The DOI for this manuscript is doi: 10.1175/JCLI-D-17-0667.1

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

Nicholas Lewis[1]

Judith Curry[2]

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

[2] Climate Forecast Applications Network, Reno, NV USA Email: curry.judith@cfanclimate.com

Abstract: Energy budget estimates of equilibrium climate sensitivity (ECS) and transient climate response (TCR) are derived based on the best estimates and uncertainty ranges for forcing provided in the IPCC Fifth Assessment Scientific Report (AR5). Recent revisions to greenhouse gas forcing and post-1990 ozone and aerosol forcing estimates are incorporated and the forcing data extended from 2011 to 2016. Reflecting recent evidence against strong aerosol forcing, its AR5 uncertainty lower bound is increased slightly. Using a 1869–1882 base period and a 2007−2016 final period, which are well-matched for volcanic activity and influence from internal variability, medians are derived for ECS of 1.50 K (5−95%: 1.05−2.45 K) and for TCR of 1.20 K (5−95%: 0.9−1.7 K). These estimates both have much lower upper bounds than those from a predecessor study using AR5 data ending in 2011. Using infilled, globally-complete temperature data gives slightly higher estimates; a median of 1.66 K for ECS (5−95%: 1.15−2.7 K) and 1.33 K for TCR (5−95%:1.0−1.90 K). These ECS estimates reflect climate feedbacks over the historical period, assumed time-invariant. Allowing for possible time-varying climate feedbacks increases the median ECS estimate to 1.76 K (5−95%: 1.2−3.1 K), using infilled temperature data. Possible biases from non-unit forcing efficacy, temperature estimation issues and variability in sea-surface temperature change patterns are examined and found to be minor when using globally-complete temperature data. These results imply that high ECS and TCR values derived from a majority of CMIP5 climate models are inconsistent with observed warming during the historical period.

1. Introduction

There has been considerable scientific investigation of the magnitude of the warming of Earth's climate from changes in atmospheric carbon dioxide (CO_2) concentration. Two standard metrics summarize the sensitivity of global surface temperature to an externally imposed radiative forcing. Equilibrium climate sensitivity (ECS) represents the equilibrium change in surface temperature to a doubling of atmospheric CO_2 concentration. Transient climate response (TCR), a shorter-term measure over 70 years, represents warming at the time CO_2 concentration has doubled when it is increased by 1% a year.

For over thirty years, climate scientists have presented a likely range for ECS that has hardly changed. The ECS range 1.5–4.5 K in 1979 (Charney 1979) is unchanged in the 2013 Fifth Assessment Scientific Report (AR5) from the IPCC. AR5 did not provide a best estimate value for ECS, stating (Summary for Policymakers D.2): "No best estimate for equilibrium climate sensitivity can now be given because of a lack of agreement on values across assessed lines of evidence".

At the heart of the difficulty surrounding the values of ECS and TCR is the substantial difference between values derived from climate models versus values derived from changes over the historical instrumental data record using energy budget models. The median ECS given in AR5 for current generation (CMIP5) atmosphere-ocean global climate models (AOGCMs) was 3.2 K, versus 2.0 K for the median values from historical-period energy budget based studies cited by AR5.

Subsequently Lewis and Curry (2015; hereafter LC15) derived, using observationally-based energy budget methodology, a median ECS estimate of 1.6 K from AR5's global forcing and heat content estimate time series, which made the discrepancy with ECS values derived from AOGCMs even larger. LC15 also derived a median TCR value of 1.3 K, well below the 1.8 K median TCR for CMIP5 models in AR5.

Considerable effort has been expended in attempts to reconcile the observationally-based ECS values with values determined using climate models. Most of these efforts have focused on arguments that the methodologies used in the energy balance model determinations result in ECS and/or TCR estimates that are biased low (e.g., Marvel et al. 2016; Richardson et al. 2016; Armour 2017).

Using a standard global energy budget approach, this paper seeks to clarify the implications for climate sensitivity (both ECS and TCR) of incorporating the most up-to-date surface temperature, forcing and ocean heat content data. Forcing and heat content estimates given in AR5 are extended from 2011 to 2016, with recent revisions to greenhouse gas forcing-concentration relationships and post-1990 tropospheric ozone and aerosol forcing changes applied and a new ocean heat content dataset incorporated. This paper also addresses a range of concerns that have been raised regarding using energy balance models to determine climate sensitivity: variability in patterns of sea-surface temperature change, non-unit forcing efficacy, temperature estimation issues and time-varying climate feedbacks.

The paper is structured as follows. The global energy budget approach is discussed in Section 2. Section 3 deals with data sources and uncertainties, Section 4 with choice of base and final periods, and methods are described in Section 5. Section 6 sets out the results, which are discussed in Section 7. Section 8 concludes.

2. Global energy budget approach

A general energy budget framework has been widely used in the estimation and analysis of climate sensitivity, such as by Armour and Roe (2011) and Roe and Armour (2011), and in AR5 (Bindoff et al. 2014). Estimation of climate sensitivity from changes in conditions between periods early and late in the industrial era has been developed by Gregory et al. (2002), Otto et al. (2013), Masters (2014), LC15 and other papers. Advantages of the energy budget approach are described by LC15; relative to less simple models that use zonally, hemispherically or land-ocean resolved data, the energy budget approach includes improved quantification of and robustness against uncertainties through use only of global mean data.

Generally, complex models are ill-suited to observationally-based climate sensitivity estimation since it may not be practicable to produce, by perturbing their internal parameters, a simulated climate system that is adequately consistent with observed variables. An increasingly popular alternative is the 'emergent constraint' approach: identifying observationally-constrainable metrics in the current climate which correlate with ECS in complex models. However, it has been shown for CMIP5 models that all such metrics are likely only to constrain shortwave cloud feedback, and not other factors controlling their ECS (Qu et al. 2017). The ability in a state-of-the-art complex model to engineer ECS over a wide range (largely arising from differing shortwave cloud feedback) by varying the formulation of convective precipitation, without being able to find a clear observational constraint that favors one version over the others (Zhao et al. 2016), casts further doubt on the emergent constraint approach.

Using a simple rather than a complex climate model also has the important advantage of transparency and reproducibility. What determines ECS and TCR in a complex model is obscure, and their estimation is affected by internal variability. The energy budget framework provides an extremely simple physically-based climate model that, given the assumptions made, follows directly from energy conservation. It has only one uncertain parameter, λ, which can be directly derived from estimates of changes in historical global mean surface temperature (hereafter 'surface temperature'), forcing and heat uptake rate.

The main assumption made by the energy budget model concerns the radiative response ΔR to a change in radiative forcing ΔF that alters positively the Earth's net downwards top-of-atmosphere (TOA) radiative imbalance N. The assumption is that, in temporal mean terms, ΔR – the change in net outgoing radiation resulting from the change in the state of the climate system caused by the forcing imposition – is linearly proportional to the forcing-induced change in surface temperature ΔT. Mathematically,

$$\Delta R = \lambda \Delta T + \mu_R \qquad (1)$$

with λ – the climate feedback parameter, representing the increase in net outgoing energy flux per degree of surface warming – constant, and μ_R a random zero-mean residual term representing internal fluctuations in the system unrelated to fluctuations in T. Together μ_R and fluctuations in R arising, through its relation to T, from internal variability in T – which will have a different signature –represent the internal variability in R. A constant λ implies it is independent of T, other aspects of the climate state, the magnitude and composition of ΔF, and the time since forcing was applied.

By conservation of energy, $\Delta N = \Delta F - \Delta R$. Therefore, in temporal mean terms, substituting using (1)

$$\lambda = (\Delta F - \Delta N) / \Delta T \qquad (2)$$

It follows from (2) that, designating the radiative forcing from a doubling of atmospheric CO_2 concentration as $F_{2\times CO2}$, once equilibrium is restored following such a doubling (implying ΔN=0),

$$\lambda = F_{2\times CO2} / \mathrm{ECS} \qquad (3).$$

Hence, substituting in (2), in general

$$\mathrm{ECS} = F_{2\times CO2} \frac{\Delta T}{\Delta F - \Delta N} \qquad (4)$$

with the CO_2 forcing component of ΔF calculated on a basis consistent with that used for $F_{2\times CO2}$. Here, N is conventionally regarded and measured as the rate of planetary heat uptake, which provides identical ΔN values to measuring its net downwards radiative imbalance. Equation (4) assumes that ΔT is entirely externally forced, but it does not imply a linear relationship between ΔN and ΔT, unlike the 'kappa' model (Gregory and Forster 2008).

We apply (4) to estimate ECS based on changes in mean values of estimates for T, F and N between well separated, fairly long base and final periods.

Being inferred from transient changes, ECS as defined in (4) is an effective climate sensitivity that embodies the assumption of a constant linear climate feedback parameter λ. Equilibrium climate sensitivity, by contrast, requires the atmosphere-ocean system (although not slow components of the climate system, such as ice sheets) to have equilibrated. Equilibrium and effective climate sensitivity will not be identical if the feedback parameter is inconstant over time or dependent on ΔF or ΔT. The behavior of CMIP5 models may provide some insight into these issues.

Throughout 140-year simulations in which CO_2 forcing is increased smoothly at 1% per annum (1pctCO2), the responses of almost all CMIP5 AOGCMs can be accurately emulated by convolving the rate of increase in forcing with the step response in their simulations in which CO_2 concentration is abruptly quadrupled (abrupt4xCO2) (Figure 1; Good et al. 2011; Caldeira and Myhrvold 2013). Such behavior strongly suggests that feedback strength in CMIP5 models generally does not change with ΔF or ΔT *per se*, at least for CO_2 forcing from up to respectively quadrupling of its preindustrial concentration and the warming reached in abrupt4xCO2 simulations after half a century or so (typically 4−5 K). Otherwise one would expect to see divergences, particularly in the first few decades of the 1pctCO2 simulation when the applicable temperature is furthest below the mean temperature of the abrupt4xCO2-derived step-emulation components. We have also investigated feedback strength in the MPI-ESM-1.2 AOGCM under differing abrupt CO_2 increases. Feedback strength is almost the same between abrupt2xCO2 and abrupt4xCO2 simulations up to at least year 150, when ΔT reaches 5 K under quadrupled CO_2.

However, in most CMIP5 AOGCMs, λ (here $-\frac{dN}{dT}$) tends to decrease a few decades into abrupt4xCO2 simulations – when N is plotted against T (a 'Gregory plot'), the slope is gentler after that time – although generally λ then remains almost constant for the rest of the simulation (Armour 2017 Fig. S1). In some cases, the decrease in λ may be linked to temperature- or time-dependent energy leakage (Hobbs et al. 2016). However, typically the decrease in λ appears to arise primarily from the strength of modeled shortwave cloud feedbacks varying with time, likely linked to evolving patterns of surface warming (Andrews et al. 2015). The decrease in λ means that effective climate sensitivity estimates derived from simulations forced by abrupt or ramped CO_2 changes tend to increase with the analysis period, although in most cases they change only modestly once a multidecadal period has elapsed. It is unclear to what extent, if any, this behavior occurs in the real climate system. Possible implications of time-varying feedbacks for historical period energy budget ECS estimation are analyzed in section 7f. Until then, ECS estimates are not distinguished according to what extent they are potentially affected by time-varying feedbacks.

ECS would also differ from the estimate provided by (4) if that were significantly affected by internal variability, or if effect on ΔT or ΔN of the composite forcing change over the estimation period differed from that of CO_2 forcing. These issues are discussed in Sections 3b, 3c, 4 and 7c. The possibility of internal variability in spatial surface temperature patterns affecting ECS estimation is discussed in Section 7a.

Transient climate response (TCR) is the increase in surface temperature (averaged over twenty years) at the time of CO_2 concentration doubling when it is increased by 1% a year, implying an almost linear forcing ramp over 70 years. Although designed as a measure of transient response in AOGCMs, TCR can be regarded as a property of the real climate system. TCR can be estimated by scaling the ratio of the response of global surface temperature to the change in forcing accruing approximately linearly over a period of circa 70 years (Bindoff et al. 2014, p.920). That is:

$$\mathrm{TCR} = F_{2\times\mathrm{CO2}} \frac{\Delta T}{\Delta F} \tag{5}$$

TCR can be estimated using (5) with a recent final period and a base period ending circa 1950. Although occurring mainly over the last 70 years, the effect on surface temperature of the development of forcing over the whole historical period (post ~1850) has been estimated to be broadly equivalent to that of a 100-year linear forcing ramp (Armour 2017). TCR may therefore also be estimated using a base period early in the historical period, with a possible marginal upwards bias since with a longer ramp period the climate system will have had more time to respond to the ramped forcing. LC15 found that estimating TCR using (5) with a recent final period and a base period either early in the historical period or of 1930−1950 provided an estimate of TCR closely consistent with its definition.

The energy budget approach has also been applied to estimate both ECS and TCR using regression over all or a substantial part of the historical period, rather than taking differences between base and final periods (Gregory and Forster 2008; Schwartz 2012). Although regression makes fuller use of available information than the two-period method, using averages over base and final periods captures much of the available information, since internal variability is high on sub-decadal timescales and total forcing has only become reasonably large relative to its uncertainty relatively recently. Moreover, handling multidecadal internal variability and volcanic eruptions poses a challenge when using regression. Gregory and Forster (2008) excluded years with significant volcanism, but subsequent years may be affected by the recovery from volcanic forcing.

It is important to use an appropriate forcing metric for energy budget sensitivity estimation. The surface temperature response to forcing from a particular agent relative to that from CO_2 (its 'efficacy': Hansen et al. 2005) is in some cases sensitive to the metric used. In such cases, efficacy is normally much closer to unity when the effective radiative forcing (ERF) metric (Sherwood et al. 2015; Myhre et al. 2014) is used rather than the common stratospherically-adjusted radiative forcing (RF) metric. Unlike ERF, the RF metric does not allow for the troposphere and land surface adjusting to the imposed forcing. Since ERF is a construct designed to fit the global radiative response as a linear function of ΔT over time scales of decades to a century (Sherwood et al. 2015), it is an appropriate metric for energy budget sensitivity estimation. References here to forcing are to ERF except where indicated otherwise. AR5 only gives estimated forcing time-series for ERF. Its best estimates of 2011 ERF differ from those of RF only for aerosols and contrails, although uncertainty ranges are generally wider for ERF than for RF.

Uncertainty in energy budget estimates of ECS and TCR from instrumental observations stems primarily from uncertainty in ΔF (LC15), which also produces most of the asymmetry in probability distributions for ECS and TCR estimates (Roe and Armour 2011). The two main contributors to uncertainty in ΔF are aerosols and, to a substantially smaller extent, well-mixed greenhouse gases (WMGG).

