Optimizing the number of convective plumes in EDMF cloud parameterization schemes using data from high-resolution LES simulations

Andrew Williams[1], Yair Cohen[2], and Tapio Schneider[2,3]

[1]St. Hilda's College, University of Oxford, UK
[2]California Institute of Technology, Pasadena, CA, USA
[3]Jet Propulsion Laboratory, Pasadena, CA, USA

Abstract

The parameterization of turbulence and convection at the sub-grid scale remains a major cause of the inter-model spread in climate predictions. The Eddy-Diffusivity/Mass-Flux (EDMF) parameterization unifies these two physical regimes by decomposing the grid box of a global climate model into two area fractions: one containing coherent updrafts and another containing the turbulent environment. To reduce computational cost, the EDMF parameterization currently neglects the variance of updraft activity and instead represents the updrafts solely by their mean values. In this project we investigate whether by including multiple updrafts in the EDMF scheme (rather than a single bulk plume, as currently used) it is possible to account for the variance of updraft activity, diagnosed from Large Eddy Simulations (LES).

To this end, we used kernel density estimation (KDE) – with N Gaussian kernels representing the multiple updrafts - to approximate the probability density function of updraft activity generated from LES for various convective regimes and domain sizes. The bandwidth of the Gaussian kernels is set the maximum bandwidth which satisfies the EDMF assumption on the Reynolds-averaged sub-grid scale variance of updraft activity. We found that for all cases tested there exists an optimal number of updrafts which minimizes the Kolmogorov-Smirnov error between the two distributions, and that this optimal number varies strongly with domain size, while also being correlated with the maximum variance of updraft activity.

Contents

1 Background and Motivation

Despite significant progress being made over previous decades of work, climate projections are still marred by large uncertainties, particularly in the Equilibrium Climate Sensitivity (ECS), a measure of the global equilibrium temperature response to a doubling of CO_2 (*Schneider et al., 2017*).

One of the key causes for this spread emanates from the representation (termed parameterization) of subgrid-scale processes such as convection and turbulence, each of which is typified by its own class of clouds with their own peculiar radiative impacts. Hence it is critical to understand how the properties of clouds (particularly low-lying equatorial clouds) will vary in response to warming so as to narrow down the uncertainty in the ECS. See *Figure 1*, below. [2]

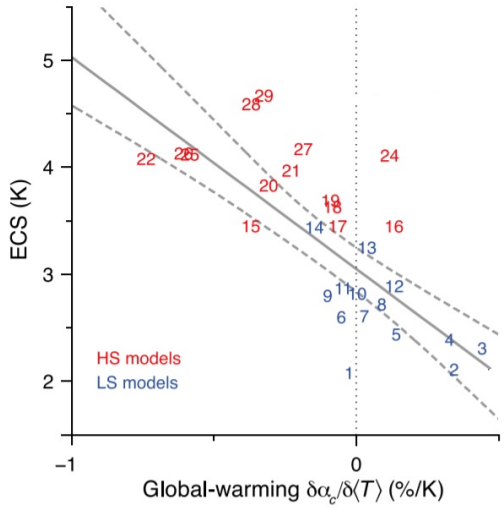

Figure 1: ECS vs global-warming tropical low cloud reflection feedback in climate models. [6]

The effect of these aforementioned subgrid-scale processes is introduced into General Circulation Models (hereafter, GCMs) using several parameterization schemes, thus introducing an artificial discontinuity into the continuous physics of turbulence and convection. The *Eddy-Diffusivity Mass-Flux (EDMF)* parameterization scheme (*Siebesma et al., 2007*) aims to unify these parameterizations (i.e. solve for both turbulence and convection in one model) by decomposing the grid box of the climate model into two area fractions, one that includes coherent updrafts and one which contains a turbulent environment (represented by a mean and variance). The time-dependent extended-EDMF scheme presented by Tan et al. (2018) might even unify boundary layer turbulence, shallow and deep convection. [5] However, any such parameterization scheme is still subjected to unknown parameters such as entrainment/detrainment rates, the optimal number of convective cells per domain size (the resolution of the GCM) etc. [4]

