The impact of recent forcing and ocean heat uptake data on estimates of climate sensitivity

Nicholas Lewis[1], Judith Curry

Supporting Information

S1: Revisions to forcing estimates and their updating to 2016

Greenhouse gas forcing

The logarithmic relationship between carbon dioxide (CO_2) atmospheric concentration and radiative forcing implied by the simplified expression used in AR5 (8.SM.3) is an approximation, derived from Myhre et al. (1998). Recent research (Byrne and Goldblatt 2014; Etminan et al. 2016) confirms that CO_2 forcing grows slightly faster than logarithmically with concentration, and provides updated estimates of the CO_2 forcing-concentration relationship. Moreover, Etminan et al. (2016; hereafter Ea16) show that methane (CH_4) forcing is significantly understated by the simplified expression used in AR5. Their expressions for CO_2, CH_4 and nitrous oxide (N_2O) radiative forcing are adopted here. The Ea16 estimates for CO_2 forcing are preferred to those in Byrne and Goldblatt 2014, which are similar but relate to instantaneous radiative forcing at the tropopause (iRF) rather than to stratospherically-adjusted forcing. Radiative forcing in 2011 using the Ea16 expressions is 0.6% higher for CO_2, 25.3% higher for CH_4 and 1.2% higher for N_2O, relative to the AR5 values, resulting in total WMGG forcing being 4.8% higher; $F_{2\times CO2}$ increases by 2.5%.

Mean annual 1750–2011 atmospheric CO_2, CH_4 and N_2O concentrations are from AR5 Table A2.1.1a, linearly interpolating pre-1955 pentadal data. The AR5 Table A2.1.2 values for 'GHG Other' have been adjusted by the difference between CH_4 and N_2O forcing calculated using alternatively the Ea16 or the AR5 expressions.

Annual mean forcing from CO_2 and N_2O for 2012–2015 was calculated using mean atmospheric concentration data from WMO (2016), rebasing the time series to match AR5 concentrations in 2011. The 2015 CO_2 concentration was updated to 2016 using the change in concentration at Mauna Loa (ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_annmean_mlo.txt). The 2015 N_2O concentration was updated to 2016 using the change in annual-mean concentration

[1] Corresponding Author: Nicholas Lewis, Bath, United Kingdom. Email: nhlewis@btinternet.com

per NOAA ERSL (ftp://ftp.cmdl.noaa.gov/hats/n2o/combined/HATS_global_N2O.txt). As the AR5 CH_4 concentrations match the NOAA ERSL values they are updated using that source (ftp://aftp.cmdl.noaa.gov/products/trends/ch4/ch4_annmean_gl.txt). Forcing from halogenated WMGGs over 2012–2016 was increased from the AR5 2011 level at a rate of 0.0145 $Wm^{-2}decade^{-1}$, the almost linear 1995–2015 growth rate per Figure 1 of WMO (2016).

Aerosol and tropospheric ozone forcing

Myhre et al (2017) revised estimated changes in aerosol and tropospheric ozone forcing over 1990–2015 using recently updated emissions data, allowing for changes in the geographical distribution of emissions. Best estimates for aerosol and ozone forcing over 1991–2015 are derived by adding their post-1990 multi-model mean changes to the AR5 1990 values. Values for 2016 are estimated by extrapolation, using the small 2011–2015 trends. Forcing changes are more positive over 1990–2011 than estimated in AR5 (marginally so for ozone). Estimated total aerosol forcing is now strongest in 1995 at −0.88 Wm^{-2}, falling to −0.78 Wm^{-2} by 2011, thereafter remaining almost stable. An allowance for uncertainty in the Myhre et al (2017) estimate for combined aerosol and ozone forcing revisions is added, randomly sampled, to the sampled ΔF values before ECS and TCR values are calculated. Being independent of other, much larger, uncertainty in ΔF, it barely increases the total uncertainty.

Stratospheric ozone forcing

A linear regression fit between KNMI MSR ozone (from KNMI Climate Explorer: https://climexp.knmi.nl) and AR5 1979–2011 stratospheric ozone forcing (slope 0.0016 Wm^{-2} Dobson $unit^{-1}$; R^2=0.63) is used to estimate post-2011 forcing changes (all < 0.01 Wm^{-2}).

Land use change, Stratospheric H_2O, Black carbon on snow & ice (BC Snow) and Contrails forcings

These minor forcings are estimated by varying them from 2011 to 2012–2016 in line with their linear trends over 2002–2011; none of them changes by more than 0.01 Wm^{-2}. BC Snow forcing is efficacy-adjusted using the AR5 2–4 efficacy range (Myhre et al. 2014, p.685), taken as 5–95%.

Solar forcing

Annually-averaged SORCE solar Total Solar Irradiance data (Kopp et al 2005; http://lasp.colorado.edu/lisird/data/sorce_tsi_24hr_l3/) was scaled as set out in AR5 Table 8:SM8. The resulting 2011–2016 values were rebased so that the 2011 forcing agreed to AR5.

Volcanic forcing

For 2012, forcing was taken as 25× the mean monthly global aerosol optical depth (AOD) values updated from Sato et al (1993), as in AR5. For 2013, 25× an update (Ridley et al 2014) of Vernier

et al. (2011) AOD data was used. For both years the estimated volcanic forcing was −0.092 Wm^{-2}. For 2014, volcanic forcing was estimated (as −0.10 Wm^{-2}) by integrating each year's WCRP CMIP6 (Eyring et al. 2016) extinction coefficients (ftp://iacftp.ethz.ch/pub_read/luo/CMIP6/CMIP_1850_2014_extinction_550nm_strat_only.nc) for 12.5+ km altitudes, scaling to match Sato 1998–2012 forcing changes and adjusting the 2012 Sato forcing by the 2012–2014 change in scaled integrated AOD. Volcanic forcing for 2015 and 2016 was set at the 2014 level, as there had been no significant change in volcanic activity since 2014.

Andersson et al. (2015) argued that radiative forcing from volcanic aerosols in the lowermost stratosphere was not included in the satellite-based AOD data from which the AR5 estimates were derived, and hence was underestimated in AR5. They found that including lowermost stratosphere aerosols would have substantially increased the AR5 0.11 Wm^{-2} estimate of mean 2008–2011 volcanic forcing. However, their Figure 4d showed a background level of volcanic forcing (applying over most of 2010) of −0.19 Wm^{-2}. The early-industrial background level is unlikely to have been significantly smaller. Although some lowermost stratosphere sulfate aerosols have been transported from the upper troposphere, the anthropogenic component of tropospheric sulfate aerosol has been estimated at 13% or less (Charlson et al. 1992; Stevens 2015). A conservative estimate of the early-industrial background level would accordingly be −0.15 Wm^{-2}. By contrast, during 1850–1900 the AR5 volcanic forcing, which is a multiple of AOD without any background level deducted, was no stronger than −0.05 Wm^{-2} in 20 years and zero in some of them. If AR5 volcanic forcing, as updated to 2016, were scaled in the ratio of mean volcanic forcing (without the background level deducted) over 2008–2011 estimated from Andersson (2015) to the AR5 estimate, but constrained to be no more positive in any year than −0.15 Wm^{-2}, the difference in mean $ERF_{Volcano}$ between 1869–1882 and 2007–2016 would change by only −0.01 Wm^{-2}, and would also be very small for the other period combinations. Therefore the AR5 $ERF_{Volcano}$ estimates have been retained, and updated on the same basis post 2011.

