Clim Dyn
DOI 10.1007/s00382-015-2936-z

Stochasticity of convection in Giga-LES data

Michèle De La Chevrotière[1,2] · Boualem Khouider[1] · Andrew J. Majda[3]

Received: 16 August 2015 / Accepted: 28 November 2015

Abstract The poor representation of tropical convection in general circulation models (GCMs) is believed to be responsible for much of the uncertainty in the predictions of weather and climate in the tropics. The stochastic multicloud model (SMCM) was recently developed by Khouider et al. (Commun Math Sci 8(1):187–216, 2010) to represent the missing variability in GCMs due to unresolved features of organized tropical convection. The SMCM is based on three cloud types (congestus, deep and stratiform), and transitions between these cloud types are formalized in terms of probability rules that are functions of the large-scale environment convective state and a set of seven arbitrary cloud timescale parameters. Here, a statistical inference method based on the Bayesian paradigm is applied to estimate these key cloud timescales from the Giga-LES dataset, a 24-h large-eddy simulation (LES) of deep tropical convection (Khairoutdinov et al. in J Adv Model Earth Syst 1(12), 2009) over a domain comparable to a GCM gridbox. A sequential learning strategy is used where the Giga-LES domain is partitioned into a few subdomains, and atmospheric time series obtained on each subdomain are used to train the Bayesian procedure incrementally. Convergence of the marginal posterior densities for all seven parameters is demonstrated for two different grid partitions, and sensitivity tests to other model parameters are also presented. A single column model simulation using the SMCM parameterization with the Giga-LES inferred parameters reproduces many important statistical features of the Giga-LES run, without any further tuning. In particular it exhibits intermittent dynamical behavior in both the stochastic cloud fractions and the large scale dynamics, with periods of dry phases followed by a coherent sequence of congestus, deep, and stratiform convection, varying on timescales of a few hours consistent with the Giga-LES time series. The chaotic variations of the cloud area fractions were captured fairly well both qualitatively and quantitatively demonstrating the stochastic nature of convection in the Giga-LES simulation.

Keywords Parameter estimation · Bayesian inference · Stochastic multicloud model · Tropical convection · General circulation models · Stochastic cumulus parameterization · Giga-LES · Markov Chain Monte Carlo · Large matrix exponential · Parallel and high performance computing

✉ Boualem Khouider
khouider@uvic.ca

Michèle De La Chevrotière
mdelachev@psu.edu

Andrew J. Majda
jonjon@cims.nyu.edu

[1] Department of Mathematics and Statistics, University of Victoria, PO BOX 3060 STN CSC, Victoria, BC, Canada

[2] Present Address: Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA

[3] Department of Mathematics and Center for Atmosphere and Ocean Sciences, Courant Institute of Mathematical Sciences, New York University, New York, NY, USA

1 Introduction

General circulation models (GCMs) generate climate predictions over a time period of seasons to years, and have a spatial resolution of $\mathcal{O}(100$ km). Clouds and convection processes, on the other hand, are organized in a hierarchy of temporal and spatial scales, ranging from individual convective cells (clouds) of one to 10 km and a few hours, to mesoscale cloud clusters of a few hundreds of kilometers

Published online: 21 December 2015

and 1–2 days, to super-clusters of a few thousands of kilometres and 5−10 days, to their planetary/intraseasonal scale envelopes such as the Madden–Julian oscillation (MJO, Madden and Julian 1972; Mapes et al. 2006; Moncrieff 2010). Because of their coarse resolution, GCMs include *parameterizations* that represent, based on some closure assumptions, the small scale effects of clouds and convection on the large-scale/resolved dynamics. Clouds and convection parameterization deficiencies have long been attributed to the failure of GCMs to adequately represent the variability associated with organized tropical convection (Hung et al. 2013; Lin et al. 2006; Moncrieff and Klinker 1997).

Purely deterministic closures, such as the Arakawa and Schubert (1974) quasi-equilibrium assumption, the moist convective adjustment idea of Manabe and Smagorinsky (1967), or the large-scale moisture convergence closure of Kuo (1974) type—and some more recent variants—, were found to inadequately represent the highly intermittent nature and variability of tropical convection. A new perspective for improving GCMs, in the last decade, came from the inclusion of stochastic parameterizations (see, for instance, the models of Lin and Neelin 2003; Buizza et al. 1999).

The stochastic multicloud model (SMCM) for tropical convection introduced by Khouider et al. (2010) (KBM10 below) is an approach based on a stochastic interacting particle lattice model where each lattice site is either occupied by a cloud of a certain type (congestus, deep, or stratiform), or is cloud free (clear sky). The transitions between the cloud types are governed by a set of probability rules that are constrained by the large-scale convective available potential energy (CAPE) and middle troposphere dryness, and are modulated by a set of seven arbitrary cloud transition timescales τ_{jk}. When local interactions between the lattice sites are ignored, a coarse-grained stochastic birth-death process is derived for the dynamical evolution of the area fractions σ_c, σ_d, and σ_s of congestus, deep, and stratiform clouds (KBM10). The coarse-graining procedure is extended to the more complex case with local interaction in Khouider (2014), under the assumption of uniform redistribution of particles within each coarse cell—i.e, a GCM grid box. The cloud area fractions in turn affect the large-scale dynamics by modulating the strength, timing and spacial distribution of the convective heating and precipitations rates.

The statistical equilibrium of σ_c, σ_d, and σ_s is critically linked to the choice of the cloud timescales τ_{jk}. In their introductory paper on the SMCM (KBM10), the authors conclude from case studies (see Table 5) that the dynamics of the stochastic lattice model and associated gridbox fractional cloudiness is very sensitive to the prescribed values of the cloud transition timescales. In another study using the SMCM, Frenkel et al. (2012) (FMK12 below) use yet a different set of parameter values (see Table 5) to study flows above the equator without rotation effects, and look at the impact of convective timescale dilation on the variability of convective coherent structures. Efforts towards a more systematic way of determining these parameters have recently been undertaken, most notably by Peters et al. (2013) (P2013 below) who visually constrained the equilibrium distribution of the multicloud area fractions to radar data (covering a 36,000 km^2 area over Darwin, Australia) to find best-fit transition timescales (see Table 5) that better represent the statistics of observed rainfall time series.

This work aims at using a rigorous statistical inference method, newly introduced by De La Chevrotière et al. (2014) (hereafter DKM14) to estimate the SMCM cloud timescale parameters from the Giga-LES dataset of Khairoutdinov et al. (2009), a large-eddy simulation (LES) of deep tropical convection over a domain comparable to a GCM gridbox. The inference method of DKM14 is basically a *static inverse problem*, in which an inference about the static model parameters τ_{jk} is obtained by assimilating time series of atmospheric data, based on the "exact" nonlinear forward SMCM (e.g. no imposed linearization). The inference model uses the Bayesian framework to formulate a posterior distribution over the model parameters, given the data. The challenge of the Bayesian method is that posterior exploration may be hard for computationally intensive forward models. In the multicloud problem, the calculation of the likelihood function requires solving a large system of differential equations (the Kolmogorov equations) as many times as there are data points, which is prohibitive both in terms of computation time and storage requirements. DKM14 uses the parallel Uniformization Method, which gives fast and scalable approximations of large sparse matrix exponentials for the solutions of systems of differential equations, without sacrificing accuracy. The high dimensional posterior distribution is sampled here using the standard Markov Chain Monte Carlo technique. As per design of the SMCM, the Bayesian inference procedure is trained using the large-scale CAPE and mid-troposphere dryness, and the subgrid-scale cloud coverage time series.

To increase information capacity, the full 205×205 km^2 Giga-LES domain is subdivided into grids, and training time series are obtained on each grid cell subdomain. The parameters are then progressively learned using a sequential learning technique tested and validated in DKM14.

The paper is organized as follows. The SMCM and the Bayesian learning procedure are presented in Sect. 2 while the Giga-LES dataset and the domain partitioning setup are discussed in Sect. 3. In Sect. 4, we process the Giga-LES

data to extract the cloud area fractions and large-scale indicator time series, which are needed as input for the Bayesian inference algorithm. The resulting distributions of the transition timescales and the convergence of the sequential learning strategy are presented in Sect. 5 together with sensitivity tests to two key parameters of the SMCM likelihood function, namely, the reference scales of CAPE and dryness. Section 5 also contains simulation results using a single column GCM based on the SMCM with the Giga-LES inferred transition timescales, which reproduces the main statistical features of the Giga-LES time series including the high intermittency and chaotic behavior of the cloud area fractions. Finally, a concluding discussion is presented in Sect. 6.

2 The model: SMCM and the Bayesian inference procedure

In this section, we describe the procedure of statistical inference for the SMCM based on the Bayesian paradigm following DKM14. As introduced in KBM10, the SMCM is essentially a multi-dimensional Markov birth-death process with immigration, which characterizes the populations or area fractions of different cloud types, evolving on the GCM subgrid scale. Transitions between the different cloud types occur at rates that are functions of the large-scale environment, modulated by a set of seven cloud transition parameters that allow for timescale adjustments (KBM10). A rigorous statistical method to estimate these cloud timescales from data was recently introduced and successfully validated using a synthetic experiment in DKM14. Here we provide a general overview of the method in Sect. 2.2, and refer the reader to DKM14 for a more in-depth discussion. We first start by reviewing in Sect. 2.1 the dynamical and physical features of the SMCM parameterization that are relevant for the Bayesian set-up.

2.1 The SMCM

The SMCM aims at representing the missing variability in GCMs due to unresolved processes of organized tropical convection (Johnson and Ciesielski 2013; Johnson et al. 1999; Mapes et al. 2006). Each GCM gridbox is overlaid with a lattice of $n \times n$ convective sites. Each of the n^2 sites is associated with a four state Markov process $(Y_t^i)_{t>0}$ that takes the values 0, 1, 2, or 3 according to whether the site is either cloud free, or occupied by one of the three cloud types. More illustrative details of the SMCM are found in earlier publications, e.g., KBM10, FMK12, and DKM14.

