

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: <https://doi.org/10.1175/JCLI-D-21-0434.1> Available at <https://centaur.reading.ac.uk/102092/>

It is advisable to refer to the publisher's version if you intend to cite from the work. See [Guidance on citing](#).

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

Publisher: American Meteorological Society

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the [End User Agreement](#).

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online



Generated using the official AMS L^AT_EX template v6.1

Does Model Calibration Reduce Uncertainty in Climate Projections?

Simon F. B. Tett^a, Jonathan M. Gregory^{b,c}, Nicolas Freychet^a, Coralie 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*

⁹ *Corresponding author: Simon Tett, simon.tett@ed.ac.uk*

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

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

39 **1. Introduction**

40 Charney et al. (1979) estimated that the equilibrium warming for doubled atmosphere CO₂
41 concentration (the Equilibrium Climate Sensitivity; ECS) is between 1.5 and 4.5K. Despite many
42 years of research, Working Group 1 of the Intergovernmental Panel on Climate Change in its Fifth
43 Assessment Report arrived at the same numerical range, though with much greater understanding
44 of the uncertainty (Stocker et al. 2013). Sherwood et al. (2020) (S2020) carried out a comprehen-
45 sive assessment of literature on climate sensitivity, and combined evidence from processes (largely
46 clouds), paleo-climate (largely the Last Glacial Maximum and mid-Pleistocene warm period), and
47 observed changes in climate. They defined an effective climate sensitivity (S) which is the ECS es-
48 timated following the linear-regression method of Gregory et al. (2004). Uncertainties in observed
49 change and paleo-climate include a considerable contribution from "pattern" uncertainty. S2020
50 reported a *likely* range of 2.3-4.5K for S. Building on this, the most recent IPCC assessment(IPCC
51 2021) reported a *likely* range of 2.5-4K for ECS with a best estimate of 3K. They also reported that
52 some models had climate sensitivities inconsistent with this range.

53 Estimates of the Transient Climate Response (TCR), which is the warming at the time of doubled
54 CO₂ in a transient simulation with CO₂ increasing by 1%/year, also have a large spread, with a
55 *likely* (66% confidence) range of 1.4-2.2K(IPCC 2021). This uncertainty has implications for the
56 global budget for CO₂ emissions required to limit temperature rise, because TCR is a factor in the
57 Transient Climate Response to Emissions (Gillett et al. 2013). A review by Knutti et al. (2017)

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

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

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

83 An under-explored issue is the role of model calibration in which model parameters are modified
84 to reduce the discrepancy between simulation and observations (Mauritsen et al. 2012). So, we
85 pose the question: how much uncertainty is there in ECS and TCR when a climate model is
86 objectively calibrated to a diverse set of large scale climatological observations? Climate models
87 are subjectively tuned to current observations (Mauritsen et al. 2012; Hourdin et al. 2017) with

88 almost all modelling groups (Hourdin et al. 2017) using the net top of atmosphere flux as a target
89 though a wide diversity of additional targets are used by different groups. Tett et al. (2013a)
90 showed that it was possible to calibrate four parameters in a climate model to top-of-atmosphere
91 (TOA) radiative flux measurements and that uncertainty in ECS was small (Tett et al. 2013b). Tett
92 et al. (2017)(T17 from hereon) built on this to show it was possible to calibrate the atmospheric
93 component (HadAM3;Pope et al. (2000)) of the venerable HadCM3 climate model (Gordon et al.
94 2000) driven by observed Sea Surface Temperatures, sea-ice and radiative forcings targeting a
95 broad set of large space and time scale atmospheric variables. We build on this work by generating,
96 using two different algorithms, two calibrated ensembles of the HadAM3 model, coupling them to
97 the HadCM3 ocean model and examining the climate response of the two ensembles. We find that
98 uncertainties in the climate response are small both at the global and regional scales suggesting
99 that the structural way in which models represent unresolved processes is key to uncertainty in
100 projections.

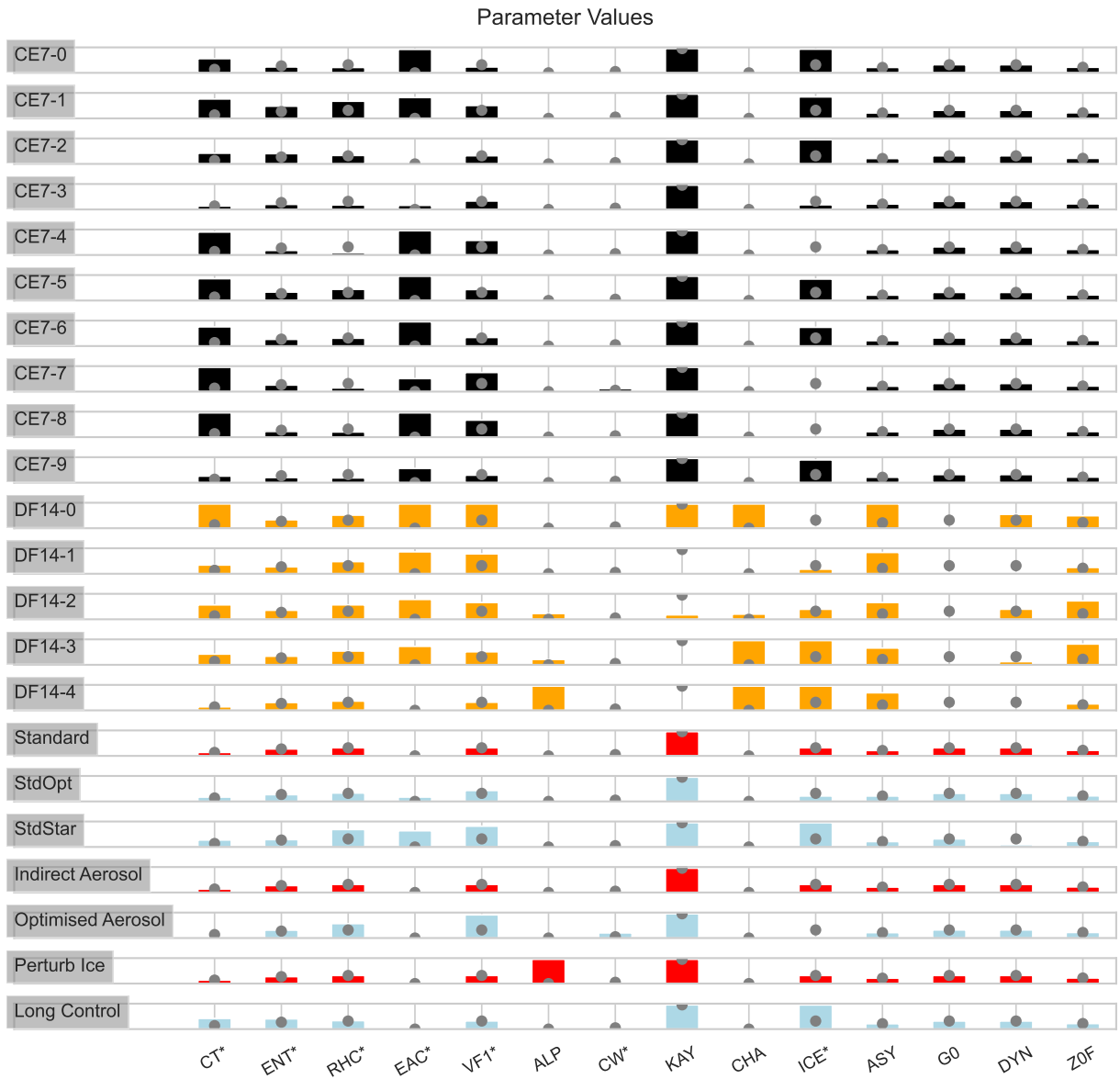
101 The rest of the paper is structured as follows. First we detail the methods used to generate the
102 ensembles and our analysis methodology. We then show results from the two ensembles, followed
103 by a set of sensitivity studies. We then report on results from a linear analysis which allows us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

105 **2. Methods**

106 *a. Calibration and Experimental Design*

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

121 We then coupled the calibrated atmospheric-model configurations to the HadCM3 ocean model,
122 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.

140 TABLE 1. Parameter descriptions and normalised perturbations used to compute Jacobians. Short names used
 141 throughout paper are the first three characters with any _ removed. Those with a * after are used in the CE7
 142 ensemble. See Yamazaki et al. (2013) and T17 for fuller description of parameters. The table shows parameter
 143 name, which process it impacts, the normalised perturbation, to 3 sig. figures, used to compute the atmospheric
 144 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
RHCRT*	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

123 (Gordon et al. 2000). With each coupled configuration we ran a **control** with unchanged CO₂, and
 124 other experiments with changes in CO₂ imposed (described below).

