Supplementary Materials

Geophysical Research Letters
Supporting Information for

Evaluating the performance of past climate model projections

Zeke Hausfather[1], Henri Drake[2], Tristan Abbott[3], Gavin Schmidt[4]

[1] Energy and Resources Group, University of California, Berkeley. 310 Barrows Hall, Berkeley, CA 94720, USA.
[2] MIT/WHOI Joint Program.
[3] MIT EAPS.
[4] NASA Goddard Institute for Space Studies, 2880 Broadway, New York, USA

Contents of this file

Data and code

A spreadsheet with tabs containing data from all of the models evaluated in this study is available here:
https://github.com/hausfath/OldModels/blob/master/Model%20data%20spreadsheet.xlsx

A public GitHub repository with code used to analyze the data, generate figures, and csv files containing the data shown in the figures is available here:
https://github.com/hausfath/OldModels

The 1000-member ensemble of observationally-informed radiative forcing estimates can be found here: https://github.com/hausfath/OldModels/tree/master/forcing_data

Observational temperature datasets can be found at the following links:
NASA GISTEMP – https://data.giss.nasa.gov/gistemp/
NOAA GlobalTemp – https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php
Hadley HadCRUT4 – http://www.metoffice.gov.uk/hadobs/hadcrut4/
Berkeley Earth – http://berkeleyearth.org/data/
Cowtan and Way – http://www-users.york.ac.uk/~kdc3/papers/coverage2013/series.html

Text S1: Detailed methods

S1.0: Additional methods notes

The choice to start the future projection period at the date of publication was made as a conservative choice to avoid any possibility of observed temperatures informing model development or parameterization. While in some cases the specific date on which the model was run prior to the paper publication is known, in most cases (particularly for earlier studies) this is not readily available. In other cases (e.g. for the IPCC AR4) models were run with projected future forcings that start well before the model was developed, which does not completely preclude the knowledge of observed temperatures in the intervening period from informing the development and tuning of parameterizations but is unlikely given the multi-year timescale of model development.

While a more complex two-layer model with ocean heat uptake would be able to better capture the relationship between forcing and temperature response than our simple implied TCR metric (Rohrschneider et al. 2019), we have purposefully chosen to avoid a situation where we are using a more complex model with its own somewhat uncertain parameterizations to evaluate the performance of historical climate models. A more complex model may also not provide an effective comparison with early climate model projections, many of which (prior to ST81) did not include ocean heat uptake dynamics. is not an optimal metric in all cases, but will provide a more accurate evaluation of model projection performance than the conventional approach of analyzing changes in temperature over time without accounting for differences in the time evolution of modeled and observationally-estimated radiative forcings.

While our analysis uses instantaneous radiative forcings either calculated from modeled CO2 concentrations or based on published data from past climate models, it does not account for differing forcing efficacies (Hansen et al 2005; Marvel et al 2016). When taken into account – based on efficacies from the GISS model – efficacy-adjusted forcings are around 3% higher at present, and representing an additional source of uncertainty in our analysis.

S1.1: Temperature vs time

To evaluate the performance of model projections against observed temperatures, the linear trend in both observations and model projections was calculated over the future projection period. An ordinary least squares approach was used to calculate the trend coefficient of all five observational temperature records over the future projection period. A first-order autoregressive model (AR1) was further used to estimate trend uncertainties, similar to the approach used in Hausfather et al. (2017).

Specifically, trend coefficients of temperature with respect to time,β, were estimated with the ordinary least squares model:

$$y_i = \beta x_i + \varepsilon_i$$

(1)

The uncertainty introduced by the choice of observational estimate was calculated from the variance of the five coefficients ($\beta_1 \ldots \beta_5$):

$$var(obs) = \sum (\beta_i - \mu)^2 / N$$

(2)

whereμ is the average of the five coefficients and $N = 5$.

An AR1 model was used to estimate the regression confidence intervals:

$$X_t = c + \rho X_t + \varepsilon_t$$

(3)

where c is a constant, ε_t is white noise, and$\rho$is the model parameter. The variance for the regression of a given observational temperature record i can be calculated by:

$$var(X_t)_i = {\sigma_e}^2 / (1 - \rho^2).$$

(4)

The ordinary least squares model provides a more physically meaningful coefficient than the AR(1) model, while the AR(1) model provides a better estimate of the variance (accounting for autocorrelation). These were calculated separately for each of the five observational temperature datasets. In cases where the confidence intervals of the regression coefficient from the AR(1) approach were smaller than those from an OLS model, the OLS coefficients were used to provide a more conservative estimate.

A mean and combined uncertainty were estimated by averaging the five coefficients and by adding the (two-sigma) coefficient uncertainty and the mean AR(1) (two-sigma) trend uncertainty in quadrature, assuming that the two are independent:

$$\bar{\beta} \pm \sqrt{4 \cdot var(obs) + 4 \cdot \overline{var(X_t)}}.$$

(5)

As the standard deviation is the square root of the variance, $\left(2 \cdot \sqrt{var}\right)^2 = 4 \cdot var$. For models, where only a single realization of projected temperatures is available, the same approach was used except with a single β and $var(X_t)$.

S1.2: Implied TCR

Implied TCR is defined as the ratio between the change in temperature and the change in external forcing over the model projection period, for both models and observations. It is referred to as ‘implied’ as it differs from the traditional definition of TCR, which is typically based on idealized experiments where CO_2 is increased by 1% per year (IPCC 2001).

When explicit external forcing values were not available, they were estimated from model greenhouse gas concentrations using the simplified radiative forcing functions from the IPCC AR5 (IPCC 2013). Forcing from a change in atmospheric concentration of CO_2 is given by:

$$\Delta F_{co_2} = 5.35 \cdot ln \frac{(P_{CO_2} + \alpha_{CO_2})}{P_{CO_2}}. \tag{7}$$

Here P_{CO_2} represents the initial concentration of CO_2 in the atmosphere when the model projection period began, while α_{CO_2} represents the additional parts per million CO_2 added through the end of 2017 (or when the model run ended if prior to 2017).