Transitions between the different cloud types, listed in Table 1, are formalized in terms of transition rates r_{jk} that are functions of the large-scale atmospheric state, and scaled by a set of timescale parameters τ_{jk}. The large-scale atmospheric state provides the three convective indicators C, C_l, and D, defined as

$$C = \frac{\text{CAPE}}{\text{CAPE}_0}, \quad C_l = \frac{\text{CAPE}_l}{\text{CAPE}_0}, \quad \text{and} \quad D = \frac{\theta_{eb} - \theta_{em}}{T_0}. \tag{1}$$

Here, CAPE and CAPE_l are the *convective available potential energies* (see Emanuel 1994), obtained by integrating the buoyancy of an adiabatically lifted parcel, over the whole and lower troposphere (see Sect. 4 and Table 2 for height levels specific to the Giga-LES study), respectively, θ_{eb} and θ_{em} are the boundary layer and mid-troposphere *equivalent potential temperatures*, and CAPE_0, T_0 are climatological reference values. D is a measure of the dryness of the mid-troposphere, i.e. $\theta_{em} \ll \theta_{eb}$ indicates a mid-troposphere that tends to be dry. The influence of C, C_l and D on the transition rates r_{jk} (see Table 1) is represented through the activation function of Arrhenius type

$$\Gamma(x) = \left\{1 - e^{-x} \quad \text{if } x > 0,\ 0 \text{ otherwise}\right\}. \tag{2}$$

In practical implementation, evolving in time each one of the n^2 microscopic Markov chains has a high computational overhead. A coarse-grained model is derived in KBM10 for the GCM grid box cloud area fractions alone, which can be easily evolved without the detailed knowledge of the microstate configuration (see also Katsoulakis et al. 2003; Khouider 2014; Khouider et al. 2003). The area fractions for congestus, deep, and stratiform clouds are given, respectively, by

Table 1 Transition rates r_{jk} given as functions of the large scale variables CAPE (C), low level CAPE (C_l), and mid troposphere dryness D, via the activation function Γ defined in (2) and modulated by the (unknown) timescale parameters τ_{jk}

Cloud transition	Transition rate
Formation of congestus	$r_{01} = \frac{1}{\tau_{01}}\Gamma(C_l)\Gamma(D)$
Decay of congestus	$r_{10} = \frac{1}{\tau_{10}}\Gamma(D)$
Conversion of congestus to deep	$r_{12} = \frac{1}{\tau_{12}}\Gamma(C)\big(1-\Gamma(D)\big)$
Formation of deep	$r_{02} = \frac{1}{\tau_{02}}\Gamma(C)\big(1-\Gamma(D)\big)$
Conversion of deep to stratiform	$r_{23} = \frac{1}{\tau_{23}}$
Decay of deep	$r_{20} = \frac{1}{\tau_{20}}\big(1-\Gamma(C)\big)$
Decay of stratiform	$r_{30} = \frac{1}{\tau_{30}}$

The transition rates r_{03}, r_{13}, r_{21}, r_{31}, and r_{32} are set zero. They represent forbidden transitions whose transition probabilities during a short period of time are negligible

$$\sigma_c^t = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{Y_t^i=1\}}, \quad \sigma_d^t = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{Y_t^i=2\}},$$
$$\sigma_s^t = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{\{Y_t^i=3\}}, \tag{3}$$

where $\mathbb{1}_{\{Y_t^i=k\}}$ is the indicator function, which takes the value one when $Y_t^i = k$ and zero otherwise. As a function of time, $(\sigma_c^t, \sigma_d^t, \sigma_s^t)$ effectively forms a three dimensional birth-death process with probability transition rules that are given in terms of the microscopic rates r_{kl}. Given the large-scale thermodynamic state, the birth-death process is easily evolved in time using Gillespie's exact algorithm (Gillespie 1975, KBM10), without significant computational overhead. However for the purpose of the Bayesian model, it is more convenient to track the cloud *populations* $N_c^t = N\sigma_c^t$, $N_d^t = N\sigma_d^t$, and $N_s^t = N\sigma_s^t$, where $N = n^2$ is the total number of microscopic sites within the coarse cell. The number of cloud free sites is given by $N_{cs}^t = N - N_c^t - N_d^t - N_s^t$.

The cloud populations N_c^t, N_d^t, and N_s^t form a three dimensional birth-death process with immigration, which is denoted by

$$(\boldsymbol{X}_t)_{t>0}, \quad \boldsymbol{X}_t = (N_c^t, N_d^t, N_s^t).$$

The probability transition matrix for this stochastic process solves the backward Kolmogorov equations (DKM14).

$$\begin{aligned}\frac{dP_{\boldsymbol{ij}}(t)}{dt} &= R_{12}^{\boldsymbol{i}} P_{\boldsymbol{i}-\boldsymbol{\varepsilon_1}+\boldsymbol{\varepsilon_2}\boldsymbol{j}}(t) + R_{23}^{\boldsymbol{i}} P_{\boldsymbol{i}-\boldsymbol{\varepsilon_2}+\boldsymbol{\varepsilon_3}\boldsymbol{j}}(t) + R_{10}^{\boldsymbol{i}} P_{\boldsymbol{i}-\boldsymbol{\varepsilon_1}\boldsymbol{j}}(t) \\ &+ R_{20}^{\boldsymbol{i}} P_{\boldsymbol{i}-\boldsymbol{\varepsilon_2}\boldsymbol{j}}(t) + R_{30}^{\boldsymbol{i}} P_{\boldsymbol{i}-\boldsymbol{\varepsilon_3}\boldsymbol{j}}(t) + R_{01}^{\boldsymbol{i}} P_{\boldsymbol{i}+\boldsymbol{\varepsilon_1}\boldsymbol{j}}(t) + R_{02}^{\boldsymbol{i}} P_{\boldsymbol{i}+\boldsymbol{\varepsilon_2}\boldsymbol{j}}(t) \\ &- \left(R_{12}^{\boldsymbol{i}} + R_{23}^{\boldsymbol{i}} + R_{10}^{\boldsymbol{i}} + R_{20}^{\boldsymbol{i}} + R_{30}^{\boldsymbol{i}} + R_{01}^{\boldsymbol{i}} + R_{02}^{\boldsymbol{i}}\right) P_{\boldsymbol{ij}}(t),\end{aligned} \tag{4}$$

with the initial conditions $P_{\boldsymbol{ij}}(0) = \delta_{\boldsymbol{ij}}$. Here $\boldsymbol{i} = (i_1, i_2, i_3)$ and $\boldsymbol{j} = (j_1, j_2, j_3)$ are triplets of non-negative integers in the range space $\mathcal{S}$ of $\boldsymbol{X}_t$, and $P_{\boldsymbol{ij}}(t)$ is the conditional probability that at time t the populations of congestus, deep and stratiform are respectively j_1, j_2, and j_3, given that there were i_1 congestus, i_2 stratiform, and i_3 deep clouds at time $t = 0$. The vectors $\boldsymbol{\varepsilon_1} = (1, 0, 0)$, $\boldsymbol{\varepsilon_2} = (0, 1, 0)$, and $\boldsymbol{\varepsilon_3} = (0, 0, 1)$ are the canonical unit vectors in $\mathbb{R}^3$. Here R_{kl} are the transition rates of the coarse-grained process. We note that only the seven admissible transitions are included in Eq. 4 (KBM10, DKM14). In the case where local interactions are ignored, all n^2 stochastic processes Y_t^i are independent and identically distributed, and the coarse transition rates satisfy (KBM10)

$$R_{kl}^{\boldsymbol{i}} = \begin{cases} i_k r_{kl}, & \text{if } k \neq 0 \\ (N - i_1 - i_2 - i_3) r_{0l}, & \text{otherwise} \end{cases}$$

where r_{kl} are the microscopic rates depending only on the exogenous factors C, C_l, D as defined in Table 1. We note that while the microscopic timescales τ_{kl} are measured in hours, the effective timescales for the coarse grained–cloud area fraction processes are given by τ_{kl}/i_k. In the extreme case where $i_k = N$, the effective transition timescale can be reduced by as much as two orders of magnitude if $N = 10 \times 10$, for example, resulting in transition times in the order of a few seconds to minutes.

If we let $P = \{P_{\boldsymbol{ij}}(t)\} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ be the matrix of transition probability functions $P_{\boldsymbol{ij}}(t)$, we may cast the Kolmogorov system in its matrix form:

$$P'(t) = RP(t), \qquad P(0) = Id, \tag{5}$$

where Id is the identity matrix of order $|\mathcal{S}|$, the cardinality of $\mathcal{S}$, and $R \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the matrix of transition rates $R_{kl}^{\boldsymbol{i}}$ (the infinitesimal generator of the birth-death process).

2.2 The Bayesian inference procedure

The SMCM simulates the evolution of the cloud populations $\boldsymbol{x} = (N_c, N_d, N_s)$[1] constrained by the large-scale atmospheric state $\boldsymbol{u} = (C, C_l, D)$. We label the corresponding sequence of observations $x_1, x_2, x_3, \ldots$ and $u_1, u_2, u_3, \ldots$ by $\boldsymbol{x}_t$ and $\boldsymbol{u}_t$, respectively. The SMCM parameterization includes seven numerical inputs (or parameters), namely the cloud convective timescales (see Table 1), which we stack in the vector

$$\boldsymbol{\theta} = (\tau_{01}, \tau_{10}, \tau_{12}, \tau_{02}, \tau_{23}, \tau_{20}, \tau_{30}).$$

While the SMCM characterizes the behavior of the future observations of $\boldsymbol{X}$ *conditional* on θ, a statistical inference method allows instead to deduce from observations $\boldsymbol{x}$ of $\boldsymbol{X}$ an *inference* about θ. A general description of this inversion is given by the Bayesian paradigm. The Bayesian approach incorporates the initial information and residual uncertainty about the model parameters $\boldsymbol{\theta}$ into a *prior* distribution $\pi(\boldsymbol{\theta})$, which is then updated by the model *likelihood* function $f(\boldsymbol{x}_t|\boldsymbol{\theta})$ to formulate a *posterior* distribution $\pi(\boldsymbol{\theta}|\boldsymbol{x}_t)$ of the parameters given the data (Robert 2007):

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_t) = \frac{f(\boldsymbol{x}_t|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{x}_t|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{6}$$

The inference is then based on the *distribution* of θ conditional on $\boldsymbol{x}$ as defined by (6). Conditioning further on $\boldsymbol{u}_t$, the posterior is given (up to a proportionality constant) as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}_t, \boldsymbol{u}_t) \propto f(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{7}$$

The likelihood function $f(\boldsymbol{x}_t|\boldsymbol{u}_t, \boldsymbol{\theta})$ is in essence an expression of the SMCM model and we refer to DKM14 for the full derivation. For series of observations $\boldsymbol{x}_{1:T} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ and $\boldsymbol{u}_{1:T} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_T)$ of length T, it is found by conditioning on past events and using the probability matrix density functions (4) for the one-step transition likelihoods (DKM14):

[1] Observations of the random variable $\boldsymbol{X}$ are here written in lower case.

$$f(\boldsymbol{x}_{1:T}|\boldsymbol{u}_{1:T},\boldsymbol{\theta}) = \prod_{t=1}^{T} f_{t-1}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{u}_{t-1},\boldsymbol{\theta})$$
$$= \prod_{t=1}^{T} \mathbb{1}^{*}_{\boldsymbol{\phi}(\boldsymbol{d}^{t-1},N_d^{t-1},N_s^{t-1})} \exp[R(\boldsymbol{u}_{t-1},\boldsymbol{\theta})h]\mathbb{1}_{\boldsymbol{\phi}(\boldsymbol{d}^{t},N_d^{t},N_s^{t})}. \quad (8)$$

Here $\mathbb{1}^{*}_{\boldsymbol{\phi}(\cdot)}$ is the transpose of $\mathbb{1}_{\boldsymbol{\phi}(\cdot)}$ ($\mathbb{1}_{\boldsymbol{\phi}(\cdot)}$ is the canonical vector in $\mathbb{R}^{|\mathcal{S}|}$ that has 1 at the index corresponding to $\phi(\cdot)$, and 0's everywhere else), and $\phi(d,b,c) = \frac{d^3}{6} + \frac{d^2}{2} + \frac{d}{3} + db - \frac{b^2}{2} + \frac{3b}{2} + c$ for non-negative integers $0 \le d \le N$, $0 \le b \le d$, and $0 \le c \le d - b$, and $d^t = N_c^t + N_d^t + N_s^t$. The function $\phi : \mathcal{S} \to \mathbb{N}$ maps each triple in $\mathcal{S}$ to a counting order (address), which is needed to automate the construction of the large matrix R. Note that the matrix exponential $\exp[R(\cdot)h]$ is the solution to the system (5) assuming that the large-scale environment varies on a longer timescale than the data sampling time interval h. In the Giga-LES inference study presented in here, cloud data is available every 15 min ($h = 15$ min) and over such period of time, the large-scale variable $\boldsymbol{u}$ is effectively approximately constant (see Fig. 4).

