Does model calibration reduce uncertainty in climate projections?

Article

Accepted Version

Tett, S. F. B., Gregory, J. M. ORCID: https://orcid.org/0000-0003-1296-8644, Freychet, N., Cartis, C., Mineter, M. J. and Roberts, L. (2022) Does model calibration reduce uncertainty in climate projections? Journal of Climate, 35 (8). pp. 2585-2602. ISSN 1520-0442 doi: 10.1175/JCLI-D-21-0434.1
Available at https://centaur.reading.ac.uk/102092/

To link to this article DOI: http://dx.doi.org/10.1175/JCLI-D-21-0434.1

Publisher: American Meteorological Society

Does Model Calibration Reduce Uncertainty in Climate Projections?

Simon F. B. Tett[a] , Jonathan M. Gregory[b,c] , Nicolas Freychet[a] , Coralia Cartis[d] , Michael J. Mineter[a] , Lindon Roberts[e]

[a] *School of Geosciences, University of Edinburgh, Edinburgh, UK*

[b] *National Centre for Atmospheric Science, University of Reading, Reading, UK*

[c] *Met Office Hadley Centre, Exeter, UK*

[d] *Mathematical Institute, University of Oxford, Oxford, UK*

[e] *Mathematical Sciences Institute, Australian National University, Canberra, Australia*

Corresponding author: Simon Tett, simon.tett@ed.ac.uk

ABSTRACT: Uncertainty in climate projections is large as shown by the likely uncertainty ranges in Equilibrium Climate Sensitivity (ECS) of 2.5-4K and in the Transient Climate Response (TCR) of 1.4-2.2K. Uncertainty in model projections could arise from the way in which unresolved processes are represented, the parameter values used, or the targets for model calibration. We show that, in two climate model ensembles which were objectively calibrated to minimise differences from observed large scale atmospheric climatology, uncertainties in ECS and TCR are about two to six times smaller than in the CMIP5 or CMIP6 multi-model ensemble. We also find that projected uncertainties in surface temperature, precipitation and annual extremes are relatively small. Residual uncertainty largely arises from unconstrained sea-ice feedbacks. The 20+ year old HadAM3 standard model configuration simulates observed hemispheric scale observations and pre-industrial surface temperatures about as well as the median CMIP5 and CMIP6 ensembles while the optimised configurations simulates these better than almost all the CMIP5 and CMIP6 models. Hemispheric scale observations and pre-industrial temperatures are not systematically better simulated in CMIP6 than in CMIP5 though the CMIP6 ensemble seems to better simulate patterns of large-scale observations than the CMIP5 ensemble and the optimised HadAM3 configurations. Our results suggest that most CMIP models could be improved in their simulation of large scale observations by systematic calibration. However, the uncertainty in climate projections (for a given scenario) likely largely arises from the choice of parametrisation schemes for unresolved processes ("structural uncertainty"), with different tuning targets another possible contributor.

SIGNIFICANCE STATEMENT: Climate models represent unresolved phenomenon controlled by uncertain parameters. Changes in these parameters impact how well a climate model simulates current climate and its climate projections. Multiple calibrations of a single climate model, using an objective method, to large scale atmospheric observations are done. These models produce very similar climate projections at both global and regional scales. An analysis which combines uncertainties in observations with simulated sensitivity to observations and climate response also has small uncertainty showing that, for this model, current observations constrain climate projections. Recently developed climate models have a broad range of abilities to simulate large scale climate with only some improvement in their ability to simulate this despite a decade of model development.

1. Introduction

Charney et al. (1979) estimated that the equilibrium warming for doubled atmosphere CO_2 concentration (the Equilibrium Climate Sensitivity; ECS) is between 1.5 and 4.5K. Despite many years of research, Working Group 1 of the Intergovernmental Panel on Climate Change in its Fifth Assessment Report arrived at the same numerical range, though with much greater understanding of the uncertainty (Stocker et al. 2013). Sherwood et al. (2020) (S2020) carried out a comprehensive assessment of literature on climate sensitivity, and combined evidence from processes (largely clouds), paleo-climate (largely the Last Glacial Maximum and mid-Pleistocene warm period), and observed changes in climate. They defined an effective climate sensitivity (S) which is the ECS estimated following the linear-regression method of Gregory et al. (2004). Uncertainties in observed change and paleo-climate include a considerable contribution from "pattern" uncertainty. S2020 reported a *likely* range of 2.3-4.5K for S. Building on this, the most recent IPCC assessment(IPCC 2021) reported a *likely* range of 2.5-4K for ECS with a best estimate of 3K. They also reported that some models had climate sensitivities inconsistent with this range.

Estimates of the Transient Climate Response (TCR), which is the warming at the time of doubled CO_2 in a transient simulation with CO_2 increasing by 1%/year, also have a large spread, with a *likely* (66% confidence) range of 1.4-2.2K(IPCC 2021). This uncertainty has implications for the global budget for CO_2 emissions required to limit temperature rise, because TCR is a factor in the Transient Climate Response to Emissions (Gillett et al. 2013). A review by Knutti et al. (2017)

of many studies which estimated ECS and TCR found that TCR was somewhat constrained by observations, and correlated with projected warming over the next few decades, while ECS has a stronger relationship with late 21st century warming (Grose et al. 2018). S2020 also reported that effective climate sensitivity was a better predictor of late 21st century warming, especially under high emission scenarios, than was TCR.

There has been hope that relating model properties outside observed change from the multi-model ensemble to properties of the observed climate (Hall and Qu 2006) or climate change might constrain future climate change ("emergent constraints"). Caldwell et al. (2018) reviewed several proposed emergent constraints, and found that many were closely related, and that only four of the constraints were consistent with the original explanations from the original author. Schlund et al. (2020) found that several emergent constraints that performed well in earlier multi-model ensembles did not perform well in the CMIP6 ensemble suggesting such constraints were not robust. Sanderson et al. (2021) argued these findings could arise from common structural assumptions in a multi-model ensemble.

Some groups have observed that the parameters used in model parameterisations are uncertain (Stainforth et al. 2005). These perturbed parameter ensembles (PPEs) have had a range of ECS values with some large (Stainforth et al. 2005) and some small(Sanderson 2011). Rowlands et al. (2012); Yamazaki et al. (2013), using variants of HadCM3(Gordon et al. 2000), found good agreement with observed climate change but very large uncertainties in future climate change. Others have also used perturbed parameter ensembles to explore potential future climate change with recent approaches by the UK's Met Office for the UKCP18 programme(Lowe et al. 2019) including constraints from forecast skill (Sexton et al. 2021; Yamazaki et al. 2021). In general, these approaches use filtering where the PPE is generated by modifying parameter values, often using a latin-hypercube design and then filtering out those models inconsistent with observations. This is computationally expensive if many of those models are inconsistent with observations.

An under-explored issue is the role of model calibration in which model parameters are modified to reduce the discrepancy between simulation and observations (Mauritsen et al. 2012). So, we pose the question: how much uncertainty is there in ECS and TCR when a climate model is objectively calibrated to a diverse set of large scale climatological observations? Climate models are subjectively tuned to current observations (Mauritsen et al. 2012; Hourdin et al. 2017) with almost all modelling groups (Hourdin et al. 2017) using the net top of atmosphere flux as a target though a wide diversity of additional targets are used by different groups. Tett et al. (2013a) showed that it was possible to calibrate four parameters in a climate model to top-of-atmosphere (TOA) radiative flux measurements and that uncertainty in ECS was small (Tett et al. 2013b). Tett et al. (2017)(T17 from hereon) built on this to show it was possible to calibrate the atmospheric component (HadAM3;Pope et al. (2000)) of the venerable HadCM3 climate model (Gordon et al. 2000) driven by observed Sea Surface Temperatures, sea-ice and radiative forcings targeting a broad set of large space and time scale atmospheric variables. We build on this work by generating, using two different algorithms, two calibrated ensembles of the HadAM3 model, coupling them to the HadCM3 ocean model and examining the climate response of the two ensembles. We find that uncertainties in the climate response are small both at the global and regional scales suggesting that the structural way in which models represent unresolved processes is key to uncertainty in projections.

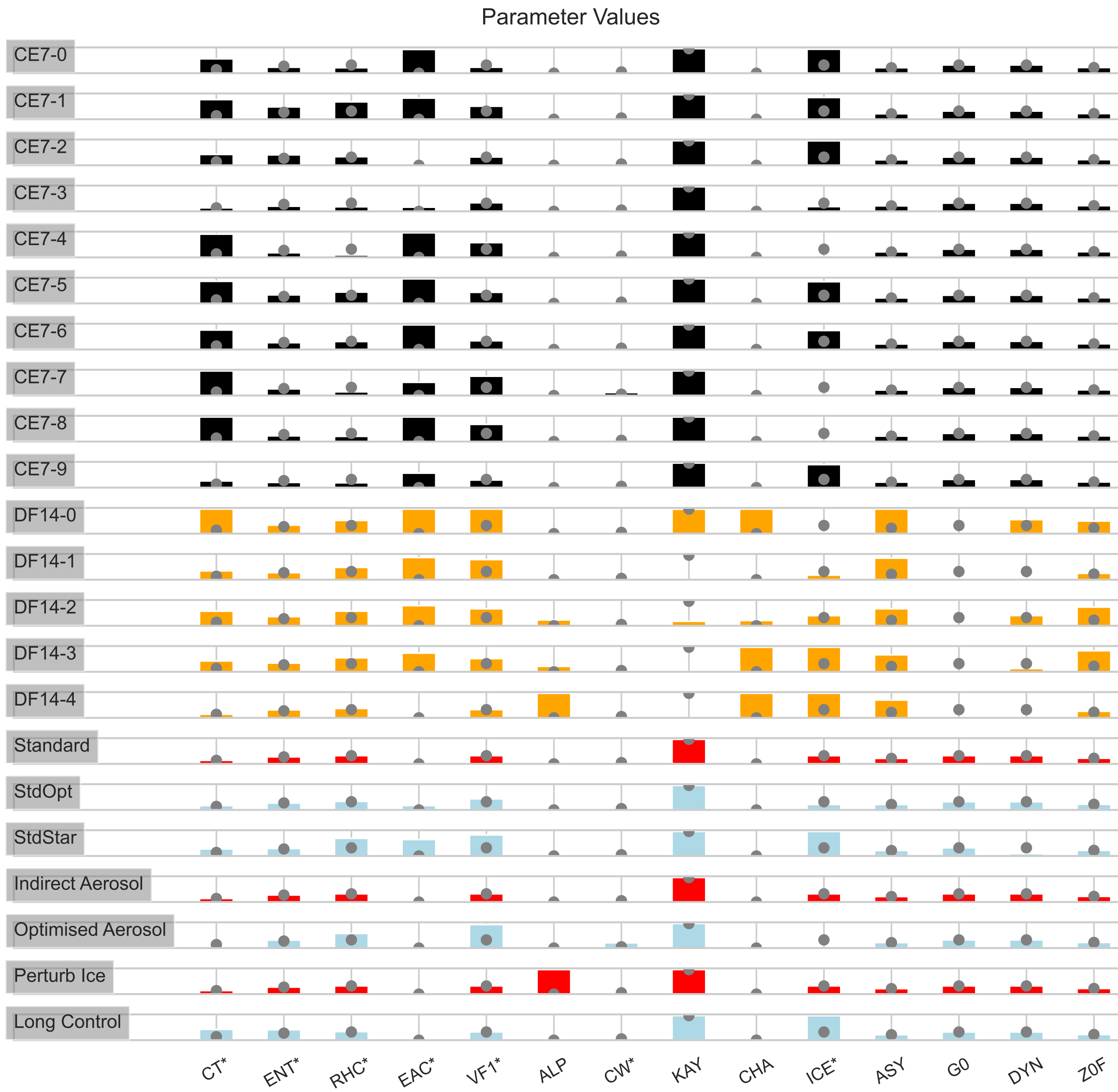
The rest of the paper is structured as follows. First we detail the methods used to generate the ensembles and our analysis methodology. We then show results from the two ensembles, followed by a set of sensitivity studies. We then report on results from a linear analysis which allows us to explore sensitivity before finally concluding.

2. Methods

a. Calibration and Experimental Design

We generated two ensembles of the HadAM3 model (Pope et al. 2000) using multiple atmospheric model simulations. The two ensembles were both calibrated to large-scale observed climate (see next paragraph for more details), each using its own algorithm. Parameter values varied across the members of both calibrated ensembles (T17 and Fig. 1) suggesting multiple, or wide and flat, minima. Several of the parameters often have values set at the expert based maxima or minima. `CW_LAND`, `KAY_GWAVE`, `CHARNOCK` & G0 in particular, show this behaviour. This suggests, for these parameters, that the expert judgement of the plausible parameter range can significantly impact the calibrated parameter values. We discuss the potential impact of this further later.

