CELLULAR AUTOMATA FOR CLOUDS AND CONVECTION *

DAAN CROMMELIN[1,2]

[1] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands

[2] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, The Netherlands

Abstract. Numerical models of the global atmosphere have spatial resolutions that are much too coarse to resolve clouds and convection processes explicitly. Because these processes play an important role in the atmosphere and climate system, they are included in numerical models by means of simplified representations, so-called parameterizations.

Traditional parameterization schemes for atmospheric convection are deterministic. To overcome the limitations of these deterministic schemes, stochastic parameterizations are being developed. The use of probabilistic cellular automata (PCA) for this application is very new and can provide a way to generate spatial patterns of convection as observed in the atmosphere. It is approached from two directions, both briefly reviewed here. In one approach, convection and other sub-grid-scale processes are represented with deterministic CA. In recent work, this is extended to PCA. In the other approach, convection is represented by means of discrete stochastic processes (finite state Markov chains) on a lattice. In most studies in this direction, there is no direct coupling between neighboring lattice nodes, however recently such couplings are considered as well. To illustrate the concept of parameterization, a frequently used test model (the L96 model) is discussed as well in this chapter.

Parameterization of atmospheric convection and clouds with PCA has several interesting mathematical aspects. One is the interactive (two-way) coupling of the PCA to a partial differential equation for large-scale atmospheric flow. The state of the PCA couples to the time evolution of the flow, and in turn the PCA rules (transition probabilities) depend on the flow state. Furthermore, for convection it is natural to consider N-state PCAs with $N > 2$ rather than a binary ($N = 2$) PCA. Finally, statistical inference can be a fruitful approach to construct the PCA rules or transition probabilities for convection. The PCA dependence on the time-evolving atmospheric flow and the large number of configurations for PCAs with $N > 2$ provide interesting challenges for such inference.

Keywords: Markov chains; Stochastic parameterization; Atmospheric convection; Statistical inference

1. Introduction

The representation of clouds and convection processes in numerical models of climate and atmosphere is of great importance. Atmospheric convection is the vertical motion of moist air and is a key element in the transportation of moisture through the atmosphere and in the hydrological cycle of the climate system. If water vapor in rising air condensates, the resulting microscopic water droplets form clouds. Further thermodynamical and physical processes such as evaporation, freezing and precipitation add to the complexity and richness of convection and cloud dynamics. The interaction of clouds with incoming solar radiation and outgoing infrared radiation (e.g. reflection) is important in the context of climate change through the mechanism of the so-called cloud-climate feedback [1].

Despite their importance, the spatial resolutions of numerical models for climate and weather prediction are too coarse to resolve clouds and convection processes explicitly [2, 3]. This is due to computational limitations: current state-of-the-art global (i.e., covering the entire earth) operational weather forecasting models can afford spatial (horizontal) resolutions on the order of 10 km, whereas the atmospheric

*chapter to appear in *Probabilistic Cellular Automata*, ed. P.-Y. Louis, F. R. Nardi (2018)

components of global climate models have even coarser resolutions (50-100 km) because they are used for simulations over much longer time intervals. The consequence is that clouds and convection must be represented in a simplified way in these global numerical models.

In atmosphere-ocean science, such simplified representations are known under the name *parameterizations*. The state of the atmosphere that can be resolved by the global numerical model is given as input to a parametrization module, which returns a contribution from convection to the overall rate of change of the model atmosphere. Let $\Psi(x,y,z,t)$ denote the state of the atmosphere at the geographical location (x,y,z) at time t. Typically, x stands for longitude, y for latitude, and z for elevation above the earth surface. In the most commonly used models, the state Ψ consists of five variables: wind velocities in three directions, temperature, and moisture. For ease of exposition, we assume that the time evolution of Ψ is governed by a nonlinear partial differential equation (PDE) (in practice, there are additional algebraic equations):

$$\frac{\partial}{\partial t}\Psi = \mathcal{N}(\Psi, \nabla\Psi) + R \tag{1.1}$$

This nonlinear PDE is derived from the Navier-Stokes equation. The variable $R(x,y,z,t)$ denotes the contribution from unresolved physical processes such as convection. Thus, it is assumed that the nonlinear differential operator $\mathcal{N}(\Psi,\nabla\Psi)$ accounts for physical and dynamical processes that can be adequately resolved in the global numerical model. As mentioned, the contributions from processes that can not be resolved in the numerical model are collected in R. In the rest of this chapter, we will focus on convection, although in practice other unresolved processes are also parameterized in global models (e.g. atmospheric gravity waves, interactions with the underlying land or ocean surface, ...).

