Continuous Structural Parameterization: A method for representing different model parameterizations within one structure demonstrated for atmospheric convection

Francis Hugo Lambert[1], Peter Challenor[1], Neil T Lewis[2], Douglas J. McNeall[3], Nathan E Owen[1], Ian Boutle[4], Hannah Christensen[2], Richard Keane[5], Alison Stirling[6], Mark J Webb[7], and Nathan J Mayne[1]

[1]University of Exeter
[2]University of Oxford
[3]Met Office Hadley Centre
[4]Met Office
[5]University of Leeds
[6]UK Met Office
[7]UK Met Office Hadley Centre

November 24, 2022

Abstract

Continuous Structural Parameterization (CSP) is a method for approximating different numerical model parameterizations of the same process as functions of the same gridscale variables. This allows systematic comparison of parameterizations with each other and observations or resolved simulations of the same process. Using the example of two convection schemes running in the Met Office Unified Model (UM), we show that a CSP is able to capture concisely the broad behavior of the two schemes, and differences between the parameterizations and resolved convection simulated by a high resolution simulation. When the original convection schemes are replaced with their CSP emulators within the UM, basic features of the original model climate and some features of climate change are reproduced, demonstrating that CSP can capture much of the important behavior of the schemes. Our results open the possibility that future work will estimate uncertainty in model projections of climate change from estimates of uncertainty in simulation of the relevant physical processes.

Continuous Structural Parameterization: A method for representing different model parameterizations within one structure demonstrated for atmospheric convection

F. H. Lambert[1], P. G. Challenor[1], N. T. Lewis[2], D. J. McNeall[3], N. Owen[4], I. A. Boutle[3], H. Christensen[2], R. J. Keane[3], N. J. Mayne[1], A. Stirling[3] and M. J. Webb[3]

[1]College of Engineering, Mathematics and Physical Science, University of Exeter
[2]Atmospheric, Oceanic and Planetary Physics, University of Oxford
[3]Met Office
[4]University of Exeter Business School, University of Exeter

Key Points:

- CSP is a method for expressing GCM parameterizations as functions of the same gridscale variables.
- The broad behavior and differences between representations of convection are captured.
- If parameterizations are replaced with their CSP emulators in a GCM, stable climates are retrieved.

Corresponding author: F. Hugo Lambert, `f.h.lambert@exeter.ac.uk`

Abstract

Continuous Structural Parameterization (CSP) is a method for approximating different numerical model parameterizations of the same process as functions of the same grid-scale variables. This allows systematic comparison of parameterizations with each other and observations or resolved simulations of the same process. Using the example of two convection schemes running in the Met Office Unified Model (UM), we show that a CSP is able to capture concisely the broad behavior of the two schemes, and differences between the parameterizations and resolved convection simulated by a high resolution simulation. When the original convection schemes are replaced with their CSP emulators within the UM, basic features of the original model climate and some features of climate change are reproduced, demonstrating that CSP can capture much of the important behavior of the schemes. Our results open the possibility that future work will estimate uncertainty in model projections of climate change from estimates of uncertainty in simulation of the relevant physical processes.

Plain Language Summary

Numerical models are used to provide estimates of future weather and climate change. The models contain “parameterizations”, which are algorithms that simulate the effect of processes too small or poorly understood to represent using physical equations. Although they are based as far as possible on physics, parameterizations are a large source of modeling uncertainty because there can be large disagreements on how best to represent a given process. The method and even the variables used by two different parameterizations may differ. It is therefore very difficult to know how different parameterizations cause numerical models to produce different results and whether the parameterizations we have are adequate and span the range of uncertainty concerning our knowledge of the processes they represent. Using the example of small-scale atmospheric convection linked to rain and thunderstorms, this paper describes a mathematical method for expressing different parameterizations within the same framework. This allows easy but formal mathematical comparison of different parameterizations and gives future work the potential to understand whether our parameterizations perform as they should in conjunction with future observations.

1 Introduction

Numerical models of weather and climate contain “parameterizations”, which are physically motivated but approximate algorithms that represent processes that cannot be simulated explicitly on the model grid. One example is atmospheric convection, which could be represented by the same equations of fluid dynamics and thermodynamics used to simulate larger-scale atmospheric dynamics, but which typically occurs below the grid-scale of contemporary climate models and some numerical weather prediction models. Another example is land surface vegetation, for which we do not even know the governing equations. The aim of parameterization is to relate the behavior of interest to resolved processes on the model grid. Parameterizations are derived semi-empirically using insights from process understanding, observations or high-resolution simulations that do capture the relevant processes explicitly but that would be too expensive to run inside a weather or climate model.

A body of literature suggests that parameterizations are the chief cause of differences between predictions of future climate change taken from different climate models (e.g. Mauritsen et al., 2012; Webb et al., 2013; Sherwood et al., 2014; Geoffroy et al., 2017). What is not known, however, is exactly how the parameterizations that we have are different from each other and whether the differences are representative of our uncertainty in the relevant processes. This poses a problem for climate prediction because it is unclear how to translate climate model output into probability distributions of possible future climate change. The difficulty arises partly because different parameterizations of the same process can have different physical bases, meaning that they may be written in terms of different equations and even different variables, and partly because it is not clear how best to write parameterizations in a way that is directly comparable to observations or resolved simulations of the same process.

Previous work has endeavored to address some of these problems. Perturbed Physics Ensembles (PPEs) are groups of general circulation model (GCM) simulations derived from one base climate model but with their uncertain parameterization parameters perturbed over the ranges of values considered possible by relevant experts (e.g. Murphy et al., 2004; Sanderson, 2011; Sexton et al., 2019). PPEs explore the uncertainty associated with one set of parameterizations systematically because the difference between different ensemble members is unambiguously defined by the differences in their parameters. However the approach is trapped within one model structure and cannot fully explore the set of plausible parameterizations. Another set of parameterizations can be introduced into the ensemble (e.g. Shiogama et al., 2013), but the ability to define systematic differences is lost.

Meanwhile, the impulse-response method of Kuang (2010); Herman and Kuang (2013) does allow systematic comparison of parameterizations in a way that is agnostic to their structure by testing the effect of idealised perturbations in the model resolved-scale variables on parameterization and then encapsulating those responses in a response matrix. As Arakawa (2004) and Herman and Kuang (2013) stated, one view is that the important question is "What does each scheme actually do [at the resolved scale]?". The internal machinery of each parameterization is secondary. This is particularly true where different parameterizations have different physical motivations, because mechanistic comparison of the internal workings of each parameterization may not then be possible. Further, because the impulse-response method is written as a function of the resolved variables only, it is possible in principle to do the same analysis for high-resolution simulations or observations of the same process, as Herman and Kuang (2013) demonstrated for atmospheric convection.

The derived response matrices must also be put in the GCM in place of the original parameterization, as was done by Kelly et al. (2017) and Mapes et al. (2019) for the impulse-response method. This is necessary if we are to demonstrate that the matrix representation captures the essence of the parameterization relevant to modeling. We can then test the effects of multiple structurally distinct schemes using one parameterization code and define and explore the unknown parameter space between them in a GCM. If the matrix representation was sufficiently accurate, then the extent to which a particular parameterized process is responsible for inter-model differences when all other model components remain the same could be determined without the expensive overhead of having to port a range of structurally different parameterizations to one GCM. As with PPEs, the systematic differences between model versions would be known and it would be possible to determine quantitatively how available parameterizations differ from one another and how well they sample the possible "structural" parameter space defined by the response matrices compared with observations or high-resolution simulations. If differences in GCM simulation of some aspect of climate change were strongly identified with parameterization of one or more processes, then over or under-sampling of regions of the relevant parameter space could be taken into account when providing projections of future climate change. This would be an alternative to rewarding each GCM in the ensemble with one vote, as is frequently done in ensemble studies of climate change (e.g. Collins et al., 2013). The over or under-sampling of regions of the structural parameter space could also assist the direction of future model development.

A variety of studies have shown the potential for "machine learning" techniques to represent complex atmospheric processes and replace traditional parameterizations running within a GCM. Krasnopolsky (2010) used a neural network to replace the radiation parameterization within the Community Atmosphere Model (CAM). Errors were comparable with the GCM's natural internal variability for a fully-coupled ocean-atmosphere simulation. The speed of simulation was also substantially accelerated compared with the case of using the original parameterization. O'Gorman and Dwyer (2018) used a random forest algorithm to parameterize convection in an idealised version of the Geophysical Fluid Dynamics Laboratory model coupled to a slab ocean and trained on data from a conventional convective parameterization. An accurate representation of both the climatological and climate change features of the original GCM containing the conventional parameterization was achieved. Rasp et al. (2018) used a neural network to represent all parameterized processes in the model atmosphere of the super-parameterized CAM (SPCAM). Super-parameterization means that there is a high-resolution simulation within each GCM gridbox and hence Rasp et al. (2018)'s neural network was effectively emulating a high-resolution explicit representation of sub-GCM gridscale processes (although processes are not shared between GCM gridboxes). It was found that the neural network parameterization provided an accurate simulation of precipitation, atmospheric heating and wave structure when compared to SPCAM and superior to the conventionally parameterized CAM. Brenowitz and Bretherton (2018) trained a neural network to emulate all sub-gridscale processes in a high-resolution simulation. It was found that using this neural network parameterization in the CAM led to a superior simulation when compared with the conventionally parameterized CAM. These studies suggest that applying machine learning techniques to parameterization will be useful for improving GCM accuracy and computational speed. Combined with impulse-response or other statistical techniques, they can also be useful for understanding how to parameterize processes (O'Gorman & Dwyer, 2018), although direct interpretation of what complex neural networks or random forest techniques are doing remains difficult.

In this paper we describe Continuous Structural Parameterization (CSP), which is a method for writing parameterizations of the same process at a given model resolution in terms of functions of the same gridscale variables, making parameterizations with distinct structures formally comparable to one another, but retaining enough skill to replace the original parameterizations in a GCM. We base our discussion around a candidate CSP for atmospheric convection derived using linear algebra. In some ways the approach is similar to the forward method of Kuang (2010); Herman and Kuang (2013). Where it differs is in the attempt to achieve efficient descriptions of parameterizations through a set of orthogonal modes most important to GCM simulation. This allows easy analysis of how parameterizations differ from one another or observations or high-resolution simulations of the same physical process. Orthogonality also allows fitting of our statistical model to output from standard GCM simulations. CSP has four broad goals:

1. Build a statistical emulator that expresses the gridscale outputs of parameterizations as simple functions of their gridscale inputs.
2. Provide low dimensional descriptions of the most important differences between parameterizations and high-resolution simulations or observations using a diagram or other easily interpretable method.
3. Replace original parameterizations with CSP statistical emulators in the GCM to assess the degree to which relevant processes are captured.
4. Test the importance of errors introduced by a given parameterization type in ensembles of models used to predict climate change.