It is easy to see that $dim = \mathcal{O}(N^3)$, and thus the size and memory requirements of R become prohibitively large with the dimensions of the cloud lattice (DKM14). A parallel version of a preconditioning technique known as the *Uniformization Method* was developed in DKM14 that allows for fast, numerically stable, and scalable approximations of large sparse matrix exponentials. The sampling of the posterior distribution is done with the Monte Carlo Markov Chain technique (see DKM14 and references therein).

3 Giga-LES dataset and sequential learning

The choice of observed data for the multicloud parameter estimation problem hinges on two major points which characterize tropical convection: (1) a fine enough resolution to capture the small (time and spacial) scale processes associated with deep convection and (2) a large enough domain in order to represent some level of multiscale organization of coherent structures. This inference study is based on the Giga-LES dataset (Khairoutdinov et al. 2009), a *large-eddy simulation* (LES) of deep tropical convection on a numerical domain comparable to a GCM grid cell. Traditionally, LES have been used to simulate turbulence and low clouds in the PBL, where the grid spacing of $\mathcal{O}$(10–100 m) is small enough to explicitly represent turbulent processes associated with large eddies occurring in the boundary layer. The "Giga-LES" is one of the very few studies that extends the technique to deep convection in the atmosphere,[2] with a grid spacing of 100 m. It can simulate deep convective cloud processes and exhibit some mesoscale organization characterized by a tri-modal vertical distribution of deep, middle, and shallow clouds similar to that often observed in the tropics (Khairoutdinov et al. 2009).

The Giga-LES dataset is a 24-h long LES of deep tropical convection over a domain of 204.8 km in both horizontal directions and about 27 km in the vertical, which uses the mean sounding and forcing observed during the GARP Atlantic Tropical Experiment (GATE) Phase III experiment over the Atlantic Inter-Tropical Convergence Zone (ITCZ) (Khairoutdinov et al. 2009). The atmospheric fields are available every 15 min at all $2048 \times 2048 \times 256$ grid points of the three dimensional space. A full description of the simulation setup, including the idealized mean GATE initial thermodynamic profiles and large–scale forcing is found in Khairoutdinov et al. (2009). Figure 1 presents a visualization of the cloud scene over the whole $204.8 \times 204.8\,\text{km}^2$ domain at hour 13. The scene illustrates complex convection activity, with individual deep clouds and mesoscale cloud systems dominated by stratiform anvils, surrounded by smaller congestus and shallow clouds.

The evolution of convection in the simulation is illustrated by the time series of the vertical profile of horizontally averaged non precipitating cloud liquid/ice condensate, in Fig. 2a. The convection activity triggers after a "spin-up" transient period of approximately 6 h, with a shallow boundary layer appearing during that period. Figure 2a shows a shallow cloud layer that gradually deepens until a burst of deep cumulus convection occurs near hour 6. A nearly steady deep cumulus regime is established by hour 12, characterized by a trimodal vertical distribution of the cloud field; namely formed by shallow and deep convective cloud maxima accompanied by a cumulus congestus maximum within the lower troposphere, i.e, near the freezing level (Khairoutdinov et al. 2009). Figure 2b presents the horizontally averaged vertical profiles of cloud water/ice and water vapor mixing ratios, and relative humidity, averaged over the last 12 h of the simulation period. The water/ice mixing ratio is of particular interest to this study, as it is used to derive the gridbox cloud area fractions $\sigma_c, \sigma_d, \sigma_s$ (see Sect. 4.2 for details).

Figure 1b shows the time evolution of horizontally averaged CAPE, low level CAPE, and midtroposphere dryness $D = (\theta_{eb} - \theta_{em})/T_0$ ($T_0 = 10$ K). The CAPE and low level CAPE were computed from the domain-averaged

[2] A simulation rerun, which uses a spatial resolution of 50 m and covers a physical domain of 86 km × 86 km × 22 km, has been carried by Loh and Austin (2015; manuscript in preparation) to study entrainment and detrainment rates.

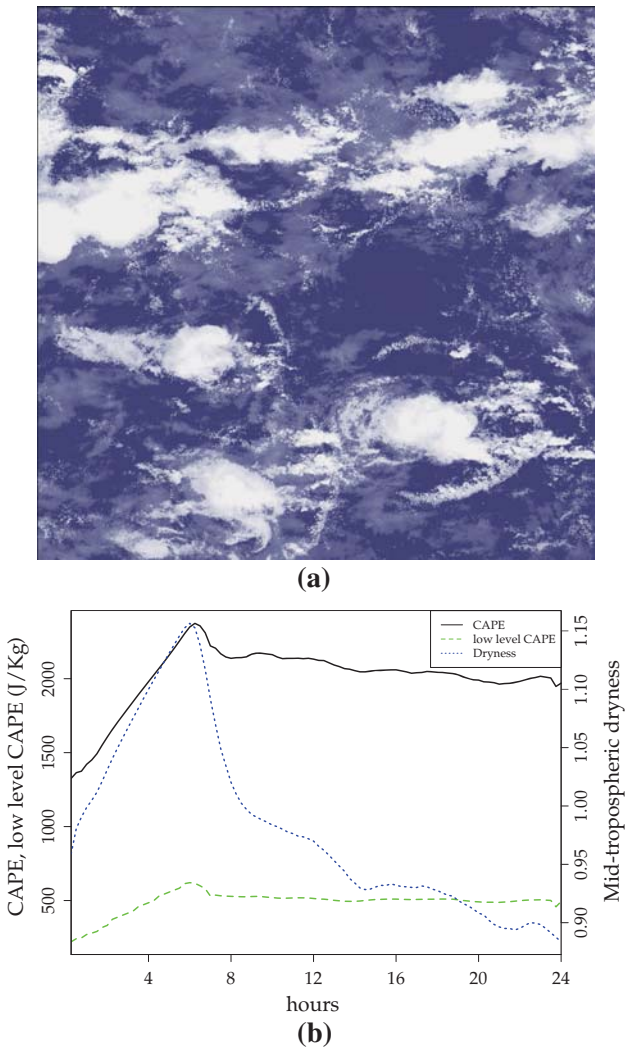


Fig. 1 **a** Image of simulated cloud scene over the area $205 \times 205\,\mathrm{km}^2$ corresponding to hour 13 of the Giga-LES simulation, obtained from visible albedo estimated from the liquid and ice water paths (from Khairoutdinov et al. 2009). **b** Time evolution of CAPE (J/kg), low level CAPE (J/kg), and midtroposphere dryness $D = (\theta_{eb} - \theta_{em})/T_0$, $T_0 = 10\,\mathrm{K}$, calculated from the horizontally averaged fields of the Giga-LES dataset. The midtroposphere dryness D measures the discrepancy between the boundary layer and midtroposphere equivalent potential temperatures, fixed at pressure levels 1000 and 500 mb, respectively

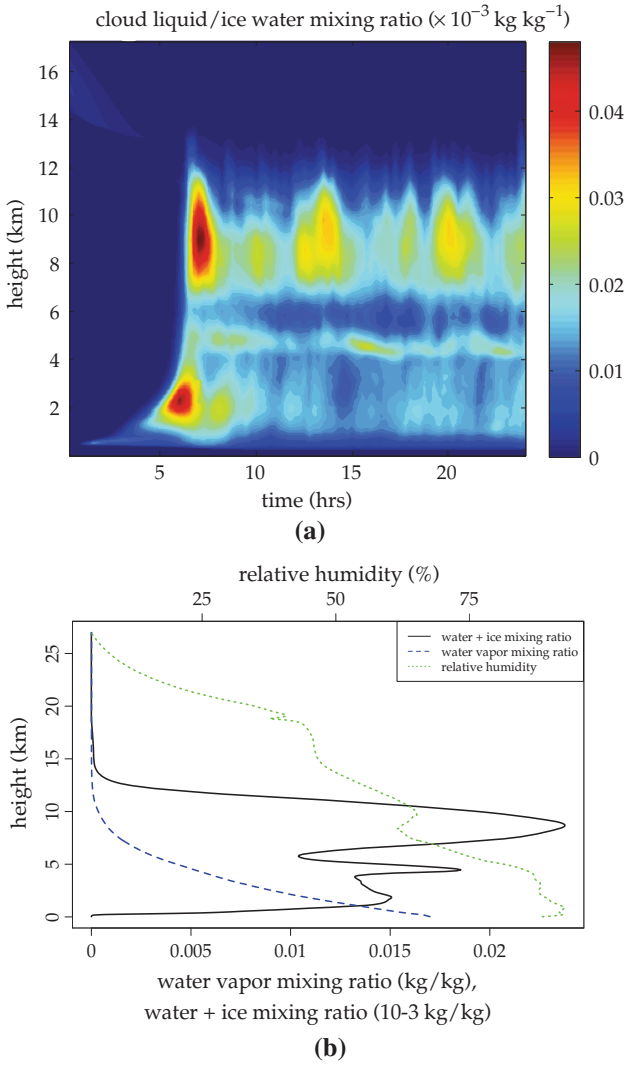


Fig. 2 **a** Evolution of horizontally averaged cloud liquid/ice water mixing ratio vertical profile. **b** Comparison of vertical profiles of horizontally averaged cloud water/ice mixing ratio (*solid black*), water vapor mixing ratio (*dashed blue*), and relative humidity (*dotted green*), averaged over the last 12 h of the Giga-LES simulation run

thermodynamic profiles assuming pseudo-adiabatic ascent with the departure point at 960 mb.

During the "spin-up" transition, the CAPE value increases by about 70 %, after which the onset of deep convection occurs and consumes a fraction of that CAPE. By hour 8, the simulation reaches a deep convection regime with an approximately steady CAPE value of 2000–2100 J/kg. The midtroposphere dryness also reaches a maximum towards the 6 h mark, indicating a moistening of the boundary layer which sets the favourable conditions for deep convection; It then gradually drops as the midtroposphere moisture content increases due to deep convection. The CAPE in the lower troposphere reaches a maximum of about 625 J/kg at the onset of convection which consumes about 16 % of that amount to stabilize at around 500 J/kg, indicating a sustained low level convection activity throughout the last 18 h of simulation.

3.1 Domain partitioning and sequential learning

The Giga-LES is a 24-h long simulation, with a time resolution of 15 min. Excluding a transient period of approximately 4–5 h, the length of the time series is between 76 and 80 data points. To increase information capacity, the full $205 \times 205\,\mathrm{km}^2$ Giga-LES domain is subdivided into

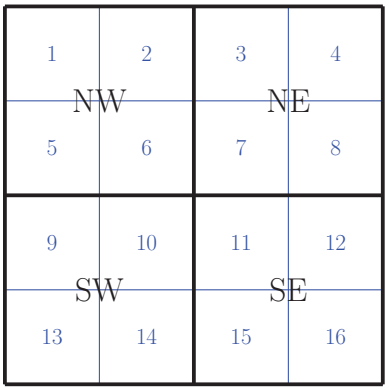


Fig. 3 Partitioning the Giga-LES domain of area $205 \times 205\,\text{km}^2$ into a 2 by 2 (*in black*) and a 4 by 4 (*in blue*) grid. The subdomains are referenced as NW,…, SE and 1,…,16 therein

M subdomains, and time series of cloud populations and large-scale convection indicators are obtained for each.

