Objectively combining climate sensitivity evidence

Climate Dynamics

Nicholas Lewis[1]

Supporting Information

S1. GCM-based estimation of F2x scaling factor and Historical forced pattern effect

A combined CMIP5 and CMIP6 model ensemble was formed as follows.

All CMIP5 models identified as having published fixed SST CO_2 ERF estimates were included, being 11 models from Vial et al (2013) Table 2, two additional models (CSIRO-Mk3-6-0 and MPI-ESM-P) from IPCC AR5 Table 9.5 (Flato et al. 2013), and two GFDL models from Paynter et al. (2018). The fixed SST CO_2 ERF estimates, were adjusted for land surface warming by dividing them by 0.949, the ratio of Smith et al. (2020)'s fixed SST CO_2 mean ERF estimates before and after such adjustment. The resulting estimates were analyzed in conjunction with CMIP5 abrupt4xCO2 simulation annual mean ΔT and ΔN data, as used for Lewis and Curry (2018) Table S2.

All CMIP6 models in common between those analyzed in Zelinka et al. (2020) and Smith et al. (2020) were included, apart from CNRM-ESM2-1. That model was excluded as, due to how the abrupt4xCO2 simulation was initialized, its stratosphere took 15 years to equilibrate. Annual mean abrupt4xCO2 simulation ΔT and ΔN data, as used for Zelinka et al. (2020), were analyzed in combination with 'ERF_ts' estimates from Smith et al. (2020) Table S1, of CO_2 ERF from a fixed SST analysis with a radiative kernel-based adjustment for land surface warming.

For a GCM, quadrupled CO_2 fixed SST simulations, with adjustment made for the effect of land surface warming ($F_{4\times CO2}^{\text{fixedSST}}$), provide an estimate of the actual ERF from a quadrupling in CO_2, which can be scaled to give estimated $F_{2\times CO2}^{\text{fixedSST}}$. The y-axis intercept from regression of annual mean ΔN data on ΔT over years 1−150 of an abrupt4xCO2 simulation provides another, usually biased downwards, CO_2 ERF estimate ($F_{4\times CO2}^{\text{regress}}$) and, post scaling, an estimate of $F_{2\times CO2}^{\text{regress}}$. The regression slope represents λ, with the x-axis intercept being S (pre-conversion from CO_2 quadrupling to doubling).

$F_{2\times CO2}^{\text{regress}}$ and $F_{2\times CO2}^{\text{fixedSST}}$ were derived for the 26 model ensemble, the estimated quadrupled CO_2 ERFs all being divided by the Meinshausen et al. (2020) $F_{4\times CO2}$ / $F_{2\times CO2}$ ratio of 2.10. The resulting ensemble-median $F_{2\times CO2}^{\text{fixedSST}}$ estimate was 3.97 Wm^{-2}, only 1% different from the 3.93 (S20: 4.00) Wm^{-2} estimate of the actual $F_{2\times CO2}$ value.

[1] Bath, United Kingdom. Email: nhlewis@btinternet.com

Values of the relevant estimates for each of the models in the ensemble are given in Table S1, along with ensemble statistics. Use of ensemble medians for $\Delta\lambda$ and $F_{2\times CO2}^{regress} / F_{2\times CO2}^{fixedSST}$ provided good prediction, across models in the ensemble, of regression-derived *S* from model $F_{2\times CO2}^{fixedSST}$ and $\lambda_{hist100}$ estimates. The resulting estimates were unbiased, and the correlation between the prediction error and *S* was low.

Table S1

Model	$F_{2\times CO2}^{fixedSST}$	$F_{2\times CO2}^{regress}$	$F_{2\times CO2}^{regress} / F_{2\times CO2}^{fixedSST}$	λ	$\lambda_{hist100}$	$\Delta\lambda^{forced}$
BNU-ESM	3.954	3.564	0.901	−0.936	−1.092	0.156
CanESM2	3.683	3.608	0.980	−1.023	−1.061	0.038
CCSM4	4.436	3.424	0.772	−1.226	−1.793	0.567
CSIRO-Mk3-6-0	3.111	2.404	0.773	−0.607	−0.983	0.376
FGOALS-s2	4.045	3.535	0.874	−0.871	−1.099	0.228
GFDL-CM3	3.512	2.910	0.828	−0.775	−1.062	0.287
GFDL-ESM2M	3.512	3.167	0.902	−1.360	−1.600	0.240
HadGEM2-ES	3.512	2.831	0.806	−0.655	−0.941	0.286
inmcm4	3.131	2.905	0.928	−1.483	−1.647	0.164
IPSL-CM5A-LR	3.251	3.044	0.936	−0.790	−0.885	0.095
MIROC5	3.984	4.129	1.036	−1.605	−1.531	-0.074
MPI-ESM-LR	4.335	3.927	0.906	−1.135	−1.328	0.193
MPI-ESM-P	4.315	4.112	0.953	−1.252	−1.358	0.106
MRI-CGCM3	3.612	3.133	0.867	−1.269	−1.553	0.284
NorESM1-M	3.482	2.957	0.849	−1.108	−1.450	0.342
CESM2	4.486	3.105	0.692	−0.633	−1.189	0.556
CNRM-CM6-1	4.014	3.467	0.864	−0.742	−0.938	0.196
EC-Earth3	4.043	3.152	0.780	−0.806	−1.151	0.345
GFDL-CM4	4.105	3.038	0.740	−0.821	−1.320	0.499
GISS-E2-1-G	3.690	3.752	1.017	−1.454	−1.428	-0.026
HadGEM3-GC31-LL	4.076	3.324	0.815	−0.629	−0.888	0.259
IPSL-CM6A-LR	4.048	3.248	0.802	−0.748	−1.041	0.293
MIROC6	3.695	3.476	0.941	−1.404	−1.544	0.140
MRI-ESM2-0	3.805	3.267	0.859	−1.097	−1.402	0.305
NorESM2-LM	4.105	3.276	0.798	−1.341	−1.880	0.539
UKESM1-0-LL	4.010	3.438	0.857	−0.673	−0.870	0.197
Median	3.969	3.271	0.861	−0.980	−1.255	0.250
Mean	3.844	3.315	0.864	−1.017	−1.271	0.254
Std Dev	0.380	0.397	0.084	−0.310	−0.296	0.166

Note: EC-Earth3's abrupt4xCO2 simulation data only extend to year 120.

Both the median and the mean of the calculated $F_{2\times CO2}^{regress} / F_{2\times CO2}^{fixedSST}$ ratios for the 26 models are 0.86; their standard deviation is 0.084. Since, from the regression line, $S = -F_{2\times CO2}^{regress} / \lambda$, while $F_{2\times CO2}$ corresponds to $F_{2\times CO2}^{fixedSST}$, this implies a 16% mean overstatement of S when using $F_{2\times CO2}$ in equation (4).

Figure S1.1 illustrates this point for the MRI-ESM2-0 CMIP6 model, which has an $F_{2\times CO2}^{regress} / F_{2\times CO2}$ ratio of 0.86 and an $F_{2\times CO2}^{fixedSST}$ of 3.81 Wm^{-2}, close to the ensemble medians. The MRI-ESM2-0 abrupt4xCO2 simulation ΔT and ΔN values and its $F_{4\times CO2}^{fixedSST}$ estimate have been rescaled by the factor required to convert its $F_{4\times CO2}^{fixedSST}$ to an $F_{2\times CO2}^{fixedSST}$ equaling the best estimate 3.93 Wm^{-2} $F_{2\times CO2}$ value.

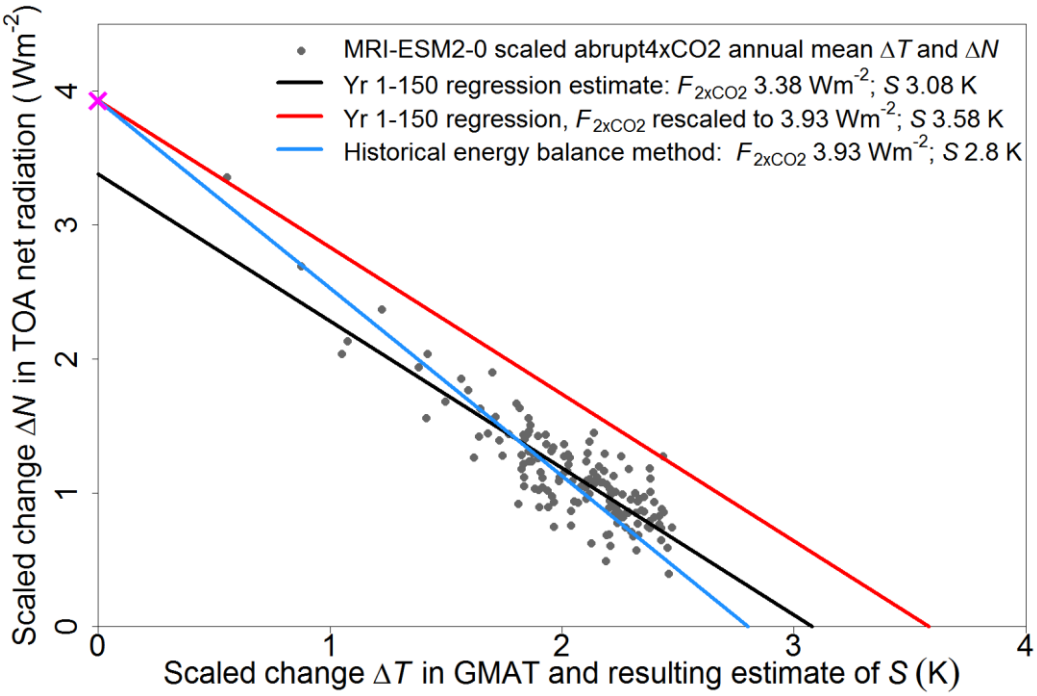

Figure S1.1. Illustration of the need to scale the actual $F_{2\times CO2}$ to avoid overestimation of S. The plot shows representative annual mean abrupt4×CO2 values. The black line shows the regression fit and resulting correct S estimate. It corresponds to use of equation (7). The red line shows the overestimation of S resulting from using equation (4), but substituting $F_{2\times CO2}$ for $F_{2\times CO2}^{regress}$, with λ estimated, in the standard way, as the regression line slope. The blue line corresponds to estimation of S_{hist} using equation (9), its slope being λ_{hist}.

The black line shows the fit from regressing simulation annual mean ΔN on ΔT (grey circles). Its slope, equating to λ, is -1.097 $Wm^{-2}K^{-1}$. The S estimate implied by the x-axis intercept is 3.08 K. The $F_{2\times CO2}$ estimate implied by the y-axis intercept, $F_{2\times CO2}^{regress}$, is 3.38 Wm^{-2}. That estimate is the $F_{2\times CO2}$ value required to obtain the correct S value when applying equation (4). It is a notional value that will not equal the actual $F_{2\times CO2}$ value, here 3.93 Wm^{-2} (shown by the magenta cross), unless

feedback is unchanging. Applying (4) using the model's actual $F_{2\times CO2}$, as estimated by $F_{2\times CO2}^{\text{fixedSST}}$, gives $S = 3.58$ K, a 16% overestimate (red line in Figure S1.1).

The difference between $F_{2\times CO2}^{\text{fixedSST}}$ and $F_{2\times CO2}^{\text{regress}}$ arises from climate feedback strength in the model abrupt4xCO2 simulation declining during the early decades, and thereafter changing little. A large increase in ΔT and large decline in ΔN (from the model's actual $F_{2\times CO2}$, as estimated by $F_{2\times CO2}^{\text{fixedSST}}$) occurs during the first decade, during which period λ is substantially stronger than later in the simulation. That behavior, which to a greater or lesser extent is almost universal in GCMs, causes the standard linear fit, which is dominated by data for the much larger number of subsequent years, to be poor in the early simulation years. The result is that $F_{2\times CO2}^{\text{regress}}$ (the *y*-axis intercept) significantly underestimates the model's actual $F_{2\times CO2}$ value.

In Table S1, $\lambda^{\text{hist.100}}$ is defined as:

$$\lambda_{\text{hist100}} = -\left(F_{2\times CO2}^{\text{fixedSST}} - 0.01\sum_{1}^{100} \Delta N / 2.10 \right) \Big/ \left(0.01\sum_{1}^{100} \Delta T / 2.10 \right) \qquad \text{(S1)}$$

while $\Delta\lambda^{\text{forced}} = \lambda - \lambda_{\text{hist.100}}$ provides an estimate of the forced historical pattern effect (S5.2.4).

That definition of $\lambda^{\text{hist.100}}$ corresponds to the slope of the blue line in Figure S1.1, the historical changes having for illustrative purposes all been scaled so that the total historical ERF change equals $F_{2\times CO2}$. This ERF change was spread over the historical period, so the time elapsed after each forcing increment until the total change is measured varies. The blue line therefore passes through a point representing an average of many grey circles (pairs of annual ΔT, ΔN values). The average used is the mean of values for the first 100 years, since summing equal annual forcing increments over that period has been found to well represent historical period energy budget estimation (Armour 2017; Lewis and Curry 2018). Taking means over the first 70 rather than 100 years would make little difference to λ_{hist}. The blue line in Figure S1.1 is therefore a fair representation of the estimation of λ_{hist} and S_{hist} in S20, as in equation (9), in the absence of any unforced historical pattern effect. Estimating *S* using equation (8) but without multiplying $F_{2\times CO2}$ by γ corresponds to adjusting the blue line's slope to that of the black line without also adjusting its *y*-intercept to equal that of the black line, results in overestimation of *S* (along the red line).

S2. Likelihood estimation in S20

As set out in the main text, S20 sample *S* uniformly and $F_{2\times CO2}$ *pro rata* to its PDF, the sample ratios providing λ samples, and for each line of evidence except Process they sample each remaining data-variable involved other than ΔT *pro rata* to its PDF[2]. They take the likelihood of each resulting

[2] S20 appear to regard all their data-variable estimates as representing (posterior) PDFs. If information about a data-variable is not represented by a PDF, but instead concerns its observed value and related error distribution, it can generally be converted into a posterior PDF for the data-variable using an Objective Bayesian approach, provided a noninformative prior distribution that is known to produce exact probability matching can be identified. Where the data-variable's estimate has a normal error distribution, with specified standard deviation and zero mean (as is the case for almost all of S20's data-variables), a uniform prior provides exact probability matching. Therefore, the data-variable's PDF will also be normally distributed, with mean equal to the data-variable's estimate and the specified error standard deviation.

multivariate sample as equal to the PDF of ΔT at the value implied by the sample's λ and data-variable values, allocating it to the S value involved. They bin the multivariate samples by their S values, and compute the S likelihood for each bin as the average of the likelihoods of the samples it contains.

