PLEASE NOTE

This manuscript

"Data-driven stochastic parametrization for subgrid-scale tropical convective area fraction"

by

Georg A. Gottwald, Karsten Peters and Laura Davies

has been submitted to Quarterly Journal of the Royal Meteorological Society on 9 August 2014 and is currently under review

16 Aug 14

Data-driven stochastic parametrization for subgrid-scale tropical convective area fraction

Georg A. Gottwald [a] * and Karsten Peters [b]* and Laura Davies [c]

[a]*School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia;* [b]*ARC Centre of Excellence for Climate System Science, School of Mathematical Sciences, Monash University, Clayton, VIC 3800, Australia;* [c]*School of Earth Sciences, University of Melbourne,VIC 3010, Australia*

*Correspondence to: E-mail: georg.gottwald@sydney.edu.au;karsten.peters@monash.edu

Observations of tropical convection from precipitation radar and the concurring large-scale atmospheric state at two locations (Darwin and Kwajalein) are used to establish an effective subgrid-scale parametrization for tropical convection. Two approaches are presented which rely on the assumption that tropical convection induces a stationary equilibrium distribution. In the first approach we parametrize convection variables such as convective area fraction or rain rates as an instantaneous random realisation conditioned on the large-scale vertical velocities according to a probability density function estimated from the observations. In the second approach convection variables are generated in a Markov process conditioned on the large-scale vertical velocity, allowing for non-trivial temporal correlations. Despite the different prevalent atmospheric and oceanic regimes at the two locations, with Kwajalein being exposed to a purely oceanic weather regime and Darwin exhibiting land-sea interaction, we establish that the empirical measure for the convective variables conditioned on the large-scale vertical velocities for the two locations are close when the respective vertical velocities are shifted by a constant amount with respect to each other. This allows us to train the stochastic models at one location and then generate time series of convective activity at the other location. The proposed stochastic subgrid-scale parametrizations adequately reproduce the statistics of the observed convective variables. Special attention is put towards capturing both the quasi-deterministic and stochastic behavior of convection for strong and weak forcing, respectively, as well as capturing the observed statistics (mean, variance, skewness). The subgrid-scale parameterisation is formulated for convective area fraction and we discuss how it may be used in future scale-independent mass-flux convection parameterisations. Copyright © 0000 Royal Meteorological Society

Key Words: tropical convection, stochastic parameterization; convective parameterization; climate models; precipitation radar; cloud base mass flux

Received . . .

Citation: . . .

1. Introduction

Despite a remarkable increase in complexity and resolution of general circulation models (GCMs), uncertainty in the understanding and the response of major atmospheric processes to anthropogenic climate change has not been satisfactorily reduced. In particular the representation of deep convection, which ultimately serves to drive the general circulation, is still associated with large uncertainties (Flato *et al.* 2013). Thus, numerical simulations of the Earth's climate are subject to considerable ambiguities.

These become especially apparent when comparing the inter-model mean and spread of hydrological-cycle related variables of the CMIP5 ensemble to observations (e.g. Jiang *et al.* 2012; Tian *et al.* 2013; Lauer and Hamilton 2013). An improved representation of fundamental atmospheric processes, such as convection, is therefore considered to be of utmost priority in the model design (Stevens and Bony 2013).

In GCMs currently used for climate projections, atmospheric convection cannot be explicitly resolved due to its subgrid-scale nature and must thus be parameterised. More than four decades ago, the pioneering works of Ooyama (1964) and Manabe *et al.* (1965) laid the foundations for the development of increasingly complex convective parameterization schemes (see Arakawa (2004) for a review and Randall (2013) for an outlook). As a result of this development, GCMs are now capable of reliably capturing the overall amount of precipitation. However, spatial distributions and variance often compare poorly to observations (e.g. Dai 2006; Pincus *et al.* 2008).

Conventional convective parameterisations tend to be of a deterministic nature and represent only the ensemble mean of the small-scale convective processes. They assume that for any given resolved large-scale state of the atmosphere-ocean system there exists a single possible response at the small-scale convective state feeding back upon the large-scale state. There is, however, a mounting body of evidence that actual observed convection does not obey deterministic relationships between large-scale variables and convective scales (e.g. Peppler and Lamb 1989; Sherwood 1999; Holloway and Neelin 2009; Stechmann and Neelin 2011; Davies *et al.* 2013; Peters *et al.* 2013). Furthermore cloud-resolving models (CRMs) revealed a high degree of variability of small-scale convective activity, contradicting a deterministic relationship between convective activity and large-scale variables (Xu *et al.* 1992; Cohen and Craig 2006; Shutts and Palmer 2007). The complex chaotic dynamics of small-scale processes is widely recognised to give rise to the observed variability. For example, Hohenegger *et al.* (2006), using an ensemble of limited-area convection permitting simulations over the European Alps, identified regions of diabatic forcing (i.e. moist convection) and the associated development of gravity waves as the main source of error growth in their simulations. A lack of variability in the high-frequency, small-scale convective processes can dynamically propagate upscale and cause GCMs to misrepresent low-frequency large-scale variability (Ricciardulli and Garcia 2000; Horinouchi *et al.* 2003). The numerical simulations and observations suggest that a stochastic approach to subgrid-scale parametrizations is needed (Palmer 2001, 2012). The recent increase of resolution of the numerical cores adds to the failure of purely deterministic parametrizations: For example, numerical 200km square grids do not contain sufficient cumulus clouds to allow for the estimation of meaningful averages (Palmer and Williams 2008), and there is a need for a stochastic resolution aware parametrization (Arakawa *et al.* 2011; Arakawa and Wu 2013).

A plethora of stochastic subgrid-scale parametrizations for convection have been developed. Buizza *et al.* (1999) applied random perturbations to the parameterised tendencies in the operational European Centre for Medium-Range Weather Forecasts (ECMWF) forecast system improving its forecast skill. Lin and Neelin (2000, 2003) introduced random perturbations to convective available potential energy (CAPE) and to the heating profile of the host convective scheme improving on the statistics of tropical intraseasonal variability. Bright and Mullen (2002) introduced random perturbations to the trigger function of the Kain and Fritsch (1990) convection scheme, and Teixeira and Reynolds (2008) randomly perturbed tendencies from a deterministic convection scheme by sampling from a normal distribution. Plant and Craig (2008) randomly sampled a distribution of convective plumes to match a required grid-box mean convective mass flux. The required convective mass flux is given by a CAPE closure under the assumption of radiative-convective-equilibrium over the domain. This scheme has been successfully applied to a limited area model-ensemble over central Europe (Groenemeijer and Craig 2012). Berner *et al.* (2005) used ideas from cellular automata to introduce stochastic forcing to the streamfunction to model the effect of mesocale convective systems. Bengtsson *et al.* (2013) developed and tested a stochastic convective parameterisation based on cellular automata via a moisture convergence closure, and showed that in a limited area model-ensemble framework over Scandinavia, the parameterisation leads to a desired increase in spread of the resolved wind field in regions of enhanced deep convection. Majda and Khouider (2002) and Khouider *et al.* (2003) drove a mass-flux convective parametrization with a stochastic model based on convective inhibition (CIN). Khouider *et al.* (2010) developed the stochastic multi-cloud model (SMCM) evolving a cloud population consisting of three cloud types associated with tropical convection (congestus, deep convective and stratiform clouds) by means of a Markovian process conditioned on the atmospheric large-scale state. The SMCM has been shown to adequately simulate tropical convection and associated wave features in a simple two-layer atmospheric model (e.g. Frenkel *et al.* 2012, 2013) and to reproduce observed convective behaviour when observation-based transition time scales between cloud-types are adopted (Peters *et al.* 2013). For a more comprehensive review on current stochastic subgrid-scale parametrizations of convection see Neelin *et al.* (2008) and Palmer and Williams (2010).

Despite providing the desired high-frequency variability stochastic subgrid-scale paramerisations are often difficult to tune and very sensitive to the choice of the parameters as shown for example by Lin and Neelin (2000, 2002, 2003). There has, however, not been much effort in alleviating this difficulty by imposing observational constraints on the parametrization. The limited availability of high-quality, long-term datasets of concurring large-scale and convective scale observations surely contributes to this omission. We list recent works in that direction. Neelin *et al.* (2008) and Stechmann and Neelin (2011) used observed relationships between column integrated water vapour (CWV) and precipitation to inform a physics-based stochastic model to simulate the onset and duration of very strong convection. Their model was able to closely represent observed CWV-precipitation relationships. Dorrestijn *et al.* (2013) used data from large-eddy simulations (LES) to design a data-driven multi-cloud model. The transitions between different cloud types are calculated using Markov chains which are conditioned on large-scale variables. This model reliably reproduces LES data, in particular when spatial couplings are incorporated. More recently Dorrestijn *et al.* (2014) have successfully employed that model on observational

data obtained in Darwin. Horenko (2011) developed a framework which allows for a purely data-based Markov chain parametrization allowing for nonstationary data to model cloud cover.

We complement here the suite of data-driven stochastic models of tropical convection by using observations to build a simple entirely observation-based stochastic model. The parametrization we propose can be built off-line and then subsequently implemented at low computational cost. An entirely observation-based model lacks the transparency of physics-based models, but is potentially more accurate. We exploit available long-term observations of the large-scale atmospheric and the concurring small-scale convective state over Darwin (Davies *et al.* 2013), complemented by a dataset over Kwajalein. The observations are used to inform stochastic models for the convective area fraction and the rain rate by treating them as both uncorrelated random variables and Markov processes, conditioned on the large-scale vertical motion at 500 hPa, ω_{500}. The stochastic parametrization can be constructed at either location and then be applied to observations of large-scale variables from the respective other location. The stochastic models are able to reproduce the observed statistics of the convective activity such as mean, variance and skewness. The underlying premise of our approach is that the stationary stochastic process relating small-scale convective activity and large-scale convergence is sufficiently universal in the sense that the stochastic model can be transferred from one geographical location to another one. Using a Kullback-Leibler information criterion for the conditional probabilities of convective activity as well as quantile regression for the observational data we establish that it is sufficient to correct for the large-scale variables by a simple linear translation to account for the respective ambient atmospheric and oceanic regimes at different locations.

