A machine learning assisted development of a model for the populations of convective and stratiform clouds

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001798

A Machine Learning Assisted Development of a Model for the Populations of Convective and Stratiform Clouds

Samson Hagos[1], Zhe Feng[1], Robert S. Plant[2], and Alain Protat[3]

[1]Atmospheric Sciences and Global Change, Pacific Northwest National Laboratory, Richland, WA, USA, [2]Department of Meteorology, University of Reading, Reading, UK, [3]Australian Bureau of Meteorology, Melbourne, Victoria, Australia

Key Points:

- A new cloud population model for characterizing the interactions between convective and stratiform clouds is developed
- The model is informed by application of machine learning on radar observations and a cloud-permitting model simulation
- The model shows that stratiform clouds act to dampen the variability in the size and number of convective cells and therefore convective mass flux variability

Correspondence to:
S. Hagos,samson.hagos@pnnl.gov

Citation:
Hagos, S., Feng, Z., Plant, R. S., & Protat, A. (2020). A machine learning assisted development of a model for the populations of convective and stratiform clouds. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001798. https://doi.org/10.1029/2019MS001798

Received 28 JUN 2019
Accepted 31 JAN 2020
Accepted article online 07 FEB 2020

Abstract Traditional parameterizations of the interaction between convection and the environment have relied on an assumption that the slowly varying large-scale environment is in statistical equilibrium with a large number of small and short-lived convective clouds. They fail to capture nonequilibrium transitions such as the diurnal cycle and the formation of mesoscale convective systems as well as observed precipitation statistics and extremes. Informed by analysis of radar observations, cloud-permitting model simulation, theory, and machine learning, this work presents a new stochastic cloud population dynamics model for characterizing the interactions between convective and stratiform clouds, with the goal of informing the representation of these interactions in global climate models. Fifteen wet seasons of precipitating cloud observations by a C-band radar at Darwin, Australia are fed into a machine learning algorithm to obtain transition functions that close a set of coupled equations relating large-scale forcing, mass flux, the convective cell size distribution, and the stratiform area. Under realistic large-scale forcing, the derived transition functions show that, on the one hand, interactions with stratiform clouds act to dampen the variability in the size and number of convective cells and therefore in the convective mass flux. On the other, for a given convective area fraction, a larger number of smaller cells is more favorable for the growth of stratiform area than a smaller number of larger cells. The combination of these two factors gives rise to solutions with a few convective cells embedded in a large stratiform area, reminiscent of mesoscale convective systems.

Plain Language Summary The work presents a stochastic cloud population model for convective and stratiform clouds for ultimate application to the development of parameterizations of those clouds in high-resolution regional and global climate models. Machine learning is applied on precipitation radar observations to derive transition functions that represent interactions between convective cells and stratiform area. The model shows that interactions with stratiform clouds limit the variability in the size and number of convective clouds. While the size distribution of the convective cells also influences the size of stratiform area. Specifically, for the same total convective area, many small convective cells are more favorable for the formation of stratiform clouds than fewer larger convective cells. Those interactions coupled with the sensitivity of convective mass flux to cell size lead to solutions with a few cells embedded in a stratiform area as in MCSs and to damped convective mass flux variability.

1. Introduction

Rapid progress in computational resources and numerical methodologies over this decade has led to the rise of operational global weather and experimental climate models with horizontal grid spacing $\leq$ 10 km (e.g., Hólm et al., 2016; Satoh et al., 2014). Despite the obvious advantages, the increased resolution makes several assumptions that typically go into traditional cumulus parameterizations unrealistic. In traditional parameterizations, cumulus convection is assumed to be in statistical equilibrium with the large-scale environment, which is slowly evolving while the cloud population is responding instantaneously and deterministically to any changes in the large-scale forcing with no explicit dependence on its own history or internal variability (Arakawa & Schubert, 1974). Such an assumption requires that a grid box contains many updrafts and therefore to be large compared to both the size of individual clouds and the typical spacing between clouds (e.g., Arakawa & Wu, 2013; Jones & Randall, 2011; Plant & Craig, 2008).

Various efforts are underway to address these issues, and Rio et al. (2019) provide a review of recent and ongoing developments. One strand has been a revival of interest of schemes with multiple cloud types (e.g., Goswami et al., 2017; Khouider et al., 2010) or cloud spectra (e.g., Neggers, 2015; Park, 2014; Plant & Craig, 2008) in attempts to circumvent the limitations of a bulk cloud approach (Plant, 2010). Such schemes require a model for the convective cloud population to represent the number and size/type of clouds. Further discussions on modeling population dynamics for the purpose of convective parameterizations of clouds can be found in Hagos et al. (2018), Khouider (2019), and Plant (2012).

In parallel with these developments in convective parameterization, there is a long and rich history of studies of cloud populations from analyses of radar and satellite observations which have provided insights into cloud-cloud and cloud-environment interactions (e.g., Gehlot & Quaas, 2012; Kumar et al., 2013; Weusthoff & Hauf, 2008). Scanning precipitation radar and geostationary satellites provide near-continuous information on the distribution of cloud sizes from different perspectives (e.g., Koren et al., 2008; Peters et al., 2009; Wood & Field, 2011). However, the direct application of insights gained from observations to the development and evaluation of parameterizations cloud population models is not straightforward partly because the observed variables (often radar reflectivity or brightness temperature, etc.) do not correspond directly to the quantities of interest in parameterizations (e.g., size and strength of updraft in a convective cell).

Recent efforts to bridge the gap include studies by Peters et al. (2017), Cardoso-Bihlo et al (2019), Hagos et al. (2018), for example. The present study may be considered as an extension of Hagos et al. (2018) in which the use of observational radar data was complemented by insights from corresponding convection-permitting simulations. Specifically, a stochastic model of convective cloud populations was developed, with the evolution of convective cell sizes being predicted from the probabilities of growth and decay obtained from analysis of the behavior of convective cell sizes from C-POL radar observations at Darwin. The role of the convection-permitting model simulations was to provide necessary information on the relationship between the cell sizes and the mass fluxes associated with them. It was shown that accounting for cloud-base mass flux to be a non-linear function of convective cell area, and allowing convective plumes to aggregate spatially, can lead to a recharge-discharge behavior under steady forcing.

Cloud-resolving model data have also been used to train deep learning algorithms, with the hope that the trained models can be used as proxies for cumulus parameterizations. Recent work on such an approach has shown some promising results. Rasp et al. (2018) and Gentine et al. (2018) trained a deep neural network and used it as a replacement for traditional subgrid parameterizations in a global general circulation model (GCM). They show that multiyear simulations reproduced the mean climate and variability including precipitation extremes. Similarly, O'Gorman and Dwyer (2018) used an ensemble of decision trees (random forest) as a convection parameterization in a GCM and showed that the GCM runs stably and accurately captures important climate statistics including precipitation extremes under both present and future climates when trained by future climate but not when trained by only present climate.

In this paper, we take a machine-learning approach in order to extend Hagos et al. (2018) by developing a model for the evolution of populations of convective cells and their associated stratiform area. The appropriate coupling of convective and stratiform cloud is an important issue in the parameterization context from several perspectives (e.g., Bechtold et al., 2008; Gerard, 2015; Gross et al., 2018; Storer et al., 2015; Thayer-Calder et al., 2015). For example, the calculations of condensate detrainment from convection, which provides a key source for stratiform cloud development (e.g., Morcrette & Petch, 2010; Tiedtke, 1993; Wilson et al., 2008) are particularly problematic in bulk convective parameterizations (Plant, 2010) and pose issues for the application of equilibrium closure conditions (Yano & Plant, 2019). The extended treatment enables an assessment to be made of the importance and the characteristics of the interactions between convective cells and stratiform area, and our model is constructed with a view towards being able to incorporate those interactions in a natural way within a GCM. The development of this model involves a hybrid of physical arguments and machine learning. Specifically, machine learning algorithms and radar data are used to obtain the transition functions that represent the interactions among the convective cells and between convective cells and the stratiform area in the cloud population model.

In section 2, the observational data and model simulations used are described. In section 3, detailed description of the modeling framework and the machine learning algorithm is presented, and interpretation of the

interactions between convective and stratiform clouds are provided in section 4. Finally, the behavior of the full cloud population model under steady and diurnally varying random large-scale forcing is documented in section 5.

2. Observations and Convection-Permitting Model Simulation

As in Hagos et al. (2018), the radar observations used in this study are obtained from the C-band polarimetric (CPOL) scanning radar located at Darwin, Australia (Kumar et al., 2013). The details of the data processing, which was also used in other previous studies over the Darwin region (e.g., Kumar et al., 2013) and over the tropical Indian Ocean region (e.g., Hagos, Feng, Burleyson, et al., 2014, Hagos, Feng, Landu, et al., 2014), are the same as in Hagos et al. (2018) except that in this study we use 15 wet seasons of CPOL data from 2002 to 2016. While the details of the data processing are discussed in the above-mentioned papers, the key points are briefly summarized in this paper.