The direct radiative forcing of a given increase of CH_4 and/or N_2O in the atmosphere can be approximated by:

$$\Delta F_{CH_4} = 0.036(\sqrt{P_{CH_4} + \beta_{CH_4}} - \sqrt{P_{CH_4}}) - f(P_{CH_4} + \beta_{CH_4},\ P_{N_2O}) + f(P_{CH_4}, P_{N_2O})$$
$$\Delta F_{N_2O} = 0.12(\sqrt{P_{N_2O} + \beta_{N_2O}} - \sqrt{P_{N_2O}}) - f(P_{CH_4}, P_{N_2O} + \beta_{N_2O}) + f(P_{CH_4}, P_{N_2O})$$

where:

$$f(M,N) = 0.47\ ln(1 + 2.01 \cdot 10^{-5}(MN)^{0.75} + 5.31 \cdot 10^{-15} M(MN)^{1.52}). \tag{8}$$

In this equation P_{CH_4} is the initial concentration of atmospheric CH_4, while β_{CH_4} is the addition being evaluated. P_{N_2O} is the initial concentration of N_2O, and β_{N_2O} is the addition being evaluated. The radiative forcing of both CH_4 and N_2O is a function of the combination of both, reflecting their interacting atmospheric chemistry.

We use a 1000-member ensemble of observationally-informed radiative forcing estimates from Forster and Dessler (2018) to account for uncertainties in forcing associated with aerosol, land albedo, and other factors that are relatively poorly observationally constrained. The ensemble members are combined with each of the five observational temperature records to regress the change in temperature against the change in radiative forcing, following the approach used in Eqs. 1-5 but substituting radiative forcing for time and using an OLS rather than AR(1) approach for trend uncertainties given the absence of a time variable needed for an autoregressive model.

Specifically, a set of 5000 $\Delta T / \Delta F_{anthro}$estimates are calculated for each model projection period across the five observational temperature estimate and 1000 radiative forcing ensemble members. The mean of these estimates is calculated, and uncertainties are estimated based on both the variation across these 5000 estimates and on the mean confidence intervals of the regression coefficients. These uncertainties and added in quadrature as the two are independent:

$$\overline{TCR_{implied}} \pm \sqrt{4 \cdot var(TCR_{implied}) + 4 \cdot \overline{var(\beta)}}$$

(9)

where $var(\beta)$ is the variance of the OLS regression in Eq. 1, but regressing temperature against anthropogenic forcing rather than time.

Models, in turn, have a single realization of ΔT and ΔF_{anthro} over their projection period, and the uncertainties are only estimated from $var(\beta)$.

S1.3: Calculating consistency and skill scores

We refer to model projections as consistent or inconsistent with observations based on a comparison of the differences between the two. Specifically, when comparing models on a temperature vs time basis, we difference the model and observation global mean surface temperature time series for each of the five observational time series. These difference series will remove any common variability between model projections and observations (Hausfather et al. 2017). Trends and trend confidence intervals for these difference series are then calculated following the approach in Eq. 5. Model projections and observations are considered consistent if the trend 95% confidence interval of the difference series is inclusive of zero, indicating that zero difference in trends cannot be ruled out.

When comparing model projections and observations on an implied TCR basis (e.g. change in temperature compared to change in forcing), using linear regressions on difference series is more problematic given the lack of shared time axis between the two. Instead, we assume that the trend uncertainties for each are independent, and add the uncertainties in quadrature to the difference in trends. Specifically, we calculate:

$$TCR_{diff} \pm \sqrt{4 \cdot var(TCR_{models}) + 4 \cdot var(TCR_{obs})}$$

where:

$$TCR_{diff} = TCR_{implied,model} - \overline{TCR_{implied,obs}}$$
$$var(TCR_{model}) = \sqrt{var(TCR_{implied,model})^2 + var(\beta_{i,model})^2}$$
$$var(TCR_{obs}) = \sqrt{var(TCR_{implied,obs})^2 + var(\beta_{i,obs})^2}.$$

(10)

Here we similarly consider model projections and observations to be consistent if the 95% confidence interval of the difference between the two is inclusive of zero. This approach produces results quite similar to those from the difference series approach used in the temperature vs time case, suggesting that the phase of internal variability in model projections and observations are largely independent.

Skill scores are calculated following the approach of Hargreaves (2010). The root-mean-squared errors of the projectioned trend, E_f, is compared to a reference technique E_{refr}, where E_{refr} is simply the assumption of temperature persistence (e.g. zero trend over time). As

Hargreaves points out, the assumption of persistence generally outperforms the extrapolation of recent trends over any given interval in the historical global mean surface temperature record, at least prior to the last few decades. This serves as a reasonable counterfactual, particularly for early 1970s and 1980s models where the modern warming trend was less apparent to researchers at the time (Broecker 1975).

Skill scores, SS, are defined as:

$$SS = 1 - \sqrt{\frac{E_f}{E_{refr}}}$$

where:

$$E_f = \left(\beta_{i,obs} - \beta_{i,model}\right)^2$$
$$E_{refr} = \left(\beta_{i,obs} - 0\right)^2 .$$

(11)

Skill score uncertainties are estimated based on calculating skill scores separately for each model projection using the five different observational temperature records (for the temperature vs time metric) and the 5000 permutations of observational temperature record and observational forcing ensemble (for the implied TCR metric). The median skill score is calculated across all available runs for each metric. This is shown rather than the mean as the absolute value nature of the skill score means that a few ensemble members with very low skill can drag the mean skill score disproportionately down.

The uncertainties shown span the 5th to 95th percentile of resulting skill scores, accounting for both uncertainties from the choice of observational record and forcing series and the trend uncertainty due to temporal variability in the underlying time series. These are calculated via a Monte Carlo approach that takes the trend coefficient uncertainties into account. For the temperature vs time metric, 100 permutations of each of the five observational temperature records are estimated, where each randomly samples a value from the Gaussian distribution of the resulting regression trend coefficients. For the implied TCR metric, 100 values were randomly sampled from the Gaussian distribution of the resulting regression trend coefficients for each of the 5000 permutations of temperature record and observationally-based forcing series.

S1.4: Temperature uncertainties at a given forcing

The combination of 5 observed temperature time series and 1000 observationally-informed forcing time series gives an ensemble of 5000 estimates of how temperature varies as a function of radiative forcing. We can not immediately estimate uncertainty in temperature as a function of forcing because the forcing data points are not co-located. Thus, we define a regular grid of forcings with fine spacing of 0.02 W/m^2 and linearly interpolate the 5000 temperature values from the annual forcings values to the fine grid. We then calculate the sample standard deviation across the 5000 member ensemble and estimate a 95% confidence interval at each

forcing value. These confidence intervals for Hansen et al. 1988 and IPCC FAR are shown by the dashed blue lines in the lower panels of Figure 3 and Figure S6, respectively.