The overall aim is not to replace conventional parameterization nor to improve GCM integration speed, but to understand our parameterizations in the context of process knowledge and provide tools for parameterization development and interpretation of climate model projections. Here we approach goals 1–3 for convective parameterization using an example CSP, following the earlier work described above and recognising that convection is believed to be one of the key processes causing model error in current GCMs (Sherwood et al., 2014; Webb et al., 2015). When our CSP emulators are run in a GCM in place of the original parameterizations, we find that basic features of climate and some features of climate change are preserved. Our results are less accurate than those achieved when machine learning techniques are applied, but we retain the ability to explain differences between parameterizations and a high resolution dataset. The remainder of the paper is organised as follows. Section 2 describes the GCM experiments that we use to build and test CSP, Section 3 presents our statistical methodology, Section 4 presents our results for both parameterized and high-resolution representations of convection, Section 5 is a discussion of the implications of our results and Section 6 presents our main conclusions.

2 Model experiments

To train and test our statistical emulators, we take data from both coarse simulations run with parameterized convection and high-resolution convection permitting simulations.

2.1 UM simulations

Our coarse simulations with parameterized convection are run using the Global Atmosphere 7.0 configuration of the Met Office Unified Model (UM) (Walters et al., 2019). The UM solves the fully compressible, deep-atmosphere, non-hydrostatic Navier-Stokes equations using a semi-implicit, semi-Lagrangian approach. Parameterizations of atmospheric radiation, boundary layer turbulence, large-scale and convective cloud and precipitation are included. The model resolution is 2.5° longitude by 2° latitude with 38 vertical levels and a timestep of 15 minutes. Two convection schemes are used in our study: the well-established Gregory-Rowntree (GR) mass-flux scheme of Gregory and Rowntree (1990) with improvements described by Walters et al. (2019), and the Lambert-Lewis (LLCS) simple moist adjustment scheme of the authors' devising described in Appendix A. The statistical emulation of these two schemes and their differences is the basis for our demonstration of CSP. The model atmosphere is coupled to a 2.5 m deep "slab" ocean with thermodynamics but no representation of ocean dynamics (Boutle et al., 2017). The model is free to find its own equilibrium state by bringing top of atmosphere radiative fluxes into balance.

A number of simplifications to the simulations were made to ease the process of coding the statistical emulators and to simplify the behavior that needs to be predicted. The UM was run in aquaplanet mode with no continents or sea ice. The sophisticated prognostic cloud scheme (PC2) and the radiative effect of clouds were switched off to simplify the relationship between GR and gridscale water. The UM's targeted diffusion parameterization was switched on, as it was found that very occasional gridpoint storms occurred when running the LLCS CSP emulator. (Gridpoint storms are large values of gridscale precipitation and upward vertical velocity that occur when physically unrealistic resolved convection arises.) Targeted diffusion disperses boundary layer water vapor to adjacent gridboxes when gridscale vertical velocity crosses a threshold (0.2 ms^{-1} in our simulations).

We run 10 year control (0.5941 g kg^{-1} atmospheric CO_2) and $4\times CO_2$ (2.3764 g kg^{-1} atmospheric CO_2) simulations for LLCS and GR, and the cases where the original convection schemes are replaced by their CSP statistical emulators between latitudes 30 °N and 30 °S and no convection scheme is used poleward of 30 ° (GREMU and LLCSEMU). (It would be preferable to run the original parameterizations poleward of 30 °, but this is technically difficult for GR. A test with LLCS shows that similar results are found for the original parameterization and no parameterization poleward of 30 ° cases (not shown).) For the original parameterization GR and LLCS cases, we also run one 30 day simulation for January and one 30 day simulation for July for which values of potential temperature, θ, and specific humidity, q, are output on every model level at every timestep directly before and after convection between 30 °N and 30 °S allowing us to collect cases that we will use to train the statistical emulator in Section 3. These simulations are spun off from January 1st and July 1st of year 5 of the relevant 10 year simulation. We also run control and $4\times CO_2$ cases for two perturbed physics setups with the original LLCS parameterization in which the value of the critical relative humidity for initiation of moist convection, r_c, is perturbed from its standard value of 0.8 to 0.7 and 0.9. All the simulations are summarised in Table 1.

2.2 Cascade high-resolution simulations

We use data derived from the 4 km convection permitting simulations of the Cascade experiment (Holloway et al., 2012). As above, the Cascade simulations are run with the UM, but the 4 km resolution allows the convective parameterization to be switched off and the explicit dynamics of the model dynamical core are used to represent convection. The expectation is that a much more faithful simulation of convection should be achieved than when a parameterization is used making Cascade a good tool to benchmark parameterizations against (e.g. Guichard et al., 2004). Christensen et al. (2018) produced a coarse-grained version of the 4km Cascade data to provide forcing data for

Table 1. Met Office Unified Model Simulations

Simulation	CO_2 [g kg^{-1}]	Convection	Training output	Length
LLCS CON	0.5941	LLCS, $r_c = 0.8$	OFF	10 years
LLCS 4×CO_2	2.3764	LLCS, $r_c = 0.8$	OFF	10 years
GR CON	0.5941	GR	OFF	10 years
GR 4×CO_2	2.3764	GR	OFF	10 years
LLCS CON $r_c = 0.7$	0.5941	LLCS, $r_c = 0.7$	OFF	10 years
LLCS 4×CO_2 $r_c = 0.7$	2.3764	LLCS, $r_c = 0.7$	OFF	10 years
LLCS CON $r_c = 0.9$	0.5941	LLCS, $r_c = 0.9$	OFF	10 years
LLCS 4×CO_2 $r_c = 0.9$	2.3764	LLCS, $r_c = 0.9$	OFF	10 years
LLCSEMU CON	0.5941	LLCS emulator	OFF	10 years
LLCSEMU 4×CO_2	2.3764	LLCS emulator	OFF	10 years
GREMU CON	0.5941	GR emulator	OFF	10 years
GREMU 4×CO_2	2.3764	GR emulator	OFF	10 years
LLCS CON January	0.5941	LLCS, $r_c = 0.8$	ON	30 days
LLCS CON July	0.5941	LLCS, $r_c = 0.8$	ON	30 days
LLCS 4×CO_2 January	2.3764	LLCS, $r_c = 0.8$	ON	30 days
LLCS 4×CO_2 July	2.3764	LLCS, $r_c = 0.8$	ON	30 days
LLCS CON $r_c = 0.7$ January	0.5941	LLCS, $r_c = 0.7$	ON	30 days
LLCS CON $r_c = 0.7$ July	0.5941	LLCS, $r_c = 0.7$	ON	30 days
LLCS 4×CO_2 $r_c = 0.7$ January	2.3764	LLCS, $r_c = 0.7$	ON	30 days
LLCS 4×CO_2 $r_c = 0.7$ July	2.3764	LLCS, $r_c = 0.7$	ON	30 days
LLCS CON $r_c = 0.9$ January	0.5941	LLCS, $r_c = 0.9$	ON	30 days
LLCS CON $r_c = 0.9$ July	0.5941	LLCS, $r_c = 0.9$	ON	30 days
LLCS 4×CO_2 $r_c = 0.9$ January	2.3764	LLCS, $r_c = 0.9$	ON	30 days
LLCS 4×CO_2 $r_c = 0.9$ July	2.3764	LLCS, $r_c = 0.9$	ON	30 days
GR CON January	0.5941	GR	ON	30 days
GR CON July	0.5941	GR	ON	30 days
GR 4×CO_2 January	2.3764	GR	ON	30 days
GR 4×CO_2 July	2.3764	GR	ON	30 days

the European Centre for Medium-Range Weather Forecasting (ECMWF) Integrated Forecasting System (IFS) single column model (SCM). The SCM was then run forced by the coarse-grained Cascade input data. This is very useful for our study because both the coarse-grained overall tendency of the Cascade data and the dynamical and parameterized tendencies of the IFS SCM were archived by Christensen et al. (2018), allowing us to construct an emulator of a high-resolution simulation of convection.

We take the coarse-grained overall tendency of the Cascade data from the last nine days of the simulation (avoiding the spin-up) over a region of the Indian Ocean (54 °E – 90 °E longitude and 21 °S – 4.5 °N latitude), subtract the radiative, boundary layer and coarse dynamical tendencies of the SCM obtaining an estimate of the remaining dynamical processes that ought to be represented by a convection scheme. The estimate is not likely to be highly accurate since Cascade was created using the Met Office UM and the SCM is an ECMWF product. Cascade data are also only archived once per hour, in contrast to the 15 minute timesteps of the SCM, meaning that the SCM may drift substantially from the Cascade state as it is re-initialised only once every four timesteps. We further average the data in the horizontal from its Christensen et al. (2018) resolution of $0.3^\circ \times 0.3^\circ$ to as a close as possible to the UM grid of 2.5° longitude by 2° latitude without horizontal interpolation but interpolated in the vertical to the UM grid to improve comparability. Given their limitiations, we analyse these data as a demonstration rather than a definitive investigation of treating high-resolution simulation of convection with CSP.

3 Statistical methodology

3.1 Linear models

In this section the statistical techniques we use to build convection emulators using training data are presented. First, we take vertical columns of potential temperature, θ, and specific humidity, q, on model levels and their respective changes across the convective timestep, $\Delta\theta$ and Δq, from a GCM or the high-resolution simulation. There are other variables that are typically inputted into and outputted from convective parameterizations, but θ and q are the most important and the data we use for this first study. The θ and q values have their mean subtracted on each level and are then converted to components of moist enthalpy $c_p\theta$ and Lq, where c_p is the specific heat capacity of dry air at constant pressure and L is the latent heat of vaporisation. This ensures that the dry and moist components of enthalpy are of similar sizes, putting dry and moist components on the same footing for statistical modeling. Similar benefits can be achieved by normalising each θ and q component by its mean and variance, but using enthalpy units has the convenient property that the sum of $c_p\Delta\theta$ and $L\Delta q$ over levels is zero as enthalpy is conserved by convection.

Placing the $c_p\theta$ and Lq values for each of the m model levels into a single vector and combining n input cases, we form the $2m \times n$ matrix $\mathbf{X}$:

$$\mathbf{X} = \begin{pmatrix} c_p\theta_{1,1} & \cdots & c_p\theta_{m,1} & Lq_{1,1} & \cdots & Lq_{m,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_p\theta_{1,n} & \cdots & c_p\theta_{m,n} & Lq_{1,n} & \cdots & Lq_{m,n} \end{pmatrix}$$

We then find the matrix of eigenvectors, $\mathbf{U}$ ($2m{\times}2m$), and their corresponding weights, $\mathbf{P}$ ($2m \times n$), so that $\mathbf{X} = \mathbf{PU}$, by taking the singular value decomposition of the covariance matrix $\mathbf{X}^T\mathbf{X}$. Similarly, columns of $c_p\Delta\theta$ and $L\Delta q$ are combined to form the output matrix, $\mathbf{Y}$ ($2m{\times}n$), which is written in terms of its eigenvectors, $\mathbf{V}$ ($2m{\times}2m$), and their corresponding weights, $\mathbf{Q}$ ($2m \times n$), such that $\mathbf{Y} = \mathbf{QV}$, by taking the singular value decomposition of $\mathbf{Y}^T\mathbf{Y}$. The aim then is to predict unknown values of output $\mathbf{Q}$ and hence $\mathbf{Y}$ from known values of the inputs $\mathbf{P}$. We predict $\mathbf{Q}$ from $\mathbf{P}$ rather than predicting $\mathbf{Y}$ from $\mathbf{X}$ because correlations between values of θ and q on different vertical levels that could cause large errors in our statistical analysis are avoided. A two-step linear statistical emulator is used that first predicts whether convection is occurring, and then when convection is predicted to occur, predicts $\Delta\theta$ and Δq on model levels. The two-step choice is helpful because convection is a rare event even in the tropical atmosphere. It is difficult to represent large numbers of cases of no or little convection, and small numbers of cases of large convection simultaneously using a linear model. Two similar steps are also used by many convection schemes, including the ones analysed in this paper.