Although most stochastic parametrizations involve CAPE, we follow Davies *et al.* (2013), Peters *et al.* (2013) and Dorrestijn *et al.* (2014) and relate the observed convective state to the observed large-scale vertical motion at 500 hPa, ω_{500}. Dorrestijn *et al.* (2014) find that the mean vertical velocity over Darwin is highly correlated with deep convection starting several hours before the onset of deep convection. This is not surprising as large scale vertical motion in the tropics is directly related to deep convection. Conditioning convective states on vertical motion raises the question of cause-and-effect ambiguities (see e.g. Arakawa 2004; Peters *et al.* 2013, for a discussion). On the one hand, convection induces large scale ascending motion through latent heating, which then facilitates further convection. On the other hand, pre-existing large scale ascending motion (or convergence) facilitates the development of convection (Hohenegger and Stevens 2013; Birch *et al.* 2014) which then further increases large scale ascending motion. We thus argue that tropical convection and large scale ascending motion are intimately linked via a positive feedback loop, limited by the available energy in the atmospheric column and its close environment. The stochastic parametrization we propose does in fact not rely on any cause-and-effect relationship between vertical velocities and convective activity such as CAF, but only utilises observed statistical features and their conditional probabilities.

We use convective area fraction (CAF) (as well as rain rate data) to characterise convective activity (cf. Dorrestijn *et al.* (2013) and Bengtsson *et al.* (2013)). Our motivation to formulate the parametrization with respect to CAF is that it can be used to close convection schemes since CAF was found to scale with domain mean rainfall. Observational and theoretical studies have illustrated, that measures of convective activity such as precipitation are linearly related to the area covered by the precipitation feature (Craig 1996; Nuijens *et al.* 2009; Yano and Plant 2012; Davies *et al.* 2013).

Parameterisations for CAF can be by construction included in the framework of resolution independent parametrizations (Arakawa *et al.* 2011; Arakawa and Wu 2013; Wu and Arakawa 2014). Current mass-flux convection schemes used in operational GCMs assume the area covered by convective updrafts to be negligible compared to the cloud-free part of a model grid box – the so-called assumption of "scale-separation". They are designed to predict changes to the environmental cloud-free air due to convection. This assumption breaks down once the resolution of the GCM becomes high enough such that the area covered by convective updrafts can occupy large parts of or even an entire grid box. Parametrizations for CAF are naturally scalable and could be used to mitigate this problem (Arakawa and Wu 2013; Wu and Arakawa 2014). Furthermore, most currently employed schemes are mass-flux schemes and need to predict the vertical mass flux at cloud base. The mass flux at cloud base could be determined by explicitly assigning an area to the convective updraft together with an updraft velocity. The effect of convection on the environment could be implemented by formulating the dependency of the vertical eddy fluxes of thermodynamic variables on updraft fraction as defined by Arakawa and Wu (2013) and Wu and Arakawa (2014) or through allowing convectively induced subsidence impact on neighbouring grid boxes (Grell and Freitas 2014).

It is pertinent to mention that although using CAF allows for a certain scale-adaptivity, an increase in resolution would prohibit to identify the large-scale environment with the grid box state. In this case spatial averaging over the region of each grid box could be used to define the environment as done in Keane and Plant (2012).

The paper is organised as follows. We introduce the observational datasets along with a comparison of convective behaviour in Darwin and Kwajalein in Section 2. We then use the data to construct the stochastic subgrid-scale convection parameterizations in Section 3. A summary of our results and an outlook to future work are provided in Section 4. In an Appendix we provide a more detailed analysis of the observational data.

2. Data

2.1. Description of the datasets of tropical convection in Kwajalein and Darwin

We utilise two datasets of observations of the large-scale vertical velocity at 500 hPa ω_{500} and of the concurring CAFs and rain rates over tropical locations, averaged to yield 6-hourly time resolution. The datasets each cover a $190 \times 190\,\mathrm{km}^2$ pentagon-shaped area centered over Darwin (Australia) and Kwajalein (Marshall Islands), respectively. The area is chosen as to represent the size of a typical GCM grid-box. The Kwajalein site is located in the tropical western Pacific and is typical for a purely tropical oceanic climate. The Darwin site on the other hand is typical for the monsoon climate of northern Australia and features

a complex topography characteristic of a coastal site. We acknowledge that by having only 6-hourly averaged data available, some characteristics of tropical convection, e.g. the diurnal cycle, are ill-resolved. Nevertheless, because both the small- and the large-scale state are self-consistent (Davies *et al.* 2013), the data are well suited for analysing relationships between them at the scales considered here.

The area-mean values of atmospheric variables are derived using the method of Xie *et al.* (2004), who employ the variational analysis approach of Zhang and Lin (1997), but use profiles of atmospheric variables from numerical weather prediction models instead of atmospheric soundings. Here, the variational analysis employs analyses from ECMWF and is constrained by observations of surface precipitation obtained from C-band polarimetric (CPOL) research radars (Keenan *et al.* 1998) and top-of-the-atmosphere radiation at both locations to reliably balance the column budgets of mass, heat, moisture and momentum. Davies *et al.* (2013) show that constraining the variational analysis by observed rainfall substantially improves the derived large-scale vertical velocities over the Darwin domain compared to using just the ECMWF analysis alone.

Over Darwin, the analysis is applied to three consecutive wet seasons (2004/2005, 2005/2006, 2006/2007), yielding a total of 1890 6-hour means. Over Kwajalein, the analysis is applied to the time period of May 2008 – Jan 2009, produced to fit into the framework of the Year Of Tropical Convection (YOTC, Waliser and Moncrieff 2008; Waliser *et al.* 2012) virtual field campaign. For Kwajalein, 1095 6-hour means are available. At both locations, the large-scale atmospheric data are complemented by data of the concurrent small-scale convective state derived from CPOL radar observations. Among other precipitation related variables, the radar observations provide rain area fractions attributable to either stratiform or convective precipitation, determined after Steiner *et al.* (1995). Here, we use the derived convective precipitation area fraction CAF as proxy for the deep convective cloud fraction. More information regarding the derivation of the datasets can be found in Davies *et al.* (2013).

The data have already provided important new insights into the behavior of tropical convection (Davies *et al.* 2013; Peters *et al.* 2013; Kumar *et al.* 2013). In particular, Peters *et al.* (2013) showed that the convective response to a range of large-scale atmospheric forcing conditions is very similar for both regions despite their distinctly different boundary conditions, *e.g.* land-sea distribution or monsoonal forcing.

2.2. *Analysis of the datasets over Kwajalein and Darwin*

To support our premise that the underlying stochastic process relating the small-scale convective activity to the large-scale variables is sufficiently independent of the geographical location, we contrast here the observed convection at Darwin and Kwajalein. Figure 1 shows CAF observed at Kwajalein and at Darwin as a function of ω_{500}. Figure 2 shows the 2D histograms of CAF and ω_{500} of the observations. The plots show strong qualitative similarities between the two locations which are suggestive of the existence of a universal stochastic subgrid-scale parametrization of CAF conditioned on the large-scale variable ω_{500}.

We remark that although CAPE is frequently used in current convection schemes, it is not well suited as a

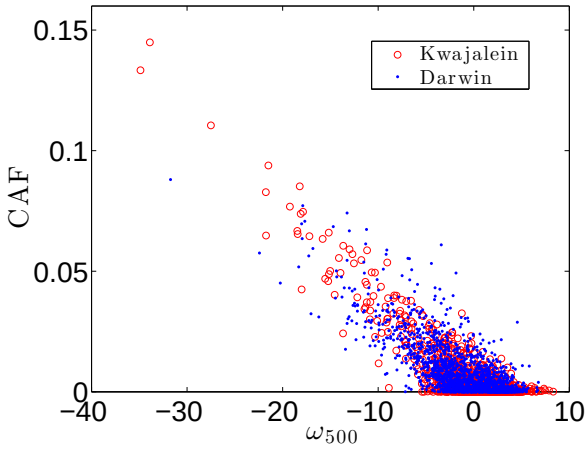

Figure 1. CAF as a function of the vertical velocities ω_{500} [hPa/hour] obtained from observations over Kwajalein (open circles, red) and Darwin (dots, blue).

macroscopic variable on which a stochastic sub-grid scale model can be conditioned on. In Figure 3 it is seen that strong rain events with large CAF do not strongly correlate with a specific range of CAPE (see also Davies *et al.* (2013)). Moreover, the observations of CAF obtained in Darwin and Kwajalein do not exhibit much similarity in their dependency on CAPE, prohibiting a "universal" approach of training the stochastic subgrid-scale model at one location and then applying it to a different location.

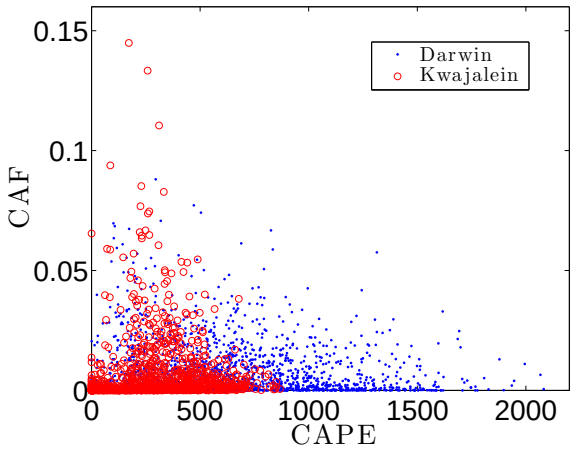

Figure 3. CAF as a function of CAPE [J/kg] obtained from observations over Kwajalein (open circles, red) and Darwin (dots, blue).