S20 use the ΔT PDF to compute each sample's likelihood. However, ΔT has no special mathematical status in their model equations. If their method were valid, it should produce the same estimated likelihood for S if ΔT were included in the sampling and a data-variable other than ΔT, instead of being sampled, had its PDF used to compute the likelihood. As Figure S2.1(a) shows, that is far from being the case for Historical evidence. The solid blue and cyan lines show the likelihoods for S and S_{hist} from S20. The dashed red and magenta lines, which closely match the blue and cyan lines, are from my emulation of their method when making ΔT the non-sampled, likelihood-determining, data-variable. The dashed orange and brown lines are from my emulation of their method when making $\Delta F_{\text{Hist}}^{\text{aerosol}}$ the non-sampled data-variable instead. The estimated likelihoods alter very substantially when the non-sampled data-variable is changed; by a factor of four for S_{hist} at $S = 10$ K. That shows S20's method is unsound, since there is no reason to favor ΔT over any other data-variable when estimating likelihood.

Moreover, a simple calculation strongly suggests that S20's likelihood for the Historical S_{hist} case is unrealistic at high S. Suppose that an increase to $S_{\text{hist}} = 10\text{K}$, from the level implied by the mode (PDF maximum point) of every data-variable, results entirely from $\Delta F_{\text{Hist}}^{\text{aerosol}}$, the dominant source of uncertainty, strengthening; a change from -0.754 to -1.916 $\text{Wm}^{-2}\text{K}^{-1}$ would be required. That would result in the likelihood falling only to 0.35 of its maximum level, approximately three times as high as S20's estimate. The likelihood should actually remain above 0.35 of its maximum (as in the brown dashed line in Figure S2.1(a)), since changes in other data-variables can achieve part of the increase in S_{hist} at a lower cost in terms of likelihood reduction than if the entire increase resulted from varying aerosol ERF.

Similarly, major likelihood differences arise for the PETM when the unsampled data-variable used for sample-set likelihood computation is changed from ΔT to ΔCO2. There is an additional problem with S20's PETM likelihood estimation: their code[3] uses a 70 ppm standard deviation for ΔCO2, ten times smaller than the intended 700 ppm value. Hence S20's PETM likelihood does not correctly reflect their method and data, as Figure S2.1(b) shows: the magenta line uses the correct standard deviation, but the dashed cyan line, which uses the erroneous standard deviation, matches S20's likelihood. Using the correct standard deviation, their method produces a substantially different likelihood when the unsampled data-variable is ΔCO2 (green line) rather than ΔT (magenta line). The dashed blue line produced by this study's integrated likelihood method lies intermediate between the S20 method lines with alternatively ΔT or ΔCO2 as the unsampled variable (all using the correct standard deviation). The integrated likelihood peak, at 2.39 K, is very close to the S value of 2.38 K implied by the modes of each of the data-variable PDFs, where the likelihood product peak occurs, as one would expect. By contrast, S20's likelihood peak occurs at the much lower value of 2.03 K.

[3] Module 'paleo_hot_synth_16Mar.R', line 93 (Webb 2020).

The ΔCO2 standard deviation error affects S20's likelihoods that involve PETM evidence; however their eventual *S* estimates do not use PETM evidence.

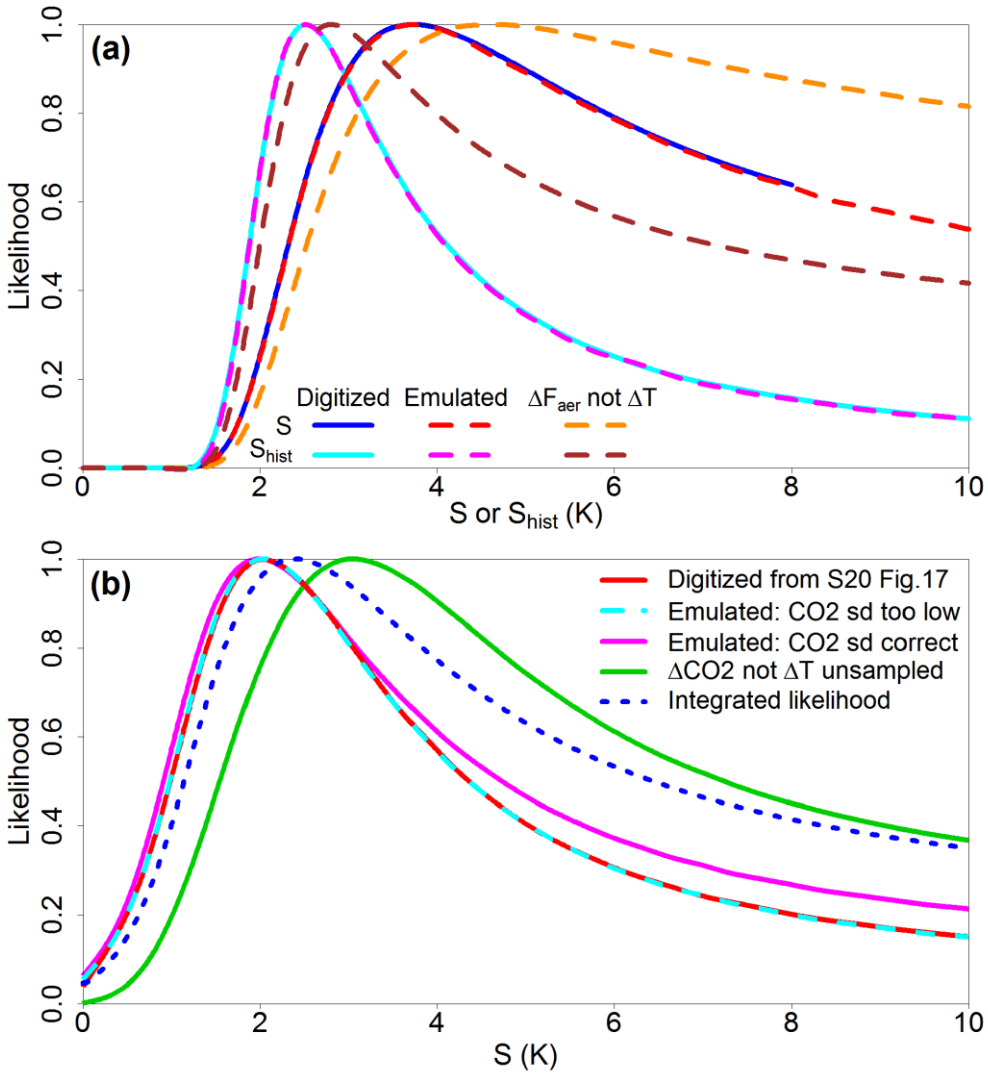

Figure S2.1. (a) Likelihood for *S* and S_{hist} from S20's Historical evidence. The solid blue and cyan lines are digitized from S20 Figure 20(b) and 11(a) ("Baseline") respectively. The dashed red and magenta lines are computed using emulation code, retaining Δ*T* as the non-sampled, likelihood-determining data-variable. The dashed orange and brown lines show the effect of exchanging the roles of Δ*T* and $\Delta F_{aerosol}$. (b) Likelihood for *S* from S20's PETM evidence. The solid red line is digitized from S20 Figure 18. The dashed cyan line, which closely matches the digitized line, uses the emulation code with Δ*T* as the non-sampled data-variable, but with the standard deviation of the ΔCO2 data-variable set to 70 ppm rather than its correct value of 700 ppm (S20 Table 9). The solid magenta line is derived as for the dashed cyan line, but using the correct ΔCO2 standard deviation. The solid green line is on the same basis but with ΔCO2 not Δ*T* being the unsampled data-variable. The dotted blue line shows the likelihood computed using this study's integrated likelihood method and S20's data.

S3. Noninformative priors: Derivation from profile likelihood, and calibration

Derivation from profile likelihood

A robust method of deriving a noninformative prior from the profile likelihood is to divide the profile likelihood into a directly sampling-derived PDF, but such a PDF is not available in combined-evidence cases. Moreover, it is useful to derive a Jeffreys' prior that is, like the profile likelihood, independent of sampling-based methods. Such a prior is derived, at each S value, as the distance $d(S)$ in whitened data space[4] that the profile-likelihood-maximizing point moves as S changes (Mosegaard and Tarantola 2002, (46); Lewis 2013b S5). This provides a direct measure of the local informativeness of the data about the parameter. The distances moved in each data dimension provide an approximation, along the profile likelihood path, to the (single-row) matrix $\mathbf{F}$ of partial derivatives of the data-variables with respect to S, while whitening of those variables is achieved by multiplying them by the square root of the data metric, that is, of the inverse data-variable covariance matrix, $\mathbf{C}^{-1}$. Thus

$$\pi_{\mathrm{JP}}(S) \propto d(S) = |(\mathbf{F}^{\mathrm{T}}\mathbf{C}^{-0.5})(\mathbf{C}^{-0.5}\mathbf{F})|^{0.5} = |\mathbf{F}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{F}|^{0.5} \qquad \text{(S2)}$$

This profile likelihood data-space movement based prior is actually the dimensionally-reducing factor that converts a PDF in data-space to a PDF in parameter (S) space (Mardia et al. 1979, their equation 2.5.16, which assumes that the data are already white).

Calibration

The priors derived from sampling-based methods should be proportional to the square root of the expected Fisher information. However, calibration to equality therewith is needed for the sum of priors in quadrature to be a Jeffreys' prior for inference from the corresponding product likelihood. Calibration is achieved by scaling all the priors to match, at their likelihood maximum, the square root of observed (Fisher) information (the negative second derivative of the log-likelihood with respect to the parameter at the observed data values). This measure is much more readily computed than expected Fisher information, for which it is a widely used proxy. They differ in that only expected Fisher information incorporates potentially but not actually observed data values. The two measures are identical for a normally distributed likelihood function, and there are arguments for preferring use of observed Fisher information (Efron and Hinckley 1978; Pawitan 2001, p.245; Brazzale et al. 2007, p.140).

The priors derived using the data-space movement method are also so calibrated. However, if the profile likelihood and data-space movement prior estimation methods are valid and work satisfactorily, the derived prior might be expected to equal the square root of observed Fisher information[5], at least where the data uncertainties are normal, and be already calibrated. At the likelihood maxima, their pre-calibration values do indeed agree very closely with the square root of observed Fisher information, validating both measures.

---

[4] With data-variables transformed to be uncorrelated, with unit variance.

[5] A non-rigorous argument for this is that, along the profile likelihood path, the second derivative of the log-profile-likelihood with respect to S equals the determinant of the second derivative matrix of that log-likelihood with respect to the data-variables (being the inverse data covariance matrix $\mathbf{C}^{-1}$ where the likelihood function is multivariate normal) with both its rows and columns multiplied by the partial derivatives of the data-variables with respect to S (being $\mathbf{F}$).

S4. Methods used to combine different lines of evidence

Some uncertain data-variables – $F_{2\times CO2}$, ζ and (when revising S20's assumptions) γ – are common to more than one of S20's statistical models, complicating the process of combining evidence. Moreover, two corresponding data-variables are treated as correlated between PETM and mPWP evidence. Excluding these, I make the reasonable assumption, as in S20, that uncertainties in estimates based on Process, Historical and LGM, mPWP and PETM Paleoclimate evidence are adequately independent, apart from a shared dependence on any data-variables in common.

Likelihoods and Jeffreys' priors are computed from each individual line of evidence using each of the methods described in Section 3.3. The relatively minor effect on S estimation of uncertainty in common data-variables is allowed for by separately estimating likelihoods and PDFs for S on a coarse grid of varying values for each of them. A grid with seven values spanning ±2 standard deviations (and hence the 95% confidence intervals) was found adequate for $F_{2\times CO2}$ and ζ; the resulting percentiles for S were within ±0.01 K of those using more values on a wider grid, or incorporating the full $F_{2\times CO2}$ or ζ uncertainty distribution. A grid of 11 values spanning ±3.33 standard deviations was used for γ.

Lines of evidence that have a data-variable in common only between them are combined first: LGM, mPWP and/or PETM, which have ζ in common, and – when S20's assumptions have been revised – Process and Historical, which then have γ in common.

At each grid value of the data-variables in common, a combined-evidence likelihood for S is computed as the product of the independent likelihoods from the individual lines of evidence involved. Since their respective Fisher informations combine additively, the Jeffreys' prior for combined-evidence therefrom is computed as the sum in quadrature of the individual calibrated Jeffreys' priors. The combined-evidence inference PDF for S is then derived, at each grid value, as the product of the thus computed combined-evidence likelihood and Jeffreys' prior, normalized to unit total probability.

Once combined-evidence gridded likelihoods and posterior PDFs have been computed, the common data-variable that relates only to the lines of evidence involved is eliminated from the PDFs by summing their values across its grid, applying weights that emulate the common data-variable's distribution. For likelihoods, the same approach is used, but those weights are multiplied by the relevant gridded PDFs, and the resulting weighted-average likelihood divided by the weighted-average PDF. Doing so emulates the sampling-based integrated likelihood derivation. Accordingly, ζ is eliminated once evidence from different branches of Paleoclimate evidence has been combined, and (if uncertain) γ is eliminated once Process and Historical evidence has been combined. Values on the $F_{2\times CO2}$ grid are retained until all lines of evidence have been combined, whereupon $F_{2\times CO2}$ is likewise eliminated. Once each common data-variable has been eliminated, a prior for S is derived by dividing the resulting PDF by the resulting likelihood. Since, by construction, this is an exactly probability matching prior, and a Jeffreys' prior provides the closest probability matching, it is a Jeffreys' prior.

The sampling-based integrated likelihood and doubled data methods cannot be used to combine mPWP and PETM evidence, due to ΔCO2 and ΔT uncertainty being regarded as correlated between

those periods. The profile likelihood method can directly probabilistically combine separate lines of evidence about *S*, taking account of any data-variables in common or that are correlated between lines of evidence. However, the optimization algorithm it uses cannot cope with combining S20's Historical evidence with all other lines of evidence. When using the profile likelihood method to combine different branches of Paleoclimate evidence, or (when an uncertain γ is specified) to combine Process and Historical evidence, computations are carried out over the grid of $F_{2\times CO2}$ values, but not over the grids of ζ or γ values.

Although these combination methods are two-stage, the final posterior PDF is effectively that from a single Bayesian step, since at the first stage likelihoods and Jeffreys' priors are computed for the partially-combined evidence.

For computational tractability, *S* is computed only over 0–20 K, on a 0.01 K grid. Most Bayesian climate sensitivity studies have imposed a lower limit of 0 K, and an upper limit of 10 K to 20 K. Samples that do not produce an *S* value in the 0–20 K range are assigned a zero *S*, but treated as being below that value, if they have an opposite sign ΔT to that of its median, and are otherwise assigned an *S* of 20 K, but treated as being above that value. Hence, probability outside 0–20 K is accounted for in sampling-derived posterior PDFs for individual lines of evidence. While the posterior PDFs for *S* may have non-negligible probability outside 0–20 K, particularly for Historical evidence, there is little advantage in using a higher upper limit, since most of that probability corresponds to an unstable climate system, with ΔR the opposite sign to the median ΔR. When two or more individual lines of evidence have been combined, a negligible amount of probability lies outside 0–20 K.

S5. Review and revision of S20 data-variable assumptions

This section reviews the evidence justifying revisions to various of S20's variable assumptions, and details each revised estimate adopted, for each line of evidence in turn. Uncertainties indicated by ± represent one standard deviation, with a normal distribution, denoted *N*(mean, standard deviation), assumed. The units of all feedback values given are $Wm^{-2}K^{-1}$.

