The Impact of Recent Forcing and Ocean Heat Uptake Data on Estimates of Climate Sensitivity

NICHOLAS LEWIS
Bath, United Kingdom

JUDITH CURRY
Climate Forecast Applications Network, Reno, Nevada

(Manuscript received 5 October 2017, in final form 9 April 2018)

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 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 an 1869–82 base period and a 2007–16 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% range: 1.05–2.45 K) and for TCR of 1.20 K (5%–95% range: 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 give slightly higher estimates: a median of 1.66 K for ECS (5%–95% range: 1.15–2.7 K) and 1.33 K for TCR (5%–95% range: 1.0–1.9 K). These ECS estimates reflect climate feedbacks over the historical period, assumed to be time invariant.

1. Introduction

There has been considerable scientific investigation of the magnitude of the warming of Earth’s climate from changes in atmospheric carbon dioxide (CO₂) 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 CO₂ concentration. Transient climate response (TCR), a shorter-term measure over 70 years, represents warming at the time CO₂ concentration has doubled when it is increased by 1% yr⁻¹.

For over 30 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 et al. 1979) is unchanged in the 2013 Fifth Assessment Report (AR5) from the IPCC (IPCC 2013). AR5 did not provide a best estimate value for ECS, stating (in the summary for policymakers, section D.2) that “No best estimate for equilibrium climate sensitivity can now be given because of a lack of agreement on values across assessed lines of evidence and studies” (p. 16).

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, hereinafter 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 and 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 provides conclusions.

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 constraining metrics in the current climate that 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. 2018). 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 $\lambda$, 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 $\Delta R$ to a change in radiative forcing $\Delta F$ that alters positively Earth’s net downward top-of-the-atmosphere (TOA) radiative imbalance $N$. The assumption is that, in temporal mean terms, $\Delta R$—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 $\Delta T$. Mathematically,

$$
\Delta R = \lambda \Delta T + \mu_R,
$$

with $\lambda$ (the climate feedback parameter, representing the increase in net outgoing energy flux per degree of surface warming) being a constant, and $\mu_R$ being 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.

Conservation of energy yields ΔN = ΔF − ΔR. Therefore, in temporal mean terms, substituting using (1) yields

\[ \lambda = (\Delta F - \Delta N) / \Delta T. \] (2)

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

\[ \lambda = F_{2 \times CO₂} / ECS. \] (3)

Hence, substituting in (2), in general

\[ ECS = F_{2 \times CO₂} \Delta T / (\Delta F - \Delta N), \] (4)

with the CO₂ 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 downward 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 \( \lambda \). 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-yr simulations in which CO₂ forcing is increased smoothly at 1% yr⁻¹ (1pctCO₂ simulation), 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 CO₂ concentration is abruptly quadrupled (abrupt4xCO₂ simulation) (Fig. 1; Good et al. 2011; Caldeira and Myhrvold 2013). Such behavior strongly...
suggests that feedback strength in CMIP5 models generally does not change with $\Delta F$ or $\Delta T$ per se, at least up to respectively the CO$_2$ forcing from a 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 $\Delta T$ reaches 5 K under quadrupled CO$_2$.

However, in most CMIP5 AOGCMs, $\lambda$ (here $-dN/dT$) tends to decrease a few decades into abrupt4xCO2 simulations—when $N$ is plotted against $T$ (a so-called Gregory plot), the slope is gentler after that time—although generally $\lambda$ then remains almost constant for the rest of the simulation (Armour 2017; see Fig. S1 in the supplemental material). In some cases, the decrease in $\lambda$ may be linked to temperature- or time-dependent energy leakage (Hobbs et al. 2016). However, typically the decrease in $\lambda$ 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 $\lambda$ 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 $\Delta T$ or $\Delta N$ 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.

TCR can be estimated 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 about 70 years (Bindoff et al. 2014, p. 920). That is,

$$\text{TCR} = \frac{\Delta T}{\Delta F}. \quad (5)$$

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-yr linear forcing ramp (Armour 2017). TCR may therefore also be estimated using a base period early in the historical period, with a possible marginal upward 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 1930–50 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 subdecadal time scales 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 $\Delta T$ 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 $\Delta F$ (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 $\Delta F$ 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. The two main contributors to uncertainty in $\Delta F$ are aerosols and, to a substantially smaller extent, well-mixed greenhouse gases (WMGG).