Recently, Tan et al. (2018) presented an extended (time dependent) EDMF parameterization which unifies boundary layer turbulence, shallow convection and deep convection. [5] This parameterization, which combines turbulence parameterization (as an *Eddy-Diffusivity (ED)*) and convective parameterization (as a *Mass-Flux (MF)*), and was tested with the assumption of a bulk plume. While this parameterization can be tested with multiple plumes, the optimal number of plumes is not yet known and may vary with different convective cases. [1]

At it's core, the EDMF is a statistical model trying to describe the probability distributions of turbulence and convection within a grid box. Such distribution from LES data is shown in *Figure 2*, below. It is clear that due to the bimodal nature of this distribution it is not possible to capture the distribution using only Mean: μ, Variance: σ^2, and Skewness: ζ, constructed from the LES data. This illustrates the need for an EDMF scheme which, through updrafts and environment, captures two different physical regimes which correspond to the different modes of the underlying distribution.

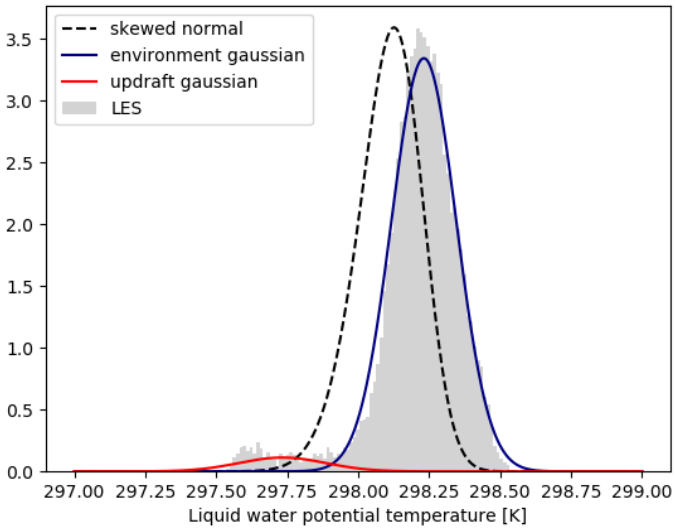

Figure 2: Comparison, RICO: LES data (histogram), LES skewed Gaussian and Updraft/Environment gaussians at 16200s, 680m.

The equation for the variance of of a scalar field ϕ in the EDMF decomposition into updraft and environment is given by the Reynolds-averaged equation:

$$\overline{\phi'\phi'} = a\overline{\phi'\phi'}^{u} + (1-a)\overline{\phi'\phi'}^{e} + a(1-a)(\overline{\phi}^{u} - \overline{\phi}^{e})^2 \tag{1}$$

Where $\overline{\phi'\phi'}$ represents the - currently unresolved - variance of the scalar field ϕ within the grid box of a GCM, $a\overline{\phi'\phi'}^{u}$ represents the variance of the updraft part of the distribution, $(1-a)\overline{\phi'\phi'}^{e}$ represents the variance of the environmental part of the distribution and $a(1-a)(\overline{\phi}^{u} - \overline{\phi}^{e})^2$ is a term which captures the added variance due to the relative separation of the means of these two updraft/environment distributions.

Currently, for reasons of mathematical simplicity and computational performance, the EDMF scheme neglects the updraft variance (the first term on the right-hand side of (1)), and as a result the updraft distribution is now represented solely by a mean value within the EDMF scheme. This is known as the *EDMF assumption*.

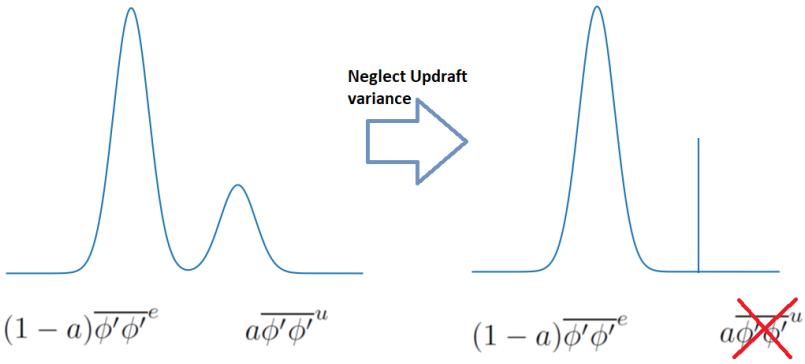

Figure 3: Schematic of the effect of neglecting updraft variance in eqn.(1).