5.1. Process evidence

5.1.1. Basis of estimation of component feedbacks

For derivation of λ from process evidence to be valid, feedbacks which appear to vary with timescale and/or cause need to be estimated from evidence consistent with the definition of λ, being forced feedbacks arising over 150-years following a abrupt increase in CO_2, and not from shorter term feedbacks such as under internal variability.

S20 discuss constraints on *S* from observations of global interannual radiation variability, which they say increase confidence that there are no significant unincluded feedbacks. Zhou et al. (2015) found that in CMIP5 models cloud feedback over 150-year abrupt4xCO2 simulations was well correlated that with from interannual internal variability, but with a regression coefficient of only 0.5. Colman and Hanson (2017) found that not only were ensemble-mean net cloud feedbacks from interannual and decadal internal variability much more positive than in abrupt4xCO2 simulations in CMIP5 models, but that separate long and short wave components had very different, compensating,

latitudinal patterns, suggesting the close correlation might be fortuitous. They also found that most non-cloud feedbacks at internal variability timescales were only weakly correlated with those at climate change timescales, and had substantially different ensemble-means. Dessler (2013) found no significant relationship exists between values in preindustrial control simulations and forced climate change scenario simulations for any feedback in the 13 CMIP3 GCM ensemble analyzed. These findings strongly suggest that feedbacks under internal variability are a poor guide to feedbacks over abrupt4xCO2 simulations, a conclusion also reached in Proistocescu et al. (2018).

Consistent with the definition of λ, S20's climate feedback estimates almost all involve regressing, usually on GMAT, changes in some variable(s) over CMIP5 and/or CMIP6 abrupt4xCO2 simulations (or in some cases using changes between their early decades and late decades). Doing so is particularly important for cloud feedback. The change in total net feedback between the early and late decades of CMIP5 GCM abrupt4xCO2 simulations is dominantly due to changing shortwave cloud feedback, linked to evolving changes in SST patterns affecting low clouds (Andrews et al. 2015). The estimates of low cloud feedbacks in the review study that S20 primarily relies on are based on observed cloud responses to changes in cloud-controlling factors (CCF), which responses are believed to be timescale invariant. Since GMAT-mediated changes in CCF in that study are consistent with those derived from changes over GCM abrupt4xCO2 simulations, the estimate of cloud feedbacks in S20 is consistent with the definition of λ.

5.1.2. Planck feedback

S20 take Planck feedback to be $N(-3.2, 0.1)$, derived principally from CMIP5 GCM simulations. Estimates from the three CMIP5-based studies they cite are highly interdependent and should be treated, together with the CMIP5 mean estimate from Zelinka et al. (2020) of −3.3, as a single estimate of −3.2. The slightly lower observational estimate taken into account by S20 (Dessler 2013) was derived from internal variability, and hence will reflect a different temperature change pattern than under CO_2 forcing. Moreover, this estimate is biased towards zero by regression dilution, since ordinary least squares regression was used and there was uncertainty in the temperature changes as well as in the radiative flux changes, with a resulting R^2 of only 0.59. Mean Planck feedback in CMIP3 models was estimated at −3.7 (Colman and Hanson 2013) on a climate change timescale, but lower on internal variability timescales. Estimated Planck feedback in the more advanced, CMIP6 models (Zelinka et. al 2020) averages −3.3, in line with that from physical expectation (S20). I adjust S20's estimate halfway towards that figure, to $N(-3.25, 0.1)$.

5.1.3. Cloud feedbacks

S20's +0.25 ± 0.16 estimate for tropical (0−30°) marine low-cloud feedback equals that in the Klein et al. (2017) review, with their uncertainty range increased substantially. Klein et al. base their estimate on the results from five studies examining the effect of changes in CCF on satellite cloud observations, standardizing their results to the same assumed sensitivity of SST and estimated inversion strength (EIS) to GMAT change. It is primarily EIS that exhibits a time-varying response to GMAT change. Klein et al. estimated a sensitivity of (tropical) EIS to GMAT of 0.14 K/K, citing two studies, the more recent of which (Qu et al. 2015) used abrupt4xCO2 simulations. The R^2 values in their regressions of EIS against SST are however mostly too small for their slopes to provide

useful sensitivity estimates; for only three of the eight GCMs does the R^2 exceed 0.4. Averaging the slopes for those three GCMs suggests a tropical EIS sensitivity to GMAT change of ~0.13 K/K. Changes between early and late periods arguably offer a more robust estimate. The mean increase in EIS between means over years 1−30 and 121−150 in abrupt4xCO2 simulations by the eight CMIP5 GCMs Qu et al. used appears to represent, allowing for the increase in area having positive EIS, a sensitivity to GMAT change somewhat below 0.14 K/K. Both $F_{2\times CO2}$ and λ estimated from mean changes between those two periods in abrupt4×CO2 simulations are fairly consistent with estimates derived by regression over years 1−150. However, the uncertainty in their estimate is large. Based on fixed SST simulations, Qu et al. estimated tropical EIS sensitivity to GMAT to be 0.2 K/K.

In all but one of the studies on which the Klein et al. (2017) estimate was based, the only CCF included were SST and EIS. Klein et al. considered SST and EIS to be the dominant CCF, but also stated that horizontal temperature advection, free-tropospheric humidity and subsidence collectively make non-negligible negative contributions to the predicted cloud feedback. Scott et al. (2020) found that, in observations, EIS and horizontal surface temperature advection were the dominant CCF for low clouds across much of the global ocean, followed by 700 hPa relative humidity, suggesting that CCF-based marine low-cloud feedback estimates using only SST and EIS might be unreliable.

S20 state that their feedback estimate is based on large-eddy simulations as well as Klein et al. (2017). While resolving turbulent eddies, large-eddy simulations still parameterize cloud microphysical processes, as well as the effects of sub-grid turbulent eddies and of radiative and surface turbulent fluxes, so may be an unreliable guide to low-cloud feedback. Perhaps reflecting such considerations, S20 make no adjustment to the Klein et al. central cloud feedback estimate to reflect the generally slightly stronger large-eddy simulation estimates.

S20 estimate mid-latitude (30−60°) low-cloud amount feedback as 0.12 ± 0.12, based partly on GCM evidence (Zelinka et al. 2016) and partly on extrapolating the Klein et al. (2017) tropical low-cloud feedback estimate, which included results from studies covering 40°S−40°N.

More recently, Myers et al. (2021) estimated 0−60° marine low-cloud feedback from observations as 0.19 ± 0.12, half S20's estimate. The authors, who include the lead and another author of Klein et al. (2017), argue that their estimate is more realistic than that in Klein et al. as they use a more comprehensive set of CCF and provide explicit evidence that trade cumulus feedback is likely to be weaker than stratocumulus feedback. They estimate changes in CCF between means over years 1−20 and 121−140 of abrupt4xCO2 simulations; $F_{2\times CO2}$ and λ estimated from mean changes between those two periods are closely consistent with estimates from regression over years 1−150.

Consistent with the Myers et al. (2021) results, Cessana and Del Genio (2021) estimate tropical low-cloud feedback, from changes in average CCF patterns over years 121-150 of abrupt4xCO2 simulations relative to preindustrial control simulations, to be only half as strong as Klein et al.'s estimate.

The Myers et al. estimate is about 0.06 lower than the corresponding marine mean low-cloud feedback in CMIP6 models (Zelinka et al. 2022). However, Mülmenstädt et al. (2021) argue, based on their model simulations, that mid-latitude, dominantly marine, low-cloud lifetime feedback is likely much more negative than in CMIP6 models.

I adopt, as being most realistic, the Myers et al. (2021) 0.19 central estimate of tropical and mid-latitude (0−60°) marine low-cloud feedback, but increase their uncertainty range to ±0.20, as in S20 (summing S20's 0−30° and 30−60° uncertainties in quadrature).

I make no changes to S20's assessments of other cloud feedbacks. However, I note that Lindzen and Choi (2021) cast doubt on the evidence, notably from Williams and Pierrehumbert (2017), relied upon by S20 that tropical anvil cloud feedback is not, as previously suggested (Lindzen and Choi 2011; Mauritsen and Stevens 2015), strongly negative.

The resulting median revised total cloud feedback estimate is 0.27 − almost double the 0.14 for nine CMIP6 GCMs that well represent observed interhemispheric warming (Wang et al. 2021).

5.1.4. Other feedback components

No changes are made to S20's assessments of non-cloud feedbacks.

5.1.5. Scaling $F_{2\times CO2}$

Since λ is derived from process evidence on a basis consistent with that from regression over 150-year abrupt4xCO2 simulations, it is necessary to adjust $F_{2\times CO2}$ to $F_{2\times CO2}^{regress}$ when applying (4), as explained in Sections 4.1 and S1. The 0.86 estimate of the $F_{2\times CO2}^{regress} / F_{2\times CO2}$ ratio derived in S1 is used, with the 0.084 standard deviation rounded up to 0.09, forming an additional data-variable, γ. This is treated as having a mean of 1 and negligible uncertainty when using S20's assumptions.

5.2. Historical evidence

Historical evidence is conventionally that arising over the period since about 1850, from which date reasonable estimates of GMST have been formed from instrumental measurements. S20 apply Eq.(1) to the disequilibrium changes in ΔT, ΔF and ΔN between averages estimated over base and final windows, respectively 1861−1880 and 2006−2018. They then adjust the resulting climate feedback value, λ_{hist}, for their estimated pattern-effect based difference between feedback applying over the historical period and that applying over 150 years after an abrupt CO_2 increase, in order to estimate λ. I consider these various elements in turn, retaining S20's window choices.

5.2.1. Temperature change

S20 estimate the change in GMST as 0.94 ± 0.07 K, measured as a blend of SST and land surface air temperature, the usual observationally-based method. Then, based on the ratio of GMAT to GMST in GCM historical simulations, they increase this to 1.03 ± 0.085 K, to obtain a GMAT estimate of ΔT. However, as discussed in Lewis and Curry (2018), observational measurements do not indicate that GMAT increases faster than GMST in the real climate system. AR6 Cross-Chapter Box 2.3 (Gulev et al. 2021) notes the conflicting evidence from models and observations, that all models share the same, potentially biased, parameterization of near surface temperature, and that theoretical understanding is limited. AR6 accordingly assesses a 90% uncertainty range for difference in GMAT warming relative to GMST warming as 0 ± 10%. I substitute this adjustment range for that used in S20.

S20's GMST estimate was infilled by kriging, which does not detect anisotropic features. Recently, a method that does detect anisotropic features was developed, with improved results (Vaccaro et al.

2021a,b). Infilling the same observational dataset as underlies S20's infilled estimate, the improved method estimates a 9% lower GMST increase. Nevertheless, I retain S20's estimate of the GMST rise, resulting in a GMAT ΔT estimate of 0.94 ± 0.095 K.

5.*2.2. Radiative imbalance change*

S20's 0.60 Wm^{-2} estimate of the change in planetary radiative imbalance equals that per AR6. However, AR6 (Gulev et al. 2021 Figure 2.26(b)) shows that, excluding series that are outliers, the AR6 0-2000m OHC estimate is middle-of-the-range in 2018 but at its bottom in 2006, hence yielding an above average increase over that period. Nevertheless, I retain S20's estimate.

5.2.3. Forcing change

The historical non-aerosol forcings time series used by S20 do not reflect the latest estimated greenhouse gas concentration−ERF relationships, nor recent assessments of some other forcings, and they have a 2008 discontinuity in volcanic forcing. I revise them in line with the estimates in AR6 (Gulev et al. 2021 Figure 2.10; dataset Smith et al. 2021), the CO_2 ERF estimates of which are consistent with the $F_{2\times CO2}$ value of 3.93 Wm^{-2} that I use. As in S20, no error in the $\Delta F_{CO2/2\times}$ term is allowed for since the fraction of $F_{2\times CO2}$ that ΔF_{CO2} represents is much less uncertain than is $F_{2\times CO2}$.

I also adjust the Black carbon on snow ERF for that part of its median assessed radiative forcing efficacy (range 2−4) that is not included in the ERF estimate (Gulev et al. 2021 7.3.4.3). The revised non-CO_2, non-aerosol ΔF is 0.33 Wm^{-2} higher than in S20, the largest differences being in volcanic and ozone ERF changes. I also adopt the AR6 non-aerosol ERF uncertainty estimates, which in total are similar to S20's. The AR6 forcing time series are also used to scale aerosol ERF estimates from 1850 to 2005−2015 to S20's base and final windows.

S20 use an unconstrained estimate of 1850 to 2005−2015 aerosol ERF from Bellouin et al. (2020; hereafter B20), which has a mean of −1.39 Wm^{-2} and 90% and 95% uncertainty ranges of respectively −3.2 to −0.4 and −3.7 to −0.3 Wm^{-2}. Its median is −1.18 Wm^{-2}. The AR6 assessment of aerosol ERF (Forster et al. 2021), adjusted to the same period, is −1.20 Wm^{-2}, almost the same median as in B20, but with a narrower, symmetrical, 5−95% uncertainty range of −1.85 to −0.55 Wm^{-2}.

B20's sophisticated method, while theoretically attractive, results in strong dependence on multiple assumptions and estimates, many of which require spatially and temporally resolved values. The resulting estimates of the weighting factors, or effective cloud fractions, applied to convert into global averages local estimates of perturbations in aerosol optical depth, and hence cloud droplet number concentration (N_d), and estimates of the sensitivity thereto of radiative changes, will be uncertain and may be substantially biased. For instance, B20's estimate of the effective cloud fraction for aerosol-driven cloud albedo changes (both c_N, before adjusting for liquid water path changes, and c_L, for that adjustment) is 0.19 to 0.29. By contrast, Stevens (2015) provides a range of evidence for a value of ~0.1, with an upper bound of 0.15, for that effective cloud fraction. Moreover, B20's estimated effective cloud fractions vary hugely between different aerosol forcing components; that from cloud area/lifetime changes is much higher, with a range of 0.59 to 1.07. That implies great sensitivity of their results to spatiotemporally-resolved estimates, which are based on snapshots of the aerosol-cloud field. As the authors say, the time-dependent nature of cloud adjustments means

that this may lead to an overestimate of the effect, or an underestimate due to undetected aerosol perturbations.

B20 central estimates for each uncertain factor in their theoretical model imply total aerosol ERF in line with their mean estimate, with approximately 30% from aerosol-radiation interactions (ERF_{ari}) and 70% from aerosol-cloud interactions (ERF_{aci}): 30% from the albedo effect (from reductions in cloud droplet sizes and via liquid water path changes) and 40% from the effect on cloud amount, in particular by affecting cloud lifetime. However, a satellite retrieval analysis involving a large sulphate aerosol anomaly over the northern Atlantic (Mallavelle et al., 2017) found that changes in cloud amount or liquid water path were undetectable.