Text S2: Climate model projection assessment

This section provides detailed descriptions of how each historical climate model projection was digitized and analyzed, including what data points were available, if and what interpolation of data was applied, and what scenarios were used. When model projections were not available in a digital form, they were digitized from published figures using the free OS X application plotDigitizer: http://plotdigitizer.sourceforge.net/

Manabe 1970

Citing results from their previously published Manabe and Wetherald 1967 model, Manabe calculates the equilibrium surface air temperature in a one-dimensional radiative convective equilibrium model for a given distribution of relative humidity. Citing an increase in surface air temperature of 2.3°C as CO_2 concentrations are doubled from 300 ppm to 600 ppm, he uses an independent prediction of external forcing to predict transient warming in 2000, relative to 1900: "suppose the concentration of CO_2 increases by about 25% from AD1900 to AD2000 as the U.N. Department of Social and Economic Affairs predicts, the resulting increase of surface temperature would be about 0.8°C". It is unclear whether he carried out additional runs of the Manabe and Wetherald model to arrive at this number or simply scaled their previously calculated ECS of 2.36°C (table 5 of Manabe and Wetherald 1967) using the logarithmic dependence of CO_2 radiative forcing on CO_2 concentrations (equation 7), which gives

$$2.36 \times log(1.25) \,/\, log(2.0) \;= 0.759 \;\simeq 0.8°C.$$

To express this prediction as a change in radiative forcing and GMST between 1970 and 2000, we assume a CO_2 concentration of 300 ppm in 1900 and 320 ppm in 1970. The 320 ppm value is consistent with other papers published at the time (Mitchell et al. 1970; Benson et al. 1970; Rascool and Schneider 1971; Sawyer 1972), though lower than our current estimate of 1970 global CO_2 concentrations (325 ppm). The referenced prediction of a 25% increase from 1900 to 2000 thus predicts 375 ppm of CO_2 in 2000. Using equation 7 to convert CO_2 concentrations into a radiative forcing, we determine a predicted increase of radiative forcing of $\Delta F = 0.85\ W/m^2$between 1970 and 2000. To determine the predicted increase in GMST between 1970 and 2000 from the increase in GMST between 1900 and 2000, we assume, as when calculating implied TCR, that a linear relationship between temperature and forcing holds. Then, the increase in GMST between 1970 and 2000 can be calculated by linearly interpolating between $T = 0ºC$at $F = 0\ W/m^2$and $T = 0.8ºC$at $F = 1.20\ W/m^2$to $T = 0.23ºC$at$F = 0.35º\ W/m^2$. The resulting changes in radiative forcing and GMST are $\Delta F_{2000-1970} = 0.85\ W/m^2$ $\Delta T_{2000-1970} = 0.57ºC$. We linearly interpolate the forcing and temperature to arrive at annual values.

Link: https://link.springer.com/chapter/10.1007%2F978-94-010-3290-2_4

Note: Manabe and Wetherald 1967 itself is not included here because it did not provide a prediction for when CO_2 would reach a given level, only for the amount of warming that would

result once that level was reached. It simply provided an equilibrium response to doubled CO_2 rather than a timeseries of transient response.

Mitchell 1970

Similar to Manabe 1970, Mitchell 1970 uses the estimate of ECS from Manabe and Wetherald 1967 and projections of CO_2 levels, implicitly assuming the system instantaneously reaches equilibrium, to determine future changes in GMST.

Mitchell states that the increase in CO_2 concentrations, "relative to a 19th century base level of 290 ppm, [...] is projected to accumulate to 11% by 1970, 15% by 1980, 20% by 1990, and 27% by 2000 A.D". We convert CO_2 concentrations into radiative forcings using equation 7, with a reference of 320 ppm in 1969. Temperatures are taken from Mitchell's statement that "temperature contribution of CO_2 changes anticipated in the future, neglecting all other mechanisms of climatic change, will consist of a further warming (above 1969 temperature levels) of about 0.1°C (0.2ºF) by 1980, 0.3ºC (0.5ºF) by 1990, and 0.5ºC (0.8ºF) by 2000 A.D". We linearly interpolate the forcing and temperature to arrive at annual values.

Link: https://link.springer.com/chapter/10.1007/978-94-010-3290-2_15

Benson 1970

Benson predicts that CO_2 concentrations will increase linearly at the contemporaneous rate of 0.7 ppm per year. Linearly extrapolating from a value of 320 ppm in 1970, he arrives at a concentration of 384 ppm in 2000. Using equation 7, we translate this into an increase in CO_2 radiative forcing of 0.98 W/m^2 from 1970 to 2000.

Using Manabe and Wetherald 1967's estimate of climate sensitivity, expressed as a warming of 0.3ºC per 10% increase in CO_2 concentrations, he finds that temperatures should increase by "about 0.6ºC" from 1970 to 2000. Presumably, he used some form of equation 7, which expresses the approximately logarithmic dependence of radiative forcing on CO_2, to get

$$\Delta T \ = \ ECS \ \times \frac{\Delta F2000 - 1970}{\Delta F2 \times} = 2.36 \ \times log(1.2) \ / \ log(2.0) \ = 0.62 \ \simeq 0.6°C.$$

We linearly interpolate the forcing and temperature to arrive at annual values.

Link https://doi.org/10.1073/pnas.67.2.898

Rasool and Schneider 1971

Rasool and Schneider (1971)'s method of projecting GMST change based on an ECS and radiative forcing is similar to the above studies but both their projected increases in CO_2 concentrations of 10% from 1971 to 2001 and their estimate $ECS = 0.8ºC$ are less than half

those of all other contemporaneous projections discussed above and below (see note below on why this disagrees so much with the Manabe and Wetherald 1967 estimate of ECS). They state: "if CO_2 is augmented by another 10 percent in the next 30 years, the increase in the global temperature may be as small as 0.1ºK". We can reproduce this calculation, following equation 7, if we assume the system is always at equilibrium and is described by a constant feedback parameter, such that

$$\Delta T = 0.8 \times log(1.1) / log(2.0) = 0.11ºC \simeq 0.1ºC.$$

We linearly interpolate the forcing and temperature to arrive at annual values.

Note: Schneider (1975) discusses the difference between the Rasool and Schneider (1971) and Manabe and Wetherald (1967)'s estimate of ECS at length, based on simulations by Manabe and Wetherald who generously replicated their simulations with Rasool and Schneider (1971)'s assumptions. The differences between their estimates can be summarized by the following: 1) Rasool and Schneider assume an isothermal stratosphere, which allows too much radiation to space in optically-thick bands as CO_2 is increased, and thus limits the amount of heating at the surface; 2) Rasool and Schneider do not include near-infrared solar absorption by water vapor and CO_2, resulting in less heating at the surface; 3) Manabe and Wetherald 1967's infrared radiation transfer scheme was less elaborate than that of Rasool and Schneider 1971, resulting in a 0.4ºC warm bias in their ECS relative to the radiation scheme used in Rasool and Schneider 1971.

