Recommended temperature metrics for carbon budget estimates, model evaluation and climate policy

Katarzyna B. Tokarska[1,2*], Carl-Friedrich Schleussner[3,4,5], Joeri Rogelj[1,6,7], Martin B. Stolpe[1], H. Damon Matthews[8], Peter Pfleiderer[3,4,5], and Nathan P. Gillett[9]

[1] Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
[2] School of Geosciences, University of Edinburgh, Edinburgh, United Kingdom
[3] Climate Analytics, Berlin, Germany
[4] Integrative Research Institute on Transformations of Human-Environment Systems (IRI THESys), Humboldt-Universität zu Berlin, Berlin, Germany
[5] Potsdam Institute for Climate Impact Research, Potsdam, Germany
[6] Grantham Institute, Imperial College London, London, United Kingdom
[7] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[8] Concordia University, Montréal, Canada
[9] Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, University of Victoria, Victoria, BC, Canada

*corresponding author: kasia.tokarska@env.ethz.ch

Recent estimates of the amount of carbon dioxide that can still be emitted while achieving the Paris Agreement temperature goals are larger than previously thought. Different temperature metrics used to estimate the observed global mean warming for the historical period affect the size of the remaining carbon budget. Here we explain the reasons behind these remaining carbon budget increases, and discuss how methodological choices of the global mean temperature metric and the reference period affect remaining carbon budget estimates. We argue that the choice of the temperature metric should depend on the domain of application. For scientific estimates of total or remaining carbon budgets, globally averaged surface air temperature estimates should be used consistently for the past and the future. However, when used to inform the achievement of the Paris Agreement goal, a temperature metric consistent with the science that was underlying and directly informed the Paris Agreement should be applied. The resulting remaining carbon budgets should be calculated using the appropriate metric or adjusted to reflect these differences among different temperature metrics. Transparency and understanding of the implications of such choices are crucial to providing useful information that can bridge the science-policy gap.

Carbon budgets provide a tool to clearly communicate that limiting global warming to a particular level implies a cap on global total CO_2 emissions[1]. Defined as the total amount of CO_2 that can be emitted while keeping global warming below a given level with some probability, carbon budgets emerge from an approximately linear relationship between warming and cumulative CO_2 emissions, known as the Transient Climate Response to cumulative CO_2 Emissions (TCRE)[2–5]. TCRE and the related carbon budgets were initially derived under idealized CO_2-only emission scenarios[2]. However, under real-world conditions, several factors complicate the simplicity and clarity of the carbon budget concept. Emissions other than CO_2 (such as methane, soot, or sulphate aerosols) also affect both global temperature and the state of carbon sinks (albeit to a smaller extent than CO_2 itself [6–9]), and hence the size of the remaining carbon budget. In addition to CO_2 emissions from fossil fuels (which are well known), CO_2 emissions from other land-use change represent a quarter of historical CO_2 emissions: these emissions are difficult to diagnose, and are subject to large uncertainty both in models[10,11] and in estimates derived from historical data based on energy and industry statistics and land-use book-keeping methods[12]. To further complicate matters, estimates of historical warming since pre-industrial times come with uncertainties due to limited observational coverage[13], instrumental uncertainty, and uncertainties associated with constructing long-term temperature datasets[14]. Global warming can also be expressed in different ways, for example, as near-surface air temperatures covering the entire globe or as a combination of sea surface temperatures over open ocean and near-surface air temperature elsewhere[15,16], averaged over locations where observations are present. Finally, inter-annual and decadal variability adds further complications[17].

Recently, several studies[18–20] and the assessment of the Special Report on Global Warming of 1.5 °C (SR1.5)[21] of the Intergovernmental Panel on Climate Change (IPCC) introduced a new approach to estimate the remaining carbon budget. These studies report model-based remaining carbon budgets for the additional warming from today until we reach 1.5 °C or 2 °C of anthropogenic warming. This was a departure from the previous approach of estimating the total carbon budget since pre-industrial times, and then reporting the remaining budget by subtracting emissions to date. The new approach in SR1.5 is a kind of bias correction, since it corrects for any inconsistencies in simulated and observed warming as a function of cumulative emissions over the historical period, and can potentially decrease uncertainties in estimates of the remaining carbon budget, especially for levels of warming relevant to the Paris Agreement[22]. Because the remaining carbon budgets for 1.5 °C or 2 °C are small, even adjustments that are limited in absolute terms result in large relative changes. For example, recent estimates of the remaining carbon budget for 1.5 °C are larger by more than a factor of two when compared to those reported in the IPCC Fifth Assessment Report (AR5)[4,23] (see Figure 2 in Ref.[24] and their Supplementary Table 2 for a comprehensive comparison of the remaining carbon budget estimates from different studies). This difference can be partly understood as a result of a higher temperature response to cumulative CO_2 emissions in the Coupled Model Intercomparison Project Phase 5 (CMIP5)[25] models used to inform the AR5 carbon budgets, compared to estimates of historical CO_2 emissions and warming[16,26]. However, recent insights related to uncertainty in the observational temperature record also suggest that part of the difference among carbon budget estimates is related to the method of calculating historical warming that is used in the analysis[27].

Here we explain the reasons why the carbon budget estimates expressed relative to a more recent reference period differ from previous ones, and separate these into differences caused by carbon cycle and temperature-driven components. We then clarify how the choice of temperature metric affects the size of remaining carbon budget estimates, and we emphasize the need for transparency and clarity about its implications. Finally, we provide recommendations for future estimates of remaining carbon budgets along with remaining challenges.

Effects underlying adjustments of the baseline

The effect of changing the baseline to a more recent period (from R to R'; Figure 1, both panels), can be separated into carbon cycle effects (arrow 1), and temperature effects (arrow 2). First, the Earth System Models (ESMs) that were used to estimate the carbon budgets reported in IPCC AR5, on average, underestimated carbon uptake (by land and ocean carbon sinks) in prescribed CO_2 concentration simulations. As a result, these models on average estimated lower cumulative CO_2 emissions over the historical period compared to CO_2 emissions estimated from independent fossil-fuel use and other data[18,19]. Updating the baseline to account for this carbon cycle bias, therefore, leads to an increase in the remaining carbon budget compared with those reported in IPCC AR5 (Figure 1 a,b, arrow 1). Second, accounting for a possible difference in warming over the historical period results in a second offset (Figure 1 a,b, arrow 2). Since the global mean temperature has already increased by about 1 °C above pre-industrial levels[28], even minor corrections arising from methodological adjustments or model biases can have a sizeable effect on the remaining 1.5 °C budget.

Remaining carbon budgets are often based on the likely (>66 % probability) TCRE range assessed by IPCC AR5[29] of 0.8 to 2.5 °C/1000 PgC (where 1 PgC = 3.67 $GtCO_2$). Several recent studies[18,19] that updated the baseline did not alter the resulting TCRE range: i.e. they used the same slope for the relationship between temperature and cumulative emissions (TCRE) before and after changing the baseline, as illustrated in schematic Figure 1a. Another approach would be to adjust the slope of TCRE relationship to align the TCRE with the lower temperature response to emissions implied by updating the baseline to a more recent period. In principle, both carbon-cycle and temperature adjustments could lead to changes in the rate of warming as a function of cumulative emissions, as illustrated in Figure 1b. Whether such an adjustment is warranted depends on the assessment of the validity of extrapolation of historical to future warming as a function of cumulative emissions. Little correlation exists between cumulative emissions at present-day warming and at 1.5 °C across the CMIP5 ensemble[19] likely due to differences in response to non-CO_2 forcing across models. Hence, we would caution against scaling simulated 1.5 °C carbon budgets based on the ratio of simulated to observed historical warming as a function of cumulative CO_2 emissions, given the important and uncertain role played by non-CO_2 forcings in historical climate change. Identifying the conditions under which the slope of TCRE would require an adjustment needs further research. Expressing carbon budgets relative to a recent reference period (e.g. using the 2006-2015 reference period instead of the pre-industrial baseline) is intended to minimize the effect of uncertainties arising from mismatches between modelled and observed cumulative CO_2 emissions and warming in the historical period. However, such adjustment of the baseline does not involve a correction for the models' processes that led to those discrepancies in the historical period.