Since B20 and AR6 were compiled, further evidence pointing to ERF_{aci} being less negative than estimated has emerged. Both B20 and AR6 assess the ERF_{aci} cloud liquid water path (LWP) adjustment as offsetting approximately 30% of RF_{aci}. Gryspeerdt, et al. (2019) found it to be twice as strong in global observations, offsetting some 60% of RF_{aci}. Glassmeier et al. (2021), studying frequently occurring non-precipitating stratocumulus clouds, find that ERF_{aci} estimates based on ship tracks are unrealistic and that the range of −0.36 to −0.011 used in B20 for the sensitivity of relative LWP to relative N_d ($\beta_{\ln L-\ln N}$) is insufficiently negative. They concluded that a realistic lower bound was −0.64 and that values of −0.3 (Possner et al. 2020) or −0.4 (Gryspeerdt et al. 2019) could be considered as possible central values. Although Gryspeerdt et al. commented that evidence from ship and volcanic aerosol perturbations pointed to LWP sensitivity to N_d being smaller than in their observations, Possner et al. found that LWP sensitivity increased with marine boundary layer depth, and that deep boundary layers were underrepresented in pollution tracks, process modeling, and in-situ studies of aerosol–cloud interactions in marine stratocumulus.

Replacing B20's range for $\beta_{\ln L-\ln N}$ with a −0.64 Wm^{-2} lower bound and a central value of −0.35 Wm^{-2} (averaging the Glassmeier et al. (2021) suggested −0.3 or −0.4 values) would change their median aerosol ERF estimate to −0.95 Wm^{-2}.

B20 state: "In reality local data typically comprise relatively small aerosol ranges and small albedo, N_d, or effective radius responses. If aci metrics or ERFs are based on aggregation of many such scenes they will tend to bias the relationships by (i) extending the range of conditions beyond the natural local fluctuations and (ii) removing the small-scale covariability between meteorology and aerosol." Glassmeier et al. (2021) point out that for sufficiently strong aerosol perturbations, the aerosol-induced cloud thinning due to a strong LWP reduction will lead to the complete dissipation of individual clouds in stratocumulus decks, implying that estimates of cloud amount sensitivity to aerosols may be too positive. Consistent with this, there is some evidence suggesting that aerosols might reduce cloud cover. In large-eddy simulations, Seifert et al. (2015) found that increased N_d, which aerosol causes, led to a reduction in cloud cover in near-equilibrium cumulus cloud fields. Lee et al. (2021) found that the COVID-19 lockdown measures in China, which caused a 40−70% reduction of aerosol emissions and aerosol optical depth, were accompanied by an increase in low-cloud cover in East Asia, both in observations and, more weakly, in simulations.

Assessments of the change in aerosol forcing over the historical period are strongly dependent on the level of preindustrial aerosols. Hamilton et al. (2018) found, using revised assessments of

preindustrial fire activity, that global model simulations predicted a 35% or 45% reduction in historical global cloud albedo forcing compared to estimates using respectively CMIP6 or AeroCom emissions data. An estimated upper limit to preindustrial fire emissions reduced the forcing by more than 90%. Forcing from aerosol-radiation interactions was also reduced, by up to 10%. Radiative forcing calculations by Liu et al. (2021), based on evidence from high-latitude ice cores, suggest that Southern Hemisphere fire emissions had a sufficiently strong decreasing trend over the past century to largely offset aerosol cooling from increasing fossil fuel and biofuel emissions.

Moreover, some of the apparent historical increase in aerosol forcing may actually represent a feedback. Paulot et al. (2020) found, in a CMIP6 model, that a stronger dependence of sea-salt emissions on temperature improved simulation of marine aerosol optical depth sensitivity to temperature and caused a negative radiative feedback that equates to a change of approximately -0.12 Wm^{-2} over the historical period.

B20 accepted that total aerosol ERF more negative than -2.0 Wm^{-2} rely on more speculative aerosol-driven cloud changes, and is inconsistent with observed changes in temperature and radiation, and as a result adopted a 5% bound for aerosol ERF of -2.0 Wm^{-2} rather than the -3.2 Wm^{-2} implied by their own analysis. S20 use the results from B20's own analysis to avoid possible circularity in estimation of *S*, on the grounds that the -2.0 Wm^{-2} constraint depends on energy-budget constraints and attribution studies. However, some of the evidence for aerosol ERF not being more negative than -2.0 Wm^{-2} appears to be independent of global energy-budget sensitivity estimation over the full historical period, which is what S20's estimation of *S* depends on. In particular, Stevens (2015) argued, from multiple lines of evidence, that under the standard assumption that the rise in global temperature is forced, the near-continuous increase in global temperature during the historical period implies that the net average forcing has been positive throughout the period, except for short periods, such as after volcanic eruptions. From that Stevens deduced that aerosol ERF cannot be more negative than -1.3 Wm^{-2}, or -1.0 Wm^{-2} based on northern hemisphere warming.

Stevens and Fiedler (2017) presented strong arguments and evidence against the claim in Kretzschmar et al. (2017) that in CMIP5 GCMs global mean aerosol ERFs as negative as -2 Wm^{-2} are still consistent with the observed Northern Hemisphere temperature increase. They showed that CMIP5 GCMs with the most negative aerosol ERFs exhibit behavior that calls their fidelity into question, such as an unrealistic pattern in aerosol radiative effects. Wang et al. (2021) made a similar finding in relation to a 30 model ensemble of CMIP6 GCMs, linking aerosol forcing with cloud feedback. The nine (B9) models with the least positive cloud feedback had a mean aerosol-mediated cloud radiative response (closely related to ERF_{aci}) of -0.05 Wm^{-2} and reproduced the observed mid to late 20th century hemispheric warming asymmetry better that the nine (T9) models with the most positive cloud feedback, which had a mean aerosol-mediated cloud radiative response of -1.02 Wm^{-2}.

Further, B20 note that requiring that each decade during the second half of the twentieth century has nonnegative total anthropogenic and natural ERF, taking into account a low efficacy of volcanic forcing, shows that it is unlikely that the aerosol forcing is more negative than -1.7 Wm^{-2}. Moreover, Golaz et al. (2019) found that an advanced GCM with historical aerosol ERF of -1.7

Wm^{-2}, tuned on the pre industrial climate, would only produce realistic GMAT projections if the aerosol forcing is scaled down to ~−0.9 Wm^{-2} (and, in addition, its climate sensitivity is halved).

Conservatively, in the light of the foregoing evidence pointing to aerosol forcing being weaker than implied by simply revising B20's $\beta_{\ln L-\ln N}$ estimate, I adopt a modestly weakened aerosol ERF estimate of −0.95 ± 0.55 Wm^{-2} over, as in B20, 1850 to 2005-15.This implies a 5–95% uncertainty range of −1.85 to −0.05 Wm^{-2}, which has the same lower bound as AR6's estimate, and is likewise symmetrical.

5.2.4. Pattern-effect based adjustment of λ_{hist} to λ

GCM-derived evidence suggests that climate feedback over the historical period (λ_{hist}) is stronger than that over 150 years following an abrupt increase in CO_2. S20 adopt an estimate for $\Delta\lambda$, being $\lambda - \lambda_{\text{hist}}$, of 0.50 ± 0.30. A non-zero $\Delta\lambda$ (a historical period pattern effect) is thought to result from differences in the evolution of those patterns over the historical period to that over 150 year abrupt4xCO2 simulations, largely through their effect on shortwave cloud feedback.

Andrews et al. (2018) estimated, using six GCMs from four modeling centers, that $\Delta\lambda$, was 0.64. Some 0.07 of this appears to be attributable to interannual fluctuations biasing their regression estimates (Lewis and Mauritsen 2021), giving a corrected estimate of 0.57. The Andrews et al. estimate of λ_{hist} was based on amipPiForcing simulations, in which the ocean component of the GCMs was absent, with an observationally-based estimate (the AMIPII dataset) of the evolving historical period pattern of SST and sea-ice used to drive the simulations, with no applied forcing changes; λ_{hist} was estimated by regressing annual mean ΔN on ΔT over 1870–2010.

The pattern effect can usefully be broken down into unforced and forced elements. The unforced historical pattern effect element reflects any effect on feedback of internal variability in SST and sea-ice patterns. The forced historical pattern effect primarily reflects changes in SST and sea-ice patterns over the course of abrupt4xCO2 simulations, to which the usual weakening of estimated λ over their course is attributed. This does assume that GCMs correctly represent the forced evolution of SST patterns, however Tang et al. (2021) find that the modeled SST pattern is an artifact due to known GCM biases.

I assess the forced historical pattern effect from GCM behavior in CO_2-forced simulations on the grounds that GCMs respond very similarly to evolving historical period ERF as they would if it had been entirely due to CO_2, that is there are no efficacy differences involved. Richardson et al. (2019) found no evidence in simulations by eleven GCMs for an efficacy effect on estimates of climate sensitivity derived from the historical period. Volcanic and black-carbon on snow forcings (which they did not investigate) are probable exceptions, but suitably chosen analysis periods can eliminate the effect of the former on historical period estimation of climate sensitivity, and the latter forcing is tiny (and its efficacy can be adjusted for). Lewis and Curry (2020) showed that in two GCMs where evolving historical forcing had been measured, it accurately matched that calculated from the sum of the evolving global ΔN, and ΔT multiplied by feedback strength estimated over the first 50 years of the GCM's abrupt4xCO2 simulation, after adjusting low efficacy volcanic forcing. Recently, Dong et al. (2021) estimated feedback over 1870-2014 in historical simulations by eight CMIP6 models for which historical ERF had been derived. They found that it was on average almost identical to (0.03 ±

0.17 $Wm^{-2}K^{-1}$ weaker than) that over a 70-year CO_2-doubling ERF ramp (1pctCO2), while, for the seven models in common, their feedback estimates are insignificantly (0.07 ± 0.17 $Wm^{-2}K^{-1}$) weaker on average than the $\lambda^{hist100}$ values in Table S1, which correspond to feedback over a 100-year CO_2 ERF ramp. None of these results suggest that any ERF efficacy difference exists over the full historical period.

The median forced difference between λ and λ_{hist} estimated from abrupt4xCO2 simulations in the 26 CMIP5 and CMIP6 model-ensemble in S1 is 0.25 ± 0.17 (Table S1), which is adopted as the estimate of the forced historical pattern effect. Tang et al. (2021)'s findings suggest a lower estimate, but no reduction is made on that account. The 0.25 estimate is marginally higher than the 0.2 assessed in S20. Deducting the adopted forced pattern effect estimate from the adjusted Andrews et al. (2018) total $\Delta\lambda$ estimate suggests that 0.32 $Wm^{-2}K^{-1}$ might be attributable to an unforced pattern effect.

However, use of the AMIPII SST dataset appears to generate unusually large estimates of the historical pattern effect, despite its global mean evolution being very similar to that in other datasets. Fueglistaler and Silvers (2021) concluded that the historical pattern effect would be smaller in simulations forced with any other SST reconstruction than AMIPII. They found that the time-varying pattern effect in amipPiForcing simulations is captured by the difference between SSTs in regions of tropical deep convection and mean tropical SST, and that the change in that difference between the pre-satellite and the satellite era is nearly three times as great for the AMIPII SSTs as for the average of five other SST reconstructions.

Substituting the HadISST1 SST dataset for AMIPII, with sea-ice evolution unchanged, weakened the historical $\Delta\lambda$ by approximately 0.35 in amipPiForcing simulations over 1871-2010 in two GCMs from different modeling centers (Lewis and Mauritsen 2020; Zhou et al. 2021).

In two GCMs, Andrews et al. (2018) found a 0.6 weakening in $\Delta\lambda$ when using the HadISST2.1 SST and sea-ice dataset in place of AMIPII (sea-ice in which is based on HadISST1 sea-ice, with minor modifications). Although the HadISST2.1 sea-ice dataset (Titchner and Rayner 2014) is no doubt imperfect (Andrews et al. 2018), its developers argue that it is an improvement on HadISST1's. However, I consider that there is too much uncertainty involved for any sea-ice related reduction to be made when estimating the unforced Historical pattern effect.

In view of the evidence that pattern effect estimates from AMIPII-based simulations are likely substantially excessive, and that the unforced element is probably minor and could potentially be negative, it is difficult to justify making a significantly positive estimate for the unforced element. However, a nominal 0.1 ± 0.25 is added to the 0.25 ± 0.17 forced pattern effect estimate, which reflects the substantial uncertainty and allows not only for any unforced pattern effect but also for the possibility that some other element of the revised Historical evidence data-variable distributions might be misestimated. Summing uncertainties in quadrature, the $\Delta\lambda$ estimate becomes 0.35 ± 0.30. This uncertainty is identical to that in S20.

5.2.5. Scaling $F_{2\times CO2}$

Having adjusted λ_{hist} to be consistent with estimates of λ derived from regression over 150-year abrupt4xCO2 simulations, it is necessary to convert $F_{2\times CO2}$ to $F_{2\times CO2}^{regress}$ when applying (4). The energy

budget estimation method used in S20 to estimate λ_{hist} divides the change in global temperature over the historical period by the excess over that period of the change in ERF over the change in TOA radiative imbalance, with S_{hist} derived by dividing the resulting λ_{hist} into $-F_{2\times CO2}$: (9).

The 0.86 ± 0.09 estimate of the $F_{2\times CO2}^{regress} / F_{2\times CO2}$ ratio adopted in S5.1.5 applies for estimating S (but not S_{hist}) from Historical evidence, as well as Process evidence.

5.3. Paleoclimate evidence

5.3.1. Conversion of ECS to S

Because of the long timescales and hence quasi-equilibrium changes involved, paleoclimate sensitivity estimates are of ECS (or ESS). As discussed in Section 2, I adjust them for the slowness of equilibrium using an estimated ECS to S ratio excess over one (ζ) of 0.135, rather than S20's 0.06. This is based on the mean estimate from sixteen long run doubled and quadrupled CO_2-forced GCM simulations investigated by Rugenstein et al. (2020), data from the FAMOUS model's abrupt4xCO2 simulation, which exhibits incipient instability, being excluded. I adopt a standard deviation for this adjustment of 0.1, which compares with 0.05 for the simulations involved, and implies a 9% probability that ζ is negative. Using S20's scaled-up standard deviation estimate for ζ of 0.2 would produce a 25% probability that S exceeds ECS. In all sixteen long run CO_2-forced GCM simulations investigated by Rugenstein et al. (2020), ζ was found to be positive, strongly suggesting that S20's uncertainty estimate substantially exaggerates the possibility that S exceeds ECS.