In order to close the system, a model for R is required. Traditionally, parameterizations are set up in a deterministic fashion, so that R is effectively a function of Ψ, without any randomness or uncertainty involved. Furthermore, it is common practice to assume that R is determined by Ψ locally in x and y (but not in z). By this we mean the following: let (x_i,y_j) be the (x,y) coordinates of the node (i,j) of the horizontal grid (or lattice) of the numerical model. We define $R_{i,j}(z,t) := R(x_i,y_j,z,t)$ and similarly for $\Psi_{i,j}(z,t)$. The "locality" assumption means that $R_{i,j} = f(\Psi_{i,j})$, i.e. R at node (i,j) is determined by Ψ at the same node (and at the same time), but not by Ψ or R at other nodes. The assumption does not involve the vertical coordinate z: the full vertical profile of $\Psi_{i,j}$ determines the full vertical profile of $R_{i,j}$. For convection, vertical nonlocal effects can be important.

Traditional convection parameterization schemes are based on physical reasoning and intuition, and they are effectively deterministic mappings $\Psi_{i,j} \mapsto R_{i,j}$ (although they are usually not formulated in such explicit manner). To overcome the limitations of these traditional schemes, stochastic parameterization schemes started to receive a lot of attention in the last 10-15 years. In these schemes, the deterministic mapping from $\Psi_{i,j}$ to $R_{i,j}$ is effectively replaced by a probabilistic one. This reflects the uncertainty about subgrid scale processes that is inevitable in numerical models with finite resolution.

Although much work has been done on developing stochastic convection parameterization schemes in the last 10-15 years, a still outstanding challenge is how such schemes can generate realistic spatial patterns for convection, with appropriate spatial correlations. The "locality" assumption discussed above translates into conditional

independence of R at different grid nodes, e.g. $R_{i,j}|\Psi_{i,j}$ and $R_{i,j+1}|\Psi_{i,j+1}$ are assumed to be uncorrelated. This is a limitation, because convection, although it is a physical process at small spatial scales, can organize into larger-scale structures (sometimes dubbed meso-scale structures), with clusters (and clusters-of-clusters) of convective elements spreading out over multiple horizontal grid nodes. Such structures are difficult to capture with parameterization schemes operating under the locality assumption [4]. Cellular automata (CA) can provide a way to generate these spatial patterns.

In this chapter we discuss the relevance and prospects of representing clouds and convection processes in atmosphere models using probabilistic cellular automata (PCA). The use of PCA for this application is very new, and is approached from two different angles. In one line of research, discussed in section 2, convection and other subgrid-scale processes are represented with the help of deterministic CA. In recent work this is extended to include PCA. In the other approach, reviewed in section 3, convection is represented by means of discrete stochastic processes (finite state Markov chains) on a horizontal lattice. In most studies in this direction, there is no direct coupling between neighboring lattice nodes, however recently such couplings are considered as well, see section 5. To clarify these ideas, in section 4 a test model is presented that is often used for designing and testing new methods for subgrid modeling. Furthermore, in section 6 it is discussed how statistical inference can contribute to determine the rules (cell transition probabilities) of a PCA for convection.

2. Convection parameterization with Cellular Automata

The proposal to use cellular automata (CA) for parameterizing the feedback from unresolved scales in numerical models of the atmosphere goes back at least to the late 1990s [6, 5]. The idea was taken up for the purpose of parameterizing the so-called backscatter of kinetic energy from unresolved scales [8, 9, 7]. A CA is used to generate dynamically evolving spatial patterns that determine patterns of kinetic energy input from unresolved scales. More specifically, if R in (1.1) is a kinetic energy source term, it is modeled as $R_{i,j}(z,t) = K(\Psi_{i,j}(z,t))S_{i,j}(t)$, i.e. as the product of an appropriate function K of Ψ and a time-evolving pattern S generated by a CA (see e.g. [7]).

The CA in these studies is a deterministic, synchronous CA, with a layer of memory (or history) added to it: a cell that "comes to life" has multiple lives $L_{\max}$ (in the abovementioned studies, $L_{\max} = 32$). Each time a cell does not meet the rules for survival, it loses one of its lives. Only neighboring cells that have the maximum amount of lives are relevant for determining whether a cell comes to life or survives (see [8] for more details). To be more precise, let us denote by $L_{i,j}(t)$ the number of lives of a cell with lattice index (i,j) at time t, so $0 \leq L_{i,j} \leq L_{\max}$. If $L_{i,j} = L_{\max}$, cell (i,j) is called "fertile". Let $M_{i,j}(t)$ be the number of fertile cells in the Moore neighborhood of (i,j), excluding (i,j) itself. Clearly, $0 \leq M_{i,j} \leq 8$. If $L_{i,j}(t) = 0$ and $M_{i,j}(t) \in \{2,3\}$, then $L_{i,j}(t+1) = L_{\max}$ ("birth"). If $L_{i,j}(t) > 0$ and $M_{i,j}(t) \in \{3,4,5\}$ then $L_{i,j}(t+1) = L_{i,j}(t)$ ("survival"). In all other cases, $L_{i,j}(t+1) = \max(L_{i,j}(t) - 1, 0)$ ("death").