Let us briefly discuss some of the particularities of the relationships between CAF and ω_{500} in Kwajalein and Darwin, as seen in Figures 1 and 2. The variance of CAF is dependent on the state ω_{500}. In particular, as already noted in Peters *et al.* (2013), the variance decreases for sufficiently negative values of ω_{500} suggesting that heavy rain events are essentially deterministic with an approximate linear dependency on ω_{500}. This is particularly evident in the Kwajalein data (cf. Figure 1). This is consistent with the results of Craig and Cohen (2006) and Cohen and Craig (2006) who show that the variance of convective activity increases with the square root of the forcing. Preliminary analysis of coarse-grained, convection

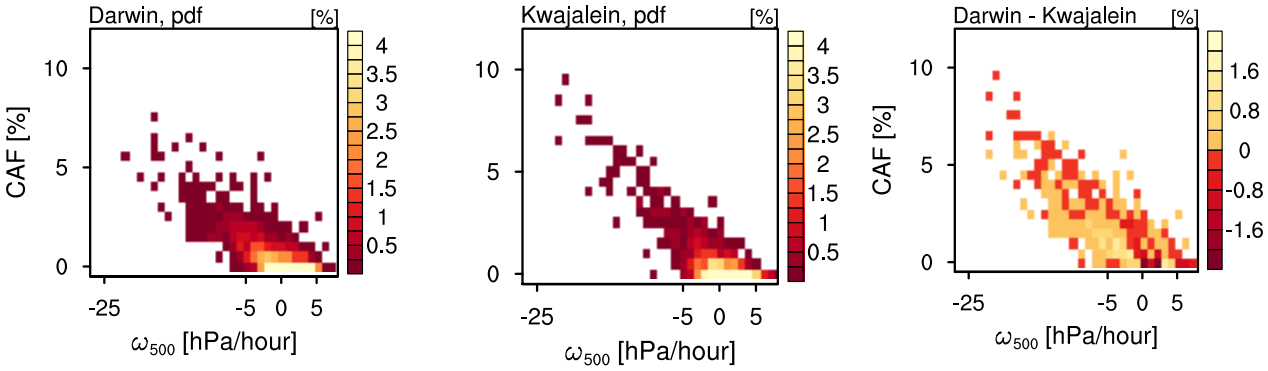

Figure 2. 2D histograms of CAF and ω_{500} obtained from observations over Darwin (left) and Kwajalein (middle). The difference of the histograms is depicted in the right most plot.

associated parameterisation tendencies obtained from the operational ECMWF forecast model reveals a similar relationship (Glenn Shutts, personal communication, 2014).

The scatterplot and the histograms in Figures 1 and 2 show that in Kwajalein much more strongly negative values of ω_{500} occur which are associated with high values of CAF exceeding 10%. This suggests that in Kwajalein stronger dynamical forcing is possible than in Darwin.

Further, for weak dynamical forcing at 500hPa with $-5 < \omega_{500} < 5$, convection over Darwin (Kwajalein) generally occurs under weakly ascending (subsiding) large-scale conditions (see Figure 2). We attribute these differences in convective behavior to the different meteorological conditions and convection initiating mechanisms determining the convective response to a given large-scale forcing in Kwajalein and Darwin. In particular, land-sea breeze induced convective organization at Darwin (diurnal cycle), and the generally more inhomogeneous surface characteristics of the Darwin domain compared to Kwajalein are expected to contribute to different convective responses given a particular forcing. Land-surface heterogeneities, such as coastlines or spatial differences in land cover, can induce subgrid-scale mesoscale circulations leading to organised convection (e.g. Pielke 2001; Rieck *et al.* 2014) which then results in mean large-scale ascent due to convective latent heating. Such significant land-surface effects on convection however are only likely to occur for relatively weak large-scale dynamical forcing (Rieck *et al.* 2014, and references therein), i.e. for $-5 < \omega_{500} < 5$ in our case. Consistent with this, we see relatively more convective activity in the range of $-5 < \omega_{500} < 0$ in Darwin compared to Kwajalein.

Figure 4 shows that CAF observed at Kwajalein and Darwin has similar autocorrelation up to lags of 12 hours. For lags longer than 12 hours, convection over Kwajalein looses memory, whereas convection over Darwin exhibits significant autocorrelation up to lags of 72 hours and features peaks corresponding to the convective diurnal cycle (every 24 hours).

In the Appendix we provide a more detailed analysis of the observations; in particular we focus on the regime classification by Pope *et al.* (2009) and on profiles of equivalent potential temperature.

The differences in convective behaviour in Darwin and Kwajalein discussed above and in the Appendix are very intriguing. A more quantitative investigation of the convective dynamics and thermodynamic at both locations

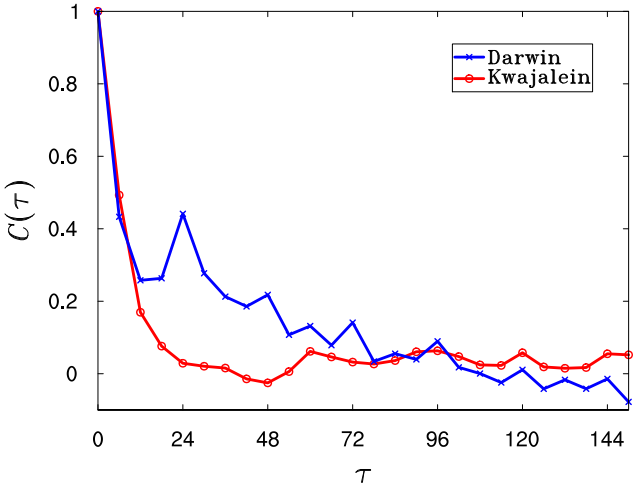

Figure 4. Temporal autocorrelation $C(\tau)$, with τ in hours, of the CAF time series for Darwin (blue crosses) and Kwajalein (red circles). For Darwin, only data snippets consisting of more than 60 time steps were used for computing $C(\tau)$.

(e.g. thermodynamics vs. dynamics) is beyond the scope of this paper and will be a subject of future work.

2.3. *Statistical universality of convective activity*

The comparison of convective behaviour in Darwin and Kwajalein above suggests that both locations feature notably different convective behaviour in terms of thermodynamics; furthermore convective initiation, triggered for example by surface fluxes, will most certainly be different at the two locations as well. In this Section we will nevertheless establish the universality of the relationship between convective activity and large-scale vertical motion which will be crucial for our stochastic parametrization schemes.

The stochastic subgrid-scale parametrizations proposed in the next Section utilise conditional probabilities such as $p(\mathrm{CAF}(t)|\omega_{500}(t))$ describing the probability of convective activity CAF occurring at time t for given vertical velocity ω_{500} at that time. We therefore now compare empirical conditional probabilities for the two locations, Darwin and Kwajalein, which we denote by p_{Darwin} and $p_{\mathrm{Kwajalein}}$, respectively. To construct the conditional probabilities we bin the $(\omega_{500}, \mathrm{CAF})$-domain into bins of size $(0.1, 0.01)$.

Assuming that the different prevailing atmospheric and oceanic regimes impact directly on the large-scale variables, we consider as a first approximation a simple translations of the large-scale vertical velocities. In particular, we show that the conditional probability functions $p_{\rm Darwin}$ and $p_{\rm Kwajalein}$ are close when the vertical velocities of Darwin are shifted as in

$$p^{\rm Kwajalein}({\rm CAF}(t)|\omega_{500}(t)) \approx p^{\rm Darwin}({\rm CAF}(t)|\omega_{500}(t)-\tau_\omega) \quad (1)$$

or analogously

$$p^{\rm Darwin}({\rm CAF}(t)|\omega_{500}(t)) \approx p^{\rm Kwajalein}({\rm CAF}(t)|\omega_{500}(t)+\tau_\omega)\,. \quad (2)$$

A standard tool to compare probability density functions is the Kullback-Leibler distance

$$D_{\rm KL}(p_{\rm Darwin}||p_{\rm Kwajalein}) = \int \log\left(\frac{p_{\rm Darwin}}{p_{\rm Kwajalein}}\right) p_{\rm Darwin}\, d{\rm CAF}\,, \quad (3)$$

with $D_{\rm KL}\geq 0$ and $D_{\rm KL}=0$ if and only if $p_{\rm Darwin}=p_{\rm Kwajalein}$ (see for example Kantz and Schreiber (1997)). Note that we have to determine a Kullback-Leibler distance for each ω_{500}-bin.

In Figure 5 we show the median of the Kullback-Leibler distance $D_{\rm KL}$ over all ω_{500}-bins as a function of the global shift τ_ω. A quadratic regression yields an optimal shift of $\tau_\omega=0.21$ where the minimum of the Kullback-Leibler distance is attained. In general, the Kullback-Leibler distance is asymmetric with $D_{\rm KL}(p||q)\neq D_{\rm KL}(q||p)$. We find, however, that $D_{\rm KL}(p_{\rm Kwajalein}||p_{\rm Darwin})$ has a minimum very close to same value of τ_ω supporting our approximation that the two conditional probability functions are related by a simple translation of the vertical velocities. We note, that due to the larger amount of available observations for Darwin ($N=1890$) when compared to Kwajalein ($N=1095$) and due to the larger support of $p_{\rm Kwajalein}$ the formulation (3) is preferred.

The universal similarities of the convective behavior at both locations can be further examined by performing a quantile regression for CAF (see for example Koeneker and Bassett (1978); Grinsted (2008)). We use a $2^{\rm nd}$ order regression and determine the conditional median for the observations of Kwajalein and Darwin. To use conditional medians rather than conditional means (as in least square regression) is advisable to eliminate the impact of the few very large rain events and other statistical outliers. The median regressions for Kwajalein and for Darwin approximately coincide if one translates the ω_{500} values of Kwajalein by $\tau_\omega=0.2$ (or those of Darwin by $-\tau_\omega=-0.2$, respectively), as seen in Figure 6. We attribute this uniform shift of the large-scale vertical velocity to the different prevailing atmospheric-oceanic regimes at the two respective locations as discussed in Section 2.2. In the following Section we shall use $\tau_\omega=0.2$.

We remark that the shift τ_ω is height dependent. We also analysed observations of the vertical velocity taken at 715 hPa; there the optimal shift for which the respective quantile regressions were closest and for which the Kullback-Leibler distance was minimal was for $\tau_\omega\approx 1.67$.

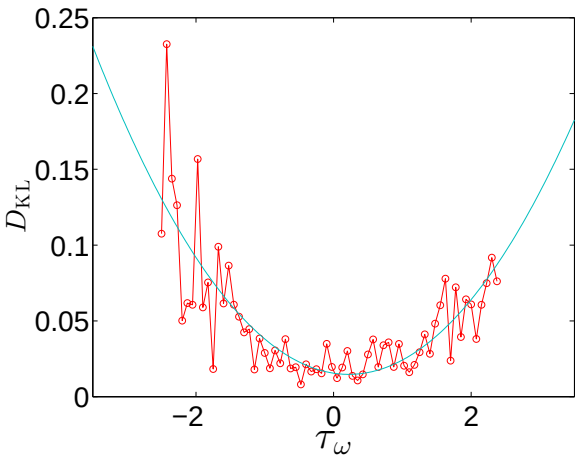

Figure 5. Kullback-Leiber distance between the conditional probability functions $p_{\rm Darwin}$ and $p_{\rm Kwajalein}$ as a function the shift τ_ω. The minimum of the quadratic least square approximation is at $\tau_\omega=0.2$.

