https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Stochastic perturbation of inputs to parametrisation schemes machine-learnt from high-resolution model variability

Helena Reid *[1] and Cyril Morcrette[1,2]

[1]Met Office, FitzRoy Road, Exeter, EX1 3PB, United Kingdom
[2]Department of Mathematics and Statistics, University of Exeter, Exeter, EX4 4QE, United Kingdom

Abstract. Stochastic parametrisation schemes represent sources of uncertainty in atmospheric model and several types of these schemes are in widespread use in general circulation models across a variety of temporal and spatial resolutions. We introduce a new stochastic scheme for use in global atmospheric models, which uses a machine learning model trained on high-resolution convection-permitting simulation data to estimate properties of the distribution of subgrid variability in potential temperature. This then informs the profile of stochastic perturbations being applied to the inputs of traditional parametrisation schemes. This scheme is tested in single column model experiments over the tropical west Pacific and is shown to improve model performance in this case.

1 Introduction

Physics-based models of the atmosphere used for weather and climate prediction discretise the atmosphere onto a three-dimensional grid and then predict the time evolution of the atmosphere by solving the dynamical equations that govern the motion of the fluid on the rotating sphere (e.g. Krishnamurti and Bounoua, 1996; Bauer et al., 2015; Gettelman and Rood, 2016). The resolution of this discretisation is limited by computational resources. On climate timescales, the horizontal grid-spacing is typically of the order of one to two hundred km; while on weather timescales for global models, it is around ten to twenty km (Brown et al., 2012). For regional weather forecasting, the grid-spacing is typically less than 5 km (e.g. Bush et al., 2023). In all cases, there are important physical processes which are unresolved by these coarse grid spacings, from the microscale (e.g. aerosol chemistry, raindrop formation) to scales comparable to the grid spacings of our highest resolution models (e.g. shallow convection). To ensure a realistic evolution of the simulated atmosphere, these processes are represented in numerical models by parametrisations, additional terms added to the primitive equations (Stensrud, 2007).

A parametrisation is usually a set of deterministic formulae which maps a set of grid-scale variables to a set of increments to those variables to represent the effects of a sub grid-scale process. However, this deterministic mapping has been criticised. Proponents of stochastic parametrisation suggest one grid-scale state is better mapped to a distribution of increments that can be randomly sampled from, rather than mapped deterministically to a single increment (which could be thought of as aiming to represent the mean of that distribution), a key justification being that this stochasticity can represent the effects of unresolved sub-grid variability (Berner et al., 2017; Palmer, 2019). Introducing stochastic schemes has been shown to provide some of the

*Corresponding author: helena.reid@metoffice.gov.uk

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

benefits of increasing resolution, at a fraction of the cost (Watson et al., 2017). Benefits have been shown across both weather and climate timescales and resolutions (Sanchez et al., 2016a; Palmer et al., 2009).

There are several types of stochastic scheme in common use. First, there are stochastically perturbed parametrisation tendency (SPPT) type schemes (Buizza et al., 1999; Palmer et al., 2009; Charron et al., 2010; Sanchez et al., 2016b). These apply multiplicative noise to the profiles of prognostic variable tendencies for some or all parametrisation schemes. The noise has spatio-temporal structure. In the horizontal, this is typically spectral Gaussian noise, while in the time dimension an autoregressive process of order 1 is assumed. This type of stochastic schemes is quite holistic: they attempt to sample the uncertainty associated with a large number of process-based parametrisation schemes in a way that is decoupled from the structural details of the schemes in question.

Secondly, there are random parameter (RP) type schemes (Bright and Mullen, 2002; Bowler et al., 2008; Grell and Freitas, 2014; Ollinaho et al., 2017)[2]. In most parametrisation schemes, there are several parameters with uncertain values. These may have been set to best guesses based on physical principles, or by some kind of optimisation that best improves model performance according to some desired metric (e.g. by using perturbed parameter ensembles (Murphy et al., 2004)). With an RP-type scheme, the values of the uncertain parameters are perturbed stochastically within a plausible range of uncertainty during the model run. A spatio-temporal structure is usually also imposed (McCabe et al., 2016; Ollinaho et al., 2017). These schemes are less holistic in terms of the overall physical process, instead they attempt to stochastically represent a single source of uncertainty (parameter uncertainty).

Thirdly, there are stochastic kinetic energy backscatter (SKEB) schemes (Shutts, 2005; Berner et al., 2009; Palmer et al., 2009; Charron et al., 2010; Tennant et al., 2011; Sanchez et al., 2016b). These stochastically inject kinetic energy back into the model, typically into the high wavenumbers (small scale) to compensate for dissipation by the numerics, dynamics, or parametrisations schemes.

Then there exists a variety of bespoke stochastic schemes that introduce stochasticity within a parametrisation to target a particular source(s) of uncertainty, for example a Betts-Miller type convection scheme (Betts and Miller, 1986) that perturbs the result of the CAPE calculation additively (Lin and Neelin, 2000). New stochastic schemes have often been developed for use in the so-called "grey zone" of km-scale resolution numerical weather prediction (NWP) models where some assumptions of scale separation between processes made in parametrisations developed for coarser resolutions break down. Examples include a grey-zone cumulus scheme that treats the cloud population in a grid box as stochastic (Sakradzija et al., 2016); and a grey-zone boundary layer scheme that uses stochastic perturbations to represent the effects of an unresolved but small and variable population of turbulent eddies, which improves NWP forecasts of convection (Puh et al., 2023).

SKEB and RP target only specific sources of uncertainty. By contrast, SPPT represents a variety of uncertainties across the whole scheme, but by design represents only multiplicative uncertainties linearly proportional to the parametrised tendency (or some other simple multiplicative relation, such as proportional to the square root of the tendency as in Shutts and Pallarès (2014)). The fraction of the overall structural uncertainty in the model which can be represented this way is unclear, though

[2]Note that in Ollinaho et al. (2017) a scheme of this type is referred to as Stochastically Perturbed Parameters, SPP, which should not to be confused with SPPT above.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

the improvements in model performance that SPPT causes (Weisheimer et al., 2014) indicate it is not negligible. However, certainly *all* uncertainty cannot be represented this way. An example of a source of uncertainty that cannot be represented in this way is now given.

By considering a spatial region of a certain size, one could imagine a thermodynamic profile with significant convective instability and some convective inhibition. This profile could be stable but just on the cusp of convecting. As a result, in the model, convection would not initiate. In the real world however, spatial variability in temperature and humidity could mean that although the spatially-averaged profile is stable, convection could actually be initiated and does develop in some part of the region. This has of course not been ignored by developers of purely deterministic convective parametrisations, and can be partially represented with a small increase in initial buoyancy of parcels over the grid box mean (Gregory and Rowntree, 1990b). However, the ideal treatment of this effect would be stochastic, recognising the uncertainty in estimating subgrid variability (Tompkins and Berner, 2008; Giles et al., 2024), and moreover, would have a means of estimating the distribution of the stochastic perturbations as a function of the local conditions. Likewise, spatial variability may have minimal effects on the lifting condensation level (LCL), level of free convection (LFC) and convective cloud top (CCT), or it may have very large effects, depending on the extent to which changes in near-surface properties on the subgrid scale are accompanied by counterbalancing changes higher up. As to which of these is the case is determined by unresolved subgrid variability, this may also be a target for some stochastic representation.

An example of the latter kind of states is illustrated in Fig. 1, showing the effect of spatial variability on parcel buoyancy. The data are from some km-scale simulations described in the next section. The profiles were chosen to be in the same geographical region and with similar CAPE, CIN and LFC when the parcel and environmental properties were calculated from high-resolution data averaged to the scale of a typical global model gridbox. As an illustration of sub-global model gridbox variability, the ascent can also be calculated by perturbing the potential temperature at each height by ± 3 standard deviations about the mean. In the left-hand panel, this has a minimal effect. Although the amount of negative buoyancy is affected, the heights of the LFC and CCT and the amount of CAPE do not change much. In the right-hand panel however, the effects of subgrid variability on the LFC, CCT and CAPE are enormous. Although an ascent following an extreme tail of temperature perturbations at each level all the way up is unlikely, Fig.1 illustrates that even with a similar sounding in a grid-box mean sense, the depth and vigour of convection will be sensitive to subgrid variability. Distinguishing between these cases using only mean state information is not trivial, but can be tackled using machine-learning approaches as will be shown later.

In the example of a case where there is considerable uncertainty about whether convection triggers, a convection scheme may predict no triggering, and no multiplicative perturbation to the increments from the convection scheme could make them anything other than zero. This is thus an example of a source of uncertainty whose effects cannot be represented by an SPPT-type scheme. Likewise, if the scheme does trigger, there may still be considerable uncertainty in the height of the convective cloud top (as in Fig. 1), but multiplicative perturbations above the scheme's predicted cloud top will again be around zero. To represent these effects, a different type of scheme is needed.

The final approach to stochastic parametrisation that will be discussed here, and the one that is used in this work, is the idea of perturbing the inputs to a parametrisation scheme (Tompkins and Berner, 2008; Giles et al., 2024).