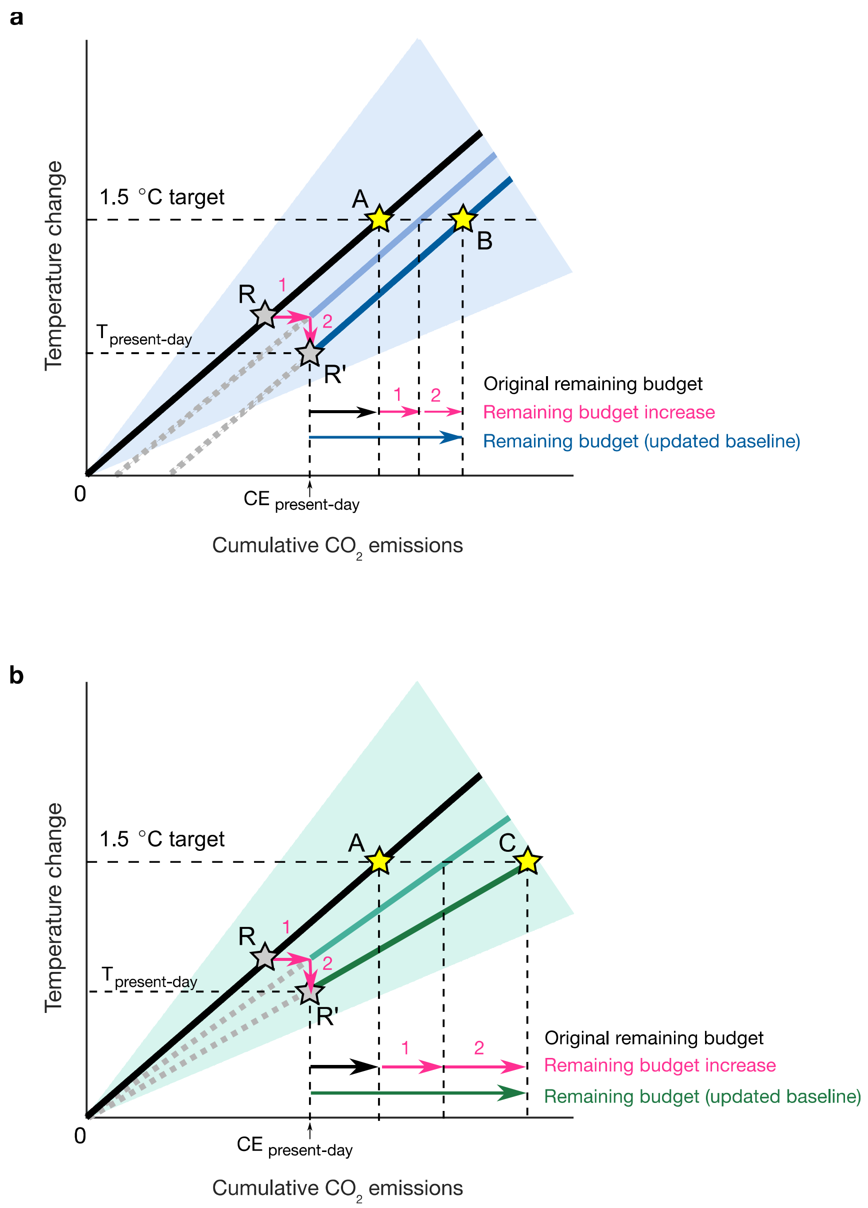


Figure 1 | Schematic representation of the effects of updating the baseline with respect to the cumulative CO_2 emissions and temperature change on estimates of the remaining carbon budget. Remaining carbon budgets after updating baseline **(a)**; and with scaling of future warming **(b)**. On either panel, Arrow 1 represents the carbon cycle effect (correction for model biases in historical CO_2 emissions); Arrow 2 represents the temperature effect (arising from the differences between modelled and observed warming). The first yellow star (A) indicates the initial carbon budget at the 1.5 °C warming level with the original reference period (R). The second yellow star (B or C) indicates the final (and larger) remaining carbon budget, calculated after updating the baseline to a present-day reference period (R'). Shaded area represents the spread of the relationship between temperature and cumulative CO_2 emissions. The present-day level of warming and cumulative CO_2 emissions is indicated by the dashed lines, as labelled, though the figure is meant for illustrative purposes only.

Temperature metric choices

While the correction for carbon cycle effects is relatively straightforward, attempts to assess consistency between warming estimates based on model output and observations have highlighted questions surrounding the choice of the method used to estimate changes in global mean temperature[30]. One way of expressing the global mean temperature is Global mean Surface Air Temperature (here referred to as GSAT), usually estimated in models by calculating the modelled global average Surface Air Temperature (SAT) – the temperature at about 2 m above the Earth's surface. By contrast, the observed global mean temperature is constructed by combining observational measurements of surface air temperature over land and sea ice (SAT) with Sea Surface Temperature (SST) measurements for open ocean locations. This blended temperature is referred to as GBST, or Global mean Blended Surface Temperature. Importantly, GBST estimates based on observational measurements do not sample the full globe. Some datasets use statistical infilling techniques to account for this and estimate the global temperature implied by nearly full observational coverage (e.g. GISTEMP[31], HadCRUT-CW[32] and Berkeley Earth[33]). Others provide estimates using only data where measurements are available (e.g. HadCRUT[34]). Estimates that use observations thus reflect the blended (SST + SAT), and in some cases masked (incomplete coverage without statistical infilling), estimates of global mean temperature. Relative to GSAT, both blending and masking in the GBST metric reduce the estimated warming[15,26], and statistical infilling might not always alleviate the masking bias when instrumental coverage is low[13]. Furthermore, both the masking and blending effects are time-dependent: (i) the observational mask will change over time as the distribution of measurements changes, and (ii) the use of SST vs SAT measurements can also change as a result of changing sea-ice coverage leading in general to more open water (and hence SST measurements) over time. This time-dependent blended-masking effect lowers warming since pre-industrial by about 0.1°C during the 10-year average reference period used in the IPCC SR1.5 report (2006-2015). This difference increases with additional warming[16,30].

To estimate remaining carbon budgets relative to a present-day reference period, an estimate of the present-day level of warming is needed in order to determine the amount of warming that is left until 1.5 °C or any other temperature level would be reached. Given a median estimate of TCRE (Refs.[4,29]), a difference in global mean temperature of 0.1 °C, either as a result of a different temperature limit or as a result of a different estimate of warming to date, would alter carbon budget estimates by about 200 $GtCO_2$ (Refs.[21,30]).

Beyond blending-masking adjustments

The multi-model mean GSAT change of the CMIP5 ensemble[25] matched well with GBST observations (HadCRUT4.6; Ref.[34]) up to the 1986-2005 period, which is the reference period used by IPCC AR5 (Ref.[35] Table 1.1 therein). However, the mean of the simulated CMIP5 GSAT warming between 1986-2005 and 2006-2015 (the updated SR1.5 reference period) lies above observation-based estimates. While the observed warming between these periods was within the range of simulated warming in the CMIP5 ensemble, the CMIP5 multi-model mean GSAT increase of 0.38 °C was larger than the GBST warming in HadCRUT4.6 of only 0.22 °C. The differences between various observation-derived GBST metrics, as well as the effect of accounting for the difference in GBST and GSAT definitions and incomplete coverage of observations, can only partly explain this difference (accounting for coverage and blending of SST and SAT reduces modelled warming to 0.33 °C, Figure 2b).

Several additional reasons have been suggested to reconcile the remaining mismatch between the multi-model mean and observations[36]. We identify three main groups of effects that might contribute to the differences between models and observations of GBST (Figure 2b). First, the SST dataset of HadCRUT4.6, HadSST3, shows a significant cooling bias from around year 2005 onwards, when compared to instrumentally homogeneous SST records from drifting buoys, Argo floats, and satellites[37]. This and other biases in the SST record have been recently addressed in HadSST4 (Ref.[38]). The increase in GBST between the two reference periods, 1986-2005 and 2006-2015, is however virtually unchanged as HadSST4 is warmer during both reference periods than HadSST3 (compared to pre-industrial baseline). The choice of the SST dataset, therefore, appears only to have a small influence on the divergence between modelled and observed warming, but uncertainties in the temperature record remain. Second, from the early 1990s, Pacific trade winds intensified, enhancing equatorial upwelling in the central and eastern Pacific. This reduced the SSTs in that region, thereby also reducing the pace of global mean temperature increase[39,40]. These effects of internal variability in the Pacific region lower the observed global mean temperature increase between the two reference periods by roughly 0.08 °C (with a range of -0.03 to -0.20 °C across published estimates), (Figure 2b, 'Pacific Variability effect' green bars). Third, a series of small-to-moderate-magnitude volcanic eruptions have led to an increase in stratospheric aerosols after the year 2004[41,42], which is neglected in CMIP5 model projections. Furthermore, CMIP5 radiative forcing projections also assume that the last solar cycle prior to 2005 is repeated in the subsequent period. As a result, the assumed recent solar forcing in the model projections is too large when compared with observations[36,41,43]. Correcting models to account for both the updated solar forcing and updated volcanic forcing, reduces the modelled global mean temperature increase between the two reference periods, but effects from revised anthropogenic tropospheric aerosols[44] are uncertain and might have reduced[43] or increased the warming[45]. Overall, the assessed studies indicate that warming changes by -0.08 to +0.02 °C from updated forcing between the two reference periods (Figure 2b, 'Updated Forced Signal effect', teal bars). The CMIP6 models[46] are forced with updated radiative forcings, and while some models indicate reduced warming in the early 21st century, explained partly by updated forcing[47], the set of available models simulates slightly more warming between the two reference periods as CMIP5. The models underwent major changes in the model physics leading to an increase in climate sensitivity[48], which might increase the warming between the two reference periods[49].