Link: http://dx.doi.org/10.1126/science.173.3992.138

Sawyer 1972

Citing Manabe 1970, he assumes an ECS of 2.4ºC. He speculates that a 25% increase in CO_2 concentrations from 319 in 1969 to 399 ppm in 2000 would lead to a warming of 0.6ºC. We are unclear how Sawyer arrived at this value, since the typical scaling would provide a temperature change of:

$$2.4 \times log(1.25) / log(2.0) = 0.77 \simeq 0.8\,ºC.$$

Given the fact that

$$2.4 \times 1.25 / 2.0 = 0.6ºC,$$

it seems possible that Sawyer mistakenly approximated CO_2 forcing as a linear function of CO_2 concentrations, resulting in a spurious underestimate of the temperature change. It is not clear to us how else Sawyer could come up with a temperature change of 0.6ºC from the cited values.

Link: https://www.nature.com/articles/239023a0

Broecker 1975

Citing the calculation of $ECS = 2.4ºC$ from the general circulation model in Manabe and Wetherald 1975 as the most reliable estimate of ECS (which coincidentally differs only slightly from their previous column-model calculations of $ECS = 2.36ºC$ in Manabe and Wetherald 1967), Broecker follows the same approach as Manabe 1970 and projects GMST changes forward to 1980, 1990, 2000, and 2010 (see his Table 1 and Figure 1), using a variant of equation 7 that gives nearly identical results. We linearly interpolate the forcing and temperature between values reported in Table 1 to arrive at annual values between 1975 and 2010.

Link: https://science.sciencemag.org/content/189/4201/460

Note: we only consider the anthropogenically forced response, ignoring projected contributions from the assumed sinusoidal cycles of natural variability, which Broecker himself later admitted were flawed (Broecker 2017).

Nordhaus 1977

The temperature response for a given trajectory of CO_2 concentrations is given by equation 7, same as all of the above, using a value of $ECS = 2ºC$, the choice of which seems to be mostly informed by the Manabe and Wetherald 1967 model but reflects the large spread of estimates in the 1970s literature (see above).

Nordhaus 1977 differs from the above models in that he calculates CO_2 trajectories based on decoupled linear economic and carbon cycle models. While Nordhaus 1977 explores scenarios with constraints on the level of allowable CO_2 concentrations in the atmosphere, we only consider the temperature time-series for the uncontrolled case, which eventually reaches CO_2 concentrations four to five times pre-industrial levels. We note that by 2020, the uncontrolled scenario results in CO_2 concentrations that, by 2020, are only slightly higher than that of a scenario where CO_2 concentrations are constrained to never go beyond double the CO_2 concentration from the year 1974.

The temperature time-series is digitized from Figure 1 while the radiative forcing is calculated according to equation 7 by digitizing the time series of carbon dioxide content in the atmosphere in Figure 9 and converting to parts per million.

Note: There appears to be a typo in Nordhaus 1977 monograph in the first footnote of page 5, which cites Manabe and Wetherald 1969 when referencing the model for the temperature response, which does not appear in the bibliography and is elsewhere cited as Manabe and Wetherald 1967.

Links: http://cowles.yale.edu/sites/default/files/files/pub/d04/d0443.pdf (long version)
https://www.jstor.org/stable/pdf/1815926 (short version)

Schneider and Thompson 1981

In contrast to all of the above, which consider the case of instantaneous thermal equilibrium, Schneider and Thompson 1981 consider the transient evolution of surface temperatures in a two-box energy balance model. When diffusive heat uptake by the deep ocean (represented by the lower box) is included, the transient warming is reduced relative to the instantaneous equilibrium case (equivalent, the small thermal inertia or short radiative relaxation timescale case). Here, we only consider the case of a diffusive timescale of 550 years for the global deep ocean, which is the scenario in Schneider and Thompson 1981 that most corresponds to modern understanding of the diffusive and advective timescales for the deep ocean circulation and is the middle of the range of diffusive timescales considered in the paper. The ratio r_c of instantaneous CO_2 concentration over the 1925 value relative is assumed to increase quadratically according to:

$$1.443\ ln(r_c)\ =\ 7.03\ \times 10^{-5}\ t^2.$$

External forcing is estimated from the CO_2 concentrations using equation 7. The temperature time series is digitized from Figure 3 for a diffusive timescale $\tau_d = 550$ years.

Link: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC086iC04p03135

Hansen et al. 1981

We consider two forcing scenarios for Hansen et al. 1981: scenario 1 ("fast growth") and scenario 2a ("slow growth" without coal phaseout). In both cases, natural gas, oil, and coal consumption increases according to a prescribed growth rate (4% and 2% for scenarios 1 and 2, respectively). While the prescribed growth rates do technically taper in time, the tapering does not come into effect until 2020 so it does not affect the results shown here. When the relatively limited gas and oil reserves are depleted by ever-increasing energy consumption, they are in principle replaced by coal though in practice none of the reserves are depleted until after 2020 in either scenario. Energy consumption in Joules is converted to ppm of CO_2 according to the conversion factors in Table 2 of Hansen et al. 1981. Hansen et al. 1981 do not discuss the potential for future carbon sinks; following them, we thus unrealistically assume all emitted CO_2 remains in the atmosphere perpetually. Some of the excess forcing due to the permanence of anthropogenic CO_2 in the atmosphere is likely offset by observed increases in other greenhouse gases, which are not included in the Hansen et al. 1981 projections. CO_2 concentrations are converted to a radiative forcing using only the terms involving a change in CO_2 concentrations from equation 9 of Hansen et al. 1981, which agrees with our equation 7 to within 2% for historical changes in CO_2 concentrations.

Temperature time series corresponding to the forcing scenarios 1 and 2a are digitized from Figure 6 of Hansen et al. 1981.

Link: https://pubs.giss.nasa.gov/abs/ha04600x.html

Hansen et al. 1988

We consider the three forcing scenarios for H88: a rapid growth scenario A, a slow growth scenario B, and is a scenario C wherein emissions are so dramatically curtailed by the year 2000 that net emissions vanish. Both annual temperature and forcing values calculated from the model were obtained from NASA Goddard Institute for Space Studies (GISS).

Link: https://pubs.giss.nasa.gov/abs/ha02700w.html

Manatabe and Stouffer 1993

Forcings are calculated by digitizing the $4xCO_2$ time series (1% increase per year) in their Figure 1a and converting to a radiative forcing using equation 7. Temperature changes are calculated by digitizing panel the time series corresponding to the $4xCO_2$ experiment in Figure 1b.