S2: Updating of heat uptake estimates and revisions to ocean heat content (OHC) components

The best estimates and uncertainty ranges for minor, non-ocean heat content components in AR5 Box 3.1, Figure 1 are extended to 2016 as follows. Land and ice heat content estimates are extrapolated at the 2000–2011 linear trend rate, scaling 2011 uncertainties pro rata. Atmospheric heat content is derived from the linear fit of 1992–2011 AR5 values to surface temperature per HadCRUT4v5; uncertainty is set at the 2007–2011 average. The sub-2000 m OHC element and uncertainty therein included in the AR5 below-700 m component (Purkey and Johnson 2010)

represent a fixed heat uptake rate of 0.068 ± 0.031 Wm^{-2} (standard error), starting in 1992. As Desbruyeres et al (2016) obtained almost identical results over 2003–2012, the AR5 1992–2011 values are extrapolated to 2012–2016.

AR5 took 0–700 m and 700–2000 m estimates from updates of respectively Domingues et al. (2008) and (by differencing its 0–2000 m and 0–700 m estimates) Levitus et al. (2012). Only the latter dataset has been updated to 2016 (https://data.nodc.noaa.gov/woa/DATA_ANALYSIS/3M_HEAT_CONTENT/DATA/basin/). However, Ishii and Kimoto (2009) 0–700 m OHC estimates are available to 2016 (http://www.data.jma.go.jp/gmd/kaiyou/english/ohc/ohc_global_en.html), as are new estimates of both 0–700 m and 700–2000 m OHC from Cheng et al. (2017). Their approach achieves annual estimates with near-global coverage earlier than the Levitus dataset, by more extensive spatial infilling. The Levitus dataset instead increases coverage by combining observations during pentads, but also provides estimates for single years. By the last pentadal estimate year (2014) their annual OHC estimates had become almost identical to the pentadal estimates, and so were used directly to extend those to 2016. In order to reduce sampling fluctuations and other sources of uncertainty associated with these two different approaches, 700–2000 m OHC is determined as the average of triadal mean estimates (up to 2015; 2016 is annual) from the Cheng dataset and pentadal estimates (up to 2014, thereafter annual) from the Levitus dataset. The same strategy is used for 0–700 m OHC except that the average includes triadal mean estimates from the Ishii and Kimoto dataset, which only covers this layer. The individual uncertainty estimates for the three 0–700 m OHC datasets differ greatly and, prior to the Argo era, appear inadequate given the differences between their estimates. Therefore, the seemingly more realistic AR5 0–700 m OHC uncertainty estimates, which are much higher before the Argo era, are used for the averaged 0–700 m OHC values. As these AR5 uncertainty estimates change little after 2005 the 2011 uncertainty value is used over 2012–2016. Much of the estimation error within the 700–2000 m layer is likely common between the datasets used, so the uncertainty estimate for 700–2000 m OHC is taken as the average of those per Cheng et al and those derived for the Levitus estimate by differencing in quadrature the 0–2000 m and 0–700 m uncertainty estimates.

To allow for coverage of the infilled 0–700 m and 700–2000 m OHC data not being quite globally complete, the averaged OHC estimates for those layers are multiplied by 1.047, the factor used by Cheng et al (2017).

The average rate of heat uptake during the final period is estimated from the annualized difference in accumulated energy between the start and end years in each period, with all start and

end year uncertainties in OHC being added in quadrature and divided by the relevant number of years. Since uncertainty for the sub-2000 m ocean is given in terms of ocean heat uptake not OHC, it is used directly (scaled down when a pre-1992 period is included). Heat uptake estimates based on regression slope are marginally higher – as is expected since 2016 experienced an El Niño peak, which increases surface temperature but reduces OHC. The inclusion of the Ishii and Kimoto data has little effect on heat uptake estimates for any of the final periods. The effect on the heat uptake estimate of using either only the Cheng dataset or only the Levitus dataset is slight, its sign depending on the final period involved.

S3: Re-estimation of the efficacy of historical forcing using GISS-E2-R simulation data

Marvel et al. 2016 (M16) computed (transient) efficacy for ERF historical forcing by comparing the mean ΔT simulated by GISS-E2-R over 1996–2005 with ERF in year 2000, the only year for which it was estimated (both relative to in preindustrial, 1850, conditions). This is not an apples-to-apples comparison, as forcing in 2000 was 5.2% higher than the 1996–2005 mean (Table S1). Moreover, averaging over 1996–2005 is unsatisfactory because the simulated temperatures in the second half of the 1990s are depressed by continuing recovery from the 1991–92 eruption of Mount Pinatubo. The depression of simulated temperatures during that period can be seen by comparing the 0.311 ratio of ensemble-mean ΔT to iRF historical ΔF (Table S1) over 1996–1999 with its mean value of 0.342 over the longer 1985–2005 period since the previous large eruption. While the mean ΔT to iRF historical ΔF ratio over 2000–2005 reverted to within 1% of the 1985-2005 mean ratio, over 1996–2005 it was 4% below it.

Downwards bias in estimated ERF efficacy due both to use of high year-2000 forcing and to the depression of 1996–1999 simulated temperatures by recovery from volcanism can be avoided by comparing the mean 2000–2005 ΔT with estimated mean 2000–2005 ERF ΔF, derived by scaling the year 2000 historical ERF estimate of 2.83 Wm^{-2} (source: https://data.giss.nasa.gov/modelforce/Fs_2000_GISS-E2.txt) by the ratio of historical mean 2000–2005 iRF to year 2000 historical iRF, giving a 2000–2005 mean ERF ΔF of 2.80 Wm^{-2}.

TCR estimated from historical ERF (as $F_{2\times CO2}\ \Delta T / \Delta F$), using the ERF $F_{2\times CO2}$ value of 4.35 Wm^{-2} (source: https://data.giss.nasa.gov/modelforce/Fs_2000_GISS-E2.txt) and mean 2000–2005 historical ERF ΔF and ΔT, is then 4.35 * 0.90 / 2.80 = 1.40 K. This value is identical to the TCR estimate of 1.4 K derived from the 1pctCO2 simulation, so the apples-to-apples estimate of the (transient) efficacy of historical ERF is exactly one.

Year	ΔT [*tas*] (K)	ΔF [iRF] (Wm^{-2})	$\Delta T / \Delta F$ (K/Wm^{-2})
1996	0.663	2.215	0.299
1997	0.696	2.307	0.302
1998	0.789	2.468	0.320
1999	0.827	2.591	0.319
2000	0.867	2.684	0.323
2001	0.879	2.708	0.325
2002	0.893	2.716	0.329
2003	0.906	2.575	0.352
2004	0.941	2.691	0.350
2005	0.922	2.570	0.359
1996-2005 mean	0.838	2.552	0.328
1996-1999 mean	0.744	2.395	0.311
2000-2005 mean	0.901	2.657	0.339
1985-2005 mean	0.686	2.006	0.342

Table S1: GISS-E2-R ensemble-mean surface temperature and iRF forcing near the end of the historical simulation, as used in M16. Source: https://data.giss.nasa.gov/modelforce/tas.Marvel_etal2015.csv; and https://data.giss.nasa.gov/modelforce/Fi_net_Miller_et_al14_upd.txt.

