteorol. Soc.*, 139:1533–1543, 2013.

[10] J.W. Bergman and P.D. Sardeshmukh. Dynamic Stabilization of Atmospheric Single Column Models. *J. Climate*, 17:1004–1021, 2004.

[11] J. Berner, F.J. Doblas-Reyes, T.N. Palmer, G. Shutts, and A. Weisheimer. Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Phil. Trans. R. Soc. A*, 366:2559–2577, 2008.

[12] L.C. Berselli, T. Iliescu, and W.J. Layton. *Mathematics of Large Eddy Simulation of Turbulent Flows*. Springer, Berlin, Heidelberg, 2006.

[13] J.A. Biello and A.J. Majda. A New Multiscale Model for the Madden-Julian Oscillation. *J. Atmos. Sci.*, 62:1694–1721, 2005.

[14] J.A. Biello and A.J. Majda. Intraseasonal multi-scale moist dynamics of the tropical troposphere. *Commun. Math. Sci.*, 8:519–540, 2010.

[15] V. Bjerknes. Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik. *Meteorol. Z.*, 21:1–7, 1904.

[16] M. Blackburn, D.L. Williamson, K. Nakajima, W. Ohfuchi, Y.O. Takashi, Y.Y. Hayashi, H. Nakamura, M. Ishiwatarj, J.L. McGregor, H. Borth, V. Wirth, H. Frank, P. Bechtold, N.P. Wedi, H.Tomita, M. Satoh, M. Zhao, I.M. Held, M.J. Suarez, M.-I. Lee, M. Watanabe, M. Kimoto, Y. Liu, Z. Wang, A. Molod, K. Rajen Dran, A. Kitoh, and Rachel Stratton. The aqua-planet experiment (APE): CONTROL SST simulation. *J. Meteor. Soc. Japan*, 91, 2013.

[17] S.J. Böing, H.J.J. Jonker, A.P. Siebesma, and W. Grabowski. Influence of the subcloud layer on the development of a deep convective ensemble. *J. Atmos. Sci.*, 69:2682–2698, 2012.

[18] S. Bony and J.L. Dufresne. Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models. *Geophys. Res. Lett.*, 32:L20806, 2005.

[19] S. Bony, B. Stevens, D.M.W. Frierson, C. Jakob, M. Kageyama, R. Pincus, T.G. Shepherd, A.H. Sobel M. Watanabe S.C. Sherwood, A.P. Siebesma, and M.J. Webb. Clouds, circulation and climate sensitivity. *Nat. Geosci.*, 8:261–268, 2015.

[20] R. Buizza, M. Milleer, and T.N. Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q.J.R. Meteorol. Soc.*, 125:2887–2908, 1999.

[21] L. Caffarelli, R. Kohn, and L. Nirenberg. Partial Regularity of Suitable Weak Solutions of the Navier-Stokes equations. *Comm. on pure and applied math.*, 35, 1982.

[22] A.C. Clement, R. Burgman, and J.R. Norris. Observational and Model Evidence for Positive Low-Level Cloud Feedback. *Science*, 325, 2009.

[23] W.D. Collins, P.J. Rasch, B.A. Boville, J.J. Hack, J.R. McCaa, D.L. Williamson, B.P. Briegleb, C.M. Bitz, S.-J. Lin, and M. Zhang. The Formulation and Atmospheric Simulation of the Community Atmosphere Model Version 3 (CAM3). *J. Climate*, 19:2144–2161, 2006.

[24] T.M. Cover and J.A. Thomas. *Elements of information theory*. 2nd edn. Hoboken, NJ: John Wiley and Sons., 1991.

[25] D. Crommelin and E. Vanden-Eijnden. Subgrid-Scale Parameterization with Conditional Markov Chains. *J. Atmos. Sci.*, 65:2661–2675, 2008.

[26] D.T. Crommelin and E. Vanden-Eijnden. Fitting timeseries by continuous-time markov chains: A quadratic programming approach. *J. Comput. Phys.*, 217(2):782–805, 2006.