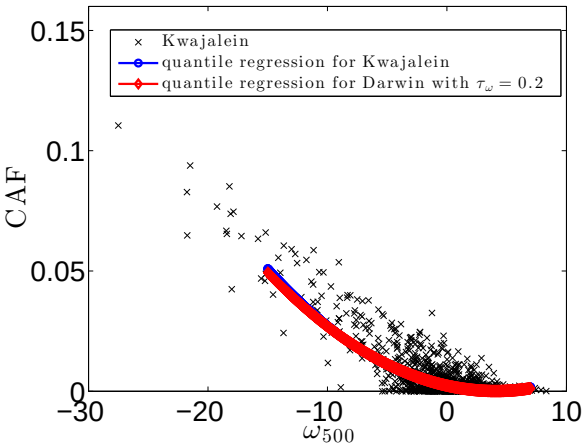

Figure 6. CAF as a function of the vertical velocities ω_{500} [hPa/hour] obtained from observations over Kwajalein (black crosses). The continuous line connecting the circles (online blue) shows the results of a $2^{\rm nd}$order $50^{\rm th}$ percentile regression. The continuous line connecting the diamonds (online red) shows the result of a $2^{\rm nd}$order $50^{\rm th}$ percentile regression for the Darwin data plotted against $\omega_{500}-0.2$.

3. Stochastic subgridscale parameterization

We will develop two stochastic subgrid-scale parametrization schemes for CAF conditioned on ω_{500}; one in which subgrid-scale convection variables such as CAF are viewed as instantaneous random variables conditioned on the current value of the large-scale vertical velocity ω_{500}, and a second approach in which the subgrid-scale variables are viewed as a conditional Markov chain taking into account non-vanishing temporal correlations of the subgrid-scale variables. The parametrisation schemes we propose model tropical convection at any location given only the information of the large-scale values of ω_{500} at a given time without any usage of the small-scale convection variables such as CAF at that time.

We are given a time series consisting of 6-hourly averaged observations of the large-scale vertical velocity at 500 hPa ω_{500} and of CAF obtained at Kwajalein and Darwin, which we denote by $\{\omega_{500_k}\}_{k=1,\cdots,N}$ and $\{y_k\}_{k=1,\cdots,N}$ with $N=1890$ for Darwin and $N=1095$

for Kwajalein, respectively (cf. Section 2). The universality argument established in Section 2.3 suggests that we can generate the stochastic model from observations of either location and apply it to the other location, respectively, if the observations of ω_{500} are corrected by a linear shift τ_ω. We present here results for both cases, but describe the methods for the situation when observations obtained in Darwin are used to train the model. To apply the resulting subgrid-scale parametrization trained in Darwin/Kwajalein, to model convection in Kwajalein/Darwin, we correct the values of ω_{500} in Kwajalein/Darwin by $\tau_\omega = \pm 0.2$ as suggested by the results in Section 2.3.

3.1. *Instantaneous conditional random variables*

In our first model convective activity is treated as a memoryless random variable conditioned on the current value of the vertical velocity ω_{500}. The algorithm for our parametrization is as follow. Let us denote by y the subgrid-scale variable, for example CAF or the rain rate. We partition the range of ω_{500} into N_ω intervals I^i_ω with $n = 1, \cdots, N_\omega$ and the range of the subgrid-scale variables into N_y intervals I^n_y with $n = 1, \cdots, N_y$. This partitions the (ω_{500}, y)-plane into $N_\omega N_y$ bins. We assume that the time series $\{\omega_{500_k}\}_{k=1,\cdots,N}$ and $\{y_k\}_{k=1,\cdots,N}$ stem from a stationary process. Coarse-grained CAF values, conditioned on the large-scale variables $\omega_{500} \in I^i_\omega$, are determined as averages over bins with

$$\bar{y}^{(n,i)} = \frac{\sum_k y_k \mathbf{1}[y_k \in I^n_y]\, \mathbf{1}[\omega_{500_k} \in I^i_\omega]}{N^{(n,i)}_y}, \qquad (4)$$

where $N^{(n,i)}_y = \sum_k \mathbf{1}[y_k \in I^n_y]\, \mathbf{1}[\omega_{500_k} \in I^i_\omega]$ is the number of y_k-values belonging to the bin defined as the intersection of the intervals I^i_ω and I^n_y. Here $\mathbf{1}[\cdot]$ denotes the indicator function with $\mathbf{1}[y_k \in I^n_y] = 1$ if $y_k \in I^n_y$ and $\mathbf{1}[y_k \in I^n_y] = 0$ otherwise. The conditional probability $P(n|i)$ of CAF y_k being in the interval I^n_y conditioned on ω_{500_k} being in the interval I^i_ω is calculated as

$$P(n|i) = \frac{\sum_k \mathbf{1}[y_k \in I^n_y]\, \mathbf{1}[\omega_{500_k} \in I^i_\omega]}{N^i_y}, \qquad (5)$$

where $N^i_y = \sum_k \mathbf{1}[\omega_{500_k} \in I^i_\omega]$ is the number of realisations of y_k for a given value of the large-scale $\omega_{500_k} \in I^i_\omega$. Note that $\sum_n P(n|i) = 1$. With probability $P(n|i)$ the subgrid-scale variable is assigned the coarse grained value $\bar{y}^{(n,i)}$.

Since Kwajalein supports convective events which are much more negative than the observations from Darwin available for the construction of the stochastic model, we use a deterministic relationship between CAF and ω_{500} for observations with $\omega_{500} < -18$ (cf. Peters *et al.* (2013)). The deterministic relationship is found by linear regression of the observations to be $\mathrm{CAF} = -0.0044\, \omega_{500} - 0.011$.

We construct the stochastic model with observations from Darwin. To test the effectiveness of the model we now apply it to observations of the large-scale vertical velocity observed in Kwajalein. We generate synthetic time series of CAF conditioned on the large-scale ω_{500} observed over Kwajalein. We partition the (ω_{500}, y)-plane into bins of size $(0.75, 0.001)$. For each value $\omega_{500_k} - 0.2 \in I^i_\omega$ we draw random variables $\bar{y}^{(n,i)}_k$ with probability $P(n|i)$ which were determined from the observations in Darwin. In Figure 7 we show the time series of the observations of CAF in Kwajalein and the corresponding synthetic time series. The stochastic model reproduces observed intermittent features of tropical convection. However, it fails to reproduce periods of sustained convection, e.g. near $t \approx 750$, and periods of sustained non-convection, e.g. near $t \approx 900$. This failure is due to our approach not incorporating any memory, despite non vanishing auto-correlations as seen in Figure 4.

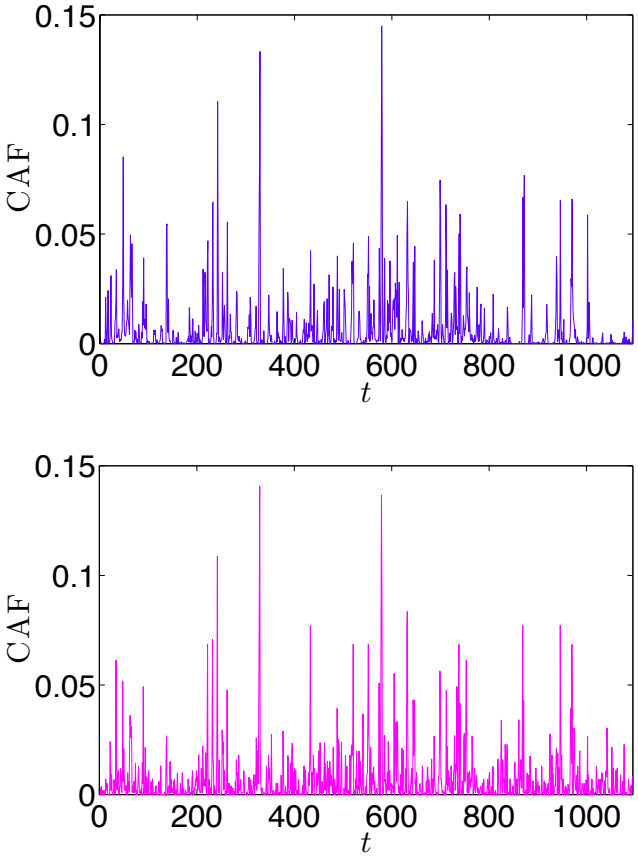

Figure 7. Time series of CAF of the observations over Kwajalein (top) and of the synthetic process conditioned on the vertical velocities ω_{500} described in Section 3.1 (bottom). The plots have a time resolution of 6 hours.

To establish a more quantitative comparison, we compare in Figure 8 the empirically determined probability density functions of CAF for the synthetic time series and the actual observations. The correspondence is remarkable for such a simple scheme. By performing averages over $1,000$ realisations of the stochastic model we have established that the first three moments of CAF in Kwajalein, the mean μ, the variance σ^2 and the skewness, are well captured by our synthetic time series. This is illustrated in Table I.

The numerical results presented in this Section used a stochastic model which was generated using the observations at Darwin and then subsequently applied to observations of large-scale vertical velocities observed at Kwajalein to produce the associated convective activity at Kwajalein. In accordance with the universality argument established in Section 2.3 we have also trained the stochastic model on the data observed at Kwajalein and applied them

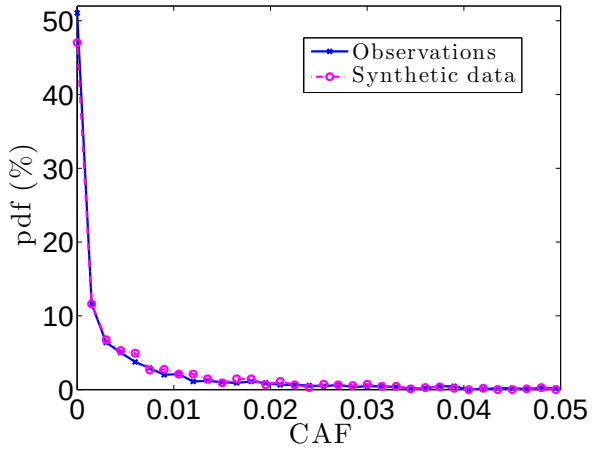

Figure 8. Empirical histogram of CAF for the observations over Kwajalein (blue) and for the synthetic process conditioned on the vertical velocities ω_{500} described in Section 3.1 (magenta).