We then coupled the calibrated atmospheric-model configurations to the HadCM3 ocean model, in a state obtained from several thousand years of coupled spinup with pre-industrial forcing



115 FIG. 1. Normalised parameters for CE7 (black), DF14 (orange), calibrated sensitivity studies(blue), and
116 uncalibrated sensitivity studies (red). Parameters named with short names in Table 1. All values are normalised
117 from 0 to 1 where 0 (1) is smallest (largest) value from expert based range. Cases are named on left with
118 number as used in Fig. 2. The grey dots show standard HadAM3/HadCM3 values. Parameters are ordered from
119 left to right by their normalised impact on ECS4. Parameters with a * after their name were used in the CE7
120 optimisation.

TABLE 1. Parameter descriptions and normalised perturbations used to compute Jacobians. Short names used throughout paper are the first three characters with any _ removed. Those with a * after are used in the CE7 ensemble. See Yamazaki et al. (2013) and T17 for fuller description of parameters. The table shows parameter name, which process it impacts, the normalised perturbation, to 3 sig. figures, used to compute the atmospheric Jacobian (Atmos), the ECS4 Jacobian (ECS4) and the T140 Jacobian (T140).

Parameter	Process	Atmos	ECS4	T140	Parameter	Process	Atmos	ECS4	T140
CT*	Cloud	0.0286	0.1	0.1	DYNDIFF	Horizontal diff.	0.111	0.5	–
EACF*	Cloud	0.1	0.2	0.2	KAY_GWAVE	Gravity wave	0.4	0.5	0.5
ENTCOEF*	Convection	0.0179	0.1	0.1	ASY_LAMBDA	Boundary Layer	1/3	0.5	–
ICE_SIZE*	Radiation	0.1	0.5	0.5	CHARNOCK	Boundary Layer	0.375	0.5	0.5
RHCRIT*	Cloud	0.0333	0.2	0.2	G0	Boundary Layer	0.267	0.5	–
VF1*	Cirrus Cloud	0.0667	0.2	0.2	Z0FSEA	Boundary Layer	0.417	0.5	–
CW_LAND*	Cloud/Precip.	0.105	0.5	0.5	ALPHAM	Sea-Ice Albedo	0.4	0.5	0.5

(Gordon et al. 2000). With each coupled configuration we ran a **control** with unchanged CO_2, and other experiments with changes in CO_2 imposed (described below).

The calibration procedure (Tett et al. 2017) chooses parameter vectors for HadAM3 to minimise the weighted squared difference between simulated control and observed climatological monthly means for March 2000 to February 2005 (inclusive), following a 16-month spinup. The calibration considered geographical fields of large-scale land air temperature (LAT), land precipitation (LP), pressure differences from the global mean (SLP), TOA outgoing longwave radiation (OLR), TOA reflected shortwave radiation (RSR), 500 hPa temperature (T500) & relative humidity (q500). For each variable, except SLP, the globe was divided into three regions and area-weighted and time means computed.The three regions considered were the Northern Hemisphere extra-tropics (NHX; latitude > 30°N), Tropics (latitude between ±30°), and the Southern Hemisphere extra-tropics (SHX; latitude < 30°S) allowing representation of different large scale climate regimes. For SLP, instead of three independent quantities, the two differences (NHX average – global average) and (Tropics average – global average) were used. Global-average TOA net radiative flux (NET, N) was included as a further constraint with a target value of 0.5 W/m^2. The atmospheric model was tuned to these 21 observations by modifying parameters (Table 1) that earlier work had used (Knight et al. 2007; Yamazaki et al. 2013; Rowlands et al. 2012).

The optimisation (Tett et al. 2017) aimed to minimize the cost-function (COST):

$$F(\mathbf{p}) = \left((\mathbf{s}-\mathbf{o})^T \mathbf{C}^{-1} (\mathbf{s}-\mathbf{o}) + \frac{1}{2\mu}(N-0.5)^2 \right) / (n+1)$$

where $\mathbf{p}$ is the vector of parameter values and $\mu = 0.01$ is a penalty weight on the net radiative balance. $\mathbf{C}$ is a covariance matrix formed by summing an estimate of observational uncertainty with twice the control variability. We do this because both the simulations and the observations are assumed to contain chaotic internally generated unforced variability with the same statistical characteristics as the control. The observational uncertainty component of $\mathbf{C}$ had all off-diagonal values set to zero. Uncertainties for OLR and RSR come from the analysis of Loeb et al. (2009), while other observations used the difference between two independent estimates (see T17 for details). n is the number of observables (20 in our case – three regions x six quantities plus two SLP values); $\mathbf{o}$ is a vector of the observed targets while $\mathbf{s}$ are the simulated values. If our estimates of observational uncertainty is reliable, and if $\mathbf{C}$ is diagonal implying F is χ^2-distributed, the 5-95% confidence range for F is 0.6–1.6.

T17 calibrated eight cases using seven parameters and a Gauss-Newton algorithm (Table 1) starting the optimisation from sets of extreme parameter values. We generated another two cases using the same algorithm to give 10 parameter sets. We call this ensemble "CE7" (indicating the number of parameters). Using a new algorithm termed Derivative Free Optimization for Least Squares (Cartis et al. 2019) (DFOLS) we generated five cases using 14 parameters (Table 1). This ensemble is called "DF14". As with CE7 these started from extreme parameter values. Unlike the Gauss-Newton algorithm, DFOLS does not explicitly compute derivatives w.r.t. parameters, instead using a local-search strategy. Finally, we generated a set of sensitivity studies (SS) (Appendix A2) *some* of which were optimised using the Gauss-Newton methodology of T17. Following T17, and to avoid selection bias, the calibrated atmosphere model was run with perturbed initial conditions, and the same boundary conditions, to compute $F(\mathbf{p})$.

All **control** simulations were ran for 180 years starting from the same well spun-up state of HadCM3. T17 (Fig 7) showed that the upper ocean adjusted quickly to the parameter changes. We repeated this calculation and find that the upper-ocean largely adjusts by year 40 though with small adjustments after that (not shown). In contrast, the deep ocean is still adjusting by year 180 of the **control** in all cases (not shown).

After 40 years of **control** simulations three simulations were carried out in which 1) CO_2 increased at a rate of 1%/year until quadrupling (**1pctCO2**); 2) was instantaneously doubled (**abrupt2xCO2**); 3) and quadrupled (**abrupt4xCO2**). The **abrupt2xCO2** and **abrupt4xCO2** cases were both integrated for 40 years while the **1pctCO2** case was ran for 140 years. We focus on the differences between the forced simulations and their **control**, especially the transient responses at 2 (TCR) and $4\times CO_2$ (T140) in **1pctCO2**, the equilibrium climate sensitivity (ECS) in **abrupt2xCO2** and the equilibrium response to $4\times CO_2$ (ECS4) in **abrupt4xCO2**. All calculations are done on the difference between forced and **control** simulation in order to correct for residual drifts. In Appendix A2 we report on a sensitivity study where we ran a control for 1000 years before starting the increased CO_2 simulations. We found only a small impact.

ECS and ECS4 were estimated by regressing net Top-of-Atmosphere (TOA) flux against global-mean temperature (Gregory et al. 2004). When obtained by this method, rather than from an equilibrium $2\times CO_2$ state, the estimated ECS is commonly called "effective climate sensitivity". Similar calculations were done for other variables to estimate the equilibrium responses at 2× and 4× CO_2. Feedback parameters for the all-sky (λ) and clear-sky (λ_C), short wave (λ_{SLW}, λ_{SWC}) and longwave (λ_{LW}, λ_{LWC}) TOA radiative fluxes were computed from the slope of the appropriate linear regression fit.

TCR was diagnosed from the **1pctCO2** simulations by fitting a 2nd order polynomial to the global-average temperature difference from the equivalent **control** simulation. We used a 2nd order polynomial to capture any deviations from a linear response at longer timescales as seen in multiple climate models (Gregory et al. 2015). The value of the fit when CO_2 doubled is our estimate of TCR. We also computed T140 (the warming at $4\times CO_2$) similarly. We also used this approach for other variables shown. As many of the **control** simulations are still warming at year 180, control values are, unless stated otherwise, taken from the value at year 180 estimated from a 2nd order polynomial fit to the data.

b. CMIP5 and CMIP6 data

We used data from CMIP5 and CMIP6 multi-model archives. CMIP5 values of ECS, TCR and T140 were taken from Gregory et al. (2015) supplemented by results from the 5th IPCC assessment report (Stocker et al. 2013) and Zelinka et al. (2020). For the CMIP6 ensemble ECS, TCR and

T140 values were taken from Ringer (2019). The ECS values in these references are actually ECS4 divided by two. For CMIP5 and CMIP6 models, the cost-function was computed from the average of all available atmospheric simulations for the same model conservatively regridded to the N48 grid of HadAM3, time-averaged and, for land values, masked by the HadAM3 land/sea mask. For Taylor diagrams(Taylor 2001) we used this regridded and masked data. The **piControl** global-average near-surface air temperature was computed from the last 100 years of the simulation. All CMIP5 and CMIP6 summary values can be found in tables 2 and 3.

c. Uncertainty

Internal variability will contribute to our estimates of the climate response. To estimate the contribution of internal variability to ECS, ECS4 and climate feedback parameters we used an ensemble of seven initial condition simulations of HadCM3 in which CO_2 was doubled and quadrupled. These simulations were all started from the same state with small perturbations and are compared against the same control simulation. To compute uncertainty in the transient and control simulations a 1000-year long control simulation of HadCM3 was used. Segments of length 140 years overlapping by 35 years were taken and a second order fit made to this timeseries. Values at year 70 and year 140 were then taken from the 2nd order fit. Variances of these values were then computed and used to estimate uncertainty from internal climate variability. For TCR, T140 and other transient values the variances were doubled as these values are computed from a difference between **1pctCO2** and **control** simulations. For simplicity the same 140-year segments were used to compute uncertainties in the **control** simulation values, although this slightly overestimates their uncertainties.

To give a qualitative estimate of how uncertain the ensembles are, we report the Coefficient of Variation (CV) as a %. CV is the standard deviation divided by the mean. When this is small then signal-to-noise is large and conversely when it is large signal-to-noise is small. The CV gives a sense of how large or small the range of model behaviour may be, but we do not estimate the uncertainty in the CV because our ensembles are too small.

TABLE 2. Summary properties for CMIP5 models. ID is the label used in Fig. 2 and other plots. N_{atmos} and N_{coup} are the sizes of the atmospheric and coupled ensembles. COST is the dimensionless value of the cost-function. Shown in K are the Equilibrium Climate Sensitivity (ECS), Transient Climate Response (TCR), Transient Climate Response (T140) at 4×CO_2 and the pre-industrial control global mean surface air temperature (GMSAT). Source shows where ECS/TCR/T140 values came from and MM Mean shows the multi-model mean of the ensemble. Other values are defined in the main text.

Model	ID	COST	N_{atmos}	ECS	TCR	T140	GMSAT	N_{coup}	Source
ACCESS1-0	a	3.1	1	3.5	2.0	4.6	287.1	1	Gregory et al. (2015)
ACCESS1.3	b	4.9	2	2.8	1.6	4.0	287.3	1	Gregory et al. (2015)
BNU-ESM	c	8.2	1	4.1	2.6	–	286.1	1	Stocker et al. (2013)
CCSM4	d	5.4	6	2.9	1.8	–	286.4	3	Stocker et al. (2013)
CESM1-CAM5	e	3.6	2	–	2.3	–	286.3	1	Stocker et al. (2013)
CMCC-CM	f	6.2	3	–	–	–	286.6	1	–
CNRM-CM5	g	5.0	1	3.2	2.1	4.5	286.4	1	Gregory et al. (2015)
CSIRO-Mk3-6-0	h	9.1	10	3.0	1.8	4.5	285.9	1	Gregory et al. (2015)
CanESM2	i	4.4	4	3.6	2.4	5.2	286.8	1	Gregory et al. (2015)
FGOALS-g2	j	6.5	1	–	1.4	–	285.5	1	Stocker et al. (2013)
FGOALS-s2	k	7.4	3	4.2	–	–	286.7	1	Zelinka et al. (2020)
GFDL-CM3	l	3.6	5	3.2	1.9	4.8	287.3	1	Gregory et al. (2015)
GISS-E2-R	m	5.2	12	2.1	1.5	–	287.6	5	Stocker et al. (2013)
HadGEM2-ES	n	3.6	6	4.3	2.5	5.4	286.8	1	Gregory et al. (2015)
IPSL-CM5A-LR	o	5.7	6	3.5	2.0	5.2	285.2	1	Gregory et al. (2015)
IPSL-CM5A-MR	p	6.2	3	3.4	2.0	5.1	286.2	1	Gregory et al. (2015)
IPSL-CM5B-LR	q	7.0	1	2.6	1.5	–	286.2	1	Stocker et al. (2013)
MIROC-ESM	r	–	–	3.5	2.2	5.6	–	–	Gregory et al. (2015)
MIROC5	s	–	–	2.1	1.5	3.7	–	–	Gregory et al. (2015)
MPI-ESM-LR	t	4.8	3	3.1	2.1	5.0	286.7	1	Gregory et al. (2015)
MPI-ESM-MR	u	4.5	3	2.9	2.0	4.8	286.9	1	Gregory et al. (2015)
MRI-CGCM3	v	–	–	2.2	1.6	4.0	–	–	Gregory et al. (2015)
NorESM1-M	w	5.9	3	2.1	1.4	3.6	286.3	1	Gregory et al. (2015)
bcc-csm1-1	x	6.1	3	2.8	1.7	–	286.9	1	Stocker et al. (2013)
bcc-csm1-1-m	y	5.5	3	2.9	2.1	–	287.1	1	Stocker et al. (2013)
inmcm4	z	4.9	1	2.0	1.3	3.0	286.1	1	Gregory et al. (2015)
MM Mean	–	5.5	–	3.0	1.9	4.6	286.5	–	–

d. Linear Uncertainty Analysis

In this subsection we explain how we compute, using a linear analysis for small perturbations, the observationally constrained distributions of ECS4, TCR and T140 for HadCM3. In essence we linearly transform observational uncertainty using Jacobians which capture the sensitivity of