Link: https://www.nature.com/articles/364215a0

IPCC First Assessment Report (FAR)

The main text of the IPCC FAR featured projections from a simple box-diffusion upwelling energy balance model (EBM) tuned to the individual climate models featured in the supplement to the report. We digitized EBM temperature valuesfrom Figure 8 in the Policymakers Summary. As the original values are unavailable in a digital form, the IPCC AR5 took a similar approach in digitizing old figures. The values we obtained through digitization were comparable to those in the AR5 WG1 Chapter 1 appendix. We chose not to directly use the digitized values reported in the AR5 as they only provided a high and low range of projections and did not include a best-estimate, and digitizing the best-estimate while relying on the digitized high and low values in the AR5 would introduce potential inconsistencies in the digitization approach.

The AR5 chose an unusual set of bounds for its reported FAR values, relying on a stringent mitigation scenario (Scenario D in Figure 9) as its lower bound and the best-estimate business-as-usual scenario as its upper bound. We instead use the values reported in Figure 8, which show a low estimate, best estimate, and high estimate of temperature change in the FAR business-as-usual scenario. The low and high estimates are used as the uncertainty bounds on the best estimate. External forcing values for the EMB were digitized from Figure 6 (also Figure A.6) using the business-as-usual scenario, with all three scenarios (low, best, and high) relying on the same underlying set of forcings.

Individual climate model projections featured in Figure S5 were obtained from the FAR supplementary materials. Climate models in the IPCC FAR from UKMET and GFDL use only CO_2 changes for future forcings. They did not have model years specified, so were aligned such

that their 1990 value was the model year in which CO_2 concentrations were closest to 1990 observations. GCMs included in the IPCC FAR employ scenarios where CO_2 or GHG forcing increases by 1% per year, while the simple energy balance models featured in the report used the IS92a scenario.

IPCC Second Assessment Report (SAR)

We digitized EBM valuesfrom Figure 19, using the 2.5ºC ECS run (including aerosols) as the best estimate and 1.5ºC/4.5ºC ECS runs as the low and high-end estimates. Similar to the FAR, the original values are unavailable in digital form and the values we obtained through digitization were comparable to those in the AR5 WG1 Chapter 1 appendix. Projected FAR EBM CO_2 concentrations were digitized from Figure 5 (IS92a scenario), while total external forcing was digitized from Figure 6.

GHG-only model runs (excluding aerosols) were used from the IPCC SAR for the individual models shown in Figure S5, as the specific aerosol forcings used differ by models and are poorly documented. SAR climate models mostly employ scenarios where CO_2 increases by 1% per year, while the simple energy balance models featured in the report used the IS92a scenario.

IPCC Third Assessment Report (TAR)

Decadal values for both temperatures, total forcings, and CO_2 used in the EBM featured in the TAR main text were obtained from Appendix I:
https://www.ipcc.ch/site/assets/uploads/2018/03/TAR-APPENDICES.pdf

These decadal values were transformed into annual estimates via linear interpolation.

Individual climate models featured in the TAR using the A2 SRES scenario were selected and shown in Figure S5, as that is the scenario with the most unique model runs available.

IPCC Fourth Assessment Report (AR4)

Coupled Model Intercomparison Project 3 (CMIP3) temperature projections featured in the AR4 were obtained from KNMI climate explorer. A1B runs were used as they were readily available, though over the 2007-2017 period differences between A1B and A2 in CMIP3 are minor. External forcing values used in the CMIP3 A1B scenario were based on those used in GISS model E, as precise forcing values used by each model are not readily available (and the differences within a given SRES scenario in forcings used between models should be small):
https://data.giss.nasa.gov/modelE/transient/dangerous.html

The AR4 best estimate shown in the paper is based on the A1B multimodel mean, while the low and high scenarios reflect the 5th and 95th percentile of the ensemble of A1B model runs for any given year.

Supplementary Figures S1-S5.

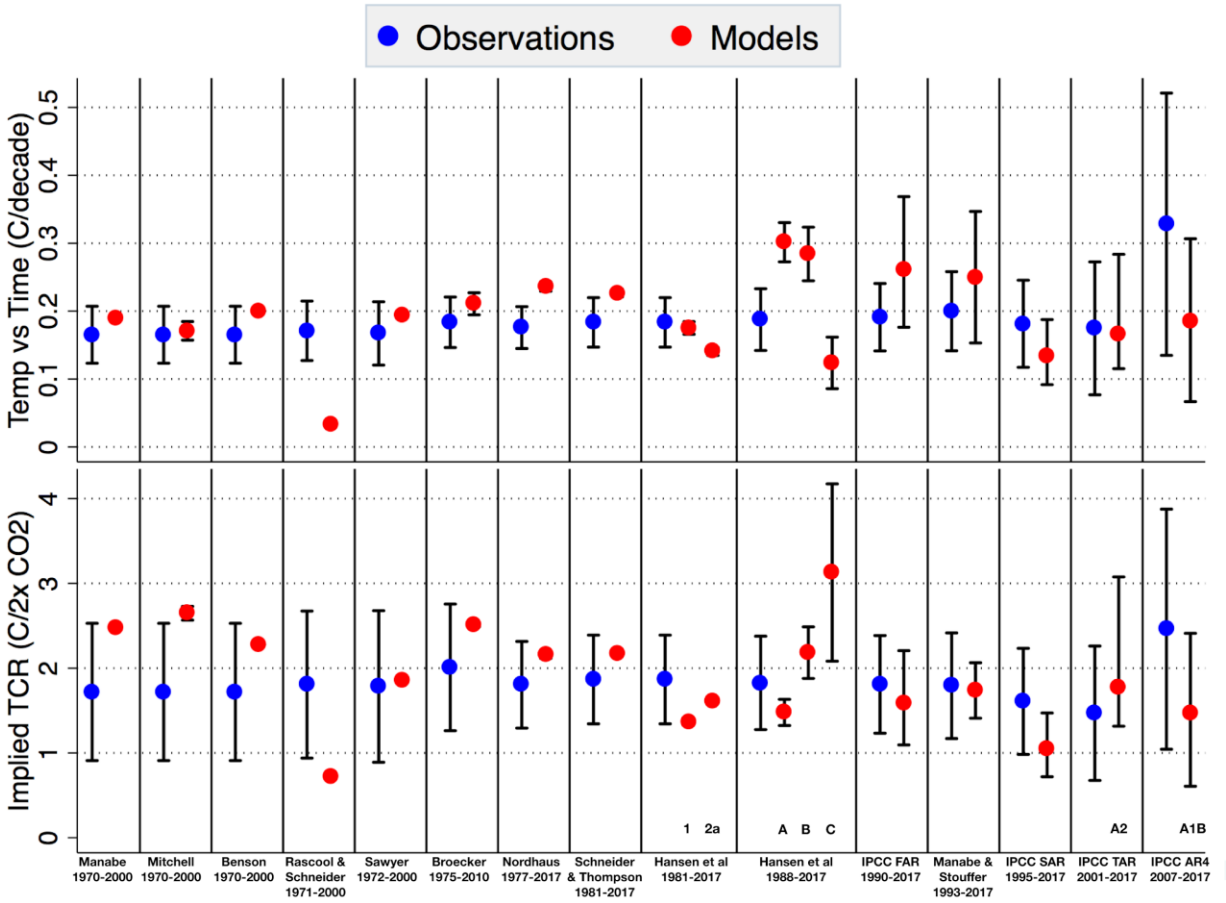

Figure S1. Comparison of trends in temperature vs time (top panel) and implied TCR (bottom panel) between observations and models over the model projection periods displayed at the bottom of the figure. As in Figure 1, but with the 2007-2017 IPCC AR4 projection included.