For M16's ERF 'equilibrium efficacy', which uses Eq. (4), it should not matter whether averaging is over 1996–2005 or 2000–2005, as heat uptake should be compensatingly higher during recovery from a volcanic eruption. However, M16 use the rate of ocean heat content change (their ΔQ) as a substitute for TOA radiative imbalance (ΔN), which in GISS-E2-R understates it by >15% over the relevant periods. Fair efficacy estimation requires use of the full TOA imbalance, as when estimating CO_2-forced ECS. Using the average ensemble-mean historical TOA radiative imbalance, relative to preindustrial, of 0.99 Wm^{-2} over 2000–2005 produces an historical ERF derived ECS estimate of 4.35 * 0.90 / (2.80 − 0.99) = 2.16 K. Performing the calculation using instead 1996–2005 averages gives an identical ECS estimate: 4.35 * 0.84 / (2.69 − 1.01) = 2.16 K. ECS estimated from CO_2 forcing with a similar time-profile to that of total historical forcing must be used as the

denominator in the efficacy calculation to obtain an apples-to-apples comparison, not (as in M16) an estimate of the model's true equilibrium climate sensitivity. Relative to the model's 2.07 K ECS estimate derived from means centered on year 100 of its 1pctCO2 simulation (Table S2), an appropriate comparative, the equilibrium efficacy estimate is 1.04.

S4: Estimation of sensitivity inferable from historical data (ECS_{hist}) and ECS in CMIP5 models

We investigate here the relationship between estimates of ECS_{hist} and ECS in CMIP5 AOGCMs. We include all models in their various versions that had archived data for preindustrial control (piControl), 1% yr^{-1} CO_2 ramping (1pctCO2) and abrupt4xCO2 simulations, save that we exclude FGOALS-g2 as its 1pctCO2 simulation results are abnormal and the p2 variants of GISS-E2-H and GISS-E2-R as their model physics is intermediate between the main (p1) and p3 physics versions. We do not use data from historical simulations as the model forcings are unknown and cannot be accurately estimated. We derive annual global mean N (*rsdt* − *rsut* − *rlut*) and T (*tas*) anomalies ΔN and ΔT by deducting the corresponding portions of linear fits for N and T over the full piControl simulation; fits over shorter periods are not recommended (Gupta et al. 2013).

As methods of estimating ECS_{hist} and ECS do not all involve the same CO_2 concentrations, we allow for the non-logarithmic aspect of CO_2 forcing (Etminan et al.2016) when carrying out all calculations, so that all the ECS_{hist} and ECS estimates reflect a doubling of CO_2 concentration. AOGCMs use simplified radiative transfer calculations and the forcing-concentration relationship for CO_2 will vary between models, but there is no reason to think that any systematic bias exists in CMIP5 models as regards the non-logarithmic element.

We use three methods of estimating ECS_{hist} in CMIP5 AOGCMs. The first method follows Armour (2017, hereafter A17) in applying Eq. (4) using mean N and T over years 85–115 of the 1pctCO2 simulations. He found that taking an average around year 100 provided the estimate most closely corresponding to estimating ECS_{hist} from historical forcing. However, we estimate $F_{2\times CO2}$ from the y-axis intercept when regressing ΔN against ΔT over years 2–10 of the abrupt4xCO2 simulations, whereas A17 regresses over years 1–5. Since the ERF metric is used for observationally-based ECS_{hist} estimates, it is not appropriate to let the mean N value in the first year of the abrupt4xCO2 simulation, when N is adjusting to the imposed abrupt forcing, significantly influence the $F_{2\times CO2}$ estimate – which it does when regressing over years 1–5. The ensemble-mean $F_{2\times CO2}$ value from regressing over years 2–10 of the abrupt4xCO2 simulations is 3.80 Wm^{-2} before

dividing by 1.046 to adjust for the non-logarithmic aspect of CO_2 forcing, and 3.63 Wm^{-2} after doing so.

The second method of estimating ECS_{hist} is the same as the first save that, instead of using actual 1pctCO2 simulation ΔN and ΔT values, it uses step-emulated versions thereof (Good et al. 2011) derived from abrupt4xCO2 N and T data. Step-emulation provides close matches to CMIP5 model 1pctCO2 simulation responses (Figure 1), particularly for T (Caldeira and Myhrvold 2013), if deviations of the CO_2 concentration-forcing relationship from a logarithmic form are incorporated, although it is less accurate if they are not (Gregory et al. 2015). Moreover, step-emulation accesses different realisations of internal variability and provides higher signal-to-noise ratios than 1pctCO2 simulation data.

The third method of estimating ECS_{hist} is to regress ΔN against ΔT over the first 50 years of the abrupt4xCO2 simulations (excluding year one, on the same grounds as for $F_{2\times CO2}$ estimation), with ECS_{hist} determined by scaling the x-axis intercept. Although abrupt4xCO2 simulations involve a less realistic forcing time-profile than 1pctCO2 simulations, the weighted-average age of forcing in year 50 of abrupt4xCO2 simulations is the same as in year 100 of 1pctCO2 simulations. This method has the advantage that no separate estimate of each model's $F_{2\times CO2}$ value is required, moreover the realisation of internal variability differs from that in the step-emulation method as it involves the first 50 rather than 115 years of the abrupt4xCO2 simulation.

We use, for each model, the mean of the values given by these three methods as the ECS_{hist} best estimate and their standard deviation as an indicator of uncertainty in it. The ensemble-mean ECS_{hist} estimates using the three methods are identical within ±0.01 K, and their ensemble-mean within-model standard deviation is 0.09 K. That indicates that the mean of the values given by the three methods provides fairly accurate ECS_{hist} estimates for CMIP5 models, with little uncertainty.

We follow Andrews et al. (2015) and A17 in estimating CMIP5 models' ECS by regressing annual mean ΔN against ΔT over years 21–150 of their abrupt4xCO2 simulations. Doing so is more justifiable than regressing over years 1–150 (Forster et al. 2013; Flato et al. 2014) where λ changes during the early decades of those simulations, and generally produces higher ECS estimates. How accurate an estimate of true ECS regressing over years 21–150 yields will depend on how similar each model's net climate feedback after year 150 is to that over years 21–150, which in most cases is not known. However, there is in any event an issue regarding the regression method. Both Andrews et al. (2015) and A17 used ordinary least squares (OLS) regression, which is only valid when there is no error noise in the regressor variable (ΔT). The presence of noise in the regressor variable

causes OLS regression to underestimate the true slope of the relationship between the variables. As the regressor variable ΔT is affected by internal variability, albeit to a lesser extent than ΔN, OLS regression slope estimates will be biased downwards, and hence ECS estimates biased upwards. We employ instead Deming regression, which given valid estimates of the error standard deviation in ΔN and ΔT will provide unbiased estimates of the regression line and hence of ECS. We use the standard deviation of detrended annual data over years 76–150 of the relevant model's abrupt4xCO2 simulation to give the required error standard deviations; using variability in detrended N and T over piControl simulations gives almost identical results. On average the resulting reduction in the ECS estimate is small (2%) compared to that from OLS regression, but for inmcm4 it is approaching 10%. The effect is, however, negligible for regressions starting from year one or two, as they involve a larger ΔT range.

The three calculated ECS_{hist} values for each model, along with their mean and standard deviation, the estimated ECS value and the ratio of the ECS estimate to the mean ECS_{hist} estimate are given in Table S2, along with their ensemble-mean and ensemble-median values. If each modeling center were given equal weight,[2] with the ACCESS1 models treated as being from the same center as HadGEM2-ES, the ensemble median ECS-to-ECS_{hist} ratio would be lower, at 1.082.