3. Data sources and uncertainties

As in LC15, forcing and heat uptake data and uncertainty estimates identical to those given in AR5 have been used unless stated otherwise. AR5 estimates represent carefully considered assessments in which many climate scientists with relevant expertise were involved, and underwent an extensive review process. Post-2011 values have insofar as possible been derived entirely from observational data, on a basis consistent with that in AR5. Trend-based extrapolation has only been used for some minor forcing and heat uptake components, except for 2016 aerosol and tropospheric ozone forcing. Only a brief discussion of the treatment of data uncertainties and internal variability is given here, since full details of our treatment can be found in LC15. This section summarizes information about the forcing, heat uptake and temperature data. Full details of changes relative to AR5 estimates for certain forcing and heat uptake components, and of the updating of all components from 2011 to 2016, are provided in the Supplementary Material (S1 and S2).

a. Forcings

ERF time series medians up to 2011 (relative to 1750) are sourced from Table AII.1.2 of AR5, with uncertainty estimates for 2011 derived from Table 8.6 and Table 8.SM.5 of AR5. The only changes to Table AII.1.2 values concern forcing from the principal WMGG, where recent revisions to forcing-concentration relationships (Etminan et al. 2016) have been incorporated throughout, and post-1990 changes in aerosol and tropospheric ozone forcing, where new estimates of their evolution based on updated anthropogenic emission data for 1990–2015 (Myhre et al. 2017) have been adopted, adding their estimated post-1990 changes to the AR5 1990 values. Recent evidence concerning volcanic forcing (Andersson et al. 2015) was considered, but no revision to AR5 estimates was found necessary (Supplementary Material S1). The principal effect of these revisions is to make methane (CH_4) forcing more positive, and post-1990 aerosol forcing less negative, than per AR5. After reaching −0.9 Wm^{-2} in 1995, $ERF_{Aerosol}$ weakens to −0.8 Wm^{-2} in 2011. The 2011 forcing uncertainty ranges are used, in conjunction with AR5 2011 medians, to specify the fractional uncertainty for each forcing constituent.

Since AR5, understanding of anthropogenic aerosol forcing ($ERF_{Aerosol}$) has improved. A number of recent studies point to total aerosol forcing being substantially weaker than the lower end of the −1.9 to −0.1 Wm^{-2} 2011 range (median −0.9 Wm^{-2}) given in AR5, primarily due to negative forcing from aerosol-cloud interactions being weaker than previously thought (Seifert et al. 2015; Stevens 2015; Gordon et al. 2016; Zhou and Penner 2017; Nazarenko et al. 2017; Lohmann 2017; Malavelle et al. 2017; Stevens et al. 2017; Fiedler et al. 2017; Toll et al. 2017). Recent evidence regarding positive aerosol forcing from absorbing carbonaceous aerosols (Wang et al. 2014, Samset et al. 2014, Wang et al. 2016, Zhang et al. 2017) is mixed, on balance suggesting it may be lower than the AR5 best estimate, but above its lower uncertainty bound in AR5. Although some post-AR5 studies (e.g. Cherian et al. 2014, McCoy et al. 2017) have reported relatively strong aerosol forcing, Stevens (2015) presented several observationally-based arguments that total aerosol forcing since preindustrial was weak, and could not be stronger than −1.0 Wm^{-2}.[1] Supporting those arguments, Zhou and Penner (2017) and Sato et al. (2018) showed that negative cloud-lifetime aerosol forcing simulated by AOGCMs was unrealistic, Bender et al. (2016) showed that the positive correlation between aerosol loading and cloud albedo displayed in most climate models is not seen in observations, and Nazarenko et al. (2017) showed that aerosol forcing was weaker when climate feedbacks were allowed for. In the light of these developments, the −1.9 Wm^{-2} model-derived lower bound for 2011 aerosol forcing in AR5 now appears too strong. We have therefore weakened it slightly to −1.7 Wm^{-2}, as in Armour (2017), making the range symmetrical about the AR5 2011 median.

Following LC15, CO_2 and 'GHG Other' forcings are combined into a single ERF_{GHG} time series, since AR5 does not distinguish between the two as regards ERF uncertainty.

[1] Substituting, for consistency, the higher WMGG forcing used in this study for that used in Stevens (2015) would slightly change its −1.0 Wm^{-2} aerosol forcing lower bound, to −1.06 Wm^{-2}, too little to weaken the argument for the proposal made here.

Uncertainty in forcing from WMGG almost entirely relates to how much forcing a given concentration of each greenhouse gas produces – uncertainty in concentrations is minor – and is likely highly correlated among WMGG. AR5 (Section 8.5.1) assumes that fractional ERF uncertainties for CO_2 applies to all WMGG and to total WMGG, implying that fractional uncertainty in $F_{2\times CO2}$ is the same as, and fully correlated with, that in ERF_{GHG}. We follow Otto et al. (2013) and LC15 in adopting this assumption. Although uncertainty in WMGG forcing is substantial, since $F_{2\times CO2}$ appears in the numerator of (4) and (5), and ΔF (to which ERF_{GHG} is by far the largest contributor) in the denominator, the effects on ECS and TCR estimation of uncertainty in forcing from WMGG cancel out to a substantial extent. Dropping the assumption of uncertainty being correlated between CO_2 and 'GHG Other' forcing would have a negligible effect on ECS and TCR estimate uncertainty ranges. That same applies if in addition the ERF-to-RF uncertainty ratio for non-CO_2 WMGG were increased from the 20%:10% ratio assumed in AR5 to 30%:10%, even if uncertainty were treated as perfectly correlated between all non-CO_2 WMGG, as in AR5.

Ozone (both Tropospheric and Stratospheric), Stratospheric H_2O (Water vapor) and Land Use Change forcings, for which uncertainty distributions can be added in quadrature, are combined into a single ERF_{OWL} forcing component series (termed $ERF_{nonGABC}$ in LC15).

The resulting forcing best estimates and uncertainties used for the main results are summarised in Table 1, for both 2011 and 2016. AR5 forcing estimates and uncertainty ranges for 2011 are also shown. Following LC15, the uncertainty ranges for solar and volcanic forcing have been widened. The revised total 1750−2011 anthropogenic forcing estimate has increased by 9% from the AR5 value; the largest contribution comes from the revision in CH_4 forcing. $F_{2\times CO2}$ has also been revised up by 2.5%, to 3.80 Wm^{-2}, which has an opposing effect on sensitivity estimation to the upward revision in total forcing. Figure 2 shows the original AR5 and revised anthropogenic forcing time-series.

LC15 concluded that volcanic forcing ($ERF_{Volcano}$) in AR5 needs to be scaled down by 40−50% in order to produce a comparable effect on surface temperature to ERF_{GHG} and other forcings. Gregory et al. (2016) likewise found that volcanic forcing produced a substantially smaller response in AOGCMs than CO_2 forcing. They quantified the effect in HadCM3, where $ERF_{Volcano}$ was smaller relative to stratospheric aerosol optical depth than per AR5 and its efficacy was also lower, implying that AR5 volcanic forcing needed to be scaled down by ~50% for use in a global energy budget model. Since there is no authority in AR5 for applying an adjustment factor, the issue is sidestepped by using base and final periods with matching mean volcanic forcing, as in LC15. The results of applying a scaling factor of 0.55 are shown where sensitivity testing of estimates to the choice of base and final periods involves mismatched volcanic forcing. Likewise, as in LC15 the AR5 Land Use Change forcing (ERF_{LUC}) series is used despite it representing only effects on surface albedo. AR5 assessed that including other effects of land use change it is about as likely as not to have caused net cooling. The effect of setting ERF_{LUC} to zero is also reported. AR5 gives an estimated efficacy range of 2−4 for the minor black carbon on snow and ice forcing (ERF_{BCsnow}), which is applied probabilistically.

b. Heat uptake

Planetary heat uptake – the rate of increase in its heat content – occurs primarily (>90%) in the ocean. The AR5 estimates for heat uptake by the atmosphere, ice, land and the deep (sub-2000 m) ocean are used unaltered up to 2011 and extended to 2016. AR5's source for 700−2000 m ocean heat content (OHC), Levitus et al. (2012), has been updated, but a new dataset (Cheng et al. 2017) is also available; the average of those two datasets is used here.

AR5's source for 0–700 m OHC has not been updated to 2016. The average of three available fully updated 0−700 m OHC datasets (Cheng et al., Levitus et al., and Ishii and Kimoto (2009)) is used instead, for all years. There are considerable divergences between OHC estimates from the various datasets, arising from differences in the data used, corrections made to it, and the mapping (infilling) methods used. Averaging results from different OHC datasets reduces the effect of errors particular to individual datasets. Over the main 1995−2011 and 1987−2011 final periods used in LC15, implementing the foregoing changes to the sourcing and calculation of OHC estimates produces slightly higher 0−2000 m ocean heat uptake (OHU) estimates than use of the original AR5 datasets. Since the mid-2000s, when the Argo floating buoy network achieved near-global coverage, OHC uncertainty has been lower. The revised estimation basis produces total heat uptake within 0.02 Wm^{-2} of the estimates by Desbruyeres et al. (2017) of 0.72 Wm^{-2} over 2006-2014 and by Johnson et al. (2016) of 0.71 W m^{-2} over 2005−2015.

As in previous energy budget studies, AOGCM simulation-derived estimates of heat uptake are used for the base periods, since OHC was not measured then. The heat uptake values used in LC15, which were derived from simulations by CCSM4 starting in AD 850 (Gregory et al. 2013), scaled by 0.60, were 0.15, 0.10 and 0.20 W m^{-2} respectively for the 1859−1882, 1850−1900 and 1930−1950 base periods. The unscaled CCSM4-derived values were consistent with the value derived by Gregory et al. (2002) from a different AOGCM. The LC15 values are adopted (taking the 1859−1882 value for 1869−1882), as are the LC15 standard error estimates, being in each case 50% of the heat uptake estimate.

The variability in total heat uptake of 0.045 Wm^{-2} for all base and final periods used in LC15, derived from the ultra-long HadCM3 (Gordon et al. 2000) control run, is also adopted. Investigation showed this to be adequate for each of the base and final periods used here.

c. Surface temperature

As in LC15, the HadCRUT4 surface temperature dataset (Morice et al. 2012) is used, updated from HadCRUT4v2 to HadCRUT4v5. Results are also presented using a globally-complete version infilled by kriging (Had4_krig_v2: Cowtan and Way 2014). The surface temperature trends over 1900−2010 are identical in both versions, with Had4_krig_v2 warming faster than HadCRUT4v5 early and late in the record.

Unlike GISStemp and NCDC MLOST (now NOAA GlobalTemp), the other two surface temperature datasets cited in AR5, HadCRUT4 extends back to 1850 rather than 1880, providing adequate data early in the historical period prior to the period of heavy volcanism from 1883 on. The warming shown by the infilled GISStemp and NOAA4.0.1 datasets between twenty-year periods early and late in their records (1880−1899 and 1997−2016) was respectively 0.85 K and 0.82 K, against 0.83 K for HadCRUT4v5 and 0.89 K for Had4_krig_v2.

Both versions of HadCRUT4 provide an ensemble of 100 temperature realizations that preserves the time-dependent correlation structure. Uncertainty in mean surface temperature for each period is calculated on a basis consistent with the applicable covariance matrix of observational uncertainty, and combined in quadrature with an estimate of inter-period internal variability in ΔT. The LC15 estimate of 0.08 K standard deviation for such internal variability is adopted; it was conservatively scaled up from 0.06 K derived from the ultra-long HadCM3 control run. Sensitivity testing in LC15 showed that a further 50% increase in internal variability in ΔT had almost no effect on uncertainty in ECS and TCR estimates.

4. Choice of base and final periods

Two-period energy budget studies have used base and final periods lasting between one and five decades. Longer periods reduce the effects of interannual and decadal internal variability, but shorter periods make it feasible to avoid major volcanism and a short final period provides a higher signal. Base and final periods should be at least a decade, to sufficiently reduce the influence of interannual variability. Volcanic forcing efficacy, relative to AR5 forcing estimates, appears to be substantially below unity, and may differ according to the location and type of eruption. Moreover, prior to the satellite (post-1978) era there are considerable uncertainties regarding the magnitude of volcanic eruptions and resulting forcing. Therefore, accurate sensitivity estimation requires estimated volcanic forcing to be matched between the base and final period, and relatively low. Likewise, initial and final periods should be well matched regarding the influence of the principal sources of interannual and multidecadal internal variability, notably ENSO and Atlantic multidecadal variability.

Atlantic multidecadal variability is often quantified by an index of detrended north Atlantic sea-surface temperatures, either including (Enfield et al. 2001) or excluding (van Oldenborgh et al. 2009) the tropics, and termed the Atlantic Multidecadal Oscillation (AMO). The internal multidecadal pattern in near-global sea-surface temperature found by Delsole et al. (2011) is very similar to Enfield et al.'s AMO index. Enfield et al. (2001) detrended relative to time, whereas van Oldenburgh detrended relative to surface temperature. While following van Oldenburgh et al. in excluding the tropics (which are more affected by ENSO state than the extratropics), we prefer detrending relative to total forcing, omitting volcanic years, in order to exclude any forced signal. Whichever definition is used, the AMO has had a quasi-periodicity of 60−70 years during the instrumental record, peaking around 1875, 1940 and 2005. When using a final period ending in 2016, to maximise the anthropogenic warming signal, matching its mean AMO state requires a base period either early in the historical period or in the mid-twentieth century.

Matching mean ENSO state for base and final period levels is not practical where a base period early in the record is used, since the mean ENSO state, as represented by the MEI index (Wolter and Timlin 1993), was lower then than in recent decades. However, the MEI index depends partly on non-detrended sea-surface temperature (SST) and could include a forced element, so use of a detrended version is arguably preferable. On that basis, there is no difficulty in matching mean ENSO state. In any event, of the natural sources of influence on sensitivity estimation considered, mean ENSO state appears to be the least influential.

LC15 used base periods of 1859−1882, 1930−1950 and 1850−1900. LC15's preferred base and final periods were 1859−1882 and 1995−2011, being the longest periods near the start and at the end of the instrumental record with low volcanic activity and with adequately matched AMO influence. As volcanic activity has remained low since 2011, the obvious choice of updated final period is 1995−2016. This includes a number of relatively cold years but also two very strong El Niños. The decade 2007−2016, which includes a mix of cold and warm years and ends with a powerful El Niño, is arguably preferable as it provides a higher ΔF and the best constrained TCR and ECS estimates. Moreover, as the Argo network was operational throughout 2007−2016, confidence in the reliability of OHU estimation is higher.

