The impact of performance filtering on climate feedbacks in a perturbed parameter ensemble

John W. Rostron¹ · David M. H. Sexton¹ · Carol F. McSweeney¹ · Kuniko Yamazaki¹ · Timothy Andrews¹ · Kalli Furtado¹ · Mark A. Ringer¹ · Yoko Tsushima¹

Received: 6 December 2019 / Accepted: 2 May 2020 / Published online: 30 May 2020
© The Author(s) 2020

Abstract

A key contribution to the latest generation of climate projections for the UK (UKCP18) was a perturbed parameter ensemble (PPE) of global coupled models based on HadGEM3-GC3.05. Together with 13 CMIP5 simulations, this PPE provides users with a dataset that samples modelling uncertainty and is ideal for use in impacts studies. Evaluations of global mean surface temperatures for this PPE have shown twenty-first century warming rates consistently at the top end of the CMIP5 range. Here we investigate one potential contributory factor to this lack of spread: that the methodology to select plausible members from a larger, related PPE of atmosphere-only experiments preferentially ruled out those predicted to have more negative climate feedbacks (i.e. lower climate sensitivities). We confirm that this is indeed the case. We show that performance in extratropical long-wave cloud forcing played a key role in this by constraining ice cloud parameters, which in turn constrained the feedback distribution (though causal links are not established). The relatively weak relationship driving this constraint is shown to arise from stronger relationships for the long-wave and short-wave cloud feedback components, which largely cancel out due to changes in tropical high clouds. Moreover, we show that the strength of these constraints is due to a structural bias in extratropical long-wave cloud forcing across the PPE. We discuss how choices made in the methodology to pick the plausible PPE members may result in an overly strong constraint when there is a structural bias and possible improvements to this methodology for the future.

Keywords Perturbed parameter ensembles · Climate feedbacks · Emergent constraints · Sensitivity analysis

1 Introduction

The most recent set of UK Climate Projections for land (UKCP18; Murphy et al. 2018) were developed to inform users about plausible twenty-first century changes to the UK climate, based on current understanding of uncertainties in future greenhouse gas concentrations, climate modelling and internal variability. UKCP18 contains an update to the set of probabilistic projections published in 2009 (UKCP09; Murphy et al. 2009), where known uncertainty sources were rigorously sampled. In UKCP18 there are now extra sources of information about climate change provided as raw model data from a set of global and regional climate models. These new datasets provide temporally, spatially and physically

coherent projections for several variables that are flexible to use, making them particularly well suited for regional impacts and adaptation studies.

The UKCP18 global projections comprise a set of 1900–2100 simulations drawn from two coupled model ensembles which sample distinct sources of uncertainty. One component is a multi-model ensemble (MME) of CMIP5 RCP8.5 simulations, which were used to sample uncertainties in modelling structures; the other is a new perturbed parameter ensemble (PPE) comprising plausible variants of a very recent configuration of the Met Office's coupled model: HadGEM3-GC3.05 (hereafter the 'coupled PPE'). The variants in the coupled PPE were generated through simultaneous perturbations made to 47 model parameters (across 7 atmospheric parameterisation schemes), chosen to sample key parametric modelling uncertainties. To account for uncertainty in carbon cycle feedbacks, each coupled PPE variant used a different prescribed CO₂ concentration pathway consistent with CMIP5 RCP8.5 emissions (Murphy

✉ John W. Rostron
john.rostron@metoffice.gov.uk

¹ Met Office Hadley Centre, FitzRoy Road, Exeter EX1 3PB, UK

et al. 2018). The unperturbed ‘standard’ PPE member used the standard RCP8.5 pathway. The combination of the CMIP5 and coupled PPE datasets was designed to sample distinct sources of modelling uncertainty, with internal variability uncertainties also being sampled in both ensembles.

The methodologies used to select the final variants also differed between the ensembles. The full set of CMIP5 models represent an ‘ensemble of opportunity’, bringing together individual variants of the most up-to-date GCMs from different modelling centres. The 13-member subset of these models used in UKCP18 were selected qualitatively based on their performance in key aspects of global and European/UK climate, along with a screening of very closely related models (Murphy et al. 2018).

The process for selecting the 15 coupled PPE variants is described in Sexton et al. (2020, submitted) and Yamazaki et al. (2020, submitted). A thorough sampling of the 47-dimensional parameter space was not feasible with coupled simulations, so based on the ideas of Karmalkar et al. (2019), Sexton et al. (2020, submitted) outline a methodology to filter the space with a PPE of computationally cheap simulations running only the atmospheric component of HadGEM3-GC3.05 at a coarser resolution (hereafter the ‘atmosphere-only PPE’). Starting from a large ensemble containing nearly 3000 members, atmosphere-only experiments were used to filter out variants with poor large-scale performance across multiple variables on 5-day and 5-year timescales (referred to hereafter as ‘performance metrics’), and additionally in their representation of North Atlantic and European climate (e.g. winter circulation, seasonal cycles of UK precipitation and 1.5 m temperature). A set of 25 of the remaining variants were selected for running as coupled simulations, as this was an affordable ensemble size given supercomputing resources. The 25 were chosen to broaden diversity in several climate responses (e.g. climate feedbacks, CO₂ and aerosol forcings), diagnosed using idealised atmosphere-only forcing experiments. This 25-member coupled PPE was reduced to 15 for use in UKCP18, following further screening of the coupled simulations themselves based on criteria such as: numerical stability; the strength of the Atlantic meridional overturning circulation (AMOC); historical trends in northern hemisphere surface air temperatures; and climatological temperature biases (Yamazaki et al. 2020, submitted).

Following the completion of the 1900–2100 simulations of the 15-member coupled PPE, assessments of global mean surface temperature (GMST) showed that future changes are distributed around a relatively high level of warming compared to the 13 CMIP5 models, with very little overlap between the two ensembles (Murphy et al. 2018; Yamazaki et al. 2020, submitted). Furthermore, the spread in GMST warming anomalies for the end of the twenty-first Century is about 2 °C for the coupled PPE, with the majority of the

variance (around 60%) driven by the differences in the prescribed CO₂ concentrations. By comparison, the spread for the CMIP5 models (where the concentrations pathways are not varied) is around 2.6 °C, suggesting a relatively narrow spread in global feedbacks for the coupled PPE.

Global feedbacks in coupled models are often assessed using idealised experiments where the CO₂ concentrations are instantly quadrupled and fixed for 150 years (Andrews et al. 2012; Forster et al. 2013). For a 25-member PPE, these idealised experiments require a control run of the same length and are expensive to run. Recognising this, Sexton et al. (2020, submitted) used a set of cheaper idealised atmosphere-only experiments with prescribed SSTs that include a 4 K global warming pattern from the CMIP5 “amipFuture” experiment (Taylor et al. 2012) to assess the spread of atmosphere processes that affect global feedbacks. Ringer et al. (2014) showed that such experiments are a good guide to the global cloud feedbacks that dominate the spread in the coupled atmosphere–ocean CO₂-forced simulations. Conceptually, the atmosphere-only runs reveal what the feedbacks are for a prescribed pattern of SST warming when coupling is turned off and sea ice is not changed. In coupled mode, there are additional processes via air–sea interactions (Hyder et al. 2018), changes in sea ice, and evolution of the SST pattern which can alter the feedbacks (Andrews et al. 2015). However, Ringer et al. (2014) showed that the atmosphere-only component provides an appreciable contribution to the overall global feedbacks.

Any reduction in the spread of climate feedbacks is potentially a contributing factor to a lack of diversity in the GMST warming anomalies in the coupled PPE. We see two occurrences of this in the generation of the 15-member coupled PPE. First, Sexton et al. (2020, submitted) shows that model performance-based filtering reduced the range of feedback values in the atmosphere-only experiments. Second, Yamazaki et al. (2020, submitted) shows that out of the 10 model variants that were ruled out after assessing the coupled simulations, six of these were amongst the seven with the most negative global feedback estimates (i.e. the smallest expected warming). In this paper, we assess the first of these—the impact of the performance filtering on the atmosphere-only component of the global feedback—recognising that we cannot assess the relative contribution of the atmosphere-only and coupling effects until we run a set of idealised coupled CO₂ forcing experiments. Still, this provides an opportunity to better understand how our method for filtering out implausible parameter combinations interacts with our aim of producing a diverse set of simulations, with a view to improving the method in the future.

We will do this by assessing how performance filtering affected distributions of the feedback components, the performance metrics responsible for producing the largest changes to these distributions, and the robustness of the

filtering methodology used in Sexton et al. (2020, submitted). The advantage of using a feedback framework is that it allows us to break down the net response into components of long-wave and short-wave (LW, SW), clear-sky and cloud feedbacks, enabling a more in-depth analysis of the processes driving the constraints. We plan to run an idealised quadrupled CO₂ experiment for each coupled PPE member. When these are complete, an analysis of these feedbacks will enable a comparison of responses estimated using the coupled and atmosphere-only PPEs, for example by clearly exposing coupled processes such as sea-ice feedbacks, which project primarily onto the SW clear-sky component.

A key aspect of this study is the concept of emergent constraints. These are based on relationships between observable present-day quantities and future responses in a variable of interest, using GCM simulations. By ruling out models with present-day quantities deemed inconsistent with observations, constraints can then be placed on the distribution of the future responses. In this paper we refer to the changes in these distributions as ‘emergent constraints’ and the relationships that underpin them as ‘emergent relationships’. This terminology may depart from other definitions of emergent constraints (e.g. Caldwell et al. 2018) but it is useful for our study as it acknowledges that an emergent relationship can exist, but that the performance filtering could be ineffective and so result in little or no constraint. Indeed, we have found examples of this (not shown).

The aim of this paper is to explore the emergent constraints applied to members of the atmosphere-only PPE of 5-year runs through the performance filtering methodology of Sexton et al. (2020, submitted). Emergent relationships and constraints have been studied widely within climate science. Examples include studies of climate feedbacks, equilibrium climate sensitivity and aerosol radiative forcings; with MMEs (e.g. Klein et al. 2013; Sherwood et al. 2014; Cherian et al. 2014) as well as PPEs (e.g. Klocke et al. 2011; Wagman and Jackson 2018; Johnson et al. 2018) being used. However, whilst they can provide a useful framework for understanding model biases and constraining responses, ongoing research has highlighted several complications. A particular issue is the reliability of the relationships: often they are found without a clear physical mechanism (particularly for combined quantities like equilibrium climate sensitivity; Caldwell et al. 2018) and lack robustness across different ensembles (Sanderson 2011; Kamae et al. 2016; Caldwell et al. 2018; Wagman and Jackson 2018). Our 5-year atmosphere-only PPE provides an ideal opportunity to further study emergent relationships and constraints and their complexities, adding value by allowing us to assess the parameters, and therefore the processes, underpinning them.

The remainder of the paper is ordered as follows: Sect. 2 provides a summary description of the design of the PPE, though most of the details are to be found in Sexton et al.

(2020, submitted). This includes the model structure on which it is based; the experiments used to diagnose climate feedbacks and present-day performance; and a description of emulation analyses used throughout the paper. Section 3 covers our main results, starting with the distributions of feedback components and the impact that performance filtering has on these, before studying in more detail the performance metrics driving the constraints, and the model processes underlying these relationships. In Sect. 4 we discuss the robustness of the constraints presented in the context of structural errors evident in the PPE, and the key lessons that can be drawn for future work with ensembles. This is followed by an overall summary in Sect. 5.

2 Models and methods

The technical details and justification of choices behind the design of the atmosphere-only PPE can be found in Sexton et al. (2020, submitted). Here, we summarise the main points that are relevant to understanding our analysis.

2.1 The base model and parameters to perturb

The base model we used for the coupled PPE is a modified version of the UK Hadley Centre Unified Model HadGEM3-GC3 (Williams et al. 2018). The atmospheric component uses 85 vertical levels and has a horizontal resolution of approximately 60 km at mid-latitudes (called ‘N216’). The ocean component has a horizontal resolution of ¼ degree. The enhanced resolution (compared with most CMIP5 models) has been shown to better capture the processes that are needed to represent variability and extremes over the North Atlantic/Europe sector (Scaife et al. 2011, 2012).

Our modifications include a subset of the changes made to the atmospheric component of the coupled model which was submitted to CMIP6 (HadGEM3-GA7.1 and HadGEM3-GC3.1, respectively; Williams et al. 2018; Walters et al. 2019). These were developed to address the strong cooling to anthropogenic aerosols seen in HadGEM3-GA7 (Mulcahy et al. 2018), with the largest impacts resulting from:

- An improvement in the cloud droplet spectral dispersion relationship used to calculate cloud effective radius (Liu et al. 2008).
- Updates to the absorption properties of black carbon and an increased spectral resolution of the look-up tables used for ARI calculations (Bond and Bergstrom 2006).
- An update to the scaling factor for natural marine emissions of dimethylsulphide (DMS) from 1.0 to 1.7. This was applied to account for missing natural emissions of organic aerosol.

