Stochastic Parameterization of Convective Area Fractions with a Multicloud Model Inferred from Observational Data

JESSE DORRESTIJN AND DAAN T. CROMMELIN

CWI, Amsterdam, Netherlands

A. PIER SIEBESMA

Royal Netherlands Meteorological Institute (KNMI), De Bilt, and Delft University of Technology, Delft, Netherlands

HARMEN J. J. JONKER

Delft University of Technology, Delft, Netherlands

CHRISTIAN JAKOB

ARC Centre of Excellence for Climate System Science, Monash University, Melbourne, Victoria, Australia

(Manuscript received 11 April 2014, in final form 11 September 2014)

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.

1. 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 cannot 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 Arakawa (2004) for an overview of the cumulus parameterization problem. Formulating moist convection parameterizations is a difficult problem: it introduces uncertainties in model predictions (e.g., Randall et al. 2003) 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., Flato et al. 2014). It has been shown by Lin et al. (2006) 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 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

Corresponding author address: Jesse Dorrestijn, CWI, P.O. Box 94079, 1090 GB, Amsterdam, Netherlands.
E-mail: j.dorrestijn@cwi.nl

DOI: 10.1175/JAS-D-14-0110.1

the choice between dynamical (e.g., vertical velocity) and thermodynamical [e.g., the convective available potential energy (CAPE), relative humidity (RH)] variables, which have been studied in a recent paper by Davies et al. (2013a). Another important issue is that if parameterizations are chosen to be deterministic functions of the resolved variables, then the subgrid response of moist convection to large-scale variations cannot 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 GCM resolutions are getting finer and finer, this issue becomes more important, because with smaller grid boxes the fluctuations around expected subgrid responses become larger. Palmer (2001) 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 been shown, for example, by Buizza et al. (1999), 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., Lin and Neelin 2000; Lin and Neelin 2003; Teixeira and Reynolds 2008; Plant and Craig 2008; Bengtsson et al. 2013) or in the shallow convection parameterization (e.g., Sakradzija et al. 2014).

Rather than relying on physical intuition or deriving parameterizations from first principles, stochastic parameterizations can be inferred directly from data. Crommelin and Vanden-Eijnden (2008) 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 (Lorenz 1996) model quite well, better than the deterministic parameterizations and the stochastic parameterizations, based on autoregressive processes, of Wilks (2005). The data-driven Markov chain model inspired Kwasniok (2012) to develop a similar model based on cluster-weighted Markov chains. In Dorrestijn et al. (2013b) the Markov chain model of Crommelin and Vanden-Eijnden (2008) was used to study stochastic parameterization of shallow convection and in Dorrestijn et al. (2013a) it was used 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 Mapes et al. (2006) and Johnson et al. (1999), is formed by multicloud models (e.g., Khouider and Majda 2006; Khouider et al. 2010; Majda et al. 2007; Frenkel et al. 2013; Peters et al. 2013). 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 coming full circle to clear sky. In the multicloud model of Dorrestijn et al. (2013a), shallow cumulus clouds are also included.

In the present paper we use high-resolution ($\sim 2.5 \times 2.5\,\mathrm{km}^2$) observational data of rainfall in combination with data defining the large-scale ($\sim 150 \times 150\,\mathrm{km}^2$) 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 Dorrestijn et al. (2013a) 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 that 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 (Gottwald et al. 2014, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*). Their stochastic models inferred from large-scale observational data also yield convective area fractions.

Our paper is organized as follows. In section 2 we explain how we use Markov chains as a foundation for our multicloud model. Then, in section 3, we give a description of the observational data and 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 we assess the skill of large-scale variables as indicators for deep convection. In section 5 we construct our model, give expected area fractions and standard deviations, and discuss scale adaptivity (i.e., the ability to adapt to the size of a GCM grid box). We give results in section 6 by comparing area fractions from the model with the observations and looking at their autocorrelation functions. In section 7 we discuss the possibilities of implementation of the stochastic model in a convection parameterization of a GCM and make some concluding remarks.

2. Markov chains

The multicloud model that we use in this study consists of Markov chains positioned on the nodes of a two-dimensional microgrid. This model setup has been

used before in Khouider et al. (2010), Dorrestijn et al. (2013a), and Peters et al. (2013). The state of each Markov chain at time t is denoted $Y_n(t)$, where n is the microgrid 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 3. We will refer to these categories as cloud types. As time evolves, the Markov chains can switch, or "make a transition," between states every $\Delta t = 10$ min. All the Markov chains on the microgrid together determine the area fractions σ_m for the various cloud types:

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

in which $\mathbf{1}$ is the indicator function ($\mathbf{1}[A] = 1$ if A is true, 0 otherwise), N is the number of microgrid nodes, and $m \in \{1, \ldots, 5\}$ is 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 10 min, 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:

$$M_b = \rho \sigma_4 w_{\mathrm{cb}}, \tag{2}$$

in which ρ is the density and w_{cb} is the vertical velocity in a deep convective updraft at cloud base (e.g., Arakawa et al. 2011; Möbis and Stevens 2012). More examples of possible applications in GCMs are given in section 7.

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 we discuss appropriate indicators. The framework of conditional Markov chains (CMCs) for parameterization was introduced by Crommelin and Vanden-Eijnden (2008).

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 Crommelin and Vanden-Eijnden (2008)]:

$$\mathrm{Prob}[Y_n(t+\Delta t) = \beta \mid Y_n(t) = \alpha, X(t) = \gamma] = \frac{T_\gamma(\alpha,\beta)}{\sum_\beta T_\gamma(\alpha,\beta)}, \tag{3}$$

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 dataset, 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 datasets corresponding to two different scales: datasets that are formed by high-resolution observations of rainfall at a scale that is equal to or smaller than the microgrid scale of the CMCs and datasets that represent the large-scale atmospheric state at the grid scale of a GCM. In the next section we introduce the high-resolution observation datasets.