TABLE 3. Summary properties for CMIP6 models with details as table 2

Model	ID	COST	N_{atmos}	ECS	TCR	T140	GMSAT	N_{coup}	Source
BCC-CSM2-MR	a	4.4	3	3.1	1.7	4.1	287.9	1	Ringer (2019)
BCC-ESM1	b	6.3	3	3.3	1.8	4.4	288.1	1	Ringer (2019)
CAMS-CSM1-0	c	7.8	2	2.3	1.7	3.8	287.3	1	Ringer (2019)
CESM2	d	5.8	1	5.2	2.1	5.1	287.2	1	Ringer (2019)
CESM2-WACCM	e	5.4	3	4.7	2.0	5.1	287.1	1	Ringer (2019)
CNRM-CM6-1	f	6.7	1	4.8	2.1	5.8	286.1	1	Ringer (2019)
CNRM-CM6-1-HR	g	–	–	4.3	2.5	5.7	–	–	Ringer (2019)
CNRM-ESM2-1	h	6.9	1	4.8	1.8	5.4	286.6	1	Ringer (2019)
CanESM5	i	–	–	5.6	2.7	6.6	–	–	Ringer (2019)
E3SM-1-0	j	–	–	5.3	3.1	7.3	–	–	Ringer (2019)
EC-Earth3	k	–	–	4.2	2.3	5.9	–	–	Ringer (2019)
EC-Earth3-Veg	l	–	–	4.3	2.6	6.1	–	–	Ringer (2019)
FGOALS-f3-L	m	8.2	3	3.0	2.1	4.8	286.1	1	Ringer (2019)
GFDL-CM4	n	–	–	3.9	2.1	5.0	–	–	Ringer (2019)
GFDL-ESM4	o	–	–	2.7	1.6	3.8	–	–	Ringer (2019)
GISS-E2-1-G	p	4.7	8	2.7	1.7	–	286.9	6	Ringer (2019)
GISS-E2-1-H	q	–	–	3.1	1.9	4.4	–	–	Ringer (2019)
GISS-E2-2-G	r	–	–	2.4	1.7	3.9	–	–	Ringer (2019)
HadGEM3-GC31-LL	s	2.9	5	5.5	2.6	6.6	286.9	1	Ringer (2019)
HadGEM3-GC31-MM	t	2.8	4	–	–	–	287.5	1	–
INM-CM4-8	u	–	–	1.8	1.3	3.1	–	–	Ringer (2019)
IPSL-CM6A-LR	v	6.0	11	4.5	2.3	5.9	285.9	2	Ringer (2019)
MCM-UA-1-0	w	–	–	3.6	1.9	4.5	–	–	Ringer (2019)
MIROC-ES2L	x	–	–	2.7	1.6	3.7	–	–	Ringer (2019)
MIROC6	y	8.2	10	2.6	1.6	3.7	288.4	1	Ringer (2019)
MPI-ESM1-2-HR	z	–	–	3.0	1.7	4.2	–	–	Ringer (2019)
MRI-ESM2-0	A	4.5	3	3.2	1.6	3.8	287.0	1	Ringer (2019)
NESM3	B	–	–	4.7	2.7	6.2	–	–	Ringer (2019)
NorESM2-LM	C	–	–	2.5	1.5	3.5	–	–	Ringer (2019)
SAM0-UNICON	D	3.7	1	3.6	2.2	4.6	286.2	1	Ringer (2019)
UKESM1-0-LL	E	3.0	1	5.3	2.8	6.6	286.5	1	Ringer (2019)
MM Mean	–	5.4	–	3.8	2.0	5.0	287.0	–	–

simulated observations and climate response to give a distribution for climate response. This allows us to compare a linear analysis with the results from the non-linear multiple calibrations and explore sensitivity to our estimate of observational uncertainty.

Assuming small perturbations and that the parameters $\mathbf{p}$ have a multi-variate Gaussian distribution ($\mathbf{p} \sim N(\mathbf{p_o}, \mathbf{C_p})$) where $\mathbf{p_0}$ are the optimised parameters, the covariance matrix ($\mathbf{C_p}$) can be computed

(T17) from:

$$\mathbf{C_p} = \mathbf{PCP}^T \tag{1}$$

where $\mathbf{P}$ is a transformation matrix = $(\mathbf{J_A}^T\mathbf{C}^{-1}\mathbf{J_A})^{-1}\mathbf{J_A}^T\mathbf{C}^{-1}$ with $\mathbf{J_A}$ the Jacobian of observational derivatives w.r.t. parameters in the atmospheric simulations estimated, in our case, using a 14-member ensemble. $\mathbf{C}$ is the observational co-variance matrix defined above. A perturbation analysis for the climate responses ($\mathbf{r} = (ECS4, TCR, T140) \sim N(\mathbf{r_o}, \mathbf{C_r})$) can be done by computing the Jacobian ($\mathbf{J_r}$) using **control**, **abrupt4xCO2** and **1pctCO2** coupled simulations for each perturbed parameter.$\mathbf{C_r} = \mathbf{J_r}\mathbf{C_p}\mathbf{J_r}^T$ where $\mathbf{r_o}$ and $\mathbf{C_r}$ are the responses from the optimised parameter settings and the response covariance matrix respectively. When computing the Jacobian for TCR and T140, only those ten parameters that had a significant impact on ECS4 were perturbed. As there are only small differences between the response of the optimised model and the standard model (Appendix A2) we approximated $\mathbf{r_o}$ and $\mathbf{p_o}$ with values for the standard HadCM3 model $\mathbf{p_s}$.

To compute the parameter perturbations, the HadSM3 simulations of Rowlands et al. (2012) were used. From the changes in ECS reported there, and assuming local linearity, the parameter changes needed to give roughly a 0.5K change in ECS were computed with a maximum normalised perturbation of 0.5 allowed (Table 1).

To keep the normalised parameters within $(0,1)$ we generated parameter vectors from the multi-variate normal distribution ($\mathbf{p} \sim N(\mathbf{p_o}, \mathbf{C_p'})$). For the small fraction of $\mathbf{p}$ where all normalised parameters were in the range $(0,1)$ we computed changes in ECS4 and T140 from $\mathbf{J}(\mathbf{p}-\mathbf{p_s})$. We generated at least 1000 realisations of $\mathbf{p}$ with normalised elements between 0 and 1 by random generation and removal of all cases where this was not so. To increase the efficiency of this process $\mathbf{C_p'}$ was computed by combining a prior distribution for the normalised elements $\mathbf{p} \sim N(\mathbf{0.5}, \mathbf{I})$ with $\mathbf{C_p}$ using Bayes theorem. The covariance and best-estimate, for ECS4 and T140, was computed from the $\mathbf{p}$ samples. Uncertainties are summarized by the standard deviation of ECS and T140 from these distributions.

This linear analysis only considers uncertainty in the perturbed parameters and does not consider structural uncertainty, nor from the error arising from HadAM3 being, on our measure, significantly different from observations.

Using this linear uncertainty approach, we can modify the observational error by changing $\mathbf{C}$ and the recomputing uncertainties in ECS4, TCR and T140. We tested the impact of forcing

$\mathbf{C}$ to be largely diagonal by, for each of the seven variables, generating the sub-matrix from the outer product of the estimated standard deviations for only this variable (which assumes perfect correlation between the three (or two) observations). These sub-matrices were composed together to form the observational error covariance matrix. Twice the internal variability covariance matrix was then added to give a different, and more correlated, estimate of $\mathbf{C}$.

We also explored the impact of the expert judgment on the parameter range by increasing the parameter range to $(-0.5, 1.5)$ and increasing the prior on the parameters to $\mathbf{p} \sim N(\mathbf{0.5}, \sqrt{(2)}\mathbf{I})$. This could lead to some unphysical parameter values but for the linear analysis this is irrelevant. We also applied this increase to `ALPHAM` alone, and all parameters except `ALPHAM`.

To test the impact of individual variables, we repeated the above analysis. We considered each of the seven variables, each with two or three observations, in turn and scaled the observational standard deviation of all other observations by 100 ("other"). This should be large enough to provide no constraint on the parameters from those observations or variable. We also repeated the analysis, but only scaled the standard deviations for the observations of that variable by 100 and left other uncertainties unmodified ("leave-out").

3. Results

In this section we first compare the calibrated HadAM3 with the atmosphere models of CMIP5 and CMIP6 with regard to their simulation of the large scale climate for 2001-2005. We then examine uncertainties in global temperature change in the two calibrated and two CMIP ensembles. We finish with a linear uncertainty analysis showing that the linear analysis of HadCM3 uncertainties has similar uncertainties to the calibrated ensembles.

a. Representation of Large-scale Climate

To assess the simulations we use the same cost function (see Methods) as T17. Both of the CMIP ensembles show a very wide distribution (top two rows of Fig. 2a) compared to both HadAM3 calibrated ensembles (third and fourth rows of Fig. 2a), with only a modest improvement in CMIP6 compared to CMIP5 (Tables 2 and 3) though there is a modest shift in the distribution to better models.

The best (worst) CMIP5 model, using our cost-function, is ACCESS-1-0 (CSIRO-Mk3-6-0) with HadGEM3-GC31-MM(FGOALS-f3-L) being the best (worst) CMIP6 models. The CE7 ensemble has a mean (range) cost-function of 4.7 (4.4-5.3) which is below and narrower than the CMIP5 ensemble, with 5.4 (3.1-9.1)(Fig. 2(a)), and the CMIP6 ensemble, with 5.4 (2.8-8.2). The DF14 ensemble has a narrower range (3.1-3.7) and a mean value (3.4) comparable with the best CMIP5 and CMIP6 atmosphere-only simulations. The standard HadAM3 configuration, with a cost-function of 4.6, is better than 17 out of 21 (10 out of 16) CMIP5 (CMIP6) AMIP simulations. This suggests that on our chosen metric that the 20+ year old HadAM3 model simulation of mean climate is comparable with the current generation of climate models. The reduction in cost function seen in the DF14 ensemble further suggests that calibration can improve the ability of models to simulate observed climate with the cases from this ensemble having cost functions close to the best models in the CMIP5/6 ensemble. However, even the minimum cost function (for HadGEM3-GC31-MM) is too large to be consistent with observations (see Methods), indicating the need for further model improvement in the CMIP6 ensemble.

Considering the simulation of the individual observational indices we find that for both the CMIP5 and CMIP6 atmospheric-only ensembles (dark-blue and blue bars, respectively, in Fig. 3), the 25-75% model range encompasses zero error, except that there is too much land precipitation in the Northern Hemisphere extra-tropics in both ensembles. However, individual models are inconsistent with observations of different quantities. All HadAM3 ensembles are inconsistent with several observational quantities, particularly land air temperature and precipitation. The DF14 ensemble has, in general, smaller errors and biases than CE7, suggesting DFOLS is a better method than the Gauss-Newton variant for calibrating atmospheric models.

We compared observational estimates of preindustrial surface temperatures with the **control** and **piControl** coupled atmosphere-ocean simulations from all four ensembles. All ensembles have broad and comparable distributions of global-average surface air temperatures; the CMIP6 ensemble has a broader range than the other three ensembles. For the CMIP5 ensemble the mean value is slightly colder than the best-estimate 19th century values (Fig. 2(b)) with about half of this ensemble being inconsistent with pre-industrial temperatures. The centre of the CMIP6 distribution is slightly warmer than the 19th century values with, also, about half the models inconsistent with the 19th century estimates. Both the CE7 and DF14 ensembles are, on average, about 0.25K warmer than those observations, while the standard HadCM3 is slightly cold. The DF14 ensemble has a narrower range than CE7 and four out of five of the members have temperatures consistent with the preindustrial temperature estimates.