The parameters are then progressively learned using a sequential learning technique described in DKM14, which in simple terms consists of a Bayesian updating scheme of the form

$$\pi_{m+1}(\theta) \propto \pi_m(\theta) \times f_{m+1}(\theta), \quad 0 \leq m \leq M-1,$$

where a prior $\pi_m(\theta)$ gets updated by a likelihood $f_{m+1}(\theta)$ to give a posterior $\pi_{m+1}(\theta)$. This gives the following recursive scheme: Starting on subdomain 1 and specifying an initial prior π_0, we run the Bayesian posterior simulator that outputs a posterior π_1, which in turn is used as a prior for subdomain 2, etc. At each step, a multivariate normal distribution is fitted to the posterior using the sample mean and sample covariance matrix. This method was successfully validated in DKM14 where the inference results based on a synthetic time series of length $5a$ were compared to the sequential analysis obtained by segmenting that time series into five contiguous subsequences of length a.

Figure 3 illustrates the two partitions that are used in this study: a 2×2 ($M = 4$) grid and a 4×4 ($M = 16$) grid, which correspond to GCM grid box sizes of 102.4 and 51.2 km, respectively. Time series for the cloud cover fractions σ_c, σ_d, and σ_s and large-scale convection indicators C, C_l, and D are derived next for the two partitions.

4 Data preprocessing

As mentioned in Sect. 2.2, the Bayesian inference procedure draws its inference on the convective timescale parameters τ_{jk} from two training sets of observations: the large-scale convection indicators (1), and the subgrid-scale area fractions σ_c, σ_d, and σ_s of congestus, deep, and stratiform clouds (see 3). The amount of convective available potential energy (CAPE) of the environment is determined from simple parcel theory (assuming pseudo-adiabatic ascent with the departure point at 960 mb; see Emanuel (1994) for more details), while the dryness D is obtained directly from the equivalent potential temperature's vertical profile. The time series for the large-scale convective indicators are presented in Sect. 4.1.

Fractional cloud area, on the other hand, is not a well defined quantity and deriving cloud fractions from experimental radar/lidar or simulated data can be done in various ways. Our calculation of the cloud area fractions σ_c, σ_d, and σ_s is based on a diagnosis of water and ice mixing ratios present within single vertical columns to identify clouds of the three types. The time series of the subgrid-scale cloud area fractions are presented in Sect. 4.2.

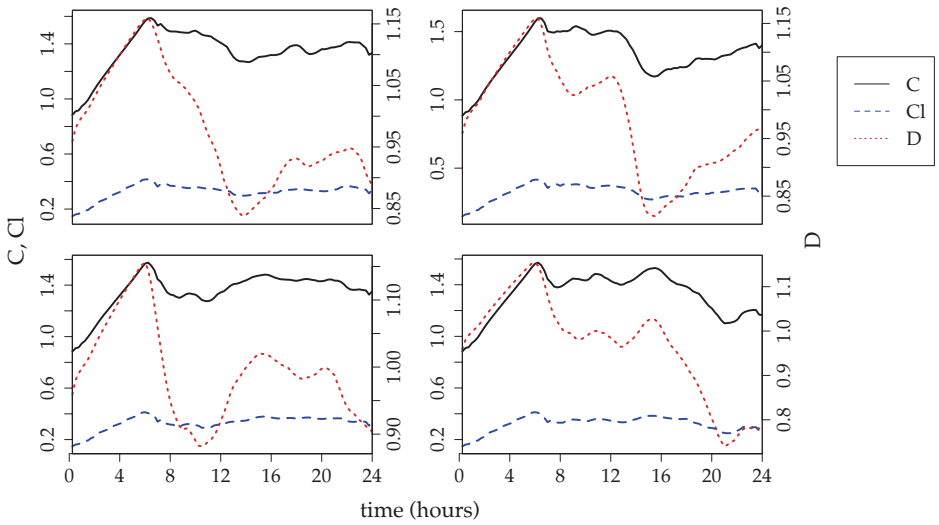
4.1 Time series of large-scale convection indicators

The time series of the large-scale convection indicators C, C_l, and D, defined in (1), are shown in Fig. 4 using the two reference values $\text{CAPE}_0 = 1500$ J/kg, and $T_0 = 10$ K for both the 2 by 2 and 4 by 4 grids. The corresponding CAPE_l and CAPE values were obtained by integrating the parcel's buoyancy over the lower and whole troposphere, respectively, whose base and top levels are fixed at the parcel's LFC and LNB (see Table 2 for values specific to the Giga-LES dataset).

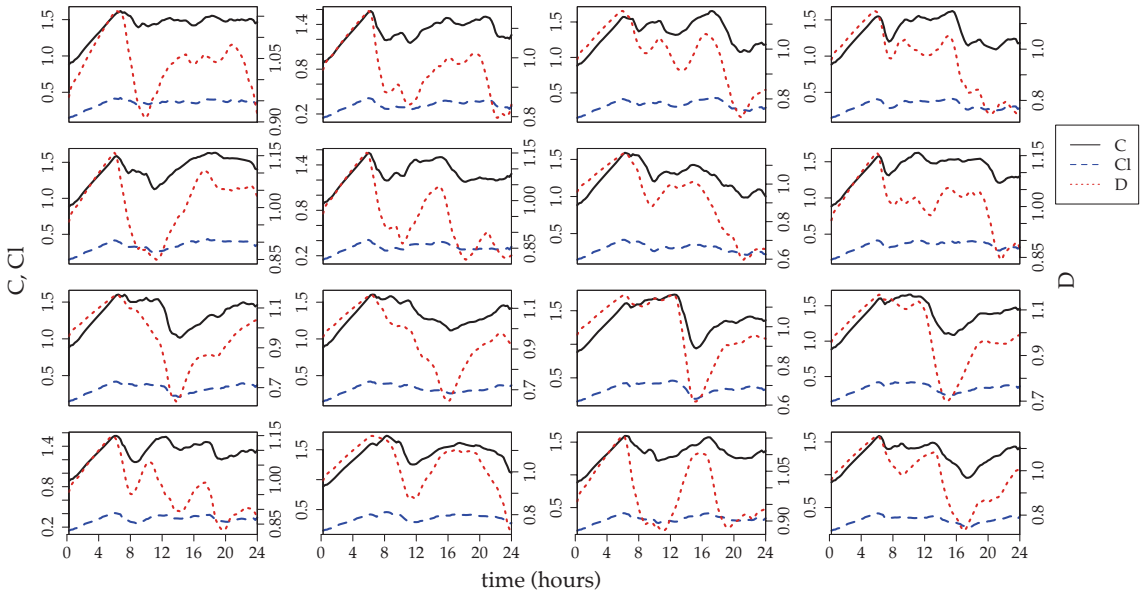
The time series of the gridded domains are qualitatively similar to those of the full domain shown in Fig. 1b: a buildup of convective energy combined to a moistening of the boundary layer takes place until an explosive transition to deep cumulus convection occurs near hour 6, which depletes the atmosphere of CAPE and moisten the midtroposphere.

4.2 Time series of cloud area fractions

The cloud fractions associated with congestus, deep and stratiform clouds are derived from the prognostic cloud water/ice mixing ratio q_n. At every grid point (i, j) of the 2048×2048 horizontal gridded domain, we consider the vertical profile of q_n and binarize it using a zero threshold to obtain a 256 level binary vector $\boldsymbol{Q}_{ij}$. Each binary vertical profile is then compared to four cloud/no cloud reference profiles: congestus $\boldsymbol{\mathcal{P}_c}$, deep $\boldsymbol{\mathcal{P}_d}$, stratiform $\boldsymbol{\mathcal{P}_s}$, and clear sky $\boldsymbol{\mathcal{P}_{cs}}$. These profiles are constructed using the *lifted convection level* (LCL) as an estimate for congestus and deep cloud base, the *freezing level* (FL) as an estimated congestus cloud top and stratiform cloud base, and the level of *neutral buoyancy* (LNB) as an estimated deep and stratiform cloud tops, as shown in Fig. 5.



(a) 2 by 2 grid.



(b) 4 by 4 grid.

Fig. 4 Time series of the large-scale convection indicators C, C_l, and D using reference values $\mathrm{CAPE}_0 = 1500$ J/kg, and $T_0 = 10$ K for the **a** 2 by 2 grid and **b** 4 by 4 grid. The dryness D measures the discrepancy between the boundary layer and midtroposphere equivalent potential temperatures, set at pressure levels 1000 and 500 mb, respectively. Note that two different scales are used to represent C (C_l) and D

The approximate height for these reference levels for the Giga-LES dataset are reported in Table 2. The reference profiles are set as

$$\boldsymbol{\mathcal{P}_c} = \mathbb{1}_{\{\mathrm{LCL}\le z\le \mathrm{FL}\}}, \boldsymbol{\mathcal{P}_d} = \mathbb{1}_{\{\mathrm{LCL}\le z\le \mathrm{LNB}\}},$$
$$\boldsymbol{\mathcal{P}_s} = \mathbb{1}_{\{\mathrm{FL}\le z\le \mathrm{LNB}\}}, \boldsymbol{\mathcal{P}_{cs}} = \mathbf{0},$$

where z is the height, $\mathbb{1}$ is an indicator vector function, and $\mathbf{0}$ the vector of zeros. A cloud type is assigned to the LES column data $\boldsymbol{Q}_{ij}$ by minimizing some misfit measure between $\boldsymbol{Q}_{ij}$ and the set of reference profiles $\mathcal{P} = \{\boldsymbol{\mathcal{P}_c}, \boldsymbol{\mathcal{P}_d}, \boldsymbol{\mathcal{P}_s}, \boldsymbol{\mathcal{P}_{cs}}\}$. Here we minimize the 2-norm of the residual vector

$$\|\boldsymbol{Q}_{ij} - \boldsymbol{P}\|_2 = \sqrt{\sum_{k=1}^{256} \left(Q_{ij}^k - P^k\right)^2},$$

where k is the vertical level, and $\boldsymbol{P} \in \mathcal{P}$. The result is a projected two-dimensional cloud lattice whose cloud area fraction time series are shown in Fig. 6 for both the 2 by 2 and

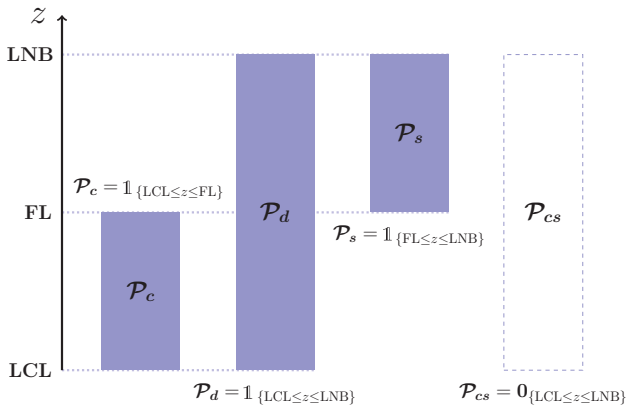


Fig. 5 Reference profiles $\mathcal{P}_c$, $\mathcal{P}_d$, $\mathcal{P}_s$, and $\mathcal{P}_{cs}$ of congestus, deep and stratiform clouds, and clear sky. The LCL, FL, and LNB approximate heights are given in Table 2

Table 2 Approximate LCL, LFC, and LNB obtained, using parcel theory, from the domain and time averaged thermodynamic fields of the Giga-LES dataset. FL is the *freezing level*, defined as the 273 K height. The LFC and LNB are determined from the sign of the parcel's buoyancy (see Emanuel 1994), and are used in the CAPE calculation

Level	Approximate height
Lifted condensation level (LCL)	355 m
Level of free convection (LFC)	455 m
Freezing level (FL)	4.4 km
Level of neutral buoyancy (LNB)	14 km

4 by 4 partitions. The time series show an interesting pattern of intermittent cloud bursts, with frequent signatures of congestus events preceding deep and stratiform events. Larger stratiform area fractions indicate strong upper level cloud condensate.