125 The calibration procedure (Tett et al. 2017) chooses parameter vectors for HadAM3 to minimise
 126 the weighted squared difference between simulated control and observed climatological monthly
 127 means for March 2000 to February 2005 (inclusive), following a 16-month spinup. The calibration
 128 considered geographical fields of large-scale land air temperature (LAT), land precipitation (LP),
 129 pressure differences from the global mean (SLP), TOA outgoing longwave radiation (OLR), TOA
 130 reflected shortwave radiation (RSR), 500 hPa temperature (T500) & relative humidity (q500). For
 131 each variable, except SLP, the globe was divided into three regions and area-weighted and time
 132 means computed. The three regions considered were the Northern Hemisphere extra-tropics (NHX;
 133 latitude > 30°N), Tropics (latitude between ±30°), and the Southern Hemisphere extra-tropics
 134 (SHX; latitude < 30°S) allowing representation of different large scale climate regimes. For SLP,
 135 instead of three independent quantities, the two differences (NHX average – global average) and
 136 (Tropics average – global average) were used. Global-average TOA net radiative flux (NET, *N*) was
 137 included as a further constraint with a target value of 0.5 W/m². The atmospheric model was tuned
 138 to these 21 observations by modifying parameters (Table 1) that earlier work had used (Knight
 139 et al. 2007; Yamazaki et al. 2013; Rowlands et al. 2012).

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

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

146 where \mathbf{p} is the vector of parameter values and $\mu = 0.01$ is a penalty weight on the net radiative
147 balance. \mathbf{C} is a covariance matrix formed by summing an estimate of observational uncertainty
148 with twice the control variability. We do this because both the simulations and the observations
149 are assumed to contain chaotic internally generated unforced variability with the same statistical
150 characteristics as the control. The observational uncertainty component of \mathbf{C} had all off-diagonal
151 values set to zero. Uncertainties for OLR and RSR come from the analysis of Loeb et al. (2009),
152 while other observations used the difference between two independent estimates (see T17 for
153 details). n is the number of observables (20 in our case – three regions x six quantities plus two
154 SLP values); \mathbf{o} is a vector of the observed targets while \mathbf{s} are the simulated values. If our estimates
155 of observational uncertainty is reliable, and if \mathbf{C} is diagonal implying F is χ^2 -distributed, the
156 5-95% confidence range for F is 0.6–1.6.

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

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

173 After 40 years of **control** simulations three simulations were carried out in which 1) CO₂
174 increased at a rate of 1%/year until quadrupling (**1pctCO2**); 2) was instantaneously doubled
175 (**abrupt2xCO2**); 3) and quadrupled (**abrupt4xCO2**). The **abrupt2xCO2** and **abrupt4xCO2**
176 cases were both integrated for 40 years while the **1pctCO2** case was ran for 140 years. We
177 focus on the differences between the forced simulations and their **control**, especially the transient
178 responses at 2 (TCR) and 4×CO₂ (T140) in **1pctCO2**, the equilibrium climate sensitivity (ECS) in
179 **abrupt2xCO2** and the equilibrium response to 4×CO₂ (ECS4) in **abrupt4xCO2**. All calculations
180 are done on the difference between forced and **control** simulation in order to correct for residual
181 drifts. In Appendix A2 we report on a sensitivity study where we ran a control for 1000 years
182 before starting the increased CO₂ simulations. We found only a small impact.

183 ECS and ECS4 were estimated by regressing net Top-of-Atmosphere (TOA) flux against global-
184 mean temperature (Gregory et al. 2004). When obtained by this method, rather than from an
185 equilibrium 2×CO₂ state, the estimated ECS is commonly called “effective climate sensitivity”.
186 Similar calculations were done for other variables to estimate the equilibrium responses at 2× and
187 4× CO₂. Feedback parameters for the all-sky (λ) and clear-sky (λ_C), short wave (λ_{SLW} , λ_{SWC})
188 and longwave (λ_{LW} , λ_{LWC}) TOA radiative fluxes were computed from the slope of the appropriate
189 linear regression fit.

190 TCR was diagnosed from the **1pctCO2** simulations by fitting a 2nd order polynomial to the
191 global-average temperature difference from the equivalent **control** simulation. We used a 2nd
192 order polynomial to capture any deviations from a linear response at longer timescales as seen
193 in multiple climate models (Gregory et al. 2015). The value of the fit when CO₂ doubled is our
194 estimate of TCR. We also computed T140 (the warming at 4×CO₂) similarly. We also used this
195 approach for other variables shown. As many of the **control** simulations are still warming at year
196 180, control values are, unless stated otherwise, taken from the value at year 180 estimated from a
197 2nd order polynomial fit to the data.

198 *b. CMIP5 and CMIP6 data*

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

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

215 *c. Uncertainty*

216 Internal variability will contribute to our estimates of the climate response. To estimate the
217 contribution of internal variability to ECS, ECS4 and climate feedback parameters we used an
218 ensemble of seven initial condition simulations of HadCM3 in which CO₂ was doubled and
219 quadrupled. These simulations were all started from the same state with small perturbations and
220 are compared against the same control simulation. To compute uncertainty in the transient and
221 control simulations a 1000-year long control simulation of HadCM3 was used. Segments of length
222 140 years overlapping by 35 years were taken and a second order fit made to this timeseries. Values
223 at year 70 and year 140 were then taken from the 2nd order fit. Variances of these values were then
224 computed and used to estimate uncertainty from internal climate variability. For TCR, T140 and
225 other transient values the variances were doubled as these values are computed from a difference
226 between **1pctCO2** and **control** simulations. For simplicity the same 140-year segments were used
227 to compute uncertainties in the **control** simulation values, although this slightly overestimates their
228 uncertainties.

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

209 TABLE 2. Summary properties for CMIP5 models. ID is the label used in Fig. 2 and other plots. N_{atmos}
 210 and N_{coup} are the sizes of the atmospheric and coupled ensembles. COST is the dimensionless value of the
 211 cost-function. Shown in K are the Equilibrium Climate Sensitivity (ECS), Transient Climate Response (TCR),
 212 Transient Climate Response (T140) at $4\times\text{CO}_2$ and the pre-industrial control global mean surface air temperature
 213 (GMSAT). Source shows where ECS/TCR/T140 values came from and MM Mean shows the multi-model mean
 214 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	–	–

234 *d. Linear Uncertainty Analysis*

235 In this subsection we explain how we compute, using a linear analysis for small perturbations,
 236 the observationally constrained distributions of ECS4, TCR and T140 for HadCM3. In essence
 237 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	–	–

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

241 Assuming small perturbations and that the parameters \mathbf{p} have a multi-variate Gaussian distribution
242 ($\mathbf{p} \sim N(\mathbf{p}_0, \mathbf{C}_p)$) where \mathbf{p}_0 are the optimised parameters, the covariance matrix (\mathbf{C}_p) can be computed

243 (T17) from:

$$\mathbf{C}_p = \mathbf{PCP}^T \quad (1)$$

244 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
245 derivatives w.r.t. parameters in the atmospheric simulations estimated, in our case, using a 14-
246 member ensemble. \mathbf{C} is the observational co-variance matrix defined above. A perturbation analysis
247 for the climate responses ($\mathbf{r} = (ECS4, TCR, T140) \sim N(\mathbf{r}_0, \mathbf{C}_r)$) can be done by computing the
248 Jacobian (\mathbf{J}_r) using **control**, **abrupt4xCO2** and **1pctCO2** coupled simulations for each perturbed
249 parameter. $\mathbf{C}_r = \mathbf{J}_r \mathbf{C}_p \mathbf{J}_r^T$ where \mathbf{r}_0 and \mathbf{C}_r are the responses from the optimised parameter settings
250 and the response covariance matrix respectively. When computing the Jacobian for TCR and
251 T140, only those ten parameters that had a significant impact on ECS4 were perturbed. As there
252 are only small differences between the response of the optimised model and the standard model
253 (Appendix A2) we approximated \mathbf{r}_0 and \mathbf{p}_0 with values for the standard HadCM3 model \mathbf{p}_s .