1.1 Research goals

The aim of this work is to use multiple updrafts - each modeled as Gaussians with a standard deviation σ chosen so as not to break the EDMF assumption, described above - to recreate the distribution of updraft activity diagnosed from high-resolution LES simulations for a range of convective cases and

domain sizes.

It is also worth noting that the updrafts in any EDMF parameterization are *statistical representations* of the observed (i.e. LES modeled) cloud activity, and that the inclusion of multiple updrafts in the scheme can be regarded as the addition of second order moments (variance) to the EDMF updraft activity as a whole.

2 Approach and Methods

2.1 Generating initial distribution of updraft activity from LES

In order to determine the number of updrafts required to recreate the distribution of updraft activity diagnosed from LES, we first fitted a skewed normal distribution to the updraft data from LES. To do this we edited the PyCLES code so as to generate 3rd-order moments of the variables we were plotting. For example, $\overline{\theta'\theta'\theta'}^u$ (Liquid water potential temperature) and $\overline{b'b'b'}^u$ (buoyancy). As a result of these additions, we were then able to calculate Mean: μ, Variance: σ^2, and Skewness: ζ.

The expressions used to calculate skewness and variance from the PyCLES data are:

$$\zeta = \overline{\phi'\phi'\phi'}/\overline{\phi'\phi'}^{3/2} \tag{2}$$

$$\sigma^2 = \overline{\phi'\phi'} \tag{3}$$

where

$$\overline{\phi'\phi'} = \overline{(\phi - \overline{\phi})^2} = \overline{\phi\phi} - \overline{\phi}^2 \tag{4}$$

and

$$\overline{\phi'\phi'\phi'} = \overline{(\phi - \overline{\phi})^3} = \overline{\phi\phi\phi} - 3\overline{\phi\phi\phi} + 2\overline{\phi}^3 \tag{5}$$

Using these values, we were able to generate a skewed normal distribution representing the PDF of updraft activity in the LES which, although imperfect, is accurate enough for use in the the context of the EDMF parameterization scheme.

2.2 Kernel Density Estimation

After generating the initial updraft PDF from LES data, we then fit N Gaussian updrafts to this distribution using a standard statistical tool known as *Kernel Density Estimation* (KDE). Throughout this work, we make use of the KDE tools in the Python package *SciKit-learn*.

To perform *Kernel Density Estimation* one must specify three input parameters:

1. The *kernel* to be used. i.e. the non-negative function which will be used N times in order to fit the distribution; in this work we use Gaussian kernels.
2. The standard deviation of each of these individual kernels, also known as their *bandwidth* (BW).
3. Sample data to fit the Kernel Density model to. In our work, this step is where we prescribe the positioning of the kernels along the x-axis.

In order to make the results of our analysis relevant for the EDMF parameterization - as currently our tool does not actually utilize the EDMF scheme directly due to the number of free parameters it is subjected to - it is necessary for us to tailor the BW and position of the updraft Gaussians so as to capture the underlying physics represented in the EDMF.

2.2.1 Bandwidth of Gaussian Kernels: Upper Bound

To physically constrain the BW upper bound of the Gaussian kernels, consider the equation for the variance of of a scalar field ϕ in the EDMF decomposition into updraft and environment, given by:

$$\overline{\phi'\phi'} = a\overline{\phi'\phi'}^u + (1-a)\overline{\phi'\phi'}^e + a(1-a)(\overline{\phi}^u - \overline{\phi}^e)^2 \tag{6}$$