The CPOL radar collects a 3-D volume of reflectivity data within a horizontal radius of 150 km at 10 min frequency. Each volume scan consists of a total of 16 sweeps at elevation angles ranging from 0.5° to 42°. As a first step, the sweep data are then gridded to a regular Cartesian grid of $(\Delta X, \Delta Y, \Delta Z) = (2.5, 2.5, 0.5)$ km. The vertical extent of the gridded data is from 0.5 to 20 km. The Steiner et al. (1995) algorithm is applied to the radar reflectivity field at 2.5 km height in order to identify the convective cells from the stratiform area they are embedded in. The convective/stratiform separation parameter settings follow Steiner et al. (1995) as the algorithm was originally designed for the CPOL radar at Darwin. Stratiform area is defined as radar reflectivity at 2.5 km > 10 dBZ that is not designated as convective echoes. Contiguous convective pixels (4-neighbor directly adjacent to the pixels) are grouped as a convective cell (Hagos, Feng, Landu, et al., 2014, 2018). The smallest cells that are considered to be resolved by the gridded CPOL radar have an area of 31.25 km^2. It is important to note that a convective cell defined using radar reflectivity is not completely filled with strong updrafts or downdrafts but is rather a proxy of where strong local vertical mass flux could occur intermittently and sporadically (Yuter & Houze, 1997). A total of 157,032 frames of CPOL volumetric data are used to construct the cloud population statistics. Figure 1 shows a snapshot of radar reflectivity at 2 km height and the convective cells (colors) and stratiform area (green) obtained from the data processing. The resulting data set contains the sizes of all the convective cells in the snapshot and the total stratiform area. For the purpose of this study, the stratiform area is assumed to be a single object for simplicity.

The convection permitting model (CPM) simulation is also described in Hagos et al. (2018). The simulation covers the monsoon period between 1 January 2006 and 28 February 2006. The Weather Research and Forecasting (WRF) model (Skamarock et al., 2008) is run at 2.5 km grid spacing. The domain covers the region between 25–5°S and 120–150°E. The identification of convective cells within the domains of fourteen "virtual radars" was done in the same way as for the observations. For each convective cell, the cloud base mass flux is calculated. Empirical relationship between mass flux and convective cell size, shown in Figure 2 a, is used in this study. The solid lines represent the relationship for the mean mass flux per cell area, and the dashed lines represent the average standard deviation, which is used to introduce a Gaussian stochasticity in the cloud population model. We also tested the effect of a more stringent definition of convective cell using the Powell et al. (2016) radar echo classification method (Figure 2b). Powell et al added additional criteria to identify weak and isolated shallow convective cores commonly observed in tropical oceanic environment. This algorithm uses the same radar reflectivity threshold and peakedness method as Steiner et al. (1995) to identify convective cores but further defines additional "isolated convective cells." Another key difference is that Powell et al. classify regions peripheral to convective cores as either "isolated convective fringe" or "uncertain" depending on the echo object area and intensity, while those peripheral regions are included as part of the convective cells in Steiner et al (1995).

Using the Powell et al. algorithm to separate convective and stratiform echoes results in smaller convective cells, primarily because the weak isolated convective echoes are identified as convective cells in that algorithm. The results shown in this study are primarily from the use of Steiner et al. (1995) algorithm to facilitate direct comparison with the Darwin CPOL radar observations, but the effect of applying Powell et al. algorithm is also investigated and briefly discussed. To that end, for each of the convective cells identified using the reflectivity from the simulation via the processing described above, the cloud-base mass flux per unit cell area was calculated. For the model simulation, we use stronger reflectivity thresholds (45 dBZ

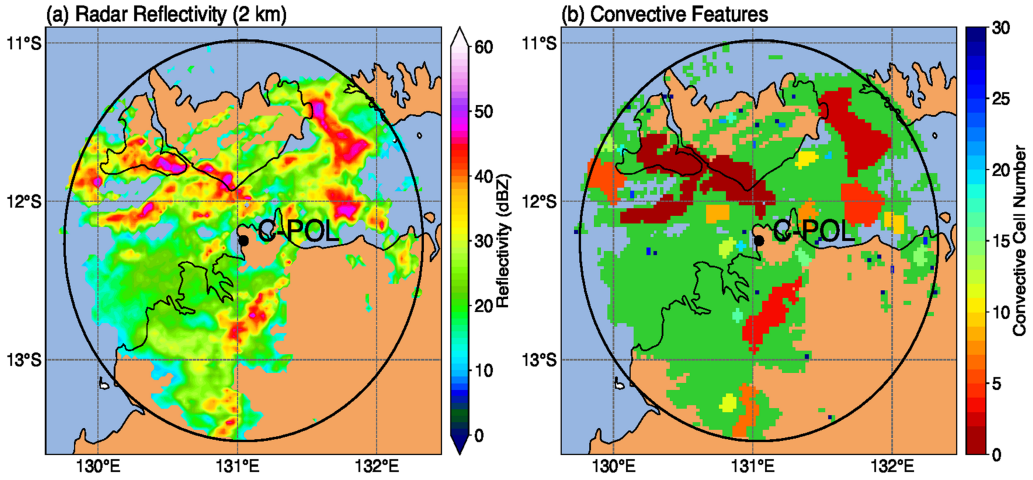

Figure 1. (a) An example radar reflectivity snapshot at 2.5 km height showing the C-Pol radar site at Darwin and (b) the convective cells (colors) embedded in the stratiform area (green).

compared to 40 dBZ for the CPOL radar data) to identify convective cells to account for the often stronger simulated radar reflectivity that could be related to uncertainties in microphysics parameterizations and/or insufficient model resolution (~5–7Δx or ~12.5–17.5 km) to resolve updrafts. The key point to note here is that the convective mass flux per cell area increases with cell size: In other words, larger cells carry more than their share of the mass flux. The empirical nonlinear relationship between the cell cloud-base mass flux M_{Bi} and cell area a_i, the mean size of convective cells in the i-th bin, shown in Figure 2a is given by piecewise linear relationship:

$$\mathrm{m_{bi}} = \lambda + \mu \mathrm{a_i} \tag{1}$$

with intercept λ and slope μ given by 2.23e–10 kg m^{-4} s^{-1} and 0.47 kg m^{-2} s^{-1} for cell size less than 1,000km^2 and 1.9e–11 kg m^{-4} s^{-1} and 0.68 kg m^{-2} s^{-1} otherwise. The solid red line represents this empirical relationship. The dashed lines represent the one standard deviation width, which has average value of 0.226 kg m^{-2} s^{-1}.

Applying the Powell et al. (2016) convective cell identification to the model simulation yields a different cell area-cloud-base mass flux relationship, although the overall trends are similar (Figure 2b). As discussed above, the Powell et al. algorithm identifies more weak and small convective echoes as "isolated convective cells" and excludes peripheral regions of larger convective cores that are otherwise included as part of "convective cells" in the Steiner et al. algorithm. These differences result in generally smaller cells and a much steeper cell area-cloud-base mass flux relationship, where the mass flux per cell area is almost doubled compared to that obtained from the Steiner et al. (1995) algorithm (Figure 2a). The slope flattens at ~500 km^2 cell area in comparison to ~1,000 km^2 for the Steiner et al. algorithm. The impact of different cell area-cloud-base mass flux relationship to the cloud population model will be discussed in section 5.

3. Development of the Cloud Population Model

3.1. Formulation

In order to construct the model, first the state of the cloud population in a given radar snapshot is defined by a vector $\boldsymbol{c}$ for convective cells size distribution at 2 km and the stratiform area by a scalar s. The elements of vector $\boldsymbol{c}$ and c_i represent the sum of the sizes of all the convective cells of size between a_i and a_{i+1}, where $a_i = ia$ for $i \in 1...n$. n is the number of bins set here to $n = 60$. a_1 is the size of the smallest convective cell, defined here as five connected pixels, and the bin spacing a is also set to the same value. Given the 2.5 km grid spacing $a_1 = 31.25$ km^2. The total convective area in the domain is then $A_c = \sum_{i=1}^{n} c_i$. As in Hagos et al. (2018), the evolution of convective area is assumed to be governed by the imbalance between destabilization

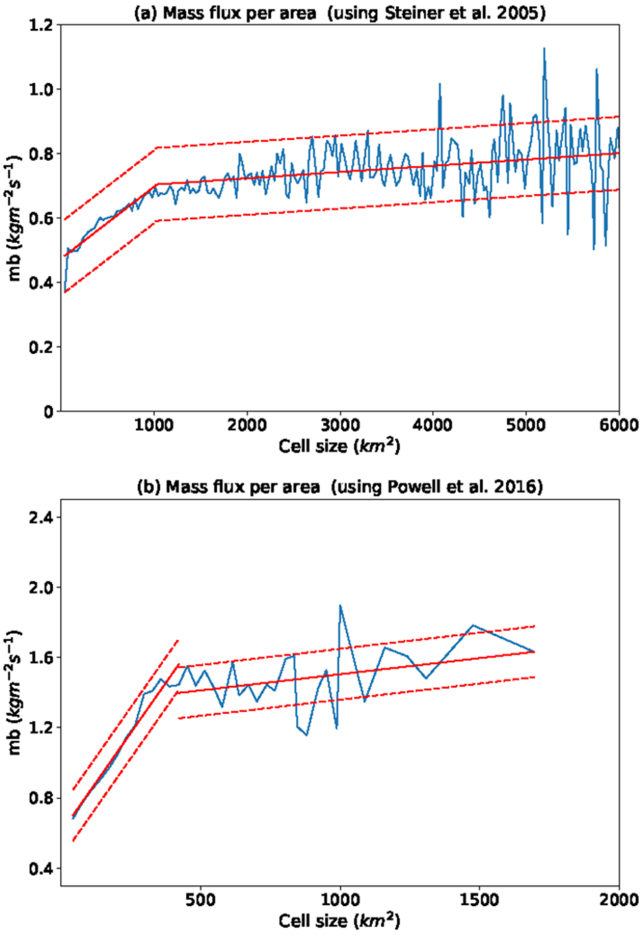

Figure 2. The relationship between cloud-base mass flux per unit area and convective cell size (a) using Steiner et al. (1995) algorithm and (b) using Powell et al. (2016) algorithm to identify convective cells. The red regression lines are used to parameterize the relationship in the cloud population model, and the dashed lines mark the average width of one standard deviation.