5.3.2. LGM

S20's ΔT_{LGM} estimate has a $N(-5, 1)$ K distribution, relative to preindustrial climate. They cite nine sources in support of their estimate, eight of which provide their own estimates, two being based on the same ice core. As S20 say that theirs is an observational estimate, where a source gives both proxy-only and model-proxy based reconstructions, it is appropriate to use the former. The sources that S20 cite, with in brackets their best observational estimate of the LGM to preindustrial GMAT difference, are: Annan & Hargreaves (2013) [−4.0 K]; Friedrich et al. (2016) [−4.6 K, after adjusting for a 0.4 K difference between preindustrial and the warmer period circa 10 ka BP that their estimate is for]; Hansen et al. (2007) [−4 K: their Fig. 2(c)]; Kohler et al. (2010) [−4.5 to −5 K, but only as 0.5 x EPICA ice-core]; MARGO (2009) [−2.7 to −3.8 K, upon dividing their SST change by alternative factors of 0.7 or 0.5]; Masson-Delmotte et al. (2010) [−4.5 K to −5 K from EPICA ice core multiplied by their estimated 0.5× factor: their Fig.7]; Rohling et al. (2012) [no global estimate given]; Schmittner et al. (2011) [−3.0 K]; Snyder (2016) [−5.0 K: 6.17 K median cooling between the averages over 0–5 ka BP and over the 19–23 ka BP LGM stable period, scaled by 4.5/5.1 to reflect the overestimation shown in their Extended Data Fig.5, and adjusted by the 0.45 K excess of 0–5 ka BP over preindustrial GSAT per the Osman et al. (2021) high temporal resolution proxy-only reconstruction]. The average of the eight estimates (taking the centers of their ranges in three cases) is 4.2 K, or 4.1 K if 50% weighting the two EPICA ice-core estimates.

Shakun et al. (2012), not cited by S20, estimates ~−3.5 K LGM ΔT relative to mid-Holocene GMST.

Moreover, temperature reconstructions based on borehole thermometry and firn properties suggest that interpretation of ice core water isotopes using modern spatial slopes overestimates last glacial maximum surface cooling in central East Antarctica (Buizert et al. 2021).

I revise S20's central LGM cooling estimate of −5 K to −4.5 K, primarily reflecting, less than fully, the −4.2 K adjusted mean ΔT_{LGM} estimate of the sources cited by S20, and increase the standard deviation estimate to 1.25 K so as to maintain the same –7 K lower bound of the 95% uncertainty range as S20's.

S20 use the single year 1850 as their preindustrial reference period for GHG concentrations, whereas for observational estimates of temperature change preindustrial generally refers to the average over 1850−1900. For consistency, the S20 GHG forcing changes should therefore use mean 1850−1900 GHG concentrations. Doing so would change the CO_2 ERF from –0.57x to –0.59x $\Delta F_{2\mathrm{xCO2}}$, as well as marginally changing the CH_4 and N_2O ERFs. However, conservatively, I do not adjust S20's LGM forcing estimates to be consistent with the LGM ΔT measure.

S20 adopt the estimate of vegetation forcing in the Kohler et al. (2010) comprehensive assessment of non-greenhouse gas LGM forcing changes, but use a central estimate of –1.0 Wm^{-2} for aerosol (dust) forcing in place of Kohler et al.'s –1.88 Wm^{-2}. This seems questionable; Friedrich and Timmermann (2020) adopt Kohler et al.'s estimate, while pointing out that estimates of its glacial-interglacial magnitude vary from ~0.33 to ~3.3 Wm^{-2}. I nevertheless accept S20's estimate of dust forcing, along with that of vegetation forcing, and their calculations of the changes in greenhouse gas ERFs.

However, S20's estimate of –3.20 Wm^{-2} combined forcing from changes in land ice sheets and the resulting sea level changes appears too small. Kohler et al. (2010) estimate –3.17 Wm^{-2} from the effects of land ice, plus a further –0.55 Wm^{-2} from the change in albedo due to the fall in sea level exposing more land surface. S20 dismiss the latter, without citing any evidence, as a "less commonly discussed factor", but Zhu and Poulsen (2021) cite seven studies that account for it. I revise S20's combined −3.20 Wm^{-2} estimate to match Kohler et al.'s –3.72 Wm^{-2}. Their uncertainty estimates are identical. PMIP3 GCM simulations of the LGM show a range from –3.6 to –5.2 Wm^{-2} for the total land ice related forcing change (Braconnot and Kageyama 2015). That is considerably stronger than in PMIP2, due mainly to use of more recent LGM ice-sheet reconstructions.

I accept the overall non-CO_2 forcing uncertainty estimate used in S20, which exceeds that calculated from its components. The resulting revised non-CO_2 ERF change estimate is $N(-6.67, 2.0)$.

S20 assume that climate feedback in equilibrium (λ') strengthens by α for every –1 K change in ΔT, resulting in the $0.5\alpha\Delta T_{\mathrm{LGM}}^2$ term in (11), reducing LGM-estimated ECS. Contrariwise, Zhu and Poulsen (2021) found that ocean feedback caused 25% higher LGM-estimated ECS. Moreover, a significant part of the reduction in mean surface air temperature at the LGM is due to ice-sheet caused increased land elevation, which would weaken λ' compared to in non-glacial climates. Although S20's $N(0.1, 0.1)$ α estimate appears questionable, I retain it.

5.3.3. mPWP

S20's observational ΔT_{mPWP} estimate has a $N(3, 1)$ K distribution, relative to preindustrial climate. The range of available proxies is limited. S20 focus on tropical SST, for which estimates have varied

considerably. S20 adopt the 1.5 K estimate in Herbert et al. (2010), a compilation focusing on arguably more reliable alkenone proxies. McClymont et al. (2020) estimated that global mean SST was ~2.3 K warmer than preindustrial, using combined Mg/Ca and alkenone proxy data. Multiplying by 1.15 (Gulev et al. 2021: AR6 2.3.1.1.1) would convert the 2.3 K SST warming to a 2.6 K estimate of GMST change. McClymont et al. estimated 43% higher warming using only alkenone proxies, at a more limited range of sites. However, much of that estimated warming appears to reflect extratropical north Atlantic proxies, treated under their reconstruction method as applying to all longitudes. Tierney et al. (2019), using a larger set of alkenone-only proxies with much greater representation in the Pacific, estimated mean tropical SST warming in the mPWP at 1.4 ± 0.25K.

S20 multiplied their tropical SST mPWP warming estimate by 2.0, scaling it up to 3.0 ± 1.0 K, on the basis that studies over the last 0.5 Ma estimated tropical SST change to be only ~50% of global mean surface air temperature change. However, since that ratio will be climate-state dependent it is appropriate to estimate it in mPWP conditions rather than during glacial cycles over the last 0.5 Ma, when ice sheet changes are likely to have increased the ratio. Extratropical proxy coverage appears too limited for reliable estimation of the ratio, however the 16 climate models involved in PlioMIP2 (Haywood et al. 2020) showed an average ratio of mPWP GMAT change to tropical SST warming of 1.65 ± 0.18.

Although the Tierney et. al (2019) 1.4 K tropical SST warming estimate appears more reliable than S20's 1.5 K, I retain the latter but multiply it by the 1.65 PlioMIP2 ratio, giving a revised GMAT ΔT_{mPWP} of 2.48 K.

S20's assumption that ΔT_{mPWP} uncertainty is ±1 K, no higher than in the LGM, seems questionable, given the much more extensive LGM proxy evidence available. I increase it to 1.25 K, thereby retaining the same 95% upper bound of the distribution as S20.

I retain the S20 estimate of the mPWP CO_2 level, being 375 ± 25 ppm, with one standard deviation uncertainty of 25 ppm, with resulting 1.32 ± 0.088 estimate of $\Delta CO2_{\mathrm{mPWP}}$, and S20's 0.4 ± 0.1 estimate for $f_{\mathrm{mPWP}}^{\mathrm{CH4}}$.

In view of the poor knowledge of mPWP changes arising from slow feedbacks, S20 estimate ESS and divide it by an uncertain factor, $(1+ f_{\mathrm{mPWP}}^{\mathrm{ESS}})$, to convert to ECS. They estimate f_{ESS} as $N(0.5, 0.25)$, reflecting the range of 1 to 2 and mean of 1.5 for the ESS/ECS ratio in the 8 model PlioMIP1 ensemble (Haywood et al. 2013).

I concur with S20's approach for estimating f_{ESS}, but substitute values from PlioMIP2, which involved 16 more advanced models and updated estimates of mPWP boundary conditions. The ESS/ECS range is 1.11 to 2.85, with a mean of 1.67 and standard deviation of 0.47. I revise the f_{ESS} estimate to $N(0.67, 0.4)$, rounding down the standard deviation, which keeps the probability of f_{ESS} being negative below 1 in 16.

5.3.4. PETM

S20's observational estimate of ΔT_{PETM}, relative to that in the early Eocene, has a $N(5, 2)$ K distribution. S20 cite estimates from various sources, ranging from 4 K to 6 K, in support of the 5 K

value. A recent study (Inglis et al. 2020) estimates a 4.6 K global mean temperature change. S20's wide uncertainty range reflects the limited number of estimates and uncertainty in the interpretation of measurements from deep in the paleorecord. I retain S20's ΔT_{PETM} distribution.

S20 assessed a $N(2400, 700)$ ppm distribution for CO_2 concentration in the PETM relative to a baseline of 900 ppm, implying a $N(1.667, 0.778)$ $\Delta CO2_{\mathrm{PETM}}$ distribution. That covers, within its 90% uncertainty range, a concentration ratio range $(1+\Delta CO2_{\mathrm{PETM}})$ of 1.39 to 3.95. The CO_2 concentration estimates considered by S20, even taking extremes of both their PETM and Eocene ranges, constrain $(1+\Delta CO2_{\mathrm{PETM}})$ within 1.4 to 5. Using instead that range would lower PETM based *S* estimates. Nevertheless, I retain S20's $\Delta CO2_{\mathrm{PETM}}$ distribution.

S20 assume a purely logarithmic relationship between CO_2 forcing and concentration. That significantly understates the change in forcing at the high PETM concentrations. I use the forcing change from 900 to 2400 ppm given by the Meinshausen et al. (2020) formula, which is 11.7% higher than that based on a logarithmic relationship. Applying that uplift matches their formula within ±2% up to 1400 ppm (two standard deviations) away from 2400 ppm. I therefore scale S20's logarithmically-based forcing by a factor of 1.117. I ignore uncertainty in the scaling, which is negligible relative to that in unscaled forcing. While Meinshausen et al. assume a fixed ratio of CO_2 ERF to stratospherically-adjusted radiative forcing, there is modeling evidence that fast adjustments become more positive at higher temperatures (Caballero and Huber 2013), which would further increase CO_2 ERF change in the PETM. I make no adjustment for this effect.

To account for forcing from changes in CH_4 concentrations, S20 apply the same 0.4 f_{CH4} factor to the CO_2 forcing change as for the mPWP, with doubled uncertainty, although noting that the tropospheric lifetime of CH_4 could be up to four times higher given sustained large inputs of CH_4 into the atmosphere (Schmidt and Shindell 2003). I retain S20's f_{CH4} distribution, although doing so may bias estimation of *S* upwards.

S20 assume that ESS for the PETM was the same as present ECS , representing uncertainty regarding this by deducting a $N(0, 0.5)$ adjustment (β) from ESS feedback when estimating ECS feedback, λ'. Assuming zero slow feedbacks in the PETM (so ESS equals ECS) may be reasonable, given the lack of evidence and the absence of major ice sheets. However, Caballero and Huber (2013) and Meraner et al. (2013) both found, in modeling studies, substantially (~50%) weaker climate feedback for climates as warm as the PETM. Zhu et al (2019) found, in a state-of-the-art GCM, that ECS was over 50% higher than in present day conditions, with little of the increase being due to higher CO_2 ERF. I therefore consider that it would be more realistic to use a positive central estimate for β. Nevertheless, I retain S20's estimate.

5.3.5. Error correlations

The proxies used to estimate the forcing and temperature at the PETM are of a similar nature to those used for the mPWP, so errors in them are likely to be correlated. S20 assume, for both variables, a 0.8 mPWP−PETM error correlation. However, even for identical proxy types, errors will depend on particular proxy samples and locations, on calibration error and on the climate state, all of which will differ, reducing mPWP−PETM error correlation.

Moreover, with the uncertainties assumed in the estimates of temperature and CO_2 change at the PETM being much greater those assumed for the mPWP, particularly for log(CO_2), it is doubtful that a correlation as high as 0.8 could arise. I consider that a correlation of 0.6 for ΔT and 0.4 for ΔCO_2 would be more realistic. However, it was found that, when combining all paleoclimate periods using either those mPWP−PETM correlations, or 0.8 correlations, the final combined-evidence estimates of S were almost identical to those when the only LGM and PETM data were combined to form the Paleoclimate evidence. Therefore, only results where the Paleoclimate evidence represents either LGM and mPWP combined, or LGM and PETM combined, are presented.

S6. Differences in Jeffreys' priors using S20s' original assumptions and the revised assumptions

The Process evidence priors should be identical, or nearly so, using the original or the revised assumptions, since they are calibrated to Fisher information, which can be expected to be fairly insensitive to most of the data-variable revisions. However, the inclusion of the $F_{2\times CO2}$ scaling factor γ, and the slightly lower $F_{2\times CO2}$, reduce the revised-evidence λ prior when it is transformed into one for S, as they scale the transformation. When the Process evidence prior using the revised assumptions is rescaled to reflect those changes, it agrees closely with the prior using S20's assumptions, both in magnitude and shape.

For Historical evidence, when the same rescaling is effected the prior derived using the revised data-variable assumptions is higher at low S, but lower at high S, than when using S20's data-variable assumptions. That is almost certainly due primarily to the very different aerosol forcing distributions involved, which represent the dominant uncertainty in Historical evidence, combined with the smaller ΔT per the revised assumptions, which will reduce the data informativeness about S.

For Paleoclimate (LGM + mPWP) evidence, the revised ζ estimate is higher than S20's. That reduces the revised data-variables' informativeness about λ, and hence both its prior and that for S, at all values. When adjusted for the ζ difference, the S priors derived using the revised data-variable assumptions are close to those derived using S20s' original assumptions. The remaining differences (of up to ±~5%) likely relate to the opposing effects of an increase in ΔT uncertainty, but a reduction in ζ uncertainty, under the revised assumptions.

S7. Effect of uniform-in-λ prior at very low S values

The reason why, where Process evidence is not used, the entire posterior probability will be located immediately above zero S is that in S-space a uniform-in-λ prior varies with S^{-2}, the integral of which tends to infinity as the lower integration limit for S approaches zero, while Historical and Paleoclimate likelihoods remain finite at $S = 0$ K. The ΔT terms which appear in all the non-Process model equations each have a non-zero probability of reaching zero, which corresponds to $\lambda = \infty$. Therefore, non-Process likelihoods and, when using a uniform-in-λ prior, non-Process posterior PDFs for λ, remain finite as $\lambda \to \infty$. That results in essentially all probability being located at arbitrarily high λ values, corresponding to near-zero S. All probability would be located symmetrically around zero S if negative S values were permitted. However, these distorting effects of a uniform-in-λ prior do not start to become noticeable until S falls to around 0.001 K.

Additional references

Andrews, T., Gregory, J.M. and Webb, M.J., 2015. The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. *Journal of Climate*, *28*(4), pp.1630-1648.

Armour, K. C., 2017: Energy budget constraints on climate sensitivity in light of inconstant climate feedbacks. Nat. Climate Change, 7, 331–335. https://doi.org/10.1038/nclimate3278

Braconnot, P., & Kageyama, M., 2015. Shortwave forcing and feedbacks in Last Glacial Maximum and mid-Holocene PMIP3 simulations. Philosophical Transactions of the Royal Society A, 373.