3. 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. These data are described in detail in Kumar et al. (2013). 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–18 April 2007, we have integer-valued CTH and rain-rate observations at 10-min time steps for a circular area with a radius of 150 km and a resolution of $2.5 \times 2.5\,\mathrm{km}^2$. In Fig. 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 silence" and is due to the 42° maximum elevation angle (May and Ballinger 2007). Therefore, we only use pixels between 25 and 97.5 km from the center of the domain. This forms an annular-shaped subdomain consisting of 4720 pixels of $2.5 \times 2.5\,\mathrm{km}^2$ corresponding to an area size of approximately $172 \times 172\,\mathrm{km}^2$. Figure 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

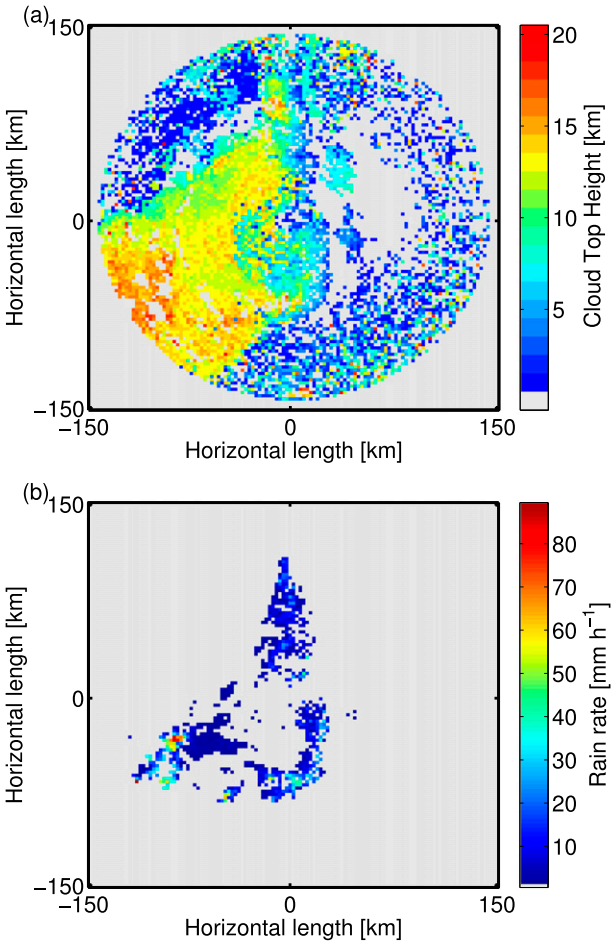

FIG. 1. (a) A snapshot of the cloud-top height derived from Darwin radar observations and (b) the corresponding rain rate.

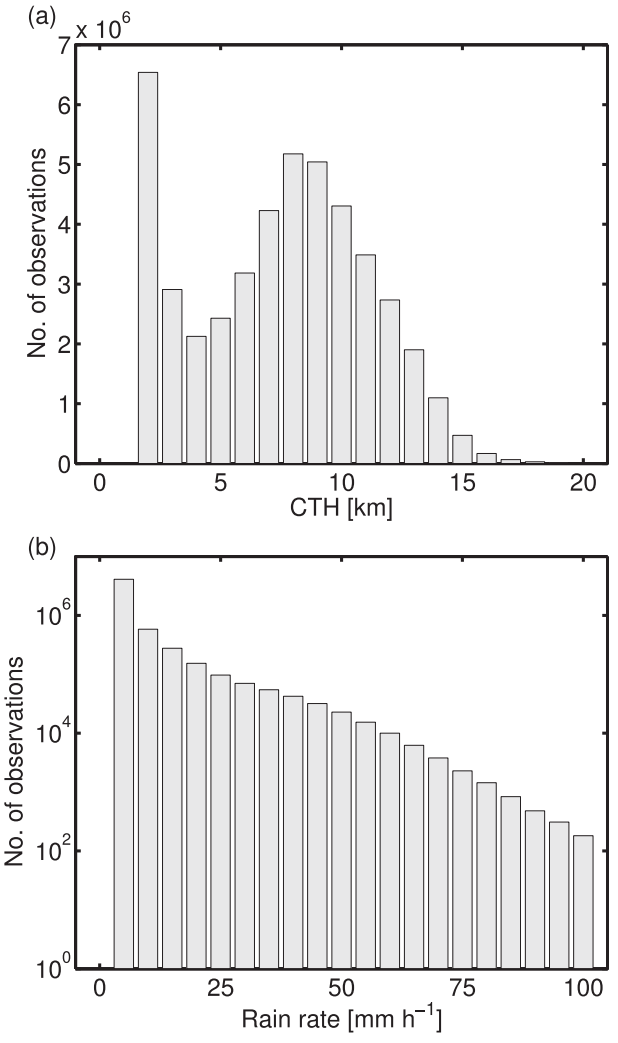

FIG. 2. Histograms of (a) the cloud-top height and (b) the rain rate observed with the Darwin radar in the periods November 2005–April 2006 and January–April 2007.

to avoid the influence of radar ground clutter. There is a bimodal 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 bimodal 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 (Johnson et al. 1999). We adopt the approach of Kumar et al. (2013), 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. 2b shows an approximately exponential distribution, so it is impossible to 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 that are described in Lang et al. (2003). 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 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 nonprecipitating clouds.

After classification, we have two-dimensional fields with discrete values (integers from 1 to 5). In Fig. 3 we

TABLE 1. Cloud-type classification using thresholds for the cloud-top height and the rain rate.

CTH (km)	Rain rate (mm h^{-1})	Classification
≥6.5	≤12	Stratiform ($m = 5$)
	>12	Deep convective ($m = 4$)
∈[1.5, 6.5)	>3	Strong congestus ($m = 3$)
	≤3	Moderate congestus ($m = 2$)
<1.5	0	Clear ($m = 1$)

give an example of a classified field, which is the classified field corresponding to the CTH and rain-rate fields shown in Fig. 1. After the classification, the observed σ_m values can be calculated according to (1), with Y_n the observed cloud type and $N = 4720$ 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%. The observed fractions are discussed in section 6 (depicted in Fig. 10) for a time period of 5 days for all cloud types and also the deep convective area fraction (Fig. 7a) for a longer period of 3 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 (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:

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

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 10 min later is 0.3540. The probability that a deep site is again a deep site 10 min later is 0.4295—much larger than the expected deep convective area fraction (at most 0.03 as can be seen Fig. 6, discussed later in this paper). This is comparable to the deep-to-deep transition probability of 0.5602 estimated from the LES dataset of Dorrestijn et al. (2013a). 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

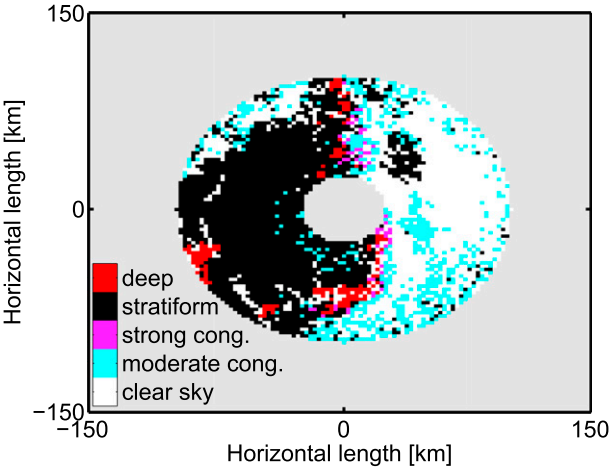

FIG. 3. Example of radar data assigned to the categories clear sky, moderate congestus, strong congestus, deep convective, and stratiform, corresponding to the CTH and rain-rate snapshots of Fig. 1.