[27] B. Cushman-Roisin and J.-M. Beckers. *Introduction to geophysical fluid dynamics: physical and numerical aspects*, volume 101. Academic Press, 2011.

[28] L. Davies, C. Jakob, K. Cheung, A. Del Genio, A. Hill, T. Hume, R.J. Keane, T. Komori, V.E. Larson, Y. Lin, B.J. Nielsen, J. Petch, R.S. Plant, M.S. Singh, X. Shi, X. Song, W. Wang, M.A. Whitall, A. Wolf, S. Xie, and G. Zhang. A single column model ensemble approach applied to the TWP-ICE experiment. *J. Geophys. Res.*, 118:6544–6563, 2013.

[29] L. Davies, C. Jakob, P. May, V.V. Kumar, and S. Xie. Relationships between the large-scale atmosphere and the small-scale convective state for Darwin, Australia. *J. Geophys. Res. Atmos.*, 118:534,11–545, 2013.

[30] M. de la Chevrotière, B. Khouider, and A.J. Majda. Calibration of the stochastic multicloud model using Bayesian Inference. *J. Sci. Comput.*, 36:B538–B560, 2014.

[31] W.C. de Rooy and A.P. Siebesma. A Simple Parameterization for Detrainment in Shallow Cumulus. *Mon. Wea. Rev.*, 136:560–576, 2008.

[32] Q. Deng, B. Khouider, and A.J. Majda. The MJO in a Coarse-Resolution GCM with a Stochastic Multicloud Parameterization. *J. Atmos. Sci.*, 72:55–74, 2015.

[33] Q. Deng, B. Khouider, and A.J. Majda. Effect of Stratiform Heating on the Planetary-Scale Organization of Tropical Convection. *J. Atmos. Sci.*, submitted.

[34] J. Dorrestijn, D.T. Crommelin, J.A. Biello, and S.J. Böing. A data-driven multi-cloud model for stochastic parametrization of deep convection. *Phil. Trans. R. Soc. A*, 371:20120374, 2013.

[35] J. Dorrestijn, D.T. Crommelin, A. Pier Siebesma, H.J.J. Jonker, and C. Jakob. Stochastic Parameterization of Convective Area Fractions with a Multicloud Model Inferred from Observational Data. *J. Atmos. Sci.*, 72:854–869, 2015.

[36] J. Dorrestijn, D.T. Crommelin, A.P. Siebesma, and H.J.J. Jonker. Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theor. Comput. Fluid Dyn.*, 27:133–148, 2013.

[37] J. Dorrestijn, D.T. Crommelin, A.P. Siebesma, H.J.J. Jonker, and F. Selten. Stochastic Convection Parameterization with Markov Chains in an Intermediate-Complexity GCM. *J. Atmos. Sci.*, 73:1367–1382, 2016.

[38] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: a review. *Commun. Comput. Phys*, 2:367–450, 2007.

[39] ECMWF. *Part IV: Physical Processes*. IFS Documentation. ECMWF, 2015. Operational implementation 12 May 2015.

[40] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. Evaluation of Climate Models. In T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 741–866. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[41] C. L.E. Franzke, T.J. O'Kane, J. Berner, P.D. Williams, and V. Lucarini. Stochastic climate theory and modeling. *Wiley Interdisciplinary Reviews: Climate Change*, 6(1):63–78, 2015.

[42] Y. Frenkel, A.J. Majda, and B. Khouider. Using the stochastic multicloud model for tropical convection. *J. Atmos. Sci.*, 69:1080–1105, 2012.

[43] Y. Frenkel, A.J. Majda, and B. Khouider. Stochastic and deterministic multicloud parameterizations for tropical convection. *Clim. Dyn.*, 41:1527–1551, 2013.

[44] D.M.W. Frierson. Convectively Coupled Kelvin Waves in an Idealized Moist General Circulation Model. *J. Atmos. Sci.*, 64:2076–2090, 2007.