However, as noted above, the EDMF parameterization neglects the first term on the RHS, so for this to hold we require that:

$$\overline{\phi'\phi'}^u = \sum_{k=1}^{N} \overline{\phi'\phi'}_k = N\ \overline{\phi'\phi'}_{individual} \ll \frac{(1-a)\overline{\phi'\phi'}^e + a(1-a)(\overline{\phi}^u - \overline{\phi}^e)^2}{a} \tag{7}$$

Or, rather:

$$\overline{\phi'\phi'}_{individual} \leq \chi \frac{(1-a)\overline{\phi'\phi'}^e + a(1-a)(\overline{\phi}^u - \overline{\phi}^e)^2}{Na} \qquad for \quad \chi \ll 1 \tag{8}$$

Where χ is a small parameter reflecting the partitioning of the total variance between updraft, environment and mass-flux contributions, thus making it negligible. Also, in the lack of any knowledge about the contribution of each individual updraft, we assume they each contribute equally to the variance of updraft activity, so $\sum_{k=1}^{N} \overline{\phi'\phi'}_k \ = \ N\ \overline{\phi'\phi'}_{individual}$.

Eqn.(7) acts as an upper bound for the variance of a single updraft out of the N updrafts such that the EDMF assumption that we can neglect the $a\overline{\phi'\phi'}$ term in eqn. (5) will hold. Hence, as $BW = \sqrt{\overline{\phi'\phi'}_{individual}}$, and by taking eqn.(7) at equality, we have an upper bound on the BW for a given N and χ.

$$BW \ \leq \sqrt{\frac{\chi}{a}\ \frac{(1-a)\overline{\phi'\phi'}^e + a(1-a)(\overline{\phi}^u - \overline{\phi}^e)^2}{N}} \tag{9}$$

Regarding the size of the "normalisation factor", χ, we also reasoned that as the EDMF scheme resolves contributions of $\mathcal{O}(a)$, where a is the area fraction of the updrafts, for a contribution to the RHS of eqn.(5) to be neglected, it must be of $\mathcal{O}(a^2)$ and hence $\chi \overset{!}{\sim} a$.

Similarly, one could argue that for $a\ \overline{\phi'\phi'}^u$ to be neglected compared to the other variance terms, $\overline{\phi'\phi'}^u$ itself has to be at a similar order of magnitude as the other terms on the RHS of eqn.(5), or at least not $\frac{1}{a}$ times larger than these terms :

$$\frac{a\ \overline{\phi'\phi'}^u}{(1-a)\overline{\phi'\phi'}^e + a(1-a)(\overline{\phi}^u - \overline{\phi}^e)^2} \ \sim \ a \tag{10}$$

Comparing with eqn.(7):

$$\longrightarrow \qquad \chi \ \overset{!}{\sim} \ a \tag{11}$$

2.2.2 Bandwidth of Gaussian Kernels: Lower bound

Similarly a physical constraint on the lower bound on bandwidth is determined by arguing that the variance of updraft activity must be greater than the variance in the environment.

To see this, consider eqn.(8) at equality and take $(1-a) \approx 1$, which is true for $a \ll 1$, it is clear that the mass flux contribution to the RHS of the numerator is always positive and hence the minimum

value of the total updraft variance is given by:

$$\overline{\phi'\phi'}^u \geq \frac{\overline{\phi'\phi'}^e}{a} \tag{12}$$

and hence:

$$BW \geq \sqrt{\frac{\chi}{a}\frac{\overline{\phi'\phi'}^e}{N}} \tag{13}$$

Finally, we can say that for $\chi \sim a$:

$$\sqrt{\frac{\overline{\phi'\phi'}^e}{N}} \leq BW \leq \sqrt{\frac{(1-a)\overline{\phi'\phi'}^e + a(1-a)(\overline{\phi}^u - \overline{\phi}^e)^2}{N}} \tag{14}$$

Giving us a relatively tight bound on BW as a function of N. In fact, one can show analytically that if the difference of the means (squared) is sufficiently close to the environmental variance, then the upper and lower bounds on BW will be very close, neglecting terms of $\mathcal{O}(a^2)$. In all of the cases and domains we tested, this condition was satisfied, and so we took the BW of the Gaussians as prescribed by eqn.(9) with $\chi \sim a$.