When comparing the ECS estimates in Table S2 with those published elsewhere that are based on simply dividing by two estimates derived from abrupt4xCO2 simulation data, the Table S2 values need to be increased by the factor of 1.046 used to adjust for the non-logarithmic element of the CO_2 forcing-concentration relationship between a doubling and a quadrupling of its initial concentration. The resulting median and mean ECS estimates for the ensemble of 31 CMIP5 models would be respectively 3.37 K and 3.46 K.

TCR estimates are also given in Table S2. They are the means of surface temperature warming at the time of CO_2 concentration doubling in the 1pctCO2 simulations (averaged over 20 years) – the formal definition of TCR – and warming at the same point when CO_2 concentration increasing at 1% per annum is step-emulated using the same model's response in its abrupt4xCO2 simulation, which accesses a different realization of internal variability. The two methods produce almost identical ensemble-mean TCR estimates. Substituting AR5 Table 9.5 TCR estimates, for models where they are available, would leave the ensemble-mean and -median values unchanged.

[2] By counting models from all centers except GISS, which has the largest number of models (four) included in the ensemble, twice (four) times where a modeling center has two (one) model(s) included, and for centers with three models included counting the model with the middle ECS-to-ECS_{hist} ratio twice.

CMIP5 model	TCR estimate (mean)	ECS_{hist} estimates by 3 methods			Mean of three ECS_{hist} estimates	St dev of three ECS_{hist} estimates	ECS estimate: Deming regression	ECS estimate/ mean ECS_{hist} estimate	ECS/ mean ECS_{hist} estimate if $ECS_{hist} \leq 2.85$
		1pctCO2: energy-budget	1pctCO2 EB: step-emulated	abrupt4x CO2 regression					
ACCESS1-0	1.91	3.28	3.07	3.13	3.16	0.11	4.20	1.33	
ACCESS1-3	1.68	2.89	2.88	2.90	2.89	0.01	4.18	1.45	
bcc-csm1-1	1.79	2.63	2.60	2.59	2.60	0.02	2.76	1.06	1.06
bcc-csm1-1-m	2.01	2.66	2.57	2.59	2.61	0.05	2.82	1.08	1.08
BNU-ESM	2.53	4.13	3.85	3.90	3.96	0.15	3.68	0.93	
CanESM2	2.27	3.36	3.34	3.35	3.35	0.01	3.67	1.09	
CCSM4	1.71	2.59	2.65	2.58	2.61	0.04	2.99	1.15	1.15
CESM1-CAM5-1-FV2	2.11	3.38	3.44	3.23	3.35	0.11	3.22	0.96	
CNRM-CM5	2.04	3.14	3.16	3.21	3.17	0.03	3.03	0.96	
CNRM-CM5-2	1.91	3.24	3.44	3.20	3.29	0.13	3.54	1.07	
CSIRO-Mk3-6-0	1.77	3.08	3.24	3.21	3.18	0.08	4.83	1.52	
FGOALS-s2	2.31	3.99	4.21	3.99	4.06	0.13	4.36	1.07	
GFDL-CM3	1.98	3.59	3.42	3.53	3.51	0.08	3.96	1.13	
GFDL-ESM2G	1.37	1.96	2.18	2.08	2.07	0.11	2.41	1.16	1.16
GFDL-ESM2M	1.47	2.29	2.28	2.26	2.28	0.02	2.42	1.06	1.06
GISS-E2-H	1.63	2.51	2.17	2.19	2.29	0.19	2.31	1.01	1.02
GISS-E2-H-p3	1.70	2.51	2.25	2.29	2.35	0.14	2.47	1.05	1.05
GISS-E2-R	1.39	2.07	1.85	1.89	1.94	0.12	2.17	1.12	1.12
GISS-E2-R-p3	1.48	2.46	2.15	2.19	2.27	0.17	2.35	1.04	1.04
HadGEM2-ES	2.42	4.24	3.96	3.94	4.04	0.17	5.27	1.30	
inmcm4	1.31	2.07	1.93	2.01	2.00	0.07	1.82	0.91	0.91
IPSL-CM5A-LR	2.05	3.77	3.76	3.79	3.77	0.02	3.99	1.06	
IPSL-CM5A-MR	2.08	3.63	3.80	3.82	3.75	0.10	4.22	1.13	
IPSL-CM5B-LR	1.54	2.25	2.24	2.34	2.28	0.06	2.54	1.12	1.12
MIROC-ESM	2.35	3.62	4.18	4.37	4.06	0.39	4.75	1.17	
MIROC5	1.62	2.30	2.55	2.52	2.45	0.13	2.54	1.03	1.03
MPI-ESM-LR	2.12	3.32	3.36	3.23	3.30	0.07	3.67	1.11	
MPI-ESM-MR	2.10	3.15	3.20	3.15	3.17	0.03	3.49	1.10	
MPI-ESM-P	2.09	3.02	3.14	3.08	3.08	0.06	3.45	1.12	
MRI-CGCM3	1.60	2.39	2.35	2.46	2.40	0.05	2.53	1.06	1.06
NorESM1-M	1.48	2.38	2.46	2.43	2.42	0.04	3.03	1.25	1.25
Median	1.91	3.02	3.07	3.08	3.08	0.08	3.22	1.09	1.06
Mean	1.86	2.97	2.96	2.95	2.96	0.09	3.31	1.12	1.08

Table S2. ECS_{hist} estimates using various methods and comparison, for each model in the ensemble, of their means with the estimated model ECS. TCR estimates are also given. All estimates, except ratios, are in K and are scaled to 2x preindustrial CO_2 forcing, using the Etminan et al. 2016 expression.

S5: Reconciliation of our CMIP5 ECS-to-ECS_{hist} ratio to that of Proistosescu and Huybers

Proistosescu and Huybers (2017, hereafter PH17) step-emulate CMIP5 AOGCM responses to AR5 historical forcing using eigenmode-decomposition models with three exponentially-damped terms having differing time-constants, fitted to the AOGCMs' abrupt4xCO2 simulation responses. They thereby derive ECS_{hist} values (which they term ICS) for each AOGCM, and calculate ECS and $F_{2\times CO2}$ values from the eigenmode-decomposition fits. However, they use inconsistent $F_{2\times CO2}$ values within their calculations, and their $F_{2\times CO2}$ values are affected by forcing adjusting from its instantaneous TOA (IRF) value towards ERF. Moreover, they use an initial forcing level that is inconsistent with the assumed initial equilibrium. For these reasons their ensemble-mean ECS-to-ICS ratio of 1.34 has a substantial upwards bias. A more detailed explanation of these issues is as follows.