Our configuration includes the first 2 of these, whilst the scaling factor for natural marine emissions of DMS is one of the parameters we perturb (Sexton et al. 2020, submitted). Other changes made for GA7.1, which we did not include in our configuration, were found to have smaller impacts on the aerosol forcing (see Table 1 in Mulcahy et al. 2018). Because we include some, but not all, of the changes for GC3.1/GA7.1, we refer to our coupled configuration as HadGEM3-GC3.05, and our atmospheric component as HadGEM3-GA7.05.

Building a large PPE of fully coupled HadGEM3-GC3.05 variants to evaluate climate performance and sub-select 25 parameter combinations would not have been practical, given the timescale and resources of the UKCP18 project. As a result, a larger atmosphere-only PPE was built based on HadGEM3-GA7.05 at a lower resolution of approximately 135 km at mid-latitudes ('N96'), under the assumption that these models would be informative about climate performance in coupled mode at the higher resolution. Under the Met Office's approach to seamless model development, assessments are carried out across a range of resolutions (Williams et al. 2018), with any significant differences being noted. Walters et al. (2019) show that for HadGEM3-GA7 only a small number of differences in performance arise between the 60 km and 135 km resolutions (for example, in temperature and humidity in the tropical tropopause layer). Additionally, Andrews et al. (2019) show that global feedbacks and climate sensitivities are very similar in the coupled configurations of HadGEM3-GC3.1 using these atmospheric resolutions.

The choice of 135 km resolution ensured that we could build a large enough ensemble to thoroughly test performance on 5-day and 5-year timescales, explore a wide range of climate responses, and ultimately allow the selection of 25 members that would stand a good chance of being plausible and diverse in their coupled configurations.

The atmosphere-only PPE was created by simultaneously perturbing 47 model parameters across seven schemes, namely: boundary layer, gravity wave drag, convection, cloud microphysics, cloud and cloud radiation, aerosol chemistry and land surface. Sexton et al. (2020, submitted) describe the extensive elicitation exercise that was carried out with modelling experts to select these parameters. An important part of this was the definition of prior probability density functions (PDFs) for the parameters, which captured the level of confidence the experts had in parameter values across their plausible ranges (see Fig. 2 in Sexton et al. 2020, submitted). These choices were largely based on the judgement of the experts, factoring in any existing targeted parameter sensitivity studies (e.g. for aerosol parameters).

A full description of the parameters perturbed is given in Sexton et al. (2020, submitted), but Table 1 gives a summary of the key parameters explored in this study. Each version of

the atmosphere-only PPE discussed in this paper contains a member with parameter values set to the unperturbed, standard values for GA7.05 configuration. We refer to this as the 'standard' member.

2.2 Experiments

A series of 5-day hindcast and 5-year climate experiments were run to diagnose model performances at the two time-scales, future climate responses, and internal variability. The 5-day hindcasts were based on the Transpose-AMIP II protocol (Williams et al. 2013) and the analysis of these runs is described Sexton et al. (2020, submitted); we do not analyse these experiments in this paper. We focus only on a subset of the 5-year experiments. These were:

1. **ATMOS experiment:** An amip-like experiment (Taylor et al. 2012) used to assess present-day climate performance, covering years 2004–2009.
2. **SSTFuture experiment:** An amipFuture-like experiment (Taylor et al. 2012) used to diagnose climate feedbacks (see below for method). SSTFuture is equivalent to ATMOS, except that a prescribed pattern of SST warming is added, with a global mean of 4 K. The pattern of SST warming is the same as that used in the CMIP5 amipFuture experiment (Taylor et al. 2012). Sea-ice is not modified from the ATMOS experiment, and so feedback estimates do not include effects from sea-ice changes.
3. **Stochastic experiments for ATMOS and SSTFuture:** For both the ATMOS and SSTFuture experiments, 32 realisations of the standard (i.e. unperturbed) model were run, which differed only in the random seed used to initialise the model's stochastic schemes. Variances across the 32 members provide estimates of internal variability, which we have assumed applies across parameter space.

Climate feedbacks can be estimated from the SSTFuture experiment by assuming the radiative response (downwards; R) to a perturbation scales linearly with changes in the global mean of surface temperature (ΔT):

$$R = \lambda \Delta T \quad (1)$$

The scaling factor, λ , is the climate feedback parameter. The patterned SST increase imposed in the SSTFuture experiment induces a radiative response, from which climate feedbacks are quantified using $\lambda = R/\Delta T$.

Our global feedback estimates were made using 5-year averages of R and ΔT , where the changes in these quantities are measured relative to the ATMOS experiment. Along with the net feedback, we also calculate short- and long-wave (SW, LW) clear-sky and cloud feedback components, using the corresponding radiative flux changes in each case.

The latter are determined using the cloud radiative effect (CRE) i.e. differences in all-sky and clear-sky fluxes.

We find the range of SW clear-sky feedback values explored by the PPE is much smaller than that for the LW clear-sky (Fig. 1). This results from the missing sea-ice albedo (primarily SW clear-sky) feedback, due to the lack of change in sea-ice between the ATMOS and SSTFuture experiments. We also find that the LW and SW clear-sky components are highly correlated ($r=0.82$), suggesting common drivers of their spread. For these reasons we choose to analyse only the sum of the LW and SW clear-sky feedbacks (the net clear-sky feedback) in this study.

2.3 Sampling the parameter space

The atmosphere-only PPE explored a total of 2915 different parameter combinations and was built in several stages, as described in Sexton et al. (2020, submitted). The parameter choices for the first 2800 members were based on Latin hypercube sampling of the prior PDFs for the 47 parameters. These members were filtered based on their performance in the 5-day hindcast experiments, leaving 442 variants.

ATMOS and SSTFuture experiments were run for these remaining variants, and the 406 members that completed both experiments provide the starting point for our analysis. We refer to this as the ‘Wave-1’ PPE (equivalent to

“ATMOS-Wave1” in Sexton et al. 2020, submitted). ATMOS and SSTFuture experiments were not run for the members ruled out using 5-day hindcast performance, so we do not have direct estimates of their feedbacks. Our analysis of the effect of performance filtering is therefore only based on the 5-year ATMOS and SSTFuture experiments.

Sexton et al. (2020, submitted) describe a methodology by which additional PPE members were selected to augment Wave-1. A second wave (‘Wave-2’) of 100 members were chosen that had a better chance of (a) not crashing, (b) passing the performance filtering and (c) sampling a diverse range of responses to future SST warming, CO₂ concentration rises and pre-industrial aerosol forcing. A further wave (‘Wave-3’) of 15 members were selected to target more negative climate feedbacks and weaker aerosol forcing. In this paper we have chosen to focus our analysis on Wave-1 only. This is because we aim to explore the constraints on feedbacks imposed by present-day climate performance, by understanding relationships between parameters, performance metrics (involving observable quantities), and feedbacks that are general across parameter space. Distributions based on Wave-1 will be as close as we can achieve to obtaining an unbiased sample of parameter combinations after filtering based on the 5-day hindcasts. In contrast, Wave-2 and Wave-3 are designed to give more weight to the tails of the distribution of feedbacks whilst maintaining

Fig. 1 Feedback components from the Wave-1 PPE are shown as box-and-whisker plots: the 406 members that completed both the ATMOS and SSTFuture experiments are shown in black, while the 41 members that passed the 5-year mean climate performance filtering (from the ATMOS experiment) are shown in red. The median,

25th and 75th percentiles, and minima and maxima are displayed in each case. Also shown are individual model values from the 32-member stochastic ensemble (green crosses) and the 11-member CMIP5 atmosphere-only ensemble (black crosses)

plausible present-day performance, and so will bias the distributions from Wave-1. Therefore, with the exception of the training of emulators (see Sect. 2.5), we do not use Wave-2 and Wave-3 in our analysis.

2.4 Filtering on present-day climate performance

As noted in the Introduction, the motivation for this study is to explore the constraints on the feedbacks in the atmosphere-only PPE, applied by performance filtering based on the ATMOS experiment. The performance of each PPE member is assessed using mean-squared-errors (MSEs) for a broad set of variables typically used to evaluate the general atmospheric performance of climate models. These include: temperature, precipitation, surface pressure, winds, all-sky and cloud radiative fluxes, and ISCCP cloud amounts (estimated in a comparable way to the ISCCP satellite; Rossow and Schiffer 1999; Bodas-Salcedo et al. 2008). In order to capture performance in different spatial regimes, the MSEs are calculated over large-scale regions (using all grid-boxes; see Eq. (2) below). For example, most variables are horizontal 2-day fields and MSEs are estimated over the following regions: global; tropical land and ocean; NH extratropical land and ocean; and SH extratropical land and ocean. To capture different temporal regimes, MSEs are calculated using both seasonal and annual means. Here, we use 925 combinations of variable, season and region and we refer to each one as a ‘performance metric’. One minor difference to the filtering described in Sexton et al. (2020, submitted) is that we do not analyse the metric for global aerosol optical depth at 550 nm. This metric required a different treatment in the filtering analysis and was found to have little effect, and so for simplicity we omit it in this paper.

The actual performance metric used is an MSE normalised by the sum of four terms representing contributions from key uncertainties that are often accounted for in filtering PPEs (e.g. Williamson et al. 2013) or in likelihood-weighting (e.g. Sexton et al. 2012). We use normalisation to ensure that each dimension in the space of model-performance metrics is fairly weighted, in accordance with an estimate of its uncertainty.

This normalised MSE (nMSE) is defined as:

$$nMSE = \frac{MSE}{(\sigma_{struc}^2 + \sigma_{meas}^2 + \sigma_{lobs}^2 + \sigma_{IV}^2)^{1/2}} = \frac{\frac{1}{n} \sum_{i=1}^{i=n} w_i (m_i - o_i)^2}{(\sigma_{struc}^2 + \sigma_{meas}^2 + \sigma_{lobs}^2 + \sigma_{IV}^2)^{1/2}} \quad (2)$$

where m_i and o_i are simulated and observed temporal mean values at each grid point in the region (where observations exist); w_i are the area weights and n is the number of grid

points in the region. The normalisation terms are based on variances of MSE estimates and are defined as follows:

- (i) σ_{struc}^2 : Structural error uncertainty. Represents the structural limitations of using approximations to real processes in the model. This term is used to down-weight the influence of variables that are poorly represented in the climate model and that therefore have large structural errors. There is no clear way to specify this term, and we do this by using the MSE value for the standard variant (but see Sect. 4.1 and Sexton et al. (2020, submitted) for a discussion of this definition).
- (ii) σ_{meas}^2 : Observational measurement uncertainty. Represents uncertainties in observational measurements and their representativeness of the modelled quantity. This is calculated using the differences between two observational datasets, if available, or as a scaling of σ_{struc}^2 if not (see Sexton et al. (2020, submitted) for details).
- (iii) σ_{IV}^2 : Modelled internal variability. This is estimated using the variance in MSE across the stochastic ensemble and is assumed to be constant across parameter space
- (iv) σ_{lobs}^2 : Real world internal variability. Represents the fact that the (real world) observations sample internal variability about a forced signal. This is assumed to be the same as σ_{IV}^2 .

The dominant term in the normalisation in Eq. (2) is the structural error uncertainty (see Sexton et al. 2020, submitted), which is also the most difficult to estimate. It is our choice of using the standard variant to specify this term that most greatly affects the results in Sect. 3. The effect of this choice is discussed in Sect. 4.

The performance filtering of the PPE was carried out using thresholding across the performance metrics. For each member, if any nMSE metric exceeded a threshold value (“T”, chosen to be 3.7; Sexton et al. 2020, submitted) then that member was ruled out. The downside of this approach is that it could accept models that only just lie within the threshold across several metrics. To counter this, a soft threshold (S) was applied whereby members could also be ruled out if n_{soft} of their metric values exceeded S. Values of 3.0 and 8 were used for S and n_{soft} respectively. However, as shown in Table 6 in Sexton et al. (2020, submitted), it was the hard threshold that dominated the performance filtering: of the 446 ATMOS Wave 1–3 members that were ruled out, 442 were ruled out by the hard threshold, while only 4 were ruled out by the soft threshold.

2.5 Emulation

Throughout this paper we make use of emulators. These are statistical models, trained on the actual simulations, which relate specific model outputs to parameter values over the 47-dimensional parameter space. The sampling for the actual simulations is sparse and we often want to explore parameter space more thoroughly and/or systematically. Emulators allow us to do this.

We build smooth, continuous models to predict feedback and performance metric MSE values across the 47-dimensional parameter using Gaussian Process emulators (Roustant et al. 2012). The statistical model is an adapted version of the one in Lee et al. (2013) and described in detail in Sexton et al. (2019). In summary this consists of a linear model, based on a subset of the parameters determined by a stepwise algorithm, and a correlation matrix of the error, which depends on a decorrelation length scale for each parameter. To account for internal variability in the data, the stochastic experiments are used to specify a nugget term, which reduces the chances of the emulator fitting to noise. We train the emulators using all the available simulations from Waves 1–3 (but note that we only sample these emulators using Wave-1-like distributions—see Sect. 2.5.1). For the performance metrics these are a set of 518 members that completed the ATMOS experiment, whilst for the feedbacks we use the 509 members that completed both the ATMOS and SSTFuture experiments.