Brazzale, A.R., Davison, A.C. and Reid, N., 2007. Applied asymptotics: case studies in small-sample statistics (Vol. 23). Cambridge University Press.

Buizert et al., 2021. Antarctic surface temperature and elevation during the Last Glacial Maximum. Science https://doi.org/10.1126/science.abd2897

Caballero, R., Huber, M., 2013. State-dependent climate sensitivity in past warm climates and its implications for future climate projections. *Proceedings of the National Academy of Sciences*, *110*(35), pp.14162-14167. www.pnas.org/cgi/doi/10.1073/pnas.1303365110

Charlson, R. J., Schwartz, S. E., Hales, J. M., Cess, R. D., Coakley, J. A., Hansen, J. E., & Hofmann, D. J. (1992). Climate forcing by anthropogenic aerosols. Science, 255(5043), 423–430. https://doi.org/10.1126/science.255.5043.423

Colman, R.A. and Hanson, L.I., 2013. On atmospheric radiative feedbacks associated with climate variability and change. *Climate dynamics*, *40*(1), pp.475-492.

Colman, R. and Hanson, L., 2017. On the relative strength of radiative feedbacks under climate variability and change. *Climate Dynamics*, *49*(5), pp.2115-2129.

Dessler, A.E., 2013. Observations of climate feedbacks over 2000–10 and comparisons to climate models. *Journal of Climate*, *26*(1), pp.333-342.

Dong, Y., Armour, K.C., Proistosescu, C., Andrews, T., Battisti, D.S., Forster, P.M., Paynter, D., Smith, C.J. and Shiogama, H., 2021. Biased estimates of equilibrium climate sensitivity and transient climate response derived from historical CMIP6 simulations. *Geophysical Research Letters*, *48*(24), p.e2021GL095778.

Efron, B.*;* Hinkley, D.V., 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher Information. *Biometrika*. **65** (3): 457–487. https://doi.org/10.1093/biomet/65.3.457.

Flato, G., J. et al., 2013: Evaluation of Climate Models. 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., et al.(eds.)]. Cambridge University Press.

Friedrich, T. and Timmermann, A., 2020. Using Late Pleistocene sea surface temperature reconstructions to constrain future greenhouse warming. *Earth and Planetary Science Letters*, *530*, p.115911. https://doi.org/10.1016/j.epsl.2019.115911

Friedrich, T., Timmermann, A., Tigchelaar, M., Elison Timm, O., & Ganopolski, A., 2016. Nonlinear climate sensitivity and its implications for future greenhouse warming. Science Advance, 2(11), e1501923. https://doi.org/10.1126/sciadv.1501923

Golaz, J.C., Caldwell, P.M., Van Roekel, L.P., Petersen, M.R., Tang, Q., Wolfe, J.D., Abeshu, G., Anantharaj, V., Asay-Davis, X.S., Bader, D.C. and Baldwin, S.A., 2019. The DOE E3SM coupled model version 1: Overview and evaluation at standard resolution. *Journal of Advances in Modeling Earth Systems*, *11*(7), pp.2089-2129. https://doi.org/10.1029/2018MS001603

Hansen, J., Sato, M., Kharecha, P., Russell, G., Lea, D. W., & Siddall, M., 2007. Climate change and trace gases. Philosophical Transactions of the Royal Society A, 365, 1925.1954. https://doi.org/10.1098/rsta.2007.2052

Haywood, A.M., Hill, D.J., Dolan, A.M., Otto-Bliesner, B.L., Bragg, F., Chan, W.L., Chandler, M.A., Contoux, C., Dowsett, H.J., Jost, A. and Kamae, Y., 2013. Large-scale features of Pliocene climate: results from the Pliocene Model Intercomparison Project. *Climate of the Past*, *9*(1), pp.191-209.

Herbert, T. D., Peterson, L. C., Lawrence, K. T., & Liu, Z., 2010. Tropical ocean temperatures over the past 3.5 million years. Science, 328(5985), 1530–1534. https://doi.org/10.1126/science.1185435

Inglis, G.N. et al., 2020. Global mean surface temperature and climate sensitivity of the early Eocene Climatic Optimum (EECO), Paleocene–Eocene Thermal Maximum (PETM), and latest Paleocene. Climate of the Past, 16(5), pp.1953-1968. https://doi.org/10.5194/cp-16-1953-2020

Kretzschmar, J., M. Salzmann, J. Mülmenstädt, O. Boucher, and J. Quaas, 2017: Comment on ‘‘Rethinking the lower bound on aerosol radiative forcing.’’ J. Climate, 30, 6579–6584, https://doi.org/10.1175/JCLI-D-16-0668.1.

Lewis, N.and Curry, J. A., 2020, "Reply to “Comment on ‘The Impact of Recent Forcing and Ocean Heat Uptake Data on Estimates of Climate Sensitivity’”" Journal of Climate Vol. 33, No. 1, pp 397, 1520-0442

Lindzen, R. S., and Choi, Y.S., 2011. On the observational determination of climate sensitivity and its implications, *Asia-Pacific J. Atmos. Sci.*, **47**, 377–390.

Lindzen, R.S. and Choi, Y.S., 2021. The Iris Effect: A Review. *Asia-Pacific Journal of Atmospheric Sciences*, pp.1-10. https://doi.org/10.1007/s13143-021-00238-1

Malavelle, F. F., Haywood, J. M., Jones, A., Gettelman, A., Clarisse, L., Bauduin, S., et al., 2017. Strong constraints on aerosol-cloud interactions from volcanic eruptions. Nature, 546(7659), 485–491. https://doi.org/10.1038/nature22974

Mardia, K. V., J. T. Kent, and J. M. Bibby, 1979: Multivariate Analysis. Academic Press, 518 pp.

MARGO, 2009. Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum. Nature Geoscience, 2, 127.132.

Masson-Delmotte, V., et al., 2010. EPICA Dome C record of glacial and interglacial intensities. Quarterly Journal of the Royal Meteorological Society, 29, 113.128.

Mauritsen, T. and Stevens, B., 2015. Missing iris effect as a possible cause of muted hydrological change and high climate sensitivity in models. *Nat. Geosci.* **8**, 346–351. doi:10.1038/ngeo2414.

Meraner, K., Mauritsen, T. and Voigt, A., 2013. Robust increase in equilibrium climate sensitivity under global warming. *Geophysical Research Letters*, *40*(22), pp.5944-5948. doi:10.1002/2013GL058118

Mosegaard, K. and A. Tarantola, 2002: Probabilistic Approach to Inverse Problems. Int'l. Handbook of Earthquake and Engineering Seismology, Volume 81A: Lee, W. H. K. et al., Int'l Assoc. Seismol. & Phys. Earth's Interior, Committee on Education, Academic Press, 933pp.

Osman, M.B., Tierney, J.E., Zhu, J., Tardif, R., Hakim, G.J., King, J. and Poulsen, C.J., 2021. Globally resolved surface temperatures since the Last Glacial Maximum. *Nature*, *599*(7884), pp.239-244. https://www.nature.com/articles/s41586-021-03984-4

Paynter, D., Frölicher, T.L., Horowitz, L.W. and Silvers, L.G., 2018. Equilibrium climate sensitivity obtained from multimillennial runs of two GFDL climate models. *Journal of Geophysical Research: Atmospheres*, *123*(4), pp.1921-1941.

Proistosescu, C., Donohoe, A., Armour, K.C., Roe, G.H., Stuecker, M.F. and Bitz, C.M., 2018. Radiative feedbacks from stochastic variability in surface temperature and radiative imbalance. *Geophysical Research Letters*, *45*(10), pp.5082-5094.

Qu, X., Hall, A., Klein, S.A. and Caldwell, P.M., 2015. The strength of the tropical inversion and its response to climate change in 18 CMIP5 models. *Climate Dynamics*, *45*(1), pp.375-396.

Richardson, T. B., Forster, P. M., Smith, C. J., Maycock, A. C., Wood, T., Andrews, T., et al., 2019. Efficacy of climate forcings in PDRMIP models. Journal of Geophysical Research: Atmospheres, 124, 12,824–12,844. https://doi.org/10.1029/2019JD030581

Rohling, E. J., Medina-Elizalde, M., Shepherd, J. G., Siddall, M., & Stanford, J. D., 2012. Sea surface and high-latitude temperature sensitivity to radiative forcing of climate over several glacial cycles. Journal of Climate, 25(5), 1635.1656. https://doi.org/10.1175/ 2011JCLI4078.1

Schmidt, G. A., & Shindell, D. T. (2003). Atmospheric composition, radiative forcing, and climate change as a consequence of a massive methane release from gas hydrates. Paleoceanography, 18(1), 1004. https://doi.org/10.1029/2002PA000757

Schmittner, A., Urban, N. M., Shakun, J. D., Mahowald, N. M., Clark, P. U., Bartlein, P. J., et al., 2011. Climate sensitivity estimated from temperature reconstructions of the last glacial maximum. Science, 334(6061), 1385.1388. https://doi.org/10.1126/science. 1203513

Scott, R.C., Myers, T.A., Norris, J.R., Zelinka, M.D., Klein, S.A., Sun, M. and Doelling, D.R., 2020. Observed sensitivity of low-cloud radiative effects to meteorological perturbations over the global oceans. *Journal of Climate*, *33*(18), pp.7717-7734. https://doi.org/10.1175/JCLI-D-19-1028.1

Seifert, A., Heus, T., Pincus, R. and Stevens, B., 2015. Large-eddy simulation of the transient and near-equilibrium behavior of precipitating shallow convection. *Journal of Advances in Modeling Earth Systems*, *7*(4), pp.1918-1937.

Shakun, J.D., Clark, P.U., He, F., Marcott, S.A., Mix, A.C., Liu, Z., Otto-Bliesner, B., Schmittner, A. and Bard, E., 2012. Global warming preceded by increasing carbon dioxide concentrations during the last deglaciation. *Nature*, *484*(7392), pp.49-54. Figure 2a. https://doi.org/10.1038/nature10915

Smith, C.J., Kramer, R.J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., Boucher, O., Dufresne, J.L., Nabat, P., Michou, M. and Yukimoto, S., 2020. Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, *20*(16), pp.9591-9618. https://doi.org/10.5194/acp-20-9591-2020

Snyder, C. W., 2016. Evolution of global temperature over the past two million years. Nature, 538, 226.228.

Stevens, B., 2015. Rethinking the lower bound on aerosol radiative forcing. J. Climate, 28, 4794–4819, doi:10.1175/JCLI-D-14-00656.1.

Stevens, B., Fiedler, S., 2017. Reply to "comment on 'rethinking the lower bound on aerosol radiative forcing'". Journal of Climate, S30(16), 6585–6589. https://doi.org/10.1175/JCLI-D-17-0034.1

Tang, T., Luo, J.J., Peng, K., Qi, L. and Tang, S.,2021. Over-projected Pacific warming and extreme El Niño frequency due to CMIP5 common biases. *National Science Review*. https://doi.org/10.1093/nsr/nwab056

Tierney, J. E., Haywood, A. M., Feng, R., Bhattacharya, T., & Otto-Bliesner, B. L., 2019. Pliocene warmth consistent with greenhouse gas forcing. Geophysical Research Letters, 46, 9136–9144. https://doi.org/10.1029/2019GL083802

Titchner, H. A., and N. A. Rayner, 2014. The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. J. Geophys. Res. Atmos., 119, 2864–2889, doi:10.1002/2013JD020316.

Vaccaro, A., Emile-Geay, J., Guillot, D., Verna, R., Morice, C., Kennedy, J. and Rajaratnam, B., 2021a. Climate Field Completion via Markov Random Fields: Application to the HadCRUT4. 6 Temperature Dataset. *Journal of Climate*, *34*(10), pp.4169-4188.

Vaccaro, A., Emile-Geay, J., Guillot, D., Verna, R., Morice, C., Kennedy, J. and Rajaratnam, B., 2021b. GraphEM-infilled HadCRUT4.6.0.0 January 1850 - December 2017 [Data set]. Zenodo. https://doi.org/10.5281/zenodo.4601616 9, accessed 3 May 2021.

Vial, J., Dufresne, J.L. and Bony, S., 2013. On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates. *Climate Dynamics*, *41*(11-12), pp.3339-3362.

Wang, C., Soden, B.J., Yang, W. and Vecchi, G.A., 2021. Compensation Between Cloud Feedback and Aerosol-Cloud Interaction in CMIP6 Models. *Geophysical Research Letters*, *48*(4), p.e2020GL091024. https://doi.org/10.1029/2020GL091024

Webb, M., 2020. Code and Data for WCRP Climate Sensitivity Assessment. Zenodo. https://doi.org/10.5281/zenodo.3945276

Williams, I.N., Pierrehumbert, R.T., 2017. Observational evidence against strongly stabilizing tropical cloud feedbacks. Geophys. Res. Lett. 44, 1503–1510

Zelinka, M.D., Zhou, C. and Klein, S.A., 2016. Insights from a refined decomposition of cloud feedbacks. *Geophysical Research Letters*, *43*(17), pp.9259-9269. https://doi.org/10.1002/2016GL069917

Zelinka, M.D., Klein, S.A., Qin, Y. and Myers, T.A., 2022. Evaluating climate models' cloud feedbacks against expert judgement. *Journal of Geophysical Research: Atmospheres*, p.e2021JD035198.

Zhou, C., Zelinka, M.D., Dessler, A.E. and Klein, S.A., 2015. The relationship between interannual and long-term cloud feedbacks. *Geophysical Research Letters*, *42*(23), pp.10-463.