While the strength of the effects is considerably uncertain, and there might be further aspects not considered here, we note that modelled and observed GBST warming between the 1986-2005 and 2006-2015 periods can be fully reconciled within the uncertainty ranges of the different contributing effects (Figure 2), and moreover we note that multi-model mean GBST warming in 2006-2015 relative to the 1850-1900 base period is very close to the best observational estimates[35]. This highlights that warming expressed in two different temperature metrics (GBST and GSAT) can be made internally consistent by carefully accounting for various effects, and used to compare models and observations for the historical period.

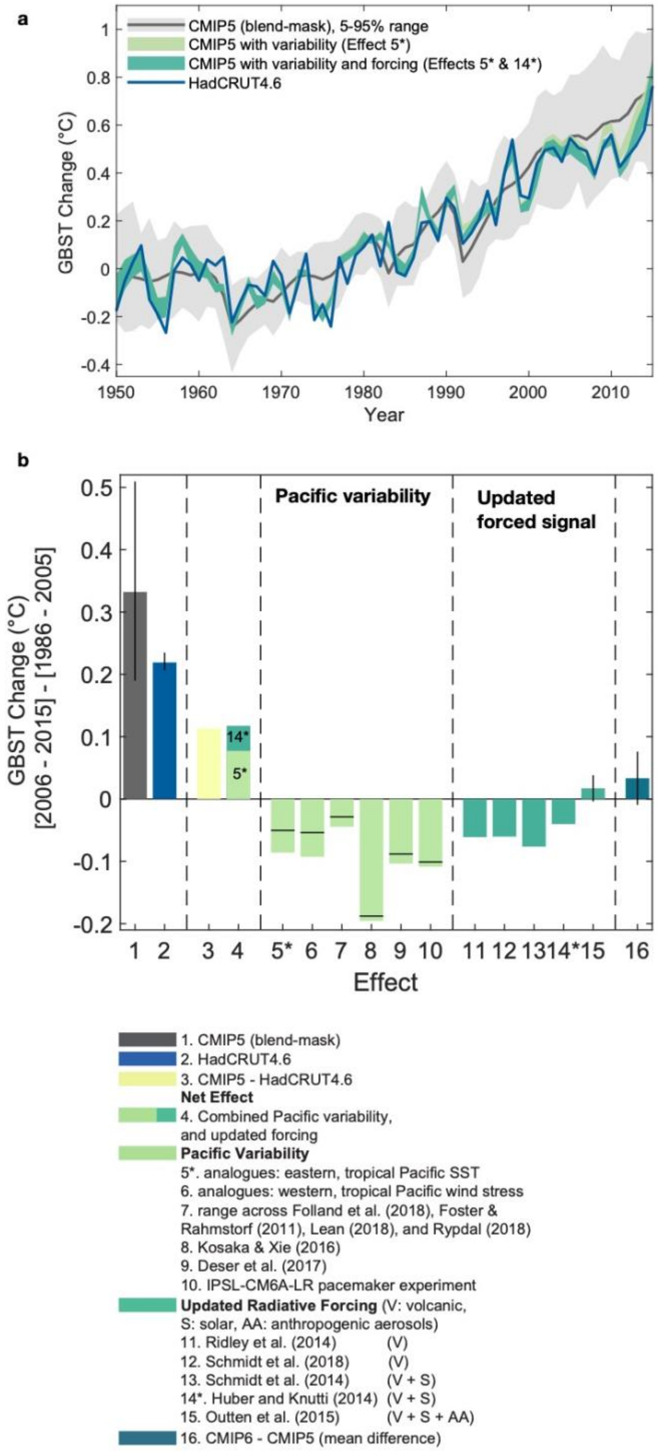


Figure 2 | Contributions to differences in recent observed and modelled warming. Time-series of modelled and observed warming (**a**), with different effects leading to adjustments in observed and modelled GBST (**b**). The length of the bars (horizontal black lines) shows upper (lower) estimates of the influence of Pacific variability on warming. The spread arises from uncertainty in both observations and the forced signal (effects 5 and 6), from missing years (effects 8 to 10), and reflects the range across four studies (effect 7). Vertical black lines indicate 5-95% uncertainty ranges. Effects indicated by an asterisk are used for the net effect shown as bar 4. The global mean temperature base period is 1961-1990 in panel (**a**), and 2006-2015 relative to 1986-2005 in panel (**b**). (See *Methods* for details and references).