The "raw" CA state or pattern at time t is determined by the pattern of the $L_{i,j}(t)$. To arrive at the pattern $S(t)$, the raw CA pattern is spatially smoothed. Note that the rules of this CA are deterministic. The rules were chosen heuristically, not inferred or derived in a rigorous way. Also, the CA evolves independently of the large-scale atmospheric state, i.e. there is no coupling of Ψ back to S.

In several studies (e.g. [12, 10, 11]), the use of CA for convection parameterization is explored, with a set-up quite similar to the kinetic energy backscatter CA schemes

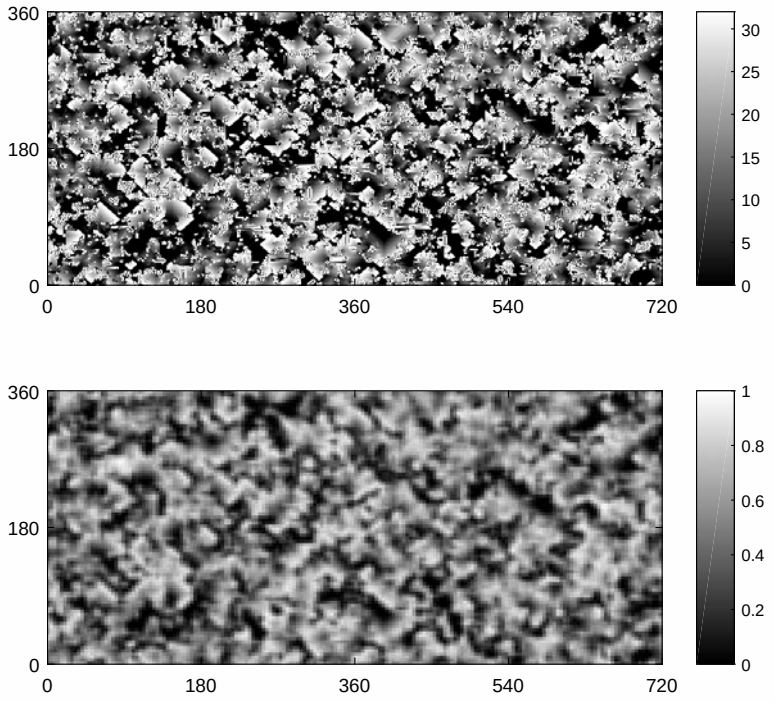

FIG. 2.1. *Example of a pattern generated by the deterministic, synchronous CA with memory as used in e.g. [7]. The top panel shows the raw pattern (number of lives) in a 720 × 360 CA. The bottom panel shows the pattern after coarse-graining, smoothing and normalizing.*

mentioned above. The CA with memory added (as described above) is the starting point, however in recent papers feedback from Ψ to S has been introduced by making the CA rules also dependent on Ψ [10, 11]. Furthermore, the CA pattern can be advected (transported horizontally) by the large-scale flow determined by Ψ. As already mentioned, the CA rules are deterministic, although elements of randomness are introduced in several of these papers by initializing the CA randomly or by adding randomly located live cells at each time step. A probabilistic version of the CA rules has also been considered [11]: if a cell meets the rule for either birth or survival, it will come to (or remain in) life with a probability smaller than one (with the deterministic rule, the probability equals one). This probability can be fixed, or made dependent on advection (i.e., on Ψ). It is reported [11] that the probabilistic, advection-dependent rule generates patterns that look more like convection than those generated with the deterministic rule.

3. Markov chains on a lattice

The CA for convection parameterization discussed in the previous section were primarily deterministic. Although some recent studies consider probabilistic extensions, the starting point is a deterministic CA. In a different line of research, the problem is approached from almost the opposite perspective: convection is parameterized using discrete stochastic processes (finite state Markov chains) on a lattice, but mostly without direct interaction or coupling between the Markov chains at neighboring lattice nodes. Thus, in this approach the starting point is stochastic and relies

on the locality assumption discussed earlier.