by large scale forcing (F) and the stabilizing effect of total cloud base mass flux M_b, with a given constant adjustment time of and mean convective mass flux per unit convective cell area $\overline{m_b}$ at cloud base.

$$\frac{dA_c}{dt} = \frac{1}{\overline{m}_b}\left(-\frac{1}{\tau_c}M_b + F\right) \tag{2}$$

That assumption is inspired by the original potential energy equation of Arakawa and Schubert (1974) and Lord and Arakawa (1980) for an ensemble of convective updrafts. The total mass cloud base flux is the sum of the contributions of all the convective cells and as noted in the last section depends on the size of convective cells.

$$M_b = \sum_{i=0}^{n} c_i m_{bi}(a_i) \tag{3}$$

The dependence of m_{bi} on a_i is approximated by the empirical relationship obtained from the convection-permitting simulation (the red regression line in Figure 2a). For a given cell size value the mean flux per cell area is obtained from the piece-wise linear relationship then that mean and the standard deviation are used to obtain a random value of mass flux per-cell area assuming Gaussian distribution within the cell size bin. Based on these simulations from Hagos et al. (2018), we also set the average mass flux per unit area $\overline{m}_b$ to 0.78 kgm^2s^{-2} and $\tau_c = 4.0$ hours.

The next step is the representation of the evolution of the convective cell size distribution and its relationship to convective area and stratiform area. Here we assume that a change in $\boldsymbol{c}$, given a change in A_c, is some function of the current state. This state is represented by $[s,\boldsymbol{c}]$ that is a vector whose first element is the stratiform area and the other elements are the convective cell size distribution as defined above. Thus,

$$\frac{d\boldsymbol{c}}{dA_c} = f_c([s,\boldsymbol{c}]). \tag{4}$$

The rate of change of stratiform area is also assumed to depend on the current state. To represent the stratiform area, we gained some insights from analysis of observations. Being primarily a result of outflow of hydrometeors from the convection in the tropics, stratiform clouds typically peak several hours after convective clouds. The rate of growth of stratiform area is therefore assumed to be a result of imbalance between formation from the convective cells and the diffusive decay as the hydrometeors mix with the environment or precipitate. The rate of decay is assumed to be proportional to the stratiform area.

$$\frac{ds(t)}{dt} = -\frac{s(t)}{\tau_s} + f_s(\boldsymbol{c}) \tag{5}$$

In order for this set of equations to be closed, the functions f_c and f_s as well as the parameter τ_s have to be specified. These two functions, hereafter referred to as convective and stratiform transition functions, represent the interactions among convective cells as well as with the stratiform area within the domain.

The full system of four equations represents a transition from one state to another through the following steps: (1) a change in A_c may occur due to an imbalance between the current mass flux and the large-scale forcing and will be reflected in a change of the convective cell size distribution (equations (2) and (4)), (2) subsequently a new mass flux is calculated from the new cell size distribution (equation (3)), while in the meantime the stratiform cloud also evolves in response to the convective cell size distribution and its own

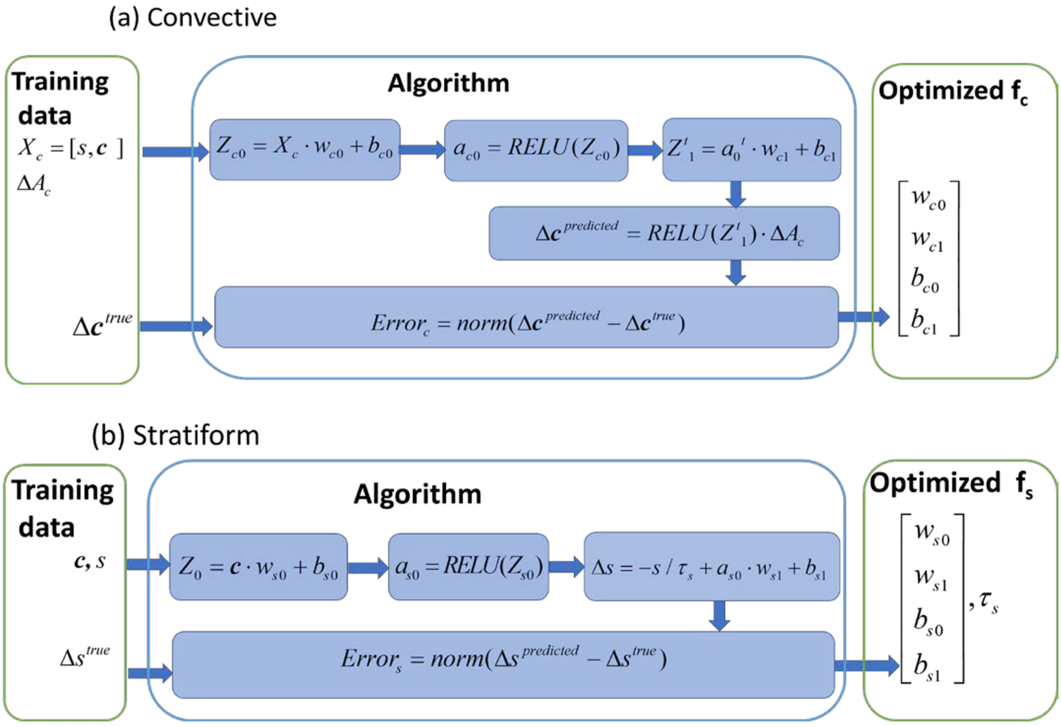

Figure 3. The machine learning algorithm constructed to derive the convective and stratiform transition functions (f_c and f_s). See text for variable definitions.

diffusion (equation (5)). In the next two subsections, the design of the machine learning algorithms used to determine these functions and the parameter τ_s and their validations will be discussed. This machine Learning Assisted Model Population of clouds will be referred to as LAMP hereafter.

3.2. The Machine Learning Algorithm

As discussed above, the radar data provide the convective cell size distribution vector **c** and stratiform area scale s every 10 min, which corresponds to about 157,000 pairs of snapshots from which the observed differences between successive snapshots $\Delta\boldsymbol{c}$ and Δs are calculated. The goal of the machine learning algorithm is to determine f_c, f_s, and the constant τ_s given $\boldsymbol{c}$,s, $\Delta\boldsymbol{c}$, and Δs and thereby to close equations (4) and (5). For both the convective and stratiform transition functions, a single nonlinear hidden layer machine learning algorithm is constructed. The hidden layer constitutes a rectified linear unit (RELU) that sets negative values in the input arrays (Z_{c0} and Z_{s0}) to zero thereby introducing nonlinearity. A graphic depiction of the algorithms is provided in Figure 3. The convective transition function is assumed to completely defined by the pairs of weight arrays (w_{c0} and w_{c1}) and bias vectors (b_{c0} and b_{c1}) and given the constraint that the error ($Error_c$), defined as the sum of the squares of the difference between elements of predicted and true Δc arrays, is minimized. Likewise, the stratiform transition function is defined by (w_{s0} and w_{s1}) and (b_{s0} and b_{s1}) that minimize the $Error_s$ defined as the element wise sum of the square of the difference between the predicted and true Δs.

In constructing a machine learning algorithm, one must balance accuracy with a simplicity that allows for interpretation. In choosing a single-layer model, we favored simplicity. The code is written in TensorFlow, and the Adaptive Moment Estimation (ADAM) optimizer (Kingma and Ba 2014) is used in the minimization. Once the training is complete the weight arrays and bias vectors are saved; f_c and f_s are constructed. As is standard in machine learning applications only half of the data is used for training, and the other half is used for evaluation. The results are rather insensitive to which half is used for training versus evaluation.