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

258 To keep the normalised parameters within (0, 1) we generated parameter vectors from the multi-
259 variate normal distribution ($\mathbf{p} \sim N(\mathbf{p}_0, \mathbf{C}'_p)$). For the small fraction of \mathbf{p} where all normalised
260 parameters were in the range (0, 1) we computed changes in ECS4 and T140 from $\mathbf{J}(\mathbf{p} - \mathbf{p}_s)$. We
261 generated at least 1000 realisations of \mathbf{p} with normalised elements between 0 and 1 by random
262 generation and removal of all cases where this was not so. To increase the efficiency of this process
263 \mathbf{C}'_p was computed by combining a prior distribution for the normalised elements $\mathbf{p} \sim N(\mathbf{0.5}, \mathbf{I})$ with
264 \mathbf{C}_p using Bayes theorem. The covariance and best-estimate, for ECS4 and T140, was computed
265 from the \mathbf{p} samples. Uncertainties are summarized by the standard deviation of ECS and T140
266 from these distributions.

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

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

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

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

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

287 **3. Results**

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

293 *a. Representation of Large-scale Climate*

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

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

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

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

349 warmer than those observations, while the standard HadCM3 is slightly cold. The DF14 ensemble
350 has a narrower range than CE7 and four out of five of the members have temperatures consistent
351 with the preindustrial temperature estimates.

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

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

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

383 *b. Climate Response*

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

396 TABLE 4. Ensemble average values for CMIP5, CMIP6, CE7, and CE14 ensembles (to two s.f.). Uncertainties
397 are one standard deviation for each ensemble (to 1 s.f.). Bracketed values are coefficient of variation rounded
398 to 1%. Standard-deviations from initial condition ensembles (ICE) for ECS/ECS4 and internal variability (IV)
399 (see methods) for TCR/T140 are also shown. Also shown are results from the linear analysis for four restricted
400 parameter cases. Sensitivity studies are shown, for ECS4 and T140, on the right. They are a strongly correlated
401 observational covariance matrix(_C), and the expert judgement parameter range doubled (x2) for all seven and
402 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%)			

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

403 The correlation between the atmosphere-only cost function and T140 (ECS4) in the CMIP5
404 ensemble is -0.15 (0.04) neither of which are significant at the 10% level. For the CMIP6 ensemble
405 the correlations are -0.46 and -0.44 for T140 and ECS4 respectively, which are just significant at
406 the 10% level. Even so these are weak correlations suggesting that the cost function applied to
407 multiple models does not provide a strong constraint. Results from our two calibrated ensembles

408 suggest that, once observational constraints have been applied, only a small uncertainty due to
409 parameter choices remains in the transient and equilibrium responses to CO₂. If this is true of
410 other models, it suggests that the much larger uncertainties shown by CMIP in TCR, TCR140, ECS
411 and ECS4 arise from the range of physical parameterisation schemes used (so-called “structural
412 uncertainty”), or from the calibration targets used, rather than from poor calibration.

413 *c. Uncertainties in Regional Climate Change*

414 Having shown that uncertainties in large-scale temperature change and climate feedbacks are
415 small, we consider the CV of regional temperature change at the 4 × CO₂ in the **1pctCO2** simula-
416 tions. These are similar, and small, in the CE7 and DF14 ensembles (Fig. 5) being between 5 and
417 10% across most of the world. Uncertainties in both ensembles are largest:

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

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

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

437 *d. Linear Uncertainty Analysis*

438 To see if our results are robust, we present a linear uncertainty analysis (see Methods). This
439 approach combines observational uncertainty estimates with the sensitivity of atmospheric simula-
440 tions and of the climate response to parameter perturbations to give an observationally constrained
441 distribution for climate response. This approach also allows us to determine which parameters are

442 constrained by the atmospheric observations, which observations constrain the response, and test
443 sensitivity to assumptions about observational uncertainty.

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

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

481 We found similar results to ECS4 for TCR (Table 4) and T140 (Fig. 7(d)). T140 mean and
482 uncertainty both increase when going from seven to ten parameters, largely due to inclusion of the

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

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

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

520 Appendix A1 explores changes in forcing from CO₂ (likely fast responses to CO₂ changes
521 rather than changes in radiative forcing) and feedbacks. We find that all-sky shortwave (λ_{SW})
522 and longwave (λ_{LW}) climate feedbacks do show large changes between the two ensembles and,
523 especially for the CE7 ensemble, within the ensemble. However, total climate feedback changes
524 are small due to near-cancellation between changes in λ_{LW} and λ_{SW} after calibration.

525 **4. Discussion and Conclusions**

526 Using two different approaches, we find that the large-scale response of HadCM3 (Gordon et al.
527 2000) to CO₂ increase is strongly constrained when the simulated control climate is objectively
528 calibrated against multiple large-scale 5-year mean atmospheric observations. Observations of
529 land precipitation, reflected shortwave radiation, and net flux provide the strongest observational
530 constraints on the model. Observational estimates of pre-industrial global-average temperature
531 give an independent test on the ability of the HadCM3 to simulate large scale climate. Most,
532 but not all, of the calibrated models are in agreement with this observation. Using the DFOLS
533 algorithm (Cartis et al. 2019) to calibrate the atmospheric component of HadCM3 (Pope et al.
534 2000) we find it is possible to produce model configurations that are in much better agreement
535 with large-scale observations than the standard configuration, and than almost all of the CMIP5
536 and CMIP6 models. For model calibration, it appears that DFOLS is better than the Gauss-Newton
537 method used in Tett et al. (2017).

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

549 observations we have identified are land precipitation, reflected solar radiation and net flux into the
550 Earth system.

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

555 In Appendix A2 we report on a series of sensitivity studies. We find that use of a short spinup
556 does not make a very large difference to our results. We also found that changes in the ice-albedo
557 hyper-parameter had little impact on the cost function but a modest impact on the model response.
558 Structural changes to the model physics through inclusion of aerosol impact on cloud properties
559 (Jones et al. 2001) had a relatively large impact on the both HadCM3's ability to simulate current
560 climate and its response to CO₂. However, we found this was due to changes in the diagnosed
561 forcing from CO₂ rather than to changes in feedbacks. We speculate that this is due to changes in
562 fast cloud feedbacks. Changes to the representation of ice-crystals in the model's radiation scheme
563 had little impact. Thus, structural changes in HadCM3 can have a significant impact on its response
564 but in a surprising way.

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

573 We believe, making the plausible assumption that there is nothing unusual about HadCM3, that
574 our results will hold for other models. Thus, for any specific model, uncertainty in climate response
575 will be small if the model parameters are calibrated against multiple observations. This may be
576 sensitive to the cancellation of SW and LW feedbacks from cloud changes seen in HadCM3. Since
577 we found no robust linear relationship between our calibration metric and climate response in
578 the CMIP5 and CMIP6 ensembles, and the changes in HadCM3 response with changes to model

579 physics, we suggest that uncertainty in the two ensembles largely arises from structural differences.
580 If so, calibrated perturbed physics ensembles (such as the UKCP18 ensemble; Lowe et al. (2019))
581 have likely too small an uncertainty range for future climate change, because they do not address
582 structural uncertainty.

583 However, the possibility that different groups have followed different calibration strategies can
584 not be ruled out as a source of uncertainty in model response to CO₂ and other forcings. Moving
585 to an objective and documented approach to model calibration rather than the current *ad hoc*
586 approach (Hourdin et al. 2017) would help understand this. Based on our results, using objective
587 methods to calibrate climate (or Earth System) models to large-scale observations is likely to
588 improve their ability to simulate current large-scale mean states, and *may* narrow the range of
589 model projections. However, it is likely that structural uncertainty arising from different choices in
590 how to parameterise unresolved processes is also important. In summary, to reduce the recalcitrant
591 uncertainty in model response to greenhouse gases and other forcings requires much more focus
592 on how models represent unresolved processes than there may have been hitherto.

593 *Acknowledgments.* We thank the two anonymous referees for their comments which improved
594 the paper. ST, CC and MM were funded by NERC (OptClim: NE/L012146/1) with simulations
595 and post-processing done on the Edinburgh Compute and Data Facility partially funded by Clima-
596 teXchange. NF was supported by the UK-China Research & Innovation Partnership Fund through
597 the Met Office Climate Science for Service Partnership(CSSP) China as part of the Newton Fund.
598 JMG was funded by the National Centre for Atmospheric Science. LR was funded by the EPSRC
599 Centre For Doctoral Training in Industrially Focused Mathematical Modelling (EP/L015803/1) in
600 collaboration with NAG Ltd. We thank the World Climate Research Programme’s working group
601 on coupled modelling for producing and making available their model output and Dr Mark Ringer
602 (Met Office) for data on CMIP6 transient and equilibrium responses.