Table I. First three moments of observed CAF for Kwajalein and of the synthetic data obtained by the subgrid-scale parametrizations conditioned on the regime-corrected values $\omega_{500}+0.2$ for the two models trained with observations from Darwin.

	μ	σ^2	skewness
observations	0.0066	$1.89\,10^{-4}$	4.27
random variable	0.0071	$1.81\,10^{-4}$	4.31
Markov chain	0.0066	$3.15\,10^{-4}$	4.21

Table II. First three moments of observed CAF for Darwin and of the synthetic data obtained by the subgrid-scale parametrization conditioned on the regime-corrected values $\omega_{500}-0.2$ for the two models trained with observations from Kwajalein.

	μ	σ^2	skewness
observations	0.0080	$1.29\,10^{-4}$	2.38
random variable	0.0077	$1.43\,10^{-4}$	2.37
Markov chain	0.0086	$2.36\,10^{-4}$	2.46

to observations of large-scale vertical velocities observed at Darwin with equal success. The results for the first three moments are shown in Table II for completeness.

We have also constructed synthetic time series of rain rate data consisting of random variables conditioned on ω_{500} and found similarly good results if the values of ω_{500} for Kwajalein are corrected by adding $\tau_\omega = 0.2$ (not shown). Further, we obtained similarly good results when parametrizing CAF conditioned on observations of the vertical velocity at 715 hPa; in this case the vertical velocities were shifted by $\tau_\omega = 1.67$.

3.2. Conditional Markov chain

The observational data obtained in Kwajalein and Darwin exhibit non-vanishing temporal autocorrelations as illustrated in Figure 4. This suggests that a more appropriate parametrization should incorporate dependencies on previous observations rather than simply conditioning on the present values of the large-scale variables. Since Kwajalein and Darwin exhibit similar values of $C(\tau)$ for a lag of one time step (6 hours), we expect a Markov model trained at one location to adequately capture the convective behaviour at the other location if conditioned on only the observations of the previous time step. As a first step towards incorporating memory one may construct a Markov chain conditioned on the previous state of the system (see, for example, Crommelin and Vanden-Eijnden (2008)) or by fitting an AR(1) process about an ω_{500}-dependent mean as in Wilks (2005). We follow here the approach proposed by Crommelin and Vanden-Eijnden (2008) for a conditional Markov chain.

To construct the Markov chain we determine a transition probability $P_{n,i}^{m,j}$ which denotes the probability for the variables (ω_{500_k}, y_k) to take values in the bin defined as the intersection of the intervals I_ω^j and I_y^m at time step k when they were in the bin defined as the intersection of the intervals I_ω^i and I_y^n at the previous time step $k-1$. To construct $P_{n,i}^{m,j}$ as a matrix we unstack the bins covering the two-dimensional (ω_{500}, y)-plane into an array of bins. The associated $N_\omega N_y \times N_\omega N_y$ transition matrix P_α^β describing transitions from bin $\alpha = i + (n-1)N_\omega$ to bin $\beta = j + (m-1)N_\omega$ is then estimated from the observations as

$$P_\alpha^\beta = \frac{T_\alpha^\beta}{\sum_{\beta=1}^{N_y N_\omega} T_\alpha^\beta}\,, \tag{6}$$

where T_α^β counts the number of transitions from the bin labelled with α to the bin labelled with β and is given by

$$T_\alpha^\beta = \sum_k \mathbf{1}[\omega_{500_{k-1}} \in I_\omega^i]\,\mathbf{1}[y_{k-1} \in I_y^n] \times \mathbf{1}[\omega_{500_k} \in I_\omega^j]\,\mathbf{1}[y_k \in I_y^m]\,.$$

To construct a Markov chain conditioned on ω_{500} taking a particular value at present time step k, we apply the transition matrix to the past given state $\alpha^\star$ at time $k-1$ to calculate $\pi_{\alpha^\star}^\beta = (0,\cdots,1,\cdots,0)P_\alpha^\beta$ where the 1 is in the $\alpha^\star$-th entry. Then we select those $L \le N_y$ bins, i.e. the non-zero coordinates of $\pi_{\alpha^\star}^\beta$, which are consistent with the current value ω_{500_k}. These L entries of $\pi_{\alpha^\star}^{\beta_l}$ with $l = 1,\cdots,L$, associated with the current value of ω_{500}, (if they exist!), do not necessarily sum up to 1 as required for a probability. Hence we renormalise as follows

$$\tilde{\pi}_{\alpha^\star}^{\beta_l} = \frac{\pi_{\alpha^\star}^{\beta_l}}{\sum_{l=1}^{L} \pi_{\alpha^\star}^{\beta_l}}\,. \tag{7}$$

The subgrid-scale variable y_k is then randomly chosen from L possible states with probability $\tilde{\pi}_{\alpha^\star}^{\beta_l}$. The assigned values corresponding to the bin labelled with β_l are coarse-grained by averaging over the bins analogously to (4).

The data sparse region of large convective activity for $\omega_{500} < -18$ is again treated with a deterministic relationship as in the instantaneous random variable model. Since the conditional Markov chain requires conditioning on the current value of ω_{500} as well as on the past observation, we are required to use larger bin-sizes to allow for transitions to be covered by the finite training set. We subdivide the (ω_{500}, y)-plane into bins of size $(0.75, 0.005)$, i.e. into 5 times larger y-bins than for the instantaneous random variable approach in Section 3.1.

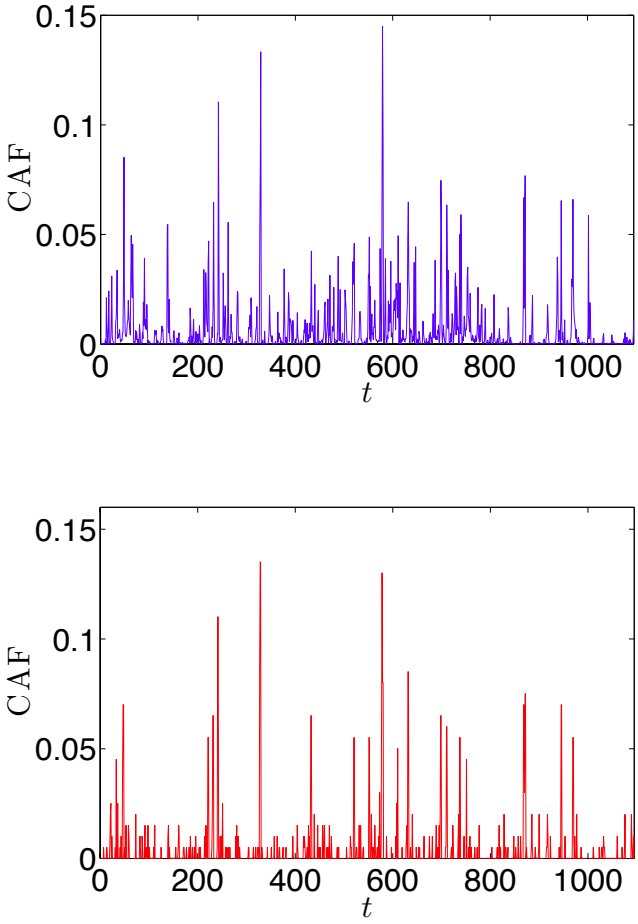

Figure 9. Time series of CAF of the observations over Kwajalein (top) and of the conditional Markov process process described in Section 3(bottom). The plots have a time resolution of 6 hours.

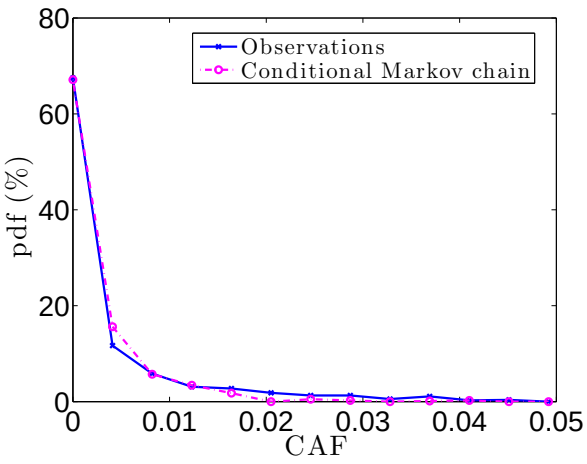

Figure 10. Empirical probability density function of CAF for the observations over Kwajalein (crosses, online blue) and for the conditional Markov chain model (circles, online magenta).

In Figure 9 we show a time series of the observations of CAF in Kwajalein and the corresponding data obtained from the conditional Markov chains which was trained with observations obtained in Darwin. Due to insufficient amount of data not all transitions could be captured leading to a shorter synthetic time series. Only approximately $3/4$ of the data points in Kwajalein can be reached by the Markov chain. Dorrestijn *et al.* (2014) have employed a Markov chain model for the data obtained in Darwin mitigating the data sparseness by i), coarse-graining the convective state into different cloud types and ii) using precipitation area fraction data at very high temporal resolution (10 minutes).

The empirical probability density functions of CAF are shown in Figure 10 with reasonable correspondence. Results of an average over $1,000$ realisations of the Markov chain for the first moments are listed in Tables I and II. Again, it is remarkable how well the statistics of the actual observations are reproduced. The variance is overestimated by the Markov chain. This may be due to the averaging of CAF within the relatively coarse bins (cf. the definition of the coarse-grained CAF values (4) which is also used in the Markov chain).