Although 1859−1882 is well matched with both 1995−2016 and 2007−2016 as regards mean volcanic forcing, and acceptably matched for mean AMO state, HadCRUT4v5 observational data sampled a particularly low proportion of the Earth's surface throughout most of the 1860s – substantially lower than both prior to 1860 and from 1869 on. During the same period, larger than usual differences arose between the original HadCRUT4v5 and the globally complete Had4_krig_v2 surface temperature estimates. Infilling through kriging is subject to greater uncertainty when observations are sparser. There is merit in using the longer 1850−1882 period, excluding all years with low (under 20% of global area) HadCRUT4v5 coverage (being 1860−1868); however, as volcanic forcing was strong (below –0.5 Wm^{-2}) over 1856−1858 those years would also need to be excluded to avoid mismatched volcanic forcing. Since the complete shorter 1869−1882 period produces essentially identical TCR and

ECS estimation we use that instead. It is well matched with the 1995−2016 and 2007−2016 final periods as regards mean volcanic forcing as well as AMO and ENSO state. The better observed 1930−1950 period is also well matched with those final periods, although its mean AMO state is stronger.

TCR and ECS estimates are also computed using much longer base and final periods. The 1850−1900 long base period, taken in AR5 to represent pre-industrial surface temperature, has substantial mean volcanic forcing. It is matched with 1980−2016, which has almost identical mean volcanic forcing and acceptably similar mean AMO and ENSO states.

Figure 3 shows variations in the three sources of natural variability discussed, along with areal coverage of HadCRUT4v5. Five year running means are shown for the MEI and AMO indexes.

5. Methods

The method used to calculate ECS and TCR is identical to that in LC15, where it is set out in detail. In summary, the main steps in deriving best estimates and uncertainty ranges for ECS and TCR for each base period and final period combination are as follows:

1) Unrevised AR5 2011 values for each forcing component (ERF_{GHG}, $ERF_{Aerosol}$, ERF_{BCsnow}, $ERF_{Contrails}$, ERF_{OWL}, ERF_{Solar} and $ERF_{Volcano}$) are sampled, using the original AR5 uncertainty distributions except for aerosol forcing. For aerosol forcing a normal distribution with unchanged −0.9 Wm^{-2} median but the revised −0.1 to −1.7 Wm^{-2} 5−95% uncertainty range is used. Where appropriate, part of fractional-type uncertainty in a forcing component (being all but any fixed element) is treated as independent between the base and final periods, and the total uncertainty is split between separate common and independent random elements before sampling. The AR5 efficacy range for ERF_{BCsnow} is applied probabilistically at this stage. After

dividing by the AR5 2011 best estimates, the (one million) samples are used to scale the period means computed from the best estimate time series (revised from AR5 where relevant), samples from the fixed elements of solar and volcanic forcing uncertainty are added and the components combined, thus deriving sampled ΔF values. The central $F_{2\times CO2}$ value is scaled in the same proportion as the central ERF_{GHG} values. This produces $F_{2\times CO2}$ samples with uncertainty realizations (proportionately) matching those for WMGG forcing.

2) Uncertainty distributions for ΔT (using the relevant 100-realizations ensemble) and for ΔN are computed, adding in quadrature the estimated uncertainties of the base and final period means and the estimated internal variability, and random samples drawn from those distributions.

3) For each sample realization of ΔT, ΔF, ΔN and $F_{2\times CO2}$, the ECS and TCR values given by equations (4) and (5) are calculated. Histograms of the sample ECS and TCR values are then computed to provide median estimates, uncertainty ranges and probability densities, treating samples where the denominator is negative as having infinitely positive sensitivities.

The estimates of ΔT, ΔF and ΔN, and their uncertainty ranges, are given in Table 2, with the relevant corresponding values from LC15 shown for comparison.

6. Results

ECS and TCR estimates based on each of the four combinations of base period – final period are presented in Table 3. The ECS estimates in this Section assume that the climate feedback parameter over the historical period, which they reflect, is a constant. That is, they measure effective climate sensitivity but assume it equals equilibrium climate sensitivity; the possible implications of relaxing this assumption are discussed in Section 7f. The relevant results from LC15 are shown for comparison. Estimates based on both original HadCRUT4v5 surface temperature data and on the globally-complete Had4_krig_v2 version are given. Probability density functions (PDFs) for these ECS and TCR estimates are presented in Figure 4.

For each source of surface temperature data, the four best (median) estimates agree closely for both ECS and TCR. Based on HadCRUT4v5 data, the best estimates are in the range 1.50−1.56 K for ECS and 1.20−1.23 K for TCR. Based on globally-complete Had4_krig_v2 data, which show greater warming, the best estimates are in the range 1.65−1.69 K for ECS and 1.27−1.33 K for TCR. Lower (5%) uncertainty bounds for ECS and TCR vary little between the four period combinations. Use of 1869−1882 as the base period and 2007−2016 as the final period provides the best-constrained, preferred, estimates, with 95% bounds for ECS and TCR of 2.45 K and 1.7 K respectively using HadCRUT4v5 (2.7 K and 1.9 K using Had4_krig_v2); the corresponding median estimates are 1.50 K and 1.20 K (Had4_krig_v2: 1.66 K and 1.33 K).

The new ECS and TCR median estimates based on HadCRUT4v5 are approximately 10% lower than those in LC15, largely due to the positive revisions to estimated CH_4 and post-1990 aerosol forcing, partly offset by the higher estimated $F_{2\times CO2}$ and (for ECS) by estimated heat uptake in the final period being a slightly higher fraction of forcing.

Results of some sensitivity analyses are shown in Table 4, with various aspects of the 1869−1882 base period, 2007−2016 final period case being modified. These analyses do not systematically explore all possible variations in choice of data, uncertainty assumptions or methodology. For clarity, only values based on HadCRUT4v5 surface temperature data are shown; fractional sensitivities are similar using Had4_krig_v2 data.

Using 1850−1882 as the base period, with low observational coverage and volcanic years excluded, produces virtually identical ECS and TCR medians and uncertainty ranges to using 1869−1882. Generally, estimates of ECS and TCR are modestly sensitive to selection of base period if no allowance is made for volcanic forcing (as estimated in AR5) having a low efficacy; when its efficacy is taken as 0.55 the ECS and TCR best estimates are little changed upon substituting 1850−1900 or 1850−1882 (all years) as the base period. Moreover, applying a volcanic forcing efficacy of 0.55 when regressing surface temperature per HadCRUT4v5 on (efficacy-adjusted) forcing over all years in 1850−2016 produces a TCR estimate of 1.19 K, almost identical to the two-period estimate. By comparison, doing so using unit volcanic efficacy gives a much lower TCR value of 0.98 K.

The residuals from regressing surface temperature per HadCRUT4v5 on efficacy-adjusted forcing over 1850−2016 with volcanic efficacy set at 0.55 have a mean over 2007−2016 only 0.01 K higher than that over 1869−1882 (0.03 K higher using Had4_krig_v2 data). For the 1995−2016 final period the corresponding excesses are similar. The tiny magnitudes of these inter-period differences indicate that both final periods are well matched with the 1869−1882 base period as regards internal variability.

Reverting the aerosol forcing 5% uncertainty bound back from –1.7 Wm^{-2} to the original AR5 –1.9 Wm^{-2} level increases the 95% bounds for ECS and TCR by respectively 0.2 K and 0.1 K; their median estimates barely change. Scaling up by 50% the uncertainty range for ERF_{WMGG} increases those bounds by 0.15 K and 0.05 K respectively, while doing so for ERF_{OWL} increases them by 0.1 K and 0.05 K respectively; scaling down these uncertainty ranges by 50% has approximately equal but opposite effects. Reducing the aerosol forcing uncertainty range by 50% reduces the 95% bound for ECS by 0.3 K to 2.15 K and that for TCR by 0.15 K to 1.55 K.

Using unrevised AR5 forcing–concentration relationship estimates for the principal WMGG and for post-1990 aerosol and tropospheric ozone forcing results in the ECS and TCR median values increasing by 0.18 K and 0.11 K respectively. The 95% uncertainty bounds for ECS and TCR increase more, by 0.8 K and 0.35 K respectively, but remain well below their levels in LC15. In contrast, computing 0−2000 m OHU using only Cheng et al., or only Levitus et al., data instead of using estimates averaged over those datasets (and, for the 0–700 m layer, the Ishii and Kimoto dataset), affects ECS best estimates by merely ±2−3%, with the 95% bound altering by ±0.1 K; TCR estimates are unaffected.

7. Discussion

Since publication of LC15, various papers have claimed that the energy budget approach and/or temperature dataset used in LC15 do not enable ECS and TCR to be determined satisfactorily from historical observations, and lead to the LC15 estimates being biased low. Here we address these critiques, as well as implications of feedback analysis and research concerning SST warming patterns.

a. Role of historical sea-surface temperature warming patterns

The pattern of observed surface warming over the historical period differs from that simulated by most CMIP5 models. Gregory and Andrews (2016) (GA16) found that feedback strength λ in simulations by two atmosphere-only models (AGCMs), HadGEM2-A and HadCM3-A, driven by observed evolving changes in SST and sea-ice, but with preindustrial atmospheric composition and other forcings fixed (amipPiForcing), was considerably higher over the historical period than in years 1–20 of abrupt4xCO2 simulations. Moreover, λ showed substantial decadal variation, being particularly large over the post-1978 period. Zhou et al. (2016) found broadly similar behavior in two other AGCMs.

We focus here on GA16's amipPiForcing simulation data from the more advanced, current generation HadGEM2 model. GA16's analysis of variation in λ (their $\tilde{\alpha}$) measured by regression over a 30-year sliding window, with small temperature changes except towards the end, is not relevant to energy budget estimation spanning much longer periods and larger changes. Moreover, GA16’s analysis method produces large variability in λ estimates when tested on pseudodata embodying a constant λ (Figure S2).

Plotting ΔR against ΔT using pentadal means, averaging-out interannual noise, and considering how averages over consecutive longer periods compare (Figure 5a), provides a more suitable assessment of the stability of feedback strength in HadGEM2-A over the historical period. Over the last 75 years, during which over 80% of the total forcing change occurred, ΔR and ΔT pentadal anomalies are clustered around the best-fit line, with means for all five 15-year sub-periods lying very close to it. There are a few pentadal points some distance from the best fit line, as one would expect from internal variability, but little evidence of fluctuating multidecadal feedback strength. The largest excursions of ΔR from the best-fit λ estimate of 1.90 $\mathrm{Wm^{-2}K^{-1}}$ were in the 20 years prior to 1925 and in the decade centered on 1980 (Figure S3).[2] The latter was responsible for the strong 1970−1995 upwards trend in 30-year regression-based λ in GA16 Figure 2(a); if the 1976−1985 ΔR values are suitably adjusted, the trend is flat from 1960 on (Figure S4). However, the anomalous ΔR values circa 1980 have only a minor effect on λ estimates derived from 15-year means: for both 1966−1980 and 1981−1995, ΔR / ΔT was only 7% lower than for 1996−2010. The early heavy volcanism (during 1883–1905) appears not to have affected the best-fit λ: the ratio of changes in R and T between 1931−1960 and 1996−2010, two volcanism-free periods, gives

[2] The first excursion is cotemporaneous with a period of strongly negative SST anomalies in the North Atlantic and reconstructed salinity anomalies in the Labrador Sea (Muller et al. 2014). The second excursion is cotemporaneous with decadal variability linked to the 1976 Pacific climate shift (Trenberth and Hurrell 1994). Both events likely arose from multidecadal internal variability; there is little evidence of either being forced.

almost the same value. Fits for each individual amipPiForcing run are very similar (negligible *y*-intercept, slopes within 5% of the 1.90 $Wm^{-2}K^{-1}$ for the ensemble-mean, $R^2 = 0.93$ versus 0.94 for the ensemble-mean, in all cases with 1906−25 data excluded). This analysis shows that HadGEM2-A displays a near constant λ of 1.9 $Wm^{-2}K^{-1}$ over the historical period when driven by observed evolving SST patterns – over 2.3× as high as the 0.82 W $m^{-2}K^{-1}$ over years 1−20 of HadGEM2-ES's abrupt4xCO2 simulation, and corresponding to an effective climate sensitivity of only 1.67 K.[3]

GA16 offered three possible explanations for feedback strength being higher over the historical period in their amipPiForcing experiments than over years 1−20 of the abrupt4xCO2 simulations. They found two of them conflicted with their calculated trends in λ, leading them to favour the importance of the third explanation, being that unforced variability strongly influenced historical variations in SST patterns. However, Zhou et al. (2016) found that if CMIP5 control simulations realistically estimate internal variability on decadal timescales, then at least part of the 1980-2005 SST trend pattern must be forced. In HadGEM2's case, under 1% of internal variability realizations simulated by CMIP5 AOGCMs would raise the ΔR value for the final 15 years of the amipPiForcing run implied by the λ value HadGEM2-ES exhibits early in its abrupt4xCO2 simulation even 30% towards its actual amipPiForcing value (Figure S5). Our finding that the relationship between pentadal ΔR and ΔT in HadGEM2-A during its amipPiForcing experiment is stable, apart from two excursions, (Figures 5a and S3) strongly points to the observed SST pattern evolution being largely forced and to much lower λ values in years 1−20 of HadGEM2-ES's abrupt4xCO2 experiment reflecting unrealistic simulated SST pattern evolution. If follows that there is no reason to believe energy budget sensitivity estimates based on changes over the full historical period are biased downwards by internal variability in SST patterns.

[3] Based on our estimated $F_{2\times CO2}$ for HadGEM2-ES of 3.18 Wm^{-2}.

Observational estimates of the relationship between ΔR and ΔT throughout the historical period are also relevant. We estimate λ using all 15-year periods in 1927−2016, as well as by regression over 1872−2016, anomalizing relative to a 1850−1884 base period. Average volcanism in 1850−1884 matches that over both 1927−2016 and 1872−2016, and when using 2007−2016 anomalies that base period gives the same λ estimate (2.29 $\mathrm{Wm^{-2}K^{-1}}$, corresponding to an ECS of 1.66 K) as per the main 2007-2016 based results with globally-complete ΔT. Until recent decades ΔR was unobserved; we approximate it by scaling ΔF *pro rata* to the observationally-estimated ΔR:ΔF ratio for 1869−1882 to 2007−2016, assuming that ΔN is proportional to ΔT over the historical period (Gregory and Forster 2008). We scale $\mathrm{ERF_{Volcano}}$ by 0.55 to adjust for its low efficacy. Our no-intercept pentadal regression fit over 1872–2016 gives λ = 2.27 $\mathrm{Wm^{-2}K^{-1}}$. Post-1926 (ΔR, ΔT) pentadal means (Figure 5b) cluster around the best-fit line, while most of the 15-year means lie almost on it.