Whether or not convection occurs is predicted using logistic regression. For the ith case in $\mathbf{P}$, an estimate of the probability that convection will occur is

$$C_i = \frac{\exp(\beta P_i)}{1 + \exp(\beta P_i)}, \tag{1}$$

where β ($2m$ component vector) are coefficients to be determined, one for each input eigenvector. Nominally, convection is expected when $C_i > 0.5$ but experience with data can lead us to shift the decision boundary in practice. If it is predicted that convection is occurring, then $\mathbf{Q}$ and hence $\Delta\theta$ and Δq on levels are predicted from a linear model:

$$\mathbf{Q} = \gamma\mathbf{P} + \epsilon, \tag{2}$$

where γ ($2m \times 2m$) are the coefficients to be determined for each output eigenvector in terms of each input vector, and ϵ is the error. Both the β and γ coefficients are predicted using ridge regression, which is a constrained variant of ordinary least squares regression that penalises large components of β and γ via a tunable coefficient λ. For example, the best estimate of γ is

$$\hat{\gamma} = (\mathbf{P}^T\mathbf{P} + \lambda\mathbf{I})^{-1}\mathbf{P}^T\mathbf{Q}.$$

Providing λ is positive and non-zero, the analysis is not very sensitive to its precise value. We use $\lambda = 10$ for logistic regression estimates of convective triggering and $\lambda = 2$ for linear regression estimates of convective strength throughout. Ridge and related techniques such as the Bayesian lasso are powerful tools for constraining regression parameters when correlations between components of the input weights permit a large range of coefficients. That should not be an issue here because the singular value decomposition almost eliminates correlations in the training data. However, experience with convecting data shows that there is still the possibility that chance correlations between small features in the input data and significant features in the output data can lead to very large coefficients and large prediction errors when the statisical model is used to predict outputs for an unseen input dataset. The ridge models used avoid these problems because large coefficients are suppressed. More details of the logistic and ridge regression methods are given by, for example, Hastie et al. (2008).

Finally, the output matrix is estimated via $\mathbf{Y} \simeq \gamma\mathbf{P}\mathbf{V}$. The mean of the training data removed in the first step is added to $\mathbf{Y}$, yielding estimates of $\Delta\theta$ and Δq. Hence, the information in the training data is encoded into the eigenvectors $\mathbf{U}$ and $\mathbf{V}$ and the coefficients β and γ. The training data can then be discarded and the statistical models tested against unseen data to assess their accuracy.

3.2 Model training and truncation

The statistical emulators for the LLCS and GR parameterizations are trained for the tropics (30 °N – 30 °S) using output from the 30 day January and July simulations described in Section 2.1. These simulations output several million cases each, of which around 2–3 % show appreciable convection. (One case is one horizontal gridpoint at one timestep.) We calculate $\Delta\theta_{trop}$, which is defined as the mean atmospheric warming between 700 and 100 hPa for each case and then choose for training the cases closest to 30000 equally spaced values of $\Delta\theta_{trop}$ from its minimum to its maximum value. This attempts to build an emulator that is equally competent at representing the full range of convective events rather than the most common ones. The largest events are rare and therefore each is typically represented on multiple occasions in the training data. A further 30000 non-convecting cases (defined as $\Delta\theta_{trop} < 0.05$ MJ m^{-2}, although results are insensitive to the precise choice) are chosen at random. For each convection scheme, we compose control emulators, which take half of their input from control January and half from control July. For standard LLCS and GR we also compose combined control – $4\times CO_2$ emulators, which take a quarter of their input from each of control January, control July, $4\times CO_2$ January, and $4\times CO_2$ July. The combined emulators show only minor differences with their control counterparts, but are useful for running emulators online in the GCM and testing the behaviour of emulated convection under climate change. The choice of 60000 cases was made as it is realistic to perform analysis on matrices of this size with available computing resources. The effect of smaller sample size is investigated in Section 4.1.

The input and output eigenvectors **U** and **V** are calculated via singular value decomposition from the 30000 equally-spaced samples and their weights **P** and **Q** calculated for all 60000 equally-spaced and non-convecting cases for each convection scheme. The γ coefficients in equation 2 are estimated from the 30000 equally-spaced cases only. Cases from the equally-spaced group deemed non-convecting ($\Delta\theta_{trop} < 0.05$ MJ m^{-2}) are then discarded, as are an equal number of non-convecting cases, leaving us with an equal number of convecting and non-convecting cases. The β coefficients in equation 1 are estimated from the remaining cases. (Optimally, a different set of input eigenvectors that also consider the non-convecting sample would be calculated to remove correlations between components of **P** when considering non-convecting data. However, in practice, the very slight benefit of doing this is outweighed by the tractability of using one set of eigenvectors.) Both β and γ are fitted using the scikit-learn python package (see Acknowledgements). The fidelity of the emulator is tested using a dataset independent from the training data.

There is then the option of truncating the matrices to improve interpretability. The eigenvectors are ordered by the proportion of variance that they represent in the training data, each representing the largest remaining fraction of variance possible after variance associated with the previous eigenvectors has been removed. In our aquaplanet simulations, it is found that the vast majority of output convective behavior can be described with relatively few eigenvectors. We typically retain two or three for discussion in the results sections and use ten when the emulators are run online as part of the GCM. Truncating the input space, on the other hand, is difficult to do because convection is a rare event and successfully predicting its occurrence and strength relies on retaining small signals in the input data. Hence, instead of truncating, we rotate β and γ back into θ,q on levels by forming the $2m$ component vector $\beta_{\theta,q} = \beta\mathbf{U}$ and the $2m \times 2m$ matrix $\gamma_{\theta,q} = \gamma\mathbf{U}$ to interpret our results. $\beta_{\theta,q}$ is the sensitivity of convective triggering to departures of θ,q from their mean values on each level. The columns of $\gamma_{\theta,q}$ are the sensitivity of each output mode $\mathbf{V}$ to departures of θ,q from their mean values given that convection is occurring. Having identified $\beta_{\theta,q}$ and $\gamma_{\theta,q}$, a new low dimensional input space that does preserve the input signals necessary to describe convection can be built. We demonstrate this in Section 4.2 and show its use for comparing multiple convection schemes.

For Cascade, after coarse-graining to 2.5° longitude by 2° latitude, only 35952 cases are available over the selected Indian Ocean region, with only 4030 showing appreciable convection with $\Delta\theta_{trop} > 0.05$ MJ m^{-2}. At 11 %, this is much more frequent than the 2–3 % of cases seen to convect in the GCMs. Nevertheless, the small amount of data available forces a change in our experimental design. The emulator is trained on 2000 cases and then evaluated for the entire dataset including the training data. The training set contains the majority of deep convecting cases, so our assessment of our ability to emulate convection should be considered preliminary as the test dataset lacks substantial independence.

4 Results

4.1 Emulation of LLCS, GR and Cascade control data

This section presents results when statistical models are fitted for the first 10 components of the output modes, $\mathbf{V}$, for each representation of convection. The most important modes of response for the control CO_2 LLCS CON, GR CON and Cascade runs

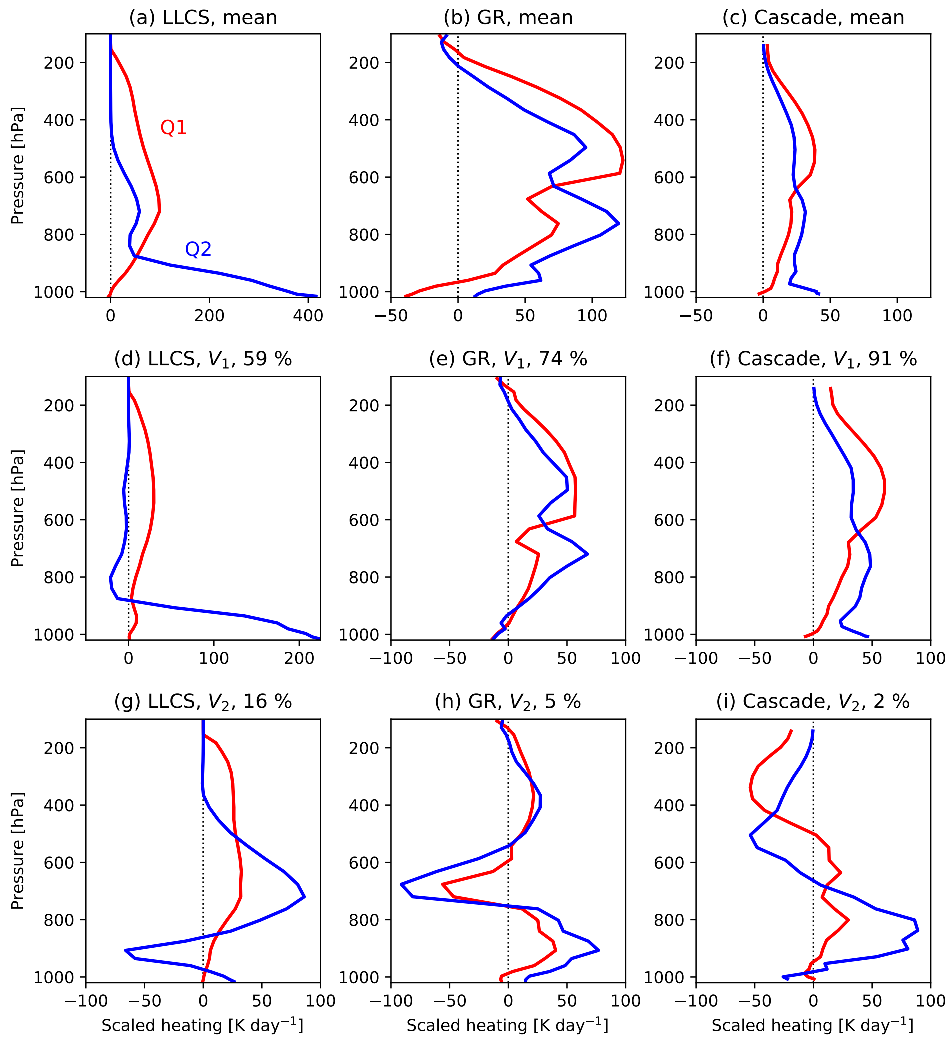

Figure 1. Main modes of convective response for LLCS (left column), GR (middle) and Cascade (right column) control cases. The top row shows the mean responses when convection is occurring, the middle row shows the first eigenvectors describing variations in convective response across the training data, V_1, and the bottom row shows the second eigenvectors, V_2. Red lines are the effective convective heating rate, Q1, and blue lines are the effective convective drying rate, Q2. Percentages in the titles of panels d-i are the proportion of output variance accounted for by each component of $\mathbf{V}$. Both are shown in temperature units of K day^{-1}, where Q2 corresponds to the latent heat of condensation associated with drying. Note the different horizontal scales in each panel.