PH17 estimate ICS values for CMIP5 AOGCMs by applying an energy-budget approach to their emulated AOGCM ΔT and ΔN responses over 1750–2011 when forced by a time-series of historical forcing *F*. For each CMIP5 AOGCM, they scale *F* by the ratio of $F_{2\times CO2}$ they derive for that AOGCM (from fitting their eigenmode-decomposition model) to the AR5 estimate of 3.7 Wm^{-2}. Using 2011 values for the scaled *F* and emulated ΔT and ΔN (their *H*), they then apply Eq. (4) and compute $\lambda = (\Delta F - \Delta N)/\Delta T$ and ICS = 3.7/λ. Figure S1 illustrates their methodology[3] for a canonical AOGCM representing PH17's ensemble-median case.[4] PH17's methodology is unsatisfactory in three respects, each of which introduces low biases into their AOGCM-based ICS values:

1. PH17 use inconsistent $F_{2\times CO2}$ values. Observational estimates use the same $F_{2\times CO2}$ value (usually 3.7 Wm^{-2}) to calculate the dominant CO_2 forcing component of *F* and to derive ICS from λ. They thereby provide ICS values correctly corresponding to a doubling of CO_2 even if the $F_{2\times CO2}$ value is inaccurate (ignoring, as PH17 does, inaccuracy arising from the non-CO_2 component of historical forcing). PH17 use different $F_{2\times CO2}$ values for these two purposes. The λ produced by their method reflects the emulated ensemble-median-based AOGCM implied $F_{2\times}$ of 4.03 Wm^{-2}. When correctly estimated by dividing λ into an $F_{2\times CO2}$

[3] For ease of illustration, prior to scaling by the GCM's $F_{2\times CO2}$/3.7, historical forcing and the responses thereto have been increased in the ratio of 3.7 Wm^{-2} to 2011 historical forcing. This leaves the slope estimates (being $-\lambda$) unchanged. They term the TOA radiative imbalance *H* instead of *N*.

[4] The ensemble-median 3-eigenmode decomposition τ_i, α_i, λ_i and ECS parameters are used, giving a derived $F_{2\times CO2}$ of 4.03 Wm^{-2}.

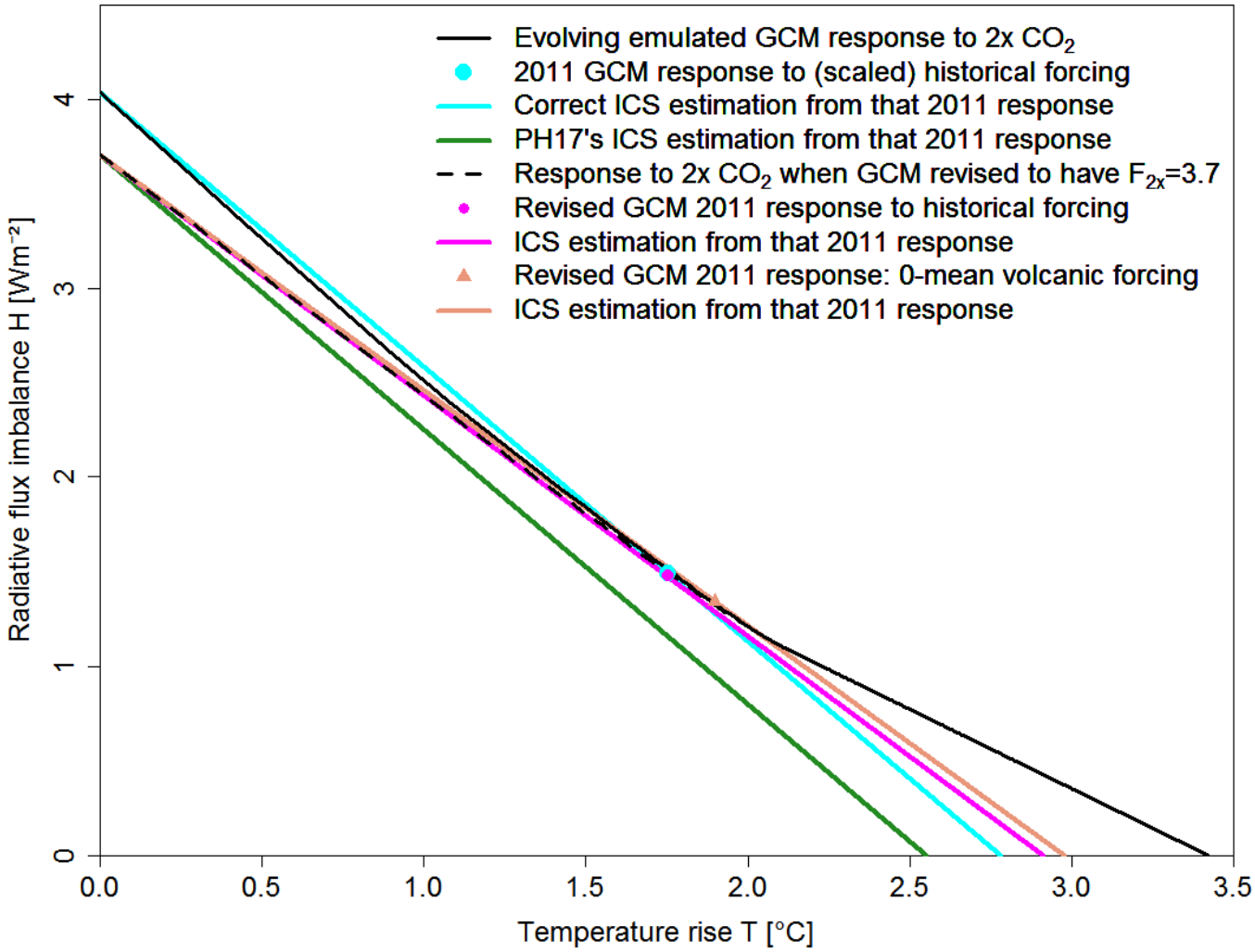

Figure S1. Biases in PH17's methodology for estimation of ICS (our ECS_{hist}). The solid black line shows the emulated evolving T and H (our N) response of a canonical AOGCM, matching their ensemble-median eigenmode-decomposition parameters, to an abrupt initial doubling of CO_2 concentration. The cyan circle represents the emulated response in 2011 when scaled[3] estimated evolving historical forcing is applied. The cyan line shows how ICS should be estimated from that 2011 response. The green line parallel to the cyan line shows how PH17 estimated ICS. The dashed black line shows how the AOGCM's response to doubled CO_2 changes when its parameters are revised to produce a more appropriate measure of its $F_{2\times}$ (3.7 Wm^{-2}). The magenta circle shows the resulting response to historical forcing in 2011, and the magenta line shows ICS estimated from that response. The orange triangle shows the response in 2011 to historical forcing with volcanic forcing re-centered to zero-mean over 1750–2011, using the same AOCGM parameters as for the magenta circle; the orange line shows ICS estimated from that response. Estimated $F_{2\times}$ and ICS/ECS values are respectively y- and x-axis intercepts, corresponding to time zero and to when equilibrium is reached.

value of 4.03 Wm^{-2}, ICS is 2.77 K (Figure S1, cyan line). However, PH17 estimate ICS by dividing λ into an $F_{2\times CO2}$ of 3.7 Wm^{-2} (green line), under-representing the AOGCM's response to a doubling of CO_2 and giving an ICS of 2.55 K – within 0.01K of PH17's ensemble-median estimate. [5]

2. While PH17's historical forcings are ERF estimates, PH17's fitted eigenmode decomposition $F_{2\times CO2}$ value represents something between IRF and ERF. Mean year one *N* of the abrupt4xCO2 simulations, which is significantly affected by stratospheric, tropospheric and other short-timescale adjustments being incomplete and should not be used in estimation of ERF, dominates fitting of the fastest component of their 3-eigenmode decomposition's $F_{2\times CO2}$ value. This effect is evident from PH17 Fig.1, where the posterior median curve is steeper to the left of the second annual average point. Estimating $F_{2\times CO2}$ ERF instead by regressing fitted *N* against *T* over years 2-10 (or 2-5) for all sample eigenmode-decomposition fits PH17 obtains to abrupt4xCO2 simulations gives an ensemble-median $F_{2\times CO2}$ estimate of 3.7 Wm^{-2}, in line with that estimated in AR5 and 8% below the 4.03 Wm^{-2} implied by the ensemble-median eigenmode-decomposition parameters. Using $F_{2\times CO2}$ = 3.7 Wm^{-2}, which reduces the curvature of the *N* versus *T* relationship (Figure S1, black lines: dashed versus solid), we estimate ICS for the ensemble-median-based AOGCM to be 2.91 K (magenta line).[6]