Emulators have been used widely in climate science (e.g. Rougier et al. 2009; Sexton et al. 2012; Williamson et al. 2013; Regayre et al. 2018; Johnson et al. 2018; Karmalkar et al. 2019). Their key advantage for our work (over the actual simulations) is that they provide predictions of model output at any point in parameter space, not just the parameter combinations sampled by the PPE itself. We can therefore train the emulators on the actual simulations and then estimate values of performance metric MSEs and feedbacks without the need for further simulations. This gives us the flexibility to sample parameter space in three different ways. Firstly, we can make more robust estimates of the effect of individual performance metrics on the performance filtering, by exploring the parameter space more thoroughly than the actual simulations allow (see Sect. 3.3). Secondly, we can carry out simple sensitivity analyses to estimate the effect of individual parameters on feedbacks and performance metrics, to elucidate their role in the emergent relationships (see Sect. 3.5).

Thirdly, we can use emulators to perform more comprehensive sensitivity analyses. In this paper we use the method of Saltelli et al. (1999) to estimate the fraction of the variance of the emulator's best-fit explained by each parameter for any model output (see Sect. 3.5). This method separates the 'total effect' of each parameter into two components:

a 'main effect', which describes the individual (potentially non-linear) effect of the parameter, and an 'interaction effect', which sum the effects of its interactions with all the other parameters.

A key limitation of emulators, however, is that they are not actual simulations—they are a best fit that will depend on many, variable-specific factors e.g. the amount of training data, the number of 'important' parameters, the validity of our assumption that internal variability is constant across parameter space, and the choices affecting the fitting of the Gaussian Process model. To assess the validity of our emulators we performed leave-one-out tests. These are described in Appendix 1 for the key model outputs analysed in this study. We find the emulators generally provide valid fits to all but a few of the members used in the training data set.

Ultimately, we are interested here in studying the constraints that were placed on the actual simulations, which may have influenced the reduced diversity seen in the coupled PPE. Therefore, we focus our analysis on the actual simulations, but use emulators as a tool to augment our understanding where it is appropriate to use them.

2.6 Emulated parameter distributions

We use a large ($N = 64,572$) set of parameter perturbations to sample our emulators, providing a dense sampling of the parameter space. This sample is based on a larger ($N = 2.5 \times 10^5$) Monte Carlo sample of the elicited (prior) parameter distributions, which was subsequently filtered using emulators to rule out parameter combinations that were likely to crash (due to model instabilities) or show poor performance in the 5-day hindcasts. The parameter distributions in the $N = 64,572$ sample broadly reflect those of the actual Wave-1 simulations. Differences may result from model crashes across the 5-year experiments, but visual inspection of the distributions showed this effect to be small. We therefore use the $N = 64,572$ Monte Carlo sampling of emulators to complement the analysis of the actual Wave-1 simulations, where appropriate, and we refer to this as the 'emulated' Wave-1 distribution.

2.7 Emulated performance constraints

In Sect. 3.3 we use emulators to identify the key performance metrics that drive changes in the feedback distributions. As part of this analysis we applied performance filtering to emulated MSE values in an equivalent way to the actual simulations (Sect. 2.4), but with the following differences:

- (i) Emulators were trained on $\log(\text{MSE})$ values in order to stabilise the variance for high MSEs.

- (ii) Thresholding was then applied according to the following equation:

$$\mu_{EM} - \alpha \Sigma_{EM}^{1/2} > \log[T(\sigma_{struc}^2 + \sigma_{meas}^2 + \sigma_{lobs}^2 + \sigma_{IV}^2)^{1/2}] \quad (3)$$

where μ_{EM} and Σ_{EM} are the mean and variance of the emulated prediction of $\log(\text{MSE})$. The $\alpha \Sigma_{EM}^{1/2}$ term is used to ensure that any threshold exceedances are statistically significant, with a value of $\alpha = 1.96$ being used (Sexton et al. 2020, submitted). Other terms in this equation are the same as in Eq. (2). An equivalent expression was used to apply the soft thresholding (with $S = 3.0$ replacing $T = 3.7$, and $n_{\text{soft}} = 8$, as for the actual simulations).

3 Results

Our aim in this section is to understand how performance filtering of the Wave-1 PPE affected its feedback component distributions (Sects. 3.2, 3.3), and to analyse the relationships and processes driving these constraints (Sects. 3.4, 3.5). To provide context for the filtering analysis, we start with a summary of the distributions of the feedback components from the Wave-1 PPE (before filtering), along with a comparison to CMIP5 atmosphere-only models.

3.1 Summary of feedbacks in the atmosphere-only PPE

Global climate feedback values from the Wave-1 PPE are summarised in Table 1 and shown in Fig. 1. Net feedback values vary between -1.80 and $-0.93 \text{ Wm}^{-2} \text{ K}^{-1}$, covering a much larger range than the stochastic ensemble, which varies between -1.30 to $-1.22 \text{ Wm}^{-2} \text{ K}^{-1}$. Similar effects are seen across all the feedback components (comparing green crosses to black box-and-whiskers in Fig. 1), indicating that the parameter perturbations are the main source of the spread in feedbacks across the PPE, rather than internal variability.

The range of net feedbacks explored in the PPE is comparable to that from a set of 11 CMIP5 atmosphere-only models, which are based on amipFuture and amip experiments (Ringer et al. 2014). The CMIP5 values sample a similar range to the PPE (-1.85 to $-1.04 \text{ Wm}^{-2} \text{ K}^{-1}$), but this similarity masks some clear differences between the ensembles. For example, Fig. 1 shows that the high end of the CMIP5 range is dominated by a single model (IPSL-CM5A-LR), with the next highest value ($-1.42 \text{ Wm}^{-2} \text{ K}^{-1}$) lying close to the PPE median. This means that over 50% of PPE members have less negative values (i.e. higher climate sensitivities) than 10 of the 11 CMIP5 models and that the PPE provides a much more thorough sampling of the upper end of the net feedbacks.

A similar, but more pronounced, situation is seen for the net CRE component where the PPE does not sample any negative feedback values, in contrast to CMIP5. This is a result of a partial cancellation of large, but oppositely signed, LW and SW CRE feedbacks, with the positive net CRE feedback values resulting from the larger magnitude in the SW CRE component. (We discuss this cancellation, which is due to the effect of high cloud, in more detail in Sect. 3.5.2.) The differences to the CMIP5 ensemble are even larger here: the stochastic ensemble lies outside of the CMIP5 ranges in both components, while there is only minimal overlap between the PPE and CMIP5 (indeed there is no overlap in the LW CRE feedbacks). This has been a feature of recent Hadley Centre models (Senior et al. 2016; Andrews et al. 2019), while Bodas-Salcedo et al. (2019) have also shown an increase (decrease) in SW (LW) CRE feedbacks in HadGEM3-GA7.1 compared to its predecessor, HadGEM3-GA6. The latter study attributed SW CRE feedback changes to new aerosol and mixed-phase cloud schemes (Mann et al. 2010; Furtado et al. 2016), both of which are included in the GA7.05 configuration, highlighting the importance of structural model changes in the representation of cloud feedbacks.

For net clear-sky feedbacks, the stochastic ensemble and the bulk of PPE members agree well with the CMIP5 ensemble. However, there is a clear tail to the PPE distribution sampling large negative feedback values that are well outside of the CMIP5 range. In the PPE, this is mainly due to the LW component, as sea ice is the same in the ATMOS and SSTfuture experiments. However, there are variations in the CMIP5 short-wave clear sky that may be due to differences in how the sea ice is treated in the amipFuture experiment.

3.2 The impact of filtering on PPE feedbacks

The 5-year climate performance filtering described in Sect. 2.4 ruled out 365 of the 406 Wave-1 PPE members, leaving 41 plausible variants. We refer to this as the ‘filtered’ Wave-1 PPE. The impact of this filtering on the feedback components is summarised by the black and red box-and-whisker plots in Fig. 1, whilst Fig. 2 shows the normalised histograms of these distributions. Median, maximum and minimum values are given in Table 2.

For the net feedback, the distribution changes show a reduction in the range across the PPE after filtering (as highlighted in Sexton et al. 2020, submitted), accompanied by a shift towards less negative feedback values (i.e. towards higher climate sensitivities). Clearer changes are seen in the cloud components, where variants with weaker negative LW CRE feedbacks and weaker positive SW CRE feedbacks are ruled out. Consequently, the distributions for these components shift to even more negative and positive feedback values, respectively, enhancing the already

Table 1 Descriptions of the perturbed parameters in the coupled and atmosphere-only PPEs which are most prominent in this analysis

Parameter	Scheme	Role	Signature effect
Deep entrainment amplitude (ent_fac_dp)	Convection	Controls the shape of the mass flux and the sensitivity of deep convection to relative humidity	Increased values lead to the reduction of convection depth and to some extent suppression of active precipitating convection
Minimum critical cloud condensate (qlmin)	Convection	The minimum value of the function that defines the maximum amount of condensate a convective parcel can hold before it is converted to precipitation	Reducing it cools the troposphere, increasing it warms the troposphere, by decreasing/increasing the amount of high cloud
Cloud erosion rate (dbsdtbs_turb_0)	Cloud and cloud radiation	Determines the rate with which un-resolved sub-grid motions mix clear and cloudy air and hence remove liquid condensate and evaporate liquid cloud fraction	Modifies the radiative properties of the cloud (especially in regions of liquid only cloud e.g. sub-tropical maritime stratocumulus) and also affects the in-cloud liquid water content and hence how the precipitation formation processes will evolve
Scaling factor applied to empirically derived one-dimensional cloud condensate variability to make it represent two-dimensional variability (two_d_fsd_factor)	Cloud and cloud radiation	Accounts for fact that cloud-condensate variability estimated from one-dimensional (constant in height) samples of CloudSat data, are likely to be an under estimate of the true two-dimensional variability at that height (Hill et al. 2015)	Changes the cloud-radiation interaction and hence the LW/SW radiative balance. Increasing this parameter from 1.4 to 1.5 reduces the amount of outgoing SW at the top of the atmosphere by around 1.2 Wm^{-2} and increases the outgoing LW radiation by 0.4 Wm^{-2}
Decorrelation scale pressure (dp_corr_strat)	Cloud and cloud radiation	Determines the vertical overlap between clouds in the sub-column in the cloud generator used to calculate the radiative impact of clouds	Alters the cloud radiative effect. High values mean the cloud is more maximally overlapped and its radiative effect reduced. Impacts are seen most clearly in convective regions
Ice fall speed (m_ci)	Cloud microphysics	Scaling factor for the ice fall speed	Increasing the fall speed will decrease ice water content
Precursor coefficient in the mass-diameter relationship for ice ($m = ai \times Dbi$) (ai)	Cloud microphysics	Changing ai has the effect of changing the density of the ice	Increasing ai will produce a narrower particle size distribution (PSD) and so the mass weighted fall speed will be lower and hence the cloud ice content should increase

The second column gives the parameterisation scheme to which the parameter belongs, and the third and fourth columns the physical role of each parameter, and the main effects of varying its value. A table of all the parameters perturbed can be found in Sexton et al. (2020, submitted)

Fig. 2 Normalised distributions of the feedback components from the Wave-1 PPE. The full distributions (406 members) are shown in black, while the filtered distributions (41 members) are shown in red.

The scales for the x-axes are the same in each plot, but the values sampled differ (depending on the typical values for each component)

Table 2 Summary of feedback component values

Feedback component [$\text{Wm}^{-2} \text{K}^{-1}$]	Wave-1 PPE [N=406]	Filtered Wave-1 PPE [N=41]	Stochastic ensemble [N=32]	CMIP5 atmosphere-only [N=11]
NET	− 1.40 [− 1.80, − 0.93]	− 1.36 [− 1.57, − 1.13]	− 1.27 [− 1.30, − 1.22]	− 1.59 [− 1.85, − 1.04]
NET clear-sky	− 1.73 [− 2.26, − 1.40]	− 1.73 [− 1.93, − 1.50]	− 2.31 [− 2.33, − 2.27]	− 1.71 [− 1.91, − 1.18]
LW clear-sky	− 2.03 [− 2.43, − 1.75]	− 2.02 [− 2.16, − 1.80]	− 1.99 [− 2.02, − 1.97]	− 2.05 [− 2.13, − 1.79]
SW clear-sky	0.29 [0.16, 0.38]	0.30 [0.21, 0.34]	0.32 [0.30, 0.34]	0.30 [0.21, 0.61]
NET CRE	0.34 [0.04, 0.66]	0.35 [0.26, 0.59]	0.41 [0.38, 0.44]	0.12 [− 0.35, 0.86]
LW CRE	− 0.59 [− 1.12, − 0.37]	− 0.67 [− 0.94, − 0.50]	− 0.70 [− 0.73, − 0.68]	− 0.01 [− 0.34, 0.40]
SW CRE	0.94 [0.50, 1.49]	1.03 [0.76, 1.44]	1.09 [1.11, 1.14]	0.12 [− 0.28, 0.85]

Median, minimum and maximum feedback component values are given for: the Wave-1 PPE (first column); the Wave-1 PPE after filtering based on 5-year mean climate performance (second column); the 32-member stochastic ensemble (third column); and an 11-member CMIP5 atmosphere-only ensemble

Wave-1 PPE and stochastic ensemble values were derived using the SSTFuture and ATMOS experiments described in Sect. 2.2, whilst the CMIP5 values are from Ringer et al. (2014), derived using amipFuture and amip experiments. Ensemble sizes (N) are given in column headers. All values are in $\text{Wm}^{-2} \text{K}^{-1}$

substantial differences to the CMIP5 models. The impact of performance filtering is also apparent in other components: for example, weakly positive net CRE feedbacks are ruled out, as are the strongly negative net clear-sky values.