The model contributions $R_{i,j}$ due to convection are functions of the vertical coordinate z, hence they live in an infinite-dimensional space. To make stochastic modeling of $R_{i,j}$ more tractable, in a number of studies this function space is effectively discretized, so that the time evolution of $R_{i,j}$ can be modeled as a finite-state Markov chain (e.g. [17, 13, 14, 15, 16]). In these studies, a small number of convective states or cloud states are chosen, and the Markov chain determines the transition probabilities of switching between the states. To reflect the dependence on $\Psi_{i,j}$, the transition probabilities are conditional on (specific functions of) $\Psi_{i,j}$.

An important quantity for convection parameterization is the so-called *convective area fraction* (CAF). Every horizontal grid node (i,j) of the global numerical model has a surface area associated with it (roughly equal to $\Delta x_i \Delta y_j$, with $\Delta x_i = x_{i+1} - x_i$ and $\Delta y_j = y_{j+1} - y_j$). The CAF is the fraction of this area that is in a convective state. In conventional, deterministic parameterizations, the CAF is fixed (e.g. at 0.03). Many stochastic approaches focus on stochastic modeling of the CAF; the resulting CAF is then used as input to calculate $R_{i,j}$ in much the same way as it is done in deterministic schemes (making use of so-called mass flux parameterization methods).

The convective/cloud states can be defined on a "microscopic" lattice, with many micro-lattice nodes pertaining to a single node of the global model grid (the "macroscopic" lattice) [17, 13, 15, 18]. The CAF for the macro-node (i,j) is given by the fraction of micro-nodes, associated with (i,j), that are in an appropriately convective state. In the simplest form there are two possible states at each node (convective and non-convective); a more complicated set-up involves more than two states. For example, in the multicloud model from [13] there are four states (three cloud states, one of which is strongly convective, and a clear sky state), and in [15] this multicloud model is extended to five states.

To formalize this, let (k,l) be the node index of the micro-grid. For every macro-node (i,j), k and l range from 1 to K and L, respectively. We define by $b(i,j,k,l)$ the state at node (k,l) of the micro-grid associated with macro-node (i,j). This state takes values in a finite set of states, $b(i,j,k,l) \in S := \{c_1, ..., c_N\}$, where $c_1, c_2, ...$ denote convective/cloud states. We denote by $\sigma_n(i,j)$ the area fractions of the various states:

$$\sigma_n(i,j) = \frac{1}{KL} \sum_{k=1}^{K} \sum_{l=1}^{L} \mathbf{1}\{b(i,j,k,l) = c_n\} \tag{3.1}$$

where $\mathbf{1}\{.\}$ is the indicator function. Suppose that c_N is a strongly convective state, the only one that contributes to the CAF. Then we simply have that the CAF for macro-node (i,j) is given by $\sigma_N(i,j)$. A mass flux parameterization scheme then takes $\sigma_N(i,j)$ as input, together with $\Psi_{i,j}$, to determine $R_{i,j}$. In this approach, the only information about the subgrid scale convection processes that enters the global numerical model (i.e., the macro-model) is $\sigma_N(i,j)$. We note that the states $b(i,j,k,l)$ evolve in time, in accordance with the Markov chain transition probabilities that are conditioned on $\Psi_{i,j}$. As a consequence, $\sigma_N(i,j)$ also changes in time.

In an alternative set-up, it is the CAF itself that is modeled with a Markov chain [16]. The CAF is discretized in multiples of 0.01 (including zero), and there is no micro-lattice involved. Furthermore, in [14] there is not even a CAF involved. The $R_{i,j}$ themselves are discretized, using a clustering algorithm, hence the states of the Markov chain correspond to entire functions of the vertical coordinate.

The transition probabilities for the Markov chain can be obtained in various ways. One approach is to rely on physical reasoning, as in [17, 13]. An alternative approach is to make use of available datasets on convection (stemming from high-resolution models, or from observations) to obtain transition probabilities through statistical inference [19, 14, 15, 18, 16].

4. Test case: the L96 model

The Lorenz '96 (L96) model [20] is an idealized model of atmospheric flow. It is used frequently as a testbed for developing new ideas and algorithms for parameterization and predictability, e.g. the Markov chain approach discussed in the previous section [19]. The model consists of a set of coupled nonlinear ordinary differential equations (ODEs), and although it was not derived from a PDE it is commonly interpreted as having spatial extent, describing an atmosphere-like dynamical system on a 1-dimensional lattice of constant latitude. The model ODEs are as follows:

$$\frac{d}{dt}X_i = X_{i-1}(X_{i+1} - X_{i-2}) - X_i + F + R_i\,, \tag{4.1a}$$

$$\frac{d}{dt}Y_{i,k} = \frac{1}{\varepsilon}\big(Y_{i,k+1}(Y_{i,k-1} - Y_{i,k+2}) - Y_{i,k} + h_y X_i\big)\,, \tag{4.1b}$$