3. PH17's method assumes a start in 1750 from equilibrium, with 1750 forcing of zero. This is unrealistic, since zero volcanic forcing (as in 1750) is 0.40 W m^{-2} more positive than the 1750–2011 mean. Multi-centennial volcanic forcing estimates give near-identical means over 1250–1750 and 1750–2011 (Crowley and Unterman 2013). To satisfy the initial equilibrium assumption, historical volcanic forcing needs to be adjusted by +0.40 W m^{-2}, giving a zero long-term mean. Doing so (Figure S1, orange line) increases the revised ensemble-median-based AOGCM ICS estimate from 2.91 K to 2.98 K.

---

[5] When continuing the integration until equilibrium (with 2011 forcing held constant), with an $F_{2\times CO2}$ of 4.03 Wm^{-2} ICS correctly reaches the AOGCM's 3.42 K ECS, whereas using PH17's method ICS only reaches 3.14 K.

[6] To achieve an $F_{2\times CO2}$ of 3.7 Wm^{-2} for the ensemble-median-based AOGCM eigenmode decomposition we set the ultrafast eigenmode parameter λ_1 (which, in ERF terms, cannot be separately estimated from abrupt4xCO2 data) equal to the fast eigenmode parameter λ_2 and then scaled both of them by 0.947.

4. PH17's ensemble-median ECS estimate of 3.42 K is 1.15x the revised ensemble-median-based AOGCM ICS estimate of 2.98 K. That ICS estimate is close to ICS computed by A17's method (revised to use $F_{2\times CO2}$ from regression over years 2–10) using step-emulated 1pctCO2 responses derived from the ensemble-median-parameters abrupt4xCO2 fitted eigenmode-decomposition. However, when ICS is calculated by that method using instead ΔT and ΔN values step-emulated directly from actual abrupt4xCO2 simulation data for each AOGCM, the ensemble-median ECS-to-ICS ratio is only 1.10x, very close to the 1.11x ratio that we estimate for PH17's ensemble using our Table S2 values. The difference is likely mainly due to the median of individual models' ECS-to-ICS ratio differing from the ECS-to-ICS ratio for the canonical ensemble-median model and to inaccuracy in the eigenmode-decomposition fits.

S6: Figures relating to section 7a – Discussion of Gregory and Andrews 2016 (GA16)

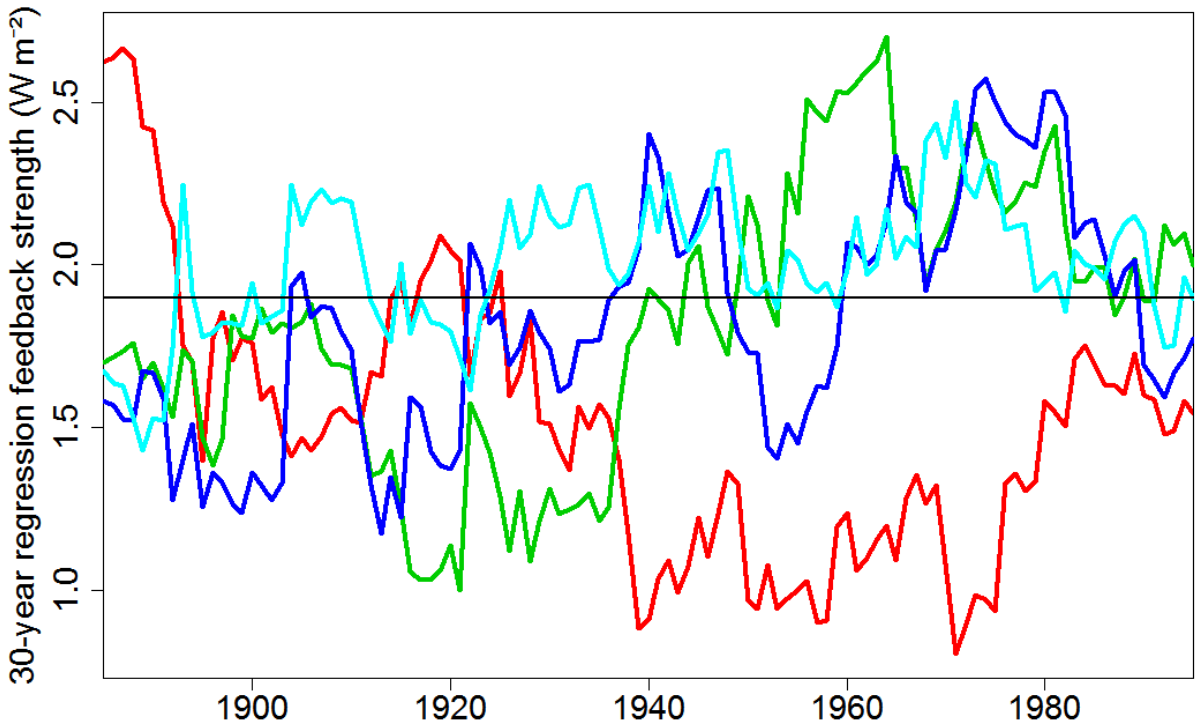

Figure S2. Variation of 30-year sliding-window regression based estimates of λ from pseudodata. HadGEM2-A ensemble-mean 1871–2010 annual T values were multiplied by a λ value of 1.9 $Wm^{-2}K^{-1}$ and N values from four non-overlapping 140-year segments of the 576-year HadGEM2-ES preindustrial control simulation were added, producing four instances of synthetic R pseudodata each embodying a constant λ value but a different realisation of internal variability in R. Internal variability in T was ignored. Regression using a 30-year sliding window was then undertaken, as in GA16. The plot shows how the resulting estimates of λ vary over time for each pseudodata instance. Note that real-world internal variability in R, treated as equal to that in the HadGEM2-ES piControl run, is being regarded as reflecting random variations in real-world SST patterns not associated with changes in T (to which variability in R is only weakly related in the piControl run). The observationally-based SST patterns used to drive the amipPiForcing simulations will reflect that variability, which should cause a broadly similar variability in R as simulated by the AGCM, common to all runs in the simulation ensemble. Long-term estimates of λ derived from the same pseudodata instances by using the first 30 years' data to calculate anomalies for the remaining 110 years and then computing a no-intercept regression fit of R anomalies on T anomalies are no more than 3% in error in any of the four cases. The same is true when estimating λ as the ratio of the means of R and T anomalies over the final 15 years.