3.3. Evaluation

To assess whether the convective and stratiform transition functions derived from the machine learning algorithm correctly represent the actual transitions, at least in a statistical mean sense, we consider their

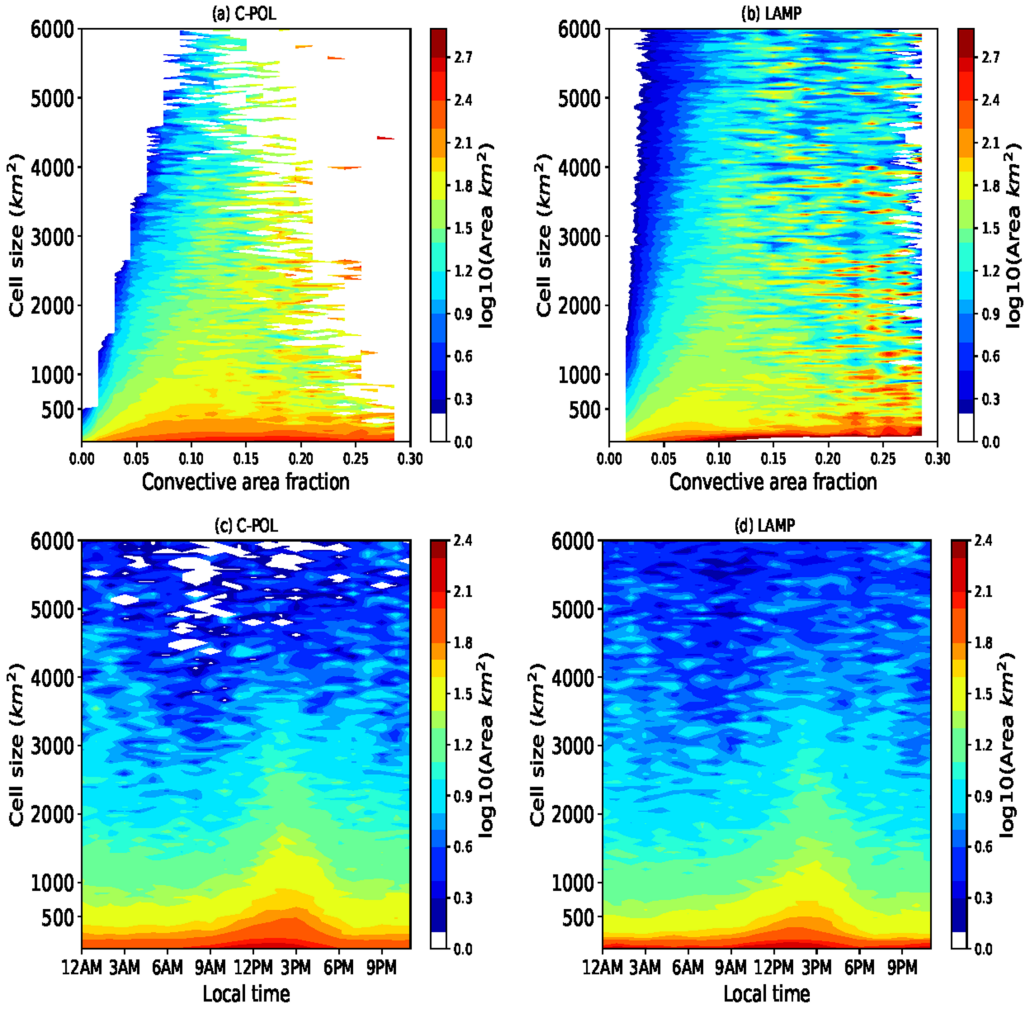

Figure 4. Comparisons of convective cell size distribution (top panels) and their diurnal cycles (bottom panels) calculated using f_c with that from C-Pol observation. The shading represent the logarithm of the total convective area contributed by the given cell size bin in the vertical axis.

representation of the subsequent state given the necessary input from the current state (equation (4)). For f_c, it is to be noted that cell size distribution at the current state was assumed to depend on the cell size distribution, stratiform area, and the change in total convective area from the previous time (10 min prior). Thus, for the second half of the data set, which was not included in the training, the cell size distribution is predicted given those variables. The predicted cell size distribution is compared with what is observed. First convective area fraction is defined as $\frac{A_c}{A_d}$, where A_d is the area of the radar domain of radius 150 km. Twenty-five convective area fraction bins of size 0.015 are constructed for which the mean of convective cell size for the radar frames that fall in those bins is calculated (shaded in Figure 4). Figure 4 shows the comparison between the predicted cell size distributions and that from the observations. Similar analysis is performed for the diurnal cycle of the cell size distribution as well, where the size distributions are partitioned by the hour. Since the cell size distribution is more or less exponential, the logarithm of the cell sizes is reported. Overall, f_c captures the size distribution well. In both the observations and LAMP, the logarithm of the convective area accounted for by a given cell size decreases approximately linearly with the cell size. In both the observations and LAMP the largest mean cell size appears at about 2 p.m. in the afternoon.

A similar analysis is performed on the stratiform transition function of LAMP, f_s. Here, as discussed above, the input variables are the stratiform area and cell size distribution from the previous time step. Figure 5 compares the distribution of the stratiform area and the diurnal cycle predicted by the LAMP to what is

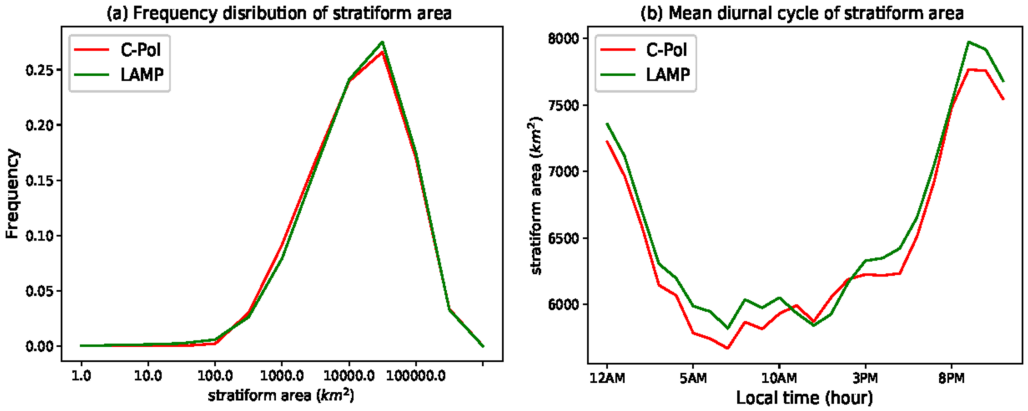

Figure 5. Comparisons of (a) stratiform area distribution and (b) their diurnal cycles calculated using f_s with that from C-Pol observation.

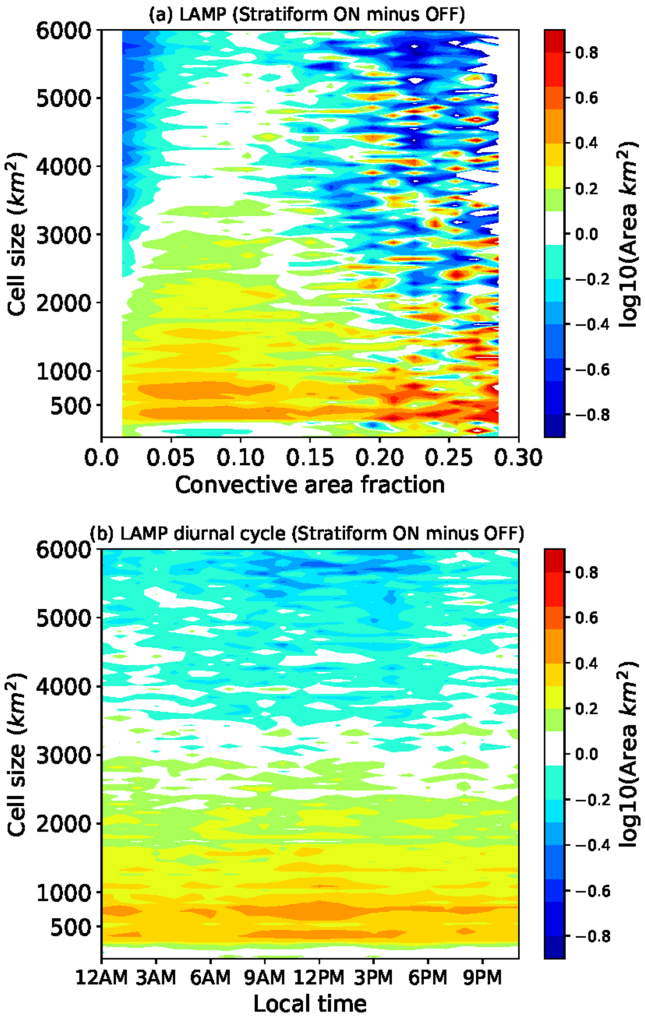

Figure 6. The differences in cell size distributions between stratiform feedback turned on and off as a function of (a) area fraction and (b) time of day.