[45] G. Gan, C. Ma, and J. Wu. *Data clustering: theory, algorithms, and applications*. SA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 3 edition, 2007.

[46] M. Gardner. Mathematical games: The fantastic combinations of John Conway's new solitaire game "life". *Scientific American*, 223, 1970.

[47] W.L. Gates. AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.*, 73:1962–1970, 1992.

[48] L. Gerard. An integrated package for subgrid convection, clouds and precipitation compatible with meso-gamma scales. *Q.J.R. Meteorol. Soc.*, 00:1–19, 2007.

[49] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics: model problems and algorithms. *Nonlinearity*, 17(6):R55, 2004.

[50] G.A. Gottwald, K. Peters, and L. Davies. A data-driven method for the stochastic parametrisation of subgrid-scale tropical convective area fraction. *Quarterly Journal of the Royal Meteorological Society*, 142(694):349–359, 2016.

[51] A.L.M. Grant. Cloud-base fluxes in the cumulus-capped boundary layer. *Q.J.R. Met. Soc.*, 127:407–421, 2001.

[52] O. Häggström. *Finite Markov chains and algorithmic applications*, volume 52. Cambridge University Press, 2002.

[53] U. Hammarstrand. Questions involving the use of traditional convection parameterization in NWP models with a higher resolution. *Tellus*, 50, 1998.

[54] J. Harlim and A.J. Majda. Test models for filtering with superparameterization. *Multiscale Model. Simul.*, 11, 2013.

[55] W. Hazeleger, X. Wang, C. Severijns, S. Stefanescu, R. Bintanja, A. Sterl, K. Wyser, T. Semmler, S. Yang, B. van den Hurk, T. van Noije, E. van der Linden, and K. van der Wiel. EC-Earth V2. 2: description and validation of a new seamless earth system prediction model. *Climate dyn.*, 39, 2012.

[56] T. Heus, C.C. van Heerwaarden, H.J.J. Jonker, A.P. Siebesma, S. Axelsen, K. van den Dries, O. Geoffroy, A.F. Moene, D. Pino, S.R. de Roode, and J. Vil'a-Guerau de Arellano. Formulation and numerical studies with the Dutch Atmospheric Large-Eddy Simulation (DALES). *Geosci. Model Dev.*, 3:415–444, 2010.

[57] C. Hohenegger and B. Stevens. Preconditioning deep convection with cumulus congestus. *J. Atmos. Sci.*, 70:448–464, 2013.

[58] J.Z. Holland and E.M. Rasmusson. Measurements of the Atmospheric Mass, Energy, and Momentum Budgets Over a 500-Kilometer Square of Tropical Ocean. *Mon. Wea. Rev.*, 101:44–55, 1973.

[59] A.A.M. Holtslag and C-H. Moeng. Eddy diffusivity and countergradient transport in the convective atmospheric boundary layer. *J. Atmos. Sci.*, 48:1690–1698, 1991.

[60] R. Honnert, V. Masson, and F. Couvreux. A diagnostic for evaluating the representation of turbulence in atmospheric models at the kilometric scale. *J. Atmos. Sci.*, 68:3112–3131, 2011.

[61] I. Horenko. On the Identification of Nonstationary Factor Models and Their Application to Atmospheric Data Analysis. *J. Atmos. Sci.*, 67:1559–1574, 2010.

[62] G.J. Huffman and D.T. Bolvin. Version 1.2 GPCP One-Degree Daily Precipitation Data Set Documentation. *WDC-A, NCDC, Asheville, NC., Data set accessed Sep. 2014*, 2013.

[63] C. Jakob. Accelerating progress in global atmosheric model development through improved parameterizations. *Bull. Am. Met. Soc.*, 91:869–875, 2010.