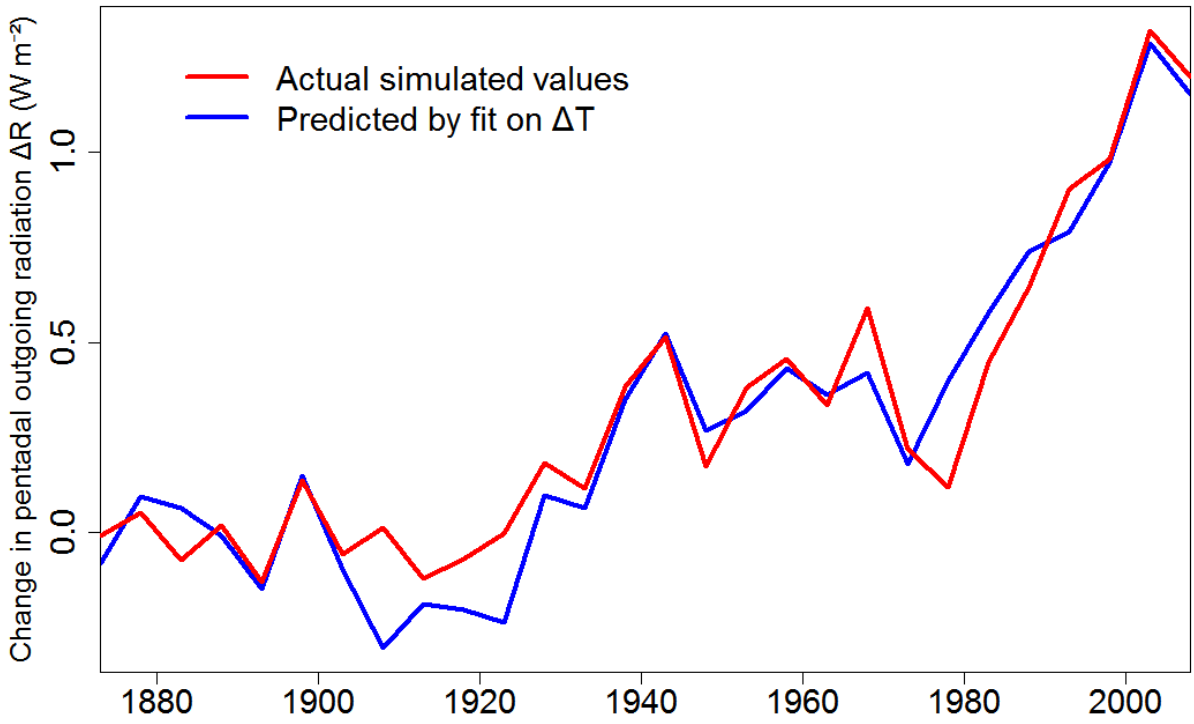

Figure S3. Net outgoing radiation ΔR in the GA16 HadGEM2-A amipPiForcing simulations: pentadal-mean anomalies relative to 1871–1900 means, averaging over the two runs in the ensemble. Actual simulated changes in R in the simulations (red) are compared with those predicted using a linear fit ($\Delta R = 0.012 + 1.898\ \Delta T$) over 1871–2010 to simulated changes in T, corresponding to a constant λ of 1.90 $Wm^{-2}K^{-1}$. It has been found previously that over most of the first quarter of the 20th century surface temperature was anomalously cool (Gillett et al 2012); it was also anomalously cool in relation to R then in both HadGEM2-A simulation runs. The fit was accordingly estimated with 1906–1925 pentads excluded, resulting in an R^2 of 0.94 and a near-zero y-intercept. The fit for each individual run has a similar slope (within ±5%), negligible y-intercept and 0.93 R^2. A no-intercept fit with all pentads included is almost identical ($\Delta R = 1.868\ \Delta T$). Removing also the 1976-1980 point, where the fit is equally poor, would have virtually no effect on the estimated fit.

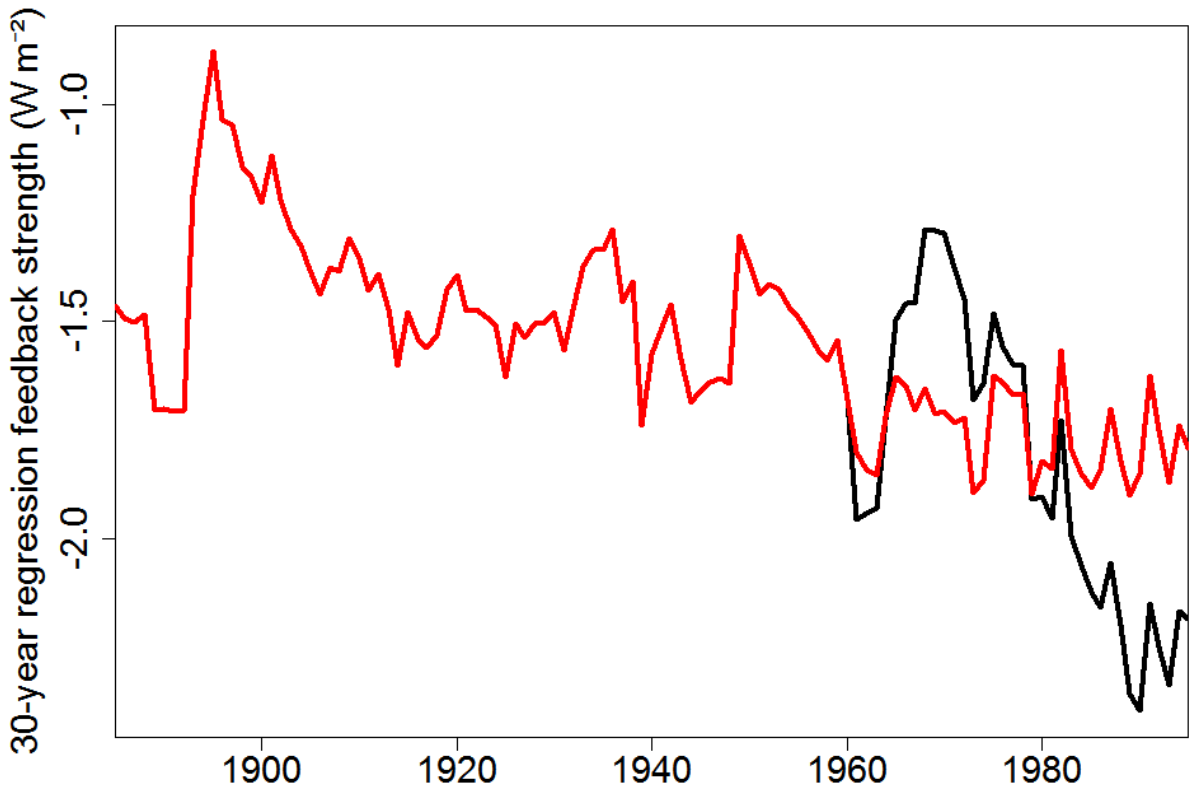

Figure S4. Variation of 30-year sliding-window regression based estimates of λ from HadGEM2-A annual 1871–2010 amipPiForcing simulation *T* and *R* ensemble-mean values: effect of adjusting 1976– 1985 *R* values. The black line is using unaltered data, and replicates the thick black line in GA16 Figure 2(c). The red line (which overlays the black line until 1960) is the same except that the value of *R* has been increased by 0.3 Wm^{-2} in each of the years 1976–1985, a period over which simulated *R* values were unusually low in relation to those predicted from a linear fit on *T* values over the whole 1871–2010 period. The resulting regression-based estimates of λ fluctuate in a narrow band after 1960.