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 artifacts 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 $N_{ij} = 4720$ 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., Press et al. 1992), 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 outside the domain at the previous time step $t - \Delta t$ or at the next time step $t + \Delta t$ owing 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 10 min 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) and when we test the CMC (section 7). 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. The large-scale data

We have data available that define 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 6-h intervals and have a vertical resolution of 40 pressure levels, from ground level to about 20-km altitude. The data have been prepared by Davies et al. (2013a), 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 are also used in Davies et al. (2013b), Peters et al. (2013), and Gottwald et al. (2014, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*). 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 Davies et al. (2013a). Then, we will use the large-scale data accordingly for conditioning the multicloud CMC model. As in Davies et al. (2013a), 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} is the level of neutral buoyancy, and p_{LFC} is the level of free convection (e.g., Siebesma 1998). We define the mean vertical velocity 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 (hPa h^{-1}), p_0 is the pressure at the surface, and p^* is the pressure level 340 hPa, chosen because the resulting $\langle\omega\rangle$ gives the highest correlation with deep convective area fractions [as calculated with (4), which 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 time series of the deep convective area fraction $\sigma_4(t)$ and the time series of the indicator $X(t)$, the normalized CCF of $X(t)$ and $\sigma_4(t)$ is

$$\text{CCF}(\tau) = \int_{-\infty}^{\infty} \tilde{X}(t+\tau)\tilde{\sigma}_4(t)\, dt \qquad (4)$$

with $\tilde{X}(t) = [X(t) - \mu_X]/\sigma_X$ (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 with respect to σ_4. As such, the CCF lies in between -1 and 1. If the maximum value of the CCF is attained at positive τ, $X(t)$ tends to follow rather than precede deep convection.

In Fig. 4 we plot the CCFs of the indicators $-\langle\omega\rangle$, CAPE, and RH with the observed deep convective area fraction for the 2005/06 period. The figure for the 2007 period is similar (not included). Before calculating the CCF, we linearly interpolate X to get its values every 10 min instead of every 6 h, 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

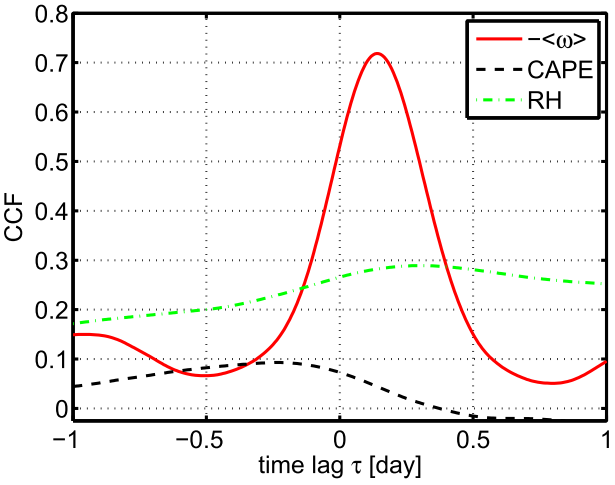

FIG. 4. Cross-correlation functions of the deep convective area fraction with $-\langle\omega\rangle$, CAPE, and RH at 640 hPa for the 2005/06 dataset.

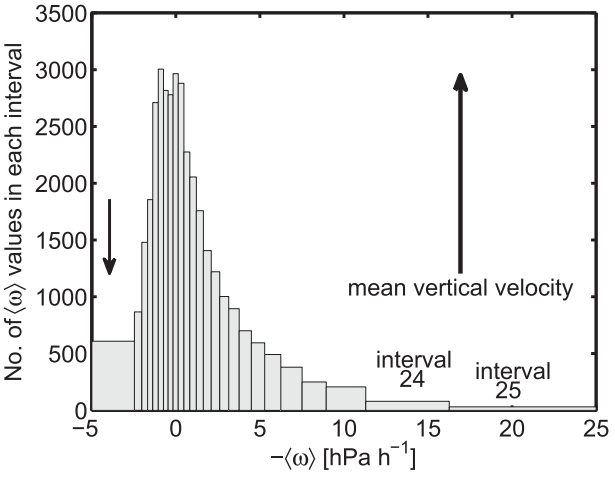

FIG. 5. Histogram of the 25 intervals of $-\langle\omega\rangle$, found by clustering the linearly interpolated $\langle\omega\rangle$ values. The first and last (twenty-fifth) intervals are open on one side. Because ω is a velocity in terms of pressure, positive $\langle\omega\rangle$ corresponds to downward mean large-scale motion and negative $\langle\omega\rangle$ corresponds to upward mean motion (as illustrated by the arrows).

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 Peters et al. (2013) 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).

To use an indicator for constructing the CMC according to (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 Dorrestijn et al. (2013b,a) and Kwasniok (2012). To give an example, in Fig. 5 we show a histogram of $\langle\omega\rangle$ discretized using 25 intervals. These intervals have been found by using the *k*-means cluster method (MacQueen 1967; Gan et al. 2007), 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 nonequidistant 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. 6 that the averaged observed deep convective and stratiform area fractions are large (around 3% and 90%, respectively) for interval number 25.

5. 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 (3). We take the period from 10 November 2005 to 15 April 2006 as the training dataset, 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 dataset is very large, containing $O(10^8)$ observations of transitions (radar images at 10-min intervals during 157 days, with 4720 pixels in each image).

In section 6 we will validate the CMCs with the test dataset, 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 dataset.

In Fig. 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. 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 that we corrected for horizontal advection (as described before in section 3).