[64] C. Jakob and A.P. Siebesma. A New Subcloud Model for Mass-Flux Convection Schemes: Influence on Triggering, Updraft Properties, and Model Climate. *Mon. Wea. Rev.*, 131:2765–2778, 2003.

[65] R.H. Johnson, T.M. Rickenbach, S.A. Rutledge, P.E. Ciesielski, and W.H. Schubert. Trimodal Characteristics of Tropical Convection. *J. Climate*, 12:2397–2418, 1999.

[66] I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*, volume 113. Springer Science & Business Media, 2012.

[67] M.A. Katsoulakis, A.J. Majda, and A. Sopasakis. Intermittency, metastability and coarse graining for coupled deterministic–stochastic lattice systems. *Nonlinearity*, 19(5):1021, 2006.

[68] M.A. Katsoulakis and P.E. Souganidis. Stochastic Ising models and anisotropic front propagation. *J. Stat. Phys.*, 87(1-2):63–89, 1997.

[69] R.A. Kerr. Clouds Appear to Be Big, Bad Player in Global Warming. *Science*, 325, 2009.

[70] I.G. Kevrekidis and G. Samaey. Equation-free multiscale computation: Algorithms and applications. *Annual review of physical chemistry*, 60:321–344, 2009.

[71] B. Khouider. A coarse grained stochastic multi-type particle interacting model for tropical convection: Nearest neighbour interactions. *Commun. Math. Sci.*, 12:1379–1407, 2014.

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

[73] B. Khouider and A.J. Majda. A simple multicloud parameterization for convectively coupled tropical waves. I. Linear Analysis. *J. Atmos. Sci.*, 63:1308–1323, 2006.

[74] B. Khouider, A.J. Majda, and A. Katsoulakis. Coarse grained stochastic models for tropical convection and climate. *Proc. Natl. Acad. Sci.*, 100:11941–11946, 2003.

[75] B. Khouider, A. St-Cyr, A.J. Majda, and J. Tribbia. The MJO and convectively coupled waves in a coarse-resolution GCM with a simple multicloud parameterization. *J. Atmos. Sci.*, 68:240–264, 2011.

[76] G.N. Kiladis, M.C. Wheeler, P. T. Haertel, K. H. Straub, and P. E. Roundy. Convectively coupled equatorial waves. *Rev. Geophys.*, 47:RG2003, 2009.

[77] F. Kucharski, F. Molteni, M.P. King, R. Farneti, I.-S. Kang, and L. Feudale. On the need of intermediate complexity general circulation models: a “SPEEDY” example. *Bull. Amer. Meteor. Soc.*, 94:25–30, 2013.

[78] J.P. Kuettner. Cloud bands in the earth’s atmosphere, Observations and Theory. *Tellus*, 23:404–425, 1971.

[79] V.V. Kumar, C. Jakob, A. Protat, P.T. May, and L. Davies. The four cumulus cloud modes and their progression during rainfall events: A C-band polarimetric radar perspective. *J. Geophys. Res. Atmos.*, 118:8375–8389, 2013.

[80] H. Kuo. On Formation and Intensification of Tropical Cyclones Through Latent Heat Release by Cumulus Convection. *J. Atmos. Sci.*, 222:40–63, 1965.

[81] F. Kwasniok. Data-based stochastic subgrid-scale parametrisation: an approach using cluster-weighted modelling. *Phil. Trans. R. Soc. A.*, 370:1061–1086, 2012.

[82] S. Lang, W.-K. Tao, J. Simpson, and B. Ferrier. Modeling of Convective-Stratiform Precipitation Processes: Sensitivity to Partitioning Methods. *J. Appl. Meteor.*, 42:505–527, 2003.

[83] J. L. Lebowitz, E. Orlandi, and E. Presutti. A particle model for spinodal decomposition. *J. Stat. Phys.*, 63(5-6):933–974, 1991.

[84] J.L. Lebowitz, E. Orlandi, and E. Presutti. Convergence of stochastic cellular automation to Burgers' equation: Fluctuations and stability. *Physica D*, 33(1):165–188, 1988.