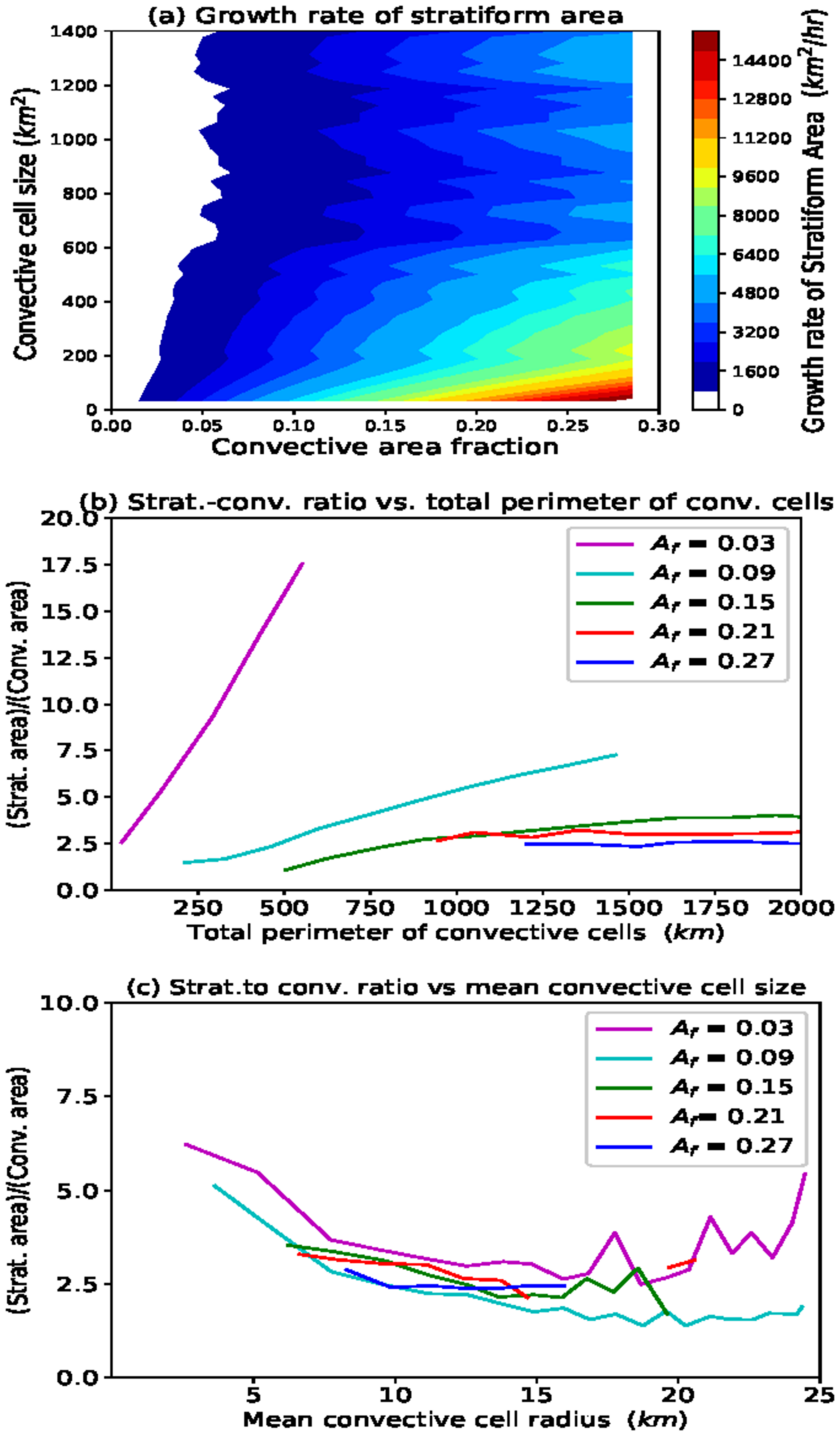

Figure 7. (a) The dependence of growth rate of stratiform area on the convective area fraction and convective cell sizes calculated using f_s from the machine learning algorithm, (b) the relationship of the ratio of stratiform area to total convective area with the sum of the perimeters of the convective cells at five values of convective area fraction from C-POL radar data and (c) same but with mean of the radii of the convective cells.

observed. The agreement suggests that the assumption of treating the stratiform area as a scalar is reasonable and allows the model to capture the evolution. The machine learning algorithm predicts the timescale τ_s to be about 7 hr, which is comparable to the lag produced between the peak convective area at 2 p.m. and peak stratiform area at about 9 p.m. (Figure 4 bottom panels vs. Figure 5).

4. Interaction Between Convective Cells and Stratiform Area

Now that we have shown that the convective and stratiform transition functions f_c and f_s capture the observed transitions well, we take a close look at the interactions they represent. As discussed above, the predicted cell size distribution is assumed to depend on stratiform area through the f_c function, but there are no preconceptions imposed on the form of that function. Plausible interaction mechanisms are based on the physical expectation that stratiform area will introduce a broad area of subsidence, and this in turn could lead to the formation of new small convective cells at the margins of the subsidence area and/or could lead to aggregation of small convective cells into larger and deeper cells (Feng et al., 2015; Rowe and Houze, 2014) via cold pool dynamics.

Now that the convective transition function is obtained, the effect of the presence of the stratiform area can be quantified by repeating the above calculation but setting the stratiform area s to zero and comparing the resulting convective cell size distribution with the actual distribution. Figure 6 shows the difference between the cell size distributions with the effect of stratiform area on (as shown in Figure 5) and that calculated with stratiform effect off (with $s = 0$). Both in the dependence on convective area fraction and the diurnal cycle, the mean cell area of large cells is reduced with the stratiform feedback present: that is, the number of large cells is reduced. As the green shadings indicate, the number of convective cells smaller than 1,500 km^2 increases by about three times ($10^{0.4}$) while that of cells larger than 4,000 km^2 is reduced by about a third ($10^{-0.4}$).

The other leg of the two-way interaction between convective cells and stratiform area is also investigated by examining the dependence of the growth rate of stratiform area on the convective cell size distribution. To that end, we examine the response of f_s to various convective cell size distributions. There are many ways that the convective cell size distribution could add up to a given convective area A_c. For our present purposes, let us assume that the convective area is composed of n number of cells with average size $A_m = A_c/n$. f_s is evaluated for the various mean cell sizes in given area fraction bins. In order to obtain a robust result, the calculation is performed 100 times for each area fraction (average cell size) by randomly selecting cell sizes within each bin. Figure 7a shows the dependence of the growth rate of stratiform area on convective area fraction and mean convective cell size. The first point to note is that the growth rate of stratiform area increases with convective area fraction. This intuitively makes sense given that convective cells are the source of the hydrometeors that constitute the stratiform area. The other more important point is that, for the same convective area fraction, the growth rate of stratiform area is larger when there is a large number of small cells (right bottom corner) in comparison to when there are fewer but larger cells (right top corner). This result can be understood physically if one considers the fact that the hydrometeors are detrained from the convective cells to form stratiform area through the perimeters of the convective cells. Therefore, a large number of small cells implies a large total perimeter through which the hydrometeors detrain from the convective cell. In other words, hydrometeors in large cells are more likely to fall within the cell. This mechanism, referred to as the "particle fountain" model of convection, is extensively discussed in Yuter and Houze (1995).

Such a physical interpretation can be tested through its implications for the dependence of stratiform area on convective area fraction and convective cell size. If indeed the hypothesis holds that stratiform area A_s is related to the perimeter of the convective cells, then it will be related to the number of cells and mean cell size as

$$A_s \sim 2\pi n \left(\frac{A_m}{\pi}\right)^{1/2}. \tag{6}$$

On the other hand, the number of cells is given by

$$n = \frac{A_c}{A_m}. \tag{7}$$

Combining equations (6) and (7) and defining R_c as the mean radius of convective cells, we obtain