We tested if the perceived shifts of the filtered Wave-1 feedback distributions, relative to the full (406-member) Wave-1 distributions, could have been caused by a random sampling of the members. We tested the null hypothesis, that the filtered distributions are random samples of the full distributions, using a bootstrapping method. For each feedback component we created 10^4 41-member samples, drawn randomly from the full Wave-1 distribution, and then performed 2-sample KS tests for each against the full Wave-1 distribution. The resulting bootstrapped KS test statistics are shown in Fig. 3, where they are compared to values found using the true filtered Wave-1 distributions. The impact of performance filtering is clearly significant for the SW and LW CRE feedbacks. Changes in the net and net CRE feedback distributions are also distinct from the random samples, but are less clear, while the result for the net clear-sky component is more marginal.

3.3 Key performance metric clusters

Having established that performance filtering drives significant changes to the PPE's feedback distributions, we now identify groups of related performance metrics which are key to these changes. We refer to these groups as 'clusters'. We use clusters, rather than focusing on individual metrics, because the full set of performance metrics (described in Sexton et al. 2020, submitted) were chosen to sample a wide range of regional and seasonal climate performances across many variables. Consequently, many of them share common driving processes and their MSE values can be highly correlated. Additionally, the effectiveness of a performance

metric is dependent on several factors (e.g. the spread in performance across the PPE, the standard model's behaviour; see Sect. 3.4.1), meaning that the number of models ruled out by these correlated metrics can vary substantially.

Although the changes for the net clear-sky feedbacks were not as clearly distinct from the sampling noise as the other components, we still include it in this analysis for completeness. Here, we use emulators (described in Sect. 2.5) to analyse the role of the performance metrics. Compared to the actual simulations, using emulators will provide a more thorough sampling of emergent relationships and will avoid any issues with noise due to small sample sizes (since only 41 actual simulations are left after filtering).

The first stage in building the clusters was to determine the individual performance metric with the largest impact on the Wave-1 distribution, for each feedback component separately. We did this using the following methodology: firstly, for every performance metric, filtering (using the 'T' threshold, described in Sect. 2.5.2) was applied to emulated Wave-1 distributions of $\log(\text{MSE})$ values, creating a constrained sample in each case. Filtered distributions of feedbacks were then generated, using each of the constrained samples, and compared to the full Wave-1 distribution using 2-sample KS tests. The performance metric resulting in the largest KS test statistic value was then selected for each feedback component separately. We refer to this as the 'principal metric' of the cluster. The cluster was then formed by selecting performance metrics for which the emulated $\log(\text{MSE})$ values correlated strongly ($r > 0.8$) with $\log(\text{MSE})$ values of the principal metric.

The first clusters and their principal metric are given in Table 3 for each feedback component. Since each cluster is representative of the collective effect of the metrics within it, we name them using shortened descriptions summarising these. We find that the performance of extratropical LW

Fig. 3 Bootstrapped KS test statistics for the Wave-1 PPE. For each feedback component, 10^4 41-member samples are drawn randomly from the full (406-member) Wave-1 distribution, and 2-sample KS tests are performed for each to compare them against the full Wave-1 distribution. These bootstrapped distributions are shown in

grey. Red lines show the KS test statistic values for the true filtered Wave-1 distributions, labelled with the corresponding percentile of the bootstrapped distribution. This gives an indication of the robustness of the impact of the performance filtering on the feedback distributions

cloud forcing (and associated) metrics provide key constraints across the feedback components. The same cluster, ‘Extratropical_LWcloudforcing’, is selected for the LW and SW CRE, and net feedbacks, whilst an almost identical cluster is also chosen for the net clear-sky feedback (‘Extratropical_LWcloudforcing-2’; see Table 4 in Appendix 2 for details on the individual metrics in these clusters). This high level

of overlap indicates that key constraints across the feedback components are driven by common processes. In contrast, Table 3 shows a distinct cluster representing tropical LW cloud forcing variables is chosen for the net clear-sky feedback component. However, we do not analyse the role of this cluster any further due to the relatively weak changes in this feedback component shown in Sect. 3.2.

Table 3 Summary of the clusters, and their principal metrics, with leading impacts on the emulated Wave-1 feedback distributions

Feedback component	Cluster #	Cluster name	Principal metric	Summary of performance metrics in cluster	Size of cluster	% of models ruled out
NET	First	Extratropical_LWcloud-forcing	lcf_nhext_ocean_Annual (Annual mean LW cloud forcing over NH extratropical oceans)	Variables: LW fluxes High mid-thickness cloud Region: Extratropics	42	27
NET clear-sky	First	Tropical_LWcloudforcing	clhmtiscpp_tropics_land_DJF (DJF mean high mid-thickness ISCCP cloud amounts over tropical land)	Variables: LW fluxes High thin cloud High mid-thick cloud Region: Tropics	60	12
NET CRE	First	Extratropical_LWcloud-forcing-2	lcf_shext_ocean_Annual (Annual mean LW cloud forcing over SH extratropical oceans)	Variables: LW fluxes High mid-thickness cloud Region: Extratropics	46	28
LW CRE	First	Extratropical_LWcloud-forcing	lcf_nhext_ocean_Annual (Annual mean LW cloud forcing over NH extratropical oceans)	Variables: LW fluxes High mid-thickness cloud Region: Extratropics	42	27
SW CRE	First	Extratropical_LWcloud-forcing	lcf_nhext_ocean_Annual (Annual mean LW cloud forcing over NH extratropical oceans)	Variables: LW fluxes High mid-thickness cloud Region: Extratropics	42	27
NET	Second	SWclear-sky_ocean	rsutcs_shext_ocean_Annual (Annual mean SW clear-sky flux over SH extratropical oceans)	Variables: SW clear-sky fluxes Regions: Tropical, extratropical and global ocean	31	13

Columns are given for: The feedback component; the cluster number; the cluster name; the principal metric; a summary of the metrics forming the cluster; the number of performance metrics in the cluster; and the percentage of emulated parameter combinations ruled out by the cluster. For comparison, the complete filtering ruled out 62% of the emulated parameter combinations

Figure 4 shows how filtering using these clusters impacts the emulated feedback distributions. For context, the full Wave-1 ($N=64,572$) distributions are shown in black, while the ‘complete’ effect of filtering using all 925 performance metrics (leaving a sample of 24,280) is shown in red. Across the feedback components, the qualitative impact of the complete filtering is mirrored by that from the first cluster (shown in blue in Fig. 4, except for the top-right panel), but the degree to which these changes are captured varies. Specifically, the first cluster explains much of the impact from the complete filtering across the cloud feedback components, particularly for the LW and SW CRE distributions.

The net feedback changes, however, are only partially captured—as evidenced by the remaining differences between the blue and red histograms for this component in Fig. 4. This is likely to be because the net feedback is a combination of more distinct processes than any of the cloud components.

To analyse these net feedback changes further, we determine a second cluster of performance metrics which is quasi-independent of the first cluster. We do this by repeating the clustering methodology described above, but after the removal of the impact of filtering by the first cluster on the feedback distribution. A full set of quasi-independent clusters could be defined this way until we run out of performance metrics. However, we limit our analysis here to the top two clusters.

The second cluster for the net feedback (SWclear-sky_ocean) comprises metrics relating to SW clear-sky fluxes over the ocean, and they include all the large-scale regions (tropics, extratropics and global) and seasons (see Table 4 in Appendix 2). The impact of combining the top two clusters on the net feedbacks is shown in the top-right panel of Fig. 4. A small but notable change in the distribution can be seen, so that it more closely reflects the total filtering than using the first cluster alone. This suggests that distinct

Fig. 4 Normalised distributions of the feedback components from the emulated Wave-1 PPE. The full distributions ($N=64,572$) are shown in black, while the distributions after filtering using all 925 performance metrics (leaving a sample of $N=24,280$) is shown in red. In all but the top-right plot, distributions filtered by the first clusters are

shown in blue. The grey shading and arrows have been added to help illustrate how much of the complete performance filtering (in red) is captured by these clusters. The top-right plot shows the impact that filtering using both the first and second clusters has on the net feedbacks

processes and relationships lie behind the selection of these two clusters.

3.4 Emergent relationships and constraints

Section 3.3 highlighted extratropical LW cloud forcing performance as an important driver of feedback changes in the Wave-1 PPE. We now explore these relationships in more detail, with the role of the parameter perturbations being considered further in Sect. 3.5.

In Fig. 5 we show the emergent relationships for each feedback component against MSEs for the principal metric of the most prominent cluster found in Sect. 3.3: the annual mean LW cloud forcing over NH extratropical oceans ('lcf_nhext_ocean_Annual'; see Table 3). We use the principal metric because it highlights the salient features of the constraints imposed by the associated cluster, due to the high correlations between the cluster's metrics. This is much simpler than trying to define a combined metric based on all the metrics in the cluster. Also, we are ultimately trying to understand the constraints placed on the actual simulations. Therefore, having established a robust relationship across parameter space using the emulators, we now focus on the actual simulations.

The impact of an emergent constraint is controlled by two factors: the effectiveness of the filtering (i.e. the number of variants ruled out) and the strength of the relationship between the performance metric and the feedback components. We consider each of these in turn below for the lcf_nhext_ocean_Annual performance metric.

3.4.1 Effectiveness of filtering by the extratropical LW cloud forcing cluster

The top-left panel of Fig. 5 shows the distribution of lcf_nhext_ocean_Annual MSE values for the Wave-1 PPE, along with the threshold value above which models are ruled out. As described in Karmalkar et al. (2019), the effectiveness of a performance metric at ruling out parts of parameter space is determined by the relative size of the spread in MSE values against the size of the normalisation term in Eq. 2. In this study, this is largely dictated by the size of the structural error, which is based on the MSE of the standard variant. Figure 5 highlights why filtering with lcf_nhext_ocean_Annual is so effective: the good performance of the standard member (see the pink lines) sets an MSE threshold value that is relatively low compared to the spread across the PPE and consequently rules out a significant number of members.

The strength of this filtering is enhanced by the fact that the standard member is in the low tail of the MSE distribution, lying at the 9th percentile. As a result, many members lie close to the MSE threshold, with a relatively high proportion of them being ruled out (34% for this metric alone).

The disparity between the standard variant and the bulk of the PPE for this metric raises a question over how confident we can be that the standard model is truly indicative of the structural uncertainty in HadGEM3-GA7.05 for extratropical LW cloud forcing. We explore this question further in Sect. 4.1.

3.4.2 Feedback constraints from the extratropical LW cloud forcing cluster

The second key aspect required for an effective emergent constraint is a clear relationship between the performance metric and the feedback components. These emergent relationships are shown in the remaining panels of Fig. 5 (with green crosses for the Wave-1 PPE). Very clear relationships can be seen for the LW and SW CRE feedbacks, where model variants with poorer performance are associated with weaker negative LW and weaker positive SW CRE feedbacks. When combined with the effective filtering by lcf_nhext_ocean_Annual, these relationships drive a key component of the constraints on the LW and SW CRE feedbacks (as seen in Sect. 3.3).

The emergent relationship for the net CRE feedback is noticeably less clear—consistent with the smaller changes in the filtered distributions seen in Sect. 3.2. This weaker relationship is a result of an anti-correlation between the LW and SW CRE feedbacks, shown in the bottom panel of Fig. 6, which results from the competing effects of high cloud on these components. (We explore this in more detail in Sect. 3.5.) Despite the weaker relationship, however, the lcf_nhext_ocean_Annual metric does still preferentially rule out lower net CRE feedbacks. This is due to a slightly stronger emergent relationship for the SW compared to the LW CRE feedback, which can be seen by comparing the bottom two panels in Fig. 5.

No discernible relationship is seen for the net clear-sky feedback (middle-left panel of Fig. 5). However, this is not surprising because a distinct cluster was chosen for this component ('Tropical_LWcloudforcing'; see Table 3). Consequently, the constraint on the net feedback (the sum of the net clear-sky and net CRE components) is qualitatively similar to that for the net CRE: a weak constraint that preferentially rules out more negative feedback values (i.e. smaller climate sensitivities). As was the case for the LW and SW CRE feedbacks, these behaviours are captured in the emulated distributions shown in Fig. 4, giving us confidence in their use in our analysis.