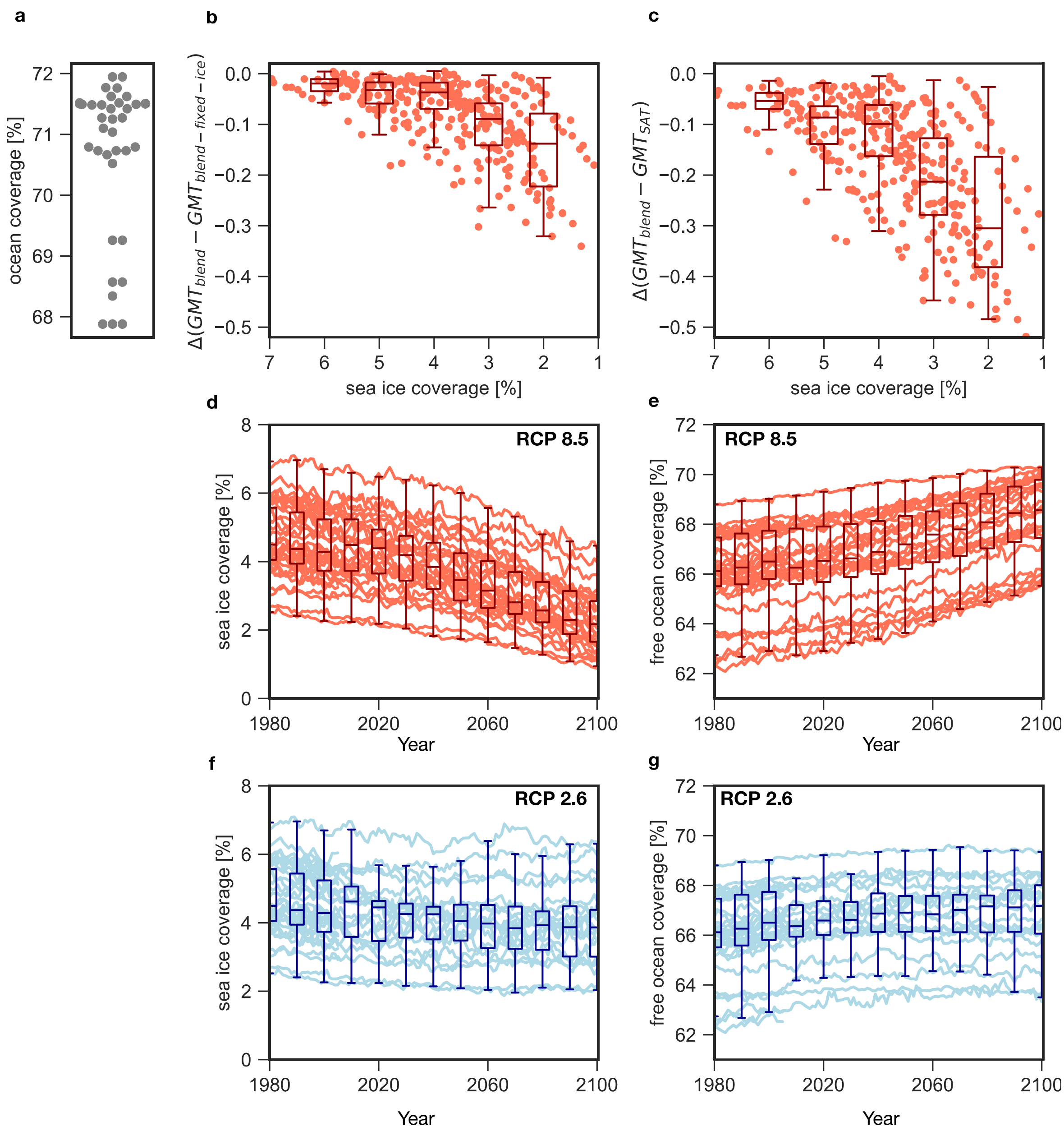


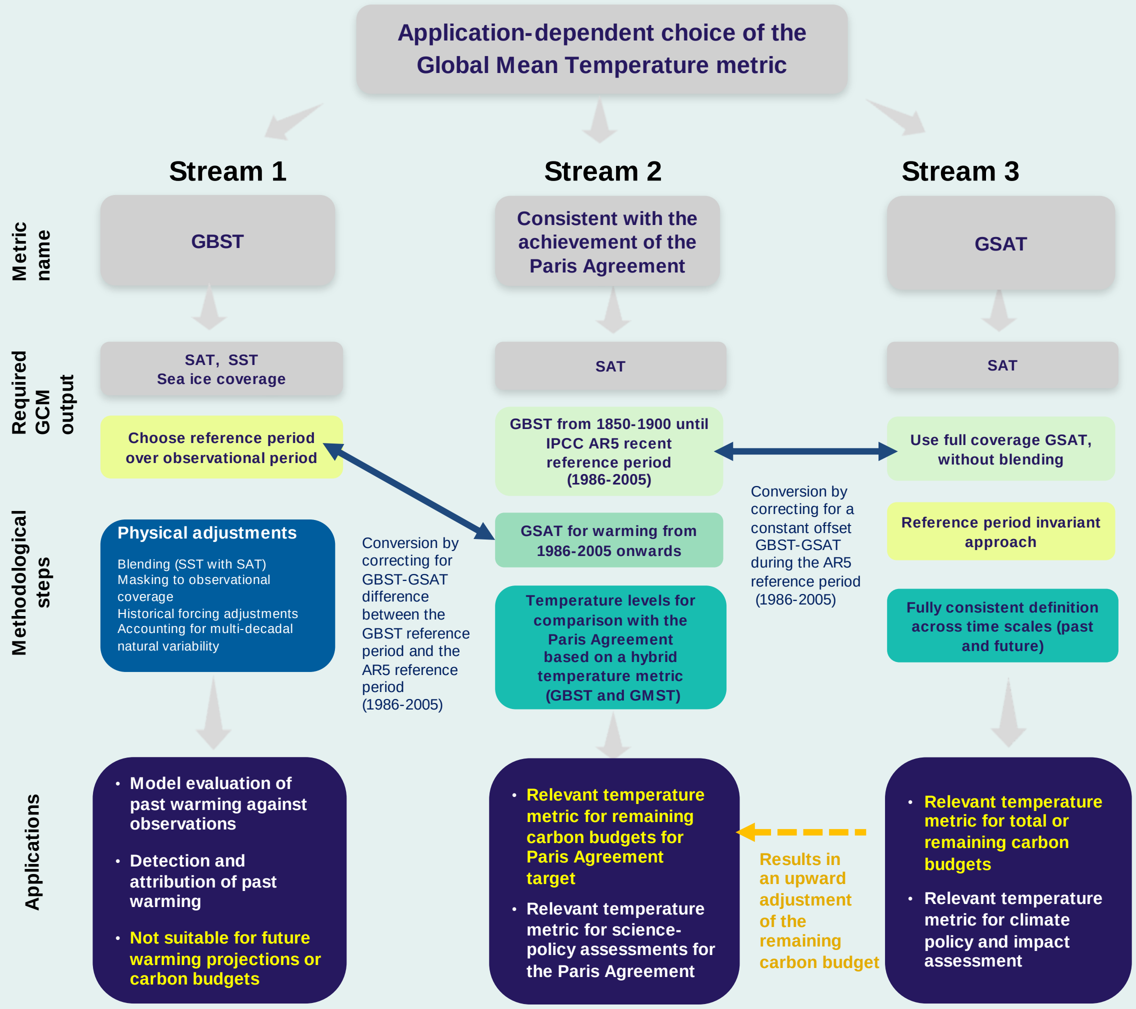
Figure 3 | Differences in ocean and sea ice coverage in CMIP5 models, and related differences between GBST and GSAT metrics, under different future emission scenarios[50] (RCP 8.5 and RCP 2.6). Swarm plot of the time-invariant, constant field defining ocean grid-cells ('sftof' CMIP variable) (**a**); the sea-ice effect, shown as a difference between GBST and GBST with fixed sea ice mask (**b**); the overall blending effect, shown as a difference between GBST and GSAT, as a function of sea ice coverage (**c**); time-series of the time evolution of sea-ice fraction in RCP 8.5 (**d**); time-series of the evolution of the free ocean area in RCP 8.5 (**e**); time-series of the time evolution of sea-ice fraction in RCP 2.6 (**f**); time-series of the evolution of the free ocean area in RCP 2.6 (**g**); *Note: In panels (**b**) and (**c**) boxplots are shown for five sea ice coverage levels: 6.5 - 5.5%, 5.5 - 4.5%, 4.5 - 3.5%, 3.5 - 2.5% and 2.5 - 1.5%. In panels (**d**) to (**g**), boxplots show interquartile ranges for 10-year time slices.*

Application and consistency

Different temperature metrics come with their respective strengths and weaknesses. A GSAT estimate will, by definition, draw from the surface air temperature field everywhere across all models. In contrast, GBST is a composite of land surface air temperature and sea surface temperature, and GBST estimates depend on the ratio of land and sea ice versus ocean across the Earth' surface. The share of free ocean coverage differs between models by about 7 percentage points (Fig. 3 e,g) due to differences in present-day sea ice (Fig. 3 d,f) and the land-sea share in the model grid (Figure 3a, *Methods*).

The land and sea ice versus ocean ratio does not only differ among models, but also among various runs from the same model due to internal variability, as well as over time as a result of differences and changes in sea-ice cover. Therefore, the GBST metric is dependent on model, time and even realisation within the model ensemble itself. Such differences complicate comparison of GBST estimates among models or even within ensemble members of the same model. Due to the combination of these challenges surrounding the GBST temperature metric, the GBST metric is not well-suited for projections of future warming levels (e.g. 1.5 °C or 2.0 °C), for which remaining carbon budgets are calculated.

Given the various possible choices regarding methods of calculating global mean temperature rise and their effect on estimates of remaining carbon budgets, we summarize recommended approaches in Box 1. We identify three main streams of application, and for each, we recommend an appropriate metric for estimating the global mean temperature level and estimate of remaining carbon budgets. These streams depend on the purpose of the application: (i) Model evaluation of global mean temperature against observations or detection and attribution analysis of global mean temperature (Box 1, Stream 1); (ii) assessments of temperature estimates and carbon budgets for the Paris Agreement goal (Box 1, Stream 2); and (iii) Assessing carbon budgets or impacts across time and for future levels of warming with a consistent definition of temperature change (Box 1, Stream 3).



Box 1| Different choices and recommendations for the use of global mean temperature metrics, depending on the application domain, illustrated in the following three Streams. The appropriate use of temperature metrics for carbon budget calculations is shown in yellow.

Stream 1, using the GBST temperature metric uniquely, allows a consistent comparison with global mean temperature estimates currently provided by observational temperature products (e.g. the HadCRUT4.6 dataset[34]). Unless observational products routinely also provide estimates of global near-surface air temperatures (GSAT), the GBST metric is so far the best choice for applications related to model evaluation of historical warming with the observations and detection and attribution[51]. However, this metric of choice for Stream 1 presents challenges when applied to future warming projections (see above discussion of Figure 3). Therefore, this metric is not recommended for calculating remaining carbon budgets (that use future warming projections).

Stream 3, using the GSAT temperature metric uniquely, provides a consistent estimate of global mean temperature increase in model simulations for both the historical period and into the future. Estimating global mean temperature increase uniquely based on GSAT with full global coverage allows achieving such consistency over time. Therefore, we recommend using GSAT as the primary temperature metric for Stream 3 applications, including remaining carbon budget calculations. This would also ensure consistency with some impact assessment studies that use model simulations from a pre-industrial baseline and use a spatially-complete temperature metric across time-scales.

Between Stream 1 and 3, lies Stream 2, with applications intending for the assessments of global mean temperature and carbon budgets to be consistent with the achievement of the Paris Agreement target. The Paris Agreement did not specify explicitly which temperature metric applies to the warming levels of 1.5 °C and well-below 2 °C. This, however, does not mean that the temperature metric is *unknown*. The temperature goal of the Paris Agreement needs to be read in the context of the accompanying decisions under the United Nations Framework Convention on Climate Change (UNFCCC) and the science as reflected in the most recent IPCC reports at the time[52]. We, therefore, propose a Paris Agreement compatible temperature metric following the approach applied in the AR5, namely a hybrid product with GBST until 1986-2005 and GSAT for warming from 1986-2005 onwards.