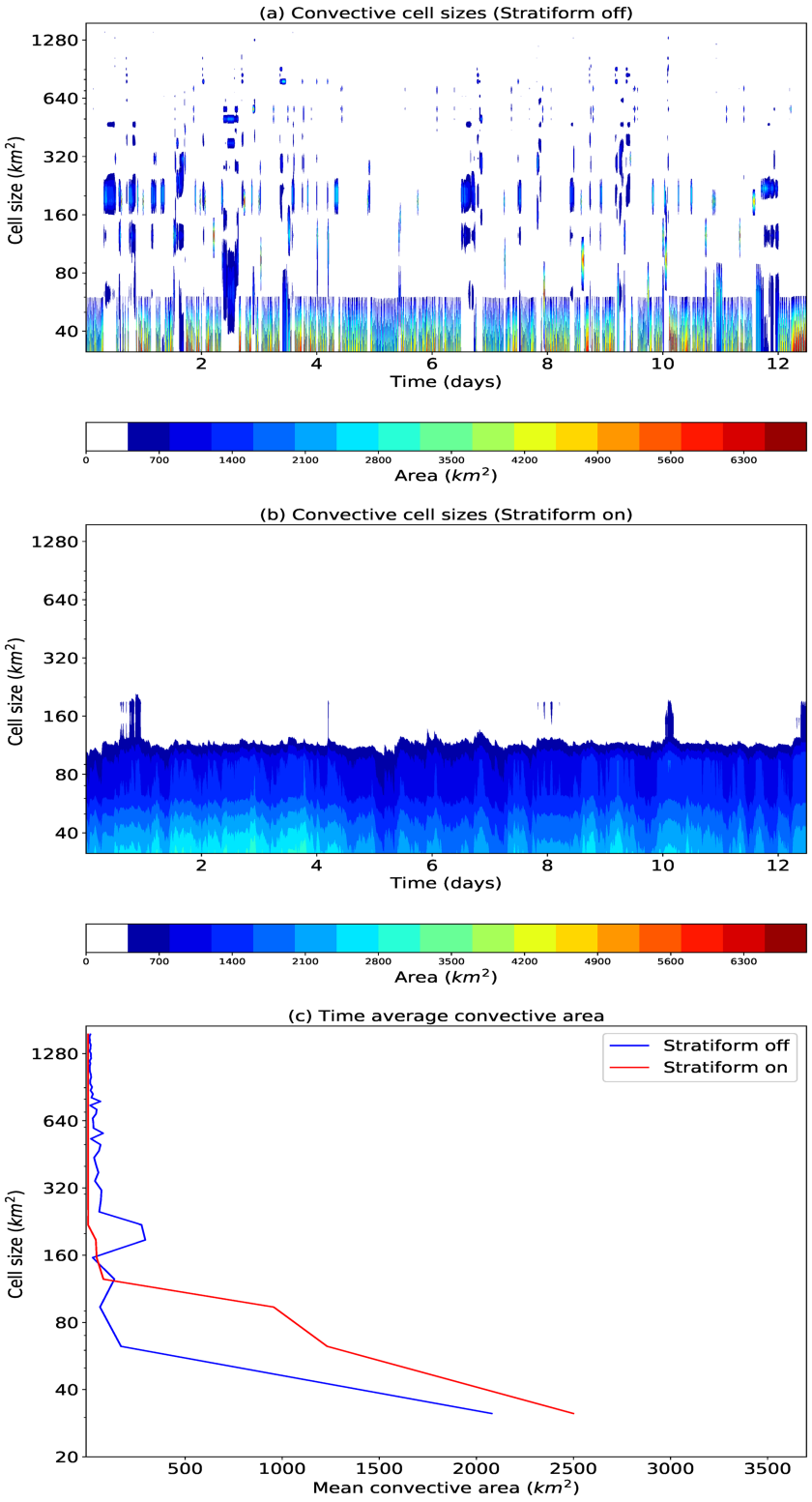

Figure 8. Convective area contributed by cells of sizes given in the vertical axis obtained from the model under steady forcing with stratiform feedback (a) off and (b) on (c) and their time averaged values.

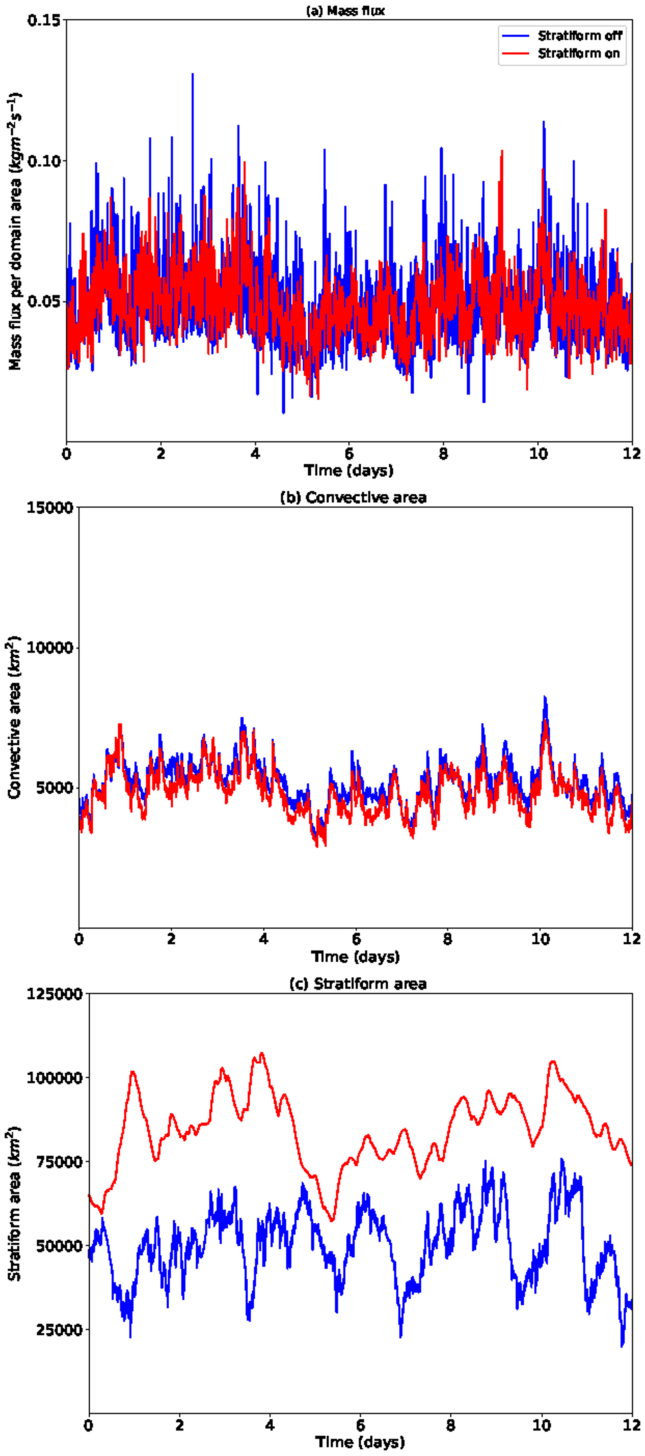

Figure 9. The response of the model to the constant mean random forcing with stratiform feedback off and on (a) mass flux, (b) total convective area, and (c) stratiform area.

$$\frac{A_s}{A_c} \sim \frac{1}{R_c}, \tag{8}$$

which we here define as stratiform to convective area ratio. To verify the relationship between stratiform to convective area ratio and the sum of the perimeters of the convective cells embedded in it and therefore

equation (8), the radar frames are grouped into five bins of mean $A_c = [0.03,0.09,0.15,0.21,0.27]$. The frames in each convective area fraction bin are further partitioned into 15 bins of total perimeter of 150 km increment. The mean of the total perimeter of the cells in each bin and the associated stratiform area are calculated for each bin. Figure 7b shows the relationship for each convective area fraction bin. Confirming our interpretations, the stratiform to convective area ratio increases with the perimeter more or less linearly, with the slope decreasing with larger convective area fraction. The slope of the relationship decreases with convective area fraction as one would expect the perimeter of the convective cells matters less as more cells are in the interior of the stratiform area and contribute less to its growth. A similar analysis is performed for the relationship between stratiform area and mean radius of convective cells. Figure 7c shows that stratiform to convective area ratio decreases with the mean of the convective cell radius for all convective area fraction bins equation (8).

5. The Model Response to Forcing

With the convective and stratiform transition functions, f_c and f_s, derived from observations and the mass flux dependence on convective cell size obtained from the convection-permitting simulations (Figure 2) now both in place, the set of equations that make up the model is closed. Therefore, one can now examine the model's response to external forcing, F. Two illustrative random forcing cases are considered in order to mimic the observed variability in forcing. The first case features a constant mean value meant to represent the low frequency quasi-uniform forcing often associated with oceanic environments. In the second case, a diurnal variability is added to the mean value to mimic a typical forcing associated with a land surface. Details of the forcing and the model behavior are discussed below.

5.1. Constant Mean Forcing

The forcing is defined to be a random fluctuation about a prescribed equilibrium mass flux M_b^{eq} as

$$F(t) = \frac{M_b^{eq}}{\tau_c} X(t), \tag{9}$$

where X(t) is a random variable with exponential distribution of mean value of $1/\lambda$

$$X \sim Exp(\lambda = 1.0). \tag{10}$$

At every time step of 10 minutes, a new value of X is introduced. The equilibrium mass flux is set to be a realistic value of 0.05 $kgm^{-2}s^{-1}$ consistent with the convection-permitting model simulation of Hagos et al. (2018). The random forcing fluctuations are not intended to represent any specific large-scale physical mechanism but rather are included to induce perturbations that then demonstrate how the character of the interactions in f_c and f_s affects the maintenance and nature of the equilibrium cloud distribution.

The model is run with stratiform cloud effects on convective clouds turned on or off (specifically turned off by setting $f_c(c,s) = f_c(c,s = 0)$) for a hundred days in order to obtain a stationary variability. Figure 8 shows the size distributions of convective cells obtained from each run. With stratiform feedback turned off, the cell sizes cover a broad range of values from a large number of very small cells but also with a substantial number of large cells. While with the full model, so that the stratiform feedback is on, the large fluctuations in cell size distribution are damped, and the cell sizes rarely exceed 250 km^2. This is consistent with the result obtained directly from the convective transition function discussed earlier (Figure 6).

This character of the stratiform feedback through the derived f_s has important implications for the variability in mass flux, convective cell size, and the stratiform area itself. Figure 9 shows the time series of mass flux. As may be expected, the mass flux variability is greatly damped as it depends on convective cell size variability. As discussed above (Figure 7), a large number of small cells are favorable for the growth of stratiform area. Therefore, even though the total convective areas are essentially the same with stratiform feedback off and on, the impact of the stratiform feedback to convective cells is to produce a larger stratiform area. Thus, under constant mean forcing, the overall effect of the interaction between convective cells and stratiform area is to produce a large number of small convective cells with a large and persistent stratiform area.