The considerable stability of observationally-based λ estimates over 1927−2016 provides further evidence that feedback strength did not fluctuate materially during the historical period, and strengthens confidence in our main results.

b. Weaknesses in the feedback analysis constraint

It has been argued that relatively well understood feedbacks (water vapor/lapse rate, albedo) imply, in the absence of evidence for cloud feedbacks being significantly negative, an upper bound on the climate feedback parameter corresponding to ECS being 2 K or higher, particularly if anvil cloud-height feedback is also included. However, an analysis of feedbacks and forcing in CMIP5 models (Caldwell et al. 2016) indicates that if diagnosed cloud feedbacks are excluded, the median implied ECS reduces from 3.4 K to 2.3 K, with ECS falling below 2 K in a quarter of the models. More fundamentally, the fact that AGCMs can generate widely varying climate feedback strength depending on the pattern of SST change (which feedback analysis does not constrain) weakens the feedbacks constraint argument.

A substantial part of the initial radiative response to CO_2 forcing may be viewed (and mathematically modeled) as reflecting a sub-decadal timescale ocean adjustment process during which ocean heat transport and SST patterns alter, negatively affecting shortwave cloud radiative effect (Andrews et al. 2015) so that R increases for a given T, thus partially counteracting the forcing independently of surface temperature increase (Williams et al. 2008; Sherwood et al. 2015; Rugenstein et al. 2016). Feedback analysis derived constraints, even if correct, do not apply to such an adjustment. Accordingly, as during the initial decade or two the radiative response partly reflects adjustments, the *apparent* climate feedback parameter may be considerably higher then than feedback analysis suggests is possible. While the (lower) underlying climate feedback parameter if not affected by adjustments, and may be time-invariant, ECS is affected.

In abrupt4xCO2 simulations, where diagnosed climate feedback strength is typically substantially greater in the first decade or two than subsequently, eigenmode decomposition of CMIP5 AOGCM responses (Proistosescu and Huybers 2017) indicates that only about one-third of the initial forcing remains once sub-decadal timescale responses are complete, and that the climate feedback parameter associated with sub-decadal timescale responses, if not regarded as partially associated with adjustment processes, ranges up to 3 $Wm^{-2} K^{-1}$.

We conclude that simple global feedback analysis cannot rule out low ECS even if global cloud feedback is ultimately positive, because radiative response, forcing adjustments and feedbacks depend on the pattern of SST warming, which may differ significantly from that simulated by AOGCMs.

c. ERF efficacy

There have been suggestions that the composite forcing during the historical period has an overall ERF efficacy below one, so that historical forcing will have produced less warming than CO_2 forcing of equal ERF magnitude (Shindell 2014; Kummer and Dessler 2014; Marvel et al. 2016). In most cases, the shortfall is attributed principally to spatially inhomogeneous negative aerosol forcing having an efficacy exceeding one. Using historical all-forcings, WMGG-only and natural forcings-only simulations by a small ensemble of CMIP5 models, Shindell estimated that aerosol ERF – combined with the much smaller ozone ERF – had an efficacy of 1.5, resulting in the (transient) efficacy of historical ERF being approximately 0.85. Kummer and Dessler showed that applying Shindell's aerosol and ozone ERF efficacy estimate increased their ECS estimate by 50%.

Marvel et al., using the GISS-E2-R model and a set of single-forcing simulation-ensembles as well as a historical all-forcings simulation-ensemble, with the applicable ERF determined from a further set of simulation-ensembles, estimated historical composite ERF to have transient and equilibrium efficacies below one; we discuss these findings below. However, they found that these shortfalls were due to solar, volcanic, ozone and (for equilibrium efficacy) WMGG ERF having an efficacy below one, with aerosol ERF having an efficacy of 1.0. Other single forcing simulation studies also indicate that aerosol ERF does not have an efficacy exceeding one (Hansen et al. 2005; Ocko et al. 2014; Paynter and Frölicher 2015; Forster 2016). Although Rotstayn et al. (2015) obtained an aerosol ERF efficacy estimate of 1.4 by regressing surface temperature change over the historical period against estimated aerosol ERF in an ensemble of CMIP5 models, their result is strongly model-ensemble dependent. Excluding an outlier model (FGOALS-s2) makes their efficacy estimate statistically indistinguishable from one.

Complicating matters, for aerosols the forcing and response may vary significantly with climate state (Miller et al. 2014; Nazarenko et al. 2017).Shindell (2014) (and thereby Kummer and Dessler 2014) and Marvel et al. (2016) estimated aerosol ERF using model simulations in which the climate state differed from that when composite historical forcing was applied, so their results are unreliable in the presence of aerosol forcing or response climate-state dependency. As Shindell differenced results from forced simulations involving different climate states and forcing combinations, his findings (and thereby Kummer and Dessler's) are particularly susceptible to bias from aerosol forcing or response climate-state dependency.

Efficacy estimates based purely on composite historical forcing may be more reliable. Marvel et al. estimated the efficacy (their transient efficacy) of composite historical instantaneous radiative forcing at the tropopause (iRF, an approximation to RF) as 1.00. Although their corresponding ERF (transient) efficacy estimate, which is more relevant to energy-budget studies, was 0.88, they derived it by comparing year-2000 forcing with mean 1996−2005 temperatures, which does not produce a satisfactory estimate. In GISS-E2-R, year 2000 forcing was higher than the 1996−2005 mean, and surface temperature in the second half of the 1990s was still depressed by recovery from the Pinatubo eruption (Table S1). Recalculating efficacy using warming over 2000−2005, scaling year-2000 historical ERF by the ratio of average 2000−2005 iRF to year-2000 iRF, raises the Marvel et al. (transient) efficacy of historical ERF to 1.00 (Supplementary Material: S3). Consistent with this, Hansen et al. (2005) estimated (transient) efficacy relative to historical ERF derived by regression as marginally above one.

Marvel et al. also derived a new efficacy metric, equilibrium efficacy, that accounts for variation in heat uptake efficiency between forcings. However, their methods also bias downwards their historical forcing equilibrium efficacy estimates. Recalculating equilibrium efficacy for historical ERF using the same mean 2000−2005 historical ERF value as for our re-estimation of transient efficacy, and the full TOA radiative imbalance rather than just its ocean heat uptake component, raises their 0.76 equilibrium efficacy estimate to 1.04 when the comparison is made with the response to CO_2-only forcing over a similar time period (Supplementary Material: S3).

Hence we conclude that assertions that historical forcing has an efficacy below one appear to be unjustified, so that the assumption of λ being independent of forcing composition holds for the change in composite forcing over the historical period (of which the volcanic component is negligible).

d. Global incompleteness of the surface temperature dataset

In principle a globally-complete surface temperature dataset is preferable, although the potential inaccuracy introduced by infilling might be greater than estimated, particularly in the early part of the record. Even during the well-observed satellite period, it is not invariably true that infilling is beneficial. ECMWF (2015) gives a global-mean comparison over 1979−2014 of 2 m air temperature for land and SST for ocean per ERA-interim (Dee et al. 2011) – generally considered the best reanalysis dataset – both on a globally-complete basis and with monthly coverage reduced to match that of HadCRUT4. The 1979−2014 linear trend of their globally-complete estimates was closely in line with that based on HadCRUT4 coverage (which equaled the actual HadCRUT4v5 trend), whereas Had4_krig_v2 shows a 9% higher trend over that period.

Nevertheless, it is more appropriate to use sensitivity estimates based on globally-complete surface temperature data for comparisons with CMIP5 model ECS and TCR values and others based on globally-complete data. We use only our Had4_krig_v2-based estimates for doing so.

e. Use of anomaly temperatures and SST versus air temperature over the oceans

Using CMIP5 model simulations, it has been claimed (Cowtan et al. 2015; Richardson et al. 2016) that even a globally-complete surface temperature estimate like Had4_krig_v2 may understate warming in global mean near-surface air temperature due to its use of SST over the ice-free ocean and of anomaly temperatures. Richardson et al. (2016) estimated a historical bias of 7−9% if real-world behavior matched that of the average CMIP5 model. They refer to the related discussion by Cowtan et al. (2015), who estimated an average bias of 7% for historical warming (their Table S1, averaging all periods with >0.2 K warming). Two causes each contributed approximately half of the 7% bias.

First, Cowtan et al. argued that temperature changes in areas becoming free of sea ice, as it shrinks, are understated due to the use of anomalies. However, CMIP5 model simulations cannot provide a realistic estimate of any resulting bias in historical warming, since most models simulate strong warming in Antarctica and a reduction in surrounding sea ice, whereas little Antarctic warming has occurred and sea ice there has actually increased. Cowtan and Way (2014: update) found that in reality the effect on temperature estimates of assuming sea ice extent was fixed (in which case no bias arises) was minimal.

Secondly, Cowtan et al. argued that in CMIP5 models SST (*tos*) warms less than ocean near-surface air temperature (*tas*), resulting on average in surface temperature warming less when SST rather than marine air temperature is used. However, CMIP5 models generally treat the ocean's skin temperature, which determines its interactions with the atmosphere, as equal to the top model ocean layer, typically 10 m deep, so that *tas* – *tos* really reflects the difference between model-simulated air temperature and ocean skin temperature. Even if the excess of near-surface air temperature increase over ocean skin temperature increase in CMIP5 models is realistic, SST, which is typically measured at 5−10 m deep, is significantly different from skin temperature and may increase faster. Observations provide alternative evidence. The HadNMAT2 (Kent et al. 2013) dataset shows a lower global trend in near-surface marine air temperature over its 1880−2010 record than does HadSST3.1.1.0, the sea-surface temperature component of HadCRUT4v5, although possible inhomogeneities mean this result is uncertain. Moreover, the 1979−2014 trend in the globally-complete ERA-interim data increases by just 2% when using background 2 m marine air temperature (calculated by the reanalysis AGCM) rather than SST (ECMWF 2015).[4] Over 1979 to July 2016 – a period in which the bulk of the historical period warming and sea-ice reduction occurred – ERA-interim shows marginally greater warming when using background marine air temperature rather than analyzed SST, but the trend is 0.17 K/decade in both cases – and lower than the 0.18 K/decade per both the SST-using HadCRUT4v5 and Had4_krig_v2 datasets (Simmons et al. 2017).

On balance the observational evidence points to past warming in global mean temperature when using near-surface air temperature everywhere being little different from when blending it with SST over the ocean. The evidence from comparing ERA-interim trends using marine air temperature and using SST, which points to approximately 2% slower warming when using SST, is perhaps most credible. However, this excess is tiny, and could be biased high by the reanalysis AGCM's behavior.

We conclude that any underestimation of past global near-surface air temperature warming arising from blending SST data over the ice-free ocean with near-surface air temperature elsewhere, as in Had4_krig_v2, is sufficiently small to be ignored (and could

[4] Digitizing the complete global averages data in the ECMWF (2015) bar graph gives a 1979−2014 trend of 0.158 K decade^{-1}, or 0.159 K decade^{-1} when masked to HadCRUT4 coverage. This data is a blend of 2 m temperature over land and SST over ocean (Paul Berrisford, ECMWF, pers. comm. 2016). A 2% higher 1979−2014 trend of 0.161 K decade^{-1} was computed using data from https://climate.copernicus.eu/sites/default/files/repository/Temp_maps/Data_for_month_8_2017_plot_3.txt. That data is for surface air temperature anomalies. Both sets of data have been adjusted by ECMWF for inhomogeneities in their source of analyzed SST.

even be negative). While incorporating an extra, multiplicative, uncertainty with a standard deviation of 4% in all the Had4_krig_v2 ΔT values might nevertheless be justified, it would not alter any ECS or TCR 5-95% uncertainty range by more than ±0.01 K.

f. ECS versus ECS$_{hist}$

The possibility that energy-budget climate sensitivity estimates based on changes over the historical period, which measure λ over that period and assume it is invariant (and which thus actually reflect an effective climate sensitivity, ECS$_{\text{hist}}$), might differ from ECS was brought up in section 2. ECS$_{\text{hist}}$ can be quantified fairly accurately in AOGCMs, their ECS estimated from centennial model response in abrupt4xCO2 simulations, and an ECS-to-ECS$_{\text{hist}}$ ratio derived.

We have calculated an ECS-to-ECS$_{\text{hist}}$ ratio for an ensemble of 31 CMIP5 models, deriving ECS by Gregory-plot regression (Gregory et al. 2004) over years 21−150 (Armour 2017) and taking the mean ECS$_{\text{hist}}$ estimate from three methods that access different realisations of model internal variability (Supplementary Material: S4). The three methods provide almost identical ensemble-mean ECS$_{\text{hist}}$ estimates (Table S2). Over the entire ensemble, ECS varies between 0.91× and 1.52× ECS$_{\text{hist}}$, the median ratio being 1.095, very close to the 1.096 ratio estimated by Mauritsen and Pincus (2017). Armour (2017) and Proistosescu and Huybers (2017) reported higher ECS-to-ECS$_{\text{hist}}$ ratios (respectively 1.26 ensemble-mean and 1.34 ensemble-median), but we find their estimation methods less satisfactory, causing quantifiable biases.

A reconciliation of the mean ECS-to-ECS$_{\text{hist}}$ ratio for CMIP5 per Armour (2017) (A17) to our 1.095 ratio is as follows. We provide a similar reconciliation for Proistosescu and Huybers (2017) in the Supplementary Material (S5).