For a direct comparison of studies using uniquely the GBST metric only (Stream 1; e.g. studies of model evaluation or detection and attribution of historical warming[51]) with the temperature metric that is consistent with the achievement of the Paris Agreement (i.e. a hybrid of GBST and GSAT metrics; Stream 2), the difference between the GBST and GSAT metrics over the period between the GBST study's reference period and the AR5 recent reference period (1986-2005) has to be accounted for (indicated by the blue arrow between Stream 1 and Stream 2). For the 2006-2015 reference period, this adjustment is about 0.16 °C and is the difference between modelled GSAT and the observed masked GBST evolution applied to the same model runs (see Methods and SR1.5 Table 1.1).

We do not recommend using GBST metric for future projections, because this would require implementing model specific and time-varying adjustments (due to changing sea-ice coverage; see Figure 3 and its discussion) to bring these estimates in line with the Paris Agreement compatible Stream 2 metric. On the other hand, for a direct comparison of results from studies using uniquely the GSAT metric (Stream 3; e.g. carbon budgets for future levels of warming) and the Paris Agreement-consistent temperature levels (Stream 2), a constant adjustment for the difference between GSAT and GBST during the 1986-2005 period (i.e. the AR5 reference period) relative to the 1850-1900 reference period in HadCRUT4 needs to be made (indicated by the blue arrow between Stream 3 and Stream 2). In the CMIP5 multi-model mean, this offset is very small (up to about 0.03 °C) compared to the 5-95% uncertainty range of the observational product (HadCRUT4 observed warming from 1850 -1900 to 1986-2005 is reported to be 0.57 to 0.66 °C, with a central estimate at 0.6 °C; Ref.[35]; Table 1.1 therein). The transition from Stream 3 to Stream 2 is independent of the chosen baseline or period of interest. For studies using CMIP5, translating results obtained with the full GSAT approach (Stream 3) to the Paris Agreement consistent metric (Stream 2) results in a constant upward adjustment of the remaining carbon budget by about 80 $GtCO_2$ (for a middle-of-the-range TCRE estimate of 1.65 °C/1000 PgC), but can depend on the precise assumptions. For studies using CMIP6 models[46], climate model emulators, or other approaches, this adjustment would need to be calculated according to those models.

Differences between temperature metrics such as GBST and GSAT were not thoroughly discussed in the literature available for the AR5, and thus could not be assessed by the IPCC before the SR1.5 was published in the year 2018. It hence cannot be expected that the 2015 Paris Agreement would be specific on the temperature metrics underlying its temperature goal. The same holds for other scientific concepts developed and assessed after the adoption of the Paris Agreement. However, the available literature at the time of AR5 can provide guidance on the metric consistent with the achievement of the Paris Agreement global mean temperature target.

The adoption of the Paris Agreement was informed by a multi-year process reviewing the temperature goal under the UNFCCC. This review process concluded in 2015 at adopting a long-term global goal under the Conference of the Parties (COP) that is identical to the Paris Agreement's Article 2.1(a)[22]. The process included a scientific arm, the so-called structured expert dialogue[52], that provided a comprehensive assessment of the impacts of climate change at 1.5 °C and 2 °C based predominantly on the IPCC AR5. The long-term temperature goal of the Paris Agreement is directly linked to this assessment and thereby the AR5 methodology[53,54]. The IPCC AR5 Working Groups 1 and 2 used GBST from 1850-1900 until the reference period 1986-2005 and GSAT for warming from the reference period onwards. We propose this temperature metric as being Paris Agreement compatible (Box 1 Stream 2). Paris Agreement compatibility is linked to the policy context and does not imply that such a hybrid temperature metric (GBST and GSAT) holds any specific scientific merit. As our scientific understanding progresses, new temperature metrics based on either new observational products or new analysis metric will become available, and could be scientifically superior. In order to not misguide policy by unintentionally shifting baselines, however, we recommend that any assessments aiming at informing the science-policy interface and the Paris Agreement should be expressed in, or at least provide a conversion to, the metric that is consistent with the achievement of the Paris Agreement (i.e. the hybrid of GBST and GSAT), presented in Stream 2, Box 1 (Refs.[24,30,53,54]). This will require conversion of temperature metrics (either in Stream 1 or Stream 3) to Stream 2 metric, illustrated in Box 1 by the two-headed arrows. Such conversion (to Stream 2) would lead to upward adjustments of carbon budgets (i.e. more allowable CO_2 emissions) calculated in Stream 3 (Box 1). This transition to Stream 2 is not exclusive to CMIP5 models, and could be applied, in principle, to any model-based temperature projections or carbon budgets that use the GSAT metric (Stream 3), and aim to report their results in the light of the Paris Agreement[22] (Stream 2).

Remaining challenges for the total carbon budget

Calculating the remaining carbon budget relative to a present-day reference period makes its estimates more accurate, as shown by recent studies[18–20] (see also Ref.[24] for a comprehensive summary of recent carbon budget estimates). However, changing the baseline to a more recent period is only a partial solution that does not address the underlying issue of discrepancies between CMIP5 models and observations in the historical period, particularly in their cumulative CO_2 emissions (as the temperature discrepancy between the models and observations can be addressed by comparing models and observations in a like for like manner). Moreover, changing the baseline does not help with constraining estimates of the total carbon budget for a given level of warming (i.e. including historical and future CO_2 emissions), which may be useful for assessing aspects of historical responsibility for past CO_2 emissions[55].

Implications for the science-policy interface

Calculating remaining carbon budgets relative to a recent reference period, rather than first calculating total carbon budgets relative to pre-industrial and then subtracting historical emissions, makes these estimates more accurate, providing a physically compelling reason to do so. However, such changes of the baseline to a more recent period also comes with political implications that one should be mindful of. Changing the reference period from pre-industrial times to the present-day shifts the focus of the study from estimating total carbon budgets and their relevance for the assessment of historical responsibilities and intergenerational or international equity, towards questions of our collective ability to avoid the exceedance of certain warming limits in line with the Paris Agreement.

Given the relevance of carbon budgets for climate policy, we recommend that methodological choices made in their estimation be fully transparent and traceable. Moreover, we recommend that assessments on the progress towards the Paris Agreement goals, including the carbon budgets for 1.5 °C, should provide a comparison to the temperature metric that is consistent with the achievement of the Paris Agreement (i.e. Stream 2 in Box 1). Due to different definitions of the temperature metrics discussed in this Perspective, carbon budgets calculated in Stream 2 are expected to be larger than carbon budgets calculated using temperature metric in Stream 3. Finally, although it may be challenging to constrain all the sources of uncertainty in estimating carbon budgets (e.g. Refs.[7,21,56–587]), the large spread in carbon budgets should not be used as an excuse to delay mitigation actions.

Ultimately, more than a decade of research on carbon budgets and the cumulative emissions framework demonstrates very clearly that reaching any global mean warming target that avoids dangerous climate change will require CO_2 emissions to be reduced to net-zero or net-negative[21] levels this century. The sooner this transition to declining emission rates begins, the smaller reliance on net-negative emissions is required in the future[21].

Correspondence and requests for materials should be addressed to K.B.T.

Acknowledgements

We are thankful for the discussions at the workshop on carbon budgets, co-organized by J.R., and attended by K.B.T., N.P.G., H.M.D., J.R., with the support of the Global Carbon Project, the CRESCENDO project, Stanford University, the University of Melbourne, and Simon Fraser University. We are grateful to E. Bush and A. Schurer for helpful insights. We thank K. Cowtan for providing the computer code for blending SAT and SST estimates. We thank I. Bethke, G. Foster, C.K. Folland, M. Huber, Y. Kosaka, J.L. Lean, K. Rypdal, and A. Schmidt for providing data used for Figure 2. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output. For CMIP the US Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

K.B.T, C-F.S., and J.R. were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 820829 (CONSTRAIN project). K.B.T. was also supported by the UK NERC-funded SMURPHs project (NE/N006143/1). C.F.S. and P.P. acknowledge support by the German Federal Ministry of Education and Research (01LN1711A).

Author contributions

C-F. S. initiated the study. K.B.T. wrote the manuscript with substantial inputs from C-F. S., J.R., M.B.S., H.D.M., and N.P.G. Figure 2 was done by M.B.S., Figure 3 was done by P.P., and the remaining figures were done by K.B.T., with suggestions from other authors. All authors participated in manuscript editing and revisions.

Competing Interests

The authors declare no competing interests.