603 *Data availability statement.* The latest version of the software used for optimisation and Jacobian
604 computations is available at <https://github.com/SimonTett/ModelOptimisation>.
605 Software used to produce the figures in this paper is available from https://github.com/SimonTett/Jclim21_calibrate while processed data is available
606 at <https://doi.org/10.7488/ds/3051>. The DFOLS software is available from
607 <https://github.com/numericalalgorithmsgroup/dfols>. TCR, T140 & ECS values
608

609 for the CMIP6 ensemble are at <https://github.com/mark-ringer/cmip6>. The full
610 multi-Tbyte dataset of HadCM3 simulations is available at doi:10.7488/84b585fc-57d2-4e5a-
611 b3a3-694f70534a02. To retrieve this data please contact SFBT.

612

613

APPENDIX

614 **A1. Drivers of Climate Response Uncertainty**

615 In this appendix we consider the drivers of the uncertainty in climate response in both ensembles.
616 We start by considering ECS4 which depends on both CO₂ forcing (including rapid adjustment) and
617 the climate feedback parameter ($ECS4 = F(4 \times CO_2)/\lambda$) with both λ and F possibly impacted by
618 changes in model parameters. We next consider the contributions of SW (λ_{SW}) and LW feedbacks
619 (λ_{LW}) to uncertainty with $\lambda = \lambda_{SW} + \lambda_{LW}$ and then similarly for clear-sky feedbacks (λ_{SWC} and
620 λ_{LWC}). To easily assess uncertainty in these joint distribution, relative to the standard model, we fix
621 one of λ , $F(4 \times CO_2)$, and λ_C to the standard values which in the plane being considered is a line.
622 Uncertainties around this line are computed by modifying ECS4, λ and λ_C to their standard value
623 $\pm 2\sqrt{2}\sigma$ where σ is the standard deviation from the 7 member initial condition ensemble. Model
624 configurations within this region have values consistent with the standard model though this may
625 arise from cancellation between processes.

626 Starting with ECS4 and forcing at $4 \times CO_2$ (Fig. A1(a)). Most of the CE7 ensemble members sit
627 inside the internal-variability confidence region suggesting no significant joint change in ECS4 and
628 forcing. All but one of the remaining CE7 members sit within the grey region suggesting that much
629 of the limited variability in ECS in this ensemble arises from cancellation between fast adjustments
630 to CO₂ forcing and feedback strengths. For the DF14 ensemble, relative to the CE7 ensemble, the
631 ensemble-mean has a smaller value of λ and a smaller forcing. The individual members of both
632 ensembles lie close to the constant ECS4 line but with different forcings and climate feedback's.
633 This suggests that internal variability in the estimation of these values produces strongly correlated
634 values (the ellipse in Fig. A1(a) is narrow and strongly oriented along the λ -F line) and that the
635 calibration process modifies feedbacks and the fast response to CO₂ such that ECS4 changes little.
636 One exception to this cancellation is the DF14-4 case which has higher TCR140 and ECS4 (Fig. 2)

637 than any of the other ensemble members. This occurs because λ is smaller than the rest of the
638 ensemble with similar CO₂ forcing.

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

665 The DF14-4 case is an outlier in that it has a weaker climate feedback strength and so higher
666 ECS₄, if fast CO₂ feedbacks do not change, than the other ensemble members. Considering the
667 all-sky SW and LW feedback strengths this case is not obviously different from the rest of the
668 ensemble. However, the clear-sky SW feedback strength is much more negative than the rest of the
669 ensemble. Several parameters from this case differ from the rest of the ensemble (Fig. 1) but one
670 parameter that has a large difference is ALPHAM. This parameter controls the albedo of sea-ice
671 and so changes in it might be expected to impact clear sky SW feedbacks.

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

681 TABLE A1. Sensitivity cases. All optimised cases started with default parameters and normalised parameter
 682 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

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

677 A2. Sensitivity Studies

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

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

690 The linear analysis and Appendix A1 suggested that the sea-ice albedo hyper-
 691 parameter(ALPHAM) might explain some of the differences between the two ensembles. To
 692 test this we carried out a set of simulations (Ic) in which ALPHAM was set to its maximum value

693 with all other parameters at their standard value. This configuration had a cost-function similar to
694 the standard model suggesting that this parameter, as expected, has little impact on the atmospheric
695 simulation. However, its control temperatures are much colder than any other case (Fig. 2 and
696 Table A1) Further, ECS4 and T140 are larger than all optimised cases consistent with the linear
697 analysis and the DF14-4 case.

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

706 To explore the role that structural uncertainty might play in our results we carried out two further
707 calibrations of HadAM3, using the Gauss Newton algorithm of T17, in which the model physics
708 was changed and then the calibrated atmospheric model coupled to the ocean model (Table A1). In
709 one (StdStar;S*) we changed the properties of ice crystals in the radiation code and then optimised
710 using the same seven parameters as CE7 plus the model diffusion hyper-parameter. In another
711 (Optimised Interactive Aerosol;IO) we added an interactive aerosol indirect effect (Jones et al.
712 2001) and optimised using the same seven parameters as used in CE7. Both calibrated models had
713 cost function values smaller than any of the CE7 ensemble members and about 15% smaller than
714 the standard model.

715 S* is very similar to the standard model though with somewhat higher values of ECS4 and T140.
716 The SW and LW all-sky feedbacks in this configuration are very different from the standard model
717 but the changes offset one another. In combination with a smaller forcing from CO₂, than the
718 standard configuration, this leads to a similar climate responds.

719 The optimised interactive aerosol configuration (IO) has T140 and ECS4 values significantly
720 below both the standard model and both calibrated ensembles (Table A1; Fig. 2). This model
721 has a significantly smaller ECS and forcing from 4×CO₂ than the standard configuration with
722 its LW and SW feedback parameters very close to the DF14 ensemble mean values. Its total

723 feedback parameter is similar to the standard configuration (Fig. A1(b)) but its diagnosed forcing
724 in **abrupt4xCO2** is much smaller than the standard configuration's value (Fig. A1(a)). It shows
725 quite dramatic changes in the SW and LW feedbacks but these cancel leading to only a small change
726 in total feedback. This model also shows changes in the clear sky feedbacks with a shift to weaker
727 clear sky feedback. Thus, the reason for the changes in T140 and ECS4 in this configuration are
728 due to relatively fast changes in the atmosphere in response to changes in CO₂ rather than changes
729 in climate feedbacks.

730 The unoptimised model with the interactive indirect aerosol scheme produces a model that has a
731 worse simulation of large scale climate, and much larger climate responses than the standard and
732 optimised aerosol configurations as well as many of the CMIP5 and CMIP6 models (Fig. 2). This
733 configuration, unlike the calibrated cases, changes the all-sky SW feedback and is also significantly
734 different clear sky feedbacks. This, in turn, suggests it is not the impact of aerosols *per say* that
735 changes the response but the calibration of other processes to produce a reasonable simulation that
736 then modify the fast response to CO₂ forcing.

737 Overall, the effect of calibration in the sensitivity studies is to generate configurations that have
738 climate responses that are similar to that of the standard configuration. This arises from near-
739 cancellation between SW and LW climate feedback strength, and then between CO₂ forcing and
740 total climate feedback strength.

741 **References**

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

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

748 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*

751 *Sciences, National Research Council*. National Academies Press, URL [http://www.nap.edu/](http://www.nap.edu/catalog/12181.html)
752 [catalog/12181.html](http://www.nap.edu/catalog/12181.html).

753 Gillett, N. P., V. K. Arora, D. Matthews, and M. R. Allen, 2013: Constraining the ratio of
754 global warming to cumulative CO₂ emissions using CMIP5 simulations. *Journal of Climate*,
755 **26 (18)**, 6844–6858, <https://doi.org/10.1175/JCLI-D-12-00476.1>, URL [https://doi.org/10.1175/](https://doi.org/10.1175/JCLI-D-12-00476.1)
756 [JCLI-D-12-00476.1](https://doi.org/10.1175/JCLI-D-12-00476.1), <https://doi.org/10.1175/JCLI-D-12-00476.1>.

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

760 Gregory, J. M., T. Andrews, and P. Good, 2015: The inconstancy of the transient climate re-
761 sponse parameter under increasing CO₂. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373 (2054)**, 20140417, [https://doi.org/](https://doi.org/10.1098/rsta.2014.0417)
762 [10.1098/rsta.2014.0417](https://doi.org/10.1098/rsta.2014.0417).