Figure 4 shows partial Taylor diagrams (Taylor (2001);"Taylor Wedges") for the seven variables used in our analysis. Focusing first on the CMIP5 and CMIP6 ensembles. For SLP we see little difference between both ensemble averages, though CMIP6 lacks the outliers seen in CMIP5. Land air temperature (LAT) is well simulated in the two ensembles. Conversely, the simulation of Land Precipitation (LP) is poorer than LAT with only modest improvement from CMIP5 to CMIP6. In the mid-Troposphere, the patterns of 500 hPa temperature (T500) from the ensembles are very similar to those observed. Mid-tropospheric relative humidity (q500) is not as well simulated as T500, though does show a modest improvement from CMIP5 to CMIP6. Finally, considering TOA radiation. OLR is reasonably well simulated in both ensembles (with some room for improvement) while RSR is not very well simulated and the CMIP6 ensemble shows a distinct improvement compared to the CMIP5 ensemble.

Considering the CE7 and DF14 HadAM3 ensembles (Figure 4). Except for SLP, and LP, the DF14 ensemble is at similar locations in the Taylor wedge as the standard model is. The CE7 ensemble for all variables is close to the standard model. For LAT, T500 and OLR calibrated, and uncalibrated, HadAM3 are comparable with the CMIP5 and CMIP6 ensembles. For SLP the DF14 ensemble improves on the uncalibrated model and is broadly consistent with the CMIP6 ensemble. For RSR, LP and q500 calibrated and uncalibrated HadAM3 are broadly consistent with the CMIP5 ensemble with somewhat worse performance than the CMIP6 ensemble.

Overall, we conclude that both calibrated ensembles are, despite the age of the HadCM3 model, comparable with the CMIP5 ensemble, and not greatly worse than the CMIP6 ensemble, in their ability to simulate observed large-scale mean observations. We also conclude that the DF14 ensemble is more realistic than the CE7 ensemble suggesting that the DFOLS algorithm is a better algorithm than the Gauss-Newton algorithm for calibrating climate models.

b. Climate Response

There is a broad range in the CMIP ensembles for T140 with ensemble-means of 4.6 (CMIP5) and 5K (CMIP6) (Fig. 2(c); Table 4). The two HadCM3 calibrated ensembles have almost

TABLE 4. Ensemble average values for CMIP5, CMIP6, CE7, and CE14 ensembles (to two s.f.). Uncertainties are one standard deviation for each ensemble (to 1 s.f.). Bracketed values are coefficient of variation rounded to 1%. Standard-deviations from initial condition ensembles (ICE) for ECS/ECS4 and internal variability (IV) (see methods) for TCR/T140 are also shown. Also shown are results from the linear analysis for four restricted parameter cases. Sensitivity studies are shown, for ECS4 and T140, on the right. They are a strongly correlated observational covariance matrix(_C), and the expert judgement parameter range doubled (x2) for all seven and ten significant parameters, only `ALPHAM` (ICERx2) and all parameters except `ALPHAM` (NoICERx2).

Ensemble	ECS	ECS4	TCR	T140	Sens Study	ECS4	T140
CMIP5	3.1 ± 0.7(21%)	6.2 ± 1(21%)	1.9 ± 0.4(19%)	4.6 ± 0.7(15%)	7PR_C	6.3 ± 0.1(2%)	4.7 ± 0.06(1%)
CMIP6	3.9 ± 1(28%)	7.8 ± 2(28%)	2 ± 0.3(17%)	5 ± 1(20%)	SigPR_C	6.5 ± 0.4(6%)	4.8 ± 0.1(3%)
CE7	3.1 ± 0.1(5%)	6.2 ± 0.4(6%)	2 ± 0.05(3%)	4.7 ± 0.2(3%)	7PRx2	6.1 ± 0.3(5%)	4.7 ± 0.2(3%)
DF14	3.1 ± 0.3(9%)	6.5 ± 0.5(8%)	2 ± 0.1(5%)	4.8 ± 0.2(5%)	SigPRx2	7 ± 1(17%)	5.1 ± 0.5(9%)
ICE	2.9 ± 0.1(4%)	6.3 ± 0.2(3%)	2.1 ± 0.03(2%)	4.7 ± 0.08(2%)	IceRx2	6.5 ± 1(16%)	4.8 ± 0.4(7%)
7PR	-	6.3 ± 0.2(3%)	2.1 ± 0.07(3%)	4.7 ± 0.1(3%)	NoIceRx2	7.1 ± 0.7(9%)	5.2 ± 0.3(6%)
14PR	-	7.1 ± 0.6(9%)	–	–			
NoIceR	-	6.3 ± 0.3(5%)	–	–			
SigPR	-	7.1 ± 0.6(9%)	2.4 ± 0.1(6%)	5 ± 0.3(5%)			

identical ensemble means, are between the two CMIP ensembles, and have similar uncertainties to one another(Table 4). In all ensembles, T140 is more than double TCR (compare stars and hexagons). This is a common feature across the CMIP5 and CMIP6 ensembles with several known mechanisms (Gregory et al. 2015). Uncertainties, summarised through standard deviations, are not much larger than internal variability for TCR in both HadCM3 calibrated ensembles (Table 4). Relative uncertainties in both ECS and TCR are very similar, and are also small in the HadCM3 ensembles, at about three to six times smaller than the CMIP ensembles. The equilibrium responses ((Fig. 2(d); Table 4) show a similar pattern to the transient responses with uncertainties in CE7 being smaller than in DF14. The calibrated ensembles have relative uncertainties at most half of the CMIP5 and CMIP6 ensembles (ECS for DF14 compared to ECS for the CMIP5 ensemble).

The correlation between the atmosphere-only cost function and T140 (ECS4) in the CMIP5 ensemble is -0.15 (0.04) neither of which are significant at the 10% level. For the CMIP6 ensemble the correlations are -0.46 and -0.44 for T140 and ECS4 respectively, which are just significant at the 10% level. Even so these are weak correlations suggesting that the cost function applied to multiple models does not provide a strong constraint. Results from our two calibrated ensembles suggest that, once observational constrains have been applied, only a small uncertainty due to parameter choices remains in the transient and equilibrium responses to CO_2. If this is true of other models, it suggests that the much larger uncertainties shown by CMIP in TCR, TCR140, ECS and ECS4 arise from the range of physical parameterisation schemes used (so-called "structural uncertainty"), or from the calibration targets used, rather than from poor calibration.

c. Uncertainties in Regional Climate Change

Having shown that uncertainties in large-scale temperature change and climate feedbacks are small, we consider the CV of regional temperature change at the $4\times CO_2$ in the **1pctCO2** simulations. These are similar, and small, in the CE7 and DF14 ensembles (Fig. 5) being between 5 and 10% across most of the world. Uncertainties in both ensembles are largest:

1. where the model shows least warming, likely because internal variability is, relative to the forced response, more important there.
2. in the Arctic likely due to large internal variability and Arctic amplification.
3. in the North Atlantic likely due to significant variability in the AMOC.

CV, in both ensembles, in zonal-mean ocean-only, land-only, annual minimum and maximum temperature surface air temperatures are also small being below 10% across most of the world (Fig. 6(a,b)). Exceptions to this are the two extreme temperature indices south of 30S and in Antarctica. CV's for mean and extreme precipitation (Fig. 6(c,d)) are also small and below 10% over most of the world. Near the equator CV values are relatively large for ocean precipitation though generally below 15%.

In summary, like the global-mean changes, the uncertainties in the calibrated ensembles are small in important characteristics of near-surface climate change.

d. Linear Uncertainty Analysis

To see if our results are robust, we present a linear uncertainty analysis (see Methods). This approach combines observational uncertainty estimates with the sensitivity of atmospheric simulations and of the climate response to parameter perturbations to give an observationally constrained distribution for climate response. This approach also allows us to determine which parameters are constrained by the atmospheric observations, which observations constrain the response, and test sensitivity to assumptions about observational uncertainty.

Perturbing parameters in the cloud and convective parametrisations (Fig. 7(a) and Table 1) has the largest impact on the simulated observations in the atmosphere-only simulations. The net TOA radiative flux (NET), tropical reflected shortwave radiation (RSR), and tropical land precipitation (LP) show the largest Jacobian values w.r.t. normalised parameter change suggesting these are key climatological observations. While, for example, Northern Hemisphere extra-tropical 500 hPa humidity is insensitive to parameter changes so provides little constraint. Many parameters, after calibration, have small uncertainties (Fig. 7(b)) showing that these parameters are strongly constrained by the observations we use. Exceptions are ALPHAM (ALP - the hyperparameter that controls the albedo of sea-ice) and CHARNOCK (CHA - a boundary layer parameter) which are unconstrained by our atmospheric model simulations and observations used. The Jacobian (Fig. 7(c)) for ECS4 and T140 shows that only a few parameters have large impact on simulated climate change. Of these, cloud and convection processes are the most important parameter uncertainties and are strongly constrained by our analysis.

Combining the parameter covariance (Fig. 7(b)) with the Jacobian of climate response (Fig. 7(c)) gives linear estimates of uncertainty (for ECS4 in red and T140 in blue in Fig. 7(d)). For the seven parameter case (7P) we find a mean and standard deviation of ECS4 similar to that from the CE7 ensemble. Using all fourteen parameters (14P) gives very large uncertainties in ECS4 (Table 4). Restricting the parameter set to the expert judgement range (see methods) slightly reduces the uncertainty range for the seven parameter case (7PR) but gives a larger ECS4 and a much narrower uncertainty range for the fourteen parameter case (14PR) than the unconstrained case (14P). Restricting to the thirteen parameters (NoIceR) excluding ALPHAM gives a mean and uncertainty in ECS4 very similar to the seven parameter cases. Overall this suggests that our results are sensitive to assumptions about the plausible range for parameters. Restricting to the ten parameters (SigP and SigPR) that had a $\geq \sigma$ impact on ECS4 gives very similar results to the fourteen parameter cases (14P and 14PR) suggesting that the other four have little effect. To compute the TCR/T140 Jacobian we restricted perturbations to only these ten parameters.

We found similar results to ECS4 for TCR (Table 4) and T140 (Fig. 7(d)). T140 mean and uncertainty both increase when going from seven to ten parameters, largely due to inclusion of the ice-albedo parameter in the analysis. Uncertainties for both TCR and T140 are comparable to the CE7 and DF14 calibrated ensembles (Table 4). To test sensitivity to our assumed observational structure, we examined the impact of producing a correlated co-variance matrix for observational error (see methods). This reduces the estimated uncertainty (Table 4) in ECS4 and T140 particularly for the SigPR case, suggesting our results are conservative. Considering the sensitivity case when the parameter range is doubled, then we find uncertainties in ECS4 and T140 increase by about 70 to 80%. This seems to largely be due to the ice hyperparameter (compare ICERx2 and NoICERx2 with SigPR) which is not well constrained with our atmosphere-only calibration simulations.

To examine if any subset of the observations are responsible for the small uncertainties we examine the standard deviations of T140 (σ_{T140}) when we increase uncertainties by a factor of 100 in all but one variable, or group of variables ("other"). We also examine the impact of increasing uncertainty in only one variable, or group of variables, by a factor of 100 ("leave-out"). We do this for the SigPR case (see Methods; Figure 8). For the "other" analysis if a variable constrains T140 we would expect σ_{T140} to change little from the All case while for the "leave-out" analysis we would expect σ_{T140} to change considerably from the All case.

We consider first the "leave-out" analysis where σ_{T140}, with the exception of the Radn and Sfc cases, is little impacted by increasing the uncertainty on other variables a hundred-fold. For this analysis leaving out individual variables gives only small changes in T140 standard deviation with removal of Land Precipitation (LP), RSR (Reflected Solar Radiation) and NET (Net flux) causing the largest, though modest, increases in σ_{T140}. In the "other" analysis the Sfc and Radn variable groups, on their own, give similar magnitudes of σ_{T140} to each other though larger than the All case. Using only single variables leads to quite large σ_{T140} values (Figure 8). Of the single variable constraints LP, SLP, RSR and NET appear to constrain the most while q500, T500 and OLR constrain T140 the least and are similar to the None analysis (where no observational constraints are applied). These results suggest that a smaller combination of variables, than the original seven, may constrain T140. After a some experimentation we found that LP, RSR and NET combined without any other variables (Best in Figure 8) lead to σ_{T140} comparable to σ_{T140} in the All analysis and is consistent with our earlier analysis of the Jacobian. Similar findings hold for ECS4 (not shown).This suggests these three variables are key, in our framework, to constraining climate response.

Appendix A1 explores changes in forcing from CO_2 (likely fast responses to CO_2 changes rather than changes in radiative forcing) and feedbacks. We find that all-sky shortwave (λ_{SW}) and longwave (λ_{LW}) climate feedbacks do show large changes between the two ensembles and, especially for the CE7 ensemble, within the ensemble. However, total climate feedback changes are small due to near-cancellation between changes in λ_{LW} and λ_{SW} after calibration.