References

1. Zickfeld, K., Eby, M., Matthews, H. D. & Weaver, A. J. Setting cumulative emissions targets to reduce the risk of dangerous climate change. *PNAS* **106**, 16129–16134 (2009).
2. Matthews, H. D., Gillett, N. P., Stott, P. A. & Zickfeld, K. The proportionality of global warming to cumulative carbon emissions. *Nature* **459**, 829–832 (2009).
3. Gillett, N. P., Arora, V. K., Matthews, D. & Allen, M. R. Constraining the Ratio of Global Warming to Cumulative CO_2 Emissions Using CMIP5 Simulations. *J. Climate* **26**, 6844–6858 (2013).
4. IPCC AR5. Stocker, T. F., D. Qin, G.-K. Plattner, L. V. Alexander, S. K. Allen, N. L. Bindoff, F.-M. Bréon, J. A. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J. M. Gregory, D. L. Hartmann, E. Jansen, B. Kirtman, R. Knutti, K. K. Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G. A. Meehl, I. I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L. D. Talley, D. G. Vaughan and S.-P. Xie (2013). Technical Summary. Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley. Cambridge, United Kingdom and New York, NY, USA, Cambridge University Press: 33-115. (2013).
5. MacDougall, A. H. The Transient Response to Cumulative CO_2 Emissions: A Review. *Curr Clim Change Rep* **2**, 39–47 (2016).

6. Gillett, N. P. & Matthews, H. D. Accounting for carbon cycle feedbacks in a comparison of the global warming effects of greenhouse gases. *Environ. Res. Lett.* **5**, 034011 (2010).
7. Tokarska, K. B., Gillett, N. P., Arora, V. K., Lee, W. G. & Zickfeld, K. The influence of non-CO_2 forcings on cumulative carbon emissions budgets. *Environ. Res. Lett.* **13**, 034039 (2018).
8. Matthews, H. D. *et al.* Estimating Carbon Budgets for Ambitious Climate Targets. *Curr Clim Change Rep* **3**, 69–77 (2017).
9. MacDougall, A. H., Zickfeld, K., Knutti, R. & Matthews, H. D. Sensitivity of carbon budgets to permafrost carbon feedbacks and non-CO_2 forcings. *Environ. Res. Lett.* **10**, 125003 (2015).
10. Arora, V. K. *et al.* Carbon–Concentration and Carbon–Climate Feedbacks in CMIP5 Earth System Models. *J. Climate* **26**, 5289–5314 (2013).
11. Arora, V. K. *et al.* Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases. *Geophysical Research Letters* **38**, (2011).
12. Quéré, C. L. *et al.* Global Carbon Budget 2017. *Earth System Science Data* **10**, 405–448 (2018).
13. Benestad, R. E., Erlandsen, H. B., Mezghani, A. & Parding, K. M. Geographical Distribution of Thermometers Gives the Appearance of Lower Historical Global Warming. *Geophysical Research Letters*
14. Cowtan, K. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. Update: COBE-SST2 based land-ocean dataset. *Unpub* (2017).
15. Cowtan, K. *et al.* Robust comparison of climate models with observations using blended land air and ocean sea surface temperatures. *Geophysical Research Letters* **42**, 6526–6534 (2015).
16. Schurer, A. P. *et al.* Interpretations of the Paris climate target. *Nature Geoscience* **11**, 220–221 (2018).
17. Kosaka, Y. & Xie, S.-P. The tropical Pacific as a key pacemaker of the variable rates of global warming. *Nat. Geosci.* **9**, 4–6 (2016).
18. Millar, R. J. *et al.* Emission budgets and pathways consistent with limiting warming to 1.5 °C. *Nature Geoscience* **10**, 741–747 (2017).
19. Tokarska, K. B. & Gillett, N. P. Cumulative carbon emissions budgets consistent with 1.5 °C global warming. *Nature Climate Change* **8**, 296–299 (2018).
20. Goodwin, P. *et al.* Pathways to 1.5 °C and 2 °C warming based on observational and geological constraints. *Nature Geoscience* **11**, 102 (2018).
21. Rogelj, J. *et al.* Mitigation pathways compatible with 1.5°C in the context of sustainable development. In: Global warming of 1.5°C. An IPCC special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [V. Masson-Delmotte, P. Zhai, H. O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J. B. R. Matthews, Y. Chen, X. Zhou, M. I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, T. Waterfield (eds.)]. In Press. (2018).
22. UNFCC. UNFCC, 2015. FCCC/CP/2015/L.9/Rev.1: Adoption of the Paris Agreement (pp. 1–32). UNFCCC, Paris, France. (2015).
23. IPCC, (2014). Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Core Writing Team, R. K. Pachauri and L. A. Meyer. Geneva, Switzerland, IPCC: 1-151. (2014).
24. Rogelj, J., Forster, P. M., Kriegler, E., Smith, C. J. & Séférian, R. Estimating and tracking the remaining carbon budget for stringent climate targets. *Nature* **571**, 335–342 (2019).
25. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An Overview of CMIP5 and the Experiment Design. *Bull. Amer. Meteor. Soc.* **93**, 485–498 (2011).
26. Richardson, M., Cowtan, K., Hawkins, E. & Stolpe, M. B. Reconciled climate response estimates from climate models and the energy budget of Earth. *Nature Climate Change* **6**, 931–935 (2016).
27. Richardson, M., Cowtan, K. & Millar, R. J. Global temperature definition affects achievement of long-term climate goals. *Environ. Res. Lett.* **13**, 054004 (2018).
28. Hawkins, E. *et al.* Estimating Changes in Global Temperature since the Preindustrial Period. *Bull. Amer. Meteor. Soc.* **98**, 1841–1856 (2017).
29. IPCC, 2013. Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis.Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. (2013).
30. Pfleiderer, P., Schleussner, C.-F., Mengel, M. & Rogelj, J. Global mean temperature indicators linked to warming levels avoiding climate risks. *Environ. Res. Lett.* **13**, 064015 (2018).

31. Lenssen, N. J. L. *et al.* Improvements in the GISTEMP Uncertainty Model. *Journal of Geophysical Research: Atmospheres* **124**, 6307–6326 (2019).
32. Cowtan, K. & Way, R. G. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Q J Roy Meteorol Soc* **140**, 1935–1944 (2014).
33. Rohde, R. *et al.* Rohde R, Muller RA, Jacobsen R, Muller E, Perlmutter S, et al. (2013) A New Estimate of the Average Earth Surface Land Temperature Spanning 1753 to 2011. *Geoinfor Geostat: An Overview 1:1.* (2013) doi:10.4172/2327-4581.1000101.
34. Morice, C. P., Kennedy, J. J., Rayner, N. A. & Jones, P. D. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres* **117**, (2012).
35. Allen, M. R. & et al. Allen, M.R., O.P. Dube, W. Solecki, F. Aragón-Durand, W. Cramer, S. Humphreys, M. Kainuma, J. Kala, N. Mahowald, Y. Mulugetta, R. Perez, M. Wairiu, and K. Zickfeld, 2018: Framing and Context. In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)]. In Press.mbridge, United Kingdom and New York, NY, USA. (2018).
36. Medhaug, I., Stolpe, M. B., Fischer, E. M. & Knutti, R. Reconciling controversies about the ‘global warming hiatus’. *Nature* **545**, 41–47 (2017).
37. Hausfather, Z. *et al.* Assessing recent warming using instrumentally homogeneous sea surface temperature records. *Science Advances* **3**, e1601207 (2017).
38. Kennedy, J. J., Rayner, N. A., Atkinson, C. P. & Killick, R. E. An Ensemble Data Set of Sea Surface Temperature Change From 1850: The Met Office Hadley Centre HadSST.4.0.0.0 Data Set. *Journal of Geophysical Research: Atmospheres* **0**,.
39. Kosaka, Y. & Xie, S.-P. Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature* **501**, 403–407 (2013).
40. England, M. H. *et al.* Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nature Climate Change* **4**, 222–227 (2014).
41. Huber, M. & Knutti, R. Natural variability, radiative forcing and climate response in the recent hiatus reconciled. *Nature Geoscience* **7**, 651–656 (2014).
42. Schmidt, A. *et al.* Volcanic Radiative Forcing From 1979 to 2015. *Journal of Geophysical Research: Atmospheres* **123**, 12,491-12,508 (2018).
43. Schmidt, G. A., Shindell, D. T. & Tsigaridis, K. Reconciling warming trends. *Nat Geosci* **7**, 158–160 (2014).
44. Myhre, G. *et al.* Multi-model simulations of aerosol and ozone radiative forcing due to anthropogenic emission changes during the period 1990–2015. *Atmospheric Chemistry and Physics* **17**, 2709–2720 (2017).
45. Outten, S., Thorne, P., Bethke, I. & Seland, Ø. Investigating the recent apparent hiatus in surface temperature increases: 1. Construction of two 30-member Earth System Model ensembles. *Journal of Geophysical Research: Atmospheres* **120**, 8575–8596 (2015).
46. Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* **9**, 1937–1958 (2016).
47. Volodin, E. & Gritsun, A. Simulation of observed climate changes in 1850–2014 with climate model INM-CM5. *Earth System Dynamics* **9**, 1235–1242 (2018).
48. The CMIP6 landscape. *Nat. Clim. Chang.* **9**, 727–727 (2019).
49. Jiménez-de-la-Cuesta, D. & Mauritsen, T. Emergent constraints on Earth’s transient and equilibrium response to doubled CO_2 from post-1970s global warming. *Nat. Geosci.* (2019) doi:10.1038/s41561-019-0463-y.
50. Vuuren, D. P. van *et al.* The representative concentration pathways: an overview. *Climatic Change* **109**, 5 (2011).
51. Bindoff, N. L. *et al.* Detection and attribution of climate change: From global to regional. Climate Change 2013: The Physical Science Basis, T. F. Stocker et al., Eds., Cambridge Univ. Press. 867–952 (2013).
52. UNFCCC. Report on the structured expert dialogue on the 2013–2015 review. FCCC/SB/2015/INF.1. (2015).
53. Rogelj, J., Schleussner, C.-F. & Hare, W. Getting It Right Matters: Temperature Goal Interpretations in Geoscience Research. *Geophysical Research Letters* **44**, 10,662-10,665 (2017).