2.2.3 Position of Gaussian Kernels

The position of the Gaussian kernels along the x-axis forms the sample space which is given as an input to the KDE package in Python, and so we have two conditions which must be met by this sample:

1. The position of the kernels must be sampled from the skewed normal distribution of (2.1) in a way which retains the Normal/Skewed Normal nature of that distribution.
2. The sampling from this distribution must also in some way be physically constrained using the assumptions in the EDMF scheme.

In light of these conditions, we chose to sub-sample the distribution of (2.1) based upon "equal-areas partitioning".

That is, we used the percent point function (PPF) of the updraft distribution in (2.1) to separate the N points equally in area under the distribution. The percent point function is the inverse of the CDF.

For example, to sub-sample 3 points from the updraft PDF, we would find the x-values at which the PPF of the updraft distribution equaled 0.25, 0.50 and 0.75. In general, to pick N points from the distribution, we would find the x-values at which the PPF of the updraft distribution was equal to $\frac{1}{N+1}$, $\frac{2}{N+1}$, ..., ..., $\frac{N}{N+1}$. See an example in *Figure 4*, below.

2.3 Error Calculation

To calculate the error between the ground truth of (2.1) and the generated KDE distribution for each point in (BW, N) space, we used the Kolmolgorov-Smirnov test, which assesses whether two underlying one-dimensional probability distributions differ by evaluating the point-by-point difference in their CDFs.

3 Results

Below we see an error contour from Rico, overlayed with the prescribed bandwidth curve. Following the error contour along this bandwidth curve we are able to draw a 1 dimensional Error[N] curve which has a clear minimum corresponding to $N_{optimal}$ for this particular domain size and convective regime.

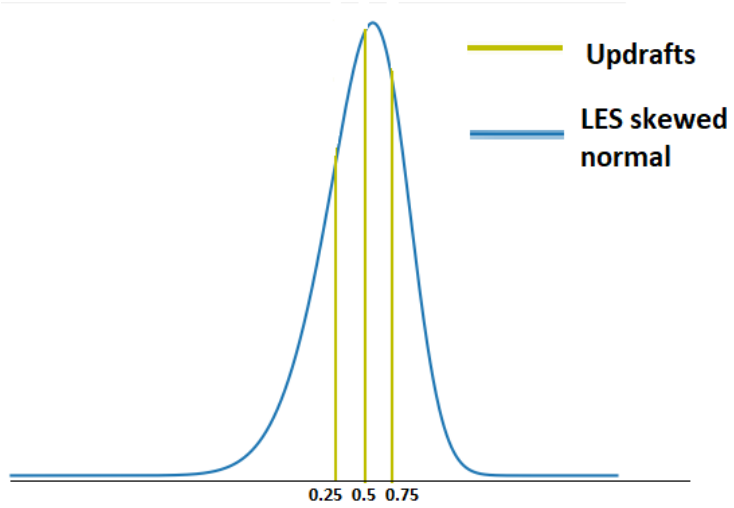

Figure 4: Positioning of the N updrafts for $N = 3$ and "equal areas" partitioning.