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 Tables 8.6 and 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 (see section S1 in the supplemental material). 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 $-0.9\,\text{W m}^{-2}$ in 1995, $\text{ERF}_{\text{Aerosol}}$ weakens to $-0.8\,\text{W m}^{-2}$ 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 ($\text{ERF}_{\text{Aerosol}}$) has improved. A number of recent studies point to total aerosol forcing being substantially weaker than the lower end of the 2011 range from $-1.9$ to $-0.1\,\text{W m}^{-2}$ (median $-0.9\,\text{W m}^{-2}$) 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 $-1.0\,\text{W m}^{-2}$. 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 $-1.9\,\text{W m}^{-2}$ model-derived lower bound for 2011 aerosol forcing in AR5 now appears too strong. We have therefore weakened it slightly to $-1.7\,\text{W m}^{-2}$, as in Armour (2017), making the range symmetrical about the AR5 2011 median.

Following LC15, CO$_2$ and other greenhouse gas (GHG) forcings are combined into a single $\text{ERF}_{\text{GHG}}$ time series, since AR5 does not distinguish between the two as regards ERF uncertainty. Uncertainty in forcing from WMGG almost entirely relates to how much forcing a given concentration of each greenhouse gas

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Forcing & ERF (W m$^{-2}$) \\
\hline
Aerosol & $-0.9$ \\
\hline
Contrail & $-0.8$ \\
\hline
Forcings & $-1.9$ \\
\hline
\end{tabular}
\caption{Summary of forcing estimates.}
\end{table}
produces—uncertainty in concentrations is minor—and is likely highly correlated among WMGG. AR5 (section 8.5.1) assumes that fractional ERF uncertainties for CO2 apply 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 $ERF_{GHC}$. 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 $\Delta F$ (to which $ERF_{GHC}$ is by far the largest contributor) in the denominator, the effects on ECS and TCR estimate of uncertainty in forcing from WMGG cancel out to a substantial extent. Dropping the assumption of uncertainty being correlated between CO2 and other GHG forcing would have a negligible effect on ECS and TCR estimate uncertainty ranges. The same would apply if in addition the ERF-to-RF uncertainty ratio for non-CO2 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-CO2 WMGG, as in AR5.

Ozone (both tropospheric and stratospheric), stratospheric water vapor ($H_2O$), and land-use (albedo) forcings, for which uncertainty distributions can be added in quadrature, are combined into a single $ERF_{O3W}$ forcing component series (termed $ERF_{nonGAB}$ in LC15).

The resulting forcing best estimates and uncertainties used for the main results are summarized 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 CH4 forcing. Also, $F_{2\times CO2}$ has been revised upward 2.5% to 3.80 W m $^{-2}$, 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.

Table 1. Components of ERF and treatment of their uncertainties (W m $^{-2}$). The three italicized rows represent the components for the row with the total ozone, water vapor, and land use. The row in boldface represents the total of the components from the previous rows. (CI is confidence interval.)

<table>
<thead>
<tr>
<th>ERF component</th>
<th>This study 1750–2016 best estimate</th>
<th>Revised 1750–2011 best estimate</th>
<th>AR5 1750–2011 best estimate and 90% CI</th>
<th>Part treated as independent</th>
<th>Added fixed uncertainty 90% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMGG</td>
<td>3.176</td>
<td>2.989</td>
<td>2.831 (2.260–3.400)</td>
<td>0%</td>
<td>—</td>
</tr>
<tr>
<td>$O_3$ (total)</td>
<td>0.392</td>
<td>0.379</td>
<td>0.350 (0.141–0.559)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Stratospheric water vapor</td>
<td>0.074</td>
<td>0.073</td>
<td>0.073 (0.022–0.124)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Land use (albedo)</td>
<td>−0.151</td>
<td>−0.150</td>
<td>−0.150 (−0.253 to −0.047)</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Total ozone, water vapor, and land use</td>
<td>0.315</td>
<td>0.302</td>
<td>0.273 (0.034–0.512)</td>
<td>50%</td>
<td>—</td>
</tr>
<tr>
<td>Aerosol (total)</td>
<td>−0.769</td>
<td>−0.777</td>
<td>−0.900 (−1.900 to −0.100; revised to −1.700 to −0.100)</td>
<td>25%</td>
<td>—</td>
</tr>
<tr>
<td>BC on snow</td>
<td>0.040</td>
<td>0.040</td>
<td>0.040 (0.020–0.090)</td>
<td>Ignored</td>
<td>—</td>
</tr>
<tr>
<td>Contrails</td>
<td>0.059</td>
<td>0.050</td>
<td>0.050 (0.020–0.150)</td>
<td>Ignored</td>
<td>—</td>
</tr>
<tr>
<td><strong>Total anthropogenic forcing</strong></td>
<td><strong>2.821</strong></td>
<td><strong>2.581</strong></td>
<td><strong>2.294 (1.134–3.334)</strong></td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Solar</td>
<td>0.021</td>
<td>0.030</td>
<td>0.030 (−0.021 to +0.081)</td>
<td>50%</td>
<td>±0.05</td>
</tr>
<tr>
<td>Volcanic</td>
<td>−0.099</td>
<td>−0.125</td>
<td>−0.125 (−0.160 to −0.090)</td>
<td>50%</td>
<td>±0.072</td>
</tr>
</tbody>
</table>