5 Results

5.1 Convergence of sequential learning

The sequential learning strategy was applied to both the 2 by 2 and 4 by 4 grids of Fig. 3 using an initial weakly informative multivariate normal prior, with the mean given by the P2013 parameter regime (see Table 5), and variance $50\,\mathrm{h}^2$. Figures 7 and 8 show the marginal posteriors for all seven parameters for the 2 by 2 and 4 by 4 grids, respectively, using the large-scale convective indicator and cloud area fraction time series given in Figs. 4 and 6.

Each posterior exploration was conducted using an ensemble of well-dispersed MCMC chains in parallel, each with a sample size of approximately 100 000. Burn in periods were removed and proposal variances were calibrated to obtain an optimal acceptance rate of 25 %. Visual diagnostics were used to monitor within-chain and in between-chain mixing, and ensure that the chains have reached equilibrium. All parallel simulations were performed on the Nestor Westgrid cluster using 72 cores.

For both partition cases, the posterior densities gradually concentrate about a mean value with a progressive reduction of the variance, and appear to converge to a limiting density. The influence of the prior is negligible: The posterior is strongly dominated by the data likelihood function right after the first learning sequence.

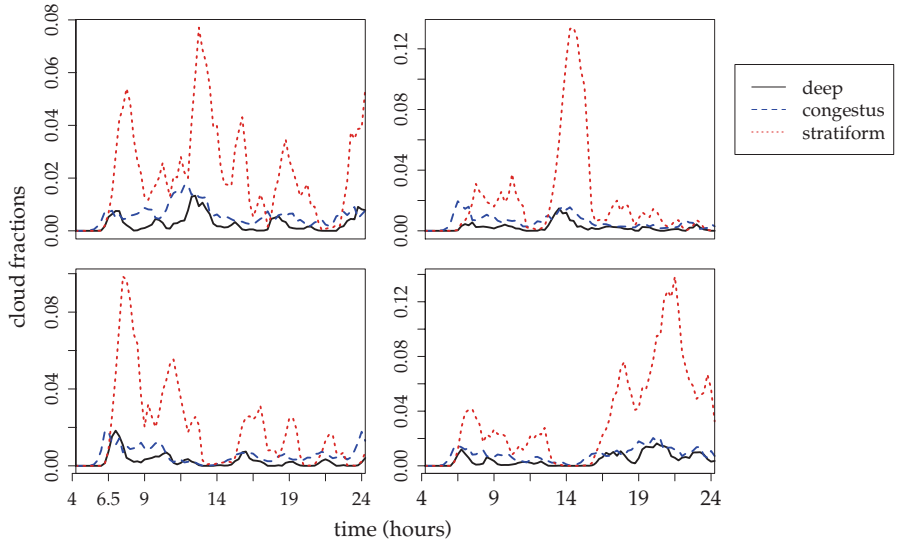
The results for the two partitions are juxtaposed in the box plot of Fig. 9 depicting the first, second, and third quartiles for both sets of posterior marginals. For most parameters, there are large discrepancies in the values for these two cases. The parameters are classified as "slow" and "fast" transitions, depending whether their inferred value is less than or greater than 1.5 h. The means and standard deviations for the two grids are reported in Table 3.

As we can see from Table 3, all the mean transition timescales appear to be smaller in the fine 4×4 partition but τ_{01} which seems to increase by roughly 15 % (from 27.686 to 31.789 h). Also the amounts by which the majority of the timescales decrease vary considerably among the τ_{ij}'s. While the physical meaning of this rather erratic behavior is hard to comprehend, it sets an interesting challenge on the way these transition timescales should actually depend on the GCM grid resolution. Nonetheless, it is interesting to note that the variance is consistently smaller for all parameters with the 4×4 partition.

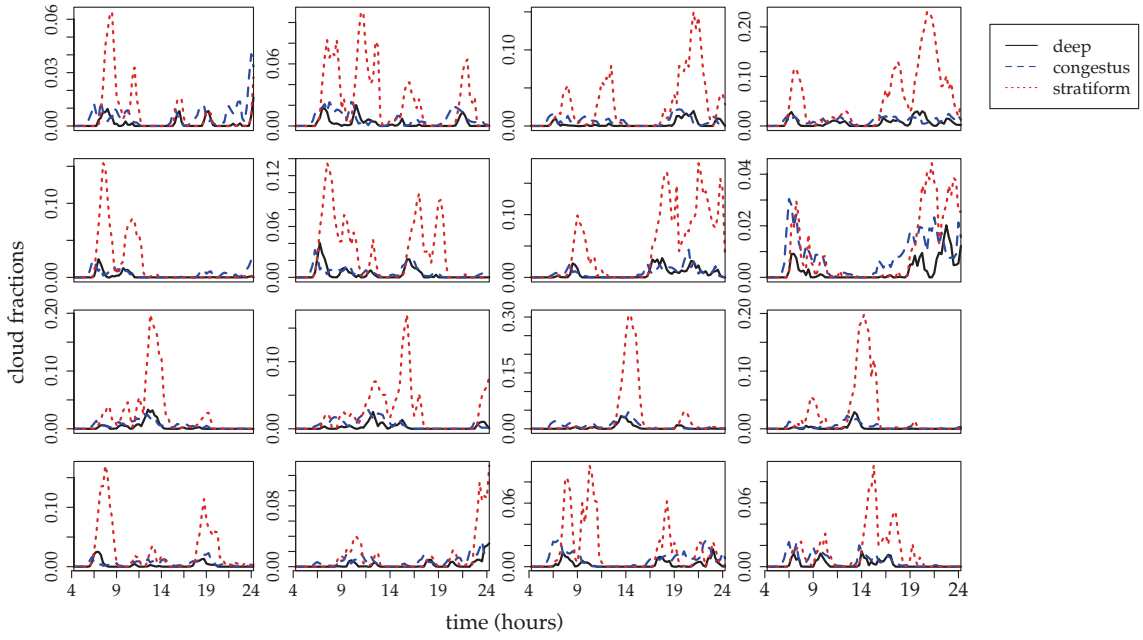
5.2 Sensitivity to the activation function parameters

Figure 10 shows a sensitivity study to the activation function parameters CAPE_0 and T_0 in the definition (1) of C, C_l, and D. It is interesting to note that varying either reference values does not have a striking effect on the convective timescale inferred values. However for some parameters the discrepancy is large, notably in the case of varying CAPE_0 for the parameters τ_{01} and τ_{02}, for which there is a gap of more than 10 h in the two sets of estimated median values, especially when comparing the small CAPE_0 value of 20 J/kg with the two larger ones (1500 and 2000 J/kg). However, some sensitivity of the same two parameters, τ_{01} and τ_{02}, can also be seen in the variation of T_0. This suggests that those two parameters, which affect directly the initiation of convection, are in fact sensitive to the large scale thermodynamics, in terms of CAPE and dryness.

From Fig. 10, we can notice that while both τ_{01} and τ_{02} decrease with the CAPE_0, the effect of variation in T_0 on the two parameters is in opposite direction; τ_{01} decreases with T_0 while τ_{02} appears to increase with T_0. This behaviour



(a) 2 by 2 grid.



(b) 4 by 4 grid.

Fig. 6 Time series of cloud area fractions for the **a** 2 by 2 grid and **b** 4 by 4 grid

is intimately tied to the way the transition rates in Table 1 have been defined in terms of CAPE and dryness. Since the intrinsic transition rates are set by the true dynamics of the LES simulation, i.e the observed data time series, they can be assumed fixed during the inference procedure. The same applies for the actual CAPE, CAPE_l and the dryness time series. Thus, as far as the inference procedure is concerned, the timescales, τ_{01} and τ_{02} in particular can be viewed as two functions of CAPE_0 and T_0. Given that the function Γ in (2) is increasing and that both CAPE_0 and T_0 appear on the denominator of C, C_l and D, respectively, it is easy to see that both τ_{01} and τ_{02} are decreasing functions of CAPE_0 while τ_{01} is decreasing with T_0 and τ_{02} is increasing with T_0, consistent with the numerical results in Fig. 10.

5.3 Single column testing

We now use the timescales inferred from the Giga-LES dataset in a simple single column climate model coupled to the SMCM parameterization to illustrate the effectiveness

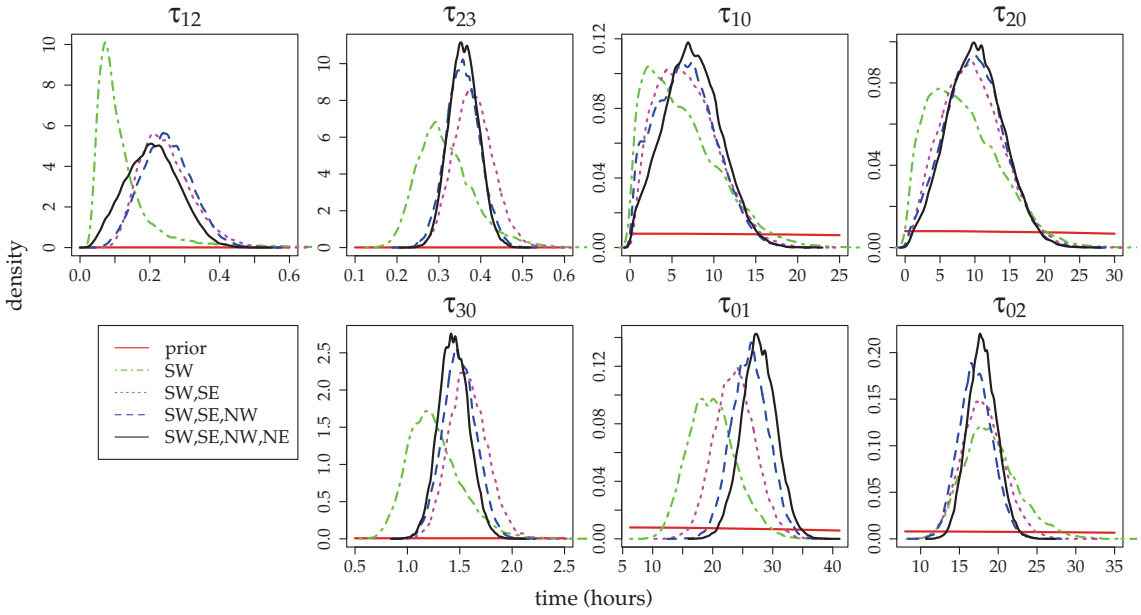


Fig. 7 Marginal posterior densities of the seven convective timescale parameters for the 2 by 2 grid, using an initial multivariate normal prior (*in red*) with mean given by the P2013 values of Table 5 and variance 50. The parameters are sequentially learned from the SW, SE, NW, and NE subdomains, with the final inference given by the *black curve*