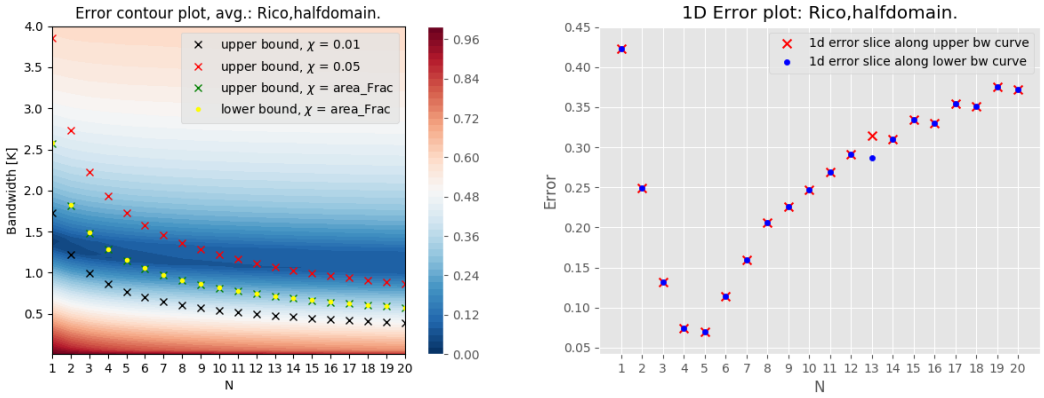

Figure 5: (Left) Error contour as a function of Bandwidth and N, with the prescribed bandwidth overlayed for various χ. (Right), 1d error plot along the $\chi = Area - Fraction$ upper bound BW curve.

3.1 Effect of domain size

The figures below illustrate how the mean updraft profiles of various LES quantities do not vary with domain size, whereas on the other hand the variance of these quantities does increase with domain size. Intriguingly, the ratios between the peak heights of the variances of these quantities seems to match (roughly) the ratio between the number of updrafts required in each situation, though more work needs to be done in order to further elucidate this link.

3.2 Effect of different convective cases

We performed our analysis on the output data from LES simulations of BOMEX and RICO, and while the value of $N_{optimal}$ changed slightly between these two cases, testing only two cases means it is difficult to deduce if these variations actually represent a correlation or just 'noise'. Regardless, it is clear that the effect of domain size is much more important in the RICO/BOMEX comparison of $N_{optimal}$ than the convective regime itself, according to our analysis.

4 Conclusions

In this work we began by recognizing the urgent need for better parameterizations of clouds and turbulence within climate models, and noted that the innately bi-modal nature of the underlying LES

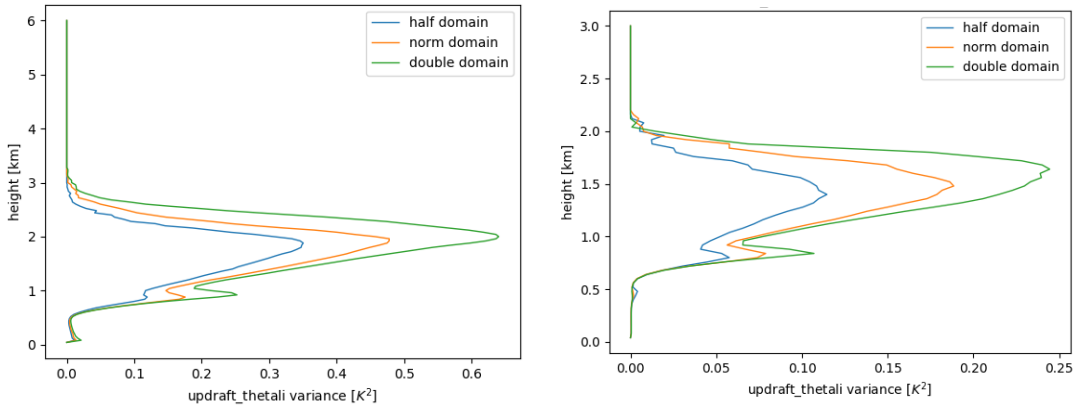

Figure 6: Variance of liquid water potential temperature (θ_l) in the updrafts for Rico (left) and Bomex (right) for different domain sizes as a function of height.