1. A17, in calculating ECS_{hist} values (there termed ECS_{infer}), estimated $F_{2\times CO2}$ from the *y*-axis intercept when regressing ΔN against ΔT over years 1−5 of abrupt4xCO2 simulations. Doing so does not provide an unbiased ERF basis estimate of $F_{2\times CO2}$, since during year one CO_2 top-of-atmosphere forcing is moving from its instantaneous value towards its ERF value as the stratosphere, troposphere and other annual- or shorter-timescale climate system components adjust to the imposed forcing independently of surface temperature increase. For example, stratospheric adjustment, which reduces forcing, takes several months to complete. When regressing over only five years, the inclusion of year one data significantly increases the mean $F_{2\times CO2}$ estimate, resulting in lower ECS_{hist} estimates. We regress over years 2−10, avoiding bias from non-fully adjusted year one data; time-variation of λ is insignificant in the first decade. A17's ensemble-mean ECS-to-ECS_{hist} ratio calculated using regression over years 2−10 to determine $F_{2\times CO2}$, but otherwise using his methods, would be 1.215.
2. A17 did not allow for the slightly faster than logarithmic relationship of CO_2 forcing to concentration (Etminan et al. 2016). There is no reason to think that CO_2 radiative forcing code in CMIP5 models does not, on average, reflect that relationship – the logarithmic relationship given in AR5 was known only to be an approximation. The effect is a 0.7% upwards bias in A17's mean ECS_{hist} estimate (which is based on ΔN and ΔT values in years 85−115 of 1pctCO2 simulations) but a 4.6% upwards bias in A17's mean ECS estimate (which is based on abrupt4xCO2 simulations). Adjusting for this bias (3.9% net) reduces A17's ECS-to-ECS_{hist} ratio estimate further, to 1.170.
3. A17 estimate ECS using OLS regression of annual-mean years 21−150 abrupt4xCO2 ΔN and ΔT values, but ΔT as well as ΔN is affected by internal variability and their

fluctuations are generally weakly correlated. Where the regressor variable contains errors, OLS regression underestimates the slope coefficient (Deming, 1943). Using Deming regression to derive unbiased ECS estimates (Supplementary Material S4), A17's ensemble-mean ECS estimate is 2.0% lower than when using OLS regression. Adjusting for this bias further reduces A17's ensemble-mean ECS-to-ECS_{hist} ratio, to 1.146.

4. We use three different methods to estimate ECS_{hist}, one being A17's method, averaging their results. For A17's ensemble our mean ECS_{hist} estimate is the same as when using only A17's method, so using our ECS_{hist} estimation basis its mean ECS-to-ECS_{hist} ratio is also 1.146.
5. A17 quote a mean ECS-to-ECS_{hist} ratio, but since the distribution is skewed it is appropriate to use the median, a robust and parameterization-independent measure, as the central estimate. The ensemble-median A17 ECS-to-ECS_{hist} ratio, using our ECS_{hist} calculation basis, is 1.115, lower than the 1.146 mean ECS-to-ECS_{hist} ratio.
6. A17 use a smaller ensemble of CMIP5 models (21 rather than our 31), which disproportionately excludes models with low ECS-to-ECS_{hist} ratios. For our ensemble, the median ECS-to-ECS_{hist} ratio using our calculation basis is 1.095.[5]

The ECS-to-ECS_{hist} ratio in CMIP5 models should vary positively with ECS_{hist} (Armour 2017); it tends to do so and is generally moderate (≤ 1.16) where ECS_{hist} is under 2.9 K, although a linear fit has little explanatory power. We derive a probabilistic estimate for ECS that reflects behavior of CMIP5 models by scaling our globally-complete Had4_krig_v2-based energy-budget ECS_{hist} estimate using CMIP5 model ECS-to-ECS_{hist} ratios, binned (0.2 K width) by ECS_{hist}. We allocate the million sample observationally-based ECS_{hist} estimates between the bins and scale them by the ECS-to-ECS_{hist} ratios of models in each bin, taking

[5] If our ensemble were equally weighted by modeling center, the median ECS-to-ECS_{hist} ratio would be 1.082.

models from the nearest bin(s) where the ECS_{hist} bin is empty and allocating samples falling in each bin equally between the applicable models. The resulting ECS median estimate is 1.76 K (5–95%:1.2–3.1 K). Scaling the median of our energy-budget ECS_{hist} estimate by the 1.06 median ECS-to-ECS_{hist} ratio for the 14 CMIP5 models with an ECS_{hist} value within its 1.15−2.7 K uncertainty range likewise produces a 1.76 K median ECS estimate. A 3.1 K 95% uncertainty bound for ECS also results if the million sample Had4_krig_v2-based ECS_{hist} estimates are scaled using the CMIP5 ensemble-median ECS-to-ECS_{hist} ratio of 1.095 with normally-distributed uncertainty added to give a 0.79–1.40 5−95% range.

The upper bound generated for ECS is not necessarily robust; the joint distribution of ECS and ECS_{hist} in CMIP5 models may not be a realistic enough measure of uncertainty in the ECS-to-ECS_{hist} ratio, nor is not known how accurately ECS can be estimated from 150 year abrupt4xCO2 simulations. If the 95% uncertainty bound for ECS_{hist} estimated using Had4_krig_v2 data, 2.7 K, were scaled up by the highest ECS-to-ECS_{hist} ratio among CMIP5 models with an ECS_{hist} below 2.85 K, the ECS upper bound would be 3.4 K. However, much of any excess of ECS over ECS_{hist} would take centuries to be realised in surface warming, with little effect on warming in 2100. Twenty-first century warming arising from future forcing increases will largely be determined by TCR, with any excess of ECS over ECS_{hist} being almost irrelevant. Even if the highest ECS-to-ECS_{hist} ratio found in CMIP5 models applied, warming in 2100 due to the past increase in forcing would be only 0.1 K greater than if ECS equaled ECS_{hist} (Mauritsen and Pincus 2017).

Observationally-based evidence of the ECS-ECS_{hist} relationship can be obtained by comparing historical-period energy budget sensitivity estimates with those based on past changes between equilibrium climate states (implying zero ΔN), using proxy paleoclimate data. However, uncertainties in forcing and temperature changes are considerably greater for past periods, particularly for more remote periods, and climate feedbacks might have been considerably different then. The most recent and best studied such change is that from the last glacial maximum (LGM) to the preindustrial Holocene. It is not obvious that ECS for the LGM transition should be lower than from preindustrial conditions, and an energy-budget approach has long been applied to estimate ECS from this period. Although Goelzer et al. (2011) found that the LGM-transition ECS could be reduced by melting ice sheets, the effect was minimal when estimated ECS was below 2.5 K.

Reasonably thorough proxy-based estimates of changes in surface temperature (Annan et al. 2013: 4.0 K; Friedrich et al. 2016: 5.0 K) and forcings (Kohler et al. 2010: total 9.5 Wm^{-2}) are available for the LGM transition. These values imply, using (4), an ECS estimate of 1.76 K, (averaging the two surface temperature increase estimates and taking $F_{2\times CO2}$ per AR5, since the WMGG forcings were derived using AR5 formulae), in line with the median obtained by scaling this study's ECS_{hist} estimate.

8. Conclusions

Using updated and revised data, we have derived ECS_{hist} and TCR estimates that are much better constrained, and slightly lower when using the same surface temperature dataset (HadCRUT4), than those in the predecessor LC15 study: 1.50 K median (5−95%: 1.05−2.45 K) for ECS_{hist} and 1.20 K median (5−95%: 0.9−1.7 K) for TCR. Using infilled, globally-complete temperature data (Had4_krig_v 2) slightly increases the new estimates, to a median of 1.66 K for ECS_{hist} (5−95%: 1.15−2.7 K) and 1.33 K for TCR (5−95%:1.0−1.90 K). We have also shown that various concerns that have been raised about the accuracy of historical period energy budget climate sensitivity estimation are misplaced. We assess nil bias from either non-unit forcing efficacy or varying SST warming patterns, and that any downwards estimation bias when using blended infilled surface temperature data is trivial. We find that high CMIP5 model-based estimates of the ratio of ECS to ECS_{hist}, the proxy for ECS that historical period based studies estimate, become far lower when calculated more appropriately. By using the ECS-to-ECS_{hist} ratios that we calculate for CMIP5 models to scale our Had4_krig_v2-based ECS_{hist} probability distribution, we derive a median estimate for ECS of 1.76 K (5−95%: 1.2−3.1 K).

Relative to LC15, most of the improvement in ECS estimation precision is due to higher greenhouse gas concentrations when using data to 2016 rather than 2011 and to the revisions to estimated CH_4 and post-1990 aerosol forcing. Forcing uncertainty remains the dominant contributor to the widths of the ECS and TCR ranges, and reducing the uncertainty in aerosol forcing would narrow them much more than reducing uncertainty in any other forcing component.

It is notable that the best estimates for both ECS and TCR are almost identical across all four combinations of base period and final period. This is consistent with a modest influence of shorter-term climate system internal variability and of measurement/estimation error on energy budget sensitivity estimates. The estimates using the 1869−1882 base period and 2007−2016 final period combination are preferred; they have the highest ΔT and ΔF values and as a result are best constrained. Moreover, with the Argo ocean-observing network fully operational throughout 2007–2016, there is also higher confidence in the reliability of the ocean heat uptake estimate when using that final period. Although HadCRUT4 observational coverage was modest during 1869−1882, the fact that TCR estimation is very similar using the higher-coverage 1930−1950 base period gives confidence in the ECS and TCR estimates using the former base period.

Over half of 31 CMIP5 models have best-estimate ECS_{hist} values of 2.9 K or higher, exceeding by over 7% our 2.7 K observationally-based 95% uncertainty bound using infilled temperature data. Moreover, a majority of the models have best-estimate TCR values above our corresponding 1.9 K 95% bound. A majority of the models also have best-estimate ECS values above our 3.1 K 95% bound. A simplified analysis (Table 5) based on considering in turn uncertainty only in ΔT, $\Delta F / F_{2\times CO2}$ and $\Delta N / F_{2\times CO2}$ (thus taking into account uncertainty in $F_{2\times CO2}$), confirms that in each case the lowest CMIP5 model TCR and ECS_{hist} values that we find to be inconsistent with observed warming imply, implausibly, that ΔT, $\Delta F / F_{2\times CO2}$ and $\Delta N / F_{2\times CO2}$ have values outside their uncertainty ranges.

The implications of our results are that high best estimates of ECS_{hist}, ECS and TCR derived from a majority of CMIP5 climate models are inconsistent with observed warming during the historical period (confidence level 95%). Moreover, our median ECS and TCR estimates using infilled temperature data imply multicentennial or multidecadal future warming under increasing forcing of only 55−70% of the mean warming simulated by CMIP5 models.

Acknowledgements. We thank Cheng Lijing for providing OHC data, updated to 2016, and three reviewers for helpful comments.

References

Andersson, S. M., et al., 2015: Significant radiative impact of volcanic aerosol in the lowermost stratosphere. *Nature communications*, **6**, 7692.

Andrews T, Gregory JM, Webb MJ, Taylor KE, 2012: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophys .Res. Lett.*, **39**:doi:101029/2012GL051607.

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.

Annan, J. D., and J. C. Hargreaves , 2013: A new global reconstruction of temperature changes at the Last Glacial Maximum. *Climate Past.*, **9**, 367–376.

Armour, K. C., and G. H. Roe, 2011: Climate commitment in an uncertain world. *Geophys. Res. Lett*. **38**:L01707.

Armour, K. C., 2017: Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. *Nature Climate Change*, **7**, 331-335.

Bender, F. A. M., A. Engström and J. Karlsson, 2016: Factors controlling cloud albedo in marine subtropical stratocumulus regions in climate models and satellite observations. *J Climate*, **29**(10), 3559-3587.

Bindoff, N. L., and Coauthors, 2014: Detection and Attribution of Climate Change: from Global to Regional. *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.

Byrne, B., and C. Goldblatt, 2014: Radiative forcing at high concentrations of well-mixed greenhouse gases. *Geophys. Res. Lett.*, **41**, 152–160, doi:10.1002/2013gl058456.

Caldeira, K., and N. P. Myhrvold, 2013: Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. *Environmental Res. Lett.* **8**.3: 034039.

Caldwell, P.M., M.D. Zelinka, K.E.Taylor, and K. Marvel, 2016: Quantifying the sources of intermodel spread in equilibrium climate sensitivity. *Journal of Climate*, ***29***(2), pp.513-524.

Charney, J. G., 1979: Carbon Dioxide and Climate: A Scientific Assessment. National Academies of Science Press, Washington, DC, 22 pp.

Cheng, L., K. Trenberth, J. Fasullo, T. Boyer, J. Abraham, and J. Zhu, 2017: Improved estimates of ocean heat content from 1960 to 2015. *Science Advances*, **3**, e1601545

Cowtan, K., and R. G. Way, 2014: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quart. J. Roy. Meteor. Soc*., ***140***(683), 1935-1944. (update at http://www.webcitation.org/6t09bN8vM; data at http://www-users.york.ac.uk/%7Ekdc3/papers/coverage2013/series.html)

Cowtan, K., and Coauthors, 2015: Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophys. Res. Lett.*, **42**, 6526–6534.

Cherian, R., J. Quaas, M. Salzmann, and M. Wild, 2014: Pollution trends over Europe constrain global aerosol forcing as simulated by climate models, *Geophys. Res. Lett*., **41**, 2176-2181, doi:10.1002/2013GL058715.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc*., **137**, 553–597

DelSole, T., M.K. Tippett, and J. Shukla, 2011: A significant component of unforced multidecadal variability in the recent acceleration of global warming. *J. Climate*, **24**, 909-926

Deming, W. E., 1943: Statistical Adjustment of Data. Wiley, NY (Dover Publications edition, 1985), 288 pp.

Desbruyeres, D., E. McDonagh, B. King, and V. Thierry, 2017: Global and Full depth Ocean Temperature Trends during the early 21 century from Argo and Repeat Hydrography. *J. Climate*, **30**(6), 1985-97.

ECMWF, 2015: ECMWF releases global reanalysis data for 2014. *European Centre for Medium-Range Weather Forecasts*. Archived at: http://web.archive.org/web/20180118202517/https://www.ecmwf.int/en/about/media-centre/news/2015/ecmwf-releases-global-reanalysis-data-2014-0.

Enfield D. B., A. M. Mestas-Nunez, and P. J. Trimble, 2001: The Atlantic Multidecadal Oscillation and its relationship to rainfall and river flows in the continental US. *Geophys. Res. Lett.* **28**:2077-2080.

Etminan, M., G. Myhre, E. J. Highwood, and K. P. Shine, 2016: Radiative forcing of carbon dioxide, methane, and nitrous oxide: A significant revision of the methane radiative forcing. *Geophys. Res. Lett.* **43**(24) doi:10.1002/2016GL071930.

Fiedler, S., B. Stevens, and T. Mauritsen, 2017: On the sensitivity of anthropogenic aerosol forcing to model-internal variability and parameterizing a Twomey effect, *J. Adv. Model. Earth Syst.*, **9**, doi:10.1002/2017MS000932.

Forster, P, 2016: Inference of Climate Sensitivity from Analysis of Earth's Energy Budget. *Annu. Rev. Earth Planet. Sci.*, **44**, doi: 10.1146/annurev-earth-060614-105156

Friedrich, T., A. Timmermann, M. Tigchelaar, O. E. Timm and A. Ganopolski, 2016: Nonlinear climate sensitivity and its implications for future greenhouse warming. *Science Advances*, **2**(11), e1501923.