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

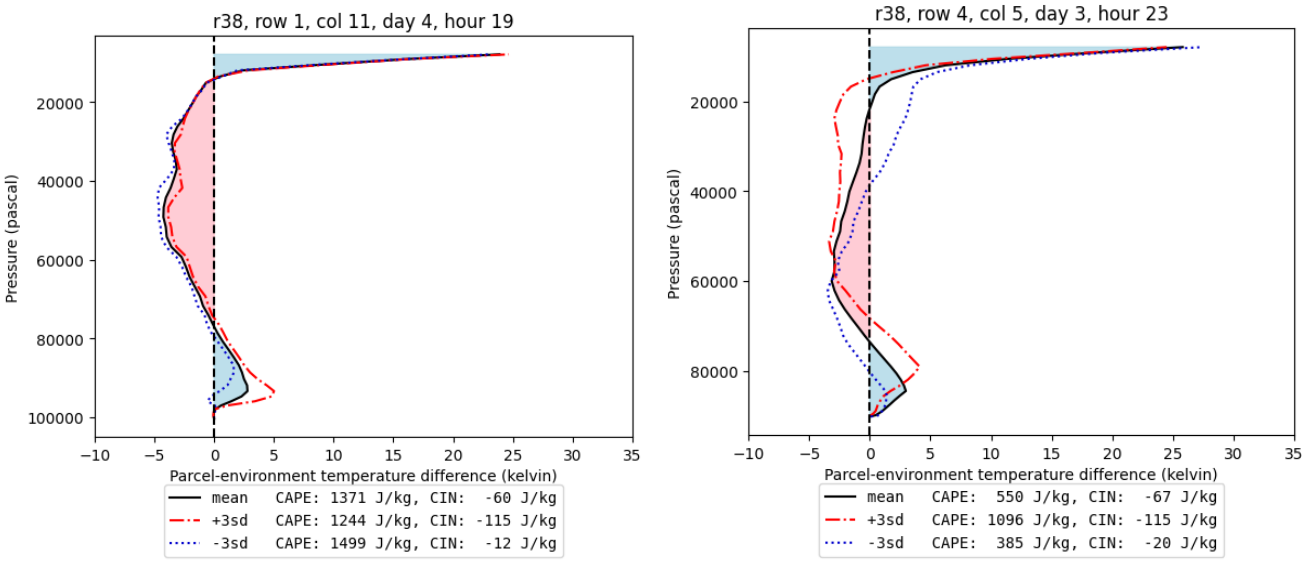

Figure 1. Profiles of temperature difference between a rising parcel and the environment. In each panel a solid black line shows this difference computed using grid-box mean properties. The subgrid standard deviation of potential temperature is computed at each height. The difference is then computed again for profiles plus and minus three standard deviations from the grid-box mean. Negative buoyancy is shaded in blue, positive buoyancy in pink. Convective Available Potential Energy (CAPE) and Convective Inhibition (CIN) are computed for each of the three profiles per panel. The dataset these were taken from is described in section 2 and is the same region used in fig. 3.

This work develops a stochastic approach that makes use of a machine learning model, trained on high-resolution atmospheric simulations to predict plausible alternative profiles, generated by perturbing grid-box mean profiles according to estimated subgrid variability. These plausible alternative profiles are then passed to the traditional process-based parametrisation schemes. The scheme is called PAPILLON, as a nod to the butterfly effect (Lorenz, 1972) and based on it generating “Plausible Alternative Profiles Intelligently Learnt from Lots of Output from NWP".

This paper is structured as follows. The methods are outlined in Section 2. Results are presented in Section 3. A discussion appears in Section 4. Concluding remarks are given in Section 5.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

2 Methods

2.1 Machine Learning Model

2.1.1 Training Data

The training data were produced using the Met Office Unified Model Nesting Suite (Webster et al., 2008). The simulations are identical to those described by Giles et al. (2024) and Morcrette et al. (2025) but are summarised here for completeness.

A global model with GA6 configuration (Walters et al., 2017) was run using a $N1280$ resolution thus gridboxes have meridional sizes of $\sim 15km$ at the equator. The global model is reinitialised from operational reanalysis every 24 hours and provides lateral boundary conditions for a set of 80 free-running Limited Area Models (LAMs) once per hour. The location of the LAMs are shown in Fig. 2.

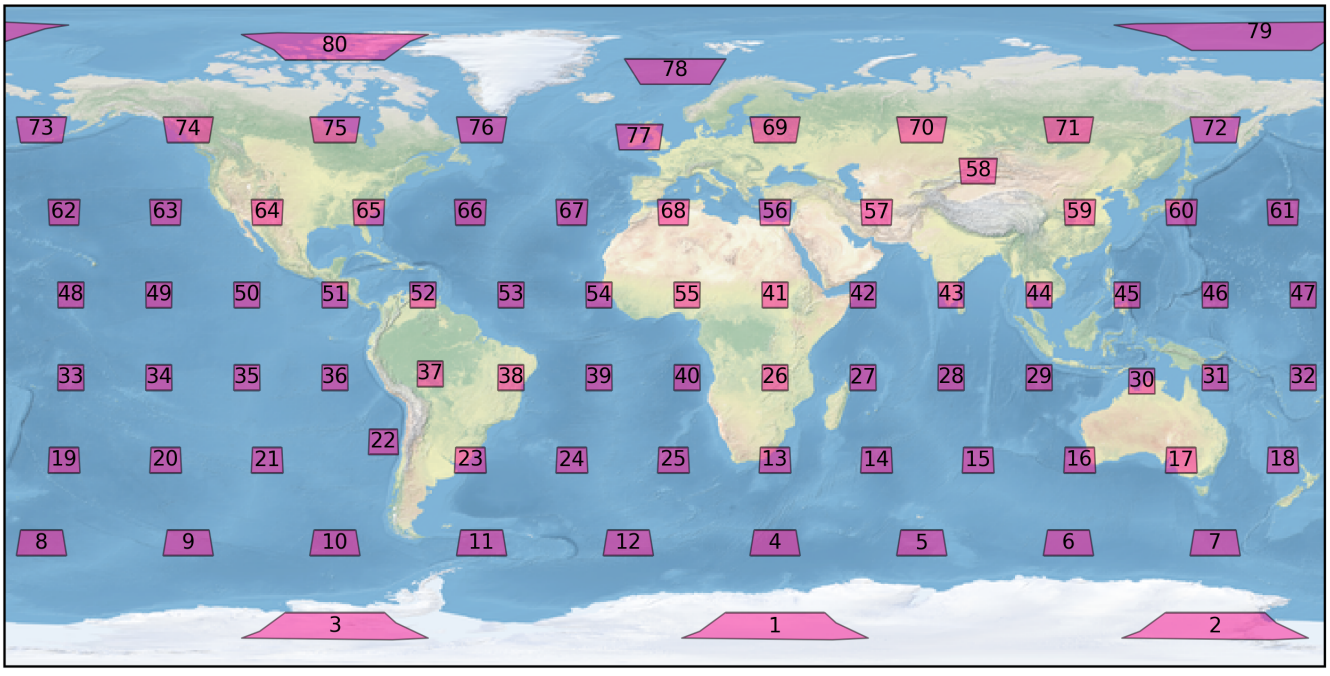

Figure 2. Location of the 80 limited area models (LAMs) used to generate the training data. Each LAM is 768 km x 768 km and uses a gridlength of 1.5 km.

The LAMs use a regional atmosphere and land configuration (RAL2-T) with a timestep of 1 minute (Bush et al., 2023). The LAM grid is a regular latitude-longitude grid with 512x512 grid points spaced $0.0135°$ apart. A rotated pole is used in each LAM's grid, such that regardless of the LAM's location on Earth, the LAM centre is at the equator on that grid, meaning the deformation of the grid boxes in the meridional direction is smaller than if the pole was not rotated. Thus grid boxes are approximately square with side lengths $dx \approx 1.5km$. The phrase "rotated pole" here refers only to the choice of coordinate

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

for discretisation on a sphere - the polar regions of Earth remain physically unchanged, solar zenith angle calculations are unaffected, etc. For both the driving global model and the nested regional model, the vertical discretisation is a height-based coordinate with 70 levels. The first 50 levels above the surface are compressed in a terrain-following manner. Level 50 is at an altitude of 18km and the model top is at 80km. Regional configurations of the UM are more routinely run with a model top at 40km but to avoid interpolation between vertical levelsets the LAMs were run with the same vertical resolution as the global model within which we wish to ultimately incorporate our stochastic parametrisation scheme. In contrast to the global model, the LAMs are only initialised from reanalysis once. No ocean model is coupled in either model, sea surface temperatures are prescribed from OSTIA (Donlon et al., 2012). The models were run for one day (30th December 2019) plus one month (January 2020). As it takes time and space for fine-scale details to spin up, the first day of data was discarded, as well as a perimeter 16 grid points wide around the edge of the LAMs, as was done in Morcrette et al. (2025).

Horizontal averaging of the $0.0135°$ (hereafter “fine") LAM data is performed to $0.54°$ resolution (i.e. 40 x 40, $dx \approx 60km$, hereafter “coarse") to produce the following coarse variables: specific humidity on model levels $\overline{q}$, temperature on model levels $\boldsymbol{T}$, pressure on model levels $\boldsymbol{P}$, land-sea mask where 0 is sea and 1 is land f_{land}, and orography h. The standard deviation of h, σ_h, in each coarse grid box is calculated from the fine orography field. The standard deviation of potential temperature for each model level in each coarse grid box, σ_θ, is calculated from the fine potential temperature field. The coarse-graining is illustrated in Fig. 3.