3.4.3 Net feedback constraint from the tropical LW cloud forcing cluster

Figure 7 shows equivalent plots to Fig. 5, but for the principal metric of the net feedback's second cluster: the annual

Fig. 5 Top left panel shows Wave-1 MSE values for the lcf_nhext_ocean_Annual principal metric, along with the GA7.05 standard model MSE (pink) and the threshold value above which all models are rejected for this metric (black). The remaining panels show emergent relationships for Wave-1 feedback components against lcf_nhext_ocean_Annual MSEs. Actual PPE data is shown as green crosses. We use univariate emulators for each of these variables to sample the relationships for different parameter settings. We sample 7

points from the 5-95th percentiles separately for ent_fac_dp (orange), m_ci (red) and ai (blue) using the emulated Wave-1 parameter distributions. Additionally, we sample m_ci and ai jointly (purple), from the 5th-95th and 95th-5th percentiles, respectively (since high ai values act in a similar sense to low m_ci values, and vice versa). All other parameters are set to their standard values. Larger MSE values are associated with higher parameter values for ent_fac_dp and m_ci, and lower values of ai

Fig. 6 Scatter plots of feedback components from the Wave-1 PPE. Actual PPE data is shown as green crosses and emulated predictions are shown as coloured circles, where the parameter sampling is the same as in Fig. 5. Note in particular the anti-correlations between

LW and SW CRE, and the net CRE and net clear-sky feedbacks which, upon addition, result in weakened relationships and weakened impacts of a_i and m_{ci} on the net CRE and net feedbacks

SW clear-sky flux over SH extratropical oceans (`'rsutcs_shext_ocean_Annual'`). The distribution of MSE values (left-hand panel) shows that the effective filtering by this metric arises for similar reasons to those for `lcf_nhext_ocean_Annual`—namely that the relative spread in MSE performance across the PPE is large compared to that of the standard member.

However, there are also clear differences: most notably that the `rsutcs_shext_ocean_Annual` MSE distribution is positively skewed. As a result, the bulk of PPE members (including the standard) are at the lower end of the MSE

range explored by the PPE, with only a poorer performing tail being ruled out. This not only means that filtering with `rsutcs_shext_ocean_Annual` metric is less effective than with `lcf_nhext_ocean_Annual` (11% vs 34% ruled-out, respectively), but also contrasts with the fact that the standard member's better performance for `lcf_nhext_ocean_Annual` placed it in the low tail of the MSE distribution. The implications of these differences for our structural uncertainty estimates are discussed further in Sect. 4.

No obvious relationship is seen between `rsutcs_shext_ocean_Annual` MSEs and net feedbacks for the Wave-1 PPE

Fig. 7 As Fig. 5, but here MSE values are shown for the principal filter of the second cluster, *rsutcs_shext_ocean_Annual*, selected based on its impact on the net feedback. Only the emergent relationship for the net feedback is shown here (right panel)

(right-hand panel of Fig. 7). This is perhaps unsurprising, given the relatively small changes for the emulated distribution (top two panels of Fig. 5) and the sparser sampling of parameter space provided by the actual Wave-1 PPE. Therefore, this second emergent constraint is based on the relationships that emerge from our emulators and the more thorough sampling of parameter space that they allow.

3.5 The role of parameters

A key advantage of PPEs is the ability to assess how parameter perturbations affect model processes and outputs. In this section we utilise this to explore the role of our parameter perturbations on: the Extratropical_LWcloudforcing cluster; on the feedback components, and finally in driving the emergent relationships described in Sect. 3.4.

3.5.1 Extratropical LW cloud forcing cluster

To assess the connection between parameters and the performance metrics, we first tested the impact of filtering on the 47 parameter distributions, using the Extratropical_LWcloudforcing cluster. We did this by performing 2-sample KS tests for each parameter, where the full Wave-1 distributions (406 members) were compared to those filtered by the cluster (223 members). We rejected the null hypothesis of no distribution change at the 5% level for only 2 parameters: *ai*, a parameter controlling the mass-diameter relationship for cloud ice; and *m_ci*, the cloud ice fall speed (see descriptions in Table 1).

Figure 8 shows the impact of this filtering on the *ai* and *m_ci* distributions. They show that low *ai* values and high *m_ci* values are preferentially ruled out. We show later that these poorly performing models have weaker LW

cloud forcing (see members with higher MSE values in the top panel of Fig. 11), which is consistent with the physical meaning of these parameters. For example, high *m_ci* values (i.e. faster ice fall speeds) drive decreased cloud ice water content, which leads to less high cloud. Consequently, the LW cloud forcing (i.e. the difference between the clear-sky and total outgoing LW flux) is reduced. This effect is especially sensitive to high cloud, since the cooler cloud-top temperatures provide a greater contrast to the surface than cloud at lower altitudes. A similar effect results from smaller *ai* values, which indirectly drive faster ice fall speeds (see Table 1).

A parameter sensitivity analysis (Saltelli et al. 1999; and described in Sect. 2.5) for the principal metric of the Extratropical_LWcloudforcing cluster (*lcf_nhext_ocean_Annual*) is shown in the right panel of Fig. 8. This also indicates important roles for *ai* and *m_ci*: together they explain 60% of the variance in these MSE values. The consistency in these results highlights the suitability of using the *lcf_nhext_ocean_Annual* metric as a proxy to understand the behaviour of the whole cluster. We do note, however, the sensitivity to other parameters e.g. *two_d_fsd_factor* (a scaling factor for cloud condensate variability) and *dp_corr_strat* (vertical cloud overlap) which explain 21% and 17% of the variance, respectively, and may also merit further investigation.

Tsushima et al. (2020, submitted) have explored the spatial sensitivity of LW cloud forcing to parameters, using the same PPE we have analysed in this paper. The parameter sensitivities of zonal mean LW cloud forcing shown in Fig. 2a of that paper are consistent with our findings that *ai* and *m_ci* play a key role in the NH extratropics, with *two_d_fsd_factor* and *dp_corr_strat* also playing a role.

Fig. 8 Left panel: Distributions of parameter values for ai and m_ci parameters. Black histograms in the left column plots show the actual Wave-1 distributions ($N=406$), while the effect of filtering using the Extratropical_LWcloudforcing cluster is shown in blue ($N=223$). Equivalent distributions for the emulated Wave-1 Monte Carlo sample are shown in the right column plots ($N=64,573$ for the black

histograms, $N=47,241$ for the blue histograms). Right panel: Sensitivities showing the fraction of variance in lcf_nhext_ocean_Annual MSE values explained by each parameter, using the sensitivity analysis described in Sect. 2.5. Sensitivities are only shown for a subset of metrics where the total effect is >0.05 for this metric, or in any of the feedback components (as shown in Fig. 9)

3.5.2 Feedback components

We show sensitivity analyses for the main parameters driving variations in the feedback components in Fig. 9. We start by considering the cloud feedbacks. Parameters such as ai, m_ci, qlmin (the minimum critical cloud condensate) and dp_corr_strat show clear impacts on the LW and SW CRE feedbacks, but their roles are diminished for the net CRE feedback, suggesting a cancellation of the LW and SW CRE effects. For m_ci and qlmin there is an almost total cancellation, whilst for ai and dp_corr_strat a residual impact remains on the net CRE feedback. The parameter with the strongest influence on net CRE feedback variations is ent_fac_dp (the deep entrainment amplitude). Whilst this does not have a strong influence on either the LW or SW CRE feedback components, its relative impact on the net CRE feedback is increased due to the strong cancellation for other parameters. Similar signatures are also seen for the parameter dbdets_turb_0 (the cloud erosion rate).

The cancellation of global LW and SW CRE feedbacks is shown explicitly in the left-hand panel of Fig. 10. There is a clear anti-correlation between the components across the Wave-1 PPE (green crosses) resulting in a relatively reduced

spread across the diagonal lines that represent the net CRE feedbacks. This reduced spread can also be seen in Fig. 1.

We highlight the role of the ai and m_ci parameters here using simple one-at-a-time (OAT) sensitivity tests. To do this we separately sample the emulated Wave-1 distributions of these two parameters at 7 equally spaced percentiles, from the 5th and the 95th percentile (inclusive). All other parameters are held to their standard (unperturbed) values. We then use emulators to predict feedback component values at these sampled points. We acknowledge that focusing only on ai and m_ci will not capture the full complexity introduced by other parameters noted above (e.g. qlmin, dp_corr_strat). However, these two parameters have the strongest influence on cloud feedbacks, and they capture the different cancellation behaviours highlighted above (as well as being key drivers of the Extratropical_LWcloudforcing cluster, shown in Sect. 3.5.1). As such they should provide a good illustration of the role of the driving parameters.

The results show a near perfect anti-correlation along lines of constant net CRE feedback for variations in m_ci ($r = -0.999$, slope = -0.882 ; red points), whilst for variations in ai (blue points) there is a small cross-diagonal component driving spread in the net CRE feedback. This

Fig. 9 Sensitivity plots showing the fraction of variance in feedback components explained by each parameter, using the analysis described in Sect. 2.5. Sensitivities are only shown for a subset of

metrics where the total effect is > 0.05 in any of the feedback components, or for the lcf_nhext_ocean_Annual MSE (as shown in Fig. 8)

is consistent with the cancelling parameter sensitivities discussed above.

We find that this cancellation is particularly strong in the tropics (middle panel of Fig. 10), where the clouds of points for the Wave-1 PPE and OAT tests are aligned with lines of constant net CRE feedback. Such a strong cancellation

is indicative of high cloud feedbacks, which are known to have compensating LW and SW CRE changes in the tropics (Kiehl 1994). The key roles for m_{ci} and ai are also consistent with high cloud changes: as described in Sect. 3.5.1, these parameters drive changes in high cloud amounts and so will likely be important for responses in high cloud amounts.

Fig. 10 LW CRE vs SW CRE feedbacks calculated using annual means globally (left-hand panel), for the tropics (middle panel) and for the extratropics (right-hand panel). Actual PPE data is shown as green crosses and emulated predictions are shown as coloured circles, where the parameter sampling is as described in the text. Diagonal

lines are constant values of net CRE feedback, with the bold diagonal line indicated a zero net CRE feedback. (Note that the left-hand panel here is equivalent to the bottom-right panel of Fig. 6, but here we only show the emulated predictions for variations in *ai* and *m_ci*)

A weaker level of cancellation is seen in the extratropics (right-hand panel of Fig. 10), with larger variations in SW CRE feedbacks relative to the LW CRE component. A possible reason for this is that the altitude of the freezing level is lower in the extratropics than in the tropics, and consequently the ‘high’ ice clouds form at a lower altitude. Here, they have less of an impact on LW CRE fluxes since the cloud-top temperatures are closer to those of the surface, but maintain a substantial impact on SW CRE fluxes through cloud albedo effects.

The OAT tests highlight that *ai* has a stronger influence in this region than *m_ci*. It is this sensitivity of the extratropical SW CRE feedback to *ai* that appears to drive the residual sensitivity seen in the global net CRE feedback. The reasons behind these regional variations in parameter sensitivities are likely to be complex, involving the details of how these parameters influence cloud height, albedo and amount, for both low and high cloud. A detailed examination of these processes is beyond the scope of this paper, so we will not explore it further here.

We note, however, that the results above are again consistent with the parameter sensitivities of zonal mean CRE responses given in Tsushima et al. (2020, submitted). Figure 2b, d in that paper show that the LW CRE and SW CRE responses are particularly sensitive to *m_ci* in the tropics and *ai* in the extratropics, which is consistent with our OAT sensitivity tests shown in the middle and right-hand panel of Fig. 10. For the net CRE response, the cancellation of the sensitivity to *m_ci* in the tropics is seen clearly in Fig. 1b of Tsushima et al. (2020, submitted), as is the residual impact of *ai* in the extratropics.

The middle-left panel of Fig. 9 shows that variations in net clear-sky feedbacks are dominated by *ent_fac_dp*, with

around 60% of the variance being explained by this parameter. As discussed above, this parameter is also the most important for net CRE feedback variations. However, for the net feedback (top panel of Fig. 9), this parameter explains a smaller fraction of the total variance. This is due to an anti-correlation in its influence on the net clear-sky and net CRE feedbacks, as shown in Fig. 6. Other key parameters for both net CRE and net clear-sky feedbacks are also important for the net feedback, highlighting that it is influenced by a wide range of processes.

3.5.3 Emergent relationships and constraints

The emergent relationships and constraints shown in Fig. 5 (and discussed in Sect. 3.4.2) result from a shared dependence on the parameters influencing the *lcf_nhext_ocean_Annual* performance metric and the feedback components. It is important to note, however, that this does not necessarily imply a causal link: that is, the performance filtering constrains parameters, which may subsequently impact the feedback distributions, but these may not be driven by the same process.

Here, we analyse these shared dependences by focusing on the two key parameters highlighted above: *ai* and *m_ci*. We carried out equivalent OAT analyses to those described for Fig. 10 for all feedback components, as well as for *lcf_nhext_ocean_Annual* log(MSE) values. The latter were subsequently transformed to MSEs, to allow comparison with the actual Wave-1 PPE. We have additionally evaluated a joint sample of *ai* and *m_ci*, due to the similarities in their effects on the emergent relationships. The results of these tests are plotted as blue, red and purple points in Fig. 5.