We see in Fig. 6a that the expected deep convective area fractions increase with increasing $\langle\omega\rangle$ interval (corresponding to increasing upward mean vertical velocities) and the area has its maximum of around 0.03 for interval number 24. Further, the strong congestus fractions in Fig. 6b increase with increasing $\langle\omega\rangle$ interval; however, for

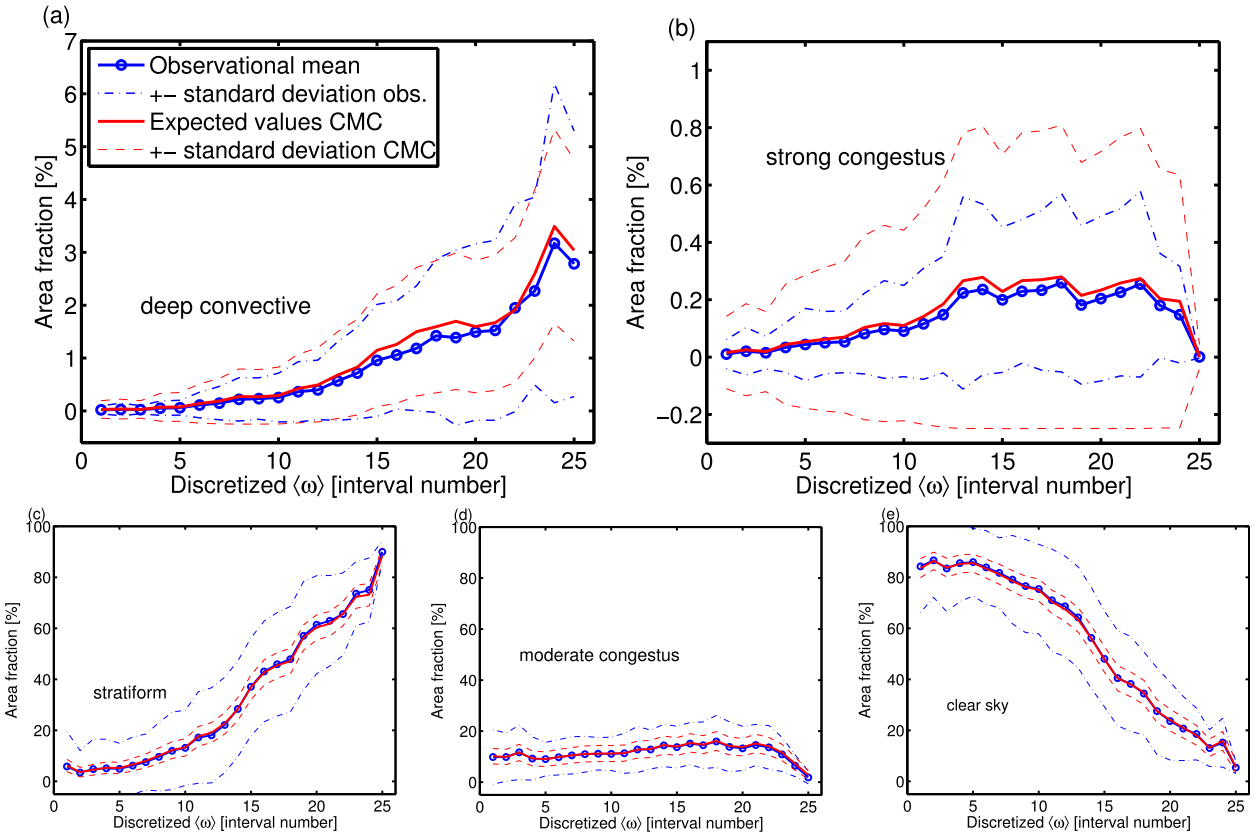

FIG. 6. Observational mean cloud-type area fractions as a function of the $\langle\omega\rangle$ intervals for the 2005/06 training period (solid line with circles) plus and minus the standard deviation (dashed–dotted line) and the CMC expected cloud-type area fractions (solid line) plus and minus the standard deviation while using $N = 100$ CMCs (dashed line). Note the different scalings on the y axes.

interval numbers 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 mean motion and 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 these $\langle\omega\rangle$ 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%, independent 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

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

in which E is the expectation. One can derive that this is equal to $\sqrt{N^{-1}p(1-p)}$, in which $p = E[\sigma_m]$ is the expected value of the fraction. Note that $E[\sigma_m]$ 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 dataset. 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 Davies et al. (2013a) 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 Arakawa et al. (2011). 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 gridbox 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 ($2.5 \times 2.5\,\text{km}^2$), 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 6, $N = 100$ versus $N = 4720$. We emphasize that the value of $N = 100$ is found during the training phase and not during the testing phase of the model.

6. Results

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

a. 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 5 using the training dataset 2005/06. We assess the model by driving the CMCs with $\langle\omega\rangle$ as observed in the other dataset (from 2007). Thus, different datasets are used for training and evaluation.

In Fig. 7a we show the deep convective area fractions as observed in the Darwin radar test dataset (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 widespread convective events, and the deep convective area fraction is exactly zero in 52% of the 10-min intervals. In Figs. 7b and 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. 7d). Since the CMC fractions are integer multiples of 0.01, we bin the Darwin observed fractions into intervals 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. 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. 8a we show the resulting CMC deep convective area fractions (cf. Fig. 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. 8b), it is indeed visible that fractions larger than 0.04 are too rare, although a fraction of 8% is reached in 1 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 $N = 69^2 = 4761$, which is (close to) the number of