4. Discussion and Conclusions

Using two different approaches, we find that the large-scale response of HadCM3 (Gordon et al. 2000) to CO_2 increase is strongly constrained when the simulated control climate is objectively calibrated against multiple large-scale 5-year mean atmospheric observations. Observations of land precipitation, reflected shortwave radiation, and net flux provide the strongest observational constraints on the model. Observational estimates of pre-industrial global-average temperature give an independent test on the ability of the HadCM3 to simulate large scale climate. Most, but not all, of the calibrated models are in agreement with this observation. Using the DFOLS algorithm (Cartis et al. 2019) to calibrate the atmospheric component of HadCM3 (Pope et al. 2000) we find it is possible to produce model configurations that are in much better agreement with large-scale observations than the standard configuration, and than almost all of the CMIP5 and CMIP6 models. For model calibration, it appears that DFOLS is better than the Gauss-Newton method used in Tett et al. (2017).

Rowlands et al. (2012) filtered perturbed physics ensemble (PPE) of flux-adjusted HadCM3 simulations to be close to observed regional trends in near-surface temperature and with a flux adjustment global-mean between $\pm 5\mathrm{Wm}^{-2}$. They found a *likely* range of 1.4-3K in near-surface changes in the 2050s driven by the SRES A1B scenario. Removing the flux-adjustment filter increased the upper limit to 3.4K. However, there are many differences between our study and theirs. Key differences are that we do not perturb the sulphur cycle and are using idealised studies to examine TCR, T140, ECS and ECS4 while they look at the response to a mixture of forcings in the mid-2050s. Our uncertainties in climate response, based on calibration to several 5-year mean observations, are considerably smaller than the approximately 40% uncertainties in the response reported by Rowlands et al. (2012). This suggests multiple large-scale observations may constrain model parameters, and thus climate response, better than observed temperature change. The key observations we have identified are land precipitation, reflected solar radiation and net flux into the Earth system.

Residual uncertainties in climate response arise from poorly constrained parameters which can have modest impact on climate feedbacks for example sea-ice. Sea-ice parameters can be successfully calibrated using decadal-scale coupled simulations(Roach et al. 2017) showing the importance of appropriate simulation design to calibrate models.

In Appendix A2 we report on a series of sensitivity studies. We find that use of a short spinup does not make a very large difference to our results. We also found that changes in the ice-albedo hyper-parameter had little impact on the cost function but a modest impact on the model response. Structural changes to the model physics through inclusion of aerosol impact on cloud properties (Jones et al. 2001) had a relatively large impact on the both HadCM3's ability to simulate current climate and its response to CO_2. However, we found this was due to changes in the diagnosed forcing from CO_2 rather than to changes in feedbacks. We speculate that this is due to changes in fast cloud feedbacks. Changes to the representation of ice-crystals in the model's radiation scheme had little impact. Thus, structural changes in HadCM3 can have a significant impact on its response but in a surprising way.

We also found a broad spread in the ability of the CMIP5 and CMIP6 multimodel ensembles to represent well the large scale 5-year mean atmospheric observations and pre-industrial temperature. Further, CMIP6 is not noticeably better than CMIP5 on the two large-scale metrics we used though does show some improvement in the simulation of patterns of 2000-2005 large-scale means. This suggests that model development, over the past decade, has not greatly improved the ability of climate models to simulate current large scale or pre-industrial climate. It is plausible that automatically calibrating many of the CMIP6 models, using state-of-the-art algorithms, would make them more in agreement with observations.

We believe, making the plausible assumption that there is nothing unusual about HadCM3, that our results will hold for other models. Thus, for any specific model, uncertainty in climate response will be small if the model parameters are calibrated against multiple observations. This may be sensitive to the cancellation of SW and LW feedbacks from cloud changes seen in HadCM3. Since we found no robust linear relationship between our calibration metric and climate response in the CMIP5 and CMIP6 ensembles, and the changes in HadCM3 response with changes to model physics, we suggest that uncertainty in the two ensembles largely arises from structural differences. If so, calibrated perturbed physics ensembles (such as the UKCP18 ensemble; Lowe et al. (2019)) have likely too small an uncertainty range for future climate change, because they do not address structural uncertainty.

However, the possibility that different groups have followed different calibration strategies can not be ruled out as a source of uncertainty in model response to CO_2 and other forcings. Moving to an objective and documented approach to model calibration rather than the current *ad hoc* approach (Hourdin et al. 2017) would help understand this. Based on our results, using objective methods to calibrate climate (or Earth System) models to large-scale observations is likely to improve their ability to simulate current large-scale mean states, and *may* narrow the range of model projections. However, it is likely that structural uncertainty arising from different choices in how to parameterise unresolved processes in also important. In summary, to reduce the recalcitrant uncertainty in model response to greenhouse gases and other forcings requires much more focus on how models represent unresolved processes than there may have been hitherto.

Acknowledgments. We thank the two anonymous referees for their comments which improved the paper. ST, CC and MM were funded by NERC (OptClim: NE/L012146/1) with simulations and post-processing done on the Edinburgh Compute and Data Facility partially funded by ClimateXchange. NF was 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. JMG was funded by the National Centre for Atmospheric Science. LR was funded by the EPSRC Centre For Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in collaboration with NAG Ltd. We thank the World Climate Research Programme's working group on coupled modelling for producing and making available their model output and Dr Mark Ringer (Met Office) for data on CMIP6 transient and equilibrium responses.

Data availability statement. The latest version of the software used for optimisation and Jacobian computations is available at `https://github.com/SimonTett/ModelOptimisation`. Software used to produce the figures in this paper is available from `https://github.com/SimonTett/Jclim21_calibrate` while processed data is available at `https://doi.org/10.7488/ds/3051`. The DFOLS software is available from `https://github.com/numericalalgorithmsgroup/dfols`. TCR, T140 & ECS values for the CMIP6 ensemble are at `https://github.com/mark-ringer/cmip6`. The full multi-Tbyte dataset of HadCM3 simulations is available at doi:10.7488/84b585fc-57d2-4e5a-b3a3-694f70534a02. To retrieve this data please contact SFBT.

APPENDIX

A1. Drivers of Climate Response Uncertainty

In this appendix we consider the drivers of the uncertainty in climate response in both ensembles. We start by considering ECS4 which depends on both CO_2 forcing (including rapid adjustment) and the climate feedback parameter (ECS4 = $F(4\times CO_2)/\lambda$) with both λ and F possibly impacted by changes in model parameters. We next consider the contributions of SW (λ_{SW}) and LW feedbacks (λ_{LW}) to uncertainty with $\lambda = \lambda_{SW} + \lambda_{LW}$ and then similarly for clear-sky feedbacks (λ_{SWC} and λ_{LWC}). To easily assess uncertainty in these joint distribution, relative to the standard model, we fix one of λ, $F(4\times CO_2)$, and λ_C to the standard values which in the plane being considered is a line. Uncertainties around this line are computed by modifying ECS4, λ and λ_C to their standard value $\pm 2\sqrt{2}\sigma$ where σ is the standard deviation from the 7 member initial condition ensemble. Model configurations within this region have values consistent with the standard model though this may arise from cancellation between processes.

Starting with ECS4 and forcing at $4\times CO_2$ (Fig. A1(a)). Most of the CE7 ensemble members sit inside the internal-variability confidence region suggesting no significant joint change in ECS4 and forcing. All but one of the remaining CE7 members sit within the grey region suggesting that much of the limited variability in ECS in this ensemble arises from cancellation between fast adjustments to CO_2 forcing and feedback strengths. For the DF14 ensemble, relative to the CE7 ensemble, the ensemble-mean has a smaller value of λ and a smaller forcing. The individual members of both ensembles lie close to the constant ECS4 line but with different forcings and climate feedback's. This suggests that internal variability in the estimation of these values produces strongly correlated values (the ellipse in Fig. A1(a) is narrow and strongly oriented along the λ-F line) and that the calibration process modifies feedbacks and the fast response to CO_2 such that ECS4 changes little. One exception to this cancellation is the DF14-4 case which has higher TCR140 and ECS4 (Fig. 2) than any of the other ensemble members. This occurs because λ is smaller than the rest of the ensemble with similar CO_2 forcing.

Internal variability does not produce strong correlations between shortwave (SW) and longwave (LW) climate feedbacks (Fig. A1(b)), but the members of both the CE7 and DF14 ensembles are aligned so that strong LW positive feedbacks are correlated with strong negative SW feedbacks. Both ensembles are significantly different from the Standard configuration. This likely arises because parameter changes modify simulated clouds and cloud feedbacks. If, in response to warming, there is a reduction in cloud cover then this will cause an increase in outgoing LW and a reduction in reflected SW. So by modifying model cloud parameters, but constraining the model to agree well with observations, we generate strong negative correlations between the SW and LW feedbacks. This is what leads to the small uncertainties in λ in CE7. DF14 shows a smaller spread in λ_{SW} and λ_{LW} suggesting that the better calibration method reduces uncertainty in these feedback parameters. Finally considering clear sky feedbacks (Fig. A1(c)), the CE7 members are largely within, or very close, to the internal variability centred on the Standard configuration suggesting no significant changes in clear sky feedbacks in this ensemble. DF14 shows a shift though no systematic change in the total clear sky LW feedback. One case (DF14-4) from this ensemble has a much more negative clear sky SW feedback than the remaining four members. The remaining ensemble members are not very different from one another with a shift to slightly larger (less amplifying) clear sky feedback parameter largely due to near cancelling changes in SW and LW clear sky feedbacks.

The DF14-4 case is an outlier in that is has a weaker climate feedback strength and so higher ECS4, if fast CO_2 feedbacks do not change, than the other ensemble members. Considering the all-sky SW and LW feedback strengths this case is not obviously different from the rest of the ensemble. However, the clear-sky SW feedback strength is much more negative that the rest of the ensemble. Several parameters from this case differ from the rest of the ensemble (Fig. 1) but one parameter that has a large difference is ALPHAM. This parameter controls the albedo of sea-ice and so changes in it might be expected to impact clear sky SW feedbacks.

Overall differences in feedbacks between the ensembles seem to arise from small changes in clear sky feedbacks and near cancellation of changes in all-sky SW and LW feedbacks arising from cloud changes. However, DF14-4 appears to be an outlier as it shows large differences, from the

TABLE A1. Sensitivity cases. All optimised cases started with default parameters and normalised parameter values for all cases. Estimated 2σ differences for ECS4 (T140) is about 0.5 (0.2) K (Table 4).

	ID	COST	ECS	ECS4	TCR	T140	GMSAT	Description
Standard	S	4.6	3.0	6.1	2.1	4.7	286.3	Standard configuration
StdOpt	SO	4.6	3.1	6.0	2.0	4.9	286.1	Optimised standard configuration
StdStar	S*	4.1	3.3	6.7	2.0	4.9	286.7	Optimised 8-parameter (7 CE7 parameters plus DYNDIFF) configuration with cloud ice properties modified
Indirect Aerosol	IA	5.9	3.5	7.4	2.2	5.5	289.1	Standard configuration with interactive indirect aerosol scheme(Jones et al. 2001) included.
Optimised Aerosol	IO	3.9	2.5	5.4	1.8	4.1	285.9	Optimised version of Indirect Aerosol.
Perturb Ice	Ic	4.8	3.5	7.3	2.2	5.5	285.2	Standard configuration with ice-albedo hyper-parameter set to maximum value.
Long Control	LC	4.9	3.1	6.7	2.0	4.7	287.9	1000-year spinup of optimised HadAM3-7#5 case.
HadAM3-7#05	–	4.9	3.2	6.8	2.0	5.0	287.5	Reference for Long Control

rest of the ensemble, in the climate feedback parameter, ECS4 and the clear-sky SW feedback parameter.

A2. Sensitivity Studies

Here we report on a series of sensitivity studies in order to understand our results. They all use the same experimental protocol described above and are shown in Figures 2 and A1. They are also summarized in Table A1.

Our protocol used a short spinup of 40 years and so we test if this impacts our results by taking a warm HadCM3 **control** case (HadAM3-7#05) and extending its **control** to 1000 years after which it warmed by a further 0.5K (Fig. 2b) (LC). This case had a T140 0.3K less ($\approx -2\sigma$) than the original case (Table A1). Impacts of 0.3K are comparable with the estimated variability in both ensembles and are not particularly large. Differences between the ECS4 and ECS values are smaller and not statistically significant, as are differences between the TCR values (Table A1). This suggests our results are not an artefact of relatively short spinup of the perturbed coupled models.

The linear analysis and Appendix A1 suggested that the sea-ice albedo hyper-parameter(ALPHAM) might explain some of the differences between the two ensembles. To test this we carried out a set of simulations (Ic) in which ALPHAM was set to its maximum value with all other parameters at their standard value. This configuration had a cost-function similar to the standard model suggesting that this parameter, as expected, has little impact on the atmospheric simulation. However, its control temperatures are much colder than any other case (Fig. 2 and Table A1) Further, ECS4 and T140 are larger than all optimised cases consistent with the linear analysis and the DF14-4 case.