are depicted in Figure 1. Shown are the mean convective responses when convection is occurring (defined as $\Delta\theta_{trop} > 0.05$ MJ m^{-2}, equivalent to a temperature change $\Delta T_{trop} \simeq$17.4 K day^{-1}) (Figure 1a-c), and the first and second eigenvectors, V_1 and V_2, that describe how convecting cases vary across the training data (Figure 1d-i). Physically, the mean responses and V_1 are identified with deep convection. For LLCS, positive V_2 is associated with stronger convection and more heating higher in the troposphere. For GR and Cascade, positive V_2 is associated with shallow convection. The first two components of $\mathbf{V}$ account for 75 % of the variance in the convecting training data for LLCS, 79 % for GR and 93 % for Cascade. In units of enthalpy change, it is found that the combined sum over vertical levels of dry and moist components of $\mathbf{V}$ is near zero for LLCS and GR, meaning that enthalpy is conserved by the convection schemes as expected. Agreement is less good for Cascade, which is unsurprising given that occurrence of convection is estimated rather than calculated explicitly. Figure 2a-c shows the corresponding mean input associated with the mean convecting case for each model control run. Figure 2d-i shows the rotated $\gamma_{\theta,q,1}$ and $\gamma_{\theta,q,2}$, which are the variations from the mean input necessary to achieve variations of size V_1 and V_2 from the mean output. Also shown are the range of responses for 1000 subsamples of the training data where 10000 cases are chosen at random without replacement and γ is recalculated (1000 subsamples of 1000 cases for Cascade). Evidently, our calculations are likely to be affected by sampling errors, especially near the surface and especially for Cascade. Results for $\beta_{\theta,q}$, which control convective triggering, are similar to $\gamma_{\theta,q,1}$ (in other words deep convection) in each case, so we omit them for brevity.

Taking Figures 1 and 2 together we can identify clear differences between the convection that occurs in the different datasets. Analysing the mean and deep convective components, V_1, it is plain that LLCS consumes far too much boundary layer moisture (Figure 1a,d) compared with GR (Figure 1b,e). LLCS convection occurs when the atmosphere is cooler and drier than GR (Figure 2a-c) and strengthens as the surface layer becomes wetter and warmer than those aloft (Figure 2d). This is in contrast to GR, where deep convection also relies on a warm atmospheric boundary layer (Figure 2e). It is more difficult to make similar arguments using the Cascade dataset perhaps due to the small sample size and limitations of the input data, Section 2.2. However, the convection realised is similar to GR if apparently weaker, although this may be due to the temporal and spatial averaging undertaken (Figure 1c,f,i).

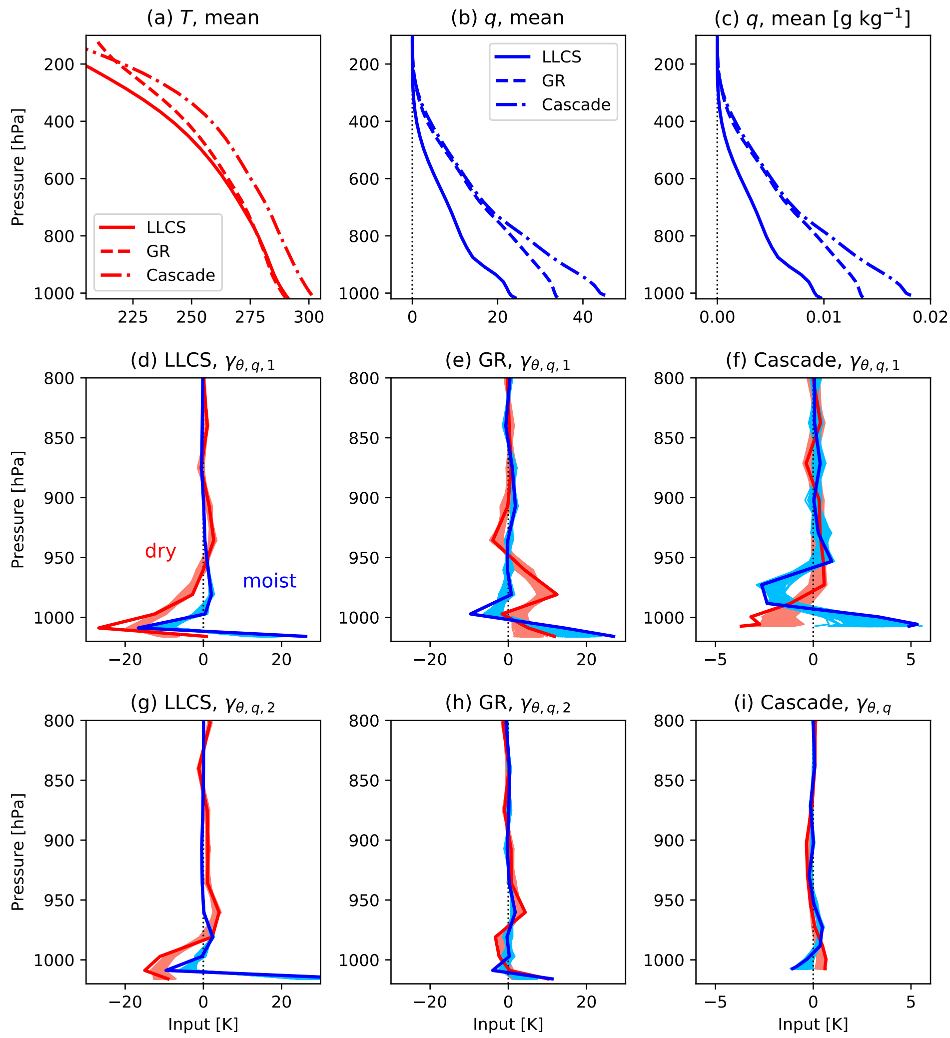

Figure 2. Inputs associated with convecting cases in the training data. The top row shows mean atmospheric profiles: (a) Temperature, T, (b-c) q. The middle and bottom rows show how anomalies from the mean input drive changes in (middle row) V_1 and (bottom row) V_2. Dry enthalpy components are red, moist components are blue. The lighter red and blue shading depicts the range of $\gamma_{\theta,q}$ for the 1000 subsampled training cases. All are shown in temperature units of K where moist components are expressed in K via the latent heat of vaporisation associated with q, as in Figure 1. The exception is panel (c) where mean q is shown in g kg^{-1}. Note the different horizontal scales on each panel and that the vertical scale only shows the region 800-1000 hPa for panels d-i. (Signals for 100-800 hPa are small on these panels.)

Table 2. Results for the LLCS and GR independent datasets and the Cascade complete dataset, which includes the training cases, for $C = 0.6$. For LLCS and GR, results are given for emulators of the control (CON) simulations and for emulators of the combined control and $4\times CO_2$ simulations. The "Convecting" and "Non-convecting" columns are the percentages and number of cases correctly identified as convecting and not convecting respectively in the simulations. R^2 is the coefficient of determination for $\Delta\theta_{trop}$ for all convecting cases (including those labelled as non-convecting by the emulator).

Simulation	Convecting	Non-convecting	R^2
LLCS CON	77 % (87584/114299)	86 % (4197723/4862341)	0.65
LLCS CON & $4\times CO_2$	76 % (88434/116200)	86 % (4203611/4860440)	0.66
LLCS CON $r_c = 0.7$	72 % (74204/102363)	86 % (4197361/4874277)	0.69
LLCS CON $r_c = 0.9$	70 % (70638/101226)	88 % (4283043/4875414)	0.65
GR CON	81 % (73430/90419)	95 % (2295750/2404466)	0.47
GR CON & $4\times CO_2$	79 % (78395/98843)	95 % (2275794/2396850)	0.50
Cascade	74 % (3035/4092)	91 % (28887/31860)	0.17

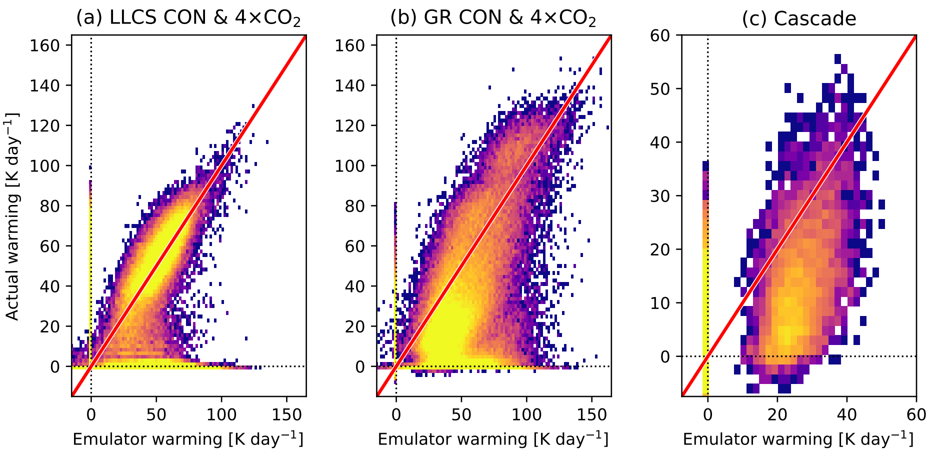

Figure 3. ΔT_{trop} for simulated versus emulated cases in the independent datasets for CON and $4\times CO_2$ (a) LLCS and (b) GR and for the complete dataset for (c) Cascade. Lighter colors indicate a higher density of cases. The red line is $y = x$.

We now test the ability of our statistical models to reproduce convection simulated in independent datasets not used for fitting. (Due to the small amount of data available, results for Cascade include the training data, so these results should be treated as preliminary.) Summary statistics for $C = 0.6$ are shown in Table 2. Overall, prediction of whether or not convection should trigger (defined as where $\Delta\theta_{trop} > 0.05$ MJ m^{-2}) is quite good, especially for GR. It is also the case that predicting both CON and 4$\times$CO$_2$ cases using one emulator does not substantially degrade performance for either LLCS or GR. At first sight, percentage results are particularly encouraging for non-convecting cases. However, because non-convecting cases are by far the majority of all cases, the number of cases that would be incorrectly predicted to convect is high. This could pose problems when using the emulator online in a GCM. The proportion of non-convecting cases correctly predicted can be increased by increasing the value of C. However, this increases the number of convecting cases that are incorrectly predicted to be non-convecting. Experience shows that $C = 0.6$ provides a balance between the convecting and non-convecting prediction errors that gives reasonable results when run online in the GCM (Section 4.3). Figure 3 shows the performance of the emulator in predicting ΔT_{trop}. Values of R^2 for convecting cases are given in Table 2. Predictions for LLCS are most accurate, followed by GR. Predictions for Cascade are weaker and show poor R^2.