2.1.2 Neural Network

A neural network is then trained to output σ_θ from the coarse variables:

$$\sigma_{\theta,ml}(\bar{X}) = G(\bar{X}, \bar{W}) \tag{1}$$

Where $\bar{X}$ is the coarse inputs $\bar{X} = (f_{land}, h, \sigma_h, \boldsymbol{q}, \boldsymbol{P}, \boldsymbol{T})$, G is inference of the neural network and $\bar{W}$ represents the weights and biases of the trained neural network. Since these are only varied during training and are constant at inference time when the network is called within an atmospheric model, we consider the trained network to be a function of $\bar{X}$ only that outputs $\sigma_{\theta,ml}$.

The neural network is a densely connected network with four hidden layers of 512 nodes each with leaky ReLU activation with a slope of 0.1 to the left of $x = 0$. The inputs are the aforementioned $\bar{X}$ which have a total length of 213, and the output is 70 model levels of σ_θ. The loss function is the Huber loss (Huber, 1964), and the Adam optimiser was used for training with an initial learning rate of 1.0×10^{-3} (Kingma and Ba, 2014). A random eighty percent of the data was used for training, ten percent for validation, and ten percent was reserved as test data. The learning rate was halved after three epochs if no improvement in the loss occurred. The model was trained until the loss showed no further improvement.

The trained model was translated to Fortran for implementation in the LFRic model using Python-to-Fortran tools based on Morcrette (2024) and Reid et al. (2025).

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

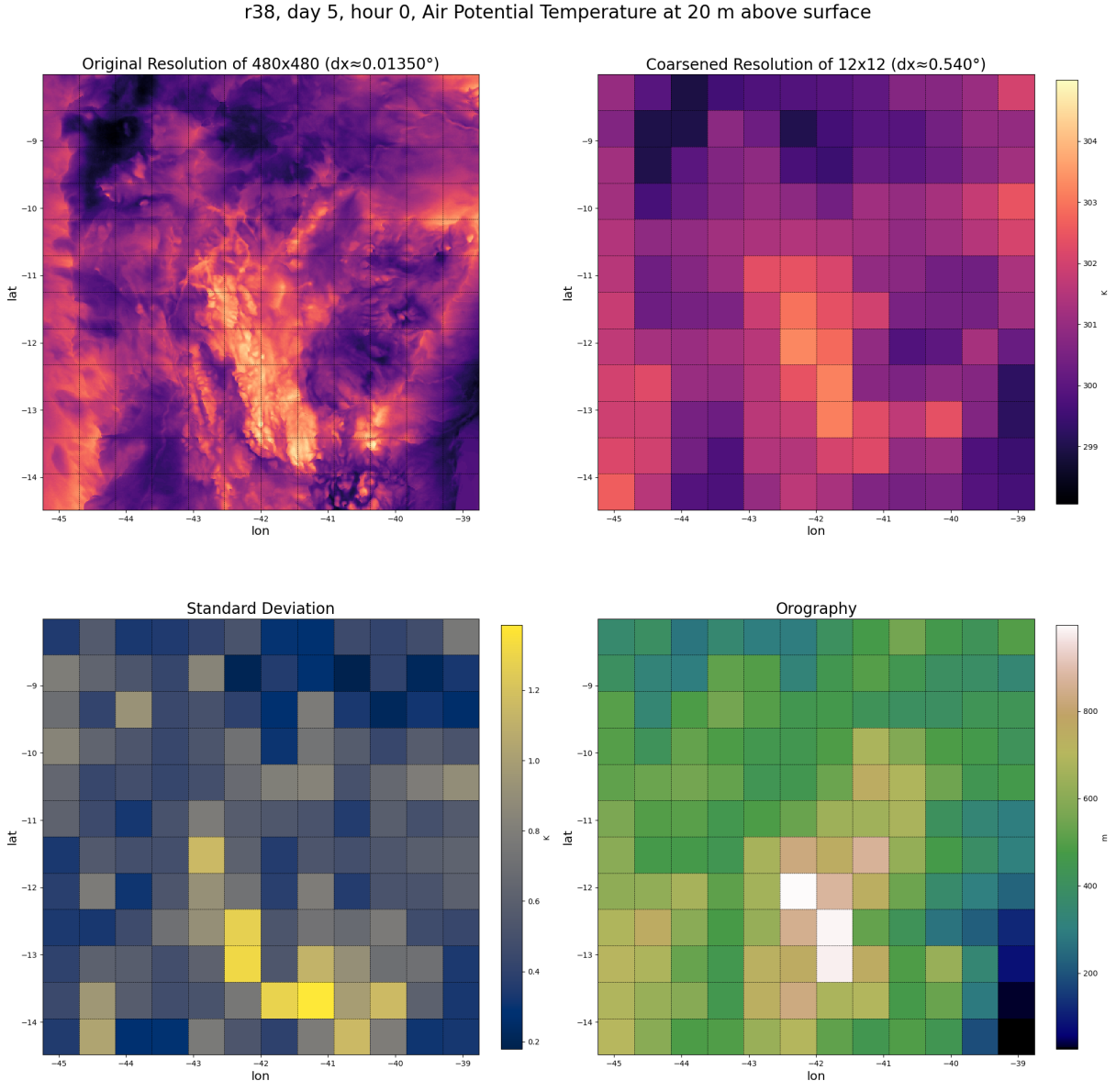

Figure 3. Illustration of the coarse-graining process using data from region 38 (eastern Brazil in Fig. 2): a) original 1.5 km near-surface level potential temperature data on 480 × 480 grid (having ignored a 16 gridpoint region around all 4 edges), b) mean data coarse-grained by a factor of 40 × 40 (i.e. 60 km × 60km) and hence now on a 12 × 12 grid, c) standard deviation of surface potential temperature on scale of 60 km × 60km, d) mean orography on scale of 60 km × 60km.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

2.2 Stochastic Parametrisation Scheme

The PAPILLON scheme is implemented into the LFRic atmospheric model (Adams et al., 2019). The specific details are as follows.

Simplex noise is a computationally efficient type of gradient noise commonly used for generating natural-looking patterns easily extensible to higher dimensions(Gustavson, 2005; Perlin, 2002).

Consider a function for generating 4-dimensional simplex noise $\eta(\bar{x},t)$, where $\bar{x}$ is a position vector in space and t is time. The simplex noise algorithm used uses Cartesian coordinates but can be straightforwardly called for any point on the atmospheric model's coordinate grid by converting the model grid coordinates to Cartesian ones. The model vertical and temporal coordinates are in metres and seconds respectively. Two free parameters are introduced, l_z and τ, which the vertical and temporal coordinates are divided by respectively before η is computed. These thus control the characteristic length scale of the noise in the vertical and temporal directions. The horizontal length scales of the noise could also be varied in a similar way, but since only single column model experiments are discussed here this is not relevant as η is called only at a fixed latitude and longitude.

The simplex noise, η, naturally produces values between -1 and 1 with a standard deviation of $\sigma_\eta \approx 0.3021$. We divide this such that the resulting distribution has a standard deviation of 1 and varies between $\pm\frac{1}{\sigma_\eta} \approx 3.31$. Temperature perturbations are applied like so:

$$\theta_{perturbed} = \theta + \frac{\eta(\phi,\lambda,z/l_z,t/\tau)\sigma_{\theta,ml}(\bar{X})}{\sigma_\eta} \tag{2}$$

Where $\theta_{perturbed}$ is the perturbed potential temperature, ϕ latitude, λ longitude, and z height above Earth's surface. Note the range of the noise distribution means that the perturbed potential temperature at any vertical level can only ever be a maximum of $\frac{1}{\sigma_\eta}$ (machine-learnt) standard deviations away from the original temperature. This is a desirable quality, as it naturally caps the perturbations, preventing them from being large enough to cause instability in the model.

These potential temperature perturbations are *not* applied to the prognostic potential temperature. They are applied to the potential temperature profile which is passed as input to one of the other parametrisation schemes in the model. That scheme then outputs tendencies which are added to the prognostic variables. Thus, the modification of the prognostic variables by the PAPILLON scheme is indirect - it modifies the inputs to other schemes, causing those other schemes to output different tendencies to what they would have output had PAPILLON not been called.

In all experiments run here, PAPILLON is called before the schemes in the "fast physics" section of the LFRic timestep, namely the parametrisations of boundary-layer mixing and moist convection (Walters et al., 2017).

2.3 Single Column Model Experiments

Single-column models are a common tool for testing new parametrisation schemes. Their one-dimensional nature allows new schemes to be tested in a simple system without complex feedbacks. They are very cheap to run and allow easier debugging (e.g. Derbyshire et al., 2004; Daleu et al., 2023; Li et al., 2025) and SCM experiments may also make use of the benefits of

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

ensembles (e.g. in the multi-model ensemble and ensemble forcing dataset used in Davies et al., 2013) commonly used in less idealised models.