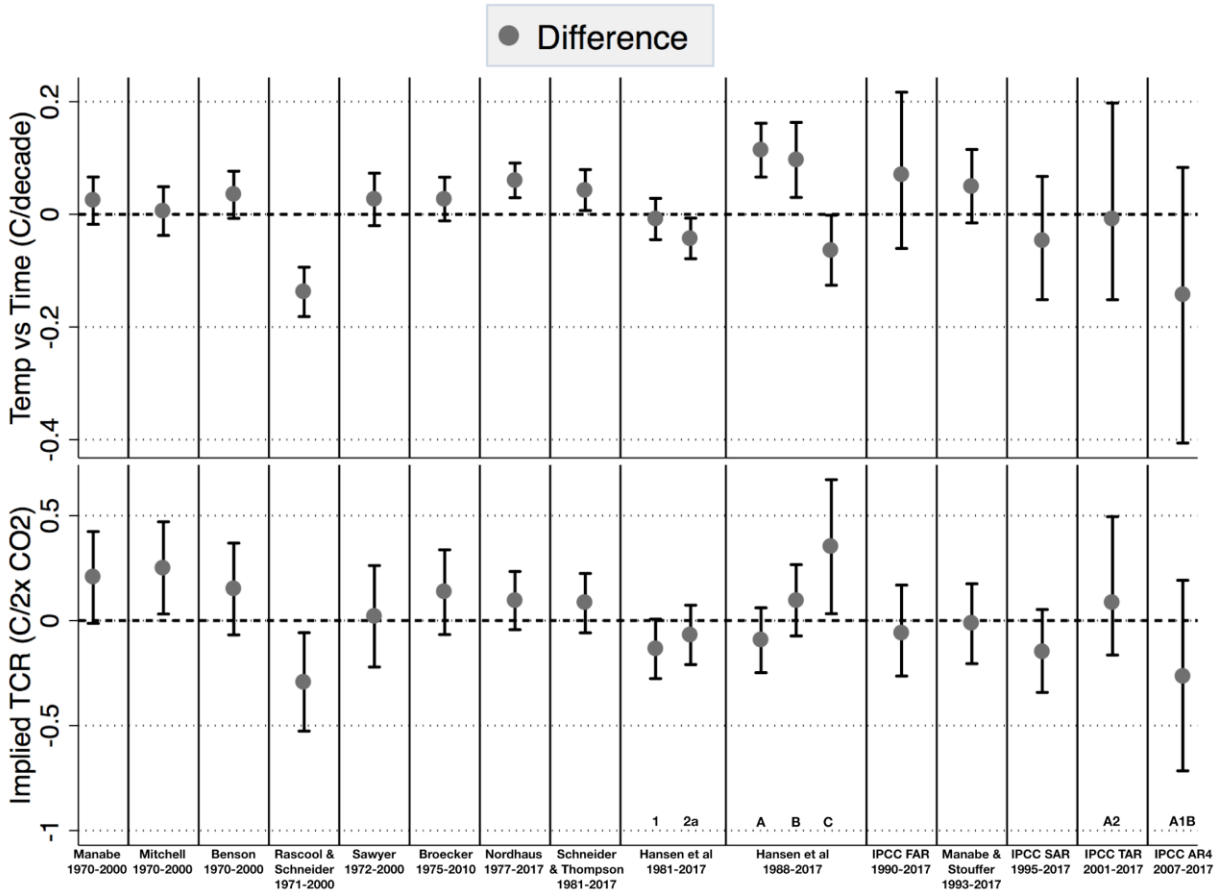

Figure S2. Difference between climate models and observations on a temperature vs time (top panel) and implied TCR (bottom panel) basis over the model projection periods displayed at the bottom of the figure. Values higher than zero indicate that the model projected more warming (or higher TCR) than observed.