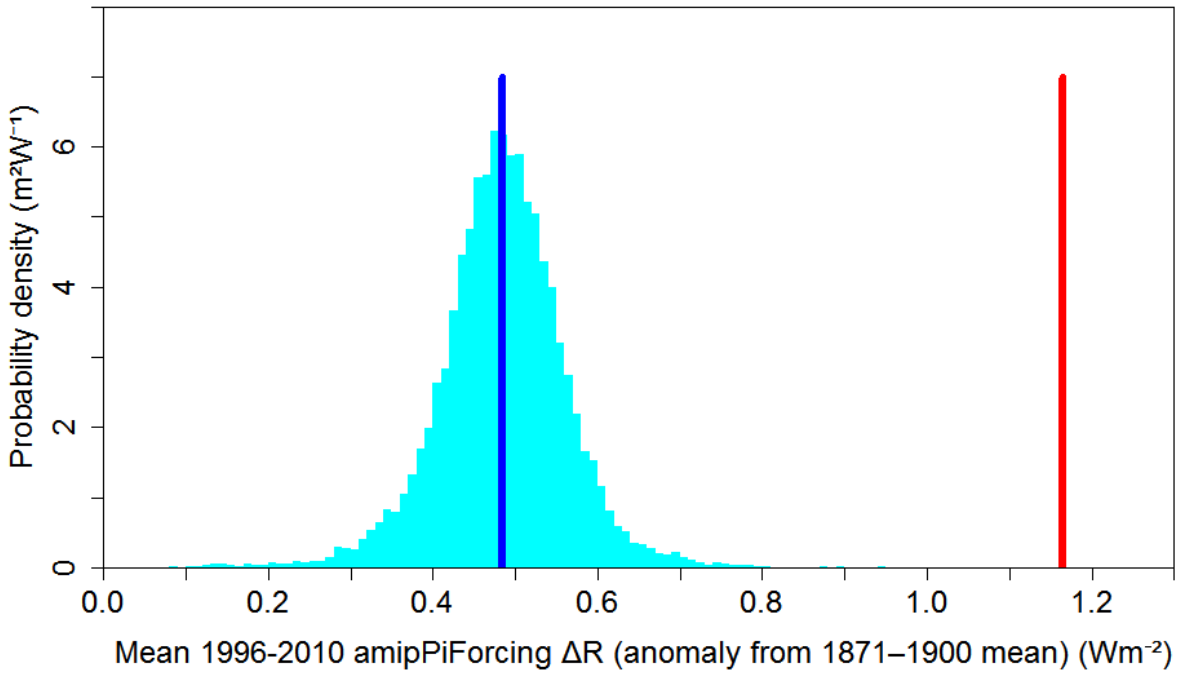

Figure S5. Comparison of actual ΔR anomalies over HadGEM2-A amipPiForcing experiment with their distribution, allowing for simulated internal variability, if λ equalled its HadGEM2-ES abrupt4xCO2 simulation years 1–20 regression estimate ($\lambda_{4\times}$) of 0.82 $Wm^{-2}K^{-1}$. The red vertical line shows the actual ensemble-mean ΔR (1.16 Wm^{-2}) for the final 15 years of the two amipPiForcing runs (which differ little). The blue vertical line shows what ΔR would have been (0.48 Wm^{-2}) if the AGCM had exhibited a climate feedback of $\lambda_{4\times}$, based on multiplying $\lambda_{4\times}$ by the amipPiForcing ensemble-mean ΔT anomaly for the final 15 years (0.59 K). The cyan histogram shows the dispersion around that $\lambda_{4\times}$-based ΔR allowing for simulated internal variability in R and T. All (>18,000) 140-year segments of detrended piControl simulation N ($\equiv -R$) and T data from 43 CMIP5 models are used to compute corresponding final 15-year ΔR_{IV} and ΔT_{IV} anomalies. For each thus generated realization of internal variability, ΔR is computed as $\lambda_{4\times}$ ($\Delta T - \Delta T_{IV}$) + ΔR_{IV}, ΔT being the ensemble-mean anomaly for the final 15 years of the amipPiForcing experiment. The distribution's 99th percentile is 0.68 Wm^{-2}. The 99th percentile is 0.67 Wm^{-2} if ΔR_{IV} and ΔT_{IV} are computed using final 15 years anomalies from all (7 million) piControl run segments 45 or more years long. All anomalies are relative to the mean over the first 30 years of the data involved.

Additional references

Andrews, T., J. M. Gregory, and M. J. Webb, 2015:The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *J. Climate*, 28.4, 1630-1648.

Charlson, R. J., et al., 1992: Climate forcing by anthropogenic aerosols. *Science* **255**.5043: 423-430.

Crowley, T. J., and M. B. Unterman, 2013: Technical details concerning development of a 1200 yr proxy index for global volcanism. *Earth Syst. Sci. Data*, 5, 187-197.

Domingues, C. M., Church J. A., White N. J., Gleckler P. J., Wijffels S. E., Barker P. M., and Dunn J. R., 2008: Improved estimates of upper-ocean warming and multi-decadal sea-level rise. *Nature* 453:1090–3.

Eyring, V., and Coauthors, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6: experimental design and organization. *Geosci. Model Dev.*, 9, 1937-1958.

Flato G, Marotzke J, et al.,2014: Evaluation of Climate Models. *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, Cambridge, 741-866.

Forster PM, Andrews T, Good P, Gregory JM, Jackson LS, Zelinka M, 2013: Evaluating adjusted forcing and model spread for historical and future scenarios in the CMIP5 generation of climate models. *J Geophys Res* 118:1139–1150.

Gillett N. P., Arora VK, Flato GM, Scinocca JF, von Salzen K, 2012: Improved constraints on 21st-century warming derived using 160 years of temperature observations. *Geophys. Res. Lett*., **39**, L01704, doi:10.1029/2011GL050226.

Gregory, J. M., Andrews, T., & Good, P., 2015: The inconstancy of the transient climate response parameter under increasing CO2. *Phil. Trans. R. Soc. A*, 373(2054), 20140417.

Gupta, A S, et al., 2013: Climate drift in the CMIP5 models. *J. Climate* 26.21, 8597-8615.

Kopp, G., Lawrence, G., and Rottman, G., 2005: The Total Irradiance Monitor (TIM): Science Results. *Solar Physics,* **230**, 1, 129-140, doi:10.1007/s11207-005-7433-9.

Myhre, G., Highwood E. J., Shine K. P., and Stordal F., 1998: New estimates of radiative forcing due to well-mixed greenhouse gases. *Geophys. Res. Lett.*, 25(14):2715-2718.

Purkey, S. G., and G. C. Johnson, 2010: Warming of Global Abyssal and Deep Southern Ocean Waters between the 1990s and 2000s: Contribution to Global Heat and Sea Level Rise Budgets. *J. Climate*, 23, 6336 – 635.

Ridley, D. A., and Coauthors, 2014: Total volcanic stratospheric aerosol optical depths and implications for global climate change. *Geophys. Res. Lett.*, 41(22), 7763-7769.

Sato, M., J.E. Hansen, M.P. McCormick, and J.B. Pollack, 1993: Stratospheric aerosol optical depth, 1850-1990. *J. Geophys. Res.* 98, 22987-22994.

Vernier, J. P., et al., 2011: Major influence of tropical volcanic eruptions on the stratospheric aerosol layer during the last decade, *Geophys. Res. Lett*., 38, L12807, doi:10.1029/2011GL047563.

WMO, 2016: WMO Greenhouse Gas Bulletin No. 12 : October 2016. Data table at http://ds.data.jma.go.jp/gmd/wdcgg/pub/global/globalmean.html.