[85] J.-L. Lin, G.N. Kiladis, B.E. Mapes, K.M. Weickmann, K.R. Sperber, W. Lin, M.C. Wheeler, S.D. Schubert, A. Del Genio, L.J. Donner, S. Emori, J.F. Gueremy, F. Hourdin, P.J. Rasch, E. Roeckner, and J.F. Scinocca. Tropical Intraseasonal Variability in 14 IPCC AR4 Climate Models. Part I: Convective Signals. *J. Climate*, 19:2665–2690, 2006.

[86] J.W.-B. Lin and J.D. Neelin. Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.*, 27:3691–3694, 2000.

[87] J.W.-B. Lin and J.D. Neelin. Considerations for stochastic convective parameterization. *J. Atmos. Sci.*, 59:959–975, 2002.

[88] J.W.-B. Lin and J.D. Neelin. Toward stochastic deep convective parameterization in general circulation models. *Geophys. Res. Lett.*, 30:1162, 2003.

[89] E.N. Lorenz. The predictability of a flow which possesses many scales of motion. *Tellus*, 21:289–307, 1969.

[90] E.N. Lorenz. Climatic predictability. *The Physical Basis of Climate and Climate Modelling. WMO GARP Publication*, 16, 1975.

[91] E.N. Lorenz. Predictability - a problem partly solved. In *Proceedings: Seminar on Predictability*, pages 1–18, ECMWF, Reading, United Kingdom, 1996.

[92] J.B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Math. Stat. Probab. 1*, pages 281–297, Statistical Laboratory of the University of California, Berkeley, 1967.

[93] A.J. Majda and B. Khouider. Stochastic and mesoscopic models for tropical convection. *Proc. Natl. Acad. Sci.*, 99:1123–1128, 2002.

[94] A.J. Majda, S.N. Stechmann, and B. Khouider. Madden–Julian oscillation analog and intraseasonal variability in a multicloud model above the equator. *Proc. Natl. Acad. Sci. USA*, 104:9919–9924, 2007.

[95] B. Mapes and R. Neale. Parameterizing convective organization to escape the entrainment dilemma. *J. Adv. Model. Earth Syst.*, 3:1–20, 2011.

[96] B. Mapes, S. Tulich, J. Lin, and P. Zuidema. The mesoscale convective life cycle: building block or prototype for large-scale tropical waves? *Dyn. Atmos. Oceans*, 42:3–29, 2006.

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

[98] P.T. May and A. Ballinger. The Statistical Characteristics of Convective Cells in a Monsoon Regime (Darwin, Northern Australia). *Mon. Weather Rev.*, 135:82–92, 2007.

[99] T.P. Meyer, F.C. Richards, and N.H. Packard. Learning algorithm for modeling complex spatial dynamics. *Phys. Rev. Lett.*, 63:1735–1738, 1989.

[100] B. Möbis and B. Stevens. Factors controlling the position of the Intertropical Convergence Zone on an aquaplanet. *J. Adv. Model. Earth Syst.*, 4:M00A04, 2012.

[101] F. Molteni. Atmospheric simulations using a GCM with simplified physical parametrizations. I: Model climatology and variability in multi-decadal experiments. *Clim. Dyn.*, 20:175–191, 2003.

[102] F. Molteni and F. Kucharski. Description of the ICTP AGCM (SPEEDY) Version 41, 2013.

[103] R.A.J. Neggers, H.J.J. Jonker, and A.P. Siebesma. Size Statistics of Cumulus Cloud Populations in Large-Eddy Simulations. *J. Atmos. Sci.*, 60:1060–1074, 2003.

[104] R.A.J. Neggers, M. Köhler, and A.C.M. Beljaars. A Dual Mass Flux Framework for Boundary Layer Convection. Part I: Transport. *J. Atmos. Sci.*, 66:1465–1487, 2009.