54. Schleussner, C.-F. *et al.* Science and policy characteristics of the Paris Agreement temperature goal. *Nature Climate Change* **6**, 827–835 (2016).
55. Matthews, H. D. *et al.* National contributions to observed global warming. *Environ. Res. Lett.* **9**, 014010 (2014).
56. Mengis, N., Partanen, A.-I., Jalbert, J. & Matthews, H. D. 1.5 °C carbon budget dependent on carbon cycle uncertainty and future non-CO_2 forcing. *Scientific Reports* **8**, 5831 (2018).
57. Lowe, J. A. & Bernie, D. The impact of Earth system feedbacks on carbon budgets and climate response. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **376**, 20170263 (2018).
58. Comyn-Platt, E. *et al.* Carbon budgets for 1.5 and 2 °C targets lowered by natural wetland and permafrost feedbacks. *Nature Geoscience* **11**, 568 (2018).

Methods

We make use of CMIP5 and CMIP6 models REF, as detailed in each sub-section regarding Figure 2 and Figure 3. The sets of models used in Figure 2 and Figure 3 are different, as described below.

Contributions of different effects to the observed and modelled warming

Figure 2 summarizes effects why observed and modelled global mean temperature might differ between the two reference periods 1986–2005 and 2006–2015. The CMIP5 ensemble is that of Ref.[15] and consists of 38 models with 86 realizations (bcc-csm1-1-m and CMCC-CESM show unphysical features in the difference between GBST and GSAT in the late 21st century and were excluded in Ref.[15], but are included here as we are interested in the period up to 2015). We first average the ensemble members of each model to then obtain the multi-model mean. Uncertainties in the observed GBST arising from SSTs is assessed by comparing the warming of the HadCRUT-CW dataset (Ref.[14]) when it is constructed using three different SST datasets: HadSST3 (Refs.[59]), COBE-SST2 (Ref.[60] and Ref.[14]), and HadSST4. With both HadSST3 and HadSST4 the GBST increase between 1986-2005 and 2006-2015 is 0.26 °C whereas it is 0.28 °C with COBE-SST2. The choice of the SST dataset has therefore only a relatively small influence on the GBST increase. GISTEMP as an alternative GBST dataset shows a warming of 0.26 °C between the two reference periods. Figure 2b bar 2 displays the 5-95% range across the 100 member HadCRUT4.6 ensemble.

We use variability analogues[41] to quantify how Pacific variability altered the warming between the two reference periods[61]. Therefore, we search for periods from 33 CMIP5 and 18 CMIP6 control simulations (29’950 model years in total) where the modelled variability agrees with the observed variability (based on the root-mean-square error between the time series over a period of 40 months, and we keep the 20 best matching analogues within each period). We standardize both the observed and modelled variability time series. The GSAT anomaly in the analogues is a measure of the contribution of the observed Pacific variability to the observed GBST evolution. To describe internal variability we take area-weighted SSTs in the Nino3.4 region (5°S–5°N, 170°W–120°W) and from a larger region in the central to eastern tropical Pacific (15°N–15°S, 180°W–90°W) using two spatially interpolated SST data sets, ERSSTv5 (Ref.[62]) and COBE-SST2. SSTs in these regions also include a forced signal that we remove prior to selecting the analogues. We estimate the forced signal by the method of Ref.[63], i.e. a linear trend over observed tropical ocean SST from 1962 to 2011, and by using the ensemble means of the CMIP5 and CMIP6 models for the respective regions. Shown in Figure 2 is the range across the resulting 12 combinations of region, SST dataset and forced signal correction. Additionally, we select analogues based on observed zonal wind stress in the western tropical Pacific over two regions (180°W–150°W, 6°S–6°N, and 150°E–150°W, 10°S–10°N) from 49 control simulations (31 CMIP5 and 18 CMIP6 models with 29'084 years). These regions are based on Ref.[40] and Ref.[64]. We take observed wind stress from two reanalyses, ERA-Interim (Ref.[65]) and MERRA2 (Ref.[66]) and in Figure 2b we display the range across the resulting four wind stress estimates.

Refs.[67,68] and Refs.[69,70] quantify the contribution of tropical Pacific variability to GBST using multiple linear regression. They describe tropical Pacific variability by the Nino3.4 and Multivariate ENSO indices[71,72] . We use an updated and modified version of Ref.[69] where a second ENSO lag term was added. Refs.[17,73] and the simulations with IPSL-CM6A-LR that follow the "Decadal Climate Prediction Project" protocol by Ref.[74], quantify the Pacific contribution to GSAT as the difference between two climate model experiments. A freely evolving initial condition ensemble forced with historical radiative forcings and a second experiment driven by the same radiative forcings, but where modelled central to eastern tropical Pacific SSTs are nudged towards observed anomalies. These so-called pacemaker experiments end in 2013 and 2014, respectively. We use the variability analogues to approximately extend the estimates to 2015. Alternatively, we assume that the complete year-to-year HadCRUT4.6 GBST variability during the missing years was caused by Pacific variability. Figure 2b shows the spread arising from these two assumptions. The pacemaker experiments indicate a larger Pacific induced global temperature decrease between the two reference periods than studies using multiple regression. This could be related to a time-scale dependence of the imprint of tropical Pacific variability on GSAT, which in climate model simulations is larger on a decadal than on an interannual time scale[17,75]. Regression models constructed on interannual variability might underestimate the Pacific influence on a decadal time scale[75]. Additionally, if and how the forced signal is removed from tropical Pacific SSTs plays a role. If it is not fully removed, the cooling from internal variability is underestimated and vice versa. The spread in Pacific contribution to the GSAT change between the two reference periods is also substantial across the pacemaker studies (Fig. 2b, effects 8 to 10) and this is probably related to how strongly the tropical Pacific variability projects onto higher latitudes on a decadal time-scale[75].

We use the forcing corrections of Refs.[41–43,45,76]. For Ref.[41] we combine the forcing corrections of updated solar variability (with PMOD) and of stratospheric aerosols (not including their correction for background stratospheric aerosols from 1960 to 1990). Ref.[43] and Ref.[45] additionally estimate the effects of updated well-mixed greenhouse gas concentrations, which is very small in both studies, and human-made tropospheric aerosols. While Ref.[43] find underestimated aerosol cooling during the first decade of the 21st century, Ref.[45] argue for overestimated aerosol cooling, presumably related to primary organic matter aerosols. For the Ref.[45] forcing correction, we only show the GSAT influence of updated solar and volcanic forcing. Refs.[42,43] downgrade the radiative forcing of the Mount Pinatubo eruption, making the 1986-2005 period warmer and thereby also decreasing the GSAT increase between the two reference periods. On the contrary, Ref.[45] increase volcanic forcing during the early reference more than from 2006 onwards, and thus increase the simulated warming between the two reference periods. This and the reduced cooling from tropospheric aerosols lead to slightly increased warming between the two reference periods compared to the control experiment with CMIP5 forcings in Ref.[45]. Different to the other forcing corrections, some internal variability is left in the estimate of Ref.[45] as it is the difference between two 30-member climate model ensembles. Figure 2b effect 15 shows the difference between the two ensemble means (with 90% confidence interval using data until 2012) and the central estimate is from assuming that the anomaly comes back to zero by 2015. Further, we display the volcanic aerosol GSAT corrections of Ref.[76] and Ref.[42] who account for volcanic aerosols in the lowermost stratosphere below 15 km which is not included in the other stratospheric aerosol corrections (for Ref.[76] we use the AERONET mean GSAT estimate which we digitized from their Figure 3b). Except for Ref.[42] that fully covers the period 2006-2015, the other studies include data until 2012/2013 and for the missing years we assume that the GSAT anomaly of stratospheric aerosols remains constant and that the adjustment from updated solar irradiance comes back to zero anomaly by 2015. Not all forcing corrections fully cover the early 1986-2005 reference period, and for missing years we assume a zero GSAT anomaly.