To see if the standard model could be further optimised using the Gauss-Newton algorithm and the impact of this optimisation was we started a Gauss-Newton optimisation using the standard parameters as initial values (T17; Table A1). This configuration had near-identical values to the standard model (Fig. 1) and differs little from the standard model (Table A1; Figures 2 andA1). The only significant changes are that this configuration is a little colder than the standard configuration. Relative to the standard configuration this optimised configuration has an increased LW feedback and more negative SW feedback which oppose one another leading to very similar net feedback. This is also the case for the clear sky feedbacks.

To explore the role that structural uncertainty might play in our results we carried out two further calibrations of HadAM3, using the Gauss Newton algorithm of T17, in which the model physics was changed and then the calibrated atmospheric model coupled to the ocean model (Table A1). In one (StdStar;S*) we changed the properties of ice crystals in the radiation code and then optimised using the same seven parameters as CE7 plus the model diffusion hyper-parameter. In another (Optimised Interactive Aerosol;IO) we added an interactive aerosol indirect effect (Jones et al. 2001) and optimised using the same seven parameters as used in CE7. Both calibrated models had cost function values smaller than any of the CE7 ensemble members and about 15% smaller than the standard model.

S* is very similar to the standard model though with somewhat higher values of ECS4 and T140. The SW and LW all-sky feedbacks in this configuration are very different from the standard model but the changes offset one another. In combination with a smaller forcing from CO_2, than the standard configuration, this leads to a similar climate responds.

The optimised interactive aerosol configuration (IO) has T140 and ECS4 values significantly below both the standard model and both calibrated ensembles (Table A1; Fig. 2). This model has a significantly smaller ECS and forcing from $4\times CO_2$ than the standard configuration with its LW and SW feedback parameters very close to the DF14 ensemble mean values. Its total feedback parameter is similar to the standard configuration (Fig. A1(b)) but its diagnosed forcing in **abrupt4xCO2** is much smaller than the standard configuration's value (Fig. A1(a). It shows quite dramatic changes in the SW and LW feedbacks but these cancel leading to only a small change in total feedback. This model also shows changes in the clear sky feedbacks with a shift to weaker clear sky feedback. Thus, the reason for the changes in T140 and ECS4 in this configuration are due to relatively fast changes in the atmosphere in response to changes in CO_2 rather than changes in climate feedbacks.

The unoptimised model with the interactive indirect aerosol scheme produces a model that has a worse simulation of large scale climate, and much larger climate responses than the standard and optimised aerosol configurations as well as many of the CMIP5 and CMIP6 models (Fig. 2). This configuration, unlike the calibrated cases, changes the all-sky SW feedback and is also significantly different clear sky feedbacks. This, in turn, suggests it is not the impact of aerosols *per say* that changes the response but the calibration of other processes to produce a reasonable simulation that then modify the fast response to CO_2 forcing.

Overall, the effect of calibration in the sensitivity studies is to generate configurations that have climate responses that are similar to that of the standard configuration. This arises from near-cancellation between SW and LW climate feedback strength, and then between CO_2 forcing and total climate feedback strength.

References

Caldwell, P. M., M. D. Zelinka, and S. A. Klein, 2018: Evaluating emergent constraints on equilibrium climate sensitivity. *Journal of Climate*, **31 (10)**, 3921–3942, https://doi.org/10.1175/JCLI-D-17-0631.1.

Cartis, C., J. Fiala, B. Marteau, and L. Roberts, 2019: Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Trans. Math. Softw.*, **45 (3)**, 32:1–32:41, https://doi.org/10.1145/3338517, URL http://doi.acm.org/10.1145/3338517.

Charney, J. G., and Coauthors, 1979: *Carbon Dioxide and Climate: A Scientific Assessment: Report of an Ad Hoc Study Group on Carbon Dioxide and Climate, Woods Hole, Massachusetts, July 23-27, 1979, to the Climate Research Board, Assembly of Mathematical and Physical*

Sciences, National Research Council. National Academies Press, URL http://www.nap.edu/catalog/12181.html.

Gillett, N. P., V. K. Arora, D. Matthews, and M. R. Allen, 2013: Constraining the ratio of global warming to cumulative CO2 emissions using CMIP5 simulations. *Journal of Climate*, **26 (18)**, 6844–6858, https://doi.org/10.1175/JCLI-D-12-00476.1, URL https://doi.org/10.1175/JCLI-D-12-00476.1, https://doi.org/10.1175/JCLI-D-12-00476.1.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood, 2000: The simulation of SST, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Clim. Dyn.*, **16**, 147–168.

Gregory, J. M., T. Andrews, and P. Good, 2015: The inconstancy of the transient climate response parameter under increasing CO_2. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373 (2054)**, 20140417, https://doi.org/10.1098/rsta.2014.0417.

Gregory, J. M., and Coauthors, 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.*, **31**, L03205, https://doi.org/10.1029/2003gl018747.

Grose, M. R., J. Gregory, R. Colman, and T. Andrews, 2018: What climate sensitivity index is most useful for projections? *Geophysical Research Letters*, **45 (3)**, 1559–1566, https://doi.org/10.1002/2017gl075742.

Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03502, https://doi.org/10.1029/2005GL025127.

Hourdin, F., and Coauthors, 2017: The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, **98 (3)**, 589–602, https://doi.org/10.1175/BAMS-D-15-00135.1, URL https://doi.org/10.1175/BAMS-D-15-00135.1, https://doi.org/10.1175/BAMS-D-15-00135.1.

IPCC, 2021: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, (In Press).

Jones, A., D. L. Roberts, M. J. Woodage, and C. E. Johnson, 2001: Indirect sulphate aerosol forcing in a climate model with an interactive sulphur cycle. *J. Geophys. Res.*, **106**, 20 293–20 310.

Knight, C. G., and Coauthors, 2007: Association of parameter, software, and hardware variation with large-scale behavior across 57,000 climate models. *Proc. Natl. Acad. Sci. U. S. A.*, **104 (30)**, 12 259–12 264, https://doi.org/10.1073/pnas.0608144104.

Knutti, R., M. A. A. Rugenstein, and G. C. Hegerl, 2017: Beyond equilibrium climate sensitivity. *Nature Geoscience*, **10**, 727, URL http://dx.doi.org/10.1038/ngeo3017.

Loeb, N. G., B. A. Wielicki, D. R. Doelling, G. L. Smith, D. F. Keyes, S. Kato, N. Manalo-Smith, and T. Wong, 2009: Toward optimal closure of the Earth's top-of-atmosphere radiation budget. *J. Clim.*, **22**, 748–765, https://doi.org/10.1175/2008JCLI2637.1.

Lowe, J. A., and Coauthors, 2019: UKCP18 science overview report. Met Office Hadley Centre, URL https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/UKCP18-Overview-report.pdf.

Mauritsen, T., and Coauthors, 2012: Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems*, **4**, M00A01, https://doi.org/10.1029/2012MS000154.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton, 2000: The impact of new physical parametrizations in the Hadley Centre climate model – HadAM3. *Clim. Dyn.*, **16**, 123–146.

Ringer, M., 2019: https://github.com/mark-ringer/cmip6.

Roach, L. A., S. F. B. Tett, M. J. Mineter, K. Yamazaki, and C. D. Rae, 2017: Automated parameter tuning applied to sea ice in a global climate model. *Climate Dynamics*, 1–15, https://doi.org/10.1007/s00382-017-3581-5.

Rowlands, D., and Coauthors, 2012: Broad range of 2050 warming from an observationally constrained large climate model ensemble. *Nature geoscience*, **5 (4)**, 256–260.

Sanderson, B. M., 2011: A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *J. Clim.*, **34**, 1362–1377, https://doi.org/10.1175/2010JCLI3498.1.

Sanderson, B. M., A. G. Pendergrass, C. D. Koven, F. Brient, B. B. B. Booth, R. A. Fisher, and R. Knutti, 2021: The potential for structural errors in emergent constraints. *Earth System Dynamics*, **12 (3)**, 899–918, https://doi.org/10.5194/esd-12-899-2021.

Schlund, M., A. Lauer, P. Gentine, S. C. Sherwood, and V. Eyring, 2020: Emergent constraints on equilibrium climate sensitivity in CMIP5: do they hold for CMIP6? *Earth System Dynamics*, **11 (4)**, 1233–1258, https://doi.org/10.5194/esd-11-1233-2020.

Sexton, D. M. H., and Coauthors, 2021: A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: part 1: selecting the parameter combinations. *Climate Dynamics*, **56 (11-12)**, 3395–3436, https://doi.org/10.1007/s00382-021-05709-9.

Sherwood, S. C., and Coauthors, 2020: An assessment of earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, **58 (4)**, https://doi.org/10.1029/2019rg000678.

Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406.

Stocker, T., and Coauthors, 2013: *IPCC, 2013: climate change 2013: the physical science basis. Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge University Press.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, **106 (D7)**, 7183–7192.

Tett, S. F. B., M. J. Mineter, C. Cartis, D. J. Rowlands, and P. Liu, 2013a: Can top of atmosphere radiation measurements constrain climate predictions? part 1: Tuning. *J. Climate*, **26**, 9348–9366, https://doi.org/10.1175/JCLI-D-12-00595.1.

Tett, S. F. B., D. J. Rowlands, M. J. Mineter, and C. Cartis, 2013b: Can top of atmosphere radiation measurements constrain climate predictions? part 2: Climate sensitivity. *J. Climate*, **26**, 9367–9383, https://doi.org/10.1175/JCLI-D-12-00596.1.

Tett, S. F. B., K. Yamazaki, M. J. Mineter, C. Cartis, and N. Eizenberg, 2017: Calibrating climate models using inverse methods: Case studies with HadAM3, HadAM3P and HadCM3. *Geoscientific Model Development*, **10**, 3567–3589, https://doi.org/10.5194/gmd-2016-305.

Williamson, D., M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki, 2013: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, **41 (7)**, 1703–1729, https://doi.org/10.1007/s00382-013-1896-4.

Yamazaki, K., D. M. H. Sexton, J. W. Rostron, C. F. McSweeney, J. M. Murphy, and G. R. Harris, 2021: A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: part 2: global performance and future changes. *Climate Dynamics*, **56 (11-12)**, 3437–3471, https://doi.org/10.1007/s00382-020-05608-5.

Yamazaki, K., and Coauthors, 2013: Obtaining diverse behaviors in a climate model without the use of flux adjustments. *JGR-Atmospheres*, **118**, 2781–2793, https://doi.org/10.1002/jgrd.50304.

Zelinka, M. D., T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor, 2020: Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, **47 (1)**, https://doi.org/10.1029/2019gl085782.