Goelzer, H., P. Huybrechts, M. F.Loutre, H. Goosse, T. Fichefet and A. Mouchet, 2011: Impact of Greenland and Antarctic ice sheet interactions on climate sensitivity. *Climate Dynamics*, **37**(5-6), 1005-1018.

Good, P., J. M. Gregory and J. A. Lowe, 2011: A step-response simple climate model to reconstruct and interpret AOGCM projections. *Geophys. Res. Lett.*, **38**(1).

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

Gordon, H., and Coauthors, 2016: Reduced anthropogenic aerosol radiative forcing caused by biogenic new particle formation. *Proceedings of the National Academy of Sciences*, **113**, 43, 12053–12058

Gregory, J. M., R. J., Stouffer, S. C. B. Raper, P. A. Stott and N. A. Rayner, 2002: An observationally based estimate of the climate sensitivity. *J. Climate*, **15**:3117–3121

Gregory, J.M., and P. M. Forster, 2008: Transient climate response estimated from radiative forcing and observed temperature change. *J. Geophys. Res.*, **113**(D23).

Gregory, J. M., and Coauthors, 2013: Climate models without pre-industrial volcanic forcing underestimate historical ocean thermal expansion. *Geophys. Res. Lett.*, **40**(8), 1600–1604.

Gregory, J. M., and Coauthors, 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophys. Res. Lett.* 31.3, L03205.

Gregory, J. M., and T. Andrews, 2016: Variation in climate sensitivity and feedback parameters during the historical period. Geophys. Res. Lett., **43**: 3911–3920.

Hansen, J., and Coauthors, 2005: Efficacy of climate forcings. *J. Geophys Res.*, **110**: D18104, doi:101029/2005JD005776

Hobbs, W., M. D. Palmer, and D. Monselesan, 2016:. An energy conservation analysis of ocean drift in the CMIP5 global coupled models. *J. Climate*, **29**(5), 1639-1653

Ishii, M., and M. Kimoto, 2009: Reevaluation of historical ocean heat content variations with time-varying XBT and MBT depth bias corrections. *J. Oceanography*, **65**, 287-299.

Johnson, G. C., J. M. Lyman and N. G. Loeb, N. G., 2016: Improving estimates of Earth's energy imbalance. *Nature Climate Change*, **6**, 639-640.

Levitus, S., and Coauthors, 2012: World ocean heat content and thermosteric sea level change (0-2000m),1955-2010. *Geophys. Res. Lett.*, **39**:L10603

Lewis, N., and J. A. Curry, 2015: The implications for climate sensitivity of AR5 forcing and heat uptake estimates. *Climate Dynamics*, **45**(3-4), 1009-1023.

Lohmann, U., 2017: Why does knowledge of past aerosol forcing matter for future climate change? *J. Geophys. Res.*, **122**, 5021–5023.

Kent, E. C., and Coauthors, 2013: Global analysis of night marine air temperature and its uncertainty since 1880: The HadNMAT2 data set. *J. Geophys. Res.*, **118**, 1281–1298

Köhler, P., R. Bintanja, H. Fischer, F. Joos, R. Knutti, G. Lohmann, and V. Masson-Delmotte, 2010: What caused Earth's temperature variations during the last 800,000 years? Data-based evidence on radiative forcing and constraints on climate sensitivity. *Quaternary Science Reviews*, **29**, 129–145.

Kummer, J. R., and A. E. Dessler, 2014: The impact of forcing efficacy on the equilibrium climate sensitivity. *Geophys. Res. Lett.* **41**, 3565-3568.

McCoy, D. T., F. A.-M. Bender, J. K. C. Mohrmann, D. L. Hartmann, R. Wood, and D. P. Grosvenor, 2017: The global aerosol-cloud first indirect effect estimated using MODIS, MERRA, and AeroCom, *J. Geophys. Res. Atmos*.,**122**, 1779–1796,

Malavelle, F.F. and Coauthors, 2017: Strong constraints on aerosol–cloud interactions from volcanic eruptions. *Nature*, **546**, 485–491.

Marvel, K., G. A. Schmidt, R. L. Miller and L. S. Nazarenko, 2016: Implications for climate sensitivity from the response to individual forcings. *Nature Climate Change*, **6**(4), 386-389.

Masters, T., 2014: Observational estimate of climate sensitivity from changes in the rate of ocean heat uptake and comparison to CMIP5 models. *Climate Dynamics*, **42**,2173-2181 DOI 101007/s00382-013-1770-4.

Mauritsen, T., and R. Pincus, 2017: Committed warming inferred from observations. *Nature Climate Change*, 7(9), nclimate3357.

Miller, R. L., and Coauthors, 2014: CMIP5 historical simulations (1850–2012) with GISS ModelE2. *J. Advances in Modeling Earth Systems*, **6**, 441_477.

Morice, C. P., J. J. Kennedy, N. A. Rayner and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**: D08101 doi:101029/2011JD017187.

Müller, W.A., and Coauthors, 2015: A twentieth-century reanalysis forced ocean model to reconstruct the North Atlantic climate variation during the 1920s. *Climate Dynamics*, **44**(7-8), 1935-1955.

Myhre, G. and Coauthors, 2014: Anthropogenic and Natural Radiative Forcing. *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

Myhre, G., and Coauthors, 2017: Multi-model simulations of aerosol and ozone radiative forcing due to anthropogenic emission changes during the period 1990–2015. *Atmos. Chemistry and Phys.*, **17**(4), 2709-2720.

Nazarenko, L., D. Rind, K. Tsigaridis, A. D. Del Genio, M. Kelley, and N. Tausnev, 2017: Interactive nature of climate change and aerosol forcing. *J. Geophys, Res*., 122, **6**, 3457-3480.

Ocko I. B., V. Ramaswamy and Y. Ming, 2014: Contrasting climate responses to the scattering and absorbing features of anthropogenic aerosol forcings. *J. Climate*, **27**, 5329–5345

Otto, A. and Coauthors, 2013: Energy budget constraints on climate response. *Nature Geosci.*, **6**, 415–416.

Paynter, D., and T. L. Frölicher, 2015: Sensitivity of radiative forcing, ocean heat uptake, and climate feedback to changes in anthropogenic greenhouse gases and aerosols, *J. Geophys. Res. Atmos*., **120**, 9837–9854, doi:10.1002/2015JD023364.

Proistosescu, C., and P. J. Huybers, 2017: Slow climate mode reconciles historical and model-based estimates of climate sensitivity. *Science Advances*, ***3***(7), e1602821.

Qi, L., Q. Li, C. He, X. Wang, and J. Huang: Effects of the Wegener–Bergeron–Findeisen process on global black carbon distribution, Atmos. Chem. Phys., 17, 7459-7479, https://doi.org/10.5194/acp-17-7459-2017, 2017.

Qu, X., and Coauthors, 2017: On the emergent constraints of climate sensitivity. *J. Climate*, doi:10.1175/JCLI-D-17-0482.1.

Richardson, M., K. Cowtan, E. Hawkins, and M. B. Stolpe, 2016: Reconciled climate response estimates from climate models and the energy budget of Earth. *Nature Climate Change*, **6**(10), 931-935.

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

Roe G. H., and K. C. Armour, 2011: How sensitive is climate sensitivity? *Geophys. Res. Lett*., **38**:L14708.

Rotstayn, L. D., M. A. Collier, D. T. Shindell, and O. Boucher, 2015: Why does aerosol forcing control historical global-mean surface temperature change in CMIP5 models?, *J. Clim.*, **28**, 6608–6625, doi:10.1175/jcli-d-14-00712.1.

Rugenstein, M. A., J. M. Gregory, N. Schaller, J. Sedláček and R. Knutti, 2016: Multiannual ocean–atmosphere adjustments to radiative forcing. *J. Climate*, **29**(15), 5643-5.

Samset, B. H., and Coauthors, 2014: Modelled black carbon radiative forcing and atmospheric lifetime in AeroCom Phase II constrained by aircraft observations. *Atmos. Chem. Phys.*, **14**,12465-12477.

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.

Sato, Y., D. Goto, T. Michibata, K. Suzuki, T. Takemura, H. Tomita, and T. Nakajima, 2018: Aerosol effects on cloud water amounts were successfully simulated by a global cloud-system resolving model. *Nature communications*, **9**(1), 985.

Seifert, A., T. Heus, R. Pincus, and B. Stevens, 2015: Large-eddy simulation of the transient and near-equilibrium behavior of precipitating shallow convection. *J. Advances in Modeling Earth Systems*, **7**, 1918–1937.

Sherwood, S.C., and Coauthors, 2015: Adjustments in the forcing-feedback framework for understanding climate change. *Bulletin of the American Meteorological Society*, ***96*(2)**, 217-228.

Shindell, D.T., 2014: Inhomogeneous forcing and transient climate sensitivity. *Nature Climate Change*, **4**, 274–277.

Simmons, A. J., and Coauthors, 2017: A reassessment of temperature variations and trends from global reanalyses and monthly surface climatological datasets. *Quart. J. Roy. Meteor. Soc.*, **143**, 101–119.

Stevens, B., 2015: Rethinking the lower bound on aerosol radiative forcing. *J. Climate*, **28**, 4794–4819.

Stevens, B., and Coauthors, 2017: MACv2-SP: a parameterization of anthropogenic aerosol optical properties and an associated Twomey effect for use in CMIP6, *Geosci. Model Dev.*, **10**, 433-452.

Toll, V., M. Christensen, S. Gass, and N. Bellouin, 2017: Volcano and ship tracks indicate excessive aerosol-induced cloud water increases in a climate model. *Geophys. Res. Lett*. doi: 10.1002/2017GL075280.

Trenberth, K.E. and J. W. Hurrell, 1994: Decadal atmosphere-ocean variations in the Pacific. *Climate Dynamics*, ***9***(6),.303-319.

van Oldenborgh, G. J., L. A. te Raa, H.A. Dijkstra, and S. Y. Philip, 2009: Frequency- or amplitude-dependent effects of the Atlantic meridional overturning on the tropical Pacific Ocean, *Ocean Sci*ence, **5**, 293-301, https://doi.org/10.5194/os-5-293-2009.

Wang, Q. Q., and Coauthors, 2014: Global budget and radiative forcing of black carbon aerosol: Constraints from pole-to-pole (HIPPO) observations across the Pacific. *J. Geophys. Res.-Atmos.*, **119**,195-206.

Wang, R., and Coauthors, 2016: Estimation of global black carbon direct radiative forcing and its uncertainty constrained by observations, *J. Geophys. Res. Atmos*., **121**, 5948–5971,

Williams, K. D., W. J. Ingram, and J. M. Gregory, 2008: Time Variation of Effective Climate Sensitivity in GCMs. *J. Climate*, **21**,5076–5090.

Wolter, K., and M. S. Timlin, 1993: Monitoring ENSO in COADS with a seasonally adjusted principal component index. *Proc of the 17th Climate Diagnostics Workshop, Norman, OK, NOAA/NMC/CAC, NSSL, Oklahoma Clim Survey, CIMMS and the School of Meteor.*, University of Oklahoma, 52-57.

Wolter, K., and M. S. Timlin, 2011: El Niño/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEIext). *Intl. J. Climatology*, **31**,1074–1087.

Zhang, Y., and Coauthors, 2017: Top-of-atmosphere radiative forcing affected by brown carbon in the upper troposphere. *Nature Geoscience,* **10**, 486–489.

Zhao, M., and Coauthors, 2016:. Uncertainty in model climate sensitivity traced to representations of cumulus precipitation microphysics. *J. Climate* **29.2**, 543-560.

Zhou, C., and J. E. Penner, 2017: Why do general circulation models overestimate the aerosol cloud lifetime effect? A case study comparing CAM5 and a CRM. *Atmos. Chem. Phys.*, **17**, 21-29

Zhou, C., M. D. Zelinka, and S. A. Klein, 2016: Impact of decadal cloud variations on the Earth's energy budget. *Nature Geoscience,* **9**, 871–874.

Table captions

Table 1 Components of ERF and treatment of their uncertainties. Units are Wm^{-2}.

Table 2 Best estimates (medians) and 5–95% uncertainty ranges for changes ΔT in global mean surface temperature, ΔF in effective radiative forcing and ΔN in total heat uptake between the base and final periods indicated. The final two lines, in italics, show comparative results for LC15 for the first two period combinations given in that paper. The values for ΔF are after probabilistically applying the AR5 efficacy range for ERF_{BCsnow}.

Table 3 Best estimates (medians) and uncertainty ranges for ECS and TCR using the base and final periods indicated. Values in roman type compute ΔT using the HadCRUT4v5 dataset; values in *italics* compute ΔT using the infilled, globally-complete Had4_krig_v2 dataset. Ranges are stated to the nearest 0.05 K. The final two lines show the comparable results from Lewis and Curry (2015) for the first two period combinations given in that paper. All these ECS estimates assume that the climate feedback parameter is a constant.

Table 4 Sensitivity of best estimates and uncertainty ranges for ECS and TCR. Ranges are stated to the nearest 0.05 K.

Table 5 Simplified, one-at-a-time analysis of data values implied by statistically inconsistent CMIP5 models. The first two rows of data show the values of each of ΔT , $\Delta F / F_{2\times CO2}$ and $\Delta N / F_{2\times CO2}$ implied by the stated ECS_{hist} and TCR values, if the remaining two of those variables each took its median value. Those ECS_{hist} and TCR values are, for each parameter, the lowest for any CMIP5 model in the ensemble that is above the 95% uncertainty bounds given by the preferred (1859–82 to 2007–16) estimates from the main analysis using the globally-complete Had4_krig_v2 dataset. The final row shows the observationally-derived uncertainty ranges the three variables. Best estimates and uncertainty ranges are derived from the same one million samples used for the main statistical analysis.

Figure captions

Fig. 1 Comparison of actual and step-emulated ensemble-mean changes from preindustrial in global surface temperature, ΔT, and TOA radiative imbalance, ΔN, in 1pctCO2 simulations. Small and large circles show respectively annual and pentadal mean actual values, blue for ΔT and green for ΔN. The red and magenta lines show respectively ΔT and ΔN values as emulated from the step-responses of the same models in abrupt4xCO2 simulations. The non-logarithmic element of the CO_2 forcing–concentration relationship (Byrne and Goldblatt 2014; Etminan et al. 2016) has been allowed for. The same ensemble of 31 CMIP5 models is used as in Table S2. The minor excess of the emulated ΔN values in the middle years is due principally to the behavior of GISS-E2 models; if their p3 versions are excluded the match for ΔN becomes almost perfect throughout, while that for ΔT remains so.