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 $ERF_{GHC}$ and other forcings. Gregory and Andrews (2016) likewise found that volcanic forcing produced a substantially smaller response in AOGCMs than CO2 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 about 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 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 (NOAA 2017), 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. 2017; NOAA 2017; JMA 2017, an update of 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 W m⁻² of the estimates by Desbruyeres et al. (2017) of 0.72 W m⁻² over 2006–14 and by Johnson et al. (2016) of 0.71 W m⁻² over 2005–15.

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 1859–82, 1850–1900, and 1930–50 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 1859–82 value for 1869–82), 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 W m⁻² for all base and final periods used in LC15, derived from the ultralong 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; Morice 2017) 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 2014a,c). 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 GISTEMP and NCDC Merged Land–Ocean Surface Temperature (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 GISTEMP and NOAA GlobalTemp version 4.0.1 datasets between 20-yr periods early and late in their records (1880–99 and 1997–2016) was respectively 0.85 and 0.82 K versus 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 interperiod 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 ultralong 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 Oldenborgh et al. detrended relative to surface temperature. While following van Oldenborgh 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 yr during the instrumental record, peaking around 1875, 1940, and 2005. When using a final period ending in 2016, to maximize the anthropogenic warming signal, matching its mean AMO state requires a base period either early in the historical period or in the midtwentieth century.

Matching the mean ENSO state for the base and final period is not practical where a base period early in the record is used, since the mean ENSO state, as represented by the multivariate ENSO index (MEI; Wolter and Timlin 1993), was lower than in recent decades. However, the MEI depends partly on nondetrended 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 1859–82, 1930–50, and 1850–1900. LC15’s preferred base and final periods were 1859–82 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 Niño events. The decade 2007–16, 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 2007–16, confidence in the reliability of OHU estimation is higher.

Although 1859–82 is well matched with both 1995–2016 and 2007–16 as regards mean volcanic forcing, it is matched with 1980–2016, which has almost identical mean volcanic forcing and acceptably similar AMO state as regards mean volcanic forcing as well as the AMO and ENSO state. The better observed 1930–50 period is also well matched with those final periods, although its shorter 1869–82 period produces essentially identical TCR and ECS estimation we use that instead. It is well matched with the 1995–2016 and 2007–16 final periods as regards mean volcanic forcing as well as the AMO and ENSO state. The better observed 1930–50 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 preindustrial 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 the AMO index.

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_BC_snow, 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 $2^{-0.9}$ W m$^{-2}$ median but the revised 5%-95% uncertainty range from $2^{-0.1}$ to $2^{1.7}$ W m$^{-2}$ 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_BC_snow 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 are combined, thus deriving sampled $\Delta F$ values. The central $F_{2xCO_2}$ value is scaled in the same proportion as the central ERF_GHG values. This produces $F_{2xCO_2}$ samples with uncertainty realizations (proportionately) matching those for WMGG forcing.

2) Uncertainty distributions for $\Delta T$ (using the relevant ensemble of 100 realizations) and for $\Delta N$ 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 $\Delta T$, $\Delta F$, $\Delta N$, and $F_{2xCO_2}$, the ECS and TCR values given by (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 $\Delta T$, $\Delta F$, and $\Delta N$, as well as 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 and 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 Fig. 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 of 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 of 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 1869–82 as the base period and 2007–16 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 and 1.9 K using Had4_krig_v2); the corresponding median estimates are 1.50 and 1.20 K (Had4_krig_v2: 1.66 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 CH4 and post-1990 aerosol forcing, partly offset by the higher estimated F2×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 1869–82 base period, 2007–16 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 1850–82 as the base period, with low observational coverage and volcanic years excluded, produces virtually identical ECS and TCR medians and uncertainty ranges to using 1869–82. 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 1850–82 (all years) as the base period. Moreover, applying a volcanic forcing efficacy of 0.55 when regressing surface

### Table 2. Best estimates (medians) and 5%–95% uncertainty ranges for changes \( \Delta T \) in global mean surface temperature, \( \Delta F \) in ERF and \( \Delta N \) 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 \( \Delta F \) are after probabilistically applying the AR5 efficacy range for ERFBCsnow.