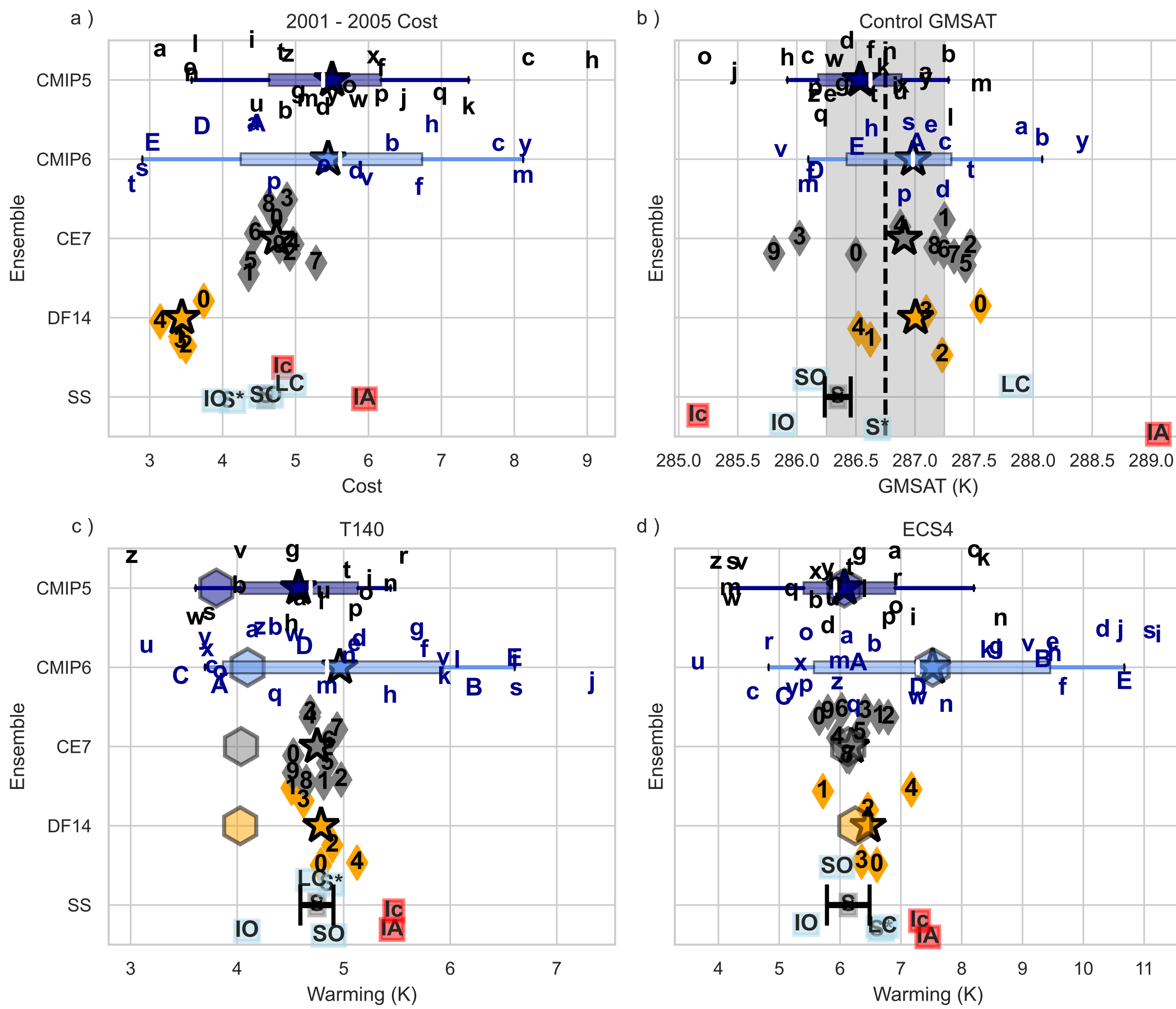


FIG. 2. Simulated values for CMIP5 (dark blue), CMIP6 (pale blue), CE7 (black) and DF14 (orange) ensembles. Also shown are sensitivity cases (SS; blue (optimised), red (unoptimised), and grey (standard configuration) boxes) described in table A1. Box and whiskers for CMIP5 and CMIP6 ensembles shows 25-75% range with whiskers extending from 5 to 95%. Stars show average value for ensemble. Y-axis in all sub-plots shows ensemble with the individual simulations having a small random offset added for presentation purposes. a: Cost values for atmosphere-only simulations. b: **Control** global average surface air temperature with vertical dashed line showing estimated observed 19th century temperature with grey shading its uncertainty range(Williamson et al. 2013). c) T140 and d) ECS4. Hexagons in c & d show ensemble average values for 2× TCR and 2× ECS. Black error bar centred on Standard HadCM3 model in b & c shows 2σ uncertainty range estimated from 1000-year long control simulation while in d shows same from 7-member initial condition ensemble. Letters for CMIP5 (black) and CMIP6 (blue) correspond to different models defined in tables 2 and 3. Numbers for CE7 and DF14 ensembles correspond to individual parameter settings (See Fig. 1 for parameter values).

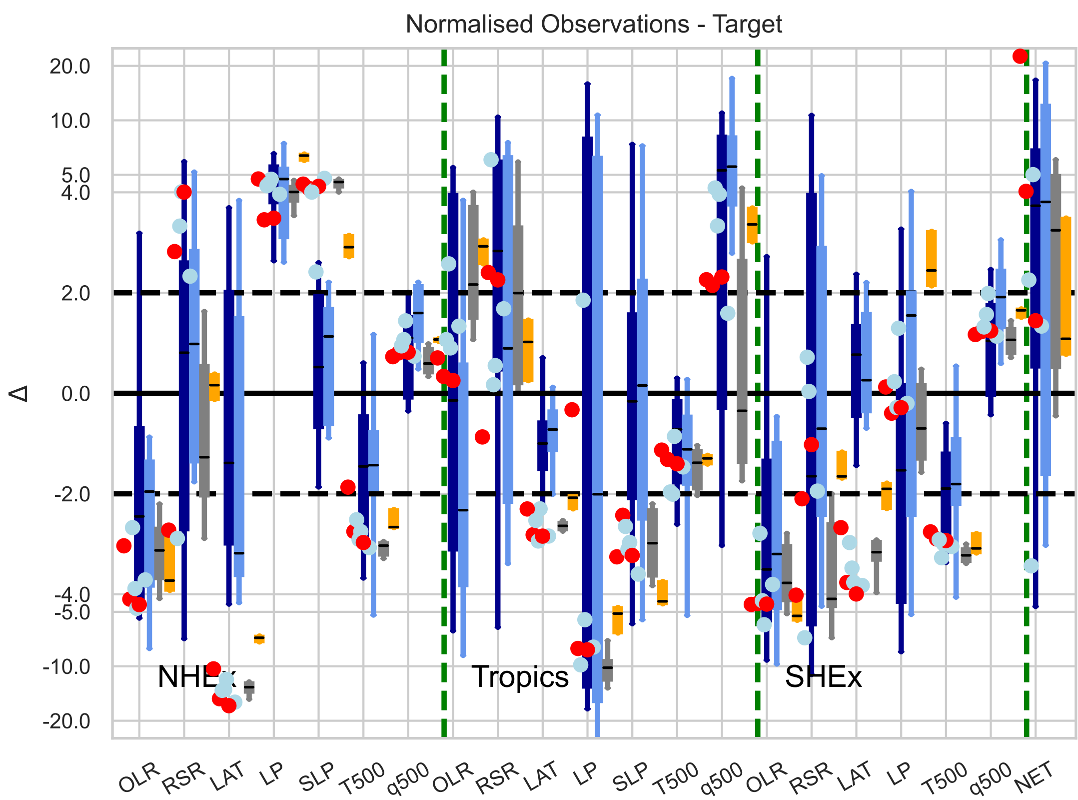


FIG. 3. Simulation minus Observations scaled by estimated error for: Northern Hemisphere extra-tropics (NHX), Tropics and Southern Hemisphere extra-tropics (SHX). Shown are land air temperature (LAT), Land Precipitation (LP), SLP difference from global-average (SLP), Reflected Shortwave Radiation (RSR), Outgoing Longwave Radiation (OLR), Temperature at 500 hPa (T500) and relative humidity at 500 hPa (q500) for CMIP5 (dark blue), CMIP6 (Blue), CE7 (black), and DF14 (Orange) atmosphere-only ensembles as box (25-75%) and whisker (5 to 95 %) plots. Contrasting horizontal lines in box plots show median value. HadAM3 sensitivity studies are shown as blue and red dots for calibrated and uncalibrated cases respectively. Horizontal dashed line at ±2 shows region of observational consistency. Scale is linear between ±4 and logarithmic outside that range.

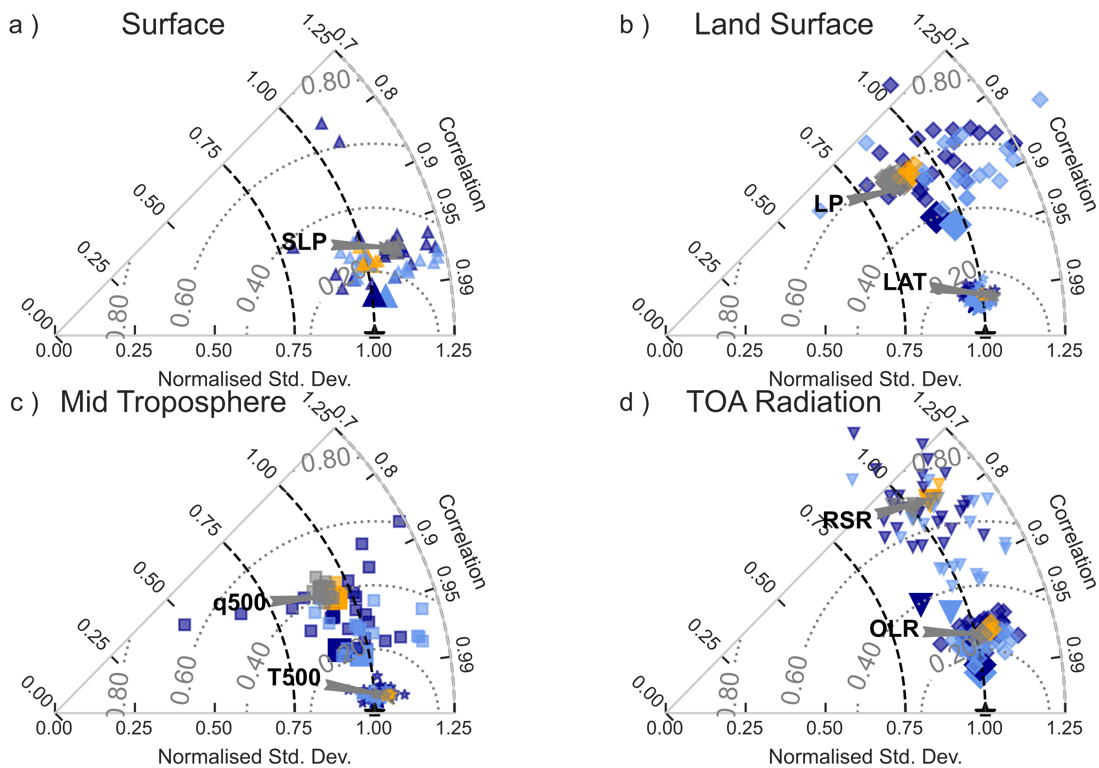


FIG. 4. Partial Taylor diagram for: a) sea level pressure (SLP;Triangles); b) land air temperature (LAT;stars) & precipitation (LP;diamonds); c) 500 hPa relative humidity (q500; squares) & temperature (T500; stars) d) TOA outgoing LW radiation (OLR; diamonds) & reflected SW radiation (RSR; upside down triangles). Shown in all plots are the CMIP5 models (dark blue), CMIP6 (pale blue), the DF14 ensemble (Orange), and the CE7 ensemble (grey). Large symbols show the multi-model average for each ensemble. The label and grey arrow points to the standard HadAM3 model. For each wedge the distance from the origin is the simulated area-weighted standard deviation normalised by the observed area-weighted standard deviation. The angle shows the correlation between observations and simulation, and dotted contour lines show normalised RMS difference.

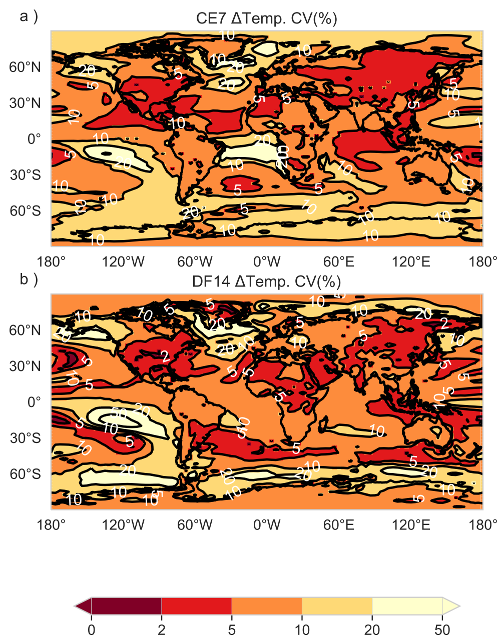


FIG. 5. Coefficient of Variation(%) for temperature change at $4 \times CO_2$ for CE7 (a) and DF (b) ensembles. Colours and contours at 0, 2, 5, 10, 20 and 50%.