764 Gregory, J. M., and Coauthors, 2004: A new method for diagnosing radiative forcing and climate
765 sensitivity. *Geophys. Res. Lett.*, **31**, L03 205, <https://doi.org/10.1029/2003gl018747>.

766 Grose, M. R., J. Gregory, R. Colman, and T. Andrews, 2018: What climate sensitivity index is
767 most useful for projections? *Geophysical Research Letters*, **45 (3)**, 1559–1566, [https://doi.org/](https://doi.org/10.1002/2017gl075742)
768 [10.1002/2017gl075742](https://doi.org/10.1002/2017gl075742).

769 Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in
770 future climate change. *Geophys. Res. Lett.*, **33**, L03 502, <https://doi.org/10.1029/2005GL025127>.

771 Hourdin, F., and Coauthors, 2017: The art and science of climate model tuning. *Bul-*
772 *letin of the American Meteorological Society*, **98 (3)**, 589–602, [https://doi.org/10.1175/](https://doi.org/10.1175/BAMS-D-15-00135.1)
773 [BAMS-D-15-00135.1](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.](https://doi.org/10.1175/BAMS-D-15-00135.1)
774 [1175/BAMS-D-15-00135.1](https://doi.org/10.1175/BAMS-D-15-00135.1).

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

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

780 Knight, C. G., and Coauthors, 2007: Association of parameter, software, and hardware variation
781 with large-scale behavior across 57,000 climate models. *Proc. Natl. Acad. Sci. U. S. A.*, **104** (30),
782 12 259–12 264, <https://doi.org/10.1073/pnas.0608144104>.

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

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

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

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

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

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

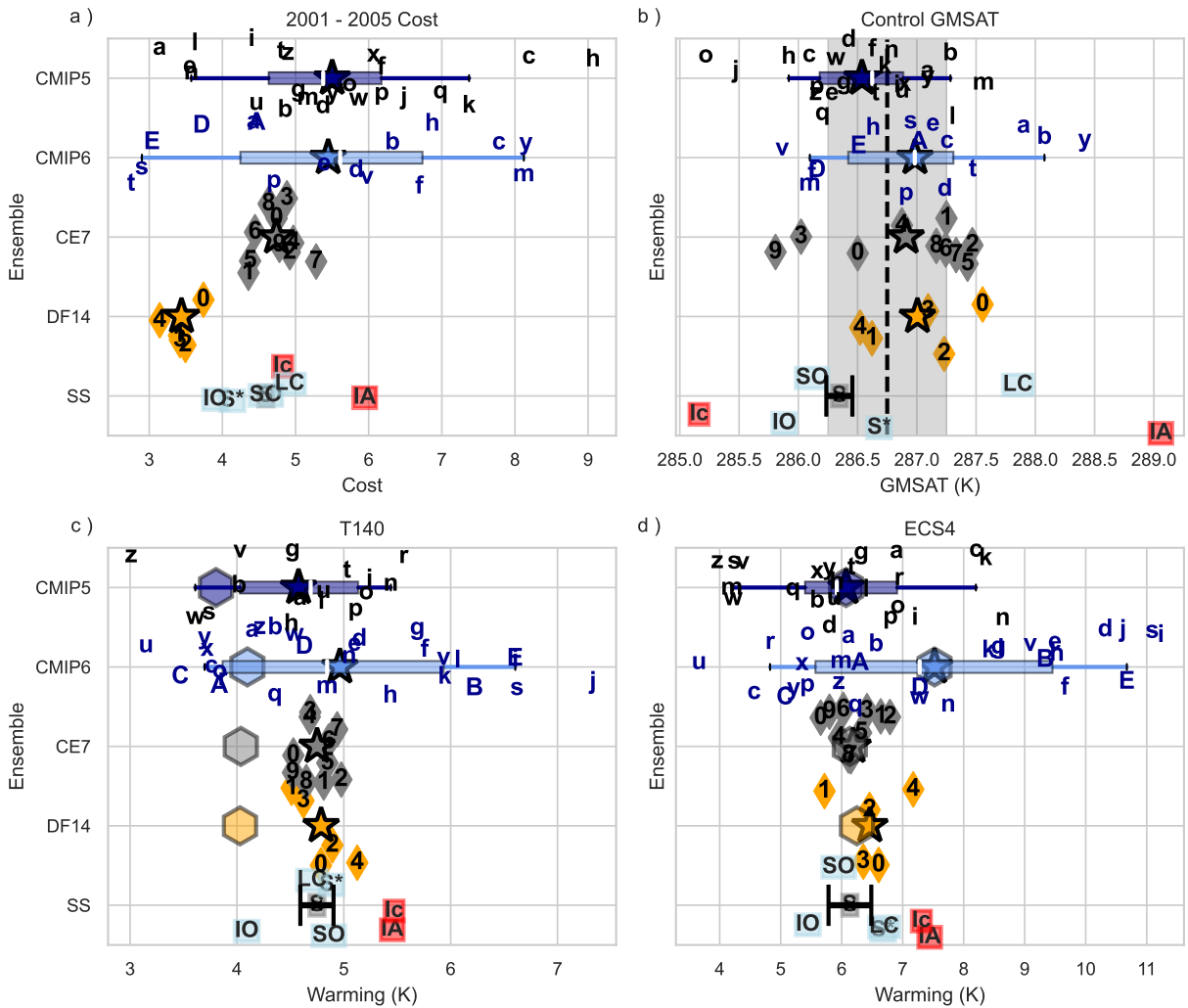
796 Roach, L. A., S. F. B. Tett, M. J. Mineter, K. Yamazaki, and C. D. Rae, 2017: Automated parameter
797 tuning applied to sea ice in a global climate model. *Climate Dynamics*, 1–15, [https://doi.org/
798 10.1007/s00382-017-3581-5](https://doi.org/10.1007/s00382-017-3581-5).

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

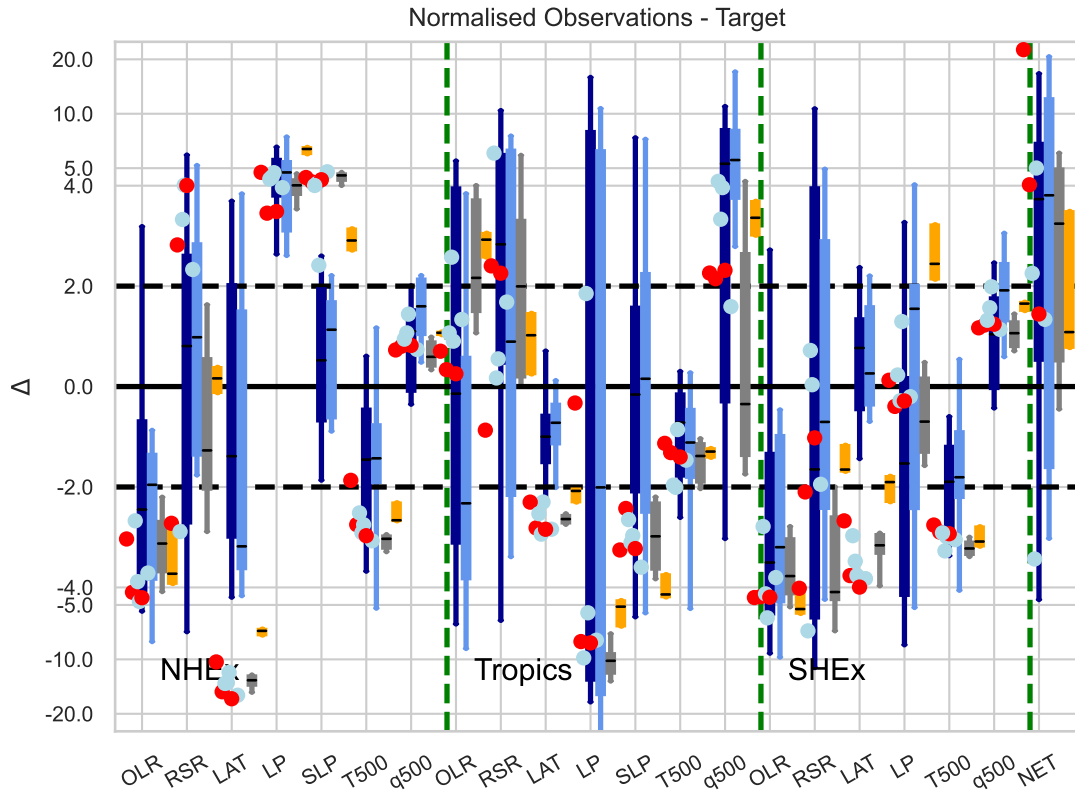
801 Sanderson, B. M., 2011: A multimodel study of parametric uncertainty in predictions of climate
802 response to rising greenhouse gas concentrations. *J. Clim.*, **34**, 1362–1377, [https://doi.org/
803 10.1175/2010JCLI3498.1](https://doi.org/10.1175/2010JCLI3498.1).