4.2 Joint analysis of LLCS, GR and Cascade control data

This subsection presents LLCS, GR and Cascade emulators built in terms of a common set of input, $\mathbf{U_C}$, and output, $\mathbf{V_C}$, eigenvectors, allowing direct comparison of values of β and γ that determine convective response to a given input. We build our joint input and output spaces from the combined LLCS CON and GR CON training data. Common output eigenvectors, $\mathbf{V_C}$, and their corresponding weights, $\mathbf{Q_C}$, are derived from the singular value decomposition of 60000 equally-spaced cases taken from the relevant January and July training runs. (We use control data because this is the only data available for the LLCS perturbed physics versions we will consider.) This is sufficient to capture the dominant behavior of convection in a few modes, as with the individual decompositions in the previous subsection. Because large numbers of input modes are important to convection in each dataset, we derive the common input modes in a slightly different way in order to obtain a small tractable set. For LLCS CON and GR CON, we form $\gamma_{\theta,q,1-3}\mathbf{X}$, where $\gamma_{\theta,q,1-3}$ are the first three regression coefficients linking anoma-

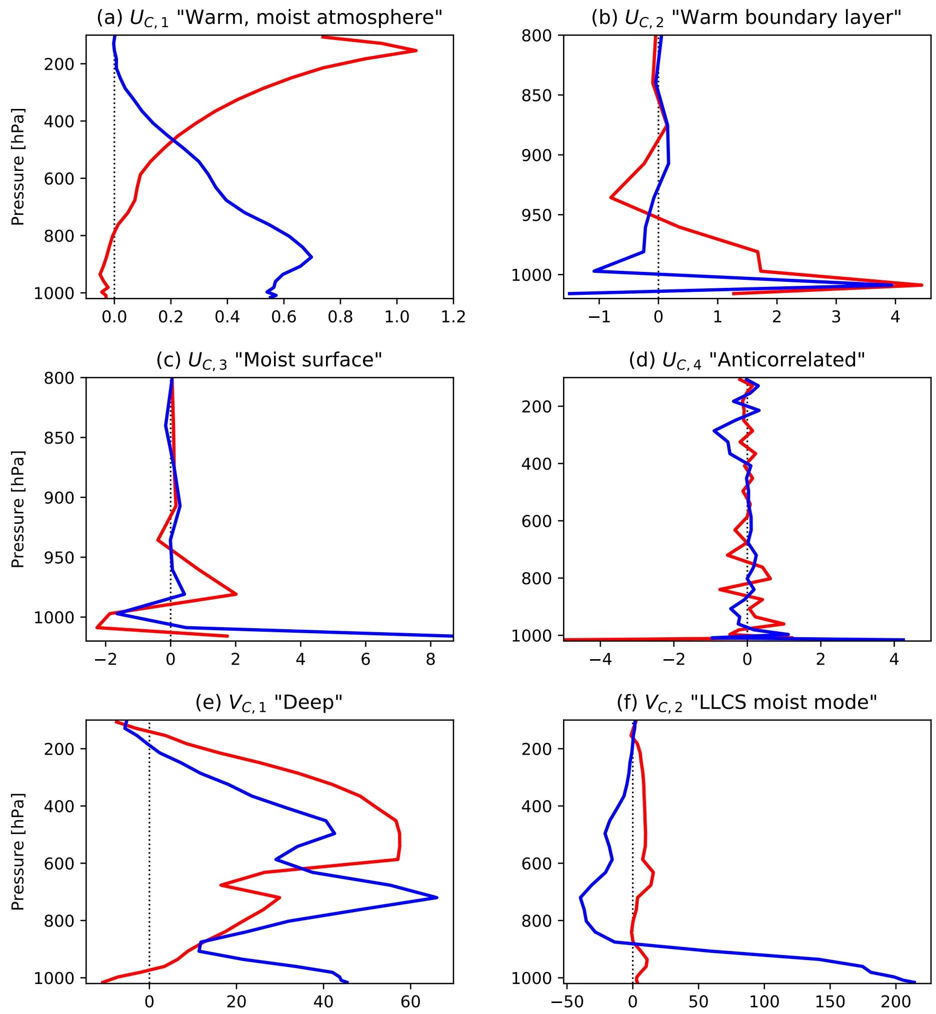

Figure 4. (a-d) The first four joint input eigenvectors for the LLCS CON and GR CON datasets. As in Figure 2, dry temperature components are red, moist components are blue in units of K. (e-f) The first two joint output eigenvectors. As in Figure 1, the red lines are the effective convective heating rate, Q1, and the blue lines are the effective convective drying rate, Q2, in units of K day^{-1}. Note the different horizontal and vertical scales on each panel, in particular the vertical scales for (b) and (c), which show the boundary layer only.

lies in θ, q inputs $\mathbf{X}$ to anomalies in outputs $\mathbf{Y}$. New θ,q input datasets containing only those data determined to be linked to convection are then written $\gamma_{\theta,q,1-3}^T \gamma_{\theta,q,1-3}\mathbf{X}$. Finally, concatenating the 30000 LLCS and 30000 GR equally-spaced training cases, we apply singular value decomposition one more time and arrive at combined input eigenvectors, $\mathbf{U_C}$. The calculation is done with respect to a common mean for the LLCS and GR datasets, allowing analysis of the effects of differences between the basic states of the two datasets.

Figure 4 shows the most important first four components of $\mathbf{U_C}$, and the first two components of $\mathbf{V_C}$. Because the analysis is done with respect to a common mean, $U_{C,1}$ reflects the warmer, moister atmosphere found when convection is occurring in GR compared with LLCS. $U_{C,2}$ is a mode with a warm boundary layer and moist near surface, $U_{C,3}$ is a very moist surface mode and $U_{C,4}$ is difficult to interpret physically but has strongly anticorrelated dry and moist components. The first output, $V_{C,1}$, is a deep convective mode similar to that seen in the individual GR decomposition; the second output, $V_{C,2}$, describes large near surface drying similar to that seen for deep convection in LLCS. $V_{C,1}$ accounts for 21 % of the output variance of LLCS, 63 % in GR and 84 % in Cascade; $V_{C,2}$ accounts for 51 % of the output variance of LLCS, 5 % in GR and 1 % in Cascade.

Statistical models that describe $\mathbf{V_C}$ in terms of $\mathbf{U_C}$ are then composed for all control training datasets, including the LLCS $r_c = 0.7$ and $r_c = 0.9$ cases. First, we estimate values of β and γ for the individual datasets as before using their original input weights, $\mathbf{P}$, to take advantage of their orthogonality, but using the common LLCS-GR outputs, $\mathbf{Q_C}$. β and γ are then rotated into the $\mathbf{U_C}$ basis by taking $\beta_C = \mathbf{U_C^T U}\beta$ and $\gamma_C = \mathbf{U_C^T U}\gamma$. Projecting the statistical models into the truncated rotated basis reduces their fidelity. The proportion of convecting and non-convecting cases correctly predicted in an independent dataset is altered to 26 % and 53 % respectively for LLCS, 84 % and 87 % for GR and 40 % and 71 % for Cascade. R^2 is reduced to 0.62 for LLCS, 0.46 for GR and 0.01 for Cascade. Hence, the rotated basis retains the ability to predict changes in convective strength in LLCS and GR presumably because these are the eigenvectors it is built from, but most other predictions are damaged, especially for triggering. Note, however, that the degradation depends on the truncation chosen. Using a larger set of eigenvectors would increase fidelity at the expense of tractability. The choice depends on the application.

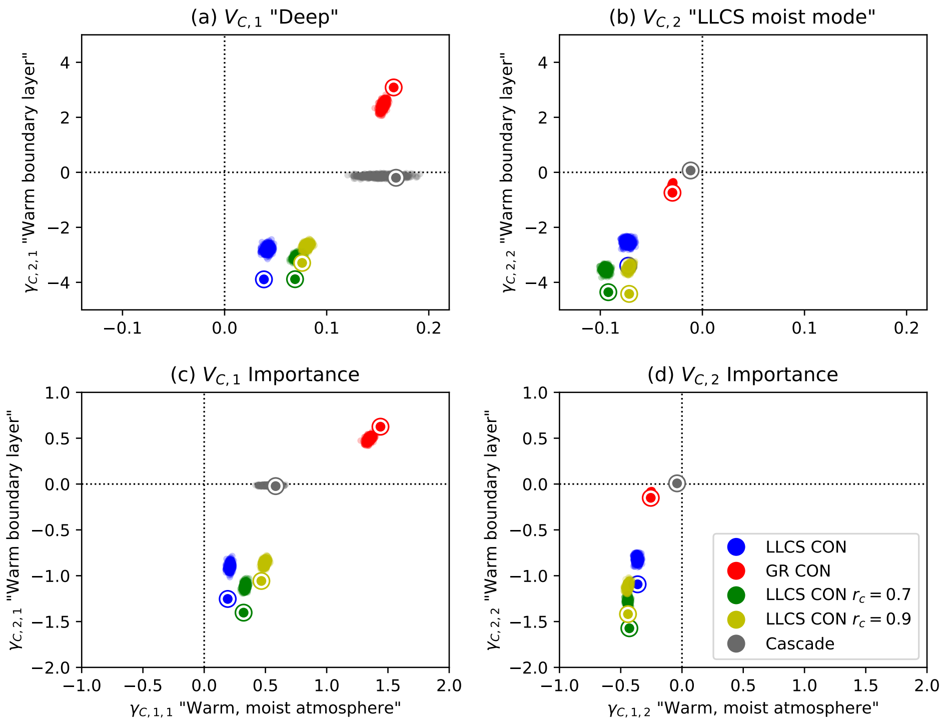

Figure 5. Sensitivity of the first two joint output modes to the first two joint input modes. (a-b) Components of γ_C. (c-d) Components of γ_C multiplied by the standard deviation of the corresponding $\mathbf{U_C}$ component, demonstrating the typical sizes of change in each component of $\mathbf{V_C}$ caused by each component of $\mathbf{U_C}$. The large bullseye circles are for the full training set of 30000 cases for each model. The spreads of smaller points are where 10000 samples have been taken for the LLCS and GR simulations and 1000 samples have been taken for Cascade.