Zhu, J., Poulsen, C.J. and Tierney, J.E., 2019. Simulation of Eocene extreme warmth and high climate sensitivity through cloud feedbacks. *Science advances*, *5*(9), p.eaax1874. https://doi.org/10.1126/sciadv.aax1874

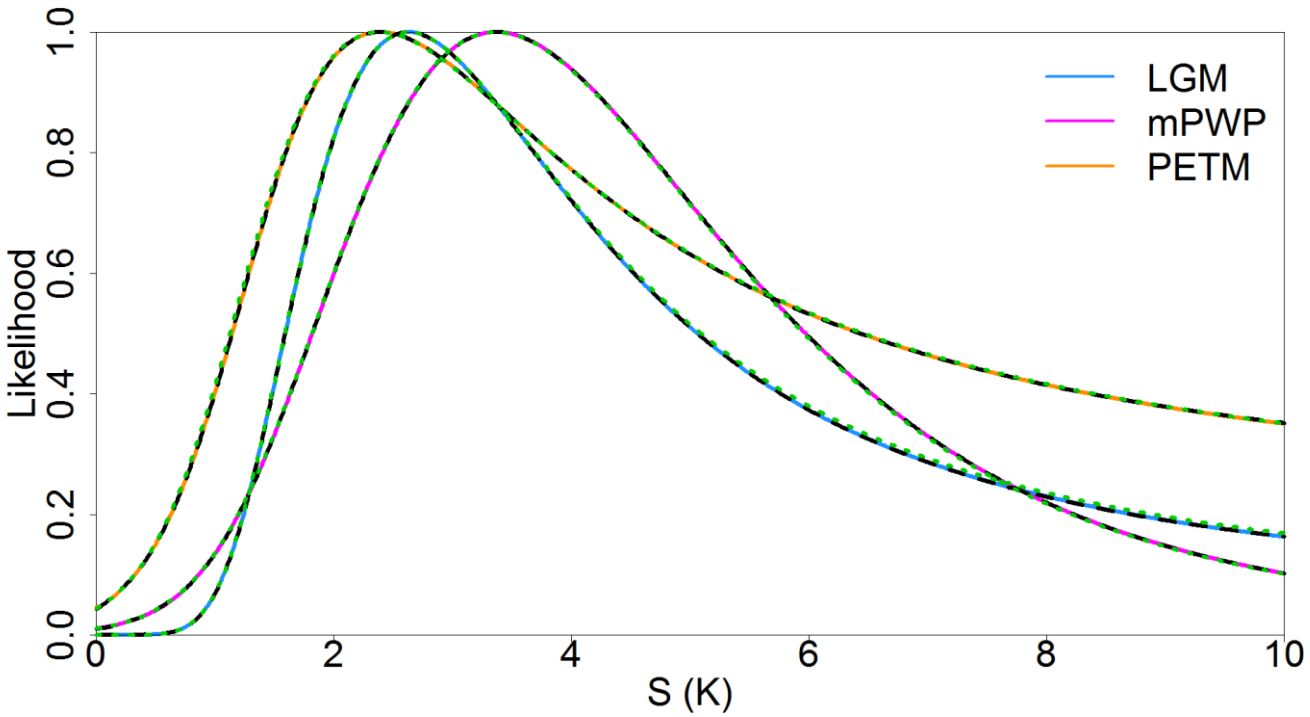

Figure S1 As for Figure 2(a) but showing likelihoods from the profile likelihood method (green dotted lines) and the doubled data method (black dashed lines) superimposed on integrated likelihoods, instead of comparisons with S20's likelihoods.

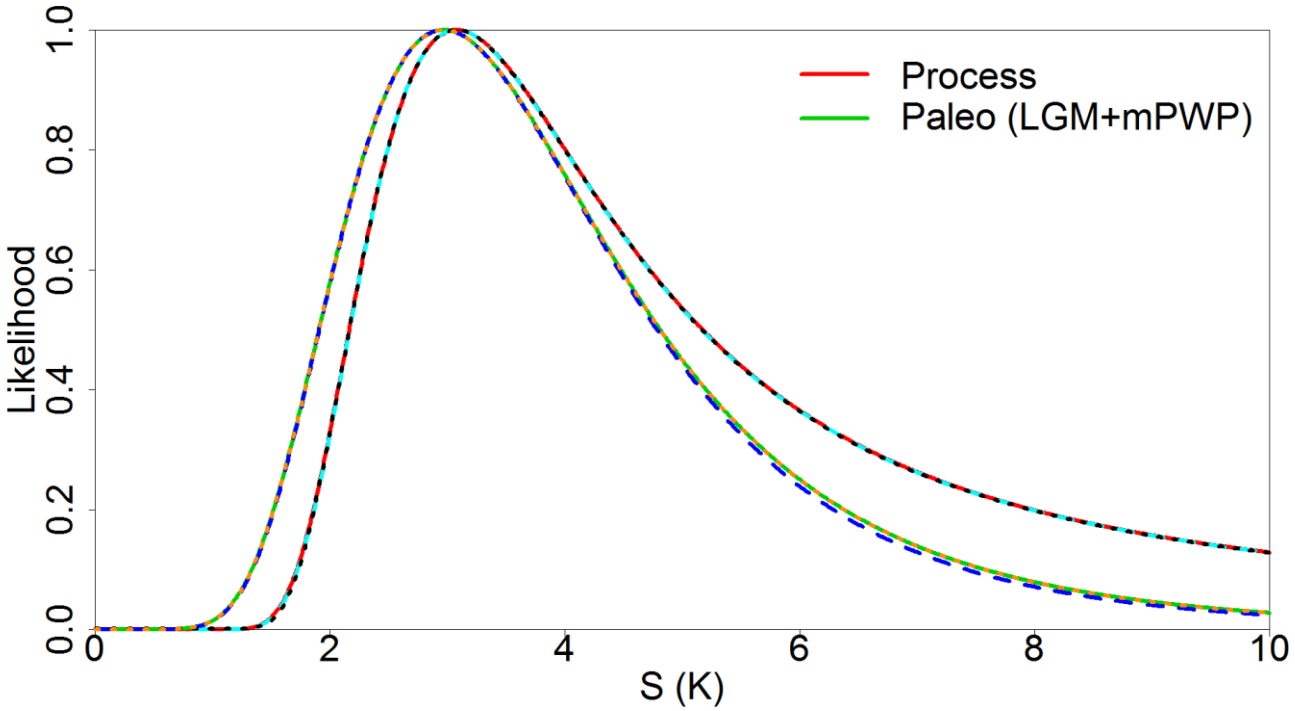

Figure S2 As for Figure 2(b) but showing likelihoods from the profile likelihood method (dashed lines: cyan for Process, blue for Paleoclimate) and the doubled data method (dotted lines: black for Process, orange for Paleoclimate) superimposed on integrated likelihoods, instead of comparisons with S20's likelihoods.

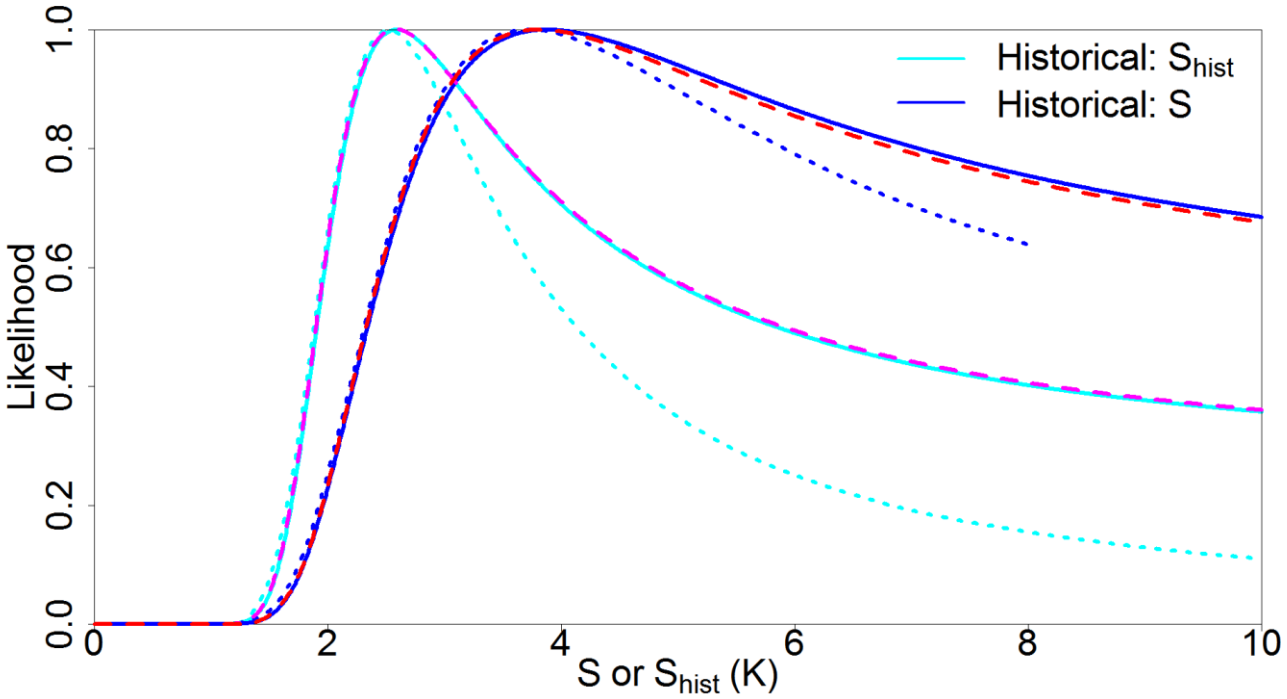

Figure S3 As for Figure 2(c) but showing also likelihoods from the profile likelihood method (dashed lines: red for *S*, magenta for S_{hist}) superimposed on integrated likelihoods, as well as comparisons with S20's likelihoods (dotted, colors as per legend).

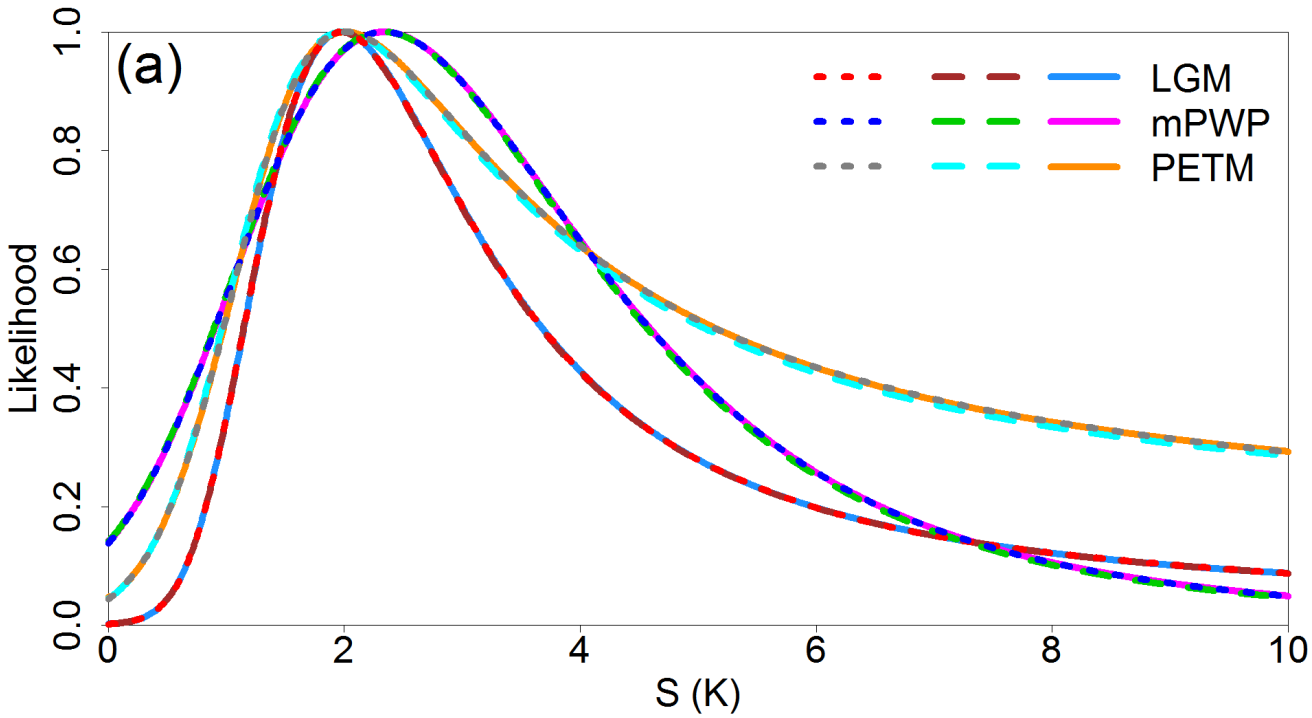

Figure S4 As for Figure 5(a) but showing also likelihoods from the profile likelihood method (dashed lines) and the doubled data method (dotted lines) instead of comparisons with likelihoods on S20's data-variable assumptions.

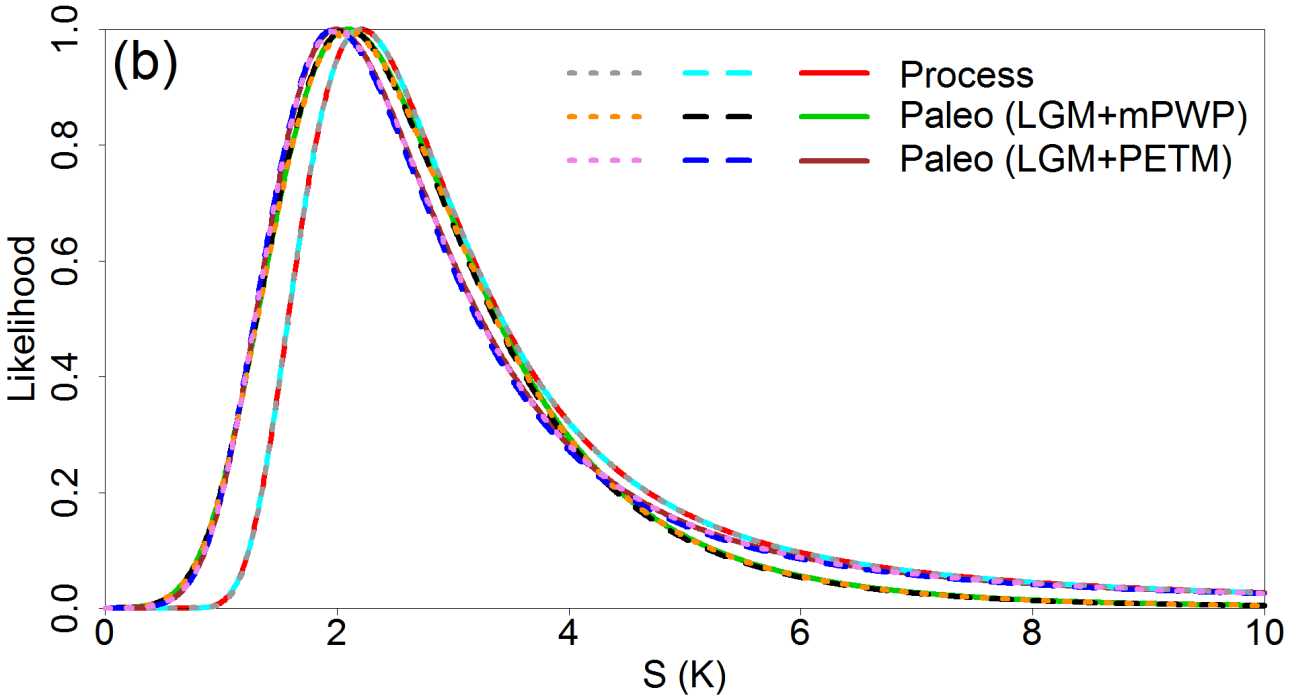

Figure S5 As for Figure 5(b) but showing also likelihoods from the profile likelihood method (dashed lines) and the doubled data method (dotted lines) instead of comparisons with likelihoods on S20's data-variable assumptions.