- 804 Sanderson, B. M., A. G. Pendergrass, C. D. Koven, F. Brient, B. B. Booth, R. A. Fisher,
805 and R. Knutti, 2021: The potential for structural errors in emergent constraints. *Earth System*
806 *Dynamics*, **12** (3), 899–918, <https://doi.org/10.5194/esd-12-899-2021>.
- 807 Schlund, M., A. Lauer, P. Gentine, S. C. Sherwood, and V. Eyring, 2020: Emergent constraints on
808 equilibrium climate sensitivity in CMIP5: do they hold for CMIP6? *Earth System Dynamics*,
809 **11** (4), 1233–1258, <https://doi.org/10.5194/esd-11-1233-2020>.
- 810 Sexton, D. M. H., and Coauthors, 2021: A perturbed parameter ensemble of HadGEM3-GC3.05
811 coupled model projections: part 1: selecting the parameter combinations. *Climate Dynamics*,
812 **56** (11-12), 3395–3436, <https://doi.org/10.1007/s00382-021-05709-9>.
- 813 Sherwood, S. C., and Coauthors, 2020: An assessment of earth's climate sensitivity using multiple
814 lines of evidence. *Reviews of Geophysics*, **58** (4), <https://doi.org/10.1029/2019rg000678>.
- 815 Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising
816 levels of greenhouse gases. *Nature*, **433**, 403–406.
- 817 Stocker, T., and Coauthors, 2013: *IPCC, 2013: climate change 2013: the physical science basis.*
818 *Contribution of working group I to the fifth assessment report of the intergovernmental panel on*
819 *climate change*. Cambridge University Press.
- 820 Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram.
821 *Journal of Geophysical Research: Atmospheres*, **106** (D7), 7183–7192.
- 822 Tett, S. F. B., M. J. Mineter, C. Cartis, D. J. Rowlands, and P. Liu, 2013a: Can top of atmosphere
823 radiation measurements constrain climate predictions? part 1: Tuning. *J. Climate*, **26**, 9348–
824 9366, <https://doi.org/10.1175/JCLI-D-12-00595.1>.
- 825 Tett, S. F. B., D. J. Rowlands, M. J. Mineter, and C. Cartis, 2013b: Can top of atmosphere
826 radiation measurements constrain climate predictions? part 2: Climate sensitivity. *J. Climate*,
827 **26**, 9367–9383, <https://doi.org/10.1175/JCLI-D-12-00596.1>.
- 828 Tett, S. F. B., K. Yamazaki, M. J. Mineter, C. Cartis, and N. Eizenberg, 2017: Calibrating
829 climate models using inverse methods: Case studies with HadAM3, HadAM3P and HadCM3.
830 *Geoscientific Model Development*, **10**, 3567–3589, <https://doi.org/10.5194/gmd-2016-305>.

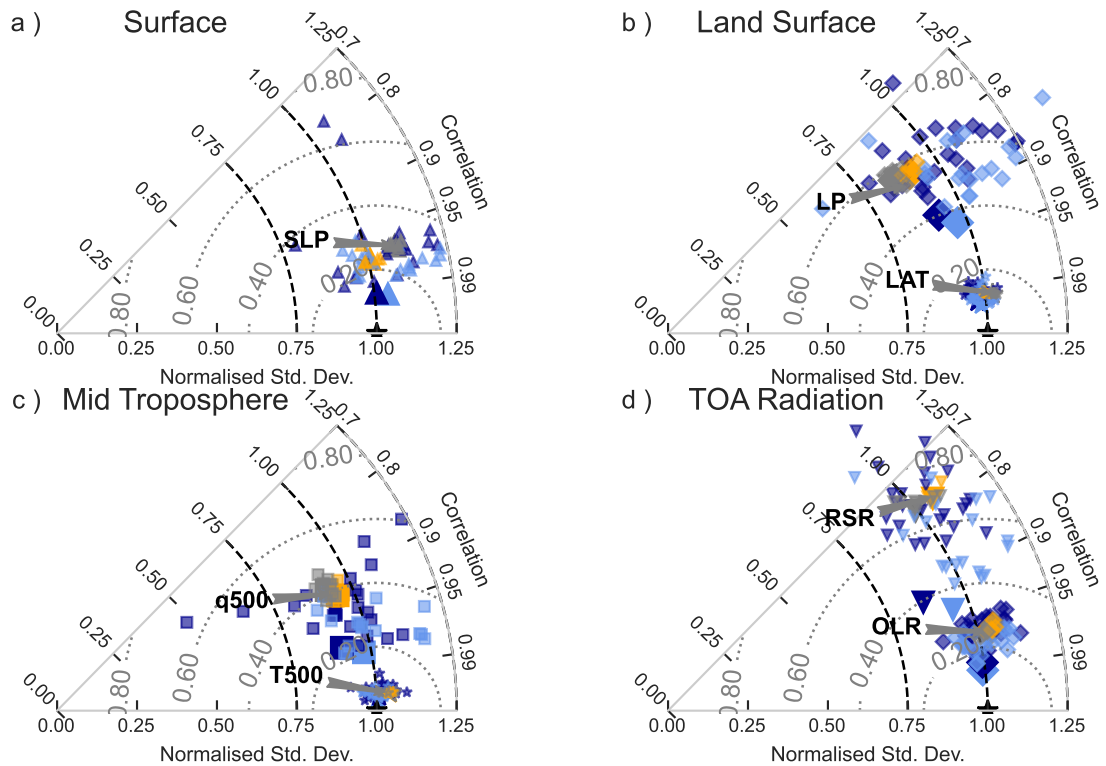
- 831 Williamson, D., M. Goldstein, L. Allison, A. Blaker, P. Challenor, L. Jackson, and K. Yamazaki,
832 2013: History matching for exploring and reducing climate model parameter space using ob-
833 servations and a large perturbed physics ensemble. *Climate Dynamics*, **41 (7)**, 1703–1729,
834 <https://doi.org/10.1007/s00382-013-1896-4>.
- 835 Yamazaki, K., D. M. H. Sexton, J. W. Rostron, C. F. McSweeney, J. M. Murphy, and G. R. Harris,
836 2021: A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections:
837 part 2: global performance and future changes. *Climate Dynamics*, **56 (11-12)**, 3437–3471,
838 <https://doi.org/10.1007/s00382-020-05608-5>.
- 839 Yamazaki, K., and Coauthors, 2013: Obtaining diverse behaviors in a climate model without the
840 use of flux adjustments. *JGR-Atmospheres*, **118**, 2781–2793, <https://doi.org/10.1002/jgrd.50304>.
- 841 Zelinka, M. D., T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein,
842 and K. E. Taylor, 2020: Causes of higher climate sensitivity in CMIP6 models. *Geophysical*
843 *Research Letters*, **47 (1)**, <https://doi.org/10.1029/2019gl085782>.



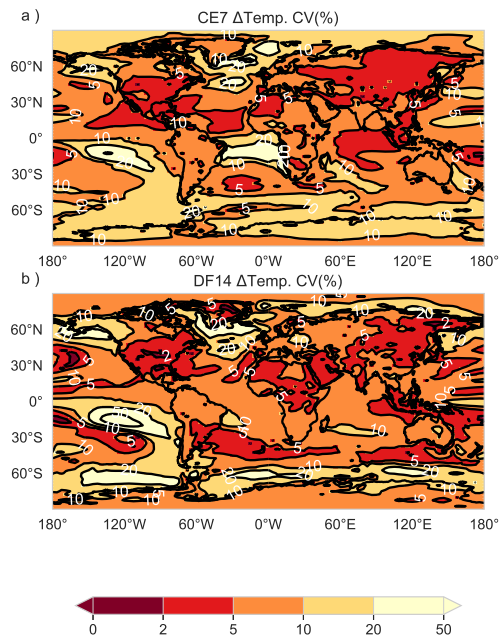
299 FIG. 2. Simulated values for CMIP5 (dark blue), CMIP6 (pale blue), CE7 (black) and DF14 (orange) ensembles.
 300 Also shown are sensitivity cases (SS; blue (optimised), red (unoptimised), and grey (standard configuration)
 301 boxes) described in table A1. Box and whiskers for CMIP5 and CMIP6 ensembles shows 25-75% range with
 302 whiskers extending from 5 to 95%. Stars show average value for ensemble. Y-axis in all sub-plots shows
 303 ensemble with the individual simulations having a small random offset added for presentation purposes. a: Cost
 304 values for atmosphere-only simulations. b: **Control** global average surface air temperature with vertical dashed
 305 line showing estimated observed 19th century temperature with grey shading its uncertainty range(Williamson
 306 et al. 2013). c) T140 and d) ECS4. Hexagons in c & d show ensemble average values for 2× TCR and 2×
 307 ECS. Black error bar centred on Standard HadCM3 model in b & c shows 2σ uncertainty range estimated from
 308 1000-year long control simulation while in d shows same from 7-member initial condition ensemble. Letters for
 309 CMIP5 (black) and CMIP6 (blue) correspond to different models defined in tables 2 and 3. Numbers for CE7
 310 and DF14 ensembles correspond to individual parameter settings (See Fig. 1 for parameter values).