4. Summary and Conclusions

In this study, we used observations of tropical deep convection and the concurring large-scale atmospheric states at two tropical locations, Darwin and Kwajalein, to design a cheap and easy-to-implement data-driven stochastic subgrid-scale parameterisation for tropical deep convection. Our approach fits within the framework of the quasi-equilibrium hypothesis introduced by Arakawa and Schubert (1974) assuming the existence of a certain degree of scale separation between the convective activity and the large-scale dynamics (but is not limited to it).

We presented two diagnostic approaches to stochastically parametrize convective activity conditioned on large-scale vertical velocity. We did not consider here the important aspect of convective initiation, but rather provide a scheme allowing to determine convective activity once triggered. The first method treated CAF as an instantaneous random variable conditioned on the current value of ω_{500}. This method suffers from neglecting non-vanishing autocorrelations present in the observations and is not able to reproduce periods of sustained convection and non-convection, for example. The second approach was built around a conditional Markov chain and incorporates auto-correlations to some degree; this method, however, requires substantially more data to train the Markov chain as it involves conditioning on the past observations as well as on the current value of ω_{500}. Given these limitations the results are very promising. It is remarkable that the marginal probability functions of CAF as well as its first three moments were reasonably well reproduced by both approaches. In general, we would expect the conditional Markov chain to provide better diagnostics than the parametrization consisting of instantaneous random variables as it accounts for memory effects. To further test the proposed parametrization schemes we will use numerical data from high-resolution cloud resolving models in future work (or larger observational data sets if they become available).

We have shown that although both locations feature differences in convective behaviour in relation to large-scale thermodynamic profiles, a universal relationship between convective activity and large-scale vertical motion at 500 hPa, ω_{500} [hPa/hour], can be exploited for constructing our data-driven stochastic parameterisations. We showed that the stochastic model was successful in reproducing important statistical features of the observations at either location if the distribution of subgrid-scale variables was shifted towards more negative/positive values of ω_{500} in

Darwin/Kwajalein, i.e. a particular CAF (or rain rate) in Darwin/Kwajalein is associated with stronger/weaker upward motion at 500 hPa compared to the original data. We presume that this is because over Darwin, subgrid-scale surface inhomogeneities, like coastlines and the presence of land surface itself, more readily lead to convective organisation and self-enforcing mechanisms (for weak large-scale dynamical forcing) compared to Kwajalein which is a purely oceanic site. To more accurately calibrate the required shifts in the vertical velocity ω_{500} and to take into account the respective atmospheric environments of different geographical locations, numerical data from high-resolution cloud resolving models could be used as a surrogate for missing observational data in future research.

We chose to parameterise mainly subgrid-scale CAF because i), it is directly related to domain mean rainfall and thus total latent heating and ii), assigning a non-zero area fraction to convective updrafts in a convection scheme relieves the problems associated with the assumption of "scale-separation" as employed in current convection schemes (e.g. Arakawa *et al.* 2011). Current mass-flux convection schemes need to predict the vertical mass flux at cloud base. Therefore, explicitly assigning an area to the convective updraft can be combined with an updraft velocity, e.g. 1 ms^{-1}, to yield the mass flux at cloud base. Such a convective scheme would be fully scalable with convective updrafts eventually covering large portions of or even entire grid-boxes. In fact, ongoing work by one of the authors shows that such an implementation yields plausible results in a full GCM. Although CAF is suited for a resolution independent parametrization, the way the observational data have been obtained involves a particular spatial scale (i.e. the $190 \times 190\,\mathrm{km}^2$ pentagon-shaped area considered here). The observations would have to be adapted for the particular resolution of the GCM.

Acknowledgements

GAG and KP acknowledge support from the Australian Research Council. We thank Mick Pope for processing the synoptic regime data. We thank Garth Tarr and Neville Weber for discussions on quantile regression. We thank Steven Sherwood, Bob Plant, Chris Holloway and Glenn Shutts for constructive comments and suggestions to an earlier version of the paper.

A. Equivalent potential temperature analysis of the observations

A detailed analysis of the influence of land-surface characteristics on convection over Darwin is presented in Kumar *et al.* (2013) who found that the influence of land-surface heterogeneities on convection depends on the prevailing synoptic regime over Darwin as defined by the classification of Pope *et al.* (2009). In four out of the five Pope-regimes, the spatial and temporal distribution of convection over Darwin shows signatures associated with land-sea breezes and the diurnal cycle over land (Kumar *et al.* 2013). This is also evident from the autocorrelation function of our observations depicted in Figure 4, which shows pronounced peaks at multiples of 24 hour time lags for Darwin. Only the so-called "deep-west" regime, which is prevalent during the active monsoon period and is close to oceanic convection regimes, shows negligible influence of the land-surface (Kumar *et al.* 2013).

This motivates contrasting the convection and concurring large-scale meteorology over Kwajalein to regime-sorted convection and large-scale meteorology over Darwin. Scatterplots and 2D histograms of CAF as function of ω_{500} as in Figures 1 and 2 sorted by synoptic regime over Darwin (not shown) exhibit convective behaviour similar as in Kwajalein (cf. Peters *et al.* 2013) in four out of five regimes. The "dry-east" regime, a trade wind regime in which dry continental air masses are advected over Darwin (Pope *et al.* 2009), which is active in less than 9% of our data and very infrequently exhibits deep convective events (cf. Kumar *et al.* 2013), does not share the systematic convective relationships of the other four regimes. The vertical velocity ω_{500} and CAF is anti-correlated with a value of 0.38 in the "dry-east" regime, whereas anti-correlations are generally larger than 0.75 in the other regimes.

One way of characterising the interplay between tropical convection and the large-scale atmospheric state is by investigating atmospheric profiles of equivalent potential temperature θ_e. Profiles of θ_e in the tropics usually exhibit a minimum $\theta_{e,\min}$ in the lower troposphere at approximately 700 hPa (Peixoto and Oort 1992). We denote by $\Delta\theta_e$ the difference between the boundary layer (or surface) value $\theta_{e,\mathrm{b}}$ and $\theta_{e,\min}$. $\Delta\theta_e$ can be interpreted as a measure of gross moist stability (Raymond 2000) and thus is suited for characterising convective versus large-scale relationships (Neelin and Held 1987): Large (small) values of $\Delta\theta_e$ occur in dry (moist) environments associated with low (large) values of moist static energy. Observations indeed show that periods of intense convection are associated with smaller values of $\Delta\theta_e$ than those associated with periods featuring less intense convection (Aspliden 1976; Lucas and Zipser 2000) (cf. Figure 12). However, caution is advised when relating $\Delta\theta_e$ and convective activity over land, because boundary layer θ_e-values over land can be very large, favouring convection despite associated large values of $\Delta\theta_e$.

Figure 11 shows mean profiles and associated standard deviations of θ_e for Darwin and Kwajalein. The Darwin dataset is sorted according to the prevailing synoptic regimes. Since Kwajalein is only subjected to a purely oceanic regime, we only consider the mean of the respective variables. Values of $\Delta\theta_e$ corresponding to Figure 11 are provided in Table III.

For Darwin, the θ_e-profiles clearly separate with respect to the five synoptic regimes. This is mainly due to the different boundary layer temperatures. Kumar *et al.* (2013) conjectured that the "deep west" regime is the most convectively active and the "dry east" is the most suppressed regime. Consistent with this, the second largest value of $\Delta\theta_e$ is found for the "dry east" regime and the lowest value for the "deep west" regime. However, the largest value of $\Delta\theta_e$ is achieved for the "shallow west" regime (cf. Table III), which Kumar *et al.* (2013) characterise as a convectively active regime. An inspection of the θ_e-profiles shown in Figure 11 suggests that this large value of $\Delta\theta_e$ results from a relatively moist boundary layer combined with a relatively dry mid troposphere (see also Kumar *et al.* (2013), Fig. 2c). In that case, diurnally forced convection, which prevails in the "shallow west" regime, can effectively

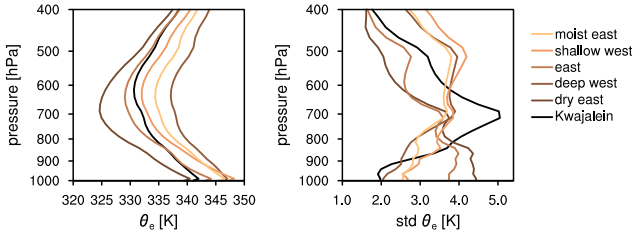

Figure 11. Profiles of mean θ_e (left) and associated sample standard deviation (right) for Darwin (composited by synoptic regime after Pope *et al.* (2009)) and for Kwajalein. The number of observations and fraction of the full sample per Pope regime are: dry east (163, 8.6%), deep west (278, 14.7%), east (210, 11.1%), shallow west (390, 20.6%) and moist east (849, 44.9%), c.f. Kumar *et al.* (2013). Kwajalein features 1095 observations.

feed on the moisture contained in the boundary layer. The "dry east" regime, on the other hand, features a relatively dry boundary layer, combined with an even drier middle troposphere, thus making it the least convectively active regime despite not exhibiting the largest value of $\Delta\theta_e$.

The shape of the mean θ_e-profile and the mean value of $\Delta\theta_e$ for Kwajalein are closest to those of the "deep west" and "moist east" regimes over Darwin. Both regimes are characterised by advection of moist, tropical air masses over Darwin (Pope *et al.* 2009), which is also found in the oceanic Kwajalein environment. The results shown in Figure 11 and Table III also somewhat support the conclusion of Kumar *et al.* (2013) that the "deep west" regime is the one most reminiscent of oceanic conditions.

Table III. $\Delta\theta_e$ in [K] as calculated from the profiles shown in Figures 11 and 12. $\Delta\theta_e$ is defined as the difference in θ_e near the surface and the θ_e-minimum in the middle troposphere. To comply with the data shown in Figure 11, we only show the mean value of $\Delta\theta_e$ for Kwajalein whereas the data for Darwin are sorted by their synoptic regime.