The LFRic Single Column Model (SCM) can be forced using large-scale advective tendencies and its outputs compared to a reference dataset. In this case, we take ERA5 data (Hersbach et al., 2020) for the entire month of November 1992 centred on latitude -1.79°, longitude 155.71°, which is in the tropical west Pacific. This time and location were chosen for two reasons. Firstly, so that we can expect the convective parametrisation to play a significant role in this area. Second, because it overlaps with the TOGA-COARE intensive observation campaign (Webster and Lukas, 1992), and we expect reanalysis errors will be lower where there is an abundance of observations. The ERA5 dataset is used to compute a time series of advective tendencies of prognostic variables into the specified SCM region from a discrete approximation of the divergence of those variables. This time series is then used as a forcing dataset, applied as increments every timestep in the SCM. The SCM also performs temporal interpolation to compensate for any differences in the timestep of the model compared to the timestep of the forcing data - the SCM timestep is 20 minutes in these experiments, while the forcing data is hourly. Thus, the forcing is smoothed out over several timesteps rather than applied suddenly every hour, but the total increment due to forcing remains the same. We force temperature, specific humidity, and zonal and meridional winds.

The same vertical discretisation is used in the SCM as in the training data for the ML model. All SCM experiments performed are over the sea. As a result, the terrain-following nature of the coordinate is not relevant and thus there is a constant, straightforward mapping of each level to a particular height above the surface, which is provided in Appendix A1 for reference.

The SCM provides an initial temperature perturbation option for generating ensembles. We make use of this to run 20-member ensembles for all experiments.

Each ensemble member consists of many 5-day SCM runs. We first initialise the SCM at the start of the first day of our forcing data, then run for 5 days. Then we initialise it at the start of the second day, and again run for 5 days. This is repeated until the end of a simulation reaches the last day of our forcing data. The SCM provides the ability to have runs overlap with one another like this so the effect of lead time can also be investigated if desired.

Thus, there are 25 5-day runs, each covering a different portion of a 1-month period, per ensemble member. These runs are repeated with perturbed initial temperatures to produce new ensemble members. There are thus 500 5-day runs in total per ensemble.

The SCM has several limitations beyond those inherent to all SCMs. First, that while initialised from reanalysis, the sea surface temperature is fixed throughout the run. The same is true for ozone concentration. Interpolation from the ERA5 vertical coordinate onto the LFRic grid is also necessary, and is performed using cubic splines.

The model control configuration is set to GA7 as described in Walters et al. (2019) with minor modifications for running in SCM mode (gravity wave drag parametrisations are switched off; the aforementioned constant SSTs and ozone). All experiment configurations are identical to the control configuration where not otherwise stated.

The SCM then outputs a comparison between the control, the experiment, and the user-supplied ground truth.

The control in our case is the LFRic SCM with no stochastic physics scheme active. Like all the experiments, the control is an ensemble of 20 members, each containing 25 5-day runs.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

We perform sensitivity tests for the two relevant free parameters τ and l_z. The temporal length scale is varied between values that imply almost no correlation in stochastic perturbations from one timestep to the next (set to 20 minutes, 1 timestep), to values that imply temporal correlation durations similar to the length of the runs (set to 5 days, same as the model run length). In keeping with the SPPT timescale in (Palmer et al., 2009) we use a default of $6h$. In addition, l_z is varied from 100m to 1000km, which again reaches both extremes (almost no vertical correlation in perturbations, to maximally correlated vertical perturbations respectively). Stochastic physics schemes such as SPPT typically assume maximal vertical correlation in perturbations in the region of the atmosphere in which they are active. We set a default of $10km$. We do not perform a full grid search of the parameter space, only sweeping a single row and column - when varying l_z, τ is set to $6h$; when varying τ, l_z is set to $10km$.

We also run an experiment comparing runs with PAPILLON to runs with an existing Met Office stochastic physics SPPT-type scheme called SPT (Sanchez et al., 2016b). A typical configuration of this SPT scheme was selected, described briefly as follows.

If the SPT perturbations would cause the convective mass flux to breach the Courant–Friedrichs–Lewy (CFL) condition, then the SPT perturbations are not applied to this column (so in SCM mode, they are skipped for this timestep). The decorrelation time for the first-order autoregressive temporal component of the forcing pattern is $20,000s$ ($\approx 5.5h$). A linear taper is applied to perturbations from $1700m$ down to $800m$, and again from $11.2km$ up to $18.0km$, such that the perturbations are zero low and high in the atmosphere. Moisture conservation in the column is enforced. Potential temperature increments are linked to specific humidity increments in a way that conserves moist static energy. Three smoothing iterations with a 1-2-1 filter are applied to the SPT perturbations. All forcing pattern standard deviation scaling coefficients are set to 1. SPT is applied to the increments from the convection, microphysics and radiation parametrisations.

We also run an experiment where the ML model was replaced with a constant profile of standard deviation of temperature, obtained by globally averaging the training data. The reason for this experiment is to isolate the effects of simply perturbing the inputs to schemes from the effects of dynamically varying the perturbations based on local conditions.

Finally, to provide a benchmark to compare the impact of our stochastic physics approach, we ran two experiments where a parametrisation was swapped for an alternative which we would expect to cause a significant change (and an overall degradation of) performance. In one case, we replace the operational PC2 cloud scheme (Wilson et al., 2008a, b) with the Smith (1990) cloud scheme (Smith, 1990), which diagnoses cloud fraction and condensate from grid-scale variables assuming a triangular probability distribution function (PDF) of subgrid saturation departure. In the other, we swap the 6a convection scheme (a mass-flux scheme based on Gregory and Rowntree (1990a) with developments as described in Walters et al. (2017)) for the Lambert-Lewis convection scheme, in which convection is simply represented as a relaxation towards a moist adiabat (Lambert et al., 2020).

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

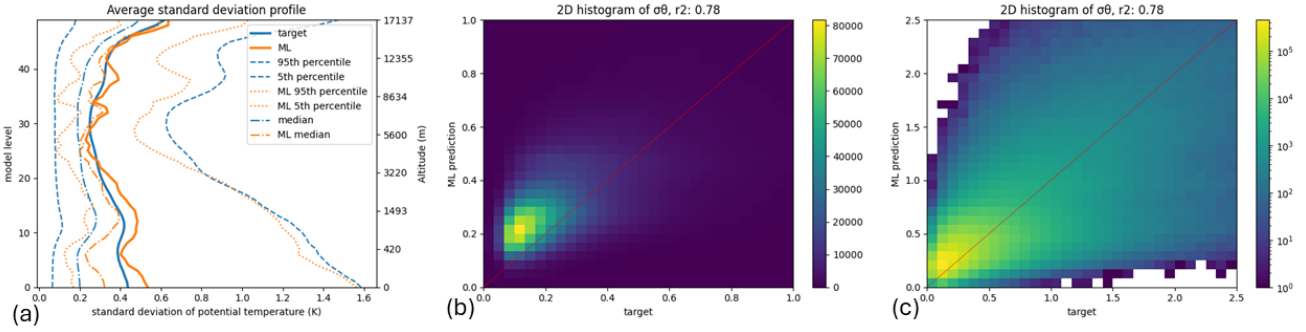

Figure 4. Evaluation of ML performance on the test dataset. a) Comparison of the mean, median, 5th and 95th percentiles of the standard deviation of potential temperature with height in the test dataset vs ML predictions. b) 2D histogram of standard deviation of potential temperature (K) as in the test dataset vs as predicted by the ML model. c) As (b), but with a logarithmic colour scale. While the ML model was trained on all model levels, performance here is evaluated only for the first 50 (approximately to the tropopause) because in the stochastic scheme described later which makes use of this ML model, we do not apply perturbations above this altitude.

3 Results

3.1 Machine Learning Model: Offline Evaluation

To illustrate the variations in the target, fig. 4a shows the profile of the standard deviation of potential temperature, σ_θ, averaged over the different locations and over time for the validation data (solid blue line). Near the surface, potential temperature has a standard deviation of around 0.45 K on average. This reduces gently with height rising up through the boundary layer, before increasing again, peaking around model level 12 ($\approx 1100m$, see table A1 for level-height mappings). The mean value of σ_θ reduces again through the free troposphere, reaching a minimum of around 0.3 K near model level 30 ($\approx 6400m$). Above that height, σ_θ increases gently again, before increasing more abruptly around level 45 as some of the samples find themselves in the stratosphere.

In fig. 4a the 5th, 50th (median) and 95th percentile profiles are also shown. The median is notably lower than the mean at all heights, indicating that σ_θ has a positively skewed distribution. Larger values of subgrid variability in θ may be an order of magnitude higher than the median.

Figure 4a also shows the prediction made by the NN. The NN mean and 95th percentile profiles agree with the target data fairly well with the latter underestimated at higher altitudes, while both the 5th percentile and median are somewhat overestimated at all heights.