[105] R.A.J. Neggers, A.P. Siebesma, G. Lenderink, and A.A.M. Holtslag. An Evaluation of Mass Flux Closures for Diurnal Cycles of Shallow Cumulus. *Mon. Wea. Rev.*, 132:2525–2537, 2004.

[106] K. Nimsaila and I. Timofeyev. Markov chain stochastic parameterizations of essential variables. *Multiscale Model. Simul.*, 8:2079–2096, 2010.

[107] J.R. Norris. *Markov chains*. Number 2008. Cambridge university press, 1998.

[108] T. N. Palmer and P. Williams. *Stochastic physics and climate modelling*. Cambridge Univ. Press, Cambridge, UK, 2010.

[109] T.N. Palmer. 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.

[110] T.N. Palmer. Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Q. J. Roy. Meteor. Soc.*, 138, 2012.

[111] G.A. Pavliotis and A. Stuart. *Multiscale methods: averaging and homogenization*. Springer Science & Business Media, 2008.

[112] J. Pergaud, V. Masson, S. Malardel, and F. Couvreux. A Parameterization of Dry Thermals and Shallow Cumuli for Mesoscale Numerical Weather Prediction. *Boundary-Layer Met.*, 132:83–106, 2009.

[113] K. Peters, C. Jakob, L. Davies, B. Khouider, and A.J. Majda. Stochastic behavior of tropical convection in observations and a multicloud model. *J. Atmos. Sci.*, 70:3556–3575, 2013.

[114] R.S. Plant and G.C. Craig. A Stochastic Parameterization for Deep Convection Based on Equilibrium Statistics. *J. Atmos. Sci.*, 65:87–105, 2008.

[115] W. H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge Univ. Press, 1992.

[116] F. Ragone, K. Fraedrich, H. Borth, and F. Lunkeit. Coupling a minimal stochastic lattice gas model of a cloud system to an atmospheric general circulation model. *Q.J.R. Meteorol. Soc.*, 2014.

[117] D. Randall, M. Khairoutdinov, A. Arakawa, and W. Grabowski. Breaking the Cloud Parameterization Deadlock. *Bull. Amer. Meteor. Soc.*, 84:1547–1564, 2003.

[118] D.A. Randall, Harshvardhan, D.A. Dazlich, and T.G. Corsetti. Interactions among Radiation, Convection, and Large-Scale Dynamics in a General Circulation Model. *J. Atmos. Sci.*, 46:1943–1970, 1989.

[119] F.C. Richards, T.P. Meyer, and N.H. packard. Extracting cellular automaton rules directly from experimental data. *Physica*, 45:189–202, 1990.

[120] K. Riemann-Campe, K. Fraedrich, and F. Lunkeit. Global climatology of convective available potential energy (CAPE) and convective inhibition (CIN) in ERA-40 reanalysis. *J. Atmos. Sci.*, 93:534–545, 2009.

[121] D.M. Romps and Z. Kuang. Nature versus nurture in shallow convection. *J. Atmos. Sci.*, 67:1655–1666, 2010.

[122] S. Sahany, J.D. Hales, and R.B. Neale. Temperature-Moisture Dependence of the Deep Convective Transition as a Constraint on Entrainment in Climate Models. *J. Atmos. Sci.*, 69:1340–1358, 2012.

[123] M. Sakradzija, A. Seifert, and T. Heus. Fluctuations in a quasi-stationary shallow cumulus cloud ensemble. *Nonlin. Processes Geophys. Discuss.*, 22:65–85, 2015.

[124] G.J. Shutts. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.*, 131:3079–3102, 2005.

[125] G.J. Shutts and T.N. Palmer. Convective Forcing Fluctuations in a Cloud-Resolving Model: Relevance to the Stochastic Parameterization Problem. *J. Clim.*, 20:187–202, 2007.

[126] A.P. Siebesma. Shallow cumulus convection. In E.J. Plate, E.E. Fedorovich, X.V. Viegas, and J.C. Wyngaard, editors, *Buoyant Convection in Geophysical Flows*, pages 441–486. Kluwer, 1998.