<table>
<thead>
<tr>
<th>Base period</th>
<th>Final period</th>
<th>( \Delta T ) HadCRUT4v5 (K)</th>
<th>( \Delta T ) Had4_krig_v2 (K)</th>
<th>( \Delta F ) (W m(^{-2}))</th>
<th>( \Delta N ) (W m(^{-2}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1869–82</td>
<td>2007–16</td>
<td>0.80 (0.65–0.95)</td>
<td>0.88 (0.73–1.03)</td>
<td>2.52 (1.68–3.36)</td>
<td>0.50 (0.25–0.75)</td>
</tr>
<tr>
<td>1869–82</td>
<td>1995–2016</td>
<td>0.73 (0.58–0.87)</td>
<td>0.79 (0.63–0.94)</td>
<td>2.26 (1.44–3.09)</td>
<td>0.49 (0.29–0.69)</td>
</tr>
<tr>
<td>1850–1900</td>
<td>1980–2016</td>
<td>0.65 (0.51–0.79)</td>
<td>0.71 (0.56–0.86)</td>
<td>2.01 (1.21–2.82)</td>
<td>0.40 (0.21–0.60)</td>
</tr>
<tr>
<td>1930–50</td>
<td>2007–16</td>
<td>0.61 (0.47–0.75)</td>
<td>0.65 (0.51–0.79)</td>
<td>1.94 (1.22–2.66)</td>
<td>0.45 (0.18–0.72)</td>
</tr>
</tbody>
</table>

LC15 estimates for comparison

| 1859–82     | 1995–2011    | 0.71 (0.56–0.86)            | —                           | 1.98 (0.99–2.86) | 0.36 (0.15–0.58) |
| 1850–1900   | 1987–2011    | 0.66 (0.52–0.81)            | —                           | 1.88 (0.92–2.74) | 0.41 (0.19–0.63) |

### Table 3. Best estimates (medians) and uncertainty ranges for ECS and TCR using the base and final periods indicated. Values in roman type compute \( \Delta T \) using the HadCRUT4v5 dataset; values in italics compute \( \Delta T \) using the infilled, globally complete Had4_krig_v2 dataset. The preferred estimates are shown in boldface. 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.

<table>
<thead>
<tr>
<th>Base period</th>
<th>Final period</th>
<th>ECS best estimate (K)</th>
<th>ECS 17%–83% range (K)</th>
<th>ECS 5%–95% range (K)</th>
<th>TCR best estimate (K)</th>
<th>TCR 17%–83% range (K)</th>
<th>TCR 5%–95% range (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1869–82</td>
<td>2007–16</td>
<td>1.50</td>
<td>1.2–1.95</td>
<td>1.05–2.45</td>
<td>1.20</td>
<td>1.0–1.45</td>
<td>0.9–1.7</td>
</tr>
<tr>
<td>1850–1900</td>
<td>1980–2016</td>
<td>1.54</td>
<td>1.2–2.15</td>
<td>1.0–2.95</td>
<td>1.23</td>
<td>1.0–1.6</td>
<td>0.95–2.0</td>
</tr>
<tr>
<td>1930–50</td>
<td>2007–16</td>
<td>1.56</td>
<td>1.2–2.15</td>
<td>1.0–3.0</td>
<td>1.20</td>
<td>0.95–1.5</td>
<td>0.9–1.9</td>
</tr>
</tbody>
</table>

LC15 results for comparison

| 1859–82     | 1995–2011    | 1.64                  | 1.25–2.45             | 1.05–4.05            | 1.33                  | 1.05–1.8              | 0.9–2.5               |
| 1850–1900   | 1987–2011    | 1.67                  | 1.25–2.6              | 1.0–4.75             | 1.31                  | 1.0–1.8               | 0.85–2.55             |
temperature per HadCRUT4v5 on (efficacy adjusted) forcing over all years from 1850 to 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 2007–16 only 0.01 K higher than that over 1869–82 (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 interperiod differences indicate that both final periods are well matched with the 1869–82 base period as regards internal variability.