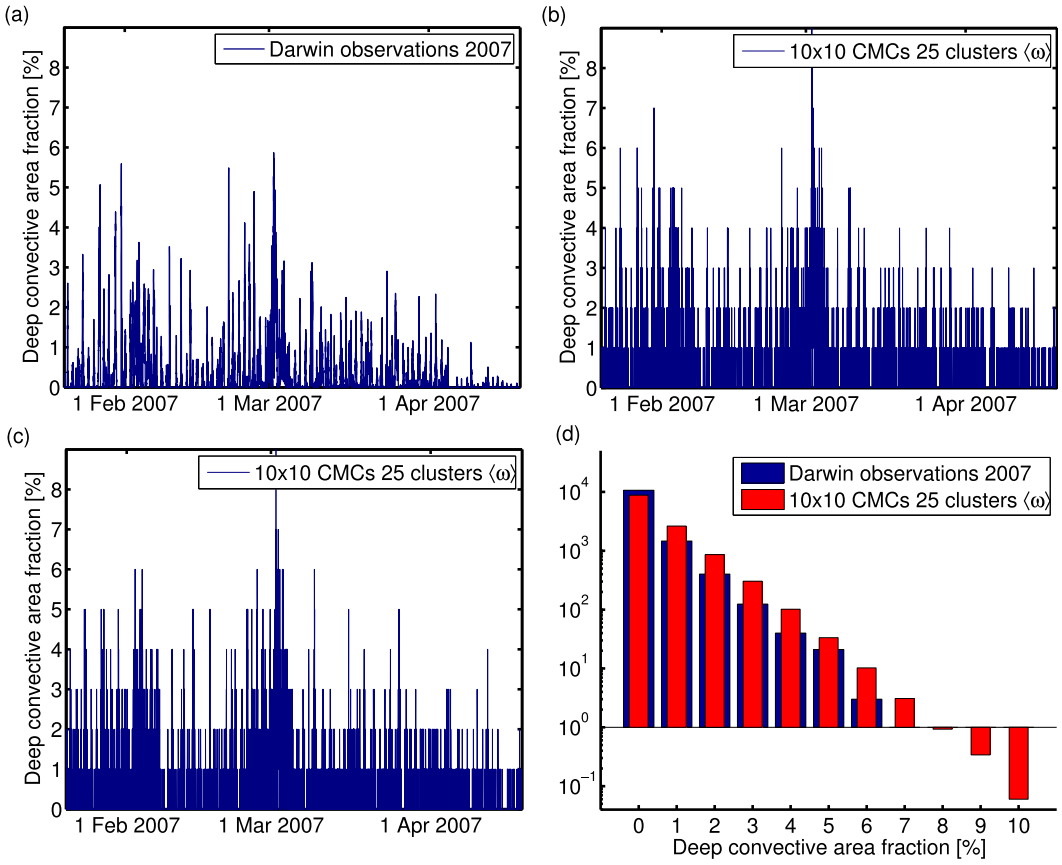

FIG. 7. (a) Deep convective area fractions observed in Darwin, (b),(c) two realizations of deep convective area fractions produced by $N = 100$ CMCs conditioned on $\langle\omega\rangle$, and (d) the corresponding histograms comparing the CMC fractions (averaged over 100 realizations) with the observed fractions (binned into intervals) on a logarithmic y axis.

radar pixels used to train the CMCs. We observe (Fig. 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 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 setup the CMCs on the 2D microlattice 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 Dorrestijn et al. (2013a) and Khouider (2014).

As a final experiment we take again $N = 100$ CMCs and $\langle\omega\rangle$ as indicator, but we interchange the roles of training dataset and test dataset. Thus, we train the CMCs with the 2007 dataset and validate using fractions for the 2005/06 period. The deep convective area fractions in the 2005/06 radar data reach higher maxima than in the 2007 dataset, with an overall maximum of about 10% (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. 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 5 days. The timing of the deep convective events produced by the CMCs is almost correct, although there is a small time lag visible in Fig. 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

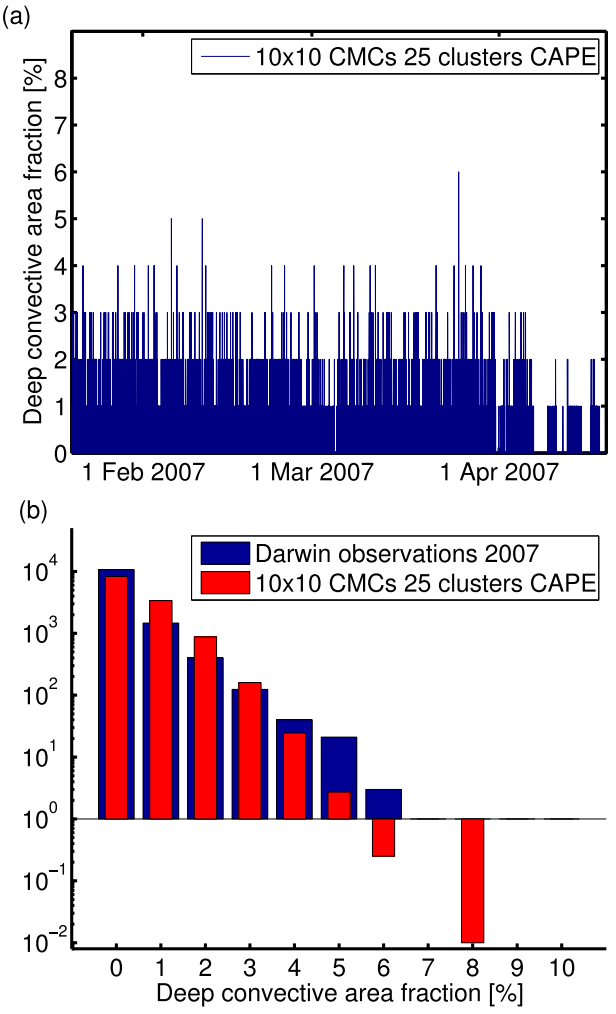

FIG. 8. Deep convective area fractions produced by $N = 100$ CMCs conditioned on CAPE and (b) the corresponding histograms in which the CMC fractions (averaged over 100 realizations) are compared to the observed fractions (binned into intervals) on a logarithmic y axis.

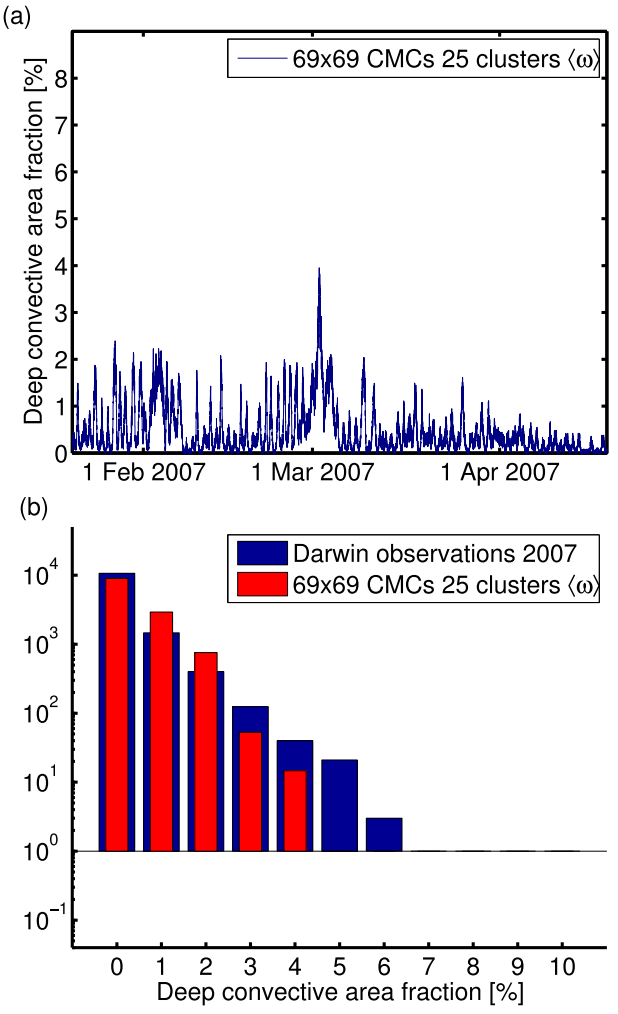

FIG. 9. Deep convective area fractions produced by $N = 69^2$ CMCs conditioned on $\langle\omega\rangle$ and (b) the corresponding histograms of the binned CMC fractions averaged over 100 realizations compared to the binned observed fractions on a logarithmic y axis.