We see from fig. 4a that the ML reproduces the expected distribution, but this is only one aspect of performance - it would be possible to predict the correct distribution of σ_θ but to predict large values whenever the true value was small and vice versa. We thus plot 2D histograms of targets vs predictions of σ_θ in fig. 4b and fig. 4c. Perfect performance would have all predictions equal to the target (red diagonal line). From the r2 score here of 0.78, the ML model is not without skill, though it is apparent

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

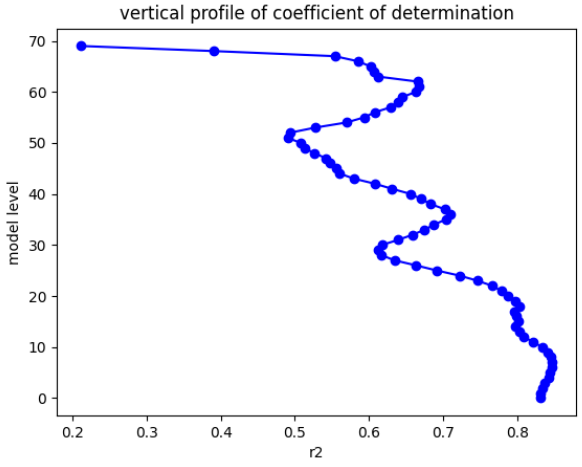

Figure 5. Coefficient of determination r^2 score of the ML model with height.

from the position of the peak of the histogram in fig. 4b that the ML model has a bias towards overpredicting σ_θ when its value is low, and when its value is high fig. 4c shows more spread in the predictions.

In fig. 5, a vertical profile of the coefficient of determination (r^2) score for model predictions compared with the test dataset is plotted. The performance in the planetary boundary layer (PBL) (level 10 being 800m, level 20 being approx 3km, see table A1 for level-height mappings) is best, with a trough in free tropospheric performance in the mixed-phase cloud region, and a minimum tropospheric performance around the height of the tropopause (level 49 is approx 17km). The performance improves again in the stratosphere, before dropping significantly above 70km (levels 69 and 70). As ultimately it was decided to not apply stochastic perturbations above the tropopause owing to this having no expected benefits for the convective and boundary layer parametrisations as they should be inactive in this region of the atmosphere, the poor performance at very high altitude was not considered a problem. The opposite is so for the good performance in the PBL, which is a desirable quality as we expect perturbations here to have a larger effect.

3.2 Perturbed profiles used in Single Column Model simulations

An obvious way to examine single-column model data is to look at time-series plots of individual variables, an analysis of which follows. There are problems with this approach however, which will be described below. Consequently, the results of the SCM simulations are synthesised into a scorecard of statistics.

Example SCM outputs of a selection of variables can be seen in fig. 6 and fig. 7. Each compares the control and experiment outputs to a source of truth, in this case the aforementioned ERA5 reanalysis. The control and experiment are in general visually more similar to one another than to the reanalysis, which is rather to be expected given they differ only in the introduction of stochastic perturbations - the entire set of physical parametrisations and forcings are otherwise identical. fig. 6 shows a time-

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

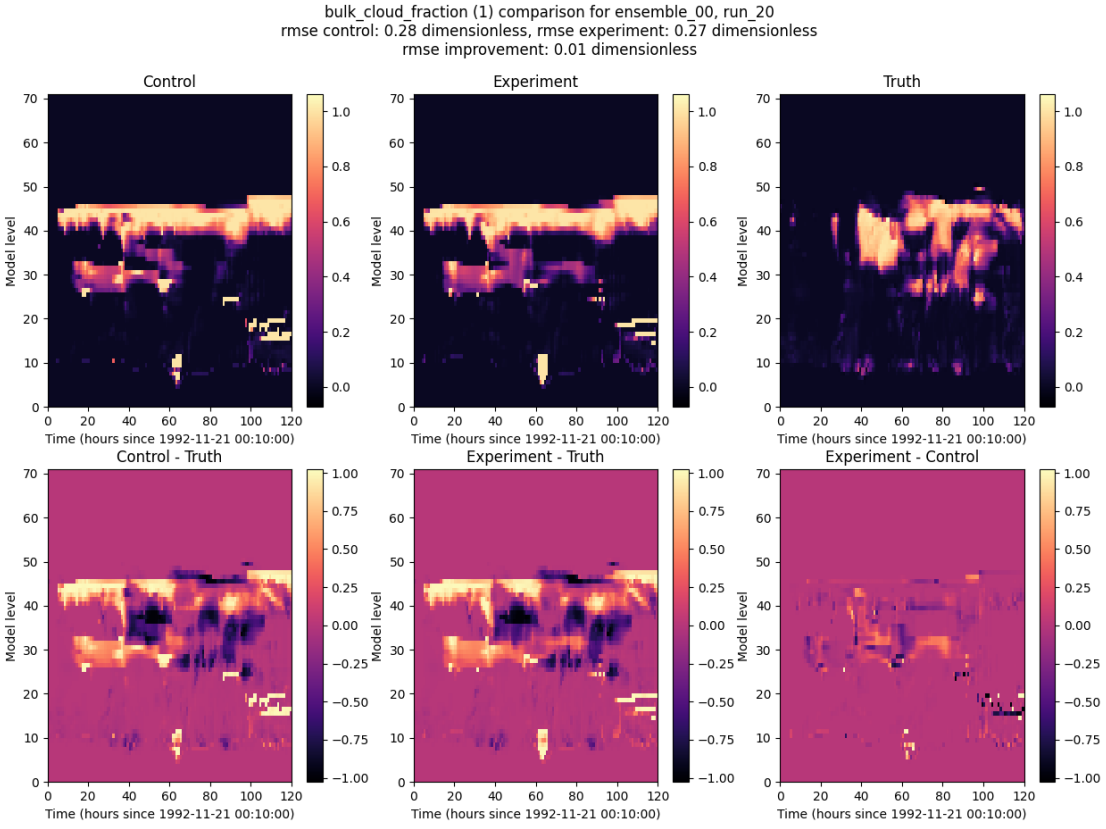

Figure 6. Top row: time-height cross-sections of bulk cloud fraction in a) the control, b) the experiment and c) the data the SCM analysis tool is treating as the ground truth, here ERA5 reanalysis. Bottom row, differences between d) control and ERA5, e) experiment and ERA5 and f) experiment and control. All plots are for a 5-day period.

height cross section of bulk cloud fraction over the course of one of the 5-days of simulation. In this example there is expansive cloud in the middle and high troposphere with more modest cloud extent in the boundary layer.

Figure 7 shows the evolution of two variables for two separate dates. Figure 7a shows the outgoing long-wave radiation (OLR) at the top of the atmosphere (TOA). In this case both the control and experimental ensembles initially over-estimate the OLR, but later on have envelopes which encompass the value from the ERA5 reference data. Figure 7b shows the ice water path (IWP). There are a few deep cloud events with peaks 40 and 60 hours into the simulations. In both cases the experimental ensemble has both a higher mean value and a broader variability which is in better, albeit not perfect, agreement with the reference data. Conversely, the experiment sees a spike in IWP at around 100 hours that is not seen in the control or reanalysis.

We must highlight the subjectivity of this analysis of specific variables at specific points in time. These examples were selected because differences between the control and experiment were very clear to see by eye. Thus they demonstrate that

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

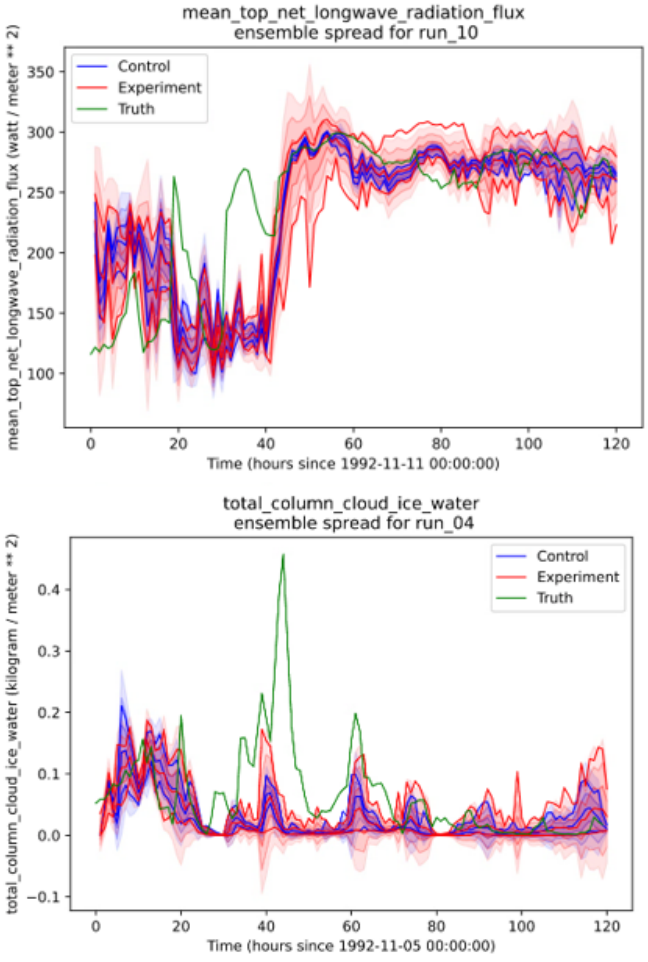

Figure 7. Example output from the SCM experiments. Here "control" refers to the control ensemble with default configuration; "experiment" refers to PAPILLON being switched on with the length parameter set to 10km and the temporal to 6h; "truth" refers to the ERA5 reference dataset. The bold lines are the minimum, mean and maximum of the variable in the ensemble. The shaded areas are the ensemble mean plus or minus one, two or three standard deviations. The top plot is of outgoing long-wave radiation (OLR) at the top of the atmosphere (TOA) on 11 Nov 1992, the bottom plot is for ice water path (IWP) on 5 Nov 1992.

noticeable differences do occur - but there are 525 such plots possible per experiment, and in general, it is difficult to discern points in time where there was an obvious advantage or disadvantage for the experiment over the control. To determine which case is more common and whether this is significant will require robust quantification. We therefore turn to an analysis of broader statistics.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

4 Discussion

To summarise all the information from the experiments, we construct a scorecard that includes all the variables output by the SCM which have a straightforward equivalent in the reanalysis dataset (fig. 8). We compute the continuous ranked probability score (CRPS, Epstein (1969)) for each of these variables across all ensembles and all timesteps, for both the control and the experiments. The rows of the scorecard are variables, the columns experiments, and values in the scorecard are then:

$$CRPS/\mu_{true} \tag{3}$$

Where $CRPS$ is the overall CRPS scores for that variable and experiment, divided by the mean value for that variable over time in the truth dataset μ_{true}.