Reverting the aerosol forcing 5% uncertainty bound back from $-1.7 \text{ W m}^{-2}$ to the original AR5 level of $-1.9 \text{ W m}^{-2}$ increases the 95% bounds for ECS and TCR by respectively 0.2 and 0.1 K; their median estimates barely change. Scaling up by 50% the uncertainty range for ERFWMGG increases those bounds by 0.15 and 0.05 K respectively, while doing so for ERFOWL increases them by 0.1 and 0.05 K respectively;

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

<table>
<thead>
<tr>
<th>Variation from 1869–82 base period, 2007–16 final period, main results for the HadCRUT4v5 case</th>
<th>ECS best estimate (K)</th>
<th>ECS 5%–95% range (K)</th>
<th>TCR best estimate (K)</th>
<th>TCR 5%–95% range (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base case: no variations.</td>
<td>1.50</td>
<td>1.05–2.45</td>
<td>1.20</td>
<td>0.9–1.7</td>
</tr>
<tr>
<td>Base period 1850–82, a omitting low cover and volcanic years. b</td>
<td>1.50</td>
<td>1.05–2.45</td>
<td>1.20</td>
<td>0.9–1.7</td>
</tr>
<tr>
<td>Base period 1850–1900; volcanic efficacy 1.0.</td>
<td>1.44</td>
<td>1.05–2.15</td>
<td>1.16</td>
<td>0.9–1.6</td>
</tr>
<tr>
<td>Base period 1850–1900; volcanic efficacy 0.55.</td>
<td>1.52</td>
<td>1.1–2.35</td>
<td>1.21</td>
<td>0.9–1.65</td>
</tr>
<tr>
<td>Base period 1850–82; volcanic efficacy 0.55.</td>
<td>1.52</td>
<td>1.1–2.4</td>
<td>1.22</td>
<td>0.9–1.7</td>
</tr>
<tr>
<td>ERF aerosol uncertainty range 5% bound as per AR5.</td>
<td>1.51</td>
<td>1.05–2.65</td>
<td>1.21</td>
<td>0.9–1.8</td>
</tr>
<tr>
<td>ERFWMGG uncertainty range scaled up by 50%.</td>
<td>1.50</td>
<td>1.05–2.6</td>
<td>1.20</td>
<td>0.9–1.75</td>
</tr>
<tr>
<td>ERFOWL uncertainty range scaled up by 50%.</td>
<td>1.50</td>
<td>1.05–2.55</td>
<td>1.20</td>
<td>0.85–1.75</td>
</tr>
<tr>
<td>ERF aerosol uncertainty range scaled down by 50%.</td>
<td>1.50</td>
<td>1.1–2.15</td>
<td>1.20</td>
<td>0.9–1.55</td>
</tr>
<tr>
<td>AR5 original ERF of aerosol and O₃ forcing. c</td>
<td>1.68</td>
<td>1.1–3.35</td>
<td>1.31</td>
<td>0.9–2.05</td>
</tr>
<tr>
<td>0–2000-m OHC based only on Cheng et al. (2017) data.</td>
<td>1.47</td>
<td>1.05–2.35</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>0–2000-m OHC based only on NOAA (2017) data.</td>
<td>1.54</td>
<td>1.05–2.55</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>ERFFLUC set to zero (increases ΔF by 0.10 W m⁻²).</td>
<td>1.43</td>
<td>1.0–2.25</td>
<td>1.16</td>
<td>0.85–1.6</td>
</tr>
</tbody>
</table>

a Heat uptake for the 1850–82 base period is set midway between those for the 1850–1900 and 1869–82 periods, or equal to the latter when low coverage and volcanic years are excluded.

b The criteria for excluding years from the 1850–82 base period as a result of low coverage or volcanism are HadCRUT4v5 areal coverage <0.2 or ERFVolcano < −0.5 W m⁻².