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. 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. 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. 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. 10e) of the CMC follow the observations quite well, but the minimum values 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.

b. *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 σ_m is

$$\mathrm{ACF}(\tau) = \int_{-\infty}^{\infty} \tilde{\sigma}_m(t+\tau)\tilde{\sigma}_m(t)\,dt, \qquad (5)$$

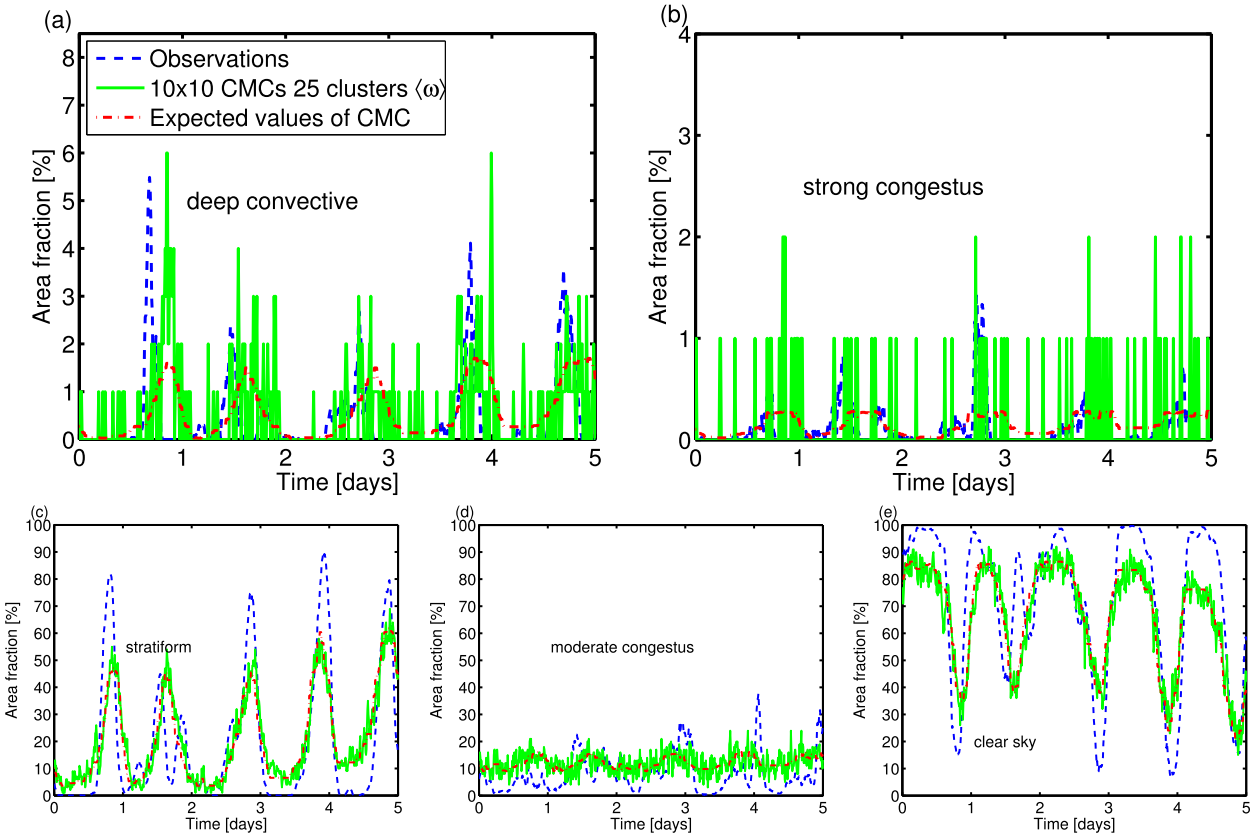

FIG. 10. Area fractions of (a) deep convective, (b) strong congestus, (c) stratiform, (d) moderate congestus, and (e) clear sky observed in Darwin (dashed line), produced by 100 CMCs (solid line) conditioned on $\langle\omega\rangle$, and the corresponding expected area fractions of the CMCs (dashed–dotted line) for a period of 5 days. Note the different scalings on the y axes.

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. 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 deep convective (Fig. 11a), strong congestus (Fig. 11b), stratiform (Fig. 11c), moderate congestus (Fig. 11d), and clear sky (Fig. 11e). Also the ACF of $\langle\omega\rangle$ is shown (dashed–dotted line). In Fig. 11a 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 3 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 dataset of 2005/06 is similar to the ACF for the 2007 dataset (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 Fig. 11b, 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 Figs. 11c and 11e we see ACFs of the CMC, which are comparable to the observational ACF only if conditioned on $\langle\omega\rangle$ and 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