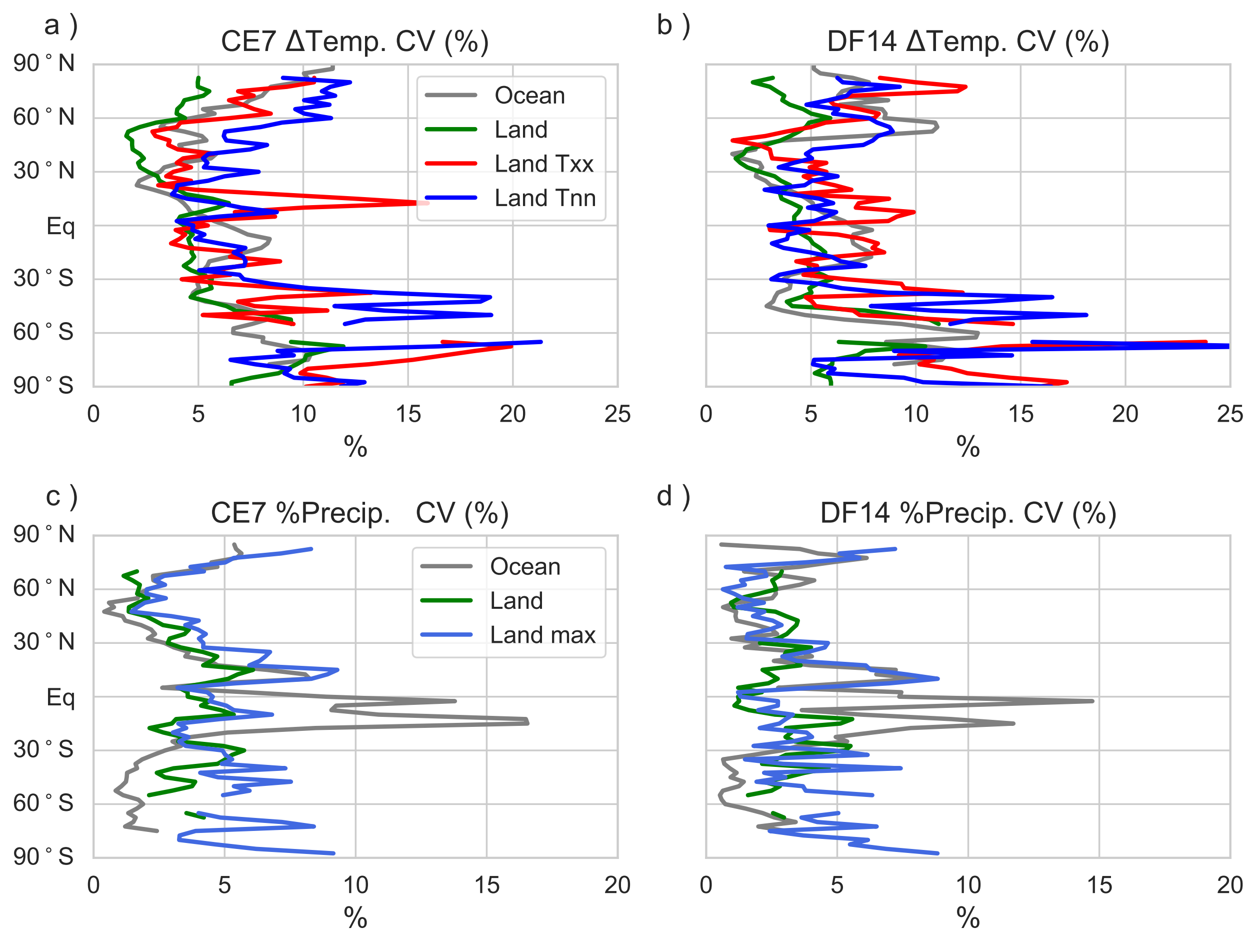


FIG. 6. Coefficient of Variation (%) of zonal-mean temperature change at $4\times CO_2$ for ocean (grey), land (green), land annual maximum (red) and land annual minimum (blue) for CE7(a) and DF14(b) ensembles. CV (%) for % change in ocean (grey), land (green) and annual maximum (dark blue) precipitation relative to **control** simulation for CE7 (c) and DF14(d) ensembles. Locations where the estimated control precipitation was less than $10^{-5}(10^{-4})$ Kg m^{-2}s^{-1} for land/ocean (annual maximum land) were ignored in the zonal-mean calculation.

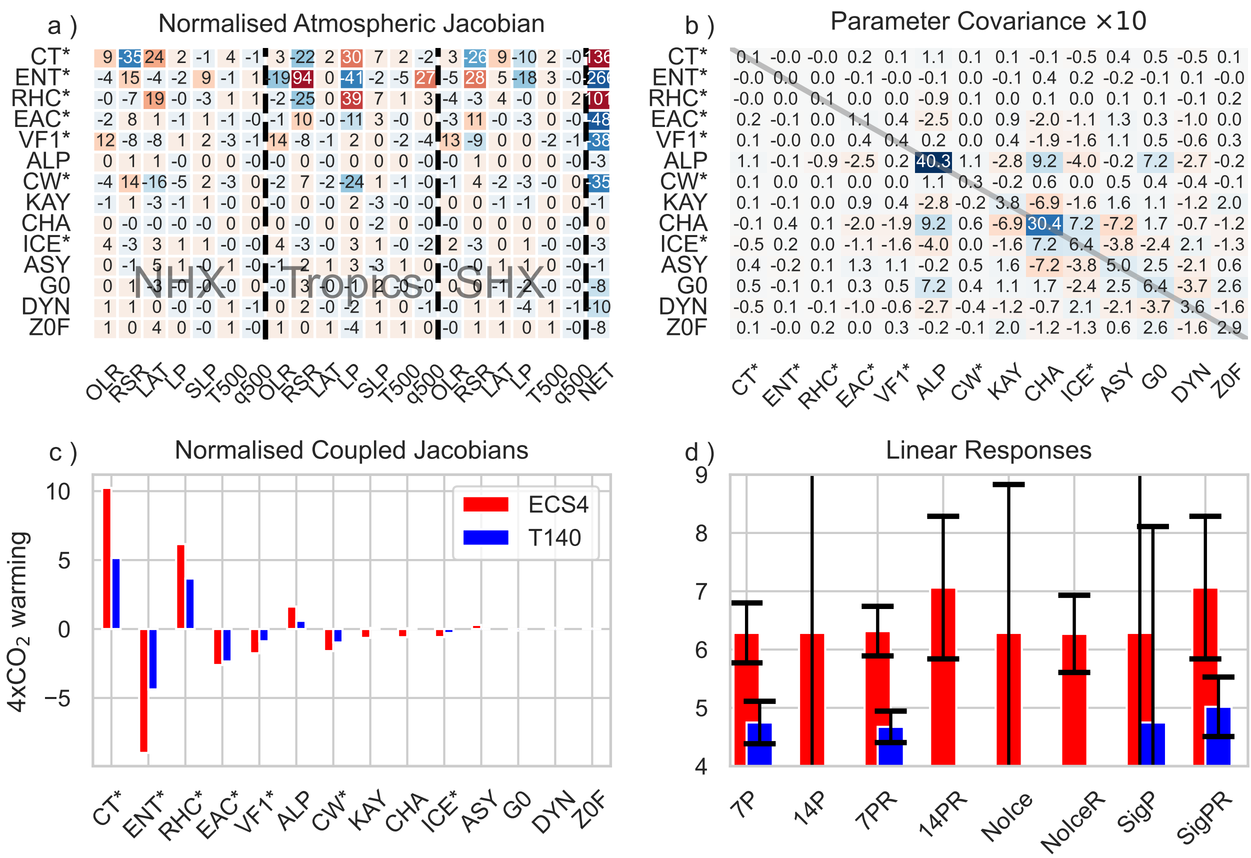
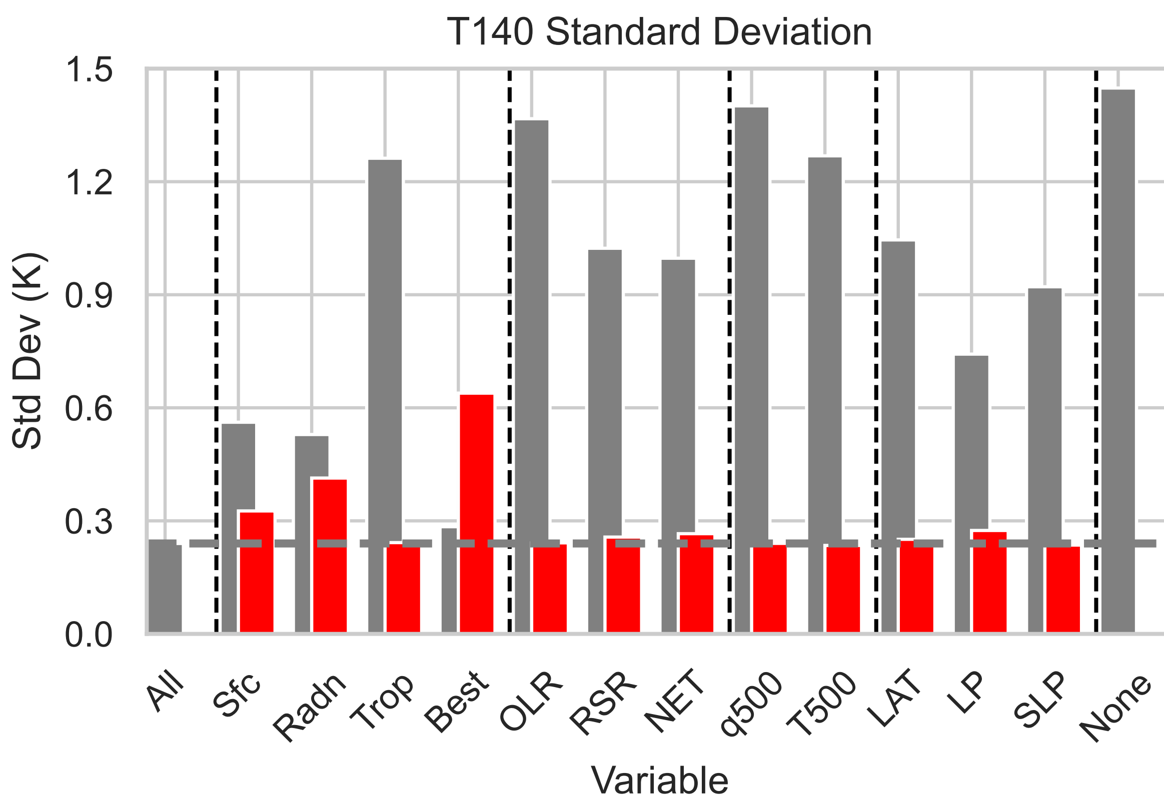


FIG. 7. a) Atmospheric Jacobian for observations normalised by their uncertainty estimates (largely observational uncertainty) where x-axis is simulated observation and y-axis is parameter. Vertical dashed lines divide up the Northern Hemisphere extra-tropics, Tropics and Southern Hemisphere extra-tropics and the global average net flux. b) 10× normalised parameter co-variance (see methods) after constraint applied. c) Jacobian for ECS4 (Red bars)& T140 (blue bars) ordered by absolute ECS4. d) Estimated ECS4 & T140 (red & blue bars; y-axis) and $\pm 2\sigma$ (error-bars) for seven parameters (7P), fourteen parameters (14P), all parameters excluding ALPHAM (NoIce), and the ten parameters with $\geq \sigma$ impact on ECS4 (sigP) in K. R appended shows when the normalised parameter values are limited to $(0, 1)$ (see methods). T140 values not shown for 14P, 14PR, NoIce, and NoIceR cases as TCR Jacobian not complete for all parameters. Both Jacobians are with respect to normalised parameter where 0 is minimum value and 1 is maximum and parameters use short names (Table 1). Parameters with * appended are the seven parameter cases.



498 FIG. 8. Standard deviation for T140. For each analysis all variables, except the named variable or group of variables, have their uncertainty increased by 100 times ("other"). This, in effect, means those observations do not constrain the parameters and T140. All is all variables, Sfc is LAT, LP and SLP, Radn is OLR, RSR and NET, while Trop is q500 and T500. Best is LP, RSR and Net and None is when all observational uncertainties are scaled. Red bars show standard deviations when only that variable, or group of variables, had its uncertainty increased by a factor of 100 ("leave-out"). Horizontal dashed line show value for All analysis while vertical dashed lines separate the variables that contribute to Sfc, Radn, and Trop groups.

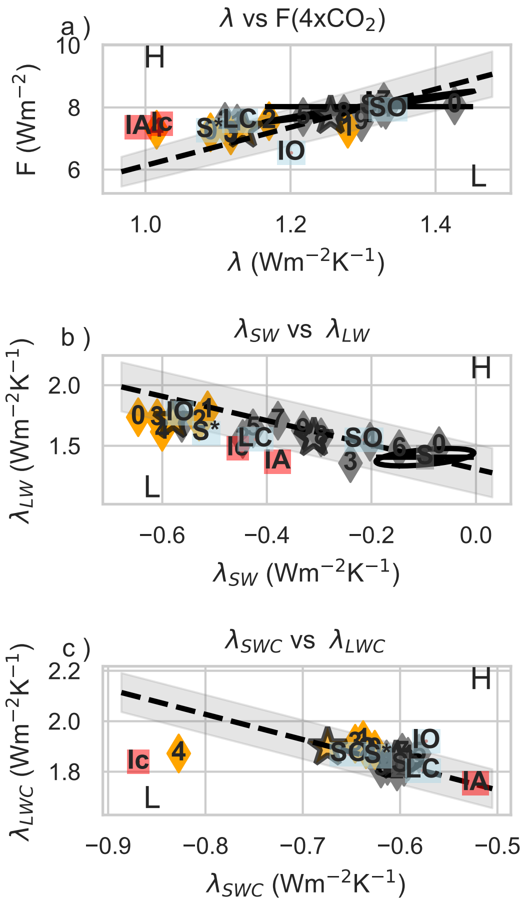


FIG. A1. Scatter plots at 4× CO_2 for CE7 (orange) and DF14 (grey) calibrated ensembles, and sensitivity studies (blue/red boxes). Stars show ensemble means. a) Forcing ($F(4\times CO_2)$ vs climate feedback (λ) ; b) SW climate feedback (λ_{SW}) vs LW climate feedback (λ_{LW}); c) Clear sky SW climate feedback (λ_{SWC}) vs clear sky LW climate feedback (λ_{LWC}). Black ellipses are centred on the Standard HadCM3 configuration and shows 2σ joint-uncertainty ellipse computed from initial condition ensemble while cross shows 2σ errors for x and y variables separately. Dashed lines show ECS4 (a), λ (b) and λ_C(c) fixed at standard values while grey region shows $\pm 2\sqrt{2}\sigma$ internal variability range around standard configuration for this parameter. H and L indicate which side of the dashed line where these values are higher or lower than standard model.