Values of γ_C that link the first two input and output eigenvectors are shown in Figure 5a,b. The effect on convection of the "warm, moist atmosphere" mode, $U_{C,1}$, per unit anomaly is weak, but its standard deviation across the training data is large, and so it plays an important role in increasing the strength of convection in all simulations through $V_{C,1}$. We judge this through "importance", which we define as a given component of γ_C multiplied by the standard deviation of the relevant component of $\mathbf{U_C}$. For all LLCS model variants, increasing $U_{C,1}$ also reduces $V_{C,2}$, reducing boundary layer drying and enhancing drying aloft. Neither GR nor Cascade show this mode very strongly, and $V_{C,2}$ is therefore not sensitive to the presence of $U_{C,1}$ or $U_{C,2}$ in their input data. Increasing the "warm boundary layer" mode, $U_{C,2}$, increases $V_{C,1}$ in GR but reduces $V_{C,1}$ in all LLCS versions. The Cascade data are largely insensitive to $U_{C,2}$. Components of $\mathbf{U_C}$ beyond $U_{C,2}$ have lower importance and contribute less to convection and intermodel difference. However, both LLCS and GR $V_{C,1}$ respond positively to the "moist surface mode", $U_{C,3}$ (not shown).

The LLCS perturbed physics versions, $r_c = 0.7$ and $r_c = 0.9$ show very similar sensitivities to standard LLCS, so we do not discuss them in detail. However, we note that zonal mean precipitation produced by LLCS $r_c = 0.9$ is more similar to GR than standard LLCS (Figure 7b). Changes in zonal mean precipitation under $4\times CO_2$ warming are more like standard LLCS, however (Figure 7c). The model simulations make it clear that LLCS can be tuned to reproduce GR zonal mean precipitation satisfactorily. However, the rotated basis shows that the fundamental sensitivities of LLCS to input are little altered by changing r_c, and it is therefore not necessarily a surprise that the climate change simulation is not improved.

Figure 6 is a comparison of the predicted convective triggering probability, β, with the actual amount of convection realised for 60000 cases from the independent datasets for LLCS CON and GR CON. Using the same method used to choose the original training data, we select 30000 convecting cases that represent the range of mean tropospheric heating and 30000 non-convecting cases at random. (Using the entire dataset swamps the parameter space with non-convecting cases, even where the percentage error in predicting the occurrence of convection is small because the number of non-convecting cases is so large, Table 2). The two-dimensional slices show which parts of the $\mathbf{U_C}$ parameter space defined by the corresponding weights $\mathbf{P_C}$ are expected to experience convection. The black idealised contours are values of β when varying components of $\mathbf{P_C}$ within the relevant plane but holding others at mean values. The blue contours are for the in-

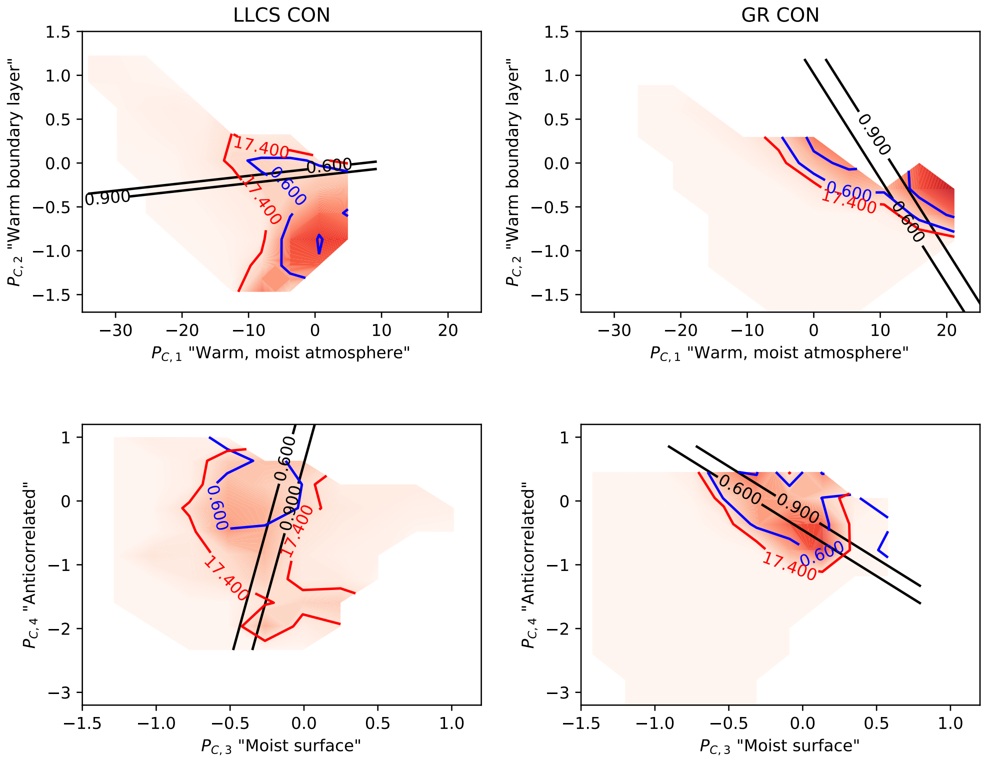

Figure 6. Convective triggering predictions compared with true simulated tropospheric warming as a function of $\mathbf{P_C}$ for 60000 cases (including 30000 convecting) from the independent datasets. Planes in the $\mathbf{P_C}$ parameter space for (top) $P_{C,1,2}$ and (bottom) $P_{C,3,4}$ for both (left) LLCS CON and (right) GR CON. Black contours are the predicted probability of triggering convection, C, when varying components of $\mathbf{P_C}$ in the plane but holding others at mean values. $C = 0.6$ is the threshold used for triggering convection in our UM simulations. The 0.9 contour is also shown to indicate which side of the 0.6 contour is expected to trigger. The blue 0.6 contours are predictions of β where all components of $\mathbf{P_C}$ are allowed to vary. Red contours and accompanying shading are values of simulated ΔT_{trop} in K day^{-1}. In our analysis the threshold for convection is $\sim$17.4 K day^{-1}. For a perfect prediction of convective triggering, the blue contour would overlay the red contour.

dependent dataset when all components of $\mathbf{P_C}$ are allowed to vary. The blue and black contours are not coincident because components of $\mathbf{P_C}$ are correlated, which occurs because the relevant singular value decomposition was done for the combined LLCS-GR training dataset and not for LLCS or GR individually.

Results for the triggering and strength of convection are complementary within the $\mathbf{U_{C,1,2}}$ plane. $P_{C,1}$ varies strongly across both the LLCS and GR datasets. More positive values of $P_{C,1}$ are associated with more triggering of convection and stronger convection in GR. In LLCS stronger convection is associated with more positive $P_{C,1}$, but its effect on triggering is apparently small (black contours) but confounded by correlations with other components of $\mathbf{P_C}$ in practice (blue contours). As with the strength of convection, the effect of $P_{C,2}$ is markedly opposite for convective triggering in GR and LLCS: more positive values of $P_{C,2}$ trigger convection in GR but suppress it in LLCS. GR responds positively to increases in both $P_{C,3}$ and $P_{C,4}$. The response of LLCS is more confused. The central region of the $P_{C,3,4}$ plane is convecting (red contours) but this is not expected purely from varying $P_{C,3}$ and $P_{C,4}$ (black contours). Correlations with other components are required.

Overall, our joint analysis shows clear differences between the different convection schemes that can be understood in simple terms. Compared with GR, LLCS condenses too much boundary layer moisture, is relatively insensitive to an important mode of warm-moist free atmosphere variation and has the wrong sign of response to boundary layer warming. This suggests pathways via which LLCS might be improved: adjustment of the scheme's ability to bring unsaturated parcels from the boundary layer into moist convection aloft could reduce boundary layer moisture consumption; a simple representation of entrainment could improve interaction with the free atmosphere. It is interesting to note that values of γ_C for Cascade have some similarity with GR, but this must be treated with caution given that the rotated model describes the Cascade data poorly. The results of this section are largely clear from the individual analyses of Section 4.1. The purpose of our example, however, is to demonstrate a low-dimensional parameter space that could express key differences between a large number of representations of convection concisely.

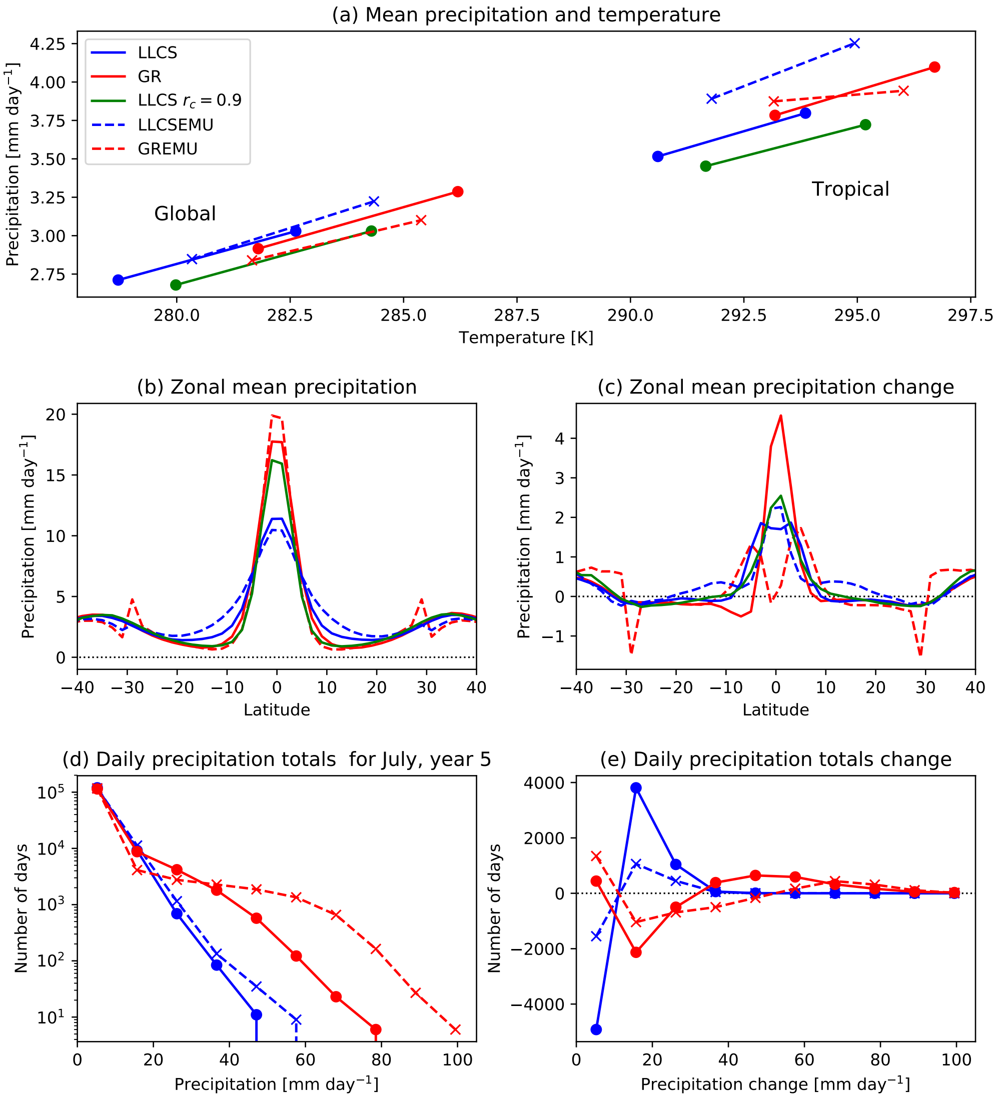

Figure 7. (a) Global and tropical mean of last five years for precipitation and temperature for control and 4×CO_2 conditions. In all cases the 4×CO_2 simulation point is above and to the right of the control simulation point. (b-e) Precipitation where convection is simulated using the original and emulated parameterizations for LLCS and GR. (b) Last five year zonal mean precipitation for the control simulations for 40 °N – 40 °S. (c) Last five year zonal mean 4×CO_2 - control precipitation change. (d) Histogram of gridbox daily precipitation totals for control July, year 5 for 30 °N - 30 °S. (Note logarithmic vertical scale on this panel.) (e) Histogram of gridbox daily precipitation 4×CO_2 - control change for July, year 5.