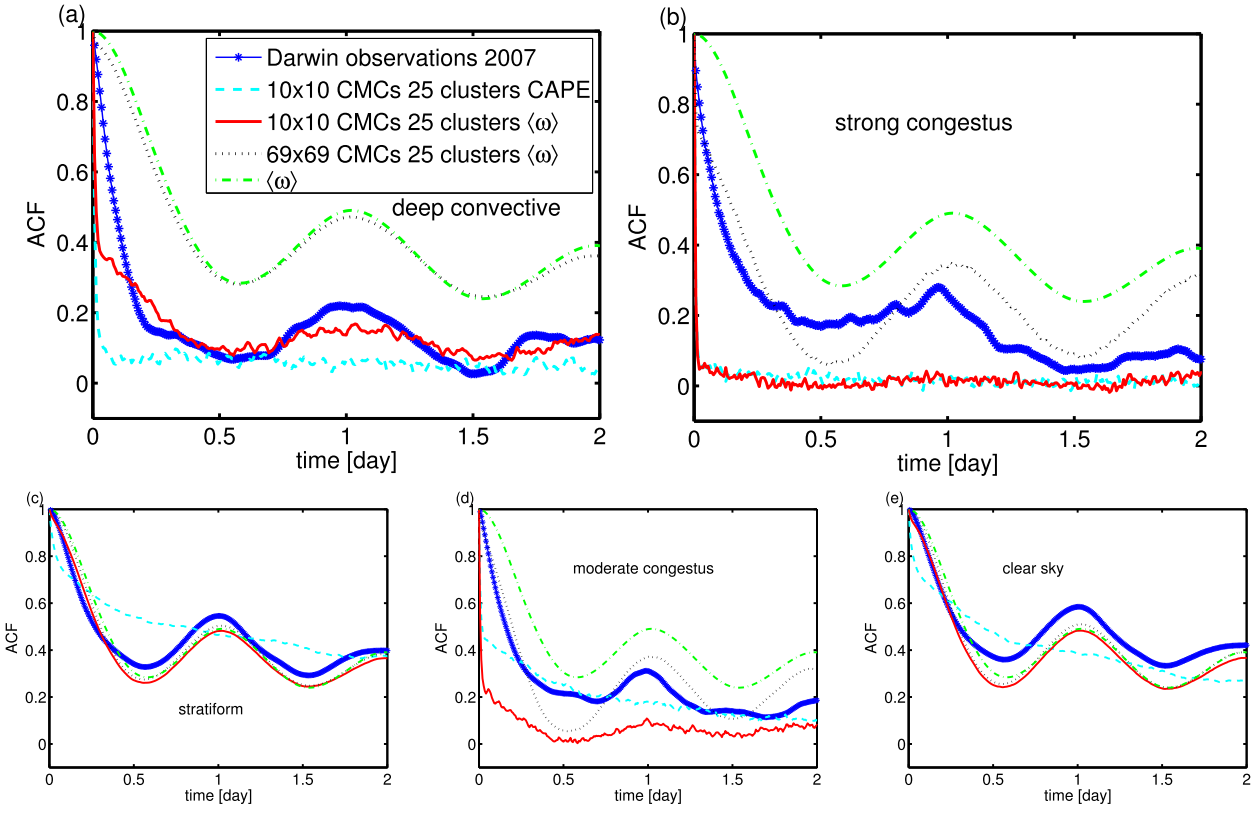

FIG. 11. Normalized ACFs of the observational area fractions (solid lines with stars), the CMC area fractions with $N = 100$ conditioned on $\langle\omega\rangle$ (solid lines) and on CAPE (dashed lines), and the ACF corresponding to 69^2 CMCs conditioned on $\langle\omega\rangle$ (dotted lines) for 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 (dashed–dotted lines).

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 nonsmooth ACFs.

7. Discussion and conclusions

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), which is a model setup similar to Khouider et al. (2010) and Dorrestijn et al. (2013a). 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 primarily focused). The vertically averaged large-scale vertical velocity $\langle\omega\rangle$ was found to be a good indicator, whereas CAPE and RH are found to be less suitable indicators. This is in agreement with the findings of Davies et al. (2013a).

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 nonuniform grids. Further, the model can be used as a start for GCMs reaching grid sizes that fall in the gray zone—that is, for grid sizes so small that subgrid convective flux terms are of the same order as the resolved flux terms (e.g., Yu and Lee 2010; Dorrestijn et al. 2013b).

In the gray zone, besides the problem that the fluxes are partly resolved and partly unresolved, the unresolved fluxes have a large standard deviation (Dorrestijn et al. 2013b). The stochastic multicloud model can produce stochastic fluctuations, resulting in a large standard deviation for the

unresolved fluxes, which 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, then 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 (Dorrestijn et al. 2013a).

The fractions produced by the multicloud model depend on the thresholds of Table 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 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, then 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, 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 0 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 (2) in section 2. 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 feedback 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 no 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 represent 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., Khouider et al. 2010) or explicitly inferring vertical heat and moisture tendency profiles from data as in Dorrestijn et al. (2013b). 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 (Dorrestijn et al. 2013a). 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 5-month period in Darwin, the multicloud model was able to adequately produce fractions for a different 3-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 5 months is long enough to train the model for Darwin. In the experiment where we interchanged training and test dataset, we found that even training on a 3-month period is enough to produce adequate fractions for the 5-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 dataset of the study of Dorrestijn et al. (2013a) was 6 h as opposed to the ±8-month period of the radar data. A simulation of 8 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. Gottwald et al. (2014, manuscript submitted to *Quart. J. Roy. Meteor. Soc.*) 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 the idea that our multicloud model could be used more globally also. 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 are used to estimate the model, and the CMC cloud-type area fractions were shown to be comparable to the observations, which is notable because we used different datasets 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 are used to produce area fractions. Because of the conditioning, memory effects that are built 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.

Acknowledgments. 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).

REFERENCES

Arakawa, A., 2004: The cumulus parameterization problem: Past, present, and future. *J. Climate,* **17,** 2493–2525, doi:10.1175/1520-0442(2004)017<2493:RATCPP>2.0.CO;2.

——, J.-H. Jung, and C.-M. Wu, 2011: Toward unification of the multiscale modeling of the atmosphere. *Atmos. Chem. Phys.,* **11,** 3731–3742, doi:10.5194/acp-11-3731-2011.

Bengtsson, L., M. Steinheimer, P. Bechtold, and J.-F. Geleyn, 2013: A stochastic parametrization for deep convection using cellular automata. *Quart. J. Roy. Meteor. Soc.,* **139,** 1533–1543, doi:10.1002/qj.2108.

Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.,* **125,** 2887–2908, doi:10.1002/qj.49712556006.

Crommelin, D., and E. Vanden-Eijnden, 2008: Subgrid-scale parameterization with conditional Markov chains. *J. Atmos. Sci.,* **65,** 2661–2675, doi:10.1175/2008JAS2566.1.

Davies, L., C. Jakob, P. May, V. Kumar, and S. Xie, 2013a: Relationships between the large-scale atmosphere and the small-scale convective state for Darwin, Australia. *J. Geophys. Res. Atmos.,* **118,** 11 534–11 545, doi:10.1002/jgrd.50645.

——, and Coauthors, 2013b: A single-column model ensemble approach applied to the TWP-ICE experiment. *J. Geophys. Res. Atmos.,* **118,** 6544–6563, doi:10.1002/jgrd.50450.

Dorrestijn, J., D. Crommelin, J. Biello, and S. Böing, 2013a: A data-driven multi-cloud model for stochastic parametrization of deep convection. *Philos. Trans. Roy. Soc. London,* **A371**, doi:10.1098/rsta.2012.0374.