$$R_i = \frac{h_x}{K}\sum_{k=1}^{K} Y_{i,k}\,. \tag{4.1c}$$

The variables $X_i(t)$ are interpreted as describing the system at large spatial scales, the $Y_{i,k}(t)$ as variables of small-scale processes. The $i \in \{1,...,I\}$ and $k \in \{1,...,K\}$ are interpreted as 1-dimensional lattice indices (i on a macro-lattice, k on a micro-lattice). The lattice has periodic boundary conditions, so that $X_i = X_{i+I}$, $Y_{i,k} = Y_{i+I,k}$ and $Y_{i,k+K} = Y_{i+1,k}$ (we note that the use of indices here differs somewhat from the convention used in e.g. [20, 19], this is done for the sake of consistency with other sections in this chapter).

The $Y_{i,k}$ evolve on a faster timescale than the X_k. The time scale separation is controlled by the parameter ε: with $\varepsilon \ll 1$ there is large scale separation, with $\varepsilon \approx 1$ there is no scale separation. Other parameters in (4.1) are the coupling strengths h_x and h_y and the forcing F. For further discussion and interpretation of these parameters we refer to [20, 19] and the references therein. In what follows we use $\varepsilon = 0.5$, $h_x = -1$, $h_y = 1$ and $F = 10$, as in [19]. Finally, the total number of variables (X_i and $Y_{i,k}$) is $I + I \times K$, examples of settings are $I = 36, K = 10$ [20] and $I = 18, K = 20$ [19].

The goal of subgrid scale parameterization, in the context of the L96 model, is to simulate the dynamics of X as generated by (4.1) as well as possible without having to simulate Y explicitly (here, X denotes the vector $(X_1,...,X_K)$ and similarly for Y). The analogy with realistic atmosphere models is that in such models it is computationally much too expensive to resolve all relevant small-scale variables (Y), it is only feasible to resolve the large-scale variables (X). For the L96 model, this requires a parameterization of the $R_i(t)$ in terms of the $X_i(t)$. The $R_i(t)$ are the quantities that provide the feedback from the small scales to the large scales, see (4.1). The parameterization (or model) for R together with the ODEs for X in (4.1a) form a system with $2I$ degrees of freedom, a large reduction compared to the $(K+1)I$ degrees of freedom in the full L96 model (4.1).

Modeling R is far from straightforward. The dynamics of Y, and hence of R, is dependent on the state of X, see (4.1b). Also, Y has its own chaotic dynamics and

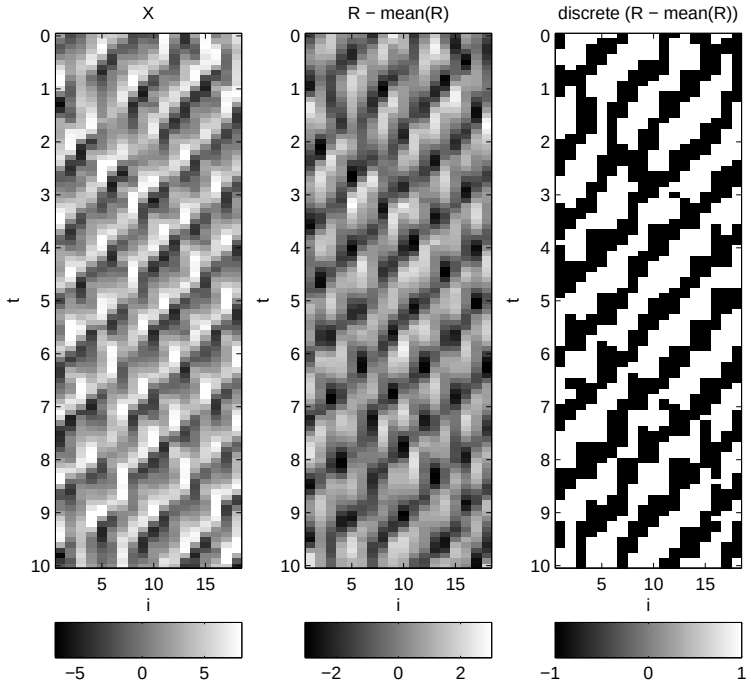

FIG. 4.1. *Example timeseries generated by the full L96 model (4.1) with parameter settings from [19]. The left panel shows $X(t)$, the middle panel shows $R(t)$ with its mean subtracted. In the right panel, $R(t)$-mean(R) is discretized into two states: -1 for negative values, +1 for positive values. In all panels, time runs from top to bottom in increments of 0.1 and the spatial index i is on the horizontal axis.*

is not simply "slaved" to X. In case of large scale separation, i.e. $\varepsilon \ll 1$, asymptotic methods such as averaging and homogenization [21] may be used to derive a reduced model for X, however in realistic atmosphere models there is no clear scale separation between resolved-scale flow and convection.