Fig. 2 Anthropogenic forcings from 1750 to 2016. All time-series that are affected by the revisions to AR5 CO_2, CH_4 and nitrous oxide forcing–concentration relationships and to post-1990 revisions to AR5 aerosol and tropospheric ozone forcing are shown separately. In some cases the Original AR5 1750–2011 time-series overlay the Revised 1750–2016 time-series prior to 2012. Unrevised anthropogenic forcing components (Stratospheric H_2O, Land use (albedo), BC on snow, Contrails) have been combined into a single Other Anthropogenic time-series. Natural forcings (Solar, Volcanic) are not shown as they have not been revised and post 2011 changes in them are very small.

Fig. 3 Natural factors that influence selection of base and final periods, and surface temperature dataset coverage, during 1850–2016. Volcanic forcing is from AR5. The AMO index comprises the residuals from regressing 25–60 N, 5–70 W HadSST3 data on total forcing with years in which volcanic forcing is < -0.5 Wm^{-2} omitted, and is scaled up by 3 times. The MEI index has been extended before 1950 using a regression fit to the MEI.ext index (Wolter and Timlin 2011), and then detrended (relative to time). The two indices are plotted as five-year centered means (three-year/one-year means for next-but-end/end years); their units are arbitrary. Annual means of HadCRUT4v5 monthly grid-cell coverage as a fraction of the Earth's surface are shown. The preferred base and final periods are shaded.

Fig. 4 Estimated probability density functions for ECS and TCR using each period combination shown in the main results. Original GMST refers to use of the HadCRUT4v5 record; Infilled GMST refers to use of the Had4_krig_v2 record. Box plots show probability percentiles, accounting for probability beyond the range plotted: 5–95 (bars at line ends), 17–83 (box-ends) and 50 (bar in box: median).

Fig. 5 Change in net outgoing radiation ΔR plotted against change in surface temperature ΔT. Blue and cyan circles show pentadal means, red squares show 15-year means. Panel **a:** average anomalies, relative to the 1871–1900 mean, from two 1871–2010 amipPiForcing simulations by HadGEM2-A. The black line shows the linear fit with pentads spanning 1906–1925 (cyan circles) excluded (see Figure S2). No-intercept fitting with all pentads included yields an almost identical fit. Plotted 15-year means are for periods ending 1950, 1965, 1980, 1995 and 2010. Panel **b**: observationally-estimated anomalies over 1872–2016 relative to the 1850–1884 mean. Forcing is as per section 3.a, with an efficacy of 0.55 applied to $\mathrm{ERF}_{\mathrm{Volcano}}$. ΔR is estimated as $(2.52 - 0.50)/2.52 * \Delta F$; this scaling is based on the ΔF and ΔN values from row one of Table 1. Had4_krig_v2 is used for ΔT. The black line shows the no-intercept linear fit to all pentadal values. Fitting with an intercept, but excluding pentads spanning 1907–1926, gives a 1% lower best-fit slope. Plotted 15-year means are for periods 1927–1941, 1942–1956, 1957–1971, 1972–1986, 1987–2001 and 2002–2016. Pre-1927 pentads are colored cyan.

ERF component	This study 1750–2016 best estimate	Revised 1750–2011 best estimate	AR5 1750–2011 best estimate and 90% CI	Part treated as independent	Added fixed uncertainty 90% CI
WMGG	3.176	2.989	2.831 (2.260–3.400)	0%	
Ozone (total)	*0.392*	*0.379*	*0.350 (0.141–0.559)*		
Stratospheric H_2O	*0.074*	*0.073*	*0.073 (0.022–0.124)*		
Land use (albedo)	*−0.151*	*−0.150*	*−0.150 (−0.253–−0.047)*		
Total OWL	0.315	0.302	0.273 (0.034– 0.512)	50%	
Aerosol (total)	−0.769	−0.777	−0.900 (−1.900–−0.100); revised to (−1.700–−0.100)	25%	
BC on snow	0.040	0.040	0.040 (0.020–0.090)	Ignored	
Contrails	0.059	0.050	0.050 (0.020–0.150)	Ignored	
Total anthropogenic	*2.821*	*2.581*	*2.294(1.134–3.334)*		
Solar	0.021	0.030	0.030 (−0.021–+0.081)	50%	±0.05
Volcanic	−0.099	−0.125	−0.125 (−0.160–−0.090)	50%	±0.072

Table 1 Components of ERF and treatment of their uncertainties. Units are Wm^{-2}.

Base period	Final period	ΔT HadCRUT4 [K]	ΔT Had4_krig_v2 [K]	ΔF [Wm^{-2}]	ΔN [Wm^{-2}]
1869–1882	2007–2016	0.80 (0.65–0.95)	0.88 (0.73–1.03)	2.52 (1.68–3.36)	0.50 (0.25–0.75)
1869–1882	1995–2016	0.73 (0.58–0.87)	0.79 (0.63–0.94)	2.26 (1.44–3.09)	0.49 (0.29–0.69)
1850–1900	1980–2016	0.65 (0.51–0.79)	0.71 (0.56–0.86)	2.01 (1.21–2.82)	0.40 (0.21–0.60)
1930–1950	2007–2016	0.61 (0.47–0.75)	0.65 (0.51–0.79)	1.94 (1.22–2.66)	0.45 (0.18–0.72)
Lewis and Curry (2015) estimates for comparison					
1859–1882	1995–2011	0.71 (0.56–0.86)	n/a	1.98 (0.99–2.86)	0.36 (0.15–0.58)
1850–1900	1987–2011	0.66 (0.52–0.81)	n/a	1.88 (0.92–2.74)	0.41 (0.19–0.63)

Table 2 Best estimates (medians) and 5–95% uncertainty ranges for changes ΔT in global mean surface temperature, ΔF in effective radiative forcing and ΔN in total heat uptake. between the base and final periods indicated. The final two lines show comparative values for LC15 for the first two period combinations given in that paper. The values for ΔF are after probabilistically applying the AR5 efficacy range for $\mathrm{ERF}_{\mathrm{BCsnow}}$.

Base period	Final period	ECS best estimate [K]	ECS 17-83% range [K]	ECS 5-95% range [K]	TCR best estimate [K]	TCR 17-83% range [K]	TCR 5-95% range [K]
1869–1882	**2007–2016**	**1.50** ***1.66***	**1.2–1.95** ***1.35–2.15***	**1.05–2.45** ***1.15–2.7***	**1.20** ***1.33***	**1.0–1.45** ***1.1–1.60***	**0.9–1.7** ***1.0–1.9***
1869–1882	1995–2016	1.56 *1.69*	1.2–2.1 *1.35–2.25*	1.05–2.75 1.15–3.0	1.22 *1.32*	1.0–1.5 *1.1–1.65*	0.85–1.85 *0.95–2.0*
1850–1900	1980–2016	1.54 *1.67*	1.2–2.15 *1.3–2.3*	1.0–2.95 *1.1–3.2*	1.23 *1.33*	1.0–1.6 *1.05–1.7*	0.85–1.95 *0.9–2.15*
1930–1950	2007–2016	1.56 *1.65*	1.2–2.15 *1.25–2.3*	1.0–3.0 *1.05–3.15*	1.20 *1.27*	0.95–1.5 *1.05–1.6*	0.85–1.85 *0.9–1.95*
Lewis and Curry (2015) results for comparison							
1859–1882	1995–2011	1.64	1.25–2.45	1.05–4.05	1.33	1.05–1.8	0.90–2.5
1850–1900	1987–2011	1.67	1.25–2.6	1.0–4.75	1.31	1.0–1.8	0.85–2.55

Table 3 Best estimates (medians) and uncertainty ranges for ECS and TCR using the base and final periods indicated. Values in roman type compute ΔT using the HadCRUT4v5 dataset; values in *italics* compute ΔT using the infilled, globally-complete Had4_krig_v2 dataset. The preferred estimates are shown in bold. Ranges are stated to the nearest 0.05 K. The final two lines show the comparable results from LC15 for the first two period combinations given in that paper. All these ECS estimates assume that the climate feedback parameter is a constant.

Variation from 1869–1882 base period, 2007–2016 final period, main results for the HadCRUT4v5 case	ECS best estimate [K]	ECS 5-95% range [K]	TCR best estimate [K]	TCR 5-95% range [K]
Base case – no variations	1.50	1.05–2.45	1.20	0.9–1.7
Base period 1850–1882[1] ex low cover & volcanic yrs[2]	1.50	1.05–2.45	1.20	0.9–1.7
Base period 1850–1900; volcanic efficacy 1.0	1.44	1.05–2.15	1.16	0.9–1.6
Base period 1850–1900; volcanic efficacy 0.55	1.52	1.1–2.35	1.21	0.9–1.65
Base period 1850–1882[1]; volcanic efficacy 0.55	1.52	1.1–2.4	1.22	0.9–1.7
$ERF_{Aerosol}$ uncertainty range 5% bound as per AR5	1.51	1.05–2.65	1.21	0.9–1.8
ERF_{WMGG} uncertainty range scaled up by 50%	1.50	1.05–2.6	1.20	0.9–1.75
ERF_{OWL} uncertainty range scaled up by 50%	1.50	1.05–2.55	1.20	0.85–1.75
$ERF_{Aerosol}$ uncertainty range scaled down by 50%	1.50	1.1–2.15	1.20	0.95–1.55
AR5 original ERF_{GHG} + >1990 aerosol & O_3 forcing[3]	1.68	1.1–3.25	1.31	0.9-2.05
0-2000 m OHC based only on Cheng et al data	1.47	1.05–2.35	n/a	n/a
0-2000 m OHC based only on Levitus/NOAA data	1.54	1.05–2.55	n/a	n/a
ERF_{LUC} set to zero (increases ΔF by 0.10 Wm^{-2})	1.43	1.0–2.25	1.16	0.85–1.6

Table 4 Sensitivity of best estimates (medians) and uncertainty ranges for ECS and TCR. Ranges are stated to the nearest 0.05 K.

[1] Heat uptake for the 1850−1882 base period is set mid-way between those for the 1850−1900 and 1869−1882 periods, or equal to the latter when low coverage and volcanic years are excluded.

[2] The criteria for excluding years from the 1850-1882 base period due to low coverage or volcanism is HadCRUT4v5 areal coverage < 0.2 or $ERF_{Volcano} < -0.5$ Wm^{-2}.

[3] AR5 original (unrevised) post-2011 tropospheric ozone and aerosol forcing are derived by extrapolation using their small 2002−11 trends.

Implied value of each variable if the other two each equal their median value	ΔT [K]	$\Delta F / F_{2\times CO2}$	$\Delta N / F_{2\times CO2}$
Parameter and lowest inconsistent value in CMIP5 model ensemble			
ECS_{hist}: 2.89 K	1.54	0.43	0.36
TCR: 1.91 K	1.26	0.46	n/a
Observational 5-95% range	0.73–1.03	0.49–0.83	0.06–0.21

Table 5 Simplified, one-at-a-time analysis of data values implied by statistically inconsistent CMIP5 models. The first two rows of data show the values of each of ΔT , $\Delta F / F_{2\times CO2}$ and $\Delta N / F_{2\times CO2}$ implied by the stated ECS_{hist} and TCR values, if the remaining two of those variables each took its median value. Those ECS_{hist} and TCR values are, for each parameter, the lowest for any CMIP5 model in the ensemble that is above the 95% uncertainty bounds given by the preferred (1859–82 to 2007–16) estimates from the main analysis using the globally-complete Had4_krig_v2 dataset. The final row shows the observationally-derived uncertainty ranges the three variables. Best estimates and uncertainty ranges are derived from the same one million samples used for the main statistical analysis.

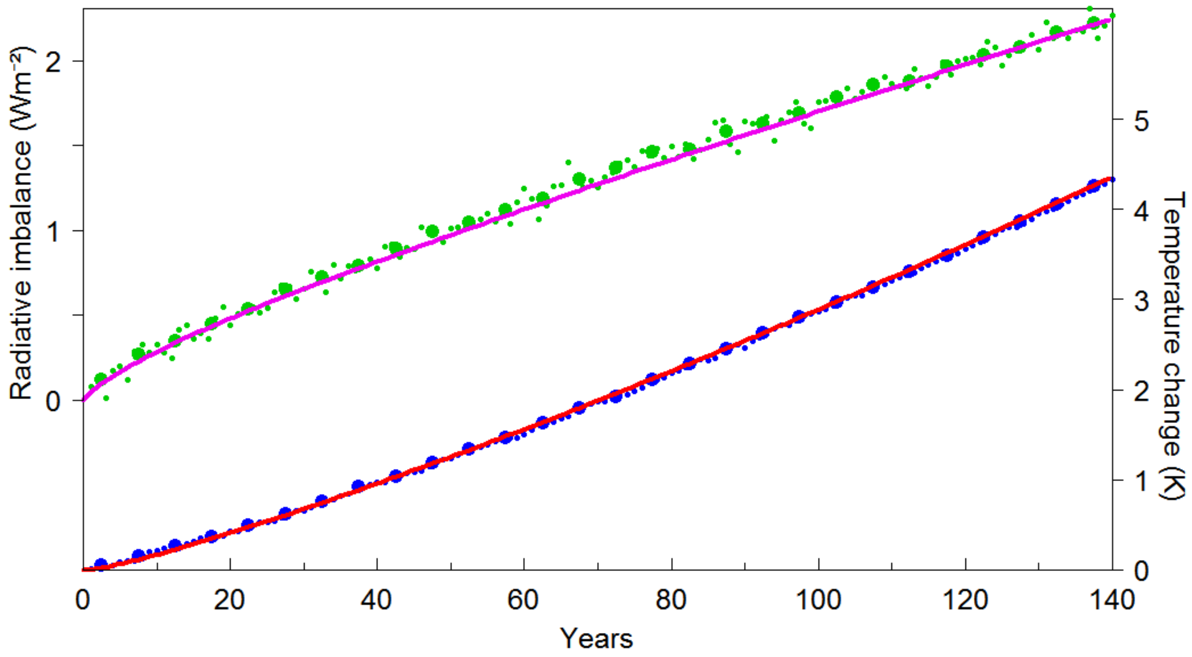

Fig. 1 Comparison of actual and step-emulated ensemble-mean changes from preindustrial in global surface temperature, Δ*T*, and TOA radiative imbalance, Δ*N*, in 1pctCO2 simulations. Small and large circles show respectively annual and pentadal mean actual values, blue for Δ*T* and green for Δ*N*. The red and magenta lines show respectively Δ*T* and Δ*N* values as emulated from the step-responses of the same models in abrupt4xCO2 simulations. The non-logarithmic element of the CO_2 forcing–concentration relationship (Byrne and Goldblatt 2014; Etminan et al. 2016) has been allowed for. The same ensemble of 31 CMIP5 models is used as in Table S2. The minor excess of the emulated Δ*N* values in the middle years is due principally to the behavior of GISS-E2 models; if their p3 versions are excluded the match for Δ*N* becomes almost perfect throughout, while that for Δ*T* remains so.