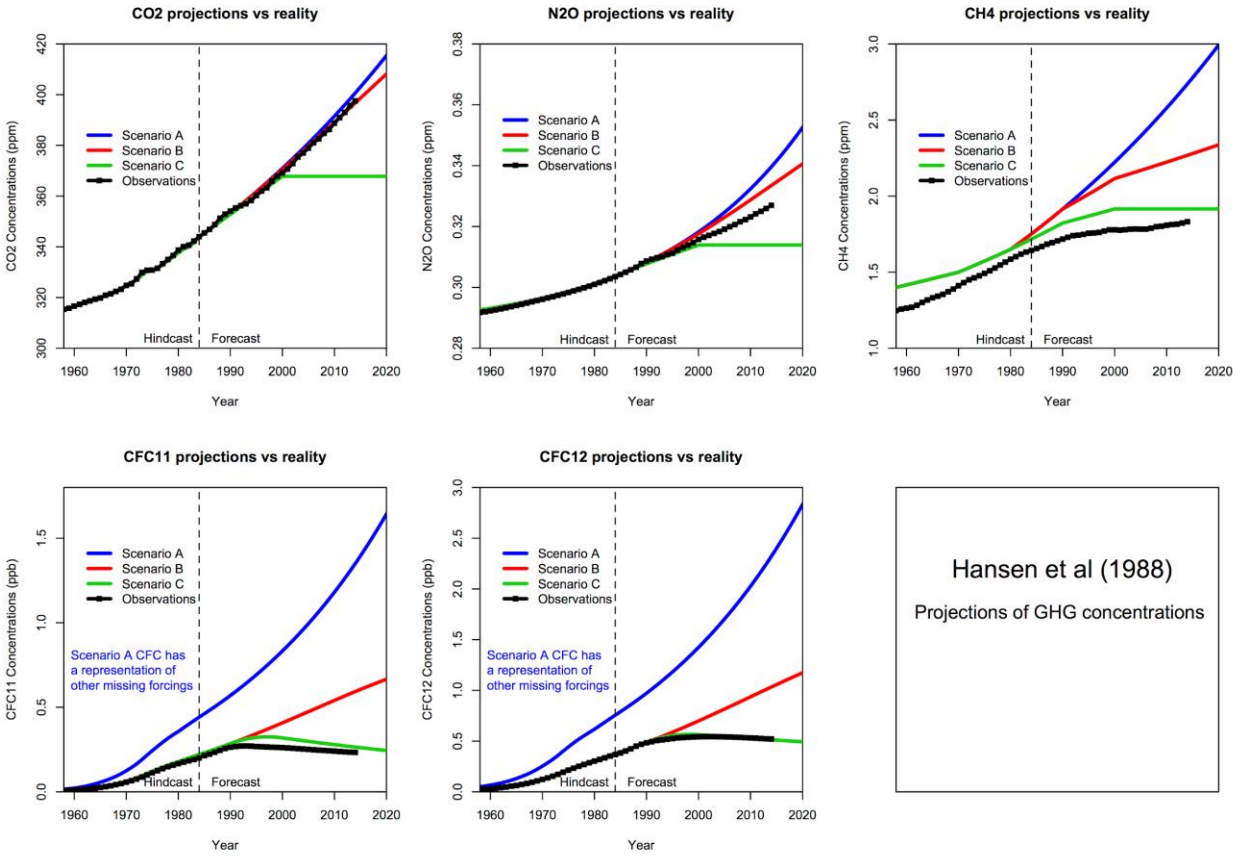

Figure S3. Greenhouse gas concentrations in Hansen et al. (1988) scenarios compared to observations.

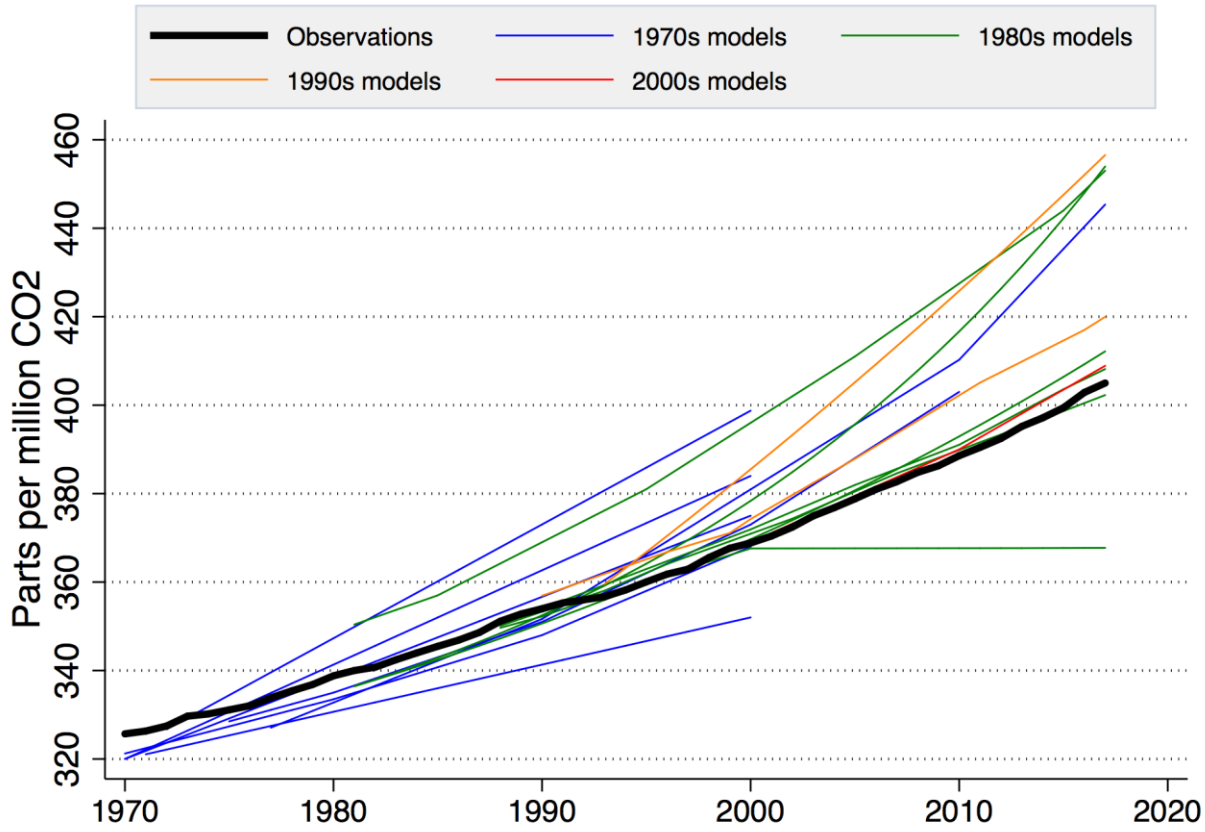

Figure S4*:* Model projected CO_2 concentrations colored by decade in which the model was published compared to observations (black). Observed CO_2 concentrations were taken from Meinshausen et al. (2017).