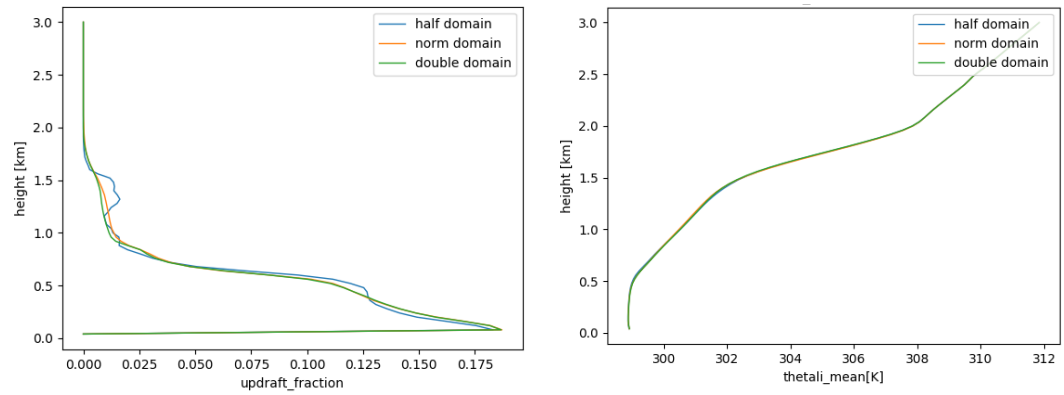

Figure 7: Mean profiles of updraft fraction (left) and liquid water potential temperature (θ_l, right) for different domain sizes as a function of height, BOMEX.

Convective case	Domain size	$N_{optimal}$
Rico	Half	5
Rico	Norm	7
Rico	Double	9
Bomex	Half	3
Bomex	Norm	7
Bomex	Double	9

Figure 8: Table 1: $N_{optimal}$ for various convective cases and domain sizes.

updraft distribution (representing two separate, but coupled, physical regimes; namely turbulent diffusion and convective upwelling) lends itself to a parameterization such as the EDMF scheme. The inadequacies of the EDMF scheme as it stands today were also discussed.

We have derived a statistical tool which takes the output from a high resolution LES simulation and gives an estimate of how many updrafts are required within the current EDMF scheme in order to allow it to recreate the variance diagnosed from the LES simulations. By minimizing the Kolmolgorov-Smirnov error between the constructed updraft distribution and the one taken from the high resolution LES simulations we were able to find that for all convective regimes and domain sizes tested there exists an $N_{optimal}$ which minimizes this error.

This error was found to vary weakly with convective cases but strongly with the domain size of the test case, but further convective regimes and domain sizes need to be checked in order to further elucidate these dependencies. It would also be prudent to investigate how the optimal number of updrafts varies with properties at the surface such as surface fluxes, shears etc, so as to form a more complete

picture of how $N_{optimal}$ varies physically. Finally, it is necessary to try and embed this statistical tool within an EDMF scheme to see what effect a varying number of updrafts has on radiative quantities such as the liquid water path and the cloud fraction.

A Reynolds decomposition of the updraft variance

In (7), we decomposed the variance of updraft activity as:

$$\overline{\phi'\phi'}^u = \sum_{k=1}^{N} \overline{\phi'\phi'}_k = N\ \overline{\phi'\phi'}_{individual}$$

However, in doing so we ignored the contribution to the variance which arises due to the distinct means of the individual updrafts, which we prescribe through the "equal-areas" partitioning as explained in 2.2.3. Here we apply a Reynolds decomposition to the updraft sub-domain in order to represent it's variance in terms of the variance of the individual updrafts and also the differences between their mean values.

For a general variable ϕ in the updraft, we can write both $\overline{\phi}$ and $\overline{\phi\phi}$ in terms of the N contributions from the individual updrafts.

$$\overline{\phi} = \sum_{i=1}^{N} A_i \overline{\phi}_i \tag{15}$$

$$\overline{\phi\phi} = \sum_{i=1}^{N} A_i \overline{\phi\phi}_i \tag{16}$$

Where $\overline{\phi}_i$ and $\overline{\phi\phi}_i$ are the means and variances of the individual updrafts within the updraft sub-domain and where A_i is the statistical *weight* we assign to each individual updraft sub-sub-domain, it plays an analogous role to the area fraction a which we have in our typical updraft/environment EDMF decomposition..