	$\Delta\theta_e$(Darwin)	$\Delta\theta_e$(Kwajalein)
regime, Fig. 11		
dry east	15.8	
deep west	9.9	11.3
east	15.1	
shallow west	16.2	
moist east	12.2	
CAF, Fig. 12		
$0 \leq \text{CAF} < 0.005$	14.5	12.2
$0.005 \leq \text{CAF} < 0.01$	12.9	9.8
$0.01 \leq \text{CAF} < 0.02$	12	9.4
$0.02 \leq \text{CAF} < 0.03$	10.5	9.1
$0.03 \leq \text{CAF} < 0.04$	9.7	8.5
$0.04 \leq \text{CAF} < 0.05$	8.8	8.1
$0.05 \leq \text{CAF}$	8.7	6.9

Figure 12 shows profiles of θ_e and associated standard deviations sorted by CAF as a proxy for convective activity (c.f. Aspliden 1976; Lucas and Zipser 2000). Sorting the observations by convective activity, one would expect that i) the smallest and largest values of $\Delta\theta_e$ occur in the convectively most active and most suppressed periods, respectively, ii) $\Delta\theta_e$ decreases with increasing convective activity and iii), that i) and ii) are universal for tropical convection independent of location.

For both locations, we find that as expected, situations featuring the least and most convective activity show the largest and smallest value of $\Delta\theta_e$, respectively (cf. Table III).

In Darwin, $\Delta\theta_e$ decreases monotonically with increasing convective activity, as expected. In Kwajalein however, convection seems to be less sensitive to the thermodynamic stratification in the range of intermediate convective activity with $\text{CAF} \in (0.005, 0.05)$. Over Darwin, $\Delta\theta_e$ decreases by 4.1 K in this range of CAF whereas Kwajalein shows a decrease of merely 1.7 K. Furthermore, $\Delta\theta_e$ is slightly smaller over Kwajalein compared to Darwin given a particular range of convective activity.

First, this implies that over Kwajalein, the atmosphere is generally less stably stratified compared to Darwin.

Second, convection with $\text{CAF} \in (0.005, 0.05)$ over Kwajalein appears less sensitive to mid-level relative humidity than convection over Darwin: We have checked that θ_e depends heavily on the ambient moisture profiles but only slightly on the temperature stratification (not shown; see also Aspliden (1976)). It is pertinent to mention that in the procedures involved in the data acquisition a concerted effort has been made to account for a reliable description of atmospheric moisture. In the variational analysis of Xie *et al.* (2004) used to obtain our data, the moisture profile of the large-scale domain is improved by incorporating observations of rainfall rather than using data from numerical weather prediction models alone (ECMWF analyses in this study). Moreover, information on atmospheric moisture as retrieved from satellite microwave and infrared observations are already assimilated into the ECMWF forecast system.

Third, our findings are consistent with the results of Pielke (2001) and Rieck *et al.* (2014) that convective organisation and self-reinforcement is enhanced in the presence of surface heterogeneities in situations when environmental conditions are less favourable for convection, i.e. those with large $\Delta\theta_e$.

We also note that profiles of ω_{500} over Kwajalein sorted by CAF show substantially stronger upward motion compared to Darwin (not shown). This suggests, along the lines of Raymond (2000), that environments featuring smaller values of $\Delta\theta_e$ and thus more neutral stability support stronger vertical motions. For weak convective activity ($\text{CAF} < 0.005$), θ_e shows substantially higher variability (in terms of sample standard deviations) below 500 hPa (approximately the freezing level in the tropics) compared to the more actively convecting periods at both locations (not shown). Over Kwajalein, this variability strongly increases from the boundary layer upwards – an effect attributable to the relatively constant moisture profile of the oceanic boundary layer. This high θ_e-variability for small CAF, as seen in Figure 12, can be related to the wide range of environmental conditions, i.e. ω_{500}, which allow for $\text{CAF} < 0.005$ as shown in Figures 1 and 2 at both locations.

References

Arakawa A. 2004. The cumulus parameterization problem: Past, present, and future. *J. Climate* **17**(13): 2493–2525.

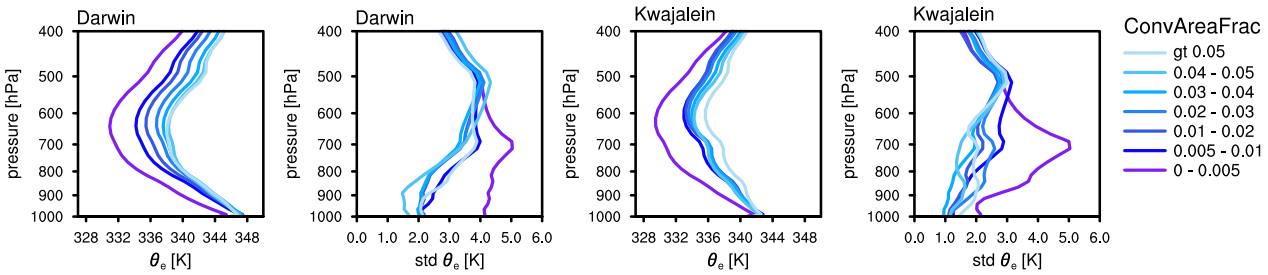

Figure 12. Profiles of mean θ_e and associated sample standard deviation for Darwin (left two panels) and Kwajalein (right two panels), sorted by observed CAF.

Arakawa A, Jung JH, Wu CM. 2011. Toward unification of the multiscale modeling of the atmosphere. *Atmos. Chem. Phys.* **11**(8): 3731–3742, doi:10.5194/acp-11-3731-2011.

Arakawa A, Schubert WH. 1974. Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *J. Atmos. Sci.* **31**(3): 674–701, doi:10.1175/1520-0469(1974)031¡0674:IOACCE¿2.0.CO;2.

Arakawa A, Wu CM. 2013. A Unified Representation of Deep Moist Convection in Numerical Modeling of the Atmosphere. Part I. *J. Atmos. Sci.* **70**(7): 1977–1992, doi:10.1175/JAS-D-12-0330.1.

Aspliden C. 1976. A classification of the structure of the tropical atmosphere and related energy fluxes. *J. Appl. Meteorol.* **15**(7): 692–697.

Bengtsson L, Steinheimer M, Bechtold P, Geleyn JF. 2013. A stochastic parametrization for deep convection using cellular automata. *Q. J. Roy. Meteor. Soc.* **139**(675): 1533–1543, doi:10.1002/qj.2108.

Berner J, Shutts G, Palmer T. 2005. Parameterising the multiscale structure of organised convection using a cellular automaton. In: *Proceedings of the ECMWF Workshop on "Representation of Sub-grid Processes Using Stochastic-dynamic Models"*. ECMWF.

Birch CE, Marsham JH, Parker DJ, Taylor CM. 2014. The scale dependence and structure of convergence fields preceding the initiation of deep convection. *Geophys. Res. Lett.* **41**(13): 4769–4776, doi:10.1002/2014GL060493.

Bright D, Mullen S. 2002. Short-range ensemble forecasts of precipitation during the southwest monsoon. *Weather Forecast.* **17**(5): 1080–1100.

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

Cohen BG, Craig GC. 2006. Fluctuations in an Equilibrium Convective Ensemble. Part II: Numerical Experiments. *J. Atmos. Sci.* **63**(8): 2005–2015, doi:10.1175/JAS3710.1.

Craig GC. 1996. Dimensional analysis of a convecting atmosphere in equilibrium with external forcing. *Q. J. Roy. Meteor. Soc.* **122**(536): 1963–1967.

Craig GC, Cohen BG. 2006. Fluctuations in an Equilibrium Convective Ensemble. Part I: Theoretical Formulation. *J. Atmos. Sci.* **63**(8): 1996–2004, doi:10.1175/JAS3709.1.

Crommelin DT, Vanden-Eijnden E. 2008. Subgrid-scale parameterization with conditional Markov chains. *Journal of the Atmospheric Sciences* **65**(8): 2661–2675.

Dai A. 2006. Precipitation characteristics in eighteen coupled climate models. *J. Climate* **19**(18): 4605–4630, doi:10.1175/JCLI3884.1.

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

Dorrestijn J, Crommelin DT, Biello JA, Böing SJ. 2013. A data-driven multi-cloud model for stochastic parametrization of deep convection. *Philos. T. Roy. Soc. A* **371**(1991, SI), doi:10.1098/rsta.2012.0374.

Dorrestijn J, Crommelin DT, Siebesma AP, Jonker H, Jakob C. 2014. A stochastic multicloud model inferred from radar data for parameterization of deep convection. *J. Atmos. Sci.* Revised.

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

Frenkel Y, Majda A, Khouider B. 2012. Using the Stochastic Multicloud Model to Improve Tropical Convective Parameterization: A Paradigm Example. *J. Atmos. Sci.* **69**: 1080–1105, doi:10.1175/JAS-D-11-0148.1.

Frenkel Y, Majda AJ, Khouider B. 2013. Stochastic and deterministic multicloud parameterizations for tropical convection. *Clim. Dyn.* **41**(5-6): 1527–1551, doi:10.1007/s00382-013-1678-z.

Grell GA, Freitas SR. 2014. A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.* **14**(10): 5233–5250, doi:10.5194/acp-14-5233-2014.

Grinsted A. 2008. quantreg.m. http://www.mathworks.com.au/ matlabcentral/ fileexchange/ 32115-quantreg-m-quantile-regression.

Groenemeijer P, Craig GC. 2012. Ensemble forecasting with a stochastic convective parametrization based on equilibrium statistics. *Atmos. Chem. Phys.* **12**(10): 4555–4565, doi:10.5194/acp-12-4555-2012.

Hohenegger C, Lüthi D, Schär C. 2006. Predictability Mysteries in Cloud-Resolving Models. *Mont. Wea. Rev.* **134**(8): 2095–2107, doi:10.1175/MWR3176.1.

Hohenegger C, Stevens B. 2013. Preconditioning deep convection with cumulus congestus. *J. Atmos. Sci.* **70**(2): 448–464, doi:10.1175/JAS-D-12-089.1.

Holloway C, Neelin J. 2009. Moisture Vertical Structure, Column Water Vapor, and Tropical Deep Convection. *J. Atmos. Sci.* **66**(6): 1665–1683, doi:10.1175/2008JAS2806.1.

Horenko I. 2011. Nonstationarity in Multifactor Models of Discrete Jump Processes, Memory, and Application to Cloud Modeling. *J. Atmos. Sci.* **68**(7): 1493–1506, doi:10.1175/2011JAS3692.1.