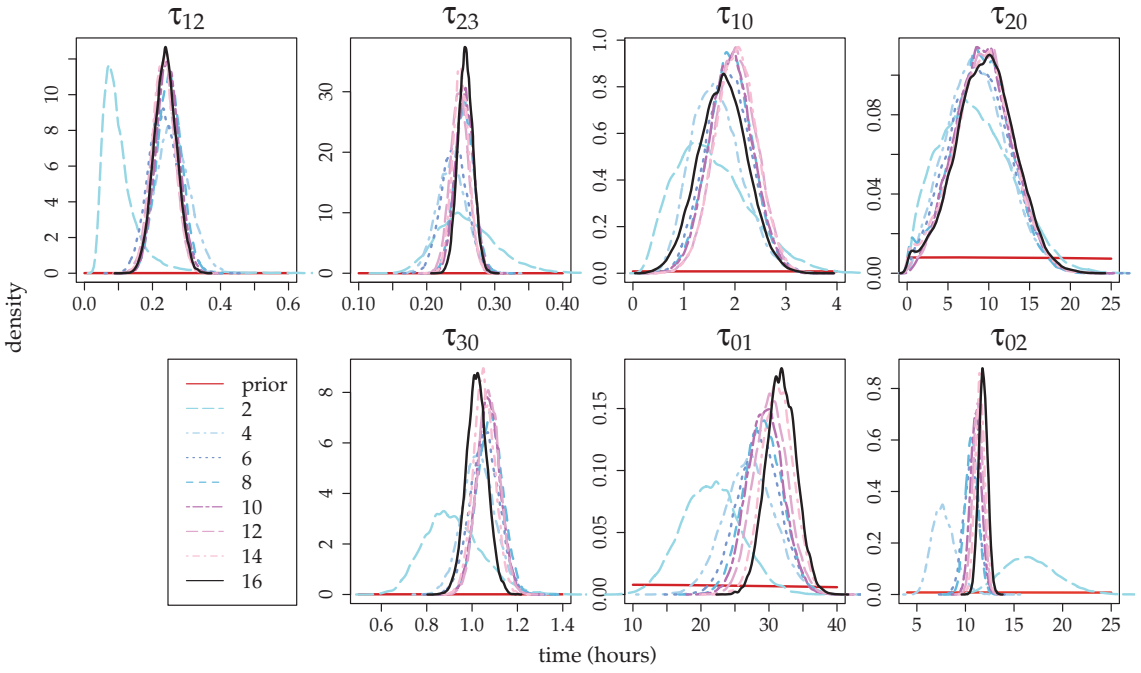


Fig. 8 Marginal posterior densities of the seven convective timescale parameters for the 4 by 4 grid, using an initial multivariate normal prior (*in red*) with mean given by the P2013 values of Table 5 and variance 50. The parameters are sequentially learned from the subdomains 1 through 16, with the final inference given by the black curve. Only even subdomains are represented for clarity

of this procedure. A somewhat more elaborate testing involving a zonally symmetric model for the monsoon meridional circulation will be presented elsewhere.

We consider the single column model used in KBM10 to test the SMCM. It consists of four diagnostic variables representing the first and second baroclinic components of

Table 3 Bayes estimates for the marginal posterior densities of Fig. 7 (2 × 2 partition) and 8 (4 × 4 partition), shown in solid black

Parameter	Mean (SD) [hours]	
	2 × 2 Partition	4 × 4 Partition
τ_{01} (Formation of congestus)	27.686 (8.233)	31.789 (4.795)
τ_{10} (Decay of congestus)	7.426 (11.155)	1.761 (0.224)
τ_{12} (Conversion of congestus to deep)	0.208 (0.006)	0.238 (0.001)
τ_{02} (formation of deep)	17.950 (3.507)	11.821 (0.211)
τ_{23} (conversion of deep to stratiform)	0.359 (0.001)	0.2570 (0.0001)
τ_{20} (decay of deep)	10.126 (15.674)	9.551 (13.146)
τ_{30} (decay of stratiform)	1.444 (0.021)	1.021 (0.002)

Mean and SD are posterior mean and standard deviation, respectively. The finest 4 × 4 partition decreases the estimated variances by as much as an order of magnitude for some of the parameters

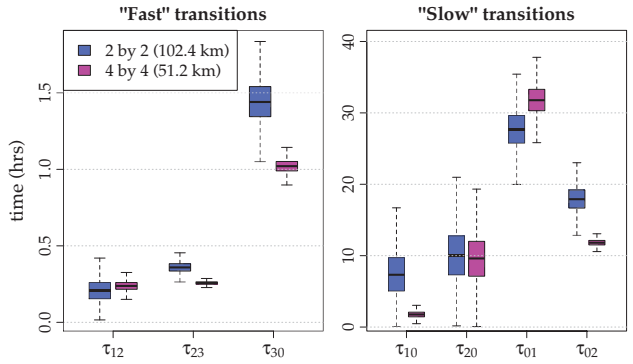


Fig. 9 *Boxplot* of the marginal posteriors shown in Figs. 7 and 8 for the 2 by 2 and 4 by 4 partitions, respectively. The *bottom and top of the boxes* represent the first and third quartiles, and the *band inside the box* is the second quartile

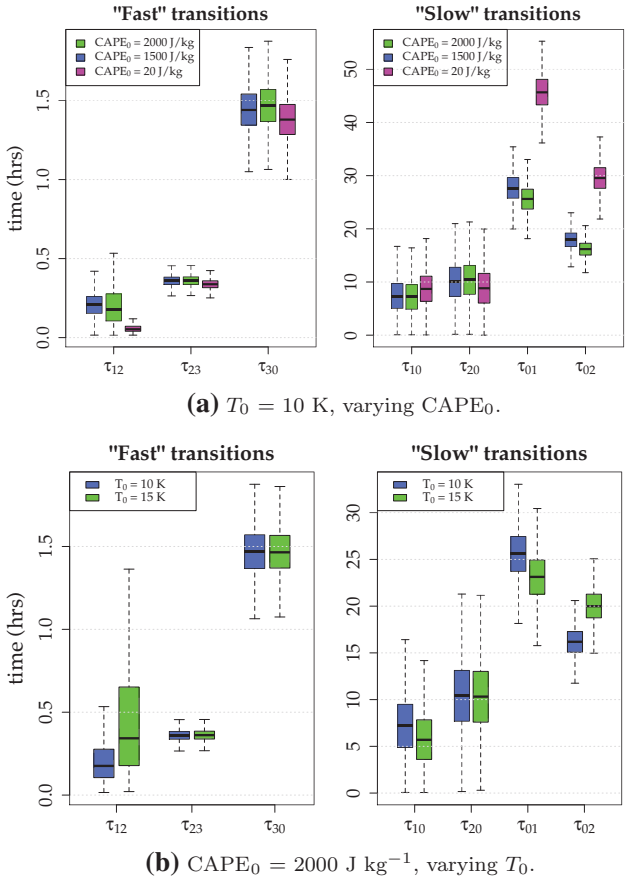


Fig. 10 *Boxplot* of the marginal posteriors for the 2 by 2 partition, **a** using $T_0 = 10$ K and varying CAPE_0, and **b** using $\text{CAPE}_0 = 2000$ J kg^{-1} and varying T_0. The *bottom and top of the boxes* represent the first and third quartiles, and the *band inside the box* is the second quartile

potential temperature, θ_1, θ_2, the vertically integrated moisture q, and the boundary layer potential temperature, θ_{eb}.

$$\begin{aligned}
\frac{d\theta_1}{dt} &= H_d + \xi_s H_s + \xi_c H_c - Q^0_{R,1} - \frac{1}{\tau_R}\theta_1,\\
\frac{d\theta_2}{dt} &= H_c - H_s - Q^0_{R,2} - \frac{1}{\tau_R}\theta_2,\\
\frac{d\theta_{eb}}{dt} &= \frac{1}{\tau_e}(\theta^*_{eb} - \theta_{eb}) - \frac{1}{h}D_c,\\
\frac{dq}{dt} &= -\frac{2\sqrt{2}}{\pi}P + \frac{1}{H_T}D_c.
\end{aligned} \tag{9}$$

Here H_d, H_c, H_s are the heating rates associated with deep convective, congestus, and stratiform clouds, and ξ_s, ξ_c are the stratiform and congestus contributions to the first baroclinic mode. The parameters $Q^0_{R,1}, Q^0_{R,2}$ are prescribed cooling rates due to long wave radiation, and τ_R and τ_e are the Newtonian cooling and surface evaporation timescales. P and D_c are respectively the precipitation rate and downdraft mass flux that serves to moisten the midtroposphere, cool and dry the boundary layer. Moreover, H_T and h are the heights of the troposphere and boundary layer, respectively, and θ^*_{eb} is the saturation equivalent potential temperature. The closure equations of the heating rates, downdrafts, and precipitation rate are listed in Table 4, together with the set of parameters and constant values used by the model, for the sake of completeness. While a detailed discussion of this model and its coupling to the SMCM is found in KBM10 and FMK12 (see also DKM14), it is worthwhile noting that the heating rates are set proportional to the cloud area fractions of their respective cloud types: an increased area fraction of deep convection, for example, yields an increased potential for deep convective heating and zero deep area fraction results in a zero deep convective heating.

In summary, the single column model in (9) is coupled to the SMCM birth-death process presented in Sect. 2.1, following KBM10, to simulate the climatology and the corresponding cloud area fraction dynamics in the G-LES

Table 4 Parameters and constants for the single column GCM coupled to the SMCM parameterization

Variable/Constant Name	Closure Equation/Value
Difference between RCE boundary layer equivalent potential temperature and its saturation value	$\theta^*_{eb} - \bar{\theta}_{eb} = 10$ K
Difference between RCE boundary layer and midtropospheric equivalent potential temperatures	$\bar{\theta}_{eb} - \bar{\theta}_{em} = 11$ K
First and second baroclinic radiative cooling rate	$Q^0_{R,1} = 1\text{K day}^{-1}$, $Q^0_{R,2}$ (determined at RCE, see KBM10)
Deep convective heating	$H_d = \left[\frac{\sigma_d}{H_m}\sqrt{\overline{\text{CAPE}}} + \frac{\sigma_d}{\overline{\sigma_d}\tau_c}(a_1\theta'_{eb} + a_2 q' - a_0(\theta'_1 + \gamma_2\theta'_2))\right]^+$
Stratiform heating	$\frac{dH_s}{dt} = \frac{1}{\tau_s}(\alpha_s\sigma_s H_d/\overline{\sigma_d} - H_s)$
Congestus heating	$H_c = \frac{\sigma_c}{H_m\alpha_c}\sqrt{\text{CAPE}_l}$
Precipitation rate	$P = \frac{2\sqrt{2}}{\pi}H_d$
Downdrafts	$D_c = m_0(1 + \mu(H_s - H_c)/Q^0_{R,1})^+(\theta_{eb} - \theta_{em})$
Mid-troposphere equivalent potential temperature	$\theta_{em} = q + \frac{2\sqrt{2}}{\pi}(\theta_1 + \alpha_2\theta_2)$
CAPE integrated over the whole troposphere	$\text{CAPE} = \left[\overline{\text{CAPE}} + R(\theta'_{eb} - \gamma(\theta'_1 + \gamma_2\theta'_2))\right]^+$
CAPE integrated over the lower troposphere	$\text{CAPE}_l = \left[\overline{\text{CAPE}} + R(\theta'_{eb} - \gamma(\theta'_1 + \gamma'_2\theta'_2))\right]^+$
Stratiform, congestus adjustment timescale	$\tau_s = 3$ h, $\tau_c = 2$ h
Newtonian cooling, surface evaporation timescale	$\tau_R = 75$ days, τ_e (determined at RCE, see KBM10)
ABL depth, free troposphere depth, mid—troposphere height	$h_b = 500$ m, $H_T = 16$ km, $H_m = 5$ km
Downdraft mass flux scale	m_0 (determined at RCE, see KBM10)
Relative contribution of θ_{eb}, q to deep convection	$a_1 = 0.5$, $a_2 = 0.5$
Contribution of θ_1 to deep convective heating anomalies	$a_0 = 2$
Contribution of θ_1 to CAPE anomalies	$\gamma = 1.7$
Contribution of θ_2 to deep convective heating anomalies	$\gamma_2 = 0.1$
Contribution of θ_2 to low level CAPE anomalies	$\gamma'_2 = 2$
CAPE constant	$R = 320\text{ J kg}^{-1}\text{ K}^{-1}$
Unit scale of temperature	$\alpha \approx 15$ K
Contribution of CAPE to stratiform, congestus heating	$\alpha_s = 0.25$, $\alpha_c = 0.1$
Value of CAPE at RCE	$Q^0_{R,1} = \overline{\sigma_d}\frac{\alpha}{H_m}\sqrt{\overline{\text{CAPE}}}$ (determined at RCE, see KBM10)
Congestus, deep, and stratiform cloud area fractions at RCE	$\overline{\sigma_c}$, $\overline{\sigma_d}$, and $\overline{\sigma_s}$ (determined at RCE, see KBM10)