——, ——, A. Siebesma, and H. Jonker, 2013b: Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theor. Comput. Fluid Dyn.,* **27,** 133–148, doi:10.1007/s00162-012-0281-y.

Flato, G., and Coauthors, 2014: Evaluation of climate models. *Climate Change 2013: The Physical Science Basis,* T. Stocker et al., Eds., Cambridge University Press, 741–866.

Frenkel, Y., A. Majda, and B. Khouider, 2013: Stochastic and deterministic multicloud parameterizations for tropical convection. *Climate Dyn.,* **41,** 1527–1551, doi:10.1007/s00382-013-1678-z.

Gan, G., C. Ma, and J. Wu, 2007: *Data Clustering: Theory, Algorithms, and Applications.* 3rd ed. SA-SIAM Series on Statistics and Applied Probability, SIAM, 466 pp.

Johnson, R., T. Rickenbach, S. Rutledge, P. Ciesielski, and W. Schubert, 1999: Trimodal characteristics of tropical convection. *J. Climate,* **12,** 2397–2418, doi:10.1175/1520-0442(1999)012<2397:TCOTC>2.0.CO;2.

Khouider, B., 2014: A coarse grained stochastic multi-type particle interacting model for tropical convection: Nearest neighbour interactions. *Commun. Math. Sci.,* **12,** 1379–1407, doi:10.4310/CMS.2014.v12.n8.a1.

——, and A. Majda, 2006: A simple multicloud parameterization for convectively coupled tropical waves. Part I: Linear analysis. *J. Atmos. Sci.,* **63,** 1308–1323, doi:10.1175/JAS3677.1.

——, J. Biello, and A. J. Majda, 2010: A stochastic multicloud model for tropical convection. *Commun. Math. Sci.,* **8,** 187–216, doi:10.4310/CMS.2010.v8.n1.a10.

Kumar, V., C. Jakob, A. Protat, P. T. May, and L. Davies, 2013: The four cumulus cloud modes and their progression during rainfall events: A C-band polarimetric radar perspective. *J. Geophys. Res. Atmos.,* **118,** 8375–8389, doi:10.1002/jgrd.50640.

Kwasniok, F., 2012: Data-based stochastic subgrid-scale parametrization: An approach using cluster-weighted modelling. *Philos. Trans. Roy. Soc. London,* **A370,** 1061–1086, doi:10.1098/rsta.2011.0384.

Lang, S., W.-K. Tao, J. Simpson, and B. Ferrier, 2003: Modeling of convective–stratiform precipitation processes: Sensitivity to partitioning methods. *J. Appl. Meteor.,* **42,** 505–527, doi:10.1175/1520-0450(2003)042<0505:MOCSPP>2.0.CO;2.

Lin, J.-L., and Coauthors, 2006: Tropical intraseasonal variability in 14 IPCC AR4 climate models. Part I: Convective signals. *J. Climate,* **19,** 2665–2690, doi:10.1175/JCLI3735.1.

Lin, J. W.-B., and J. D. Neelin, 2000: Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.,* **27,** 3691–3694, doi:10.1029/2000GL011964.

——, and ——, 2003: Toward stochastic deep convective parameterization in general circulation models. *Geophys. Res. Lett.,* **30,** 1162, doi:10.1029/2002GL016203.

Lorenz, E., 1996: Predictability—A problem partly solved. *Proc. Seminar on Predictability,* Reading, United Kingdom, ECMWF, 40–58.

MacQueen, J., 1967: Some methods for classification and analysis of multivariate observations. *Proc. Fifth Berkeley Symp. Mathematical Statistics and Probability,* Berkeley, CA, Vol. 1, Statistical Laboratory of the University of California, Berkeley, 281–297.

Majda, A. J., S. N. Stechmann, and B. Khouider, 2007: Madden–Julian oscillation analog and intraseasonal variability in a multicloud model above the equator. *Proc. Natl. Acad. Sci. USA,* **104,** 9919–9924, doi:10.1073/pnas.0703572104.

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

May, P., and A. Ballinger, 2007: The statistical characteristics of convective cells in a monsoon regime (Darwin, Northern Australia). *Mon. Wea. Rev.,* **135,** 82–92, doi:10.1175/MWR3273.1.

Möbis, B., and B. Stevens, 2012: Factors controlling the position of the Intertropical Convergence Zone on an aquaplanet. *J. Adv. Model. Earth Syst.,* **4,** M00A04, doi:10.1029/2012MS000199.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.,* **127,** 279–304, doi:10.1002/qj.49712757202.

Peters, K., C. Jakob, L. Davies, B. Khouider, and A. Majda, 2013: Stochastic behavior of tropical convection in observations and a multicloud model. *J. Atmos. Sci.,* **70,** 3556–3575, doi:10.1175/JAS-D-13-031.1.

Plant, R., and G. Craig, 2008: A stochastic parameterization for deep convection based on equilibrium statistics. *J. Atmos. Sci.,* **65,** 87–105, doi:10.1175/2007JAS2263.1.

Press, W. H., S. Teukolsky, W. Vetterling, and B. Flannery, 1992: *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 994 pp.

Randall, D., M. Khairoutdinov, A. Arakawa, and W. Grabowski, 2003: Breaking the cloud parameterization deadlock. *Bull. Amer. Meteor. Soc.,* **84,** 1547–1564, doi:10.1175/BAMS-84-11-1547.

Sakradzija, M., A. Seifert, and T. Heus, 2014: Fluctuations in a quasi-stationary shallow cumulus cloud ensemble. *Nonlin. Processes Geophys. Discuss.,* **1,** 1223–1282, doi:10.5194/npgd-1-1223-2014.

Siebesma, A., 1998: Shallow cumulus convection. *Buoyant Convection in Geophysical Flows,* E. Plate et al., Eds., Kluwer, 441–486.

Teixeira, J., and C. Reynolds, 2008: Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Mon. Wea. Rev.,* **136,** 483–496, doi:10.1175/2007MWR1870.1.

Wilks, D., 2005: Effects of stochastic parameterizations in the Lorenz '96 system. *Quart. J. Roy. Meteor. Soc.,* **131,** 389–407, doi:10.1256/qj.04.03.

Yu, X., and T.-Y. Lee, 2010: Role of convective parameterization in simulations of a convection band at grey-zone resolutions. *Tellus,* **62A,** 617–632, doi:10.1111/j.1600-0870.2010.00470.x.