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

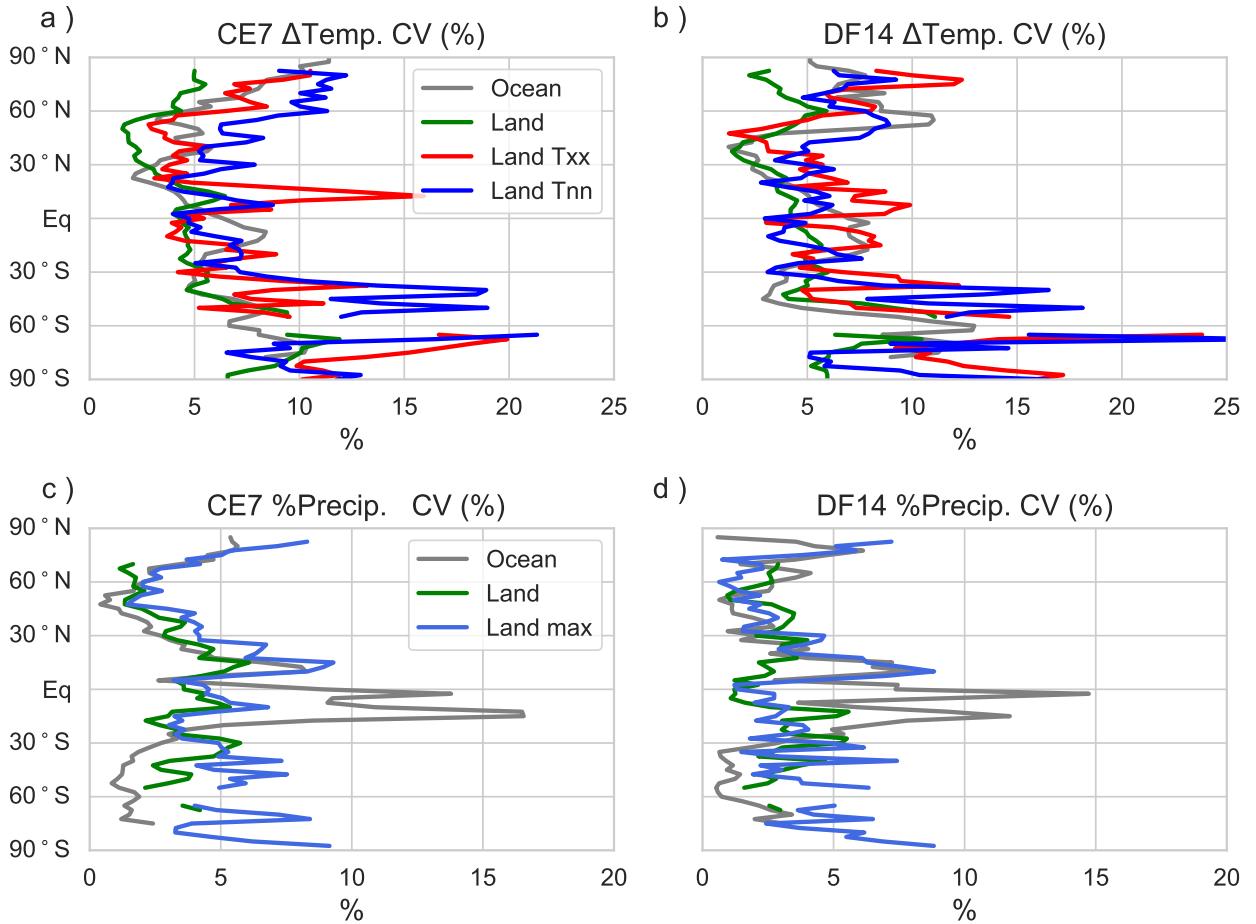


363 FIG. 4. Partial Taylor diagram for: a) sea level pressure (SLP; Triangles); b) land air temperature (LAT; stars) &
 364 precipitation (LP; diamonds); c) 500 hPa relative humidity (q500; squares) & temperature (T500; stars) d) TOA
 365 outgoing LW radiation (OLR; diamonds) & reflected SW radiation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

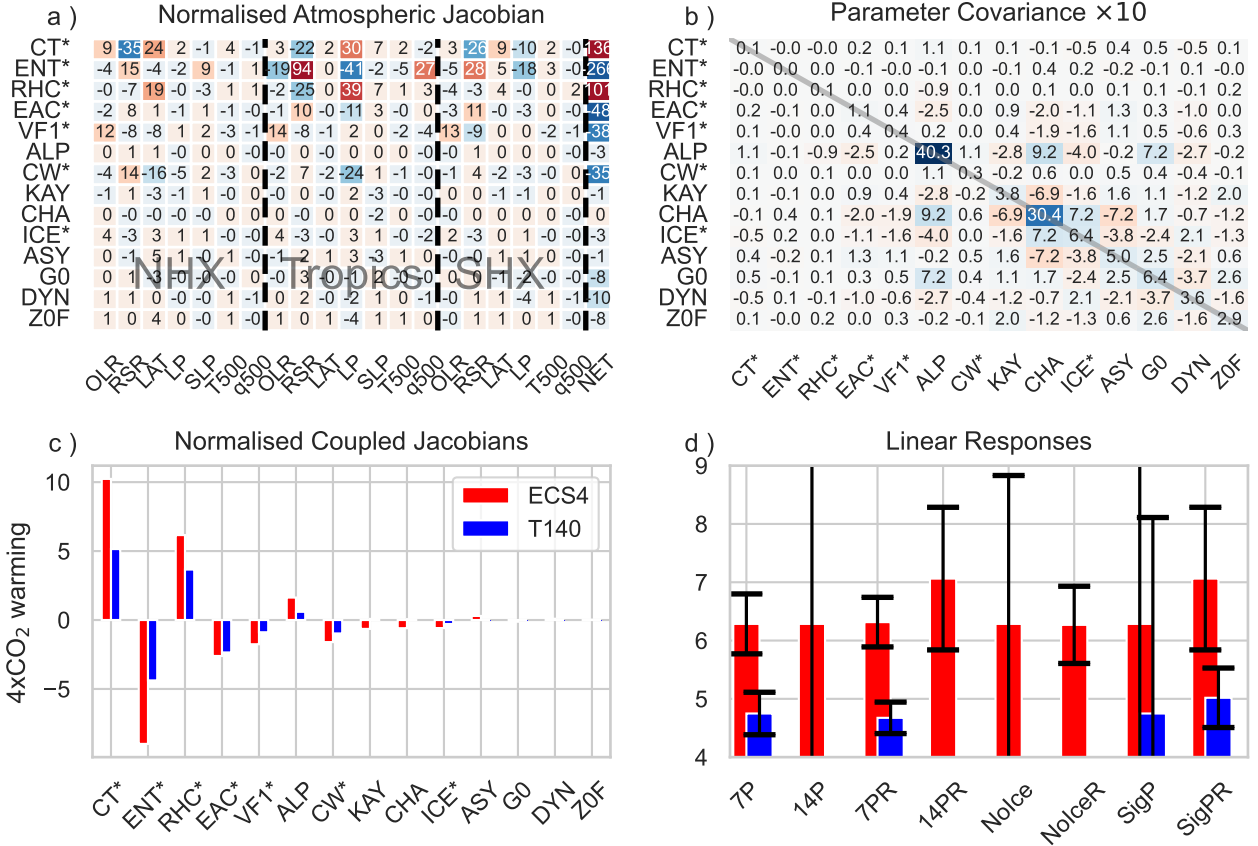


422 FIG. 5. Coefficient of Variation(%) for temperature change at $4 \times \text{CO}_2$ for CE7 (a) and DF (b) ensembles.