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

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-and-whisker plots show probability percentiles, accounting for probability beyond the range plotted: 5%–95% (bars at line ends), 17%–83% (box ends), and 50% (bar inside box: median).
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 resulted in the ECS and TCR median values increasing by 0.18 and 0.11 K respectively. The 95% uncertainty bounds for ECS and TCR increase more, by 0.8 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. (2017) or only NOAA (2017) data instead of using estimates averaged over those datasets [and, for the 0–700-m layer, the JMA (2017) 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, hereinafter GA16) found that feedback strength $\lambda$ 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 simulation), was considerably higher over the historical period than in years 1–20 of abrupt4xCO2 simulations. Moreover, $\lambda$ 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 $\lambda$ (their $\bar{\lambda}$) measured by regression over a 30-yr sliding window, with small temperature changes except toward 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 $\lambda$ estimates when tested on pseudodata embodying a constant $\lambda$ (Fig. S2 in the supplemental material).

Plotting $\Delta R$ against $\Delta T$ using pentadal means, averaging-out interannual noise, and considering how averages over consecutive longer periods compare (Fig. 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, $\Delta R$ and $\Delta T$ pentadal anomalies are clustered around the best-fit line, with means for all five 15-yr subperiods 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 $\Delta R$ from the best-fit $\lambda$ estimate of 1.90 W m$^{-2}$ K$^{-1}$ were in the 20 years prior to 1925 and in the decade centered on 1980 (Fig. S3 in the supplemental material). The latter was responsible for the strong 1970–95 upward trend in 30-yr regression-based $\lambda$ in GA16’s Fig. 2a; if the 1976–85 $\Delta R$ values are suitably adjusted, the trend is flat from 1960 on (Fig. S4 in the supplemental material). However, the anomalous $\Delta R$ values circa 1980 have only a minor effect on $\lambda$ estimates derived from 15-yr means: for both 1966–80 and 1981–95, $\Delta R/\Delta T$ was only 7% lower than for 1996–2010. The early heavy volcanism (during 1883–1905) appears not to have affected the best-fit $\lambda$: the ratio of changes in $R$ and $T$ between 1931–60 and 1996–10, two volcanism-free periods, gives almost the same value. Fits for each individual amipPiForcing run are very similar (negligible y intercept, slopes within 5% of the 1.90 W m$^{-2}$ K$^{-1}$ for the ensemble mean, $R^2 = 0.93$ vs 0.94 for the ensemble mean, in all cases with 1906–25 data excluded). This analysis shows that HadGEM2-A displays a near-constant $\lambda$ of 1.9 W m$^{-2}$ K$^{-1}$ over the historical period when driven by observed evolving SST patterns—over 2.3 times 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.

GA16 offered three possible explanations for feedback strength being higher over the historical period in

---

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 (Müller et al. 2015). 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.

3 Based on our estimated $F_{2,CO2}$ for HadGEM2-ES of 3.18 W m$^{-2}$.
their amipPiForcing experiments than over years 1–20 of the abrupt4xCO2 simulations. They found two of them conflicted with their calculated trends in $\lambda$, leading them to favor the importance of the third explanation, namely that unforced variability strongly influenced historical variations in SST patterns. However, Zhou et al. (2016) found that CMIP5 control simulations realistically estimate internal variability on decadal time scales, 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 $\Delta R$ value for the final 15 years of the amipPiForcing run implied by the $\lambda$ value HadGEM2-ES exhibits early in its abrupt4xCO2 simulation even 30% toward its actual amipPiForcing value (Fig. S5 in the supplemental material). Our finding that the relationship between pentadal $\Delta R$ and $\Delta T$ in HadGEM2-A during its amipPiForcing experiment is stable, apart from two excursions, (Figs. 5a and S3), strongly points to the observed SST pattern evolution being largely forced and to much lower $\lambda$ values in years 1–20 of HadGEM2-ES’s abrupt4xCO2 experiment reflecting unrealistic simulated SST pattern evolution. It follows that there is no reason to believe that energy budget sensitivity estimates based on changes over the full historical period are biased downward by internal variability in SST patterns.

Observational estimates of the relationship between $\Delta R$ and $\Delta T$ throughout the historical period are also relevant. We estimate $\lambda$ using all 15-yr periods in 1927–2016, as well as by regression over 1872–2016, anomalizing relative to the 1850–84 base period. Average volcanism in 1850–84 matches that over both 1927–2016 and 1872–2016, and when using 2007–16 anomalies that base period gives the same $\lambda$ estimate (2.29 W m$^{-2}$ K$^{-1}$, corresponding to an ECS of 1.66 K) as per the main 2007–16 based results with globally complete $\Delta T$. Until recent decades $\Delta R$ was unobserved; we approximate it by scaling $\Delta F_{\text{pro rata}}$ over the historical period (Gregory and Forster 2008). We scale $\text{ERF}_{\text{Volcano}}$ by 0.55 to adjust for its low efficacy. Our no-intercept pentadal regression fit over 1872–2016 gives $\lambda = 2.27$ W m$^{-2}$ K$^{-1}$. Post-1926 ($\Delta R, \Delta T$) pentadal means (Fig. 5b) cluster around the best-fit line, while most of the 15-yr means lie almost on it.