Figure 4.1 shows an example of data (timeseries) generated by numerical integration of the full L96 model (4.1), i.e. generated by simulating Y as well as X. The parameter settings $(I,K,\varepsilon,h_x,h_y,F) = (18,20,0.5,-1,1,10)$ are those from [19]. The left panel shows the time evolution of the vector $X(t)$, the middle panel that of $R(t)$ with its mean subtracted. A simple 2-state discretization, in which each $R_i(t)$-mean(R) is mapped to either +1 or -1 depending on its sign, is shown in the right panel.

Although the behavior shown in figure 4.1 is chaotic, wave-like structures can be seen to travel through the spatial domain, not only in X but also in R. A parameterization should capture these noisy space-time patterns of R and their dependence on the patterns of X. Under the locality assumption discussed earlier, a stochastic parameterization for R consists of I copies of a scalar stochastic process for R_i conditioned on X_i. In the Markov chain approach, the parameterized R_i can take on only a finite number of values.

5. From Markov chains to PCA

The conditional Markov chain (CMC) lattice models described in the previous

sections do not involve direct interactions between Markov chains at neighboring lattice nodes. However, the CMC states at neighboring nodes are not independent, due to the coupling to $\Psi_{i,j}$. In the case of a microlattice, the chains governing the time evolution of $b(i,j,k,l)$ and $b(i,j,k\pm1,l\pm1)$ are conditioned on the same $\Psi_{i,j}$. If the Markov chains are only defined at the level of the macro-lattice, then $b(i,j)$ and $b(i\pm1,j\pm1)$ are correlated because $\Psi_{i,j}$ and $\Psi_{i\pm1,j\pm1}$ are coupled through the PDE model (1.1).

Notwithstanding these indirect couplings between Markov chains at neighboring lattice nodes, there may be reason to couple them more directly. For example, it was demonstrated [15] that such direct coupling can strongly enhance the variance of the area fractions $\sigma_n(t)$. Recently, a detailed investigation into the limitations of the locality assumption for capturing large-scale coherence in convection modeling was presented [4].

Let $\{k,l\}$ denote the neighborhood of the microlattice node (k,l) (e.g. the Moore neighborhood, or the von Neumann neighborhood). If we generalize the CMC model to include dependencies on the neighborhood, while also retaining the (local) dependence on the macrostate Ψ, we arrive at a model characterized by the following transition probabilities for the cloud states $b(i,j,k,l,t)$:

$$b(i,j,k,l,t+\Delta t)\,|\,b(i,j,k,l,t),\Psi(i,j,t),b(i,j,\{k,l\},t) \tag{5.1}$$

Clearly, this "conditional PCA" is a rich model with many possible scenarios. There are N possible states for $b(i,j,k,l,t)$, so that with the Moore neighborhood there are N^9 different configurations for $b(i,j,k,l,t)$ and $b(i,j,\{k,l\},t)$ together in case of a 2-dimensional lattice. The dependence on $\Psi(i,j,t)$ makes the number of possible configurations even higher. To control these possibilities, it is nearly inevitable to impose certain structures on the model. How to do this is largely an open question. In [15], some ad-hoc choices were made to control the number of parameters that determine the transition probabilities for the conditional PCA. Controlling the parameters in a more systematic way is still a challenge.

The CMC lattice model from [13] was recently generalized to include interactions between neighboring cells on the micro-lattice [22]. The transition probabilities (PCA rules) are designed and motivated from physical intuition, similar to [13]. Energies (or interaction potentials) are assigned to all possible combinations of two neighboring cell states. For a given configuration of the lattice model, the sum of the potentials of all interactions present in that configuration determines a Hamiltonian energy. The transition rates for the individual cells are functions of this Hamiltonian. As the system state (configuration) evolves over time, so do the Hamiltonian and the transition rates. Only nearest neighbors are taken into account (i.e., Moore or von Neumann neighborhood), as it is argued that these are physically the most relevant [22].