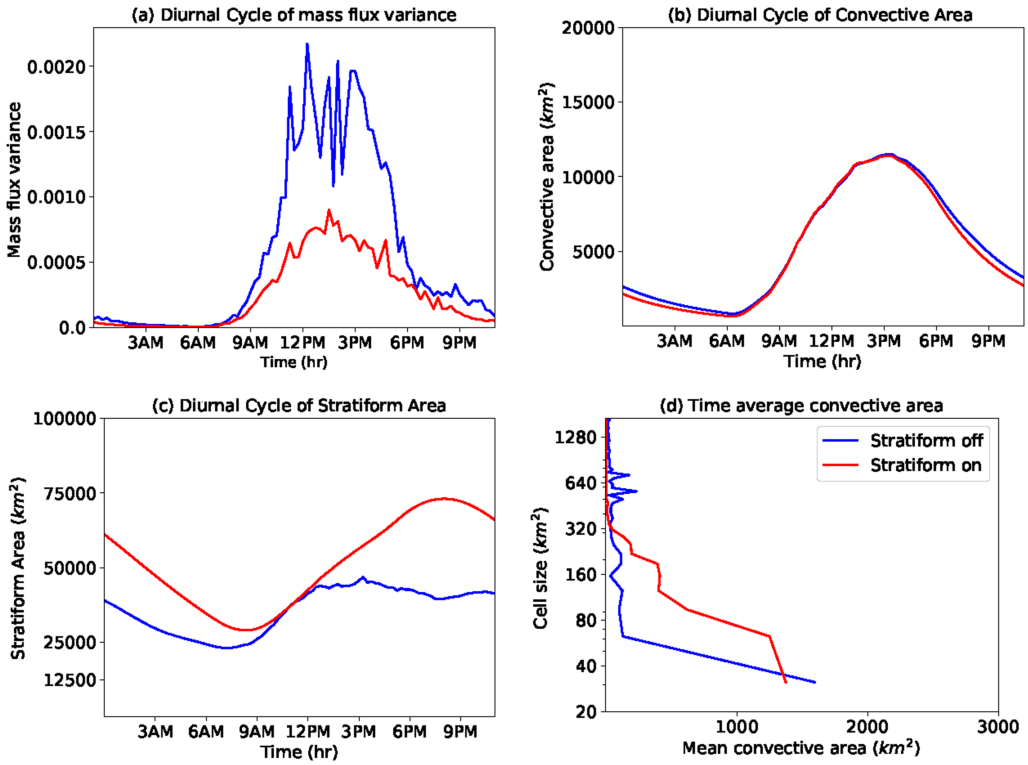

Figure 10. The response of the model to a random forcing of diurnally varying mean value with stratiform feedback off and on (a) mass flux, (b) total convective area, (c) stratiform area, and (d) the mean of convective area contributed by the give convective cell size bin.

5.2. Response to Diurnally Varying Forcing

Finally, the response to a random forcing with diurnally varying mean value is considered. The mean forcing is aimed to mimic the diurnal cycle over land. It is a half sine function with its peak at noon and is zero between 6 p.m. and 6 a.m. local time. Again, the amplitude is modulated by a random factor X as defined above, and the long-term mean mass flux is set to 0.05 kgm^2s^{-1}. Figure 10 shows the responses with stratiform feedback turned on and off. The first point to note is that the observed lag in mass flux is realistically captured (see Figure 10a) in that the peak mass flux is about 3 hours after noon (time of maximum forcing). The strong moderating effect from the stratiform area interactions is readily apparent in the weaker variance in mass flux (Figure 10b), as well as its effects in producing a smaller number of both very large and very small cells while having little impact on the overall convective area (Figures 10c and 10d).

The nonlinear relationship between mass flux and stratiform area are critical components of the interactions between convective and stratiform area obtained from this analysis, and thus, results are sensitive to the representation of these relationships. For example, the above analysis was performed with a steeper cell area-mass flux relationship obtained from the Powell et al. (2016) algorithm (Figure 2b), and the damping effect of mass flux variability is found to be stronger as mass flux, and its sensitivity to cell size is stronger (not shown). Thus, actual implementation of such a model in parameterization might require a more precise empirical relationship, for example, one that is obtained from large domain LES simulations and mass-flux-based definition of convective cells.

6. Conclusion

In order to understand the processes that govern the interactions between convective cells and the associated stratiform area, and thereby accurately represent the frequency distributions of cloudiness and precipitation in climate models, this work presents a model for the population dynamics of clouds. The key assumptions in the model include the following: (i) The growth rate of convective cell area is controlled by the imbalance between large-scale forcing and damping by mass flux; (ii) the growth rate of convective cells is a function of

the cell size distribution and the stratiform area in the current state and the change in convective area fraction; (iii) the growth rate of stratiform area is related to the production by convective cells and the diffusive decay of existing stratiform area; and (iv) the mass flux associated with convective cells is a stochastic nonlinear function of convective cell size. This stochastic nonlinear function is obtained from a convection-permitting model simulation.

The two functions that represent transitions in convective cell sizes and stratiform area are determined by a machine learning algorithm trained by precipitation radar observations. The response of the functions to idealized forcing shows that interactions with stratiform clouds damp the variability in size and number of convective cells. Furthermore, for a given convective area fraction, a larger number of smaller cells is more favorable for the development of the stratiform area than a small number of large cells. These two factors lead to a large stratiform area (i.e., MCS like features) under a steady forcing.

The nature of the transition functions is consistent with forms of behavior that are observed in conditions of convection organization where the appearance of larger cells and the resulting cold pool dynamics can lead to the formation of small cells (Feng et al., 2015), which could potentially aggregate to form larger precipitating cells as their number increases (Rowe and Houze 2014). Moreover, the sensitivity that was found for the stratiform area to the number and sizes of convective cells is consistent with the "particle fountain" conceptual model proposed by Yuter and Houze (1995). That conceptual model implies the stratiform area is related to the sum of the perimeters of the convective cells embedded in it. We show that indeed for a given convective area fraction the stratiform area is approximately linearly related to the sum of the perimeters of the convective cells. This result also implies for the same convective area fraction a large number of small cells is more favorable for stratiform area than a smaller number of large cells.

Our cloud population model has been devised with a view towards implementation in a climate model, as a possible extension of a spectral or multitype cumulus parameterization. The function f_c provides a basis for modifying the determination of a (e.g., Plant & Craig, 2008; Wagner and Graf, 2010; Zhang & McFarlane, 1995) convective population distribution dependent on the stratiform area. In turn, a consistent coupling of the convective population to the stratiform cloud area determination, as expressed through f_s, could be straightforwardly incorporated as a deep-convective source term in a prognostic cloud scheme such as Tiedtke (1993) or PC2 (Wilson et al., 2008). As for any data-driven approach, one of course needs to keep in mind the limitations arising from the scope of the data: here that our cloud population model is constructed based on radar observations and convection-permitting model simulation over the tropics. A parameterization drawing on the cloud population model presented here could be thought of as a hybrid of physically based arguments and of machine learning and therefore promises a potentially acceptable balance between accuracy and interpretability. This study emphasizes in particular the importance of a careful coupling between convective and stratiform clouds for determining the cell size distribution and suggests a route towards accomplishing that. Implementation of a machine learning assisted parameterization based on this cloud population model into an atmospheric model is ongoing, and the results will be reported in future manuscripts.

Acknowledgments
This research is based on work supported by the U.S. Department of Energy Office of Science Biological and Environmental Research as part of the Atmospheric Systems Research (ASR) Program. Computing resources for the model simulations are provided by the National Energy Research Scientific Computing Center (NERSC). Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RLO1830. CPOL radar work has been supported by the ASR program through Grant DE-SC0014063. The convective cloud population data derived from the simulation and C-POL radar observational data are available online (https://portal.nersc.gov/project/cpmmjo/LAMP/). The authors would also like to acknowledge the contribution from Michael Whimpey (Bureau of Meteorology), who has produced the CPOL data set.

References

Arakawa, A., & Schubert, W. H. (1974). Interaction of a cumulus cloud ensemble with the large-scale environment, Part I. *Journal of the Atmospheric Sciences*, *31*, 674–701.

Arakawa, A., & Wu, C. (2013). A unified representation of deep moist convection in numerical modeling of the atmosphere. Part I. *Journal of the Atmospheric Sciences*, *70*, 1977–1992.

Bechtold, P., Köhler, M., Jung, T., Doblas-Reyes, F., Leutbecher, M., Rodwell, J., et al. (2008). Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, *134*, 1337–1351.

Cardoso-Bihlo, E., Khouider, B., Schumacher, C., & De La Chevrotiere, M. (2019). Using radar data to calibrate a stochastic parameterization of organized convection. *Journal of Advances in Modeling Earth Systems*, *11*(6), 1655–1684. https://doi.org/10.1029/2018MS001537

Feng, Z., Hagos, S., Rowe, A. K., Burleyson, C. D., Martini, M. N., & de Szoeke, S. P. (2015). Mechanisms of convective cloud organization by cold pools over tropical warm ocean during the AMIE/DYNAMO field campaign. *Journal of Advances in Modeling Earth Systems*, *7*(2), 357–381. https://doi.org/10.1002/2014MS000384

Gehlot, S., & Quaas, J. (2012). Convection-climate feedbacks in ECHAM5 general circulation model: A Lagrangian trajectory perspective of cirrus cloud life cycle. *Journal of Climate*, *25*, 5241–5259.

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018GL078202

Gerard, L. (2015). Bulk mass-flux perturbation formulation for a unified approach of deep convection at high resolution. *Monthly Weather Review*, *143*, 4038–4063.