Writing the total variance of the updraft sub-domains as $\overline{\phi'\phi'} = \overline{\phi\phi} - \overline{\phi}^2$, we can (15) and (16) to write the total updraft variance $\overline{\phi'\phi'}$ as:

$$\overline{\phi'\phi'} = \overline{\phi\phi} - \overline{\phi}^2 \tag{17}$$

$$\overline{\phi'\phi'} = \sum_{i=1}^{N} A_i \overline{\phi\phi}_i - (\sum_{i=1}^{N} A_i \overline{\phi}_i)(\sum_{j=1}^{N} A_j \overline{\phi}_j) \tag{18}$$

$$\overline{\phi'\phi'} = \sum_{i=1}^{N} A_i \overline{\phi\phi}_i - \sum_{i=1}^{N} \{(A_i \overline{\phi}_i)^2 + {\sum_{\substack{0<j\leq N \\ i\neq j}}}' (A_i A_j \overline{\phi}_i \overline{\phi}_j)\} \tag{19}$$

Where the $\sum'$ simply denotes a conditional summation.

Noting that $\overline{\phi\phi}_i = \overline{\phi'\phi'}_i \overline{\phi}_i^2$, and grouping terms, we arrive at:

$$\overline{\phi'\phi'}^u = \sum_{i=1}^{N} (A_i \overline{\phi'\phi'}_i + A_i(1 - A_i)\overline{\phi}_i^2 - {\sum_{\substack{0<j\leq N \\ i\neq j}}}' (A_i A_j \overline{\phi}_i \overline{\phi}_j)) \tag{20}$$

Now, inserting this new expression for the updraft variance into (7), and requiring that $A_i = \frac{1}{N} \quad \forall i$, as the A_is must sum to unity and additional scaling arguments show that $A_i \sim \mathcal{O}(\frac{1}{N})$, we arrive at the new upper bound on bandwidth (assuming the bandwidth of each individual bandwidth is the same):

$$BW \leq \sqrt{\frac{\chi}{a}(1-a)\overline{\phi'\phi'}^{e} + a(1-a)(\overline{\phi}^{u} - \overline{\phi}^{e})^2 - \sum_{i=1}^{N}\left[\frac{1}{N^2}((N-1)\overline{\phi}_i^2 - \sum_{\substack{0<j\leq N \\ i\neq j}}{}'(\overline{\phi}_i\overline{\phi}_j)\right]} \quad (21)$$

Again, as before a similar form can be found for the lower bound, but these are essentially equal, and so (21) can be taken as the prescribed bandwidth along which the error is calculated.

When this new scheme is implemented however, the Kolmolgorov-Smirnov error is found to be roughly constant as a function of N (changing only by 10^{-3} between $N = 0$ and $N = 20$) suggesting that when this new bandwidth formula follows a contour of constant error in the error map.

References

[1] Pressel, K.G., Kaul, C. M., Schneider, T., Tan, Z., & Mishra, S. (2015): Large eddy simulation in an anelastic framework with closed water and entropy balances. *J. Adv. Model. Earth Syst.*,,7, 1425-1456, https://doi:10.1002/ 2015MS000496.

[2] Schneider, T., Teixeira, J., Bretherton, C., Florent, B., Pressel, K., Schär, C., & Siebesma, A. P. (2017)., Climate goals and computing the future of clouds. *Nature Climate Change*, 7.

[3] Schneider, T., Lan, S., Stuart, A., & Teixeira, J. (2017):, Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophysical Research Letters,* 44, 12,396 12,417. https://doi.org/10.1002/2017GL076101

[4] Siebesma, P. A., Soares, P. M. M. & Teixeira, J., (2007): A Combined Eddy-Diffusivity Mass-Flux Approach for the Convective Boundary Layer. *Journal of Atmospheric Sciences,* 64, 1230-1248.

[5] Tan, Z., C. M. Kaul, K. G. Pressel, Y. Cohen, T. Schneider, & J. Teixeira, (2018): An extended eddy-diffusivity mass-flux scheme for unified representation of subgrid-scale turbulence and convection. *Journal of Advances in Modeling Earth Systems*, in press.

[6] Brient, F. & T. Schneider, (2016): Constraints on Climate Sensitivity from Space-Based Measurements of Low-Cloud Reflection. *J. Climate*, 29 5821-5835.

[7] vanZanten, M.C. et al. (2011): Controls on precipitation and cloudiness in simulations of trade-wind cumulus as observed during RICO. *J. Adv. Model. Earth Syst.*,,3, https://doi:10.1029/2011MS000056