Key roles for a_i and m_{ci} on the emergent relationship for the LW and SW CRE feedbacks are clearly seen, with high m_{ci} and low a_i values being associated with poorer performance (higher MSE values), weaker negative LW and weaker positive SW CRE feedbacks. The joint sample (purple dots) also suggest a role for a_i and m_{ci} in the net CRE and net feedback relationships, but these are clearly affected by the cancellation effects described in Sect. 3.5.2.

As noted above, these associations do not necessarily imply causal links. The emergent relationship for the LW CRE feedback component provides an example of this: it is the performance of LW cloud forcing in the extratropics which constrains a_i and m_{ci} but, as shown in Fig. 10, the influence of these parameters on the LW CRE feedbacks comes mainly from the tropics i.e. the key processes are not co-located. There may be more potential to explore a causal link for the net CRE feedbacks, since it is the influence of these parameters (particularly a_i) in the extratropics which appears to be important for this component. However, given the potential complexities of the cloud processes in this region (as noted in Sect. 3.5.2), such an analysis is beyond the scope of this paper.

Other parameters highlighted in Sect. 3.5.2 may also play a role, but we note here the case of ent_fac_dp , for which we also carried out an OAT test (orange points in Fig. 5). Whilst several feedback components are influenced by this parameter (e.g. net CRE and net clear-sky), it does not drive variations in $lcf_nhext_ocean_Annual$ MSE values and so has little impact on the constraints.

4 Discussion

The results presented in Sect. 3 have shown how the performance constraints applied in Sexton et al. (2020, submitted), combined with emergent relationships, ultimately lead to lower feedback values being ruled out from the atmosphere-only PPE. Here we discuss some of the key aspects of these results, and the lessons that can be drawn from them, namely: an assessment of the robustness of our most important emergent constraints; a discussion of the key advantages and limitations of studying emergent relationships and constraints using different experimental designs; and the implications of this work in the context of the coupled PPE used in the UKCP18 project, and more generally for future studies using climate model ensembles.

4.1 Sensitivity of the feedback constraints to structural error uncertainty

As discussed in Sect. 3.4, the effectiveness of filtering by a performance metric is partly controlled by the size of the normalisation term in Eq. (2), and the dominant term here

is the structural error uncertainty. This term is included to down-weight the impact of the poorest performing variables on the filtering. In Sect. 3.3 we identified two key clusters of performance metrics (Extratropical_LWcloud-forcing and SWclear-sky_ocean) that together explained a large fraction of the constraint on the climate feedback parameter. In Sect. 3.4 we showed the first cluster was most effective because the standard variant outperforms most PPE members. In contrast, the standard variant for the second cluster is closer to, and therefore more representative of, the typical performance of the PPE.

One way to further explore the structural deficiencies in the model is to analyse mean biases for the principal metrics. The filtering methodology we have used is based on normalised MSE metrics, which cannot be used directly to assess the sign of the errors and, consequently, the extent to which the PPE has a structural bias. However, MSEs can be expressed as the sum of the squared mean bias and the MSE of the centred pattern. Figure 11 shows mean bias versus MSE values for the principal metrics, for the Wave-1 PPE (green crosses) along with a curve (in black) where the MSE of the centred pattern is fixed to the standard model's value. The close correspondence between the PPE members and this curve suggests that the MSEs are dominated by the squared mean bias term. Consequently, in these cases, we can use these mean bias values to make statements about the signs of the errors dictating the MSEs.

Strikingly, for both principal metrics there is a clear systematic shift of the PPE away from a zero-bias. 98.8% of PPE members display negative mean bias values for $lcf_nhext_ocean_Annual$, whilst 99.8% have positive biases for $rsutcs_shext_ocean_Annual$. Similar systematic offsets are seen across the respective cluster members (not shown). Clear evidence for these structural errors is also shown in Fig. 12, where widespread systematic biases in present-day LW cloud forcing are seen across the PPE.

While the PPE clearly exposes a structural error for $rsutcs_shext_ocean_Annual$, the fact that the standard variant's MSE (and mean bias) is typical of the bulk of the PPE indicates that its error is representative of the PPE. On the other hand, the better performance of the standard member for $lcf_nhext_ocean_Annual$ is not representative of many of the PPE members.

There are two contrasting ways to view the standard variant. First, it could be that this standard variant is in a narrow region of parameter space that truly represents more accurate model variants for some performance metrics. In this case, the poorer performers in the PPE represent sampling of less accurate model variants and filtering these out would be appropriate. On the other hand, it could be that the standard variant is not representative of the bulk of the PPE due to tuning. If this is indeed the

Fig. 11 Top panel: Mean bias vs MSE for the `lcf_nhext_ocean_Annual` principal metric. Green crosses are for the Wave-1 PPE, with the GA7.05 standard model highlighted with a pink star. Emulated MSE predictions for this standard model (pink line) and a modal model (described in the text; blue line) are given, along with the threshold MSE threshold value (black line). Bottom panel: Equivalent to the top panel, but for the principal metric of the second cluster for the net feedback: `rsutcs_shext_ocean_Annual`

case, it will have resulted in an MSE threshold that may be too low, compounded further by the fact that the bulk of PPE members have ‘poorer’ performance and lie near this threshold. If this is the case for `lcf_nhext_ocean_Annual`, then the constraint applied by its cluster (Extratropical_LWcloudforcing) is arguably too strong. Perhaps then, more central parts of the PPE distributions of MSE may provide a fairer reflection of the structural uncertainty than the standard variant. These cases represent two extremes and it is unlikely that either applies entirely to our key clusters.

So, was the standard variant an appropriate choice to represent the structural uncertainty? There is no single definition of this term. In PPE studies that account for structural uncertainty (e.g. Sexton et al. 2012; Williamson et al.

2013), the term has been included to represent how model imperfections affect different variables, and to down-weight the impact of the poorly performing ones. One approach for formalising a definition has been based around the idea that there exists a best model within parameter space (e.g. Rougier 2007), and the performance of that model represents structural uncertainty. In the spirit of the ‘best model’ assumption, Sexton et al. (2020, submitted) opted for the standard member as the single variant that represented the best choice of model development experts.

This was felt to be more appropriate than some potential alternatives. For example, using the best MSE for each performance metric would not be based on a single model, but on different PPE members, and the value would have to be updated if a better member was later run and added to the PPE.

Fig. 12 Annual mean bias fields for LW cloud forcing for: the GA7.05 standard model (left panel) and the Wave-1 PPE mean (right panel). Hatching indicates where >95% of ensemble members have the same sign of bias. Both panels highlight widespread negative LW

cloud forcing biases in the PPE, whilst the better performance for the standard model in the NH extratropics, relative to the PPE mean (shown in the top panel of Fig. 11) is supported by the smaller biases in this region in the left panel

Another option would be to use the best MSE from the CMIP5 models; however, this would not have provided the desired relative measure of the structural error within the single-modelling framework of HadGEM3-GA7.05. Another alternative could be based on the ideas of McNeall et al. (2016), who expose structural errors in a climate model by showing that no single region of parameter space can minimise errors in a given quantity simultaneously across the globe. However, it is not immediately clear how this idea could be developed to provide the structural uncertainty term in Eq. (2).

A potential single-model alternative would be to use a ‘modal’ variant to define the structural error uncertainty. In this case parameters are set to the mode of their prior PDFs, as defined in the elicitation exercise described in Sexton et al. (2020, submitted). As such, the modal model represents the ‘maximum likelihood’ model according to the modelling experts’ prior beliefs. Some distributions are flat-topped, so this model is not uniquely defined, but this could be addressed for example by using the average parameter value over the flat-topped domain. Like the standard variant, the modal variant represents a potential ‘best choice’ model from experts, but one that should not be subject to tuning and therefore should be more representative of the centre of the PPE distributions.

We have not run a modal model simulation; however, we can illustrate it by predicting its performance using emulators. The blue lines in Fig. 11 show these emulated MSE predictions for `lcf_nhext_ocean_Annual` and `rsutcs_shext_ocean_Annual`, and they broadly reflect the points made above: the modal model shows a clear degradation in performance relative to the standard variant for `lcf_nhext_ocean_Annual`, whilst for `rsutcs_shext_ocean_Annual`, the two models are very similar (in fact the modal model is predicted

to be amongst the best for this metric). So, had we used the modal model for the structural error term in the normalised MSE metric, filtering with the Extratropical_LWcloudforcing cluster would have been much less effective, and the constraints on the feedbacks would have been weakened.

The choice of the standard variant was consistent with the best model assumption; however, we have shown that the strength of constraints can be sensitive to this choice. There is a possibility that any tuning of the standard variant could bias its performance to lower MSEs and that these are unrepresentative of the true structural error in the PPE. Whether this is the case with the first cluster is not clear, but this possibility is now a consideration for future PPEs, where the choice over how to account for structural error uncertainty will depend on the level of risk that the scientists designing the PPE associate with this issue.

4.2 Emergent relationships and constraints in a single model framework

In Sect. 3 we found strong emergent relationships between LW cloud forcing metrics and SW and LW CRE feedback components, with weaker relationships seen for the net CRE and net feedbacks. The exploration of these relationships presented in Sects. 3.4 and 3.5 was possible due to our experimental design: specifically in that we have control over the choice of the parameter perturbations in the PPE. For example, we used Latin hypercube sampling to effectively explore the parameter space for the actual simulations, in order to pick out the key emergent relationships and to train emulators. The emergent relationships were explored in more detail using the emulators, through a denser sampling of parameter space than was

possible with the actual simulations. PPEs also allow us to identify relationships between model outputs and the parameters being perturbed. We have exploited this, again using emulators, through sensitivity and one-at-a-time tests (Sect. 3.5) designed to pick out the key parameters involved in the emergent relationships. Our analyses in Sects. 3.4 and 3.5 exposed the highly complex nature of these relationships and constraints, involving multiple processes and cancellations, which ultimately produced a relatively weak constraint for the net feedback. Additionally, we have highlighted the fact that these emergent relationships do not necessarily imply causality: the performance filtering constrains the parameters, which subsequently impacts the feedback distributions, but these may not be driven by the same processes.

PPEs also provide an ideal framework in which to identify structural model errors, by exploring how the different variants (sampling expert elicited priors for the parameters) are distributed around observations. As we have shown above, structural errors are implied where the PPE is robustly positively or negatively biased, whilst potential model tuning can be identified in variables where the standard model differs significantly from the bulk of PPE members. This is again a strength of PPEs in general and facilitates straightforward interpretation of the impact of structural errors on the robustness of constraints.

There are, however, some key limitations of using PPEs for studying emergent relationships and constraints. Perhaps the most important is that they are based on a single model. This leads to an ambiguity as to whether emergent relationships found using PPEs are applicable to other models, or if they are particular to the specific model and parameter space being explored (e.g. Shiogama et al. 2014). Furthermore, as we have shown, emergent constraints derived using a single-model framework can be sensitive to the idiosyncrasies of that model: its structure, structural errors and the level of model tuning can all have an impact on the constraints.

This limitation can be addressed to some extent using multi-model ensembles (MMEs), which provide a framework to study emergent relationships and constraints across a diverse range of model structures. Indeed, multi-model ensembles have been used extensively to identify many emergent constraints (e.g. Klein and Hall 2015; Caldwell et al. 2018). However, recent studies (e.g. Caldwell et al. 2018; Wagman and Jackson 2018) have shown that these are not always robust across different MMEs, highlighting the complexities involved in interpreting relationships emerging across a variety of model structures, sampling a wide range of model processes. Our study has highlighted further complexities. For example, each MME member will have its own set of key driving processes, cancellations and structural errors. We have shown that,

in a single-model framework, these factors can have a substantial impact on emergent relationships and constraints. Without a PPE around each MME member we cannot expose these complexities. We would therefore caution against straightforward interpretations of emergent constraints arising from MMEs.

4.3 Lessons for projections work

This study has been carried out in the context of an effort to understand the contributing factors driving the lack of diversity in 21st Century warming seen in the coupled PPE (Yamazaki et al. 2020, submitted), which fed into projection information provided by the UKCP18 project (Murphy et al. 2018). In the Introduction we identified two potential causes of a reduction in the spread of feedbacks in the generation of this 15-member coupled PPE: one was from the 5-year mean climate performance filtering (Sexton et al. 2020, submitted; and studied in this paper); the other resulted from the filtering of the initial 25-members of the coupled PPE (Yamazaki et al. 2020, submitted). The relative extent to which these had an impact on the diversity will depend on how much the atmosphere-only component of the feedback contributes to the feedback in the coupled model. Ringer et al. (2014) showed evidence that the atmosphere-only components make an appreciable contribution in CMIP5. In contrast, Yamazaki et al. (2020, submitted) showed that most of the variance in the GMST warming anomalies for the coupled PPE was driven by differences in prescribed CO₂ concentrations, with only 10% being explained by the feedbacks diagnosed in the atmosphere-only experiments. It will be important, therefore, to directly assess how coupled effects influence the feedbacks, and further experiments will be required to evaluate these. For example, abrupt4xCO₂ and 1% experiments could be performed to explicitly test the traceability in climate feedbacks.