[127] A.P. Siebesma, C.S. Bretherton, A. Brown, A. Chlond, J. Cuxart, P.G. Duynkerke, H. Jiang, M. Khairoutdinov, D. Lewellen, C.-H. Moeng, E. Sanchez, B. Stevens, and D.E. Stevens. A Large-Eddy Simulation Intercomparison Study of Shallow Cumulus Convection. *J. Atmos. Sci.*, 60:1201–1219, 2003.

[128] A.P. Siebesma and J.W.M. Cuijpers. Evaluation of parametric assumptions for shallow cumulus convection. *J. Atmos. Sci.*, 52:650–666, 1995.

[129] A.P. Siebesma and A.A.M. Holtslag. Model Impacts of Entrainment and Detrainment Rates in Shallow Cumulus Convection. *J. Atmos. Sci.*, 53:2354–2364, 1996.

[130] A.P. Siebesma, P.M.M. Soares, and J. Teixeira. A Combined Eddy-Diffusivity Mass-Flux Approach for the Convective Boundary Layer. *J. Atmos. Sci.*, 64:1230–1248, 2007.

[131] P.K. Smolarkiewicz, L.G. Margolin, and A.A. Wyszogrodzki. A class of nonhydrostatic global models. *J. Atmos. Sci*, 58:349–364, 2001.

[132] P.M.M. Soares, P.M.A. Miranda, A.P. Siebesma, and J. Teixeira. An eddy-diffusivity/mass-flux parametrization for dry and shallow cumulus convection. *Q.J.R. Meteorol. Soc.*, 130:3365–3383, 2004.

[133] E.A. Spiegel and G. Veronis. On the Boussinesq approximation for a compressible fluid. *The Astrophysical Journal*, 131:442, 1960.

[134] S.N. Stechmann and J.D. Neelin. A stochastic model for the transition to strong convection. *J. Atmos. Sci.*, 68:2955–2970, 2011.

[135] R.B. Stull. *An Introduction to Boundary Layer Meteorology*. Springer Netherlands, 1988.

[136] J. Teixeira and C.A. Reynolds. Stochastic Nature of Physical Parameterizations in Ensemble Prediction: A Stochastic Convection Approach. *Mon. Wea. Rev.*, 136:483–496, 2008.

[137] M. Tiedtke. A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, 117:1779–1800, 1989.

[138] A.M. Tompkins. Organization of tropical convection in low vertical wind shears: the role of cold pools. *J. Atmos. Sci.*, 58:1650–1672, 2001.

[139] W.H. Tsai. Moment-Preserving Thresholding: A New Approach. *Comput. Vis. and Image Process.*, 29:377–393, 1985.

[140] R. van Driel and H.J.J. Jonker. Convective Boundary Layers Driven by Nonstationary Surface Heat Fluxes. *J. Atmos. Sci.*, 68:727–738, 2011.

[141] R. Vautard, K.C. Mo, and M. Ghil. Statistical significance test for transition matrices of atmospheric markov chains. *J. Atmos. Sci.*, 47(15):1926–1931, 1990.

[142] N. Verheul and D. Crommelin. Data-driven stochastic representations of unresolved features in multiscale models. *Comm. Math. Sci, to appear*.

[143] M. Wheeler and G.N. Kiladis. Convectively Coupled Equatorial Waves: Analysis of Clouds and Temperature in the Wavenumber-Frequency Domain. *J. Atmos. Sci.*, 56:374–399, 1999.

[144] D.S. Wilks. Effects of stochastic parameterizations in the Lorenz '96 system. *Q.J.R. Meteorol. Soc.*, 131:389–407, 2005.

[145] D. Williams. *Probability with Martingales*. Cambridge university press, 1991.

[146] C.-M. Wu, B. Stevens, and A. Arakawa. What controls the transition from shallow to deep convection? *J. Atmos. Sci.*, 66:1793–1806, 2009.