4.3 UM simulations with emulated convection

10 year control and $4\times CO_2$ UM simulations with emulated convection were run for GR and LLCS $r_c = 0.8$ using the combined control – $4\times CO_2$ emulators (LLCSEMU CON and $4\times CO_2$, and GREMU CON and $4\times CO_2$, Table 1). All simulations have a stable equilibrium climate and reproduce broad features of the original parameterized runs (LLCS CON and $4\times CO_2$, and GR CON and $4\times CO_2$) with reasonable fidelity. Figure 7a shows values of global and tropical mean precipitation and temperature in the original and emulator parameterization simulations. Control emulator simulations are biased with respect to the corresponding original simulations in the global mean by 1.6 K and 0.1 mm day^{-1} for LLCSEMU, and -0.1 K and -0.1 mm day^{-1} for GREMU. Tropical mean biases are 1.2 K and 0.4 mm day $^{-1}$ for LLCSEMU and -0.04 K and 0.1 mm day^{-1} for GREMU. $4\times CO_2$ - control climate change is quite well-simulated in LLCSEMU. Climate change is more disappointing for GREMU, particularly in the tropics, where precipitation increases at only 0.7 % K^{-1} tropical mean temperature change compared with original GR values of 2.3 % K^{-1}.

More detailed precipitation statistics are shown in Figure 7b-e. Zonal mean precipitation in the LLCSEMU and GREMU control runs is quite reasonable and clearly captures the difference between LLCS and GR (Figure 7b). $4\times CO_2$ - control zonal mean changes are fair for LLCSEMU, but disappointing for GREMU (Figure 7c). The sharp features in panels b and c seen at 30 °N – 30 °S in LLCSEMU and GREMU occur because the convection emulator is switched off poleward of 30°. Results for convective precipitation only are very similar (not shown). Figure 7d and e are histograms of gridbox total daily precipitation for July in year 5 of the simulations and $4\times CO_2$ - control changes. LLCSEMU totals are satisfactory, while GREMU tends to predict too many heavy precipitating events and too few light precipitating events. $4\times CO_2$ - control changes show the correct sense of change for both LLCSEMU and GREMU: more lighter events tend to occur in LLCS, while heavier events increase at the expense of lighter events in GR. The emulated changes tend to be too weak for both LLCSEMU and GREMU, however, particularly for lighter events.

Overall, the online LLCSEMU and GREMU results are encouraging. The model is stable and equilibrium climate is close to LLCS and GR, although LLCSEMU rains too much in the subtropics and GREMU has too many heavy precipitation days. Climate change simulations are reasonable, although changes in zonal mean precipitation in GR are disappointing and changes in daily precipitation totals are too weak in both models.

5 Discussion

Our analysis achieves each of goals 1–3 set out for CSP in the introduction to at least some degree. We have demonstrated that statistical emulators of two GCM convection schemes and a high-resolution dataset can have skill in predicting the onset and magnitude of atmospheric convection. The representation is quite approximate, but could surely be improved. To form a CSP, a framework need only provide a structure that represents a group of parameterizations and the differences between them smoothly and unambiguously by providing as much orthogonality between modes as possible. A straightforward improvement to our CSP would be to introduce higher order and cross terms into the regression calculations using discrete orthogonal polynomials. We could also introduce more variables into the analysis, although we note that past work has found θ and q to be satisfactory for analysing both model output and observed effects of convection (Yanai et al., 1973; Johnson et al., 2016; Mapes et al., 2019). Another framework entirely is evolutionary genetic programming, which uses Darwinian evolution to produce models from combinations of simple functions (e.g. (Makkeasorn et al., 2008)).

We also showed that a rotated, reduced input space allows us to describe the most important differences between different representations of convection more easily and might assist in future model development. Care must be taken in the analysis as the reduced input and output spaces lose skill in predicting aspects of convection. In our demonstration, representation of triggering was particularly affected perhaps because we built an input space based on modes known to control the strength of convection. There is a balance between emulator skill and tractability that is set by the degree of truncation of our input and output spaces. We may compose as many representations as we like, each optimised for a different purpose. A key advantage of our approach over others is that it is possible in principle to define eigenvectors that allow estimation of the relationship between the most important inputs and outputs without contamination from linear correlations between variables. A good basis for many applications might be derived from observations or high-resolution simulations that explicitly resolve convection. We did not attempt this because the Cascade high-resolution dataset was small and our ability to represent it was limited. The inaccuracy of our Cascade emulator may stem from a combination of having too few cases to fit to and the fact that different parameterization schemes were used in the original CASCADE simulations and the SCM experiments. Alternatively, it may come from fundamental limitations of our technique. Defining the convection that should be parameterized and separating it cleanly from other processes is difficult in resolved simulations. It is even more challenging for observations as explored by Mapes et al. (2019) for the impulse-response method, although their analysis did yield useful conclusions regarding the sensitivity of observed versus resolved simulation of convection to q.

While one major goal for CSP is to develop metrics for model development, another is to develop emulators with sufficient fidelity that they can be run within a GCM. Success in this goal would mean that the emulators reproduce their targets well enough that we might explore the parameter space of possible parameterization schemes online within a GCM. Our GCM simulations that run the LLCS and GR emulators interactively show stable equilibrium climates with broadly similar characteristics to GCMs run with the original parameterizations. This is encouraging. Nevertheless, some aspects of the CSP emulator performance are disappointing, particularly for climate change where emulator simulations tend to respond too weakly. Performance is certainly weaker than that achieved with the random forest technique of O'Gorman and Dwyer (2018). Random forest or other machine learning representations of a range of convection schemes may themselves be analysed with a linear model, but our complete emulators have an unambiguous relationship with each other and with the results they achieve when applied within a GCM. Hence, further work that improves our emulators would be useful.

Our emulators are deterministic – a given input always leads to the same output. However, a body of recent work suggests that performance can be improved in some cases through making parameterizations "stochastic" by adding noise to the parameterization output (e.g. Lin & Neelin, 2003; Plant & Craig, 2008). It is trivial to introduce this extension to our statistical emulators by perturbing their parameters. When applied to a high-resolution model or observed dataset, CSP is also well-adapted to discovering the range of outputs that occur for a given set of coarse-grained inputs, potentially providing new routes to building stochastic parameterizations.

If a future study is to build emulators good enough to probe the effect of the range of possible convective parameterization on gross features of future climate change, then it needs to engage with clouds and cloud radiative effects. A move to a more realistic land and ocean configuration may not be necessary in the first instance, however, as it has been demonstrated that global mean temperature sensitivity to increased atmospheric CO_2 concentration in comprehensive land-ocean-atmosphere GCMs is well-related to that in corresponding aquaplanet simulations (Ringer et al., 2014).

6 Conclusion

Using the example of convection, we describe Continuous Structural Parameterization (CSP), which is a method for writing different representations of the same subgridscale process as functions of the same gridscale variables. It is found that CSP can represent two convection schemes implemented within the Met Office Unified Model (UM) with reasonable fidelity. When emulated convection is implemented within the UM, the GCM produces a stable equilibrium climate with features broadly similar to the case where the original convection scheme is used.

Using our CSP, key differences between parameterization schemes can be expressed concisely within a new parameter space that is agnostic to model structure and offers the possibility of comparison with high-resolution models of convection or observations. Here, a CSP representation of a high-resolution dataset taken from the Cascade experiment has some success, even though the dataset is small and not optimally designed for our purposes. Further CSP development is necessary and a large high-resolution dataset designed specifically for emulation is needed to produce cleaner results. Nevertheless, our work suggests that CSP can assist parameterization development both by indicating realistic areas of the relevant parameter space and by providing parameterization prototypes directly. Our long term goal is that CSP can assist ensemble prediction of climate change by highlighting how the set of model parameterization we have relate to our true uncertainty in physical processes.

Appendix A Lambert-Lewis Convection Scheme

The Lambert-Lewis Convection Scheme (LLCS) is a simple but flexible adjustment scheme that has been used for simulating the atmospheres of terrestrial planets and for testing new GCM versions at the Met Office. LLCS has similarities to the simplified Betts-

Miller scheme (Betts, 1986; Frierson, 2007), but also some significant differences. In contrast to Betts-Miller, triggering of convection is based on dry and moist stability arguments, and purely dry convection with no condensation is possible. The scheme first evaluates whether or not convection should be triggered in a given model vertical column, then constructs new preliminary "plume" vertical profiles of θ and q in which convective instability is removed, before applying an adjustment timescale that relaxes the entire vertical column towards the new state while conserving enthalpy and moisture.

A1 Triggering

Starting from the surface, LLCS searches for the lowest unstable model level, k. Dry triggering occurs if $\theta_{k+1} < \theta_k$, meaning that a test parcel from level k perturbed upwards to level $k+1$ would find itself to be less dense than its surroundings and be expected to rise. Moist triggering occurs if $r_c q_{sat,k} < q_k$, where r_c is the critical relative humidity parameter, and q_{sat} is the saturation specific humidity. In this case, a test parcel on level k is expected to saturate in-situ, leading to condensation and convective heating. Normally, $r_c < 1$ meaning that the criterion is satisfied when the atmosphere is unsaturated at gridscale. The rationale is that a model column whose mean specific humidity is $r_c q_{sat,k}$ will contain some supersaturated regions able to trigger convection. $r_c = 0.8$ is the default value. If the dry trigger is satisfied but the moist trigger is not, then moist convection can still be triggered on a higher level, l, if $r_c q_{sat,l} < q_k$. This occurs if the triggered dry convective event reaches level l. The value of $q_{sat,l}$ used is that before any dry convective adjustments have taken place.

A2 Convective adjustment

Once convection is triggered, a preliminary profile is established through convective adjustment. Where dry convection is triggered, θ_{k+1} is adjusted so that the preliminary value of θ_{k+1}, $\theta_{k+1,p} = \theta_k$. Dry convection continues upwards providing that the new value of $\theta_{k+1,p}$ satisfies $\theta_{k+2} < \theta_{k+1,p}$. Moisture is mixed upwards by setting $q_{k+i,p} = q_k$, where i is the ith level above k.