Our study has highlighted further lessons around producing a plausible and diverse coupled PPE based on cheaper, atmosphere-only simulations, and the experimental design required to do this. Whilst the approach for choosing promising atmosphere-only variants from Sexton et al. (2020, submitted) seemed successful in terms of plausible performance, we have shown that diversity can be limited by how structural model errors affect the filtering out of implausible variants in the presence of emergent relationships that may be specific to that model. In our case, a combination of factors has resulted in what is probably too strong a constraint on the climate feedbacks, namely: the presence of a structural error in extratropical LW cloud forcing; and our choice to use the standard member to define the structural error. The latter implies that there is potential for the performance filtering to be sensitive to choices made in model tuning. We have suggested that a modal model variant is an alternative which should be insensitive to tuning.

Many of these factors result from working in a single-model PPE framework. To mitigate these issues when providing projections to users, a sensible approach is to pool ensembles that explore modelling uncertainties in different ways e.g. combining PPEs and MMEs, as was done in UKCP18. A more thorough, though much more ambitious, approach would be to create a multi-model ensemble, with PPEs being built around each member of the MME. This would allow a more complete sampling of the parametric uncertainties across MME members and provide users of climate projections with a more honest assessment of the range of plausible future climates in current climate models. Such a multi-model PPE would also be an invaluable resource for those studying uncertainties and emergent relationships and constraints in climate models more broadly, where clarity in the impacts of the peculiarities of each multi-model member is crucial.

5 Summary

In this paper we have explored a potential contributing factor to the HadGEM3-GC3.05 PPE's lack of diversity in twenty-first century GMST warming rates—namely, that the atmosphere-only-based performance filtering led to a suppression of diversity in climate feedbacks. This might have fed through into the temperature responses in the coupled models, but more idealised coupled abrupt4xCO₂ experiments are required to better understand this. This study has provided insight into the effect that a particular design choice, the specification of the structural error uncertainty, has on the strength of the filtering. We have found that:

- The performance filtering applied by Sexton et al. (2020, submitted) drives significant changes in the distributions of climate feedbacks (with the exception of the net clear-sky component, which shows marginal differences).
- A large component of these feedback changes is driven by performance metrics associated with extratropical LW cloud forcing. These constraints are particularly strong for the SW and LW CRE feedbacks, with cancellation between these components—largely due to the effect of tropical high cloud—resulting in weaker constraints for the net CRE and net feedbacks.
- The SW and LW CRE feedback constraints are strongly controlled by parameters driving the ice content for high cloud. One of these (the ice fall speed, m_{ci}) is particularly influential in the tropics but, because of the cancellation, shows little control over the net cloud and net feedback constraints. Another—controlling the density of ice crystals (ai)—has a residual effect on SW extratropical feedbacks and appears to drive some of the net cloud and net feedback constraints, but the picture is very com-

plex, with other parameters having an increasing impact on these constraints.

- We have highlighted the issues around establishing causality in these relationships, noting that while model performances place constraints on parameters, which then have an impact on the feedbacks, these are not necessarily driven by the same processes.
- The GA7.05 standard (unperturbed) variant, which is used to represent the structural error uncertainty in the normalisation of performance metrics, is found to outperform over 90% of PPE members in its extratropical LW cloud forcing. Combined with the fact that nearly all members have too weak forcing (a structural bias), there is a risk that the resulting constraint may have been made stronger by model tuning. We suggest the use of a modal variant will reduce this risk in the generation of future PPEs.
- Using sensitivity analyses, we have shown how the highly complex nature of emergent relationships and constraints can be exposed in the single-model framework of a PPE. Such complexities cannot be fully appreciated using MMEs, and we caution against straightforward interpretations of emergent constraints arising from these ensembles.

Acknowledgements This work and its contributors (John W. Rostron and Kalli Furtado) were supported by the UK-China Research & Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. David M. H. Sexton, Carol F. McSweeney, Kuniko Yamazaki, Timothy Andrews, Mark A. Ringer and Yoko Tsushima were supported by the Met Office Hadley Centre Climate Programme funded by BEIS and Defra. We thank Cyril Morcrette for his advice on descriptions of GA7 cloud scheme parameters.

Funding Funding was provided by Newton Fund (415000024199), Department for Business, Energy and Industrial Strategy (GA01101).

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix 1: Leave-one-out tests for emulators

We carried out leave-one-out (LOO) tests to assess the reliability of the emulators we have used in this study. LOO tests are performed by removing a member from the training

Fig. 13 Results from the leave-one-out tests for emulators of the principal metric of the Extratropical_LWcloudforcing cluster (lcf_nhext_ocean_Annual; top-left panel) and feedback components (remaining panels). For the former we show the transformed log(MSE) values which were used in training the emulator. For the feedbacks we use the original units ($\text{Wm}^{-2}\text{K}^{-1}$). Fitted values are found by removing a PPE member from the training dataset, refitting the emulator and

dataset, refitting the emulator (trained on the remaining members) and comparing the predicted model output to the actual value for that member. This is done for each member in turn.

predicting the model output. These are compared to the actual values for the model output. This is done for each member in the sample in turn. Results are shown for Waves 1–3, which were used to train the emulators: for lcf_nhext_ocean_Annual these were the 518 members that completed the ATMOS experiments; for the feedback components these were the 509 members that completed both the ATMOS and SSTFuture experiments.

The results of these LOO test for the feedback components and the principal metric of the Extratropical_LWcloudforcing cluster (lcf_nhext_ocean_Annual) are shown in Fig. 13. They

show that the emulators generally provide good fits to the actual data, with R^2 values between 0.69 and 0.88.

We find that between 20–30 members lie outside of the emulators' 95% confidence intervals, which is consistent with expectations for the sample sizes analysed.

For some variables the emulated predictions become less reliable in the tails of the distributions. For example, we see fitted values are over-predicted for the most negative values of the net clear-sky feedbacks, and this is also seen in the net feedback. This is to be expected, as

the predictions are more reliant on extrapolation in these regions of parameter space. This bias is also seen for low $\log(\text{MSE})$ values for `lcf_nhext_ocean_Annual`; however we note that since these are the better performing members, this should not affect the filtering.

Appendix 2: Performance metric clusters

See Table 4.

Table 4 Individual performance metrics that make up each of the clusters given in Table 3

Cluster name	Performance metrics in cluster	
Extratropical_LWcloudforcing	<code>rlut_nhext_ocean_Annual</code> <code>rlut_nhext_ocean_MAM</code> <code>rlut_nhext_ocean_SON*</code> <code>rlut_nhext_land_Annual</code> <code>rlut_shext_ocean_Annual</code> <code>rlut_shext_ocean_DJF</code> <code>rlut_shext_ocean_MAM</code> <code>lcf_glober_whole_Annual</code> <code>lcf_glober_whole_DJF*</code> <code>lcf_glober_whole_MAM</code> <code>lcf_glober_whole_JJA*</code> <code>lcf_glober_whole_SON</code> <code>lcf_nhext_ocean_Annual</code> <code>lcf_nhext_ocean_DJF</code> <code>lcf_nhext_ocean_MAM</code> <code>lcf_nhext_ocean_JJA</code> <code>lcf_nhext_ocean_SON</code> <code>lcf_nhext_land_Annual</code> <code>lcf_nhext_land_DJF</code> <code>lcf_nhext_land_MAM</code> <code>lcf_nhext_land_JJA</code>	<code>lcf_nhext_land_SON</code> <code>lcf_shext_ocean_Annual</code> <code>lcf_shext_ocean_DJF</code> <code>lcf_shext_ocean_MAM</code> <code>lcf_shext_ocean_JJA</code> <code>lcf_shext_ocean_SON</code> <code>rlds_glober_whole_Annual</code> <code>rlds_glober_whole_DJF</code> <code>rlds_glober_whole_JJA</code> <code>clhmtisccp_globerR_whole_MAM*</code> <code>clhmtisccp_nhext_ocean_Annual</code> <code>clhmtisccp_nhext_ocean_DJF</code> <code>clhmtisccp_nhext_ocean_MAM</code> <code>clhmtisccp_nhext_ocean_JJA*</code> <code>clhmtisccp_nhext_ocean_SON</code> <code>clhmtisccp_shext_ocean_Annual</code> <code>clhmtisccp_shext_ocean_DJF</code> <code>clhmtisccp_shext_ocean_MAM</code> <code>clhmtisccp_shext_ocean_JJA</code> <code>clhmtisccp_shext_ocean_SON</code> <code>clhmtisccp_nhext_land_DJF</code>
Extratropical_LWcloudforcing-2	<code>rlut_nhext_ocean_Annual</code> <code>rlut_nhext_ocean_MAM</code> <code>rlut_nhext_land_Annual</code> <code>rlut_shext_ocean_Annual</code> <code>rlut_shext_ocean_DJF</code> <code>rlut_shext_ocean_MAM</code> <code>scf_glober_whole_Annual†</code> <code>scf_glober_whole_DJF†</code> <code>scf_glober_whole_SON†</code> <code>scf_shext_ocean_Annual†</code> <code>scf_shext_ocean_DJF†</code> <code>scf_shext_ocean_SON†</code> <code>lcf_glober_whole_Annual</code> <code>lcf_glober_whole_MAM</code> <code>lcf_glober_whole_SON</code> <code>lcf_nhext_ocean_Annual</code> <code>lcf_nhext_ocean_DJF</code> <code>lcf_nhext_ocean_MAM</code> <code>lcf_nhext_ocean_JJA</code> <code>lcf_nhext_ocean_SON</code> <code>lcf_nhext_land_Annual</code> <code>lcf_nhext_land_DJF</code> <code>lcf_nhext_land_MAM</code>	<code>lcf_nhext_land_JJA</code> <code>lcf_nhext_land_SON</code> <code>lcf_shext_ocean_Annual</code> <code>lcf_shext_ocean_DJF</code> <code>lcf_shext_ocean_MAM</code> <code>lcf_shext_ocean_JJA</code> <code>lcf_shext_ocean_SON</code> <code>rlds_glober_whole_Annual</code> <code>rlds_glober_whole_DJF</code> <code>rlds_glober_whole_MAM†</code> <code>rlds_glober_whole_JJA</code> <code>rlds_glober_whole_SON†</code> <code>rlds_nhext_ocean_MAM†</code> <code>clhmtisccp_nhext_ocean_Annual</code> <code>clhmtisccp_nhext_ocean_DJF</code> <code>clhmtisccp_nhext_ocean_MAM</code> <code>clhmtisccp_nhext_ocean_SON</code> <code>clhmtisccp_shext_ocean_Annual</code> <code>clhmtisccp_shext_ocean_DJF</code> <code>clhmtisccp_shext_ocean_MAM</code> <code>clhmtisccp_shext_ocean_JJA</code> <code>clhmtisccp_shext_ocean_SON</code> <code>clhmtisccp_nhext_land_DJF</code>

Table 4 (continued)

Cluster name	Performance metrics in cluster	
Tropical_LWcloudforcing	rlut_glober_whole_DJF rlut_glober_whole_MAM rlut_glober_whole_JJA rlut_glober_whole_SON rlut_tropics_ocean_Annual rlut_tropics_ocean_DJF rlut_tropics_ocean_MAM rlut_tropics_ocean_JJA rlut_tropics_ocean_SON rlut_tropics_land_Annual rlut_tropics_land_DJF rlut_tropics_land_MAM rlut_tropics_land_JJA rlut_tropics_land_SON rlut_nhext_ocean_JJA lcf_glober_whole_Annual lcf_glober_whole_DJF lcf_glober_whole_MAM lcf_glober_whole_JJA lcf_glober_whole_SON lcf_tropics_ocean_Annual lcf_tropics_ocean_DJF lcf_tropics_ocean_MAM lcf_tropics_ocean_JJA lcf_tropics_ocean_SON lcf_tropics_land_Annual lcf_tropics_land_DJF lcf_tropics_land_MAM lcf_tropics_land_JJA lcf_tropics_land_SON	clhmtisccp_globerR_whole_Annual clhmtisccp_globerR_whole_DJF clhmtisccp_globerR_whole_MAM clhmtisccp_globerR_whole_JJA clhmtisccp_globerR_whole_SON clhmtisccp_tropics_ocean_Annual clhmtisccp_tropics_ocean_DJF clhmtisccp_tropics_ocean_SON clhmtisccp_tropics_land_Annual clhmtisccp_tropics_land_DJF clhmtisccp_tropics_land_MAM clhmtisccp_tropics_land_JJA clhmtisccp_tropics_land_SON clhmtisccp_nhext_land_JJA clhtnisccp_globerR_whole_Annual clhtnisccp_globerR_whole_DJF clhtnisccp_globerR_whole_MAM clhtnisccp_globerR_whole_JJA clhtnisccp_globerR_whole_SON clhtnisccp_tropics_ocean_Annual clhtnisccp_tropics_ocean_DJF clhtnisccp_tropics_ocean_MAM clhtnisccp_tropics_ocean_JJA clhtnisccp_tropics_ocean_SON clhtnisccp_tropics_land_Annual clhtnisccp_tropics_land_DJF clhtnisccp_tropics_land_MAM clhtnisccp_tropics_land_JJA clhtnisccp_tropics_land_SON
SWclear-sky_ocean	rsutcs_glober_whole_Annual rsutcs_glober_whole_DJF rsutcs_glober_whole_MAM rsutcs_glober_whole_JJA rsutcs_glober_whole_SON rsutcs_tropics_ocean_Annual rsutcs_tropics_ocean_DJF rsutcs_tropics_ocean_MAM rsutcs_tropics_ocean_JJA rsutcs_tropics_ocean_SON rsutcs_nhext_ocean_Annual rsutcs_nhext_ocean_DJF rsutcs_nhext_ocean_MAM rsutcs_nhext_ocean_JJA rsutcs_nhext_ocean_SON rsutcs_shext_ocean_Annual	rsutcs_shext_ocean_DJF rsutcs_shext_ocean_MAM rsutcs_shext_ocean_JJA rsutcs_shext_ocean_SON rsdscs_nhext_ocean_Annual rsdscs_nhext_ocean_DJF rsdscs_nhext_ocean_MAM rsdscs_nhext_ocean_SON rsdscs_shext_ocean_Annual rsdscs_shext_ocean_MAM rsdscs_shext_ocean_JJA rsdscs_shext_ocean_SON rsuscs_nhext_ocean_SON rsuscs_shext_ocean_MAM rsuscs_shext_ocean_JJA