The considerable stability of observationally based $\lambda$ 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 and 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 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 CO₂ forcing may be viewed (and mathematically modeled) as reflecting a subdecadal time scale 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 than feedback analysis suggests is possible. While the (lower) underlying climate feedback parameter is 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 subdecadal time scale responses are complete, and that the climate feedback parameter associated with subdecadal time scale responses, if not regarded as partially associated with adjustment processes, ranges up to 3 W m⁻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 CO₂ 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 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. (2016) 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 differentiated 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. (2016) 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 in the supplemental material). Recalculating efficacy using warming over 2000–05, scaling year-2000 historical ERF by the ratio of average 2000–05 iRF to year-2000 iRF, raises the Marvel et al. (transient) efficacy of historical ERF to 1.00 (see section S3 in the supplemental material). Consistent with this, Hansen et al. (2005) estimated (transient) efficacy relative to historical ERF derived by regression as marginally above one.

Marvel et al. (2016) also derived a new efficacy metric, called equilibrium efficacy, that accounts for variation in heat uptake efficiency between forcings. However, their methods also bias downward their historical forcing equilibrium efficacy estimates. Recalculating equilibrium efficacy for historical ERF using the same mean 2000–05 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 CO2-only forcing over a similar time period (see section 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 closest 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. (2015) 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 (2014b) found that in reality the effect on temperature estimates of assuming sea ice extent was fixed (in which case no bias arises) was minimal.

Second, Cowtan et al. (2015) 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 Hadley Centre Night Marine Air Temperature version 2 (HadNMAT2; Kent et al. 2013) dataset shows a lower global trend in near-surface marine air temperature over its 1880–2015 period than does the Hadley Centre SST version 3.1.1.0 (HadSST3.1.1.0) dataset, 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–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\(^{-1}\) in both cases, and lower than the 0.18 K decade\(^{-1}\) 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 even be negative). While incorporating an extra multiplicative uncertainty with a standard deviation of 4% in all the Had4_krig_v2 \(\Delta T\) values might nevertheless be justified, it would not alter any ECS or TCR 5%–95% uncertainty range by more than \(\pm 0.01\) K.

---

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. These data are a blend of 2-m temperature over land and SST over ocean (Paul Berrisford, ECMWF, 2016, personal communication). 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. Those data are for surface air temperature anomalies. Both sets of data have been adjusted by ECMWF for inhomogeneities in their source of analyzed SST.

**f. ECS versus ECS\(_{\text{hist}}\)**

The possibility that energy-budget climate sensitivity estimates based on changes over the historical period, which measure \(\lambda\) 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. In AOGCMs, ECS\(_{\text{hist}}\) can be quantified fairly accurately, 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 realizations of model internal variability (see section S4 in the supplemental material). The three methods provide almost identical ensemble-mean ECS\(_{\text{hist}}\) estimates (Table S2 in the supplemental material). Over the entire ensemble, ECS varies between 0.91 and 1.52 times ECS\(_{\text{hist}}\), the median ratio being 1.095, very close to the 1.096 ratio estimated by Mauritzen and Pincus (2017), Armour (2017) and Proistosescu and Huybers (2017) reported higher ECS-to-ECS\(_{\text{hist}}\) ratios (respectively a 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, hereinafter A17) to our 1.095 ratio is as follows. We provide a similar reconciliation for Proistosescu and Huybers (2017) in section S5 of the supplemental material.