The CMIP6 models are forced with updated radiative forcing until 2014 (we extrapolate until 2015 by repeating the warming of the previous year), but as also model physics changed, and the set of models is not the same, the difference in GSAT increase compared to CMIP5 cannot solely be attributed to changes in radiative forcings. The CMIP6 ensemble of historical simulations consists of (number of members in parentheses) BCC-CSM2-MR (3), BCC-ESM1 (3), CAMS-CSM1-0 (2), CanESM5 (50), CESM2 (11), CESM2-WACCM (3), CNRM-CM6-1 (10), CNRM-ESM2-1 (5), E3SM-1-0 (5), EC-Earth3 (6), EC-Earth3-Veg (4), FGOALS-g3 (3), GFDL-CM4 (1), GFDL-ESM4 (1), GISS-E2-1-G (20), GISS-E2-1-G-CC (1), GISS-E2-1-H (10), HadGEM3-GC31-LL (4), IPSL-CM6A-LR (32), MIROC6 (10), MIROC-ES2L (3), MRI-ESM2-0 (5), NESM3 (5), NorCPM1 (30), NorESM2-LM (1), SAM0-UNICON (1), and UKESM1-0-LL (6). We compare the CMIP6 ensemble mean with the CMIP5 mean for GSAT (with RCP8.5 from 2006 onwards) and estimate the uncertainty of the difference in the ensemble means using Welch's t-test (Figure 2b shows the 90% confidence interval). Overall, the warming simulated by the CMIP6 ensemble mean between the two reference periods is slightly higher than that of the CMIP5 ensemble (Figure 2b).

For the net effect, we combine the Pacific variability estimated by analogues from the central to eastern tropical Pacific with the CMIP5 mean removed and averaged across ERSSTv5 and COBE-SST2 (for Figure 2a we show the range across all combinations of SST-based analogues), and the updated radiative forcing of Ref.[41]. We, however, stress that this only one possible combination and that the individual components are rather uncertain. There might be further effects not accounted for by our analysis, such as Atlantic multidecadal variability but which effect on GSAT is probably small during the period examined[77]. Also, forcing and variability corrections are estimated for GSAT and not GBST, which might cause a small bias.

Differences in the ocean and sea ice coverage, and related differences between GBST and GSAT

Figure 3 displays global free ocean fraction and the influence of changes in sea ice coverage on the difference between GBST and GSAT. Free ocean coverage is the area fraction of ocean cells in each model subtracted by sea ice coverage. While the number of ocean cells is constant sea ice coverage declines with global warming. In the computation of GBST surface air temperatures are taken over land and sea ice and surface ocean temperatures are used for ocean cells. In grid-cells partially covered by sea ice surface air and ocean temperatures are blended respective to the sea ice fraction. We follow Ref.[15] for the computation of GBST and GBST with fixed sea ice.

Fixed sea ice coverage is based on monthly sea ice coverage between 1961-2014: cells that have not been covered in that period (and in the respective month) are considered as sea ice free, the remaining cells are considered as fully covered by sea ice. Figure 3 includes 28 CMIP5 models: ACCESS1-0, ACCESS1-3, CCSM4, CESM1-BGC, CMCC-CMS, CMCC-CM, CSIRO-Mk3-6-0, CanESM2, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-H-CC, GISS-E2-H, GISS-E2-R-CC, GISS-E2-R, HadGEM2-CC, HadGEM2-ES, IPSL-CM5A-LR, IPSL-CM5A-MR, IPSL-CM5B-LR, MIROC-ESM-CHEM, MIROC-ESM, MIROC5, MPI-ESM-LR, MPI-ESM-MR, MRI-CGCM3, NorESM1-ME, and NorESM1-M.

Transitions between GBST and the Paris-consistent method

The magnitude of the first arrow in Box 1 between Stream 1 and Stream 2 (i.e. the difference between the GBST and Paris-consistent temperature method for 2006-2015) is based on the values from the IPCC SR1.5 Table 1.1 (Ref.[35]). It is calculated as the difference between the CMIP5 GSAT for the period 1850–1900 to 2006–2015 and the CMIP5 GSAT for the period 1850–1900 to 1986–2005, minus the difference between HadCRUT4.6 for the period 1850–1900 to 2006–2015 and HadCRUT4.6 for the period 1850–1900 to 1986–2005. Using values from Table 1.1 (Ref.[73]) results in: (0.99-0.62)-(0.84-0.60) = 0.13 °C, or more precisely, taking the values in brackets directly from column 4 (i.e., directly the GBST change from 1986-2005 to 2006-2015) of Table 1.1 results in: 0.38-0.22 = 0.16 °C. (Note the difference between these two estimates comes from rounding).

Data availability

The Cowtan and Way GBST datasets with different SST reconstructions are available at:
HadCRUT4.6 data is available at:
GISTEMPv4 is available at: https://data.giss.nasa.gov/gistemp/.
COBE-SST2 and ERSSTv5 data is provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at https://www.esrl.noaa.gov/psd/data/gridded/.
ERA-Interim is available at: https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim.
MERRA2 was downloaded from: https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/.
CMIP5 and CMIP6 model output is available at: http://pcmdi9.llnl.gov/.
CESM1 pacemaker experiments are available at: https://www.earthsystemgrid.org/.

Methods References

59. Kennedy, J. J., Rayner, N. A., Smith, R. O., Parker, D. E. & Saunby, M. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research: Atmospheres* **116**, (2011).
60. Hirahara, S., Ishii, M. & Fukuda, Y. Centennial-Scale Sea Surface Temperature Analysis and Its Uncertainty. *J. Climate* **27**, 57–75 (2013).
61. Stolpe, M. B., Cowtan, K., Medhaug, I. & Knutti, R. Pacific Variability Reconciles Observed and Modelled Global Mean Temperature Increase since 1950. *in preparation* (2019).
62. Huang, B. *et al.* Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons. *J. Climate* **30**, 8179–8205 (2017).
63. Turkington, T., Timbal, B. & Rahmat, R. The impact of global warming on sea surface temperature based El Niño–Southern Oscillation monitoring indices. *International Journal of Climatology* **39**, 1092–1103 (2019).
64. Saenko, O. A., Fyfe, J. C., Swart, N. C., Lee, W. G. & England, M. H. Influence of tropical wind on global temperature from months to decades. *Clim Dyn* **47**, 2193–2203 (2016).
65. Dee, D. P. *et al.* The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q J Roy Meteorol Soc* **137**, 553–597 (2011).
66. Gelaro, R. *et al.* The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Climate* **30**, 5419–5454 (2017).
67. Folland, C. K., Boucher, O., Colman, A. & Parker, D. E. Causes of irregularities in trends of global mean surface temperature since the late 19th century. *Science Advances* **4**, (2018).
68. Lean, J. L. Observation-based detection and attribution of 21st century climate change. *Wiley Interdisciplinary Reviews: Climate Change* **9**, e511 (2018).
69. Foster, G. & Rahmstorf, S. Global temperature evolution 1979–2010. *Environ. Res. Lett.* **6**, 044022 (2011).
70. Rypdal, K. The Life and Death of the Recent Global Surface Warming Hiatus Parsimoniously Explained. *Climate* **6**, 64 (2018).
71. Trenberth, K. E. The Definition of El Niño. *Bull. Amer. Meteor. Soc.* **78**, 2771–2778 (1997).
72. Wolter, K. & Timlin, M. S. El Nino/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). *Int J Climatol* **31**, 1074–1087 (2011).
73. Deser, C., Guo, R. & Lehner, F. The relative contributions of tropical Pacific sea surface temperatures and atmospheric internal variability to the recent global warming hiatus. *Geophysical Research Letters* **44**, 7945–7954 (2017).
74. Boer, G. J. *et al.* The Decadal Climate Prediction Project (DCPP) contribution to CMIP6. *Geoscientific Model Development* **9**, 3751–3777 (2016).
75. Wang, C.-Y., Xie, S.-P., Kosaka, Y., Liu, Q. & Zheng, X.-T. Global Influence of Tropical Pacific Variability with Implications for Global Warming Slowdown. *Journal of Climate* **30**, 2679–2695 (2017).
76. Ridley, D. A. *et al.* Total volcanic stratospheric aerosol optical depths and implications for global climate change. *Geophysical Research Letters* **41**, 7763–7769 (2014).
77. Haustein, K. *et al.* A Limited Role for Unforced Internal Variability in Twentieth-Century Warming. *J. Climate* **32**, 4893–4917 (2019).