Examining first the sanity check runs, where the cloud or convection schemes were replaced with simpler parametrisations. One can see the expected general degradation of performance in both cases, particularly in the convection scheme replacement. These experiments show the impact of PAPILLON is comparable in magnitude to the impact of swapping one parametrisation scheme for another - not a negligible effect.

The experiment with SPT unexpectedly turned out to degrade performance across the board compared with the control. This was unexpected as SPT is known to improve performance in both regional and global scales (Weisheimer et al., 2014; Sanchez et al., 2016a; Jankov et al., 2017). We recognise the SPT scheme is here being used outside the environment it was developed and tested in, which was 3D rather than 1D. A possible explanation for the different effects in the SCM may involve the lack of dynamics for the SPT scheme's effects to interact with. The SPT scheme's large perturbations, of similar order to the scheme tendencies themselves, may provide benefit only through the interaction of the effects of these perturbations across many grid points. That is, in a single column, SPT may push its state very far away from the state that would be reached by applying the unknown true tendency due to subgrid processes, but in a 3D run errors in different locations are able to compensate for one another. For example, if one grid box has the mid troposphere significantly dried out by a particularly extreme SPT perturbation, but adjacent grid boxes did so less, then advection and diffusion could quickly smooth the model state back out.

The experiment where the ML model was replaced with a constant profile also showed this improvement in all variables, but to a lesser degree than most of the configurations that used ML. The magnitude of the improvement between this non-ML configuration and the control is similar in size to that between ML configurations and the non-ML configuration. This suggests the effects of dynamically scaling the perturbations based on local conditions are significant, its effects being of comparable magnitude to the introduction of the perturbations.

Crucially, for all experiments where PAPILLON was active, variables showed universal improvement in CRPS.

The experiments varying the vertical length scale show a peak in CRPS improvement for most variables in the 1 to 3 km range. A possible explanation for a km-scale length being optimal is that the greatest impact on the convection scheme may be had when the length scale is large enough to perturb the whole boundary layer coherently (which would have a big impact on the initial parcel properties and trajectory), but not so large that the perturbations in the rest of the troposphere are always of

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

All experiments relative_crps relative to mean

Variable	smith	llcs	spt	no_ml	100m	1km	3km	10km	1000km	20min	90min	3h	6h	15h	5day	ctrl
total column water	31.23	41.67	30.82	30.16	29.22	29.59	29.85	29.48	29.31	30.03	29.60	29.46	29.48	29.99	29.92	30.50
total column water vapour	31.35	40.88	30.88	30.24	29.26	29.64	29.91	29.55	29.38	30.11	29.68	29.54	29.55	30.06	29.98	30.57
total column cloud liquid water	227.44	630.23	286.07	216.04	221.97	216.25	207.92	204.55	206.78	210.58	204.22	202.10	204.55	217.23	218.54	235.39
total column cloud ice water	192.80	552.50	210.15	181.50	174.59	174.44	174.99	177.99	179.95	185.21	179.51	178.43	177.99	184.21	183.99	188.19
total cloud cover	45.05	45.48	45.64	41.42	40.95	40.20	40.50	41.15	41.20	41.79	41.37	41.02	41.15	41.50	41.39	43.17
low cloud cover	272.93	250.23	283.70	231.89	247.01	224.32	217.81	221.23	221.96	228.32	225.07	220.08	221.23	224.49	226.25	248.16
medium cloud cover	421.90	482.65	510.14	351.94	342.63	334.11	332.33	338.87	339.59	360.91	340.95	335.87	338.87	350.73	349.09	391.00
high cloud cover	58.84	59.40	65.05	57.53	56.27	55.40	55.86	57.32	57.83	59.04	58.14	57.69	57.32	56.99	57.12	60.08
precipitation flux	302.26	274.77	369.47	288.40	276.48	270.68	273.12	276.61	276.27	296.91	279.05	275.36	276.61	291.75	292.46	310.47
precipitation convective mean rate	316.71	276.90	393.98	302.01	282.93	276.88	282.29	286.84	286.20	315.14	292.14	286.86	286.84	303.96	304.63	328.07
precipitation large scale mean rate	338.89	880.40	369.86	320.35	314.38	311.55	310.62	312.88	314.67	312.49	308.55	308.99	312.88	317.17	317.92	328.48
mean top net longwave radiation flux	17.15	22.20	17.90	15.38	14.81	14.69	14.86	15.16	15.31	15.81	15.32	15.17	15.16	15.53	15.56	15.96
mean surface net longwave radiation flux	30.58	36.54	34.12	30.79	29.95	30.36	30.60	30.22	30.23	30.75	30.36	30.13	30.22	30.25	30.14	31.71
mean surface downward longwave radiation	4.17	5.01	4.63	4.22	4.08	4.16	4.20	4.14	4.13	4.20	4.15	4.12	4.14	4.14	4.12	4.33
mean surface net shortwave radiation flux	22.89	38.62	26.88	21.04	19.97	19.65	19.64	20.19	20.46	20.93	20.34	20.28	20.19	20.99	21.07	22.34
mean surface downward shortwave radiatio	22.60	38.36	26.56	20.78	19.75	19.41	19.40	19.93	20.20	20.67	20.08	20.02	19.93	20.72	20.81	22.08
mean surface direct shortwave radiation flu	38.18	53.69	44.39	34.93	32.14	31.80	32.38	33.52	34.14	35.14	34.02	33.76	33.52	34.65	34.89	37.17
mean surface latent heat flux	76.41	95.03	77.63	74.79	58.67	63.84	70.50	69.66	69.22	73.79	72.15	70.82	69.66	67.67	67.26	75.41
10 metre u wind component	1423.97	1130.35	1436.59	1402.67	1299.21	1345.41	1376.36	1383.73	1382.90	1411.86	1399.45	1389.81	1383.73	1371.08	1364.34	1414.97
10 metre v wind component	247.57	233.03	248.10	243.70	233.12	236.97	238.86	240.28	240.31	244.80	242.45	241.12	240.28	240.55	239.83	245.90
temperature at 2m	0.59	0.47	0.60	0.57	0.50	0.53	0.56	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.58

Figure 8. Each column represents an experiment and each row a variable of interest. Each cell gives the CRPS score for that variable in that experiment, normalised by dividing all the values in each row by the time-mean value of that variable in the truth dataset. The cells are coloured according to difference from the control, with better CRPS in blue and worse CRPS scores in orange. From left to right, the first three experiments are benchmarks (where the default cloud scheme was replaced with the Smith scheme, the default convection scheme with the LLCS, and the SPT stochastic scheme was switched on, respectively); in the fourth PAPILLON was switched on with the ML component replaced with a constant profile; the fifth through ninth are the experiments where the length scale parameter was varied; the tenth through fifteenth are those where the temporal scale parameter was varied; and the final column is the control.

the same sign to those in the boundary layer (which would reduce the variety of possible effects the perturbations could have on parcel ascent).

The experiments varying the temporal length scale show a peak in improvement for most variables in the 3 to 6 hour range. This is also the timescale other stochastic schemes typically use. A possible explanation for this is that this stochastic timescale is similar to the typical lifetimes for tropical cloud system (e.g. Hagos et al. (2013)).

5 Conclusion

A new stochastic parametrisation approach has been developed and tested in single-column model experiments. The parametrisation stochastically perturbs the input profile of potential temperature seen by the convection and boundary-layer mixing schemes. The perturbations are generated by first calling a NN to generate a profile of standard deviation of potential temperature, and then multiplying this by a noise field which has correlations on spatial and temporal length scales specified by the user. The NN is trained on convection-permitting regional simulations run at various locations around the globe and then coarse-grained spatially to the size of a global model gridbox. The NN performs best at predicting the variance in potential tem-

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

perature in the planetary boundary layer. This is advantageous as both schemes the perturbations are applied to are particularly sensitive to perturbations there.

The effects of PAPILLON are found to be comparable in magnitude to replacing one of the parametrisation schemes with a completely different scheme. Running the scheme without the ML model, that is by perturbing the profile using a fixed profile of standard deviation of temperature is found to also improve CRPS scores. But crucially, switching on the ML model which dynamically varies the profile of standard deviation of potential temperature based on local conditions is found to approximately double these improvements. The optimal vertical length scale is found for most variables to be around 1 to 3km. The optimal temporal length scale is found to be approximately 3 to 6 hours. An improvement in CRPS for all variables is seen in all experiments in which PAPILLON was switched on.