Horinouchi T, Pawson S, Shibata K, Langematz U, Manzini E, Giorgetta MA, Sassi F, Wilson RJ, Hamilton K, de Grandpré J, Scaife AA. 2003. Tropical Cumulus Convection and Upward-Propagating Waves in Middle-Atmospheric GCMs. *J. Atmos. Sci.* **60**(22): 2765–2782, doi:10.1175/1520-0469(2003)060¡2765:TCCAUW¿2.0.CO;2.

Jiang JH, Su H, Zhai C, Perun VS, Del Genio A, Nazarenko LS, Donner LJ, Horowitz L, Seman C, Cole J, Gettelman A, Ringer MA, Rotstayn L, Jeffrey S, Wu T, Brient F, Dufresne JL, Kawai H, Koshiro T, Watanabe M, Lcuyer TS, Volodin EM, Iversen T, Drange H, Mesquita MDS, Read WG, Waters JW, Tian B, Teixeira J, Stephens GL. 2012. Evaluation of cloud and water vapor simulations in CMIP5 climate models using NASA "A-Train" satellite observations. *J. Geophys. Res.* **117**: D14 105, doi:10.1029/2011JD017237.

Kain JS, Fritsch JM. 1990. A One-Dimensional Entraining/Detraining Plume Model and Its Application in Convective Parameterization. *J. Atmos. Sci.* **47**(23): 2784–2802, doi:10.1175/1520-0469(1990)047¡2784:AODEPM¿2.0.CO;2.

Kantz H, Schreiber T. 1997. *Nonlinear Time Series Analysis*. Cambridge University Press: Cambridge.

Keane RJ, Plant RS. 2012. Large-scale length and time-scales for use with stochastic convective parametrization. *Q. J. Roy. Meteor. Soc.* **138**(666): 1150–1164, doi:10.1002/qj.992.

Keenan T, Glasson K, Cummings F, Bird T, Keeler J, Lutz J. 1998. The BMRC/NCAR C-band polarimetric (C-Pol) radar system. *J. Atmos. Ocean Tech.* **15**(4): 871–886.

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

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

Koeneker R, Bassett G. 1978. Regression quantiles. *Econometrica* **46**(1): 33–50.

Kumar VV, Protat A, May PT, 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. *Mon. Wea. Rev.* **141**(4): 1358–1374, doi:10.1175/MWR-D-12-00160.1.

Lauer A, Hamilton K. 2013. Simulating Clouds with Global Climate Models: A Comparison of CMIP5 Results with CMIP3 and Satellite Data. *J. Climate* **26**(11): 3823–3845, doi:10.1175/JCLI-D-12-00451.1.

Lin J, Neelin J. 2000. Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.* **27**(22): 3691–3694.

Lin J, Neelin J. 2002. Considerations for a stochastic convective parametrisation. *J. Atmos. Sci.* **59**: 959–975.

Lin J, Neelin J. 2003. Toward stochastic deep convective parameterization in general circulation models. *Geophys. Res. Lett* **30**(4): 1162, doi:10.1029/2002GL016203.

Lucas C, Zipser EJ. 2000. Environmental variability during TOGA COARE. *J. Atmos. Sci.* **57**(15): 2333–2350.

Majda A, Khouider B. 2002. Stochastic and mesoscopic models for tropical convection. *Proc.Nat. Acad. Sci.* **99**(3): 1123–1128.

Manabe S, Smagorinsky J, Strickler RF. 1965. Simulated climatology of a general circulation model with a hydrologic cycle. *Mon. Wea. Rev.* **93**(12): 769–798.

Neelin JD, Held IM. 1987. Modeling Tropical Convergence Based on the Moist Static Energy Budget. *Mon. Wea. Rev.* **115**(1): 3–12.

Neelin JD, Peters O, Lin JWB, Hales K, Holloway CE. 2008. Rethinking convective quasi-equilibrium: observational constraints for stochastic convective schemes in climate models. *Philos. T. R. Soc. A* **366**(1875): 2579–2602, doi:10.1098/rsta.2008.0056.

Nuijens L, Stevens B, Siebesma AP. 2009. The environment of precipitating shallow cumulus convection. *J. Atmos. Sci.* **66**(7): 1962–1979, doi:10.1175/2008JAS2841.1.

Ooyama K. 1964. A dynamical model for the study of tropical cyclone development. *Geofis. Int.* **4**: 187–198.

Palmer T. 2001. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. Roy. Meteor. Soc.* **127**(572): 279–304.

Palmer T. 2012. Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Q. J. Roy. Meteor. Soc.* **138**: 841–861, doi:10.1002/qj.1923.

Palmer T, Williams P. 2008. Introduction. stochastic physics and climate modelling. *Philos. T. R. Soc. A* **366**(1875): 2419–2425, doi:10.1098/rsta.2008.0059.

Palmer T, Williams P. 2010. *Stochastic Pphysics and Climate Modelling*. Cambridge University Press.

Peixoto J, Oort A. 1992. *The Physics of Climate*. Springer Verlag GmbH.

Peppler RA, Lamb PJ. 1989. Tropospheric static stability and central North American growing season rainfall. *Mon. Wea. Rev.* **117**(6): 1156–1180.

Peters K, Jakob C, Davies L, Khouider B, Majda AJ. 2013. Stochastic Behavior of Tropical Convection in Observations and a Multicloud Model. *J. Atmos. Sci.* **70**(11): 3556–3575, doi:10.1175/JAS-D-13-031.1.

Pielke RA. 2001. Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. *Rev. Geophys.* **39**(2): 151–177.

Pincus R, Batstone C, Hofmann R, Taylor K, Glecker P. 2008. Evaluating the present-day simulation of clouds, precipitation, and radiation in climate models. *J. Geophys. Res.* **113**: D14 209, doi:10.1029/2007JD009334.

Plant R, Craig G. 2008. A stochastic parameterization for deep convection based on equilibrium statistics. *J. Atmos. Sci.* **65**(1): 87–105.

Pope M, Jakob C, Reeder M. 2009. Regimes of the North Australian wet season. *J. Clim.* **22**(24): 6699–6715, doi:10.1175/2009JCLI3057.1.

Randall DA. 2013. Beyond deadlock. *Geophys. Res. Lett.* **40**(22): 5970–5976, doi:10.1002/2013GL057998.

Raymond DJ. 2000. Thermodynamic control of tropical rainfall. *Q. J. Roy. Meteor. Soc.* **126**(564): 889–898.

Ricciardulli L, Garcia RR. 2000. The Excitation of Equatorial Waves by Deep Convection in the NCAR Community Climate Model (CCM3). *J. Atmos. Sci.* **57**(21): 3461–3487, doi:10.1175/1520-0469(2000)057¡3461:TEOEWB¿2.0.CO;2.

Rieck M, Hohenegger C, van Heerwaarden C. 2014. The Influence of Land Surface Heterogeneities on Cloud Size Development. *Mon. Wea. Rev.* doi:10.1175/MWR-D-13-00354.1.

Sherwood S. 1999. Convective Precursors and Predictability in the Tropical Western Pacific. *Mon. Wea. Rev.* **127**(12): 2977–2991.

Shutts G, Palmer T. 2007. Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate* **20**(2): 187–202.

Stechmann SN, Neelin JD. 2011. A Stochastic Model for the Transition to Strong Convection. *J. Atmos. Sci.* **68**(12): 2955–2970, doi:10.1175/JAS-D-11-028.1.

Steiner M, Houze R, Yuter S. 1995. Climatological characterisation of three-dimensional storm structure from operational radar and rain gauge data. *J. Appl. Meteorol.* **34**(9): 1978–2007.

Stevens B, Bony S. 2013. What Are Climate Models Missing? *Science* **340**(6136): 1053–1054, doi:10.1126/science.1237554.

Teixeira J, Reynolds C. 2008. Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Mon. Wea. Rev.* **136**(2): 483–496.

Tian B, Fetzer EJ, Kahn BH, Teixeira J, Manning E, Hearty T. 2013. Evaluating CMIP5 models using AIRS tropospheric air temperature and specific humidity climatology. *J. Geophys. Res.* **118**(1): 114–134, doi:10.1029/2012JD018607.

Waliser D, Moncrieff M. 2008. The Year of Tropical Convection (YOTC) science plan: A joint WCRP-WWRP/THORPEX international initiative. Technical Report WMO/TD No. 1452, WCRP - 130, WWRP/THORPEX - No 9, WMO, Geneva, Switzerland.

Waliser DE, Moncrieff MW, Burridge D, Fink AH, Gochis D, Goswami BN, Guan B, Harr P, Heming J, Hsu HH, Jakob C, Janiga M, Johnson R, Jones S, Knippertz P, Marengo J, Nguyen H, Pope M, Serra Y, Thorncroft C, Wheeler M, Wood R, Yuter S. 2012. The "Year" of Tropical Convection (May 2008–April 2010): Climate Variability and Weather Highlights. *Bull. Amer. Meteor. Soc.* **93**(8): 1189–1218, doi:10.1175/2011BAMS3095.1.

Wilks DS. 2005. Effects of stochastic parametrizations in the Lorenz '96 system. *Quarterly Journal of the Royal Meteorological Society* **131**(606): 389–407.

Wu CM, Arakawa A. 2014. A Unified Representation of Deep Moist Convection in Numerical Modeling of the Atmosphere: Part II. *J. Atmos. Sci.* **71**(6): 2089–2103.

Xie S, Cederwall RT, Zhang M. 2004. Developing long-term single-column model/cloud system–resolving model forcing data using numerical weather prediction products constrained by surface and top of the atmosphere observations. *J. Geophys. Res.* **109**: D01 104, doi:10.1029/2003JD004045.

Xu KM, Arakawa A, Krueger SK. 1992. The Macroscopic Behavior of Cumulus Ensembles Simulated by a Cumulus Ensemble Model. *J. Atmos. Sci.* **49**(24): 2402–2420, doi:10.1175/1520-0469(1992)049¡2402:TMBOCE¿2.0.CO;2.

Yano JI, Plant R. 2012. Finite departure from convective quasi-equilibrium: periodic cycle and dischargerecharge mechanism. *Q. J. Roy. Meteor. Soc.* **138**(664): 626–637, doi:10.1002/qj.957.

Zhang M, Lin J. 1997. Constrained variational analysis of sounding data based on column-integrated budgets of mass, heat, moisture, and momentum: Approach and application to ARM measurements. *J. Atmos. Sci.* **54**(11): 1503–1524.