Goswami, B., Khouider, B., Phani, R., Mukhopadhyay, P., & Majda, A. J. (2017). Implementation and calibration of a stochastic multicloud convective parameterization in the NCEP climate forecast system (CFSv2). *Journal of Advances in Modeling Earth Systems*, *9*(3), 1721–1739. https://doi.org/10.1002/2017MS001014

Gross, M., Wan, H., Rasch, P. J., Caldwell, P. M., Williamson, D. L., Klocke, D., et al. (2018). Physics–dynamics coupling in weather, climate, and earth system models: Challenges and recent progress. *Monthly Weather Review*, *146*, 3505–3544.

Hagos, S., Feng, Z., Burleyson, C. D., Lim, K.-S. S., Long, C. N., Wu, D., & Thompson, G. (2014). Evaluation of convection-permitting model simulations of cloud populations associated with the Madden-Julian Oscillation using data collected during the AMIE/DYNAMO field campaign. *Journal of Geophysical Research – Atmospheres*, *119*(21), 12,052–12,068. https://doi.org/10.1002/2014JD022143

Hagos, S., Feng, Z., Landu, K., & Long, C. N. (2014). Advection, moistening, and shallow-to-deep convection transitions during the initiation and propagation of Madden-Julian Oscillation. *Journal of Advances in Modeling Earth Systems*, *6*(3), 938–949. https://doi.org/10.1002/2014MS000335

Hagos, S., Feng, Z., Plant, R. S., Houze, R. A., & Xiao, H. (2018). A stochastic framework for modeling the population dynamics of convective clouds. *Journal of Advances in Modeling Earth Systems*, *10*, 448–465. https://doi.org/10.1002/2017MS001214

Hólm, E., Forbes, R., Lang, S., Magnusson, L., & Malardel, S. (2016). New model cycle brings higher resolution. *ECMWF Newsletter*, *147*, 14–19.

Jones, T. R., & Randall, D. A. (2011). Quantifying the limits of convective parameterizations. *Journal of Geophysical Research*, *116*, D08210. https://doi.org/10.1029/2010JD014913

Khouider, B. (2019). *Models for tropical climate dynamics: Waves, clouds and precipitation*. Switzerland: Springer Nature.

Khouider, B., Biello, J., & Majda, A. (2010). A stochastic multicloud model for tropical convection. *Communications in Mathematical Sciences*, *8*, 187–216.

Kingma, D. P. & Ba J. (2014). Adam: A method for stochastic optimization. arXiv:1412.6980v9.

Koren, I., Oreopoulos, L., Feingold, G., Remer, L. A., & Altaratz, O. (2008). How small is a small cloud? *Atmospheric Chemistry and Physics*, *8*, 3855–3864.

Kumar, V. V., Jakob, C., Protat, A., May, P. T., & Davies, L. (2013). The four cumulus cloud modes and their progression during rainfall events: A C-band polarimetric radar perspective. *Journal of Geophysical Research – Atmospheres*, *118*(15), 8375–8389. https://doi.org/10.1002/jgrd.50640

Kumar, V. V., Jakob, C., Protat, A., Williams, C. R., & May, P. T. (2015). Mass-flux characteristics of tropical cumulus clouds from wind profiler observations at Darwin, Australia. *Journal of the Atmospheric Sciences*, *72*, 1837–1855.

Kumar, V. V., Protat, A., May, P. T., Jakob, C., Penide, G., Kumar, S., & Davies, L. (2013). On the effects of large-scale environment and surface types on convective cloud characteristics over Darwin, Australia. *Monthly Weather Review*, *141*, 1358–1374.

Lord, S. J., & Arakawa, A. (1980). Interaction of a cumulus ensemble with the large-scale environment. Part II. *Journal of the Atmospheric Sciences*, *37*, 2677–2692.

Morcrette, C. J., & Petch, J. C. (2010). Analysis of prognostic cloud scheme increments in a climate model. *Quarterly Journal of the Royal Meteorological Society*, *136*, 2061–2073.

Neggers, R. A. J. (2015). Exploring bin-macrophysics models for moist convective transport and clouds. *Journal of Advances in Modeling Earth Systems*, *7*, 2079–2104.

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018MS001351

Park, S. (2014). A unified convection scheme (UNICON). Part I: Formulation. *Journal of the Atmospheric Sciences*, *71*, 3902–3930.

Peters, K., Crueger, T., Jakob, C., & Möbis, B. (2017). Improved MJO-simulation in ECHAM6.3 by coupling a Stochastic Multicloud Model to the convection scheme. *Journal of Advances in Modeling Earth Systems*, *9*(1), 193–219. https://doi.org/10.1002/2016MS000809

Peters, O., Neelin, J. D., & Nesbitt, S. W. (2009). Mesoscale convective systems and critical clusters. *Journal of the Atmospheric Sciences*, *66*, 2913–2924.

Plant, R. S. (2010). A review of the theoretical basis for bulk mass flux convective parameterization. *Atmospheric Chemistry and Physics*, *10*, 3529–3544.

Plant, R. S. (2012). A new modelling framework for statistical cumulus dynamics. *Philosophical Transactions of the Royal Society of London A*, *370*, 1041–1060.

Plant, R. S., & Craig, G. C. (2008). A stochastic parameterization for deep convection based on equilibrium statistics. *Journal of the Atmospheric Sciences*, *65*, 87–105.

Powell, S. W., Houze, R. A. Jr., & Brodzik, S. R. (2016). Rainfall-type categorization of radar echoes using polar coordinate reflectivity data. *Journal of Atmospheric and Oceanic Technology*, *33*(3), 523–538.

Rasp, S., Pritchard, M. S. & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences USA*, *115*, 9684-9689, https://doi.org/10.1073/pnas.1810286115

Rio, C., Del Genio, A. D., & Hourdin, F. (2019). Ongoing breakthroughs in convective parameterization. *Current Climate Change Reports*, *5*(2), 95–111.

Rowe, A. K., & Houze, A. Jr. (2014), Microphysical characteristics of MJO convection over the Indian Ocean during DYNAMO. *Journal of Geophysical Research Atmospheres*, *119*, 2543-2554, https://doi.org/10.1002/2013JO020799

Satoh, M., Tomita, H., Yashiro, H., Miura, H., Kodama, C., Seiki, T., et al. (2014), The non- hydrostatic icosahedral atmospheric model: Description and development *Progress in Earth and Planetary Science*, *1*(18), 1-32.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill D. O., Barker D. M., Duda M. G., et al. (2008). A description of the advanced research WRF Version 3. NCAR Technical Note, NCAR/TN-475 + STR. Boulder, CO: Mesoscale and Microscale Meteorology Division, National Center for Atmospheric Research

Steiner, M., Houze, R. A. Jr., & Yuter, S. E. (1995). Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data. *Journal of Applied Meteorology*, *34*, 1978–2007.

Storer, R. L., Griffin, B. M., Höft, J., Weber, J. K., Raut, E., Larson, V. E., et al. (2015). Parameterizing deep convection using the assumed probability density function method. *Geoscientific Model Development*, *8*, 1–19.

Thayer-Calder, K., Gettelman, A., Craig, C., Goldhaber, S., Bogenschutz, P. A., Chen, C.-C., et al. (2015). A unified parameterization of clouds and turbulence using CLUBB and subcolumns in the Community Atmosphere Model. *Geoscientific Model Development*, *8*, 3801–3821.

Tiedtke, M. (1993). Representation of clouds in large-scale models. *Monthly Weather Review*, *121*, 3040–3061.

Wagner, T.M. & Graf, H. (2010), An ensemble cumulus convection parameterization with explicit cloud treatment. *Journal of the Atmospheric Sciences*, *67*, 3854-3869. https://doi.org/10.1175/2010JAS3485.1

Weusthoff, T., & Hauf, T. (2008). The life cycle of convective-shower cells under post-frontal conditions. *Quarterly Journal of the Royal Meteorological Society*, *134*, 841–857.

Wilson, D. R., Bushell, A. C., Kerr-Munslow, A. M., Price, J. D., & Morcrette, C. J. (2008). A prognostic cloud fraction and condensation scheme. I: Scheme description. *Quarterly Journal of the Royal Meteorological Society*, *134*, 2093–2107.

Wood, R., & Field, P. R. (2011). The distribution of cloud horizontal sizes. *Journal of Climate*, *24*, 4800–4816.

Yano, J., & Plant, R. S. (2019). Why Does Arakawa and Schubert's Convective Quasi-Equilibrium Closure Not Work? Mathematical Analysis and Implications. *Journal of the Atmospheric Sciences*. https://doi.org/10.1175/JAS-D-19-0165.1

Yuter, S. E., & Houze, R. A. (1995). Three-dimensional kinematic and microphysical evolution of Florida cumulonimbus. Part III: Vertical Mass Transport, Mass Divergence, and Synthesis. *Monthly Weather Review*, *123*, 1964–1983.

Yuter, S. E., & Houze, R. A. (1997). Measurements of Raindrop Size Distributions over the Pacific Warm Pool and Implications for Z–R Relations. *Journal of Applied Meteorology*, *36*(7), 847–867.

Zhang, G. J., & McFarlane, N. A. (1995). Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model. *Atmosphere-Ocean*, *33*, 407–446.