6. Statistical inference for PCA

To obtain the rules or transition probabilities of a PCA for clouds and convection, several of the papers mentioned previously rely on physical intuition and heuristics, e.g. [11, 22]. An alternative approach is to infer these rules from available datasets. Such data can come from two sources: observations / measurements of the real physical atmosphere, and numerical simulations with high-resolution models. Regarding the latter, we note that it is possible to do fairly realistic simulations of convection processes (although the detailed physics of e.g. the involved phase-changes (ice - water vapor- liquid water) and ice microphysics are still challenging). However, these

simulations require extremely high model resolution, so that they are restricted in practice to limited spatial domains and short time intervals (e.g., 24 hours on a 100 km by 100 km horizontal domain). It will be many years before such high resolution simulations become feasible for the global atmosphere on climate timescales (years, decades and longer). Hence, the need to parameterize convection will persist for many more years.

With a dataset of sufficient spatial and temporal resolution, a pre-processing step is needed to assign a discrete state to all the lattice nodes at every time step. In previous sections it was discussed how clouds and convection can be modeled by defining a few cloud states (e.g. deep, stratiform, clear sky). The step of classifying the states at the lattice nodes, i.e. of deciding in what cloud or convective state a cell is, is nontrivial. However, we will not discuss it here further as it is primarily a matter of physical insight.

Once the space-time patterns of the discrete states are extracted from the dataset, one can attempt to fit a PCA to these patterns by means of statistical inference of the PCA rules or transition probabilities. There are two aspects to this inference task: selecting the neighborhood, and identifying the rules. In previous sections we have mainly focused on neighborhoods that are one step deep in space and time (e.g., the neighborhood for $R_i(t+\Delta t)$ consisting of $\{R_{i-1}(t), R_i(t), R_{i+1}(t)\}$ in case of a 1-dimensional lattice). Larger neighborhoods, either in space or time, may give better results but can also lead to overfitting, hence selecting the neighborhood is part of the inference problem. Furthermore, inferring the PCA rules with a given neighborhood is equally nontrivial.

Various methods have been developed for neighborhood selection and rule identification, see e.g. [23, 24, 25, 26]. The focus in these studies is mostly on binary systems, i.e. CA with two states. However, for PCA modeling of convection more than two states are typically used, as discussed in previous sections. A method for N-state systems proposed in [27] has not yet been used for convection PCA identification.

A major complication for inferring a PCA for convection (or other subgrid processes) from data is the influence of the large-scale state. As already discussed, Ψ and R in (1.1) are two-way coupled, so the behavior of R is dependent on Ψ. How to infer a PCA for R that is dependent on Ψ, with R and Ψ both evolving in time, is an open question and a challenging research topic. It is assumed here that timeseries data of both R and Ψ are available to infer the PCA. It may be fruitful to consider Ψ as a time-dependent covariate for R, although strictly speaking, Ψ does not evolve independently of R, see (1.1).

In the Markov chain approach discussed in section 3, the dependence on the large-scale state is considered in several papers. In [19], the inference of transition probabilities for R conditional on X from L96 model data is carried out through a straightforward extension of maximum likelihood estimation. This procedure is also used in e.g. [15, 16, 18]. In [28] a Bayesian approach is proposed to estimate parameters of the multicloud model from [13]. It is mentioned in [28] that this approach can be extended to the multicloud model with neighbor interaction as proposed in [22].

7. Summary and conclusion

Modeling of atmospheric convection and clouds is an emerging application for PCA that entails several interesting mathematical challenges. An important aspect is the interactive (two-way) coupling to a PDE for large-scale atmospheric flow, see equation (1.1). The state of the PCA for R couples to the time evolution of the large-scale flow state Ψ through (1.1), and at the same time the PCA rules (transition

probabilities) depend on Ψ. Furthermore, in most studies so far where convection is modeled as a discrete process, more than two states are used, so it is natural to consider N-state PCAs with $N > 2$ rather than a binary ($N = 2$) PCA.

As discussed in sections 2 and 3, PCA modeling for convection emerges from two different research directions. The use of deterministic CA for modeling convection and other subgrid processes has been pursued for more than ten years, see section 2, however the extension of this approach to stochastic modeling (i.e., to PCA) is quite recent (e.g. [11]). Markov chain lattice models for convection, discussed in section 3 have also been studied for a while. These models are stochastic from the outset, but they usually do not include interactions between neighboring cells. Such interactions were added recently [22].

Deriving the PCA rules or transition probabilities from first principles is very challenging for convection. Besides physical intuition and heuristics, statistical inference can be a fruitful approach to construct these rules. A major challenge for inference is the fact that a PCA for convection should be dependent on the large-scale state Ψ. Some work has been done to include this dependence in the Markov chain lattice models, but the generalization to PCA has hardly been explored yet.

In section 4, the L96 model was discussed, an often used idealized model for experimenting with subgrid scale parameterizations. This would be a suitable model for testing and validating new ideas and algorithms to tackle the challenges summarized here.