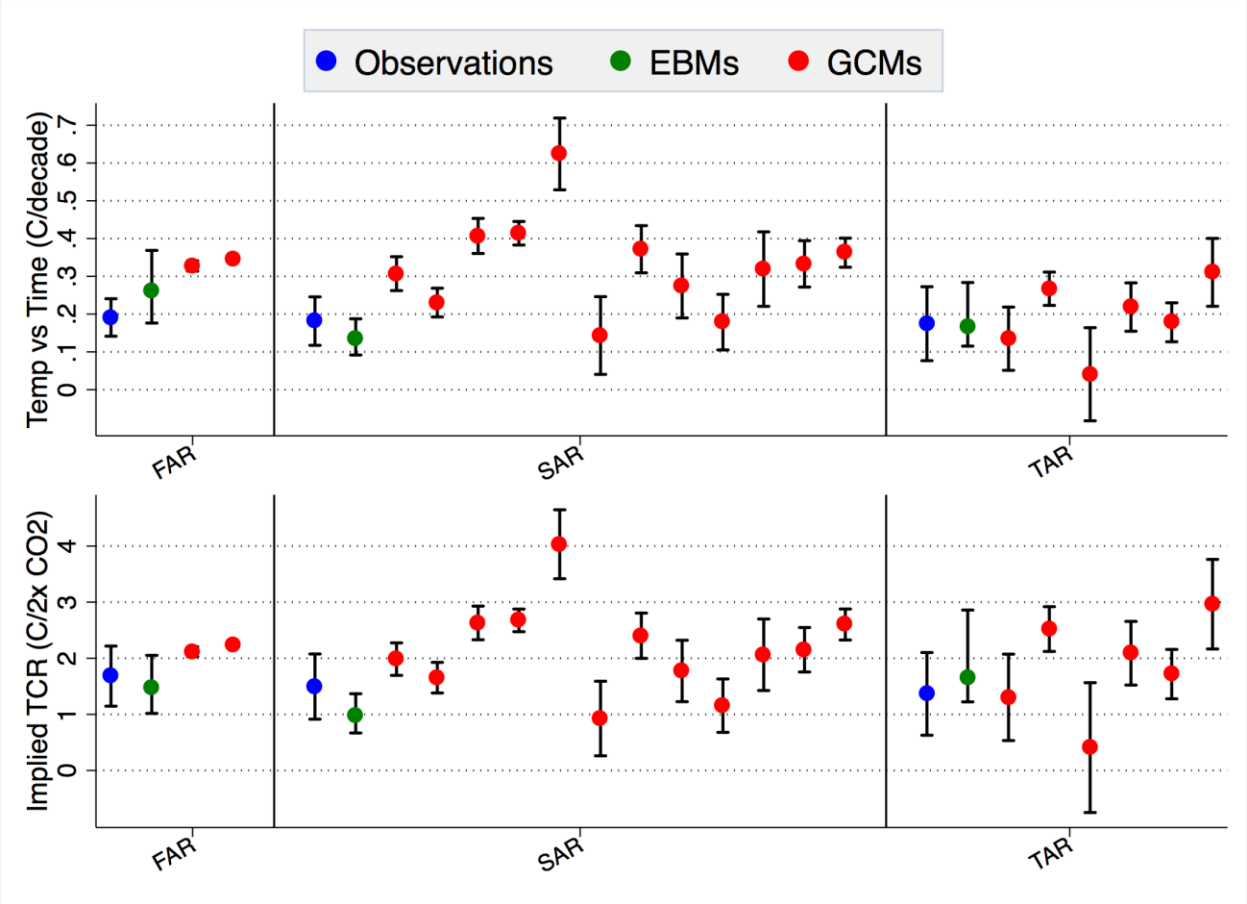

Figure S5. Comparison of trends in temperature vs time (top panel) and implied TCR (bottom panel) between observations and models included in the first three IPCC assessment reports over model future projection periods. Main-text projections based on simple energy balance models are shown in green (those are also included in Figure 1).

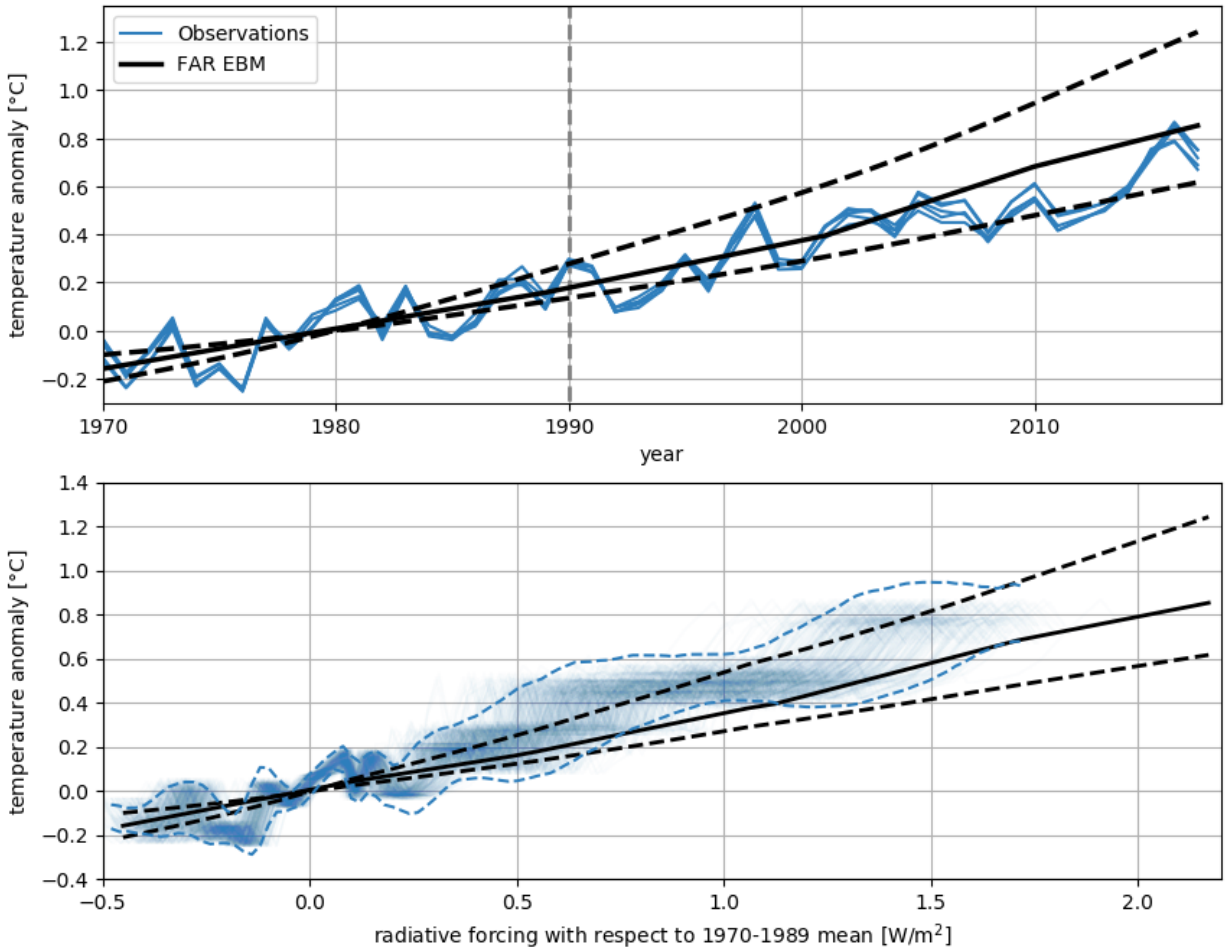

Figure S6: IPCC FAR projections compared with observations on a temperature vs. time basis (top) and temperature vs forcing (bottom). The dashed grey line in the top panel represent the start of the future projection period. The probability distribution in the lower panel represents the 5000 combinations of the 5 temperature observation products and the 1000 ensemble members of estimated forcings. Anomalies for both temperature and forcing are shown relative to a 1970-1989 pre-future-projection baseline.