In future work the machine-learning model could be developed further. This could be by making use of one-dimensional convolutional neural networks in the vertical to better capture the importance of atmospheric stability variations on subgrid temperature variability. In addition, the amount of training data could be increased by making use of km-scale simulation from sources beyond the simulations described here. This could include data from previously completed km-scale such as those performed as part of the DYAMOND project (Stevens et al., 2019).

The clearest next step for future work would be to evaluate this stochastic scheme in a full three-dimensional atmospheric model (either LFRic in 3D mode, or in a different general circulation model) to examine its applicability to less idealised scenarios than those presented here.

Author contributions. Helena Reid: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing (original draft preparation), Writing (review and editing). Cyril Morcrette: Conceptualisation, Funding acquisition, Supervision, Writing (review and editing).

Competing interests. The authors declare that they have no conflict of interest.

Code and data availability. The LFRic atmospheric model code can be found https://github.com/MetOffice/lfric_apps/ (UK Met Office, 2025). The ENNUF software used for translating ML models written in Python to Fortran can be found https://github.com/MetOffice/ennuf (Reid et al., 2025). A portion of the high resolution model output used as training data can be found https://doi.org/10.5281/zenodo.13332843 (Morcrette, 2025).

Acknowledgements. The authors acknowledge support from the Met Office and AI4PEX (UK Research and Innovation (UKRI); Horizon Europe, Grant/Award Numbers: 10114295,10103109, 10093450).

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Appendix A: Model level to height conversion

The mapping between model levels used in all experiments in this paper and heights above the surface is given in table A1.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Level	Height (m)	Level	Height (m)
0	0	36	9121
1	20	37	9622
2	53	38	10137
3	100	39	10667
4	160	40	11213
5	233	41	11775
6	320	42	12355
7	420	43	12954
8	533	44	13575
9	660	45	14221
10	800	46	14895
11	953	47	15602
12	1120	48	16348
13	1300	49	17137
14	1493	50	17980
15	1700	51	18884
16	1920	52	19861
17	2153	53	20923
18	2400	54	22087
19	2660	55	23369
20	2933	56	24789
21	3220	57	26371
22	3520	58	28141
23	3833	59	30130
24	4160	60	32371
25	4500	61	34904
26	4853	62	37771
27	5220	63	41022
28	5600	64	44712
29	5993	65	48902
30	6400	66	53659
31	6820	67	59060
32	7253	68	65187
33	7700	69	72133
34	8160	70	80000
35	8634		

Table A1. Levels and corresponding heights in meters

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

References

Adams, S., Ford, R., Hambley, M., Hobson, J., Kavčič, I., Maynard, C., Melvin, T., Müller, E., Mullerworth, S., Porter, A., Rezny, M., Shipway, B., and Wong, R.: LFRic: Meeting the challenges of scalability and performance portability in Weather and Climate models, Journal of Parallel and Distributed Computing, 132, 383–396, doi:10.1016/j.jpdc.2019.02.007, 2019.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, doi:10.1038/nature14956, 2015.

Berner, J., Shutts, G. J., Leutbecher, M., and Palmer, T. N.: A Spectral Stochastic Kinetic Energy Backscatter Scheme and Its Impact on Flow-Dependent Predictability in the ECMWF Ensemble Prediction System, Journal of the Atmospheric Sciences, 66, 603 – 626, https://doi.org/10.1175/2008JAS2677.1, 2009.

Berner, J., Achatz, U., Batte, L., Bengtsson, L., Cámara, A. d. l., Christensen, H. M., Colangeli, M., Coleman, D. R., Crommelin, D., Dolaptchiev, S. I., et al.: Stochastic parameterization: Toward a new view of weather and climate models, Bulletin of the American Meteorological Society, 98, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1, 2017.

Betts, A. K. and Miller, M. J.: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets, Quarterly Journal of the Royal Meteorological Society, 112, 693–709, https://doi.org/10.1002/qj.49711247308, 1986.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E.: The MOGREPS short-range ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 134, 703–722, https://doi.org/10.1002/qj.234, 2008.

Bright, D. R. and Mullen, S. L.: Short-Range Ensemble Forecasts of Precipitation during the Southwest Monsoon, Weather and Forecasting, 17, 1080 – 1100, https://doi.org/10.1175/1520-0434(2002)017<1080:SREFOP>2.0.CO;2, 2002.

Brown, A., Milton, S., Cullen, M., Golding, B., Mitchell, J., and Shelly, A.: Unified modeling and prediction of weather and climate: a 25 year journey, Bull. Amer. Meteor. Soc., 93, 1865–1877, doi:10.1175/BAMS-D-12-00018.1, 2012.

Buizza, R., Milleer, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 125, 2887–2908, https://doi.org/10.1002/qj.49712556006, 1999.

Bush, M., Boutle, I., Edwards, J., Finnenkoetter, A., Franklin, C., Hanley, K., Jayakumar, A., Lewis, H., Lock, A., Mittermaier, M., Mohandas, S., North, R., Porson, A., Roux, B., Webster, S., and Weeks, M.: The second Met Office Unified Model–JULES Regional Atmosphere and Land configuration, RAL2, Geoscientific Model Development, 16, 1713–1734, doi:10.5194/gmd-16-1713-2023, 2023.

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N., Mitchell, H. L., and Michelin, L.: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System, Monthly Weather Review, 138, 1877 – 1901, https://doi.org/10.1175/2009MWR3187.1, 2010.

Daleu, C. L., Plant, R. S., Stirling, A. J., and Whitall, M.: Evaluating the CoMorph-A parametrization using idealized simulations of the two-way coupling between convection and large-scale dynamics, Quarterly Journal of the Royal Meteorological Society, 149, 3087–3109, doi:10.1002/qj.4547, 2023.

Davies, L., Jakob, C., Cheung, K., Genio, A. D., Hill, A., Hume, T., Keane, R. J., Komori, T., Larson, V. E., Lin, Y., Liu, X., Nielsen, B. J., Petch, J., Plant, R. S., Singh, M. S., Shi, X., Song, X., Wang, W., Whitall, M. A., Wolf, A., Xie, S., and Zhang, G.: A single-column model ensemble approach applied to the TWP-ICE experiment, Journal of Geophysical Research: Atmospheres, 118, 6544–6563, doi:10.1002/jgrd.50450, 2013.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Derbyshire, S. H., Beau, I., Bechtold, P., Grandpeix, J.-Y., Piriou, J.-M., Redelsperger, J.-L., and Soares, P. M. M.: Sensitivity of moist convection to environmental humidity, Quarterly Journal of the Royal Meteorological Society, 130, 3055–3079, doi:10.1256/qj.03.130, 2004.

Donlon, C. J., Martin, M., Stark, J. D., Roberts-Jones, J., Fiedler, E., and Wimmer, W.: The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system, Remote Sensing of Environment, 116, 140–158, doi :10.1016/j.rse.2010.10.017, 2012.

Epstein, E. S.: A Scoring System for Probability Forecasts of Ranked Categories, Journal of Applied Meteorology (1962-1982), 8, 985–987, http://www.jstor.org/stable/26174707, 1969.

Gettelman, A. and Rood, R. B.: Demystifying Climate Models, Springer Berlin, Heidelberg, doi:10.1007/978-3-662-48959-8, 2016.

Giles, D., Briant, J., Morcrette, C. J., and Guillas, S.: Embedding machine-learnt sub-grid variability improves climate model precipitation patterns, Commun Earth Environ, 5, doi:10.1038/s43247-024-01885-8, 2024.

Gregory, D. and Rowntree, P. R.: A mass flux convection scheme with representation of cloud ensemble characteristics and stability-dependent closure, Mon. Wea. Rev., 118, 1483–1506, 1990a.

Gregory, D. and Rowntree, P. R.: A Mass Flux Convection Scheme with Representation of Cloud Ensemble Characteristics and Stability-Dependent Closure, Monthly Weather Review, 118, 1483 – 1506, https://doi.org/10.1175/1520-0493(1990)118<1483:AMFCSW>2.0.CO;2, 1990b.

Grell, G. A. and Freitas, S. R.: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling, Atmospheric Chemistry and Physics, 14, 5233–5250, 2014.

Gustavson, S.: Simplex noise demystified, 2005.

Hagos, S., Feng, Z., McFarlane, S., and Leung, L. R.: Environment and the Lifetime of Tropical Deep Convection in a Cloud-Permitting Regional Model Simulation, Journal of the Atmospheric Sciences, 70, 2409–2425, https://doi.org/10.1175/jas-d-12-0260.1, 2013.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, doi:10.1002/qj.3803, 2020.

Huber, P. J.: Robust Estimation of a Location Parameter, The Annals of Mathematical Statistics, 35, 73–101, http://www.jstor.org/stable/2238020, 1964.

Jankov, I., Berner, J., Beck, J., Jiang, H., Olson, J. B., Grell, G., Smirnova, T. G., Benjamin, S. G., and Brown, J. M.: A Performance Comparison between Multiphysics and Stochastic Approaches within a North American RAP Ensemble, Monthly Weather Review, 145, 1161 – 1179, doi:10.1175/MWR-D-16-0160.1, 2017.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, arXiv, https://arxiv.org/abs/1412.6980, 2014.

Krishnamurti, T. and Bounoua, L.: An Introduction to Numerical Weather Prediction Techniques, CRC Press, 1st edn., doi:10.1201/9781315137285, 1996.