Acknowledgements DC is financially supported by the Netherlands Organisation for Scientific Research (NWO) through the Vidi project *Stochastic models for unresolved scales in geophysical flows.*

REFERENCES

[1] Stephens, G. L. (2005). Cloud feedbacks in the climate system: A critical review. *Journal of climate*, 18, 237-273.

[2] Arakawa, A. (2004). The cumulus parameterization problem: Past, present, and future. *Journal of Climate*, 17, 2493-2525.

[3] Randall, D., Khairoutdinov, M., Arakawa, A., and Grabowski, W. (2003). Breaking the cloud parameterization deadlock. *Bulletin of the American Meteorological Society*, 84, 1547-1564.

[4] Tan, J., Jakob, C., and Lane, T. P. (2015). The Consequences of a Local Approach in Statistical Models of Convection on its Large-Scale Coherence. *Journal of Geophysical Research: Atmospheres*, 120, 931-944.

[5] Palmer T.N. (2001) A nonlinear dynamical perspective on model error: a proposal for non-local stochastic-dynamic parameterization in weather and climate prediction models. *Q.J.R. Meteorol. Soc.* **127**, 279-304

[6] Palmer, T. N. (1997). On parametrizing scales that are only somewhat smaller than the smallest resolved scales, with application to convection and orography. In *Proceedings of the ECMWF workshop on New insights and approaches to convective parametrization*, 328-337.

[7] Berner J, Doblas-Reyes FJ, Palmer TN, Shutts G, Weisheimer A. (2008) Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Phil. Trans. Roy. Soc. A* **366**: 2561-2579

[8] Shutts, G. (2004). A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems. Technical Memorandum 449, ECMWF.

[9] Shutts, G. (2005) A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.,* 131, 3079-3102

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

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

[12] Berner, J., Shutts, G., and Palmer, T. (2005). Parameterising the multiscale structure of organised convection using a cellular automaton. In *ECMWF Workshop on Representation of Sub-grid Processes Using Stochastic-dynamic Models*, 129-139.

[13] Khouider B, Biello J, Majda AJ. (2010) A stochastic multicloud model for tropical convection. *Commun. Math. Sci***8**, 187-216

[14] Dorrestijn J., Crommelin DT, Siebesma AP, Jonker HJJ. (2013) Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theor. Comput. Fluid Dyn.* **27**, 133-148

[15] Dorrestijn J, Crommelin DT, Biello, JA, Böing, SJ. (2013) A data-driven multicloud model for stochastic parameterization of deep convection. *Phil. Trans. Roy. Soc. A* **371**(1991):20120374

[16] Gottwald, G.A., Peters, K., Davies, L. (2016). A data-driven method for the stochastic parametrisation of subgrid-scale tropical convective area fraction. *Q. J. R.Meteorol. Soc.,* 142, 349-359

[17] Majda, A. J., and Khouider, B. (2002). Stochastic and mesoscopic models for tropical convection. *Proceedings of the National Academy of Sciences*, 99, 1123-1128.

[18] Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., Jonker, H. J. J., and Jakob, C. (2015). Stochastic parameterization of convective area fractions with a multicloud model inferred from observational data. *Journal of the Atmospheric Sciences*, 72, 854-869.

[19] Crommelin, D., and Vanden-Eijnden, E. (2008). Subgrid-scale parameterization with conditional Markov chains. *Journal of the Atmospheric Sciences*, 65, 2661-2675.

[20] Lorenz, E. N. (1996). Predictability - a problem partly solved. In *Proceedings of the 1995 ECMWF seminar on Predictability*, ECMWF, Reading, UK, 118.

[21] Pavliotis, G., and Stuart, A. (2008). *Multiscale methods: averaging and homogenization.* Springer.

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

[23] Richards, F. C., Meyer, T. P., and Packard, N. H. (1990). Extracting cellular automaton rules directly from experimental data. *Physica D: Nonlinear Phenomena*, 45, 189-202.

[24] Adamatzky, A. I. (1994). *Identification of cellular automata.* CRC Press.

[25] Billings, S. A., and Yang, Y. (2003). Identification of probabilistic cellular automata. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 33, 225-236.

[26] Sun, X., Rosin, P. L., and Martin, R. R. (2011). Fast rule identification and neighborhood selection for cellular automata. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41, 749-760.

[27] Guo, Y., Billings, S. A., and Coca, D. (2008). Identification of N-state spatio-temporal dynamical systems using a polynomial model. *International Journal of Bifurcation and Chaos*, 18, 2049-2057.

[28] De La Chevrotiere, M., Khouider, B., and Majda, A. J. (2014). Calibration of the stochastic multicloud model using Bayesian inference. *SIAM Journal on Scientific Computing*, 36, B538-B560.