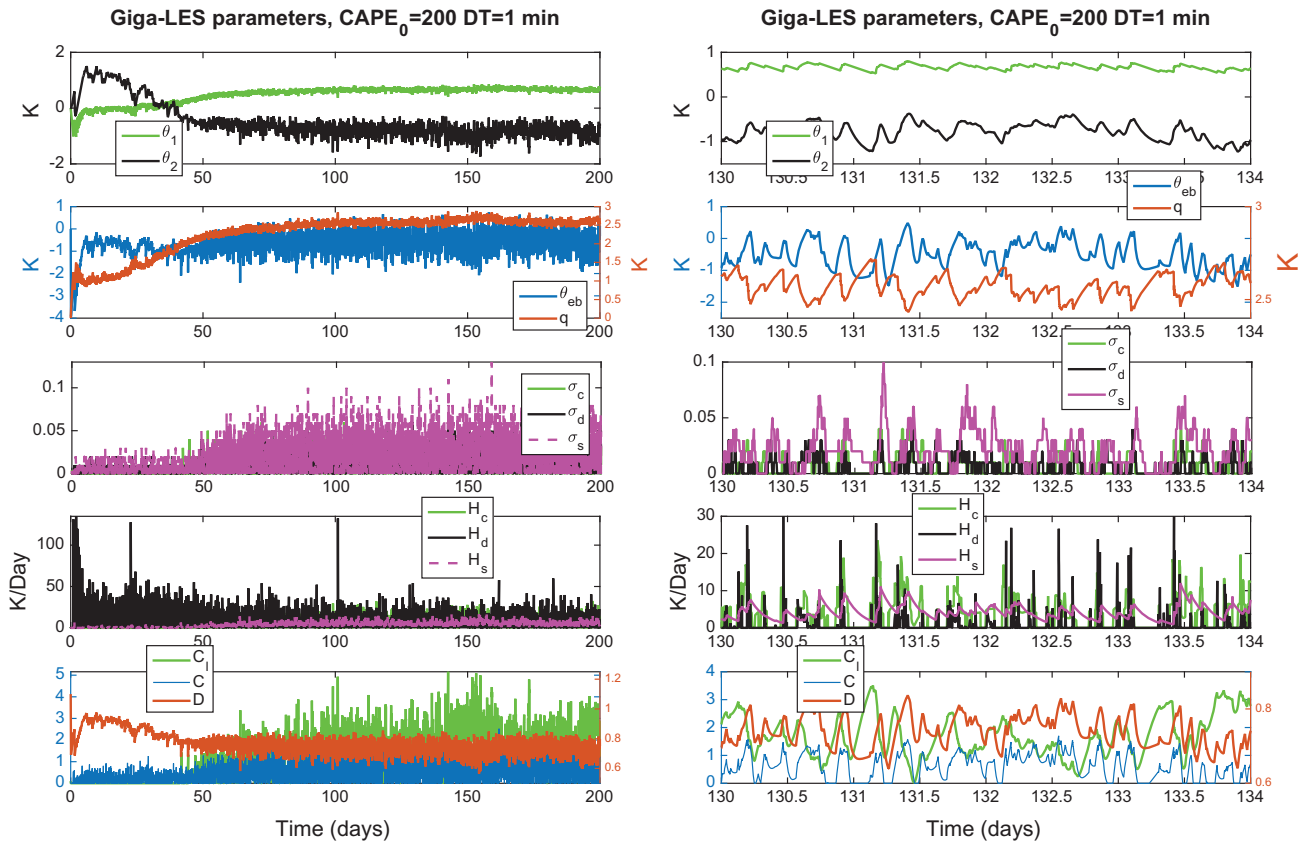
The overbars indicate the radiative-convective equilibrium (RCE) values, while the primes indicate the deviation from the RCE. The subscripts b and m correspond to atmospheric boundary layer (ABL) and mid-troposphere values, respectively

regime, as reported in Table 5. We note that the 4×4 partition parameters, corresponding to a GCM resolution of 50 km, are used mainly because they have an overall smaller variance, according to Table 3.

As in KBM10, the deterministic ODEs in (9) are integrated with a third order Adams-Bashforth and the stochastic SMCM is simulated by the exact algorithm of Gillespie (1975). The integration is carried over a period of 200 days to allow convergence toward a statistical radiative convective equilibrium.

In Fig. 11, we plot the solution times series of the prognostic, climate model variables $(\theta_1, \theta_2, \theta_{eb}, q)$ and the heating rates H_c, H_d, H_s as well as the simulated cloud area fractions time series and the associated large scale predictors, CAPE, CAPE_l and dryness. After a transient period of about 50 days, the solution enters its statistical steady state. The times series in Fig. 11a exhibit some interesting chaotic dynamics, reminiscent of tropical convection. From the closeup plots in Fig. 11b, we can see some coherence between the oscillations in large scale variables and the fluctuations of the area fractions suggesting some kind of resonance between the stochastic and the deterministic models as observed in KBM10. More importantly, although chaotic, the area fractions display a clear pattern characterized by clear sky periods followed by congestus activity which in turns triggers deep convection followed by extended stratiform events.

It is worthwhile noting that while the Giga-LES inference was conducted with a value of $\text{CAPE}_0 = 1500$ J/kg, the simulations in Fig. 11 are obtained with $\text{CAPE}_0 = 200$ J/kg. The simulation with $\text{CAPE}_0 = 1500$ J/kg is numerically unstable, probably because of the extra-stiffness



(a) Full simulation. **(b)** Zoom on a 4 day period.

Fig. 11 Simulation of the single column GCM in (9) using an integration time step of 1 min, and the values $\mathrm{CAPE}_0 = 200\,\mathrm{J}$, $T_0 = 10\,\mathrm{K}$. The closure equations, parameter et constant values used for the simulation are reported in Table 4. *Left panels* (**a**) show the full simulation time series and the *right panels* (**b**) display a zoom in on a short 4 day period

Table 5 Cloud timescale parameter values used in Khouider et al. (2010) (KBM10, 2 cases), Frenkel et al. (2012) (FMK12), and Peters et al. (2013) (P2013; from Darwin dataset using scaled CAPE)

	G-LES	KBM10	FMK12	P2013
Parameter	Transition time (h)			
τ_{01} (Formation of congestus)	31.789	1, 3	1	1
τ_{10} (Decay of congestus)	1.761	5, 2	1	1
τ_{12} (Conversion of congestus to deep)	0.238	1, 2	1	3
τ_{02} (Formation of deep)	11.821	2, 5	3	4
τ_{23} (Conversion of deep to stratiform)	0.257	3, 0.5	3	0.13
τ_{20} (Decay of deep)	9.552	5, 5	3	5
τ_{30} (Decay of stratiform)	1.021	5, 24	5	5

The mean estimated values obtained using the 4 by 4 partition are reported as "G-LES"

induced by the large CAPE_0 value. While at first glance this seems inconsistent, it is important to note that firstly, from Fig. 11b, the simulated cloud area fractions are in rough agreement with the Giga-LES time series, both qualitatively and quantitatively, and secondly, the CAPE approximation used by the simple single column model (9) is very distinct from the accurate calculation performed for the Giga-LES inference, based on the parcel lifting method. One would expect some discrepancy in performance by some parameters depending whether one is using one CAPE calculation or another. The (large scale) dynamical models are also very different.

The large scale variables exhibit periods of CAPE and moisture buildup during the clear sky and congestus activity periods, respectively, as in the Giga-LES simulation of Figs. 4 and 6. CAPE (and low level CAPE) is consumed quickly during the convective active phase sending the system to its next suppressed (clear sky) phase. The suppressed periods last between 6 and 12 h, as seen around times 131 days and around 133 days, for example. Beside this intermittency, the simulated dynamics occurs at timescales of a few hours, comparable to that of the Giga-LES time series in Fig. 4. Also, the cloud area fractions fluctuate within ranges of values around a few percents, with the

stratiform one dominating by almost 10 %, consistent the Giga-LES time series in Fig. 6.

As can be seen, from the time series of q and dryness, D, on the second and forth panels, respectively, the moisture variations are weak when compared to those of θ_{eb} and CAPE. However, a close look at the second panel, shows that, although small, the variations in q are consistent with our physical intuition as it shows clear moistening trends during the suppressed phase, around 131 days and 132.5 days, for example.

6 Concluding discussion

In this study we applied the rigorous Bayesian inference method developed in DKM14 to calibrate the SMCM (KBM10). More precisely we learn the values of some of its most sensitive parameters, namely the timescales, τ_{jk}, from the data. These timescale parameters modulate the rates at which a convective cell of the SMCM stochastic Markov lattice switches from one cloud state (congestus, deep, stratiform, or cloud free) to another. In the coarse-grained version of the model, the lattice cloud coverage is governed by a birth-death process whose equilibrium probability distribution is largely determined by the timescales τ_{jk}. Several studies (FMK12, KBM10, De La Chevrotière and Khouider (2015, in preparation) using the SMCM coupled with an idealized GCM show that the dynamics of the lattice cloud area fractions and the associated large-scale flow circulation is extremely sensitive to the choice of the cloud transition timescales.

The data used here for the inference is the Giga-LES dataset of Khairoutdinov et al. (2009), a 24-h large-eddy simulation of deep tropical convection over a large horizontal domain of 205×205 km^2. The simulation covers a wide range of scales of motion, from turbulent eddies to mesoscale circulations, in a domain comparable to a typical grid cell size in a GCM. One of the very few studies that applies the LES technique to deep convection in the atmosphere, the Giga-LES captures the multiscale organization of convection and exhibits a tri-modal cloud distribution of deep, middle, and shallow clouds similar to that observed in the tropics (Khairoutdinov et al. 2009). This makes the Giga-LES a suitable dataset for the SMCM cloud timescale parameter estimation problem.

The Bayesian procedure draws inference about the model parameters based on a posterior distribution of the parameters given cloud coverage data, constrained on the large-scale convective state. The posterior distribution is obtained by updating a prior distribution by a model likelihood function. The likelihood function is essentially given as the product of hundreds of thousands of large sparse matrix exponentials, which are approximated using a parallel version of a preconditioning technique known as the Uniformization Method, developed in DKM14.