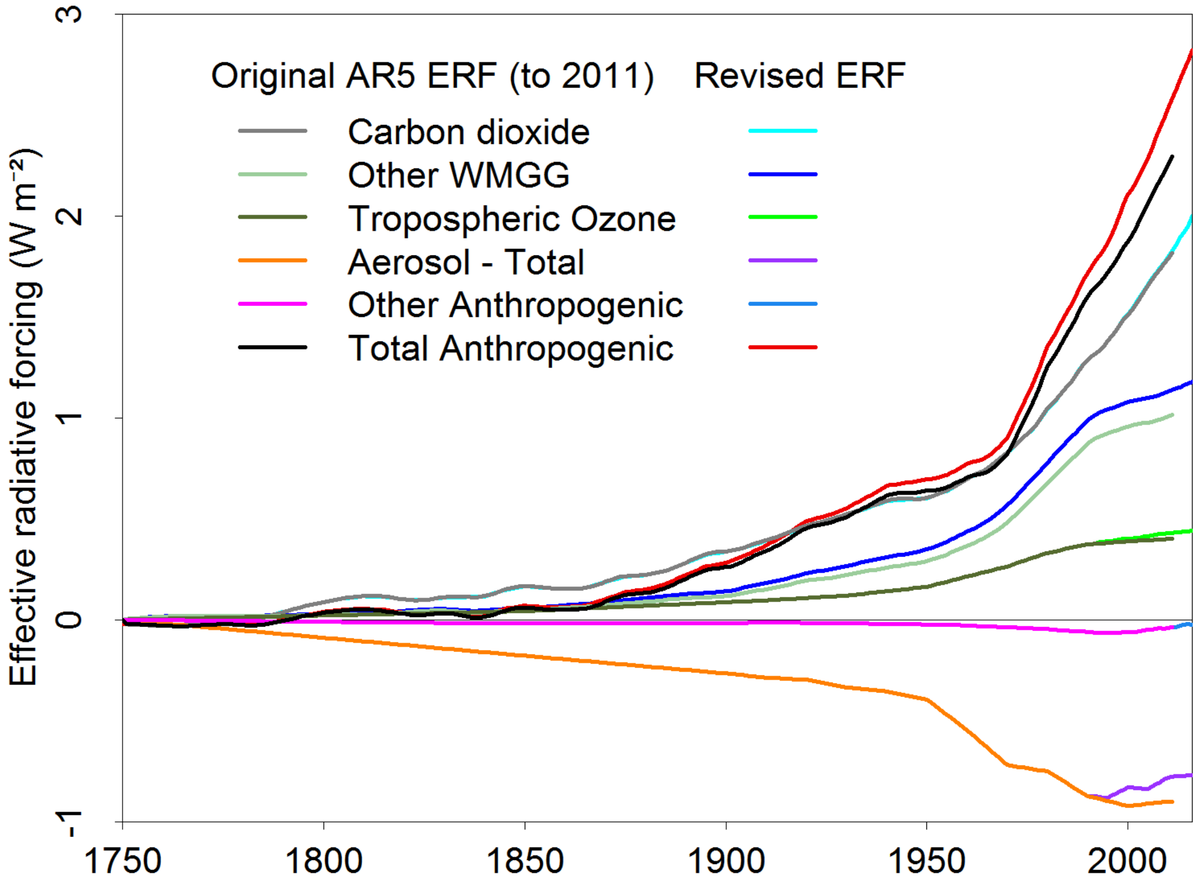

Fig. 2 Anthropogenic forcings from 1750 to 2016. All time-series that are affected by the revisions to AR5 CO_2, CH_4 and nitrous oxide forcing–concentration relationships and to post-1990 revisions to AR5 aerosol and tropospheric ozone forcing are shown separately. In some cases the Original AR5 1750–2011 time-series overlay the Revised 1750–2016 time-series prior to 2012. Unrevised anthropogenic forcing components (Stratospheric H_2O, Land use (albedo), BC on snow, Contrails) have been combined into a single Other Anthropogenic time-series. Natural forcings (Solar, Volcanic) are not shown as they have not been revised and post 2011 changes in them are very small.

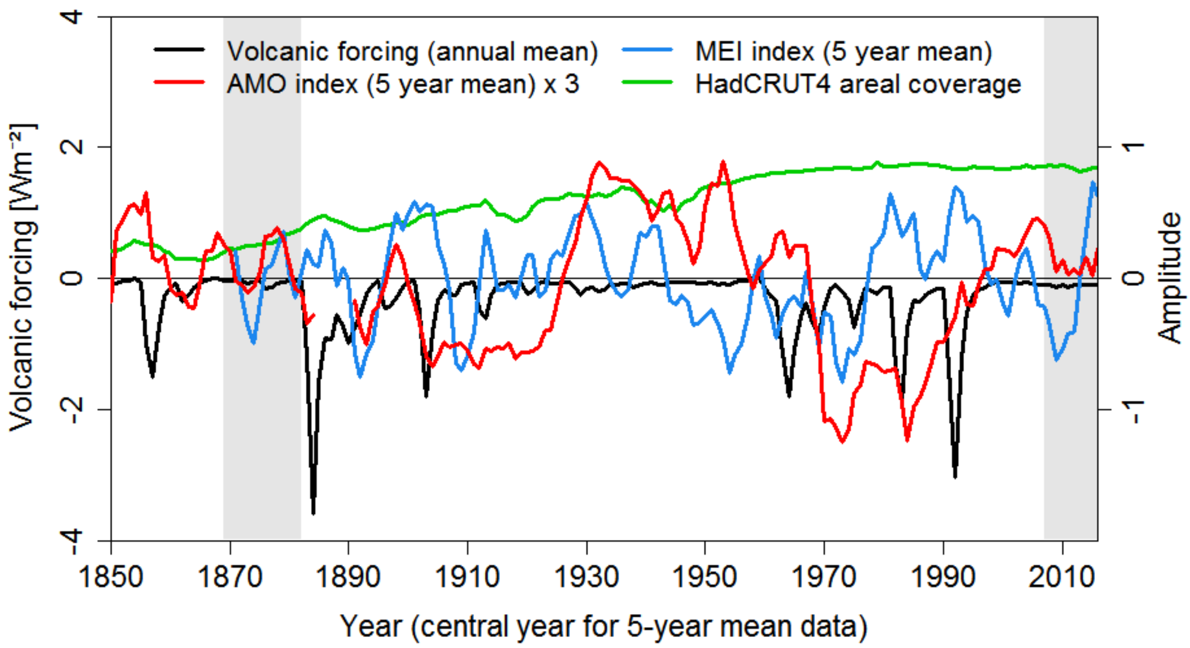

Fig. 3 Natural factors that influence selection of base and final periods, and surface temperature dataset coverage, during 1850–2016. Volcanic forcing is from AR5. The AMO index comprises the residuals from regressing 25–60 N, 5–70 W HadSST3 data on total forcing with years in which volcanic forcing is < −0.5 Wm^{-2} omitted, and is scaled up by 3 times. The MEI index has been extended before 1950 using a regression fit to the MEI.ext index (Wolter and Timlin 2011), and then detrended (relative to time). The two indices are plotted as five-year centered means (three-year/one-year means for next-but-end/end years); their units are arbitrary. Annual means of HadCRUT4v5 monthly grid-cell coverage as a fraction of the Earth's surface are shown. The preferred base and final periods are shaded.

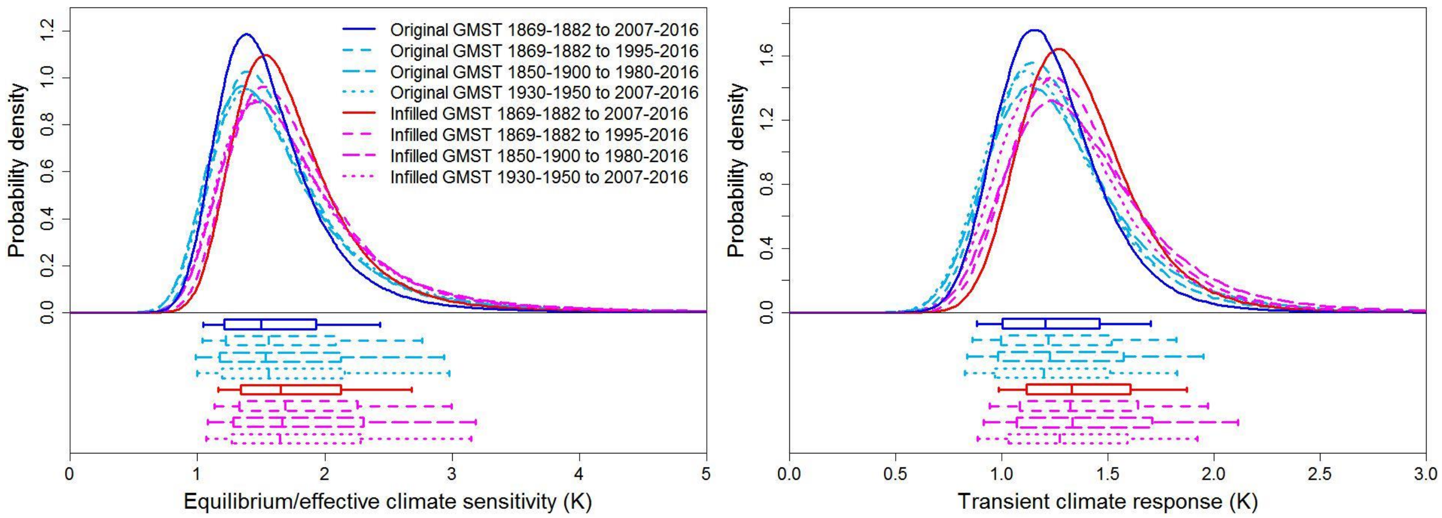

Fig. 4 Estimated probability density functions for ECS and TCR using each period combination shown in the main results. Original GMST refers to use of the HadCRUT4v5 record; Infilled GMST refers to use of the Had4_krig_v2 record. Box plots show probability percentiles, accounting for probability beyond the range plotted: 5–95 (bars at line ends), 17–83 (box-ends) and 50 (bar in box: median).

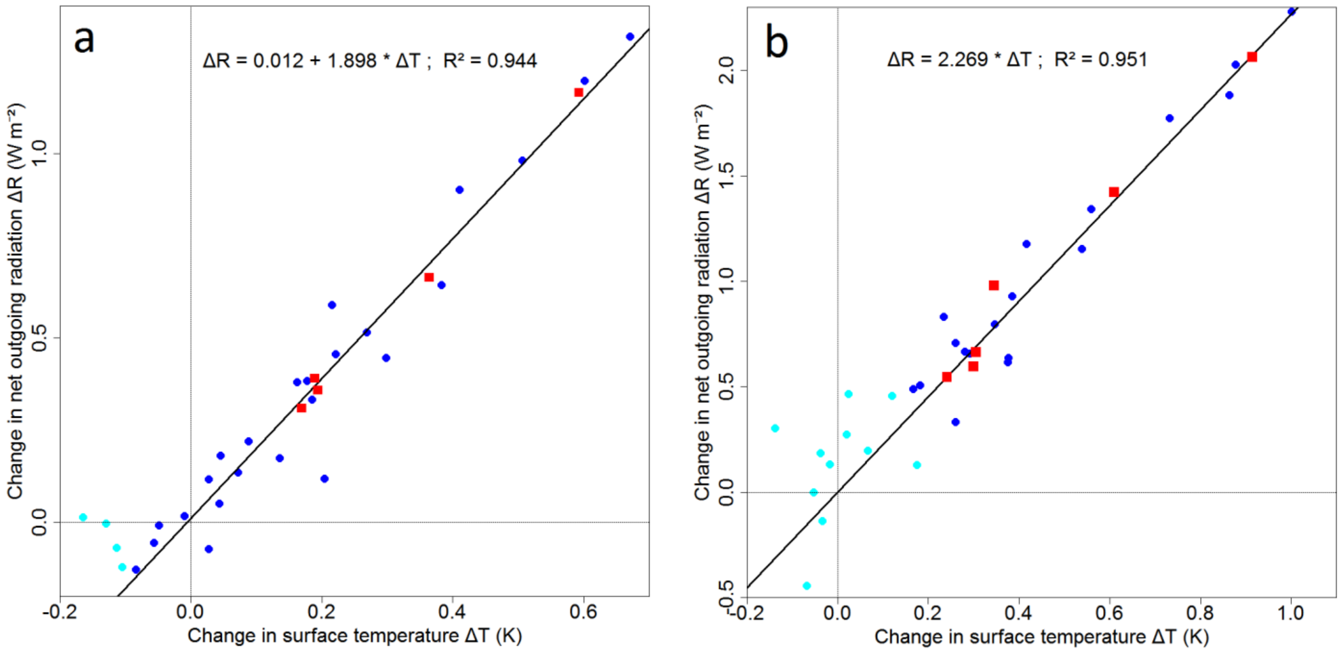

Fig. 5 Change in net outgoing radiation Δ*R* plotted against change in surface temperature Δ*T*. Blue and cyan circles show pentadal means, red squares show 15-year means. Panel **a:** average anomalies, relative to the 1871–1900 mean, from two 1871–2010 amipPiForcing simulations by HadGEM2-A. The black line shows the linear fit with pentads spanning 1906–1925 (cyan circles) excluded (see Figure S2). No-intercept fitting with all pentads included yields an almost identical fit. Plotted 15-year means are for periods ending 1950, 1965, 1980, 1995 and 2010. Panel **b**: observationally-estimated anomalies over 1872–2016 relative to the 1850–1884 mean. Forcing is as per section 3.a, with an efficacy of 0.55 applied to $ERF_{Volcano}$. Δ*R* is estimated as (2.52 − 0.50)/2.52 * Δ*F*; this scaling is based on the Δ*F* and Δ*N* values from row one of Table 2. Had4_krig_v2 is used for Δ*T*. The black line shows the no-intercept linear fit to all pentadal values. Fitting with an intercept, but excluding pentads spanning 1907–1926, gives a 1% lower best-fit slope. Plotted 15-year means are for periods 1927–1941, 1942–1956, 1957–1971, 1972–1986, 1987–2001 and 2002–2016. Pre-1927 pentads are colored cyan.