If moist convection is triggered on level k, then levels above k involved in the convective event are adjusted to the moist pseudoadiabat:

$$\Gamma_{ps} = \frac{g(1+r_v)\left(1+\frac{Lr_v}{R_d T}\right)}{c_p + r_v c_{pv} + \frac{L_v^2 r_v(\eta + r_v)}{R_d T^2}},$$

where r_v is the mass-mixing ratio of water vapor, L is the latent heat of vaporisation of water vapor, R_d is the gas constant for dry air and $\eta \simeq 0.622$ is the ratio of the dry air and water vapor gas constants. Preliminary q is set to its saturation value, $q_k = q_{sat,k}$, on each level that moist convection is occurring including the bottom level unless $q_{k+i,p} > q_{sat,k+i,p}$. Similar to dry convection, moist convection continues upwards if the new value of $\theta_{k+1,p}$ derived from the pseudoadiabat satisfies $\theta_{k+i+1} < \theta_{k+i,p}$.

If the dry or moist convective event terminates below the highest model level, then subsequent levels are tested to determine whether another event can trigger in the same vertical column. Note that LLCS does not consider the freezing level and assumes that all condensation and precipitation is liquid.

A3 Relaxation timescale and conservation

Recognising that evolution to a new stable profile is not instantaneous, the original input θ and q are relaxed towards the preliminary values, θ_p and q_p via

$$\Delta\xi_r = (\xi_p - \xi)\left[1 - \exp\left(-\frac{t_{step}}{\tau}\right)\right],$$

where ξ_r represents either θ_r or q_r, subscript r corresponds to values after the relaxation timescale has been applied, t_{step} is the GCM timestep (1200 seconds in our experiments) and τ is a relaxation timescale, a free parameter of the scheme. The standard value used in our simulations is the "pure mixing timescale" of 3600 seconds of Tompkins and Craig (1998).

Moisture and enthalpy are then conserved within each separate convective event in the column. First, moisture is adjusted so that the total mass of water vapor within each convective event, $M_{q,r}$, is the same as in the input, M_q, less the amount of water condensed to produce latent heating, M_L, by adjusting specific humidity via

$$q_{f,k} = \left(\frac{M_q - M_L}{M_{q,r}}\right) q_{r,k},$$

where subscript f refers to final calculated values. M_L is outputted by the scheme as precipitation at the surface, thus conserving the moist component of enthalpy. This is done on all convecting levels of a given event including dry convection below the level at which condensation first occurs. Hence the scheme has the tendency to eliminate large amounts of boundary layer moisture, producing behavior that arguably should be simulated via the UM boundary layer scheme. This feature may be revised in future versions, but is probably useful for suppressing the occurrence of gridpoint storms.

Dry enthalpy must be conserved to take account of heat added to the column during dry adjustment. As for moist enthalpy, this includes all levels of convective events that begin as dry adjustments that then trigger moist events above the bottom level. For each level, implied dry heating is written $\Delta Q_d = M_k c_p (T_{d,k} - T_k)$, where M_k is the total mass of the level, T_k the initial temperature and $T_{d,k}$ is the implied temperature change if latent heating is neglected (equal to the entire convective adjustment for events with no moist component). The final temperature change ΔT_f is calculated by subtracting ΔQ_d from the relaxation value ΔT_r uniformly per unit mass:

$$\Delta T_f = \Delta T_r - \frac{c_p \Sigma_k \Delta Q_d}{\Sigma_k M_k}.$$

Final output θ_f is calculated via

$$\theta_f = \theta_k + \Delta T_f \left(\frac{p_0}{p} \right)^{\kappa},$$

where $p_0 = 1000$ hPa and $\kappa = \frac{R_d}{c_p}$.

Acknowledgments

FHL thanks John C. H. Chiang and the University of California, Berkeley Department of Geography for their hospitality and the use of their facilities during a period of study leave used to work on this project. We thank Steven Boeing and Jonathan Fieldsend for useful discussions. The scikit-learn package is available from *http://scikit-learn.org/*. This work was partly supported by Science and Technology Facilities Council Consolidated Grant (ST/R000395/1). All supporting data will be uploaded to the University of Exeter public repository Open Research Exeter before acceptance.

References

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

Betts, A. K. (1986). A new convective adjustment scheme. Part I: Observational and theoretical basis. *Q. J. R. Meteorol. Soc.*, *112*, 677–691.

Boutle, I. B., Drummond, B., Manners, J., Mayne, N. J., Goyal, J., Lambert, F. H., … Earnshaw, P. D. (2017). Exploring the climate of Proxima B with the Met Office Unified Model. *Astron. and Astrophys.*, *601*, A120. doi: 10.1051/0004-6361/201630020

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.*, *45*, 6289–6298. doi: 10.1029/2018GL078510

Christensen, H. M., Dawson, A., & Holloway, C. E. (2018). Forcing single-column models using high-resolution model simulations. *J. Adv. Mod. Earth Sys.*, *10*, 1833 1857. doi: 10.1029/2017MS001189

Collins et al. (2013). Long-term climate change: Projections, commitments and irreversibility. In T. Stocker et al. (Eds.), *Climate Change 2013: The Physical Science Basis* (pp. 1029–1136). Cambridge Univ. Press.

Frierson, D. M. W. (2007). The dynamics of idealized convection schemes and their effect on the zonally averaged tropical circulation. *J. Atmos. Sci.*, *64*, 1959–1976.

Geoffroy, O., Sherwood, S. C., & Fuchs, D. (2017). On the role of the stratiform cloud scheme in the inter-model spread of cloud feedback. *J. Adv. Mod. Earth Sys.*, *9*, 423–437.

Gregory, D., & Rowntree, P. (1990). A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure.

Guichard, F., Petch, J. C., Redelsperger, J., Bechtold, P., Chaboureau, J., Cheinet, S., … Tomasini, M. (2004). Modelling the diurnal cycle of deep precipitating convection over land with cloudresolving models and singlecolumn models. *Q. J. R. Meteorol. Soc.*, *130*, 3139–3172.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference and prediction.* Springer. (739 pp)

Herman, M. J., & Kuang, Z. (2013). Linear response functions of two convective parameterization schemes. *J. Adv. Mod. Earth Sys.*, *5*, 510–541. doi: 10.1002/jame.20037

Holloway, C. E., Woolnough, S. J., & Lister, G. M. S. (2012). Precipitation distributions for explicit versus parametrized convection in a large-domain high-resolution tropical case study. *Q. J. R. Meteorol. Soc.*, *138*, 1692–1708.

Johnson, R. H., Ciesielski, P. E., & Rickenbach, T. M. (2016). A further look at Q_1 and Q_2 from TOGA COARE. *Meteorological Monographs*, *56*. doi: 10.1175/AMSMONOGRAPHS-D-15-0002.1

Kelly, P., Mapes, B. E., Hu, I., Song, S., & Kuang, Z. (2017). Tangent linear superparameterization of convection in a 10-layer global atmosphere with calibrated climatology. *J. Adv. Mod. Earth Sys.*. doi: 10.1002/2016MS000871

Krasnopolsky, V. M. (2010). Accurate and fast neural network emulations of model radiation for the NCEP coupled climate forecast system: Climate simulations and seasonal predictions. *Mon. Weather Rev.*, *138*, 1822–1842. doi: 10.1175/2009MWR3149.1

Kuang, Z. (2010). Linear response functions of a cumulus ensemble to temperature and moisture perturbations and implications for the dynamics of convectively coupled waves. *J. Atmos. Sci.*, *67*, 941–962. doi: 10.1175/2009JAS3260.1

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

Makkeasorn, A., Chang, N. B., & Zhou, X. (2008). Short-term streamflow forecasting with global climate change implications – A comparative study between genetic programming and neural network models. *J. Hydrol.*, *352*, 336–354.

Mapes, B. E., Chandra, A. S., Kuang, Z., Song, S., & Zuidema, P. (2019). Estimating convection's moisture sensitivity: an observation-model synthesis using AMIE-DYNAMO field data. *J. Atmos. Sci.*, *76*, 1505–1520. doi: 10.1175/JAS-D-18-0127.1

Mauritsen, T., Stevens, B., Roeckner, E., Crueger, T., Esch, M., Giorgetta, M., ... Tomassini, L. (2012). Tuning the climate of a global model. *J. Adv. Mod. Earth Sys.*, *4*. doi: 10.1029/2012MS000154

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. J. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*, 768–772.

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *J. Adv. Mod. Earth Sys.*, *10*. doi: 10.1029/2018MS001351

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

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci.*, *115*, 9684–9689. doi: 10.1073/pnas.1810286115

Ringer, M. A., Andrews, T., & Webb, M. J. (2014). Global-mean radiative feedbacks and forcing in atmosphere-only and coupled atmosphere-ocean climate change experiments. *Geophys. Res. Lett.*, *41*, 4035–4042.

Sanderson, B. M. (2011). A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J. Clim.*, *24*, 1362–1377.

Sexton, D. M. H., Karmalkar, A. V., Murphy, J. M., Williams, K. D., Boutle, I. A., Morcrette, C. J., … Vosper, S. B. (2019). Finding plausible and diverse variants of a climate model. Part 1: establishing the relationship between errors at weather and climate time scales. *Clim. Dyn.*, *53*, 989–1022.

Sherwood, S. C., Bony, S., & Dufresne, J. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*, 37–42. doi: 10.1038/nature12829

Shiogama, H., Watanabe, M., Ogura, T., Yokohata, T., & Kimoto, M. (2013). Multiparameter multiphysics ensemble (MPMPE): a new approach exploring the uncertainties of climate sensitivity. *Atmos. Sci. Lett.*, *15*, 97–102. doi: 10.1002/asl2.472

Tompkins, A. M., & Craig, G. C. (1998). Time-scales of adjustment to radiative-convective equilibrium in the tropical atmosphere. *Q. J. R. Meteorol. Soc.*, *124*, 2693–2713.

Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., … Zerroukat, M. (2019). The Met Office unified model global atmosphere 7.0/7.1 and jules global land 7.0 configurations. *Geosci. Model Dev.*, *12*(5), 1909–1963. doi: 10.5194/gmd-12-1909-2019

Webb, M. J., Lambert, F. H., & Gregory, J. M. (2013). Origins of differences in climate sensitivity, forcing and feedback in climate models. *Clim. Dyn.*, *40*, 677–707.

Webb, M. J., Lock, A. P., Bretherton, C. S., Bony, S., Cole, J. N. S., Idelkadi, A., … Zhao, M. (2015). The impact of parametrized convection on cloud feedback.

Phil .Trans. Roy. Soc. A, *373*. (20140414) doi: 10.1098/rsta.2014.0414

Yanai, M., Esbensen, S., & Chu, J. (1973). Determination of bulk properties of tropical cloud clusters from large-scale heat and moisture budgets. *J. Meteorol. Soc. Jpn.*, *73*, 291–304.