Lambert, F. H., Challenor, P. G., Lewis, N. T., McNeall, D. J., Owen, N., Boutle, I. A., Christensen, H. M., Keane, R. J., Mayne, N. J., Stirling, A., and Webb, M. J.: Continuous Structural Parameterization: A Proposed Method for Representing Different Model Parameterizations Within One Structure Demonstrated for Atmospheric Convection, Journal of Advances in Modeling Earth Systems, 12, doi:10.1029/2020MS002085, 2020.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Li, W., D'Amico, D., Bernardet, L., Xue, L., Dudhia, J., Shin, H. H., Firl, G., Sun, S., Harrold, M., Nance, L. B., Ek, M., and Chu, Y.: Demonstrating Hierarchical System Development With the Common Community Physics Package Single-Column Model: A Case Study Over the Southern Great Plains, Meteorological Applications, 32, e70 073, doi:10.1002/met.70073, 2025.

Lin, J. W.-B. and Neelin, J. D.: Influence of a stochastic moist convective parameterization on tropical climate variability, Geophysical Research Letters, 27, 3691–3694, https://doi.org/10.1029/2000GL011964, 2000.

Lorenz, E. N.: Predictability: Does the Flap of a Butterfly's Wings in Brazil Set off a Tornado in Texas?, in: Proceedings of the American Association for the Advancement of Science, Washington address = Washington, D.C., 1972.

McCabe, A., Swinbank, R., Tennant, W., and Lock, A.: Representing model uncertainty in the Met Office convection-permitting ensemble prediction system and its impact on fog forecasting, Quarterly Journal of the Royal Meteorological Society, 142, 2897–2910, https://doi.org/10.1002/qj.2876, 2016.

Morcrette, C.: CRMML dataset, https://doi.org/10.5281/zenodo.13332843, accessed: 2025-12-17, 2025.

Morcrette, C., Cave, T., Reid, H., da Silva Rodrigues, J., Deveney, T., Kreusser, L., Van Weverberg, K., and Budd, C.: Scale-Aware Parameterization of Cloud Fraction and Condensate for a Global Atmospheric Model Machine-Learned From Coarse-Grained Kilometer-Scale Simulations, Journal of Advances in Modeling Earth Systems, 17, doi:10.1029/2024MS004651, 2025.

Morcrette, C. J.: ENNUF: Easy Neural Networks for Use in Fortran [Software], https://doi.org/10.5281/zenodo.12581295, 2024.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of Modelling Uncertainties in a Large Ensemble of Climate Change Simulations, Nature, 430, 768–772, https://doi.org/10.1038/nature02771, 2004.

Ollinaho, P., Lock, S.-J., Leutbecher, M., Bechtold, P., Beljaars, A., Bozzo, A., Forbes, R. M., Haiden, T., Hogan, R. J., and Sandu, I.: Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble, Quarterly Journal of the Royal Meteorological Society, 143, 408–422, https://doi.org/10.1002/qj.2931, 2017.

Palmer, T. N.: Stochastic weather and climate models, Nature Reviews Physics, 1, 463–471, https://doi.org/10.1038/s42254-019-0062-2, 2019.

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., Steinheimer, M., and Weisheimer, A.: Stochastic parametrization and model uncertainty, Tech. rep., ECMWF Reading, UK, https://www.ecmwf.int/sites/default/files/elibrary/2009/11577-stochastic-parametrization-and-model-uncertainty.pdf, 2009.

Perlin, K.: Improving noise, ACM Trans. Graph., 21, 681–682, https://doi.org/10.1145/566654.566636, 2002.

Puh, M., Keil, C., Gebhardt, C., Marsigli, C., Hirt, M., Jakub, F., and Craig, G. C.: Physically based stochastic perturbations improve a high-resolution forecast of convection, Quarterly Journal of the Royal Meteorological Society, 149, 3582–3592, doi:10.1002/qj.4574, 2023.

Reid, H., da Silva Rodrigues, J., Xirouchaki, T., and Morcrette, C.: ENNUF: Easy Neural Network for Use in Fortran, https://github.com/MetOffice/ennuf, accessed: 2025-12-17, 2025.

Sakradzija, M., Seifert, A., and Dipankar, A.: A stochastic scale-aware parameterization of shallow cumulus convection across the convective gray zone, Journal of Advances in Modeling Earth Systems, 8, 786–812, https://doi.org/10.1002/2016MS000634, 2016.

Sanchez, C., Williams, K. D., and Collins, M.: Improved stochastic physics schemes for global weather and climate models, Quarterly Journal of the Royal Meteorological Society, 142, 147–159, doi:10.1002/qj.2640, 2016a.

Sanchez, C., Williams, K. D., and Collins, M.: Improved stochastic physics schemes for global weather and climate models, Quarterly Journal of the Royal Meteorological Society, 142, 147–159, https://doi.org/10.1002/qj.2640, 2016b.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Shutts, G.: A kinetic energy backscatter algorithm for use in ensemble prediction systems, Quarterly Journal of the Royal Meteorological Society, 131, 3079–3102, https://doi.org/10.1256/qj.04.106, 2005.

Shutts, G. and Pallarès, A. C.: Assessing parametrization uncertainty associated with horizontal resolution in numerical weather prediction models, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 372, 20130284, https://doi.org/10.1098/rsta.2013.0284, 2014.

Smith, R. N. B.: A scheme for predicting layer cloud and their water content in a general circulation model, Quart. J. Roy. Meteor. Soc., 116, 435–460, 1990.

Stensrud, D. J.: Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models, Cambridge University Press, 2007.

Stevens, B., Satoh, M., Auger, L., and et al.: DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains, Progress in Earth and Planetary Science, 6, 61, https://doi.org/10.1186/s40645-019-0304-z, 2019.

Tennant, W. J., Shutts, G. J., Arribas, A., and Thompson, S. A.: Using a Stochastic Kinetic Energy Backscatter Scheme to Improve MOGREPS Probabilistic Forecast Skill, Monthly Weather Review, 139, 1190 – 1206, https://doi.org/10.1175/2010MWR3430.1, 2011.

Tompkins, A. M. and Berner, J.: A stochastic convective approach to account for model uncertainty due to unresolved humidity variability, Journal of Geophysical Research: Atmospheres, 113, doi:10.1029/2007JD009284, 2008.

UK Met Office: LFRic Atmospheric Model, https://github.com/MetOffice/lfric_apps/, accessed: 2025-12-17, 2025.

Walters, D. N., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P.-L., Wilkinson, J., Willett, M., Woolnough, S., and Xavier, P.: The Met Office Unified Model Global Atmosphere 6.0/6.1 and JULES Global Land 6.0/6.1 configurations, Geosci. Model Dev., 10, 1487–1520, doi:10.5194/gmd-10-1487-2017, 2017.

Walters, D. N., Baran, A., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Manners, J., Morcrette, C., Mulcahy, J., Sanchez, C., Smith, C., Stratton, R., Tennant, W., Tomassini, L., Weverberg, K. V., Vosper, S., Willett, M., Browse, J., Bushell, A., Dalvi, M., Essery, R., Gedney, N., Hardiman, S., Johnson, B., Johnson, C., Jones, A., Mann, G., Milton, S., Rumbold, H., Sellar, A., Ujiie, M., Whitall, M., Williams, K., and Zerroukat, M.: The Met Office Unified Model Global Atmosphere 7.0 and JULES Global Land 7.0 configurations, Geosci. Model Dev., 12, 1909–1963, doi: 10.5194/gmd-2017-291, 2019.

Watson, P. A. G., Berner, J., Corti, S., Davini, P., von Hardenberg, J., Sanchez, C., Weisheimer, A., and Palmer, T. N.: The impact of stochastic physics on tropical rainfall variability in global climate models on daily to weekly time scales, Journal of Geophysical Research: Atmospheres, 122, 5738–5762, doi:10.1002/2016JD026386, 2017.

Webster, P. J. and Lukas, R.: TOGA COARE: The Coupled Ocean–Atmosphere Response Experiment, Bulletin of the American Meteorological Society, 73, 1377 – 1416, doi:10.1175/1520-0477(1992)073<1377:TCTCOR>2.0.CO;2, 1992.

Webster, S., Uddstrom, M., Oliver, H., and Vosper, S.: A high resolution modelling case study of a severe weather event over New Zealand, Atmos. Sci. Let., 9, 119–128, doi: 10.1002/asl.172, 2008.

Weisheimer, A., Corti, S., Palmer, T., and Vitart, F.: Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 372, 20130290, https://doi.org/10.1098/rsta.2013.0290, 2014.

Wilson, D. R., Bushell, A. C., Kerr-Munslow, A. M., Price, J. D., and Morcrette, C. J.: PC2: A prognostic cloud fraction and condensation scheme. I: Scheme description, Quart. J. Roy. Meteor. Soc., 134, 2093–2107, 2008a.

https://doi.org/10.5194/egusphere-2025-6312
Preprint. Discussion started: 20 March 2026

Wilson, D. R., Bushell, A. C., Kerr-Munslow, A. M., Price, J. D., Morcrette, C. J., and Bodas-Salcedo, A.: PC2: A prognostic cloud fraction and condensation scheme. II: Climate model simulations, Quart. J. Roy. Meteor. Soc., 134, 2109–2125, 2008b.