[147] J.C. Wyngaard. Toward Numerical Modeling in the "Terra Incognita". *J. Atmos. Sci.*, 61:1816–1826, 2004.

[148] J.C. Wyngaard and C-.H. Moeng. Parameterizing turbulent diffusion through the joint probability density. *Boundary-Layer Met.*, 60:1–13, 1992.

[149] X. Yu and T.-Y. Lee. Role of convective parameterization in simulations of a convection band at grey-zone resolutions. *Tellus A*, 62:617–632, 2010.

[150] C. Zhang. Madden-Julian Oscillation. *Rev. Geophys.*, 43:RG2003, 2005.

Curriculum Vitae

Jesse Dorrestijn was born on 17 April 1982 in 's-Hertogenbosch in the Netherlands. He went to school in Ede and Hilversum. Then, in 2000, he started his studies in mathematics at Utrecht University. In 2004 he moved to Rome, for one year, also to study mathematics at La Sapienza. In 2005 he returned to the Netherlands to do a master program in mathematics. His thesis was written on Benford's Law in probability theory and he received his master's degree from Utrecht University with distinction in 2008. Between 2008 and 2010 he worked in Amersfoort as a tutor for secondary school children and gave trainings for the students' final exams.

In 2010, he started his PhD research on stochastic convection parameterization at the National Research Institute for Mathematics and Computer Science (CWI) in Amsterdam, working together with scientists at KNMI and Delft University of Technology. He attended several national and international workshops and conferences and gave talks at for instance the Max Planck Institute for Meteorology in Hamburg and the World Weather Open Science Conference in Montreal. He also presented a poster at the AGU in San Francisco and he won the poster prize at the Dutch Mathematical Congress, organized at the University of Twente, Enschede.

List of Publications

4. **J. Dorrestijn and D.T. Crommelin and A.P. Siebesma and H.J.J. Jonker and F. Selten**, *Stochastic Convection Parameterization with Markov Chains in an Intermediate-Complexity GCM*, J. Atmos. Sci., **73**, 1367–1382 (2016).

3. **J. Dorrestijn and D.T. Crommelin and A.P. Siebesma and H.J.J. Jonker and C. Jakob**, *Stochastic parameterization of convective area fractions with a multicloud model inferred from observational data*, J. Atmos. Sci. **72**, 854–869 (2015).

2. **J. Dorrestijn and D.T. Crommelin and J.A. Biello and S.J. Böing**, *A data-driven multi-cloud model for stochastic parametrization of deep convection*, Phil. Trans. R. Soc. A. **317**, 20120374 (2013).

1. **J. Dorrestijn and D.T. Crommelin and A.P. Siebesma and H.J.J. Jonker**, *Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data*, Theor. Comput. Fluid Dyn. **27**, 133–148 (2013).

Acknowledgment

I would like to express my sincere gratitude to my daily supervisor Prof. dr. Daan Crommelin (leader of the Scientific Computing research group at CWI and affiliated to the University of Amsterdam (UvA)) and weekly supervisors Prof. dr. Pier Siebesma (KNMI and affiliated to Delft University of Technology) and Prof. dr. Harm Jonker (Delft University of Technology). I have been able to learn a lot and reach goals thanks to their great efforts and help. I am also grateful to Prof. dr. Joseph Biello (Department of Mathematics, University of California, Davis, CA, USA) visiting CWI, dr. Steef Böing, former PhD student at Delft University of Technology, Prof. dr. Christian Jakob (ARC Centre of Excellence for Climate System Science, Monash University, Melbourne, Australia) and dr. Frank Selten (KNMI). I would like to thank dr. Keith Myerscough, former PhD student at CWI, and drs. Alkor Zelle for carefully reading and correcting the manuscript.

Finally, I thank my friends and family for a lot of support.

Jesse Dorrestijn
Amsterdam, March 2016