1) A17, in calculating ECS\(_{\text{hist}}\) values (therein termed ECS\(_{\text{inter}}\)), estimated \(F_{2\times\text{CO2}}\) from the y-axis intercept when regressing \(\Delta N\) against \(\Delta T\) over years 1–5 of abrupt4xCO2 simulations. Doing so does not provide an unbiased ERF basis estimate of \(F_{2\times\text{CO2}}\), since during year 1 the CO\(_2\) top-of-the-atmosphere forcing is moving from its instantaneous value toward its ERF value as the stratosphere, troposphere, and other annual- or shorter-time scale 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-1 data significantly increases the mean \(F_{2\times\text{CO2}}\) estimate, resulting in lower ECS\(_{\text{hist}}\) estimates. We regress over years 2–10, avoiding bias from not fully adjusted year-1 data; time-variation of \(\lambda\) is insignificant in the first decade. A17’s ensemble-mean ECS-to-ECS\(_{\text{hist}}\) ratio calculated using regression over years 2–10 to
determine $F_{2\times CO_2}$, 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 the 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% upward bias in A17’s mean ECS$_{hist}$ estimate (which is based on $\Delta N$ and $\Delta T$ values in years 85–115 of 1pctCO$_2$ simulations) but a 4.6% upward bias in A17’s mean ECS estimate (which is based on abrupt4xCO$_2$ 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 ordinary least squares (OLS) regression of annual-mean years 21–150 abrupt4xCO$_2$ $\Delta N$ and $\Delta T$ values, but $\Delta T$ as well as $\Delta N$ 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 1985). Using Deming regression to derive unbiased ECS estimates (see section S4 in the supplemental material), 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$_{hist}$ (A17); it tends to do so and is generally moderate ($\approx 1.16$) 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 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% range: 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 5%–95% range of 0.79–1.40.

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 it known how accurately ECS can be estimated from 150-yr abrupt4xCO$_2$ 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 realized 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 $\Delta N$), 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
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 $\Delta T$, $\Delta F/F_{2\times CO_2}$, and $\Delta N/F_{2\times CO_2}$ 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 (from 1859–82 to 2007–16) estimates from the main analysis using the globally complete Had4_krig_v2 dataset. The bottom row shows the observationally derived uncertainty ranges for the three variables. Best estimates and uncertainty ranges are derived from the same one million samples used for the main statistical analysis.

<table>
<thead>
<tr>
<th>Parameter and lowest inconsistent value in CMIP5 model ensemble</th>
<th>$\Delta T$ (K)</th>
<th>$\Delta F/F_{2\times CO_2}$</th>
<th>$\Delta N/F_{2\times CO_2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ECS$_{hist}$: 2.89 K</td>
<td>1.54</td>
<td>0.43</td>
<td>0.36</td>
</tr>
<tr>
<td>TCR: 1.91 K</td>
<td>1.26</td>
<td>0.46</td>
<td>—</td>
</tr>
<tr>
<td>Observational 5%–95% range</td>
<td>0.73–1.03</td>
<td>0.49–0.83</td>
<td>0.06–0.21</td>
</tr>
</tbody>
</table>

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% range: 1.05–2.45 K) for ECS$_{hist}$ and 1.20-K median (5%–95% range: 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% range: 1.15–2.7 K) and 1.33 K for TCR (5%–95% range: 1.0–1.9 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 downward 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% range: 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 CH4 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 and/or estimation error on energy budget sensitivity estimates. The estimates using the 1869–82 base period and 2007–16 final period combination are preferred; they have the highest $\Delta T$ and $\Delta F$ values and as a result are best constrained. Moreover, with the Argo ocean-observing network fully operational throughout 2007–16, 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 1869–82, the fact that TCR estimation is very similar using the higher-coverage 1930–50 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 $\Delta T$, 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 [4.0 K in Annan and Hargreaves (2013) and 5.0 K in Friedrich et al. (2016)] and forcings (total 9.5 W m$^{-2}$; Köhler et al. 2010) 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 $F_{2\times CO_2}$ per AR5, since the WMGG forcings were derived using AR5 formulas), in line with the median obtained by scaling this study’s ECS$_{hist}$ estimate.
ΔF/ΔF\textsubscript{2\textsubscript{CO\textsubscript{2}}}, and ΔN/ΔF\textsubscript{2\textsubscript{CO\textsubscript{2}}} (thus taking into account uncertainty in \(F_2 CO_2\)) confirms that in each case the lowest CMIP5 model TCR and ECS\textsubscript{bias} values that we find to be inconsistent with observed warming imply, implausibly, that ΔT, ΔF/ΔF\textsubscript{2\textsubscript{CO\textsubscript{2}}}, and ΔN/ΔF\textsubscript{2\textsubscript{CO\textsubscript{2}}} have values outside their uncertainty ranges.

The implications of our results are that high best estimates of ECS\textsubscript{bias}, 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.

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

REFERENCES


——, and ——, 2014b: Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Update: Temperature reconstruction by domain: Version 2.0 temperature series. 9 pp., http://www.webcitation.org/609b9N8vM.