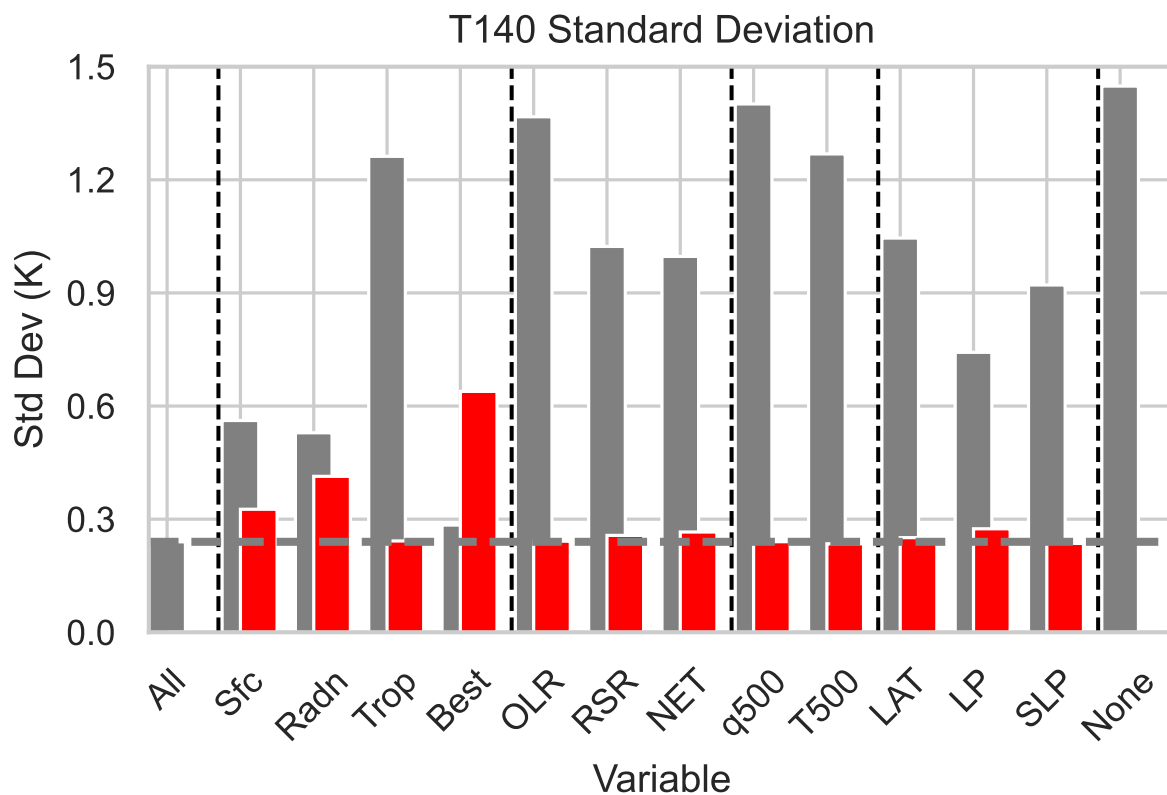
423 Colours and contours at 0, 2, 5, 10, 20 and 50%.



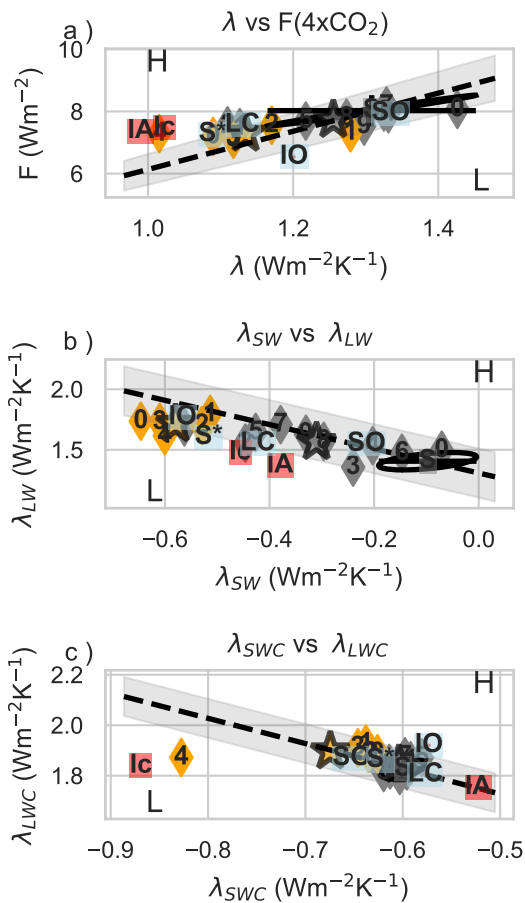
430 FIG. 6. Coefficient of Variation (%) of zonal-mean temperature change at $4\times\text{CO}_2$ for ocean (grey), land
 431 (green), land annual maximum (red) and land annual minimum (blue) for CE7(a) and DF14(b) ensembles. CV
 432 (%) for % change in ocean (grey), land (green) and annual maximum (dark blue) precipitation relative to **control**
 433 simulation for CE7 (c) and DF14(d) ensembles. Locations where the estimated control precipitation was less
 434 than 10^{-5} (10^{-4}) $\text{Kg m}^{-2}\text{s}^{-1}$ for land/ocean (annual maximum land) were ignored in the zonal-mean calculation.



457 FIG. 7. a) Atmospheric Jacobian for observations normalised by their uncertainty estimates (largely observa-
 458 tional uncertainty) where x-axis is simulated observation and y-axis is parameter. Vertical dashed lines divide up
 459 the Northern Hemisphere extra-tropics, Tropics and Southern Hemisphere extra-tropics and the global average
 460 net flux. b) 10 \times normalised parameter co-variance (see methods) after constraint applied. c) Jacobian for ECS4
 461 (Red bars) & T140 (blue bars) ordered by absolute ECS4. d) Estimated ECS4 & T140 (red & blue bars; y-axis)
 462 and $\pm 2\sigma$ (error-bars) for seven parameters (7P), fourteen parameters (14P), all parameters excluding ALPHAM
 463 (NoIce), and the ten parameters with $\geq \sigma$ impact on ECS4 (sigP) in K. R appended shows when the normalised
 464 parameter values are limited to (0, 1) (see methods). T140 values not shown for 14P, 14PR, NoIce, and NoIceR
 465 cases as TCR Jacobian not complete for all parameters. Both Jacobians are with respect to normalised parameter
 466 where 0 is minimum value and 1 is maximum and parameters use short names (Table 1). Parameters with *
 467 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
 499 variables, have their uncertainty increased by 100 times (“other”). This, in effect, means those observations do
 500 not constrain the parameters and T140. All is all variables, Sfc is LAT, LP and SLP, Radn is OLR, RSR and
 501 NET, while Trop is q500 and T500. Best is LP, RSR and Net and None is when all observational uncertainties
 502 are scaled. Red bars show standard deviations when only that variable, or group of variables, had its uncertainty
 503 increased by a factor of 100 (“leave-out”). Horizontal dashed line show value for All analysis while vertical
 504 dashed lines separate the variables that contribute to Sfc, Radn, and Trop groups.



639 FIG. A1. Scatter plots at $4\times\text{CO}_2$ for CE7 (orange) and DF14 (grey) calibrated ensembles, and sensitivity
 640 studies (blue/red boxes). Stars show ensemble means. a) Forcing ($F(4\times\text{CO}_2)$) vs climate feedback (λ); b)
 641 SW climate feedback (λ_{SW}) vs LW climate feedback (λ_{LW}); c) Clear sky SW climate feedback (λ_{SWC}) vs clear
 642 sky LW climate feedback (λ_{LWC}). Black ellipses are centred on the Standard HadCM3 configuration and shows
 643 2σ joint-uncertainty ellipse computed from initial condition ensemble while cross shows 2σ errors for x and y
 644 variables separately. Dashed lines show ECS4 (a), λ (b) and λ_C (c) fixed at standard values while grey region
 645 shows $\pm 2\sqrt{2}\sigma$ internal variability range around standard configuration for this parameter. H and L indicate which
 646 side of the dashed line where these values are higher or lower than standard model.