Because of the limited number of data points (the resolution of 15 min yields a time series of length 96), the Giga-LES numerical domain is partitioned into a 2 by 2 and 4 by 4 grids, and data time series obtained on each grid cell are used to train the Bayesian procedure incrementally following the sequential learning strategy introduced in DKM14. Provided each subdomain is statistically self-similar, this technique can potentially increase the information capacity by a factor equal to the number of grid cells.

The cloud area fraction time series are obtained following a simple scheme in which each vertical column of the horizontal domain is binarized using a zero threshold of the total cloud condensate (cloud water and ice mixing ratio) and compared to reference profiles for the four SMCM cloud/no cloud states. These reference profiles are constructed using the lifted convection level (LCL), freezing level, and level of neutral buoyancy (LNB): e.g. the LCL and LNB are used as estimated cloud base and cloud top of the deep cloud, respectively.

The large-scale convective state is based on three indicators, the convective available energy (CAPE) integrated over the whole and lower troposphere, C and C_l, and the mid-troposphere dryness D. These are calculated from the horizontally averaged temperature fields, averaged over the last 20 h of simulation. The large-scale variables C, C_l, and D are defined in terms of scaling "activation" parameters CAPE_0 and T_0, which are additional parameters to the model.

The sequential Bayesian procedure is trained using the cloud area fractions and large-scale convection indicator time series obtained using both a 2×2 and a 4×4 grids. The high-dimensional posterior distributions are explored using the Markov Chain Monte Carlo standard technique, and posterior marginal densities for each of the seven parameters are plotted as the parameters are sequentially learned from one subdomain to the next. For a given grid, the sequence of densities appear to reach a limiting distribution for all seven parameters. However when the two grids are compared, not all parameter values agree. In fact, the second quartile of τ_{12} and τ_{20} only agree within one quartile deviation (see Fig. 9).

In terms of the mean values, all timescale parameters, except from τ_{01}, seem to decrease as the partitioning is refined (i.e by going from 2 by 2 to 4 by 4), and the degree by which each parameter decreases varies considerably among the τ_{ij}'s. This constitutes an important challenge on how these timescales scale with the GCM grid resolution. FMK12 introduced an ad hoc parameter τ_{grid} that uniformly scales the τ_{ij}'s proportionally to GCM grid coarsening. Although, the procedure seems to be effective overall (FMK12), the present results constitute a serious challenge

to such simple strategies of changing timescale with grid resolution. While a proper scaling method remains to be found, these results suggest that a safe choice of a spatial scale for the parameter inference is the one given by the GCM resolution. The fact that τ_{01} is the one parameter that trends in an opposite direction is consistent with the recent work by Deng et al. (2015), who found that, in an aqua-planet GCM simulation with a warm pool forcing, substantially rising the value of τ_{01}, from 5 to 40 h, was necessary in order to produce acceptable MJO simulations. This is somewhat consistent with the value for τ_{01} on the order of 30 h inferred here.

It has to be noted here that timescales of 30 or 40 h are too much for cloud dynamics, especially for a simulation that lasted only 24 h. However, one has to bear in mind that the τ_{kl}'s modulate the transition rates of the microscopic processes, not the bulk area fractions. The coarse grained rates are actually compounded with the number of sites that are susceptible for the corresponding transition; In the extreme case, where all underlying sites are clear sky, for example, the effective timescale for congestus formation is given by τ_{01}/N, where N is the total number of sites within the coarse cell. Thus, for $N = 10 \times 10$, overall, the effective transition times are on the order of minutes, not hours.

The inferred parameters were tested in the context of a simple single column GCM, with crude vertical resolution, based on the first and second baroclinic modes and a boundary layer approach, coupled to the stochastic multicloud model (KBM10, DKM14). As shown in Fig. 11, the coupled single column GCM-SMCM model produced chaotic dynamics in both the cloud area fractions and the large scale dynamics, consistent with the intermittent dynamics of tropical convection and with the expected coherence between the large scale and the stochastic cloud area fractions. Clear sky periods characterized by CAPE build up are followed by congestus events that serve to moisten the environment and trigger deep convection. Deep convection leads stratiform clouds and together deplete CAPE and send the system to a new suppressed or clear sky period. Both the timescales at which these transitions occur and the range of values between which the area fractions oscillate are consistent with the Giga-LES time series used to infer the transition timescales. Although small, the moisture variations are also consistent with physical intuition, as characterized by moisture building during the suppressed phases and rapid decrease when convection starts.

In De La Chevrotière and Khouider (2015), the SMCM is coupled to a zonally symmetric model to study the Hadley-Monsoon dynamics. The simulations of the meridional mean circulation and waves show complex nonlinear interactions between the stochastic area fractions and the large-scale flow that are sensitive to the choice of the convective timescales. Interestingly, the Giga-LES transition timescales, inferred here, are found to be superior to other ad hoc choices, such as the ones used in KBM10, FMK12 or P2013 listed in Table 5, in terms of reproducing both the right amount of wave variability and displaying a mean meridional flow and heating structure; The Giga-LES parameters yielded results that are more consistent with the monsoon circulation, exhibiting, for example, a clear monsoon trough characterized by a drop in low-level pressure, westerly winds, and cyclonic vorticity in lower troposphere surmounted by positive vorticity, in addition to a single Hadley cell rising in the summer hemisphere and sinking in the winter hemisphere. The success of the Giga-LES parameters in this experiment suggests some consistency of tropical cloud dynamics but this is not enough to claim universality of the SMCM's cloud transition timescales throughout the tropics. Since the GATE experiment used as the Giga-LES mean sounding and forcing took place over the tropical Atlantic, one has to be cautious when drawing conclusions for cloud systems in other geographical locations, for instance within the MJO envelop. Further similar studies using other tropical field experiments such as the Tropical Ocean Global Atmosphere–Coupled Ocean Atmosphere Response Experiment (Webster and Lukas 1992) or Dynamics of the Madden–Julian Oscillation (Yoneyama et al. 2013) need to be conducted and compared to the present results.

Acknowledgments This work is part of M.D.'s Ph.D. dissertation. The research of B.K. is supported in part by a Natural Sciences and Engineering Research Council of Canada Discovery Grant and the Indian Institute for Tropical Meteorology National Monsoon Mission initiative. M. D. is partially supported through these Grants as a graduate student fellow. The parallel/high performance computing required for this research was enabled by WestGrid (www.westgrid.ca) and Compute Canada Calcul Canada's (www.computecanada.ca) infrastructures. The authors would like to thank Belaid Moa, Computing Specialist at Compute Canada/Westgrid, for his expertise and assistance, and Michael Waite for providing them with a CAPE calculation algorithm.

References

Arakawa A, Schubert WH (1974) Interaction of a cumulus cloud ensemble with the large-scale environment, part I. J Atmos Sci 31(3):674–701

Buizza R, Milleer M, Palmer T (1999) Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. Q J R Meteorol Soc 125(560):2887–2908

De La Chevrotière M, Khouider B (2015, in preparation) A zonally symmetric model for the summer monsoon circulation with a stochastic multicloud convective parametrization

De La Chevrotière M, Khouider B, Majda AJ (2014) Calibration of the stochastic multicloud model using bayesian inference. SIAM J Sci Comput 36(3):B538–B560

Deng Q, Khouider B, Majda A, Ajayamohan (2015) Effect of stratiform heating on the planetary-scale organization of tropical convection. J Atmos Sci (In press). doi: 10.1175/JAS-D-15-0178.1

Emanuel KA (1994) Atmospheric convection. Oxford University Press, Oxford

Frenkel Y, Majda AJ, Khouider B (2012) Using the stochastic multicloud model to improve tropical convective parameterization: a paradigm example. J Atmos Sci 69(3):1080–1105

Gillespie DT (1975) An exact method for numerically simulating the stochastic coalescence process in a cloud. J Atmos Sci 32(10):1977–1989

Hung MP, Lin JL, Wang W, Kim D, Shinoda T, Weaver SJ (2013) MJO and convectively coupled equatorial waves simulated by CMIP5 climate models. J Clim 26:6185–6214

Johnson R, Ciesielski PE (2013) Structure and properties of Madden–Julian oscillations deduced from DYNAMO sounding arrays. J Atmos Sci. 70:3157–3179. doi: 10.1175/JAS-D-13-065.1

Johnson RH, Rickenbach TM, Rutledge SA, Ciesielski PE, Schubert WH (1999) Trimodal characteristics of tropical convection. J Clim 12(8):2397–2418

Katsoulakis MA, Majda AJ, Vlachos DG (2003) Coarse-grained stochastic processes for microscopic lattice systems. Proc Nat Acad Sci 100(3):782–787

Khairoutdinov MF, Krueger SK, Moeng CH, Bogenschutz PA, Randall DA (2009) Large-eddy simulation of maritime deep tropical convection. J Adv Model Earth Syst 1:Art. #15, pp 13

Khouider B (2014) A coarse grained stochastic multi-type particle interacting model for tropical convection: nearest neighbour interactions. Commun Math Sci 12(8):1379–1407

Khouider B, Majda AJ, Katsoulakis MA (2003) Coarse-grained stochastic models for tropical convection and climate. Proc Nat Acad Sci 100(21):11,941–11,946

Khouider B, Biello J, Majda AJ (2010) A stochastic multicloud model for tropical convection. Commun Math Sci 8(1):187–216

Kuo HL (1974) Further studies of the parameterization of the influence of cumulus convection on large-scale flow. J Atmos Sci 31(5):1232–1240

Lin JL, Kiladis GN, Mapes BE, Weickmann KM, Sperber KR, Lin W, Wheeler MC, Schubert SD, Del Genio AD, Donner LJ, Emori S, Gueremy JF, Hourdin F, Rasch PJ, Roeckner E, Scinocca JF (2006) Tropical intraseasonal variability in 14 IPCC AR4 climate models part I: convective signals. J Clim 19:2665–2690

Lin JWB, Neelin JD (2003) Toward stochastic deep convective parameterization in general circulation models. Geophys Res Lett 30(4)

Madden RA, Julian PR (1972) Description of global-scale circulation cells in the tropics with a 40–50 day period. J Atmos Sci 29(6):1109–1123

Manabe S, Smagorinsky J (1967) Simulated climatology of a general circulation model with a hydrologic cycle. Mon Weather Rev 95:769–798. doi: 10.1175/1520-0493(1965)093<0769:SCOAGC>2.3.CO;2

Mapes B, Tulich S, Lin J, Zuidema P (2006) The mesoscale convection life cycle: building block or prototype for large-scale tropical waves? Dyn Atmos Oceans 42(1):3–29

Moncrieff MW (2010) The multiscale organization of moist convection and the intersection of weather and climate. In: Climate dynamics: why does climate vary? pp 3–26

Moncrieff MW, Klinker E (1997) Organized convective systems in the tropical western pacific as a process in general circulation models: a toga coare case-study. Q J R Meteorol Soc 123(540):805–827

Peters K, Jakob C, Davies L, Khouider B, Majda AJ (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

Robert C (2007) The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer, New York

Webster PJ, Lukas R (1992) TOGA COARE: the coupled ocean-atmosphere response experiment. Bull Am Meteorol Soc 73:1377–1416

Yoneyama K, Zhang C, Long C (2013) Tracking pulses of the Madden–Julian oscillation. Bull Am Meteorol Soc 94:1871–1891. doi:10.1175/BAMS-D-12-00157.1