Note the large amount of overlap between metrics in the Extratropical_LWcloudforcing and Extratropical_LWcloudforcing-2 clusters: of the 42 and 46 metrics in these clusters (respectively), 37 are the same. Those which are unique to these clusters are marked with a ‘*’ (for Extratropical_LWcloudforcing) or a ‘†’ (for Extratropical_LWcloudforcing-2)

References

- Andrews T, Andrews MB, Bodas-Salcedo A et al (2019) Forcings, feedbacks, and climate sensitivity in HadGEM3-GC3.1 and UKESM1. *J Adv Model Earth Syst* 11:4377–4394. <https://doi.org/10.1029/2019MS001866>
- Andrews T, Gregory JM, Webb MJ (2015) The Dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *J Clim* 28:1630–1648. <https://doi.org/10.1175/JCLI-D-14-00545.1>
- Andrews T, Gregory JM, Webb MJ, Taylor KE (2012) Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophys Res Lett* 39:L09712. <https://doi.org/10.1029/2012GL051607>
- Bodas-Salcedo A, Webb MJ, Brooks ME et al (2008) Evaluating cloud systems in the Met Office global forecast model

- using simulated CloudSat radar reflectivities. *J Geophys Res* 113:D00A13. <https://doi.org/10.1029/2007JD009620>
- Bodas-Salcedo A, Mulcahy JP, Andrews T et al (2019) Strong dependence of atmospheric feedbacks on mixed-phase microphysics and aerosol–cloud interactions in HadGEM3. *J Adv Model Earth Syst*. <https://doi.org/10.1029/2019MS001688>
- Bond TC, Bergstrom RW (2006) Light absorption by carbonaceous particles: an investigative review. *Aerosol Sci Technol* 40:27–67. <https://doi.org/10.1080/02786820500421521>
- Caldwell PM, Zelinka MD, Klein SA (2018) Evaluating emergent constraints on equilibrium climate sensitivity. *J Clim* 31:3921–3942. <https://doi.org/10.1175/JCLI-D-17-0631.1>
- Cherian R, Quaas J, Salzmann M, Wild M (2014) Pollution trends over Europe constrain global aerosol forcing as simulated by climate models. *Geophys Res Lett* 41:2176–2181. <https://doi.org/10.1002/2013GL058715>
- Forster PM, Andrews T, Good P et al (2013) Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *J Geophys Res Atmos* 118:1139–1150. <https://doi.org/10.1002/jgrd.50174>
- Furtado K, Field PR, Boutle IA et al (2016) A physically based subgrid parameterization for the production and maintenance of mixed-phase clouds in a general circulation model. *J Atmos Sci* 73:279–291. <https://doi.org/10.1175/JAS-D-15-0021.1>
- Hill PG, Morcrette CJ, Boutle IA (2015) A regime-dependent parameterization of subgrid-scale cloud water content variability. *Q J R Meteorol Soc* 141:1975–1986. <https://doi.org/10.1002/qj.2506>
- Hyder P, Edwards JM, Allan RP et al (2018) Critical Southern Ocean climate model biases traced to atmospheric model cloud errors. *Nat Commun* 9:3625. <https://doi.org/10.1038/s41467-018-05634-2>
- Johnson JS, Regayre LA, Yoshioka M et al (2018) The importance of comprehensive parameter sampling and multiple observations for robust constraint of aerosol radiative forcing. *Atmos Chem Phys* 18:13031–13053. <https://doi.org/10.5194/acp-18-13031-2018>
- Kamae Y, Shiogama H, Watanabe M et al (2016) Lower-tropospheric mixing as a constraint on cloud feedback in a multiparameter multiphysics ensemble. *J Clim* 29:6259–6275. <https://doi.org/10.1175/JCLI-D-16-0042.1>
- Karmalkar AV, Sexton DMH, Murphy JM et al (2019) Finding plausible and diverse variants of a climate model. Part II: development and validation of methodology. *Clim Dyn* 53:847–877. <https://doi.org/10.1007/s00382-019-04617-3>
- Kiehl JT (1994) On the observed near cancellation between longwave and shortwave cloud forcing in tropical regions. *J Clim* 7:559–565. [https://doi.org/10.1175/1520-0442\(1994\)007%3c0559:OTONCB%3e2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007%3c0559:OTONCB%3e2.0.CO;2)
- Klein SA, Hall A (2015) Emergent constraints for cloud feedbacks. *Curr Clim Chang Rep* 1:276–287. <https://doi.org/10.1007/s40641-015-0027-1>
- Klein SA, Zhang Y, Zelinka MD et al (2013) Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator. *J Geophys Res Atmos* 118:1329–1342. <https://doi.org/10.1002/jgrd.50141>
- Klocke D, Pincus R, Quaas J (2011) On constraining estimates of climate sensitivity with present-day observations through model weighting. *J Clim* 24:6092–6099. <https://doi.org/10.1175/2011JCLI4193.1>
- Lee LA, Pringle KJ, Reddington CL et al (2013) The magnitude and causes of uncertainty in global model simulations of cloud condensation nuclei. *Atmos Chem Phys* 13:8879–8914. <https://doi.org/10.5194/acp-13-8879-2013>
- Liu Y, Daum PH, Guo H, Peng Y (2008) Dispersion bias, dispersion effect, and the aerosol–cloud conundrum. *Environ Res Lett*. <https://doi.org/10.1088/1748-9326/3/4/045021>
- Mann GW, Carslaw KS, Spracklen DV et al (2010) Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the UKCA composition–climate model. *Geosci Model Dev* 3:519–551. <https://doi.org/10.5194/gmd-3-519-2010>
- McNeill D, Williams J, Booth B et al (2016) The impact of structural error on parameter constraint in a climate model. *Earth Syst Dyn* 7:917–935. <https://doi.org/10.5194/esd-7-917-2016>
- Mulcahy JP, Jones C, Sellar A et al (2018) Improved aerosol processes and effective radiative forcing in HadGEM3 and UKESM1. *J Adv Model Earth Syst* 10:2786–2805. <https://doi.org/10.1029/2018MS001464>
- Murphy JM, Sexton DMH, Jenkins GJ, Boorman PM, Booth BBB, Brown CC, Clark RT, Collins M, Harris GR, Kendon EJ, Betts RA, Brown SJ, Howard TP, Humphrey KA, McCarthy MP, McDonald RE, Stephens A, Wallace RA (2009) UK Climate projections science report: climate change projections. Met Office Hadley Centre, Exeter
- Murphy JM, Harris GR, Sexton DMH, et al (2018) UKCP18 Land Projections: Science Report
- Regayre LA, Johnson JS, Yoshioka M et al (2018) Aerosol and physical atmosphere model parameters are both important sources of uncertainty in aerosol ERF. *Atmos Chem Phys* 18:9975–10006. <https://doi.org/10.5194/acp-18-9975-2018>
- Ringer MA, Andrews T, Webb MJ (2014) Global-mean radiative feedbacks and forcing in atmosphere-only and coupled atmosphere–ocean climate change experiments. *Geophys Res Lett* 41:4035–4042. <https://doi.org/10.1002/2014GL060347>
- Rossow WB, Schiffer RA (1999) Advances in understanding clouds from ISCCP. *Bull Am Meteorol Soc* 80:2261–2287. [https://doi.org/10.1175/1520-0477\(1999\)080%3c2261:AIUCFI%3e2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080%3c2261:AIUCFI%3e2.0.CO;2)
- Rougier J (2007) Probabilistic inference for future climate using an ensemble of climate model evaluations. *Clim Change* 81:247–264. <https://doi.org/10.1007/s10584-006-9156-9>
- Rougier J, Sexton DMH, Murphy JM, Stainforth D (2009) Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments. *J Clim* 22:3540–3557. <https://doi.org/10.1175/2008JCLI2533.1>
- Roustant O, Ginsbourger D, Deville Y (2012) DiceKriging, DiceOptim: Two R Packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *J Stat Softw* 51:1–5. <https://doi.org/10.18637/jss.v051.i01>
- Saltelli A, Tarantola S, Chan KP-S (1999) A Quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* 41:39–56. <https://doi.org/10.1080/00401706.1999.10485594>
- Sanderson BM (2011) A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J Clim* 24:1362–1377. <https://doi.org/10.1175/2010JCLI3498.1>
- Scaife AA, Copsey D, Gordon C et al (2011) Improved Atlantic winter blocking in a climate model. *Geophys Res Lett* 38:L23703. <https://doi.org/10.1029/2011GL049573>
- Scaife AA, Spanghel T, Fereday DR et al (2012) Climate change projections and stratosphere–troposphere interaction. *Clim Dyn* 38:2089–2097. <https://doi.org/10.1007/s00382-011-1080-7>
- Senior CA, Andrews T, Burton C et al (2016) Idealized climate change simulations with a high-resolution physical model: HadGEM3-GC2. *J Adv Model Earth Syst* 8:813–830. <https://doi.org/10.1002/2015MS000614>
- Sexton DMH, Murphy JM, Collins M, Webb MJ (2012) Multivariate probabilistic projections using imperfect climate models part I: outline of methodology. *Clim Dyn* 38:2513–2542. <https://doi.org/10.1007/s00382-011-1208-9>
- Sexton DMH, Karmalkar AV, Murphy JM et al (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. <https://doi.org/10.1007/s00382-019-04625-3>
- Sexton DMH, McSweeney CF, Rostron JW et al (2020) A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: Part 1: selecting the parameter combinations (**submitted**)
- Sherwood SC, Bony S, Dufresne J-L (2014) Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature* 505:37–42. <https://doi.org/10.1038/nature12829>
- Shiogama H, Watanabe M, Ogura T et al (2014) Multi-parameter multi-physics ensemble (MPMPE): a new approach exploring the uncertainties of climate sensitivity. *Atmos Sci Lett* 15:97–102. <https://doi.org/10.1002/asl2.472>
- Taylor KE, Stouffer RJ, Meehl GA (2012) An Overview of CMIP5 and the experiment design. *Bull Am Meteorol Soc* 93:485–498. <https://doi.org/10.1175/BAMS-D-11-00094.1>
- Tsushima Y, Ringer MA, Martin G et al (2020) Investigating physical constraints on climate feedbacks using a perturbed parameter ensemble. *Clim Dyn*. <https://doi.org/10.1007/s00382-020-05318-y>
- Wagman BM, Jackson CS (2018) A test of emergent constraints on cloud feedback and climate sensitivity using a calibrated single-model ensemble. *J Clim* 31:7515–7532. <https://doi.org/10.1175/JCLI-D-17-0682.1>
- Walters D, Baran AJ, Boutle I et al (2019) The Met Office unified model global atmosphere 7.0/7.1 and JULES global land 7.0 configurations. *Geosci Model Dev* 12:1909–1963. <https://doi.org/10.5194/gmd-12-1909-2019>
- Williams KD, Bodas-Salcedo A, Déqué M et al (2013) The transpose-AMIP II experiment and its application to the understanding of southern ocean cloud biases in climate models. *J Clim* 26:3258–3274. <https://doi.org/10.1175/JCLI-D-12-00429.1>
- Williams KD, Copsey D, Blockley EW et al (2018) The Met Office Global Coupled model 3.0 and 3.1 (GC3.0 and GC3.1) configurations. *J Adv Model Earth Syst* 10:357–380. <https://doi.org/10.1002/2017MS001115>
- Williamson D, Goldstein M, Allison L et al (2013) History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Clim Dyn* 41:1703–1729. <https://doi.org/10.1007/s00382-013-1896-4>
- Yamazaki K, Sexton DMH, Rostron JW et al (2020) A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: Part 2: global performance and future changes (**submitted**)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.