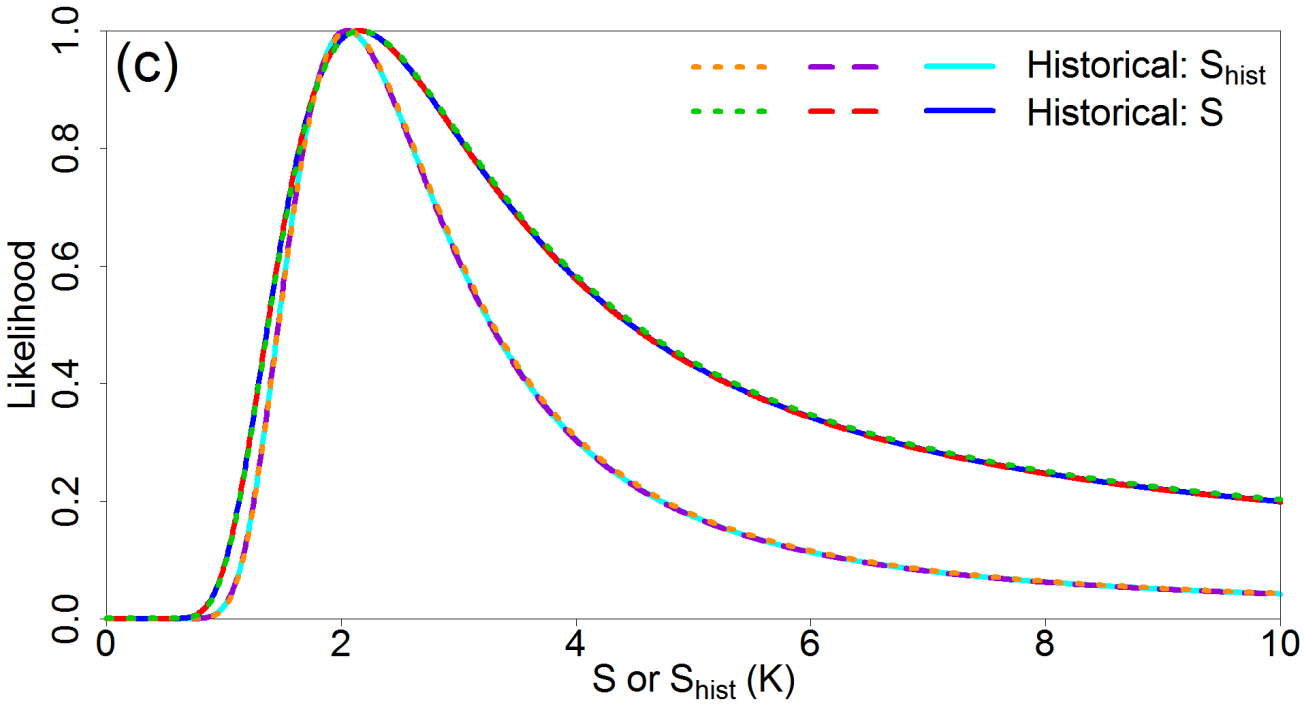

Figure S6 As for Figure 5(c) but showing also likelihoods from the profile likelihood method (dashed lines) and the doubled data method (dotted lines) instead of comparisons with likelihoods on S20's data-variable assumptions.

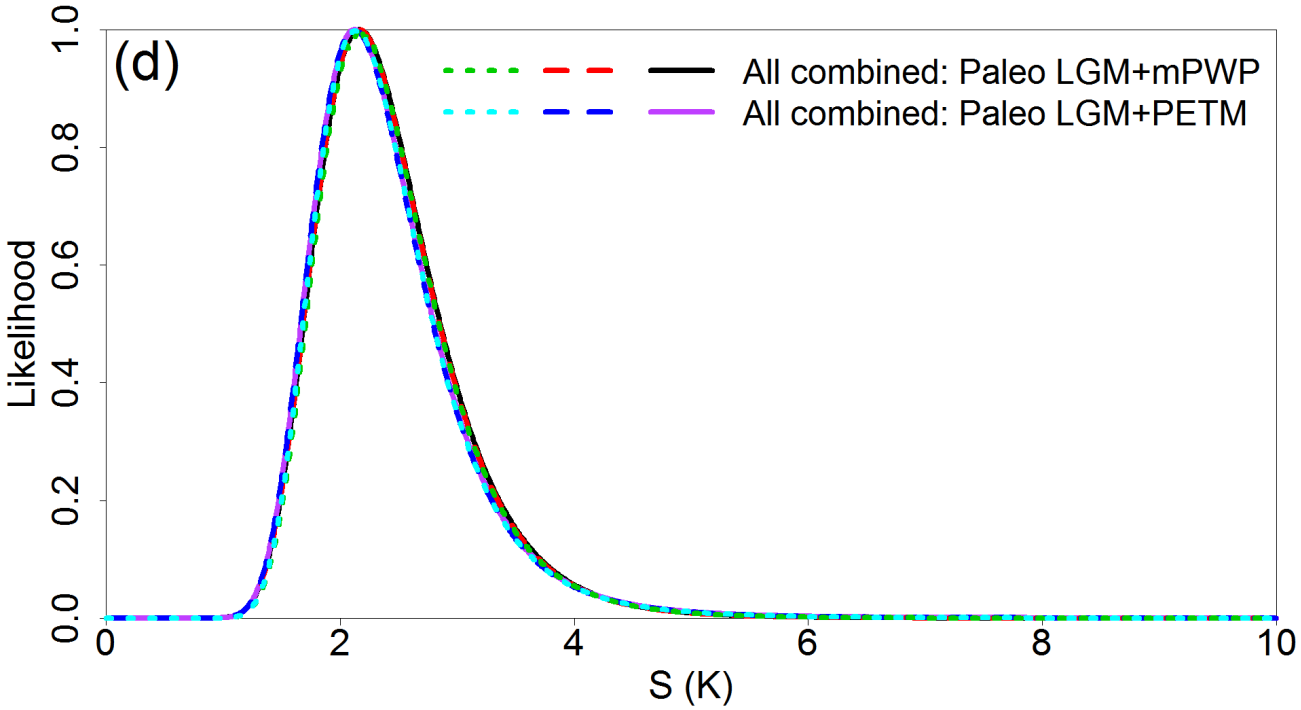

Figure S7 As for Figure 5(d) but showing also likelihoods from the profile likelihood method (dashed lines) and the doubled data method (dotted lines) instead of comparisons with likelihoods on S20's data-variable assumptions.

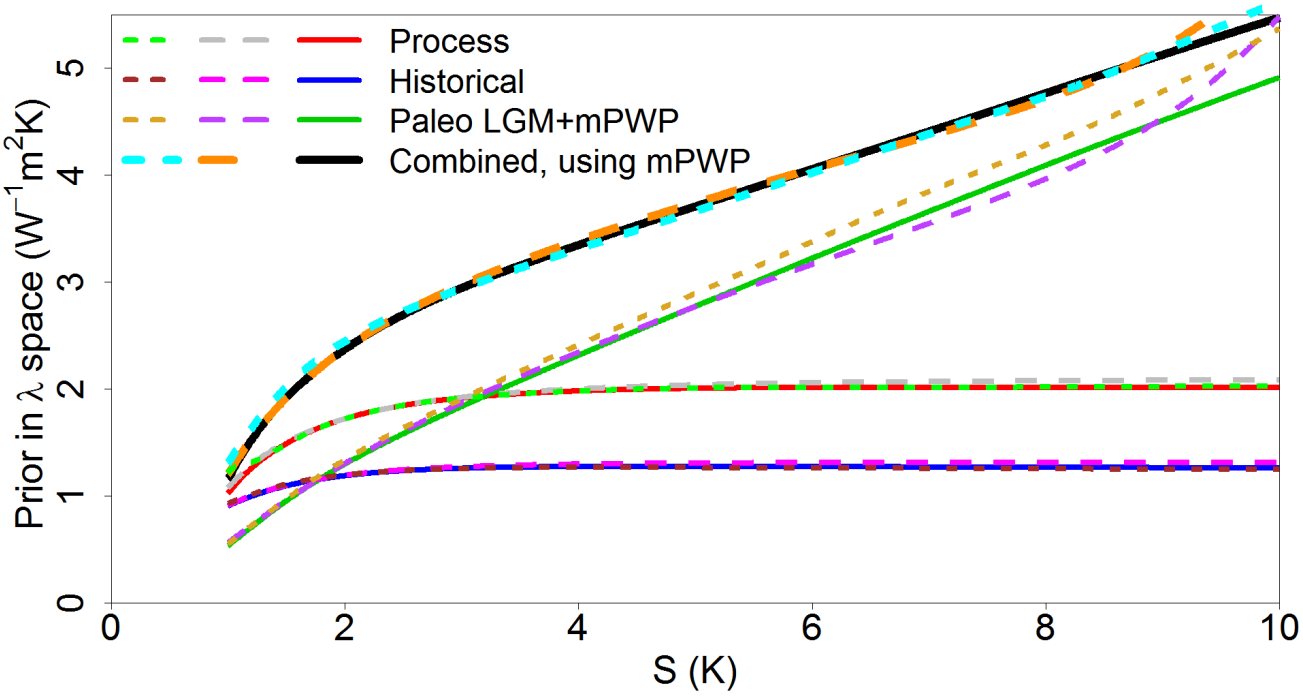

Figure S8 As for Figure 6(b), for clarity without the line for Paleoclimate LGM+PETM, but showing also transformed priors from the profile likelihood method (dashed lines) and the doubled data method (dotted lines).

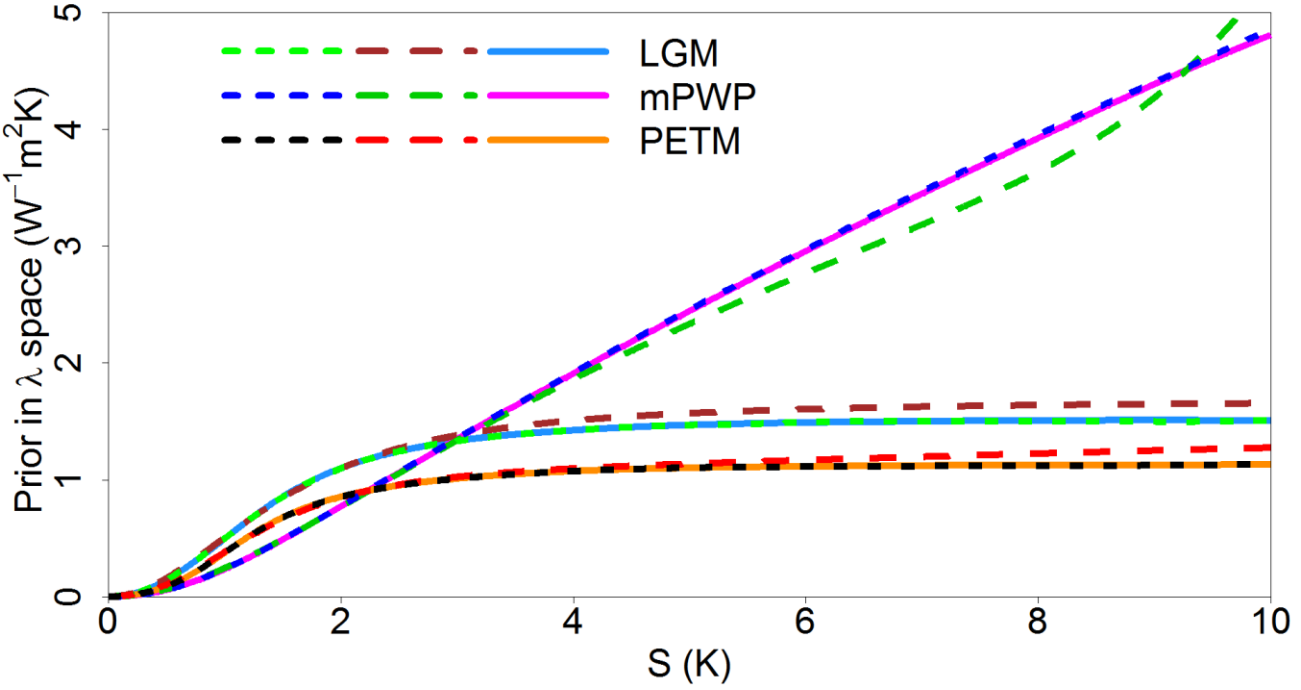

Figure S9 As for Figure 6(c), but showing also transformed priors from the profile likelihood method (dashed lines) and the doubled data method (dotted lines).

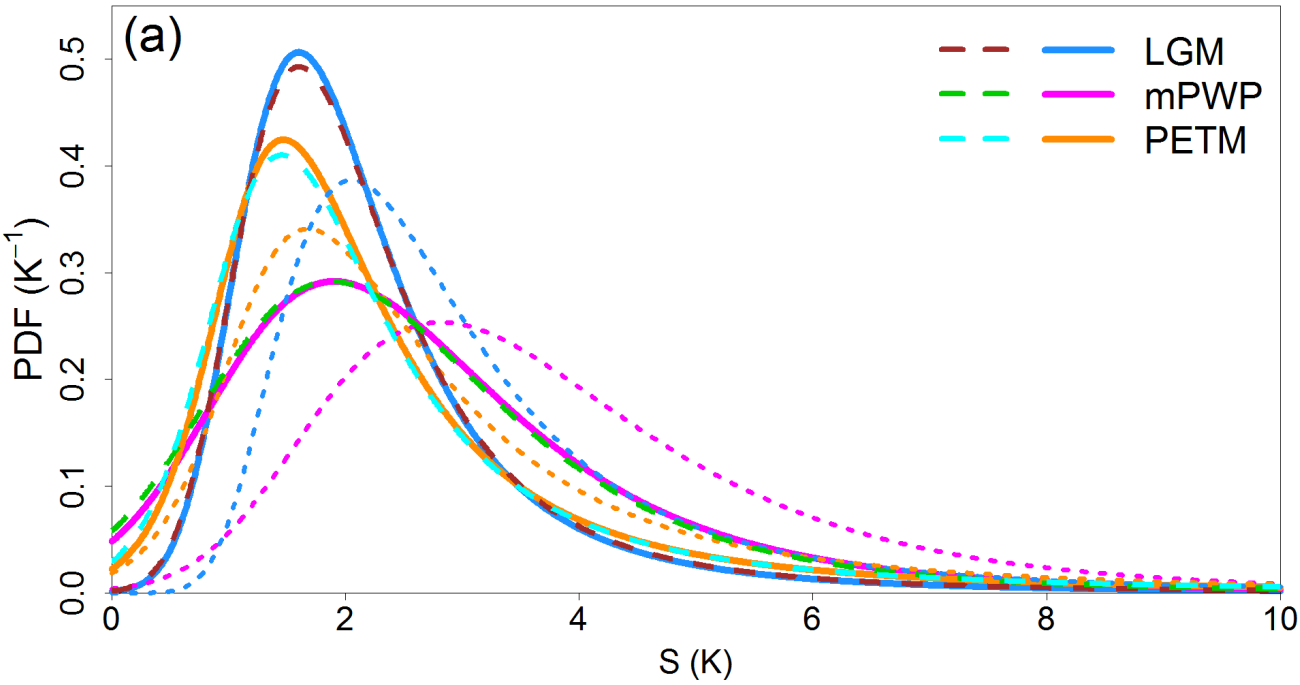

Figure S10 As for Figure 7(a), but showing also PDFs computed using the profile likelihood method and its data-space movement prior (dashed lines). No lines are shown using the doubled data method since for individual lines of evidence it uses the same sampling-derived PDF as does the primary integrated likelihood method. Dotted lines show comparatives based on S20's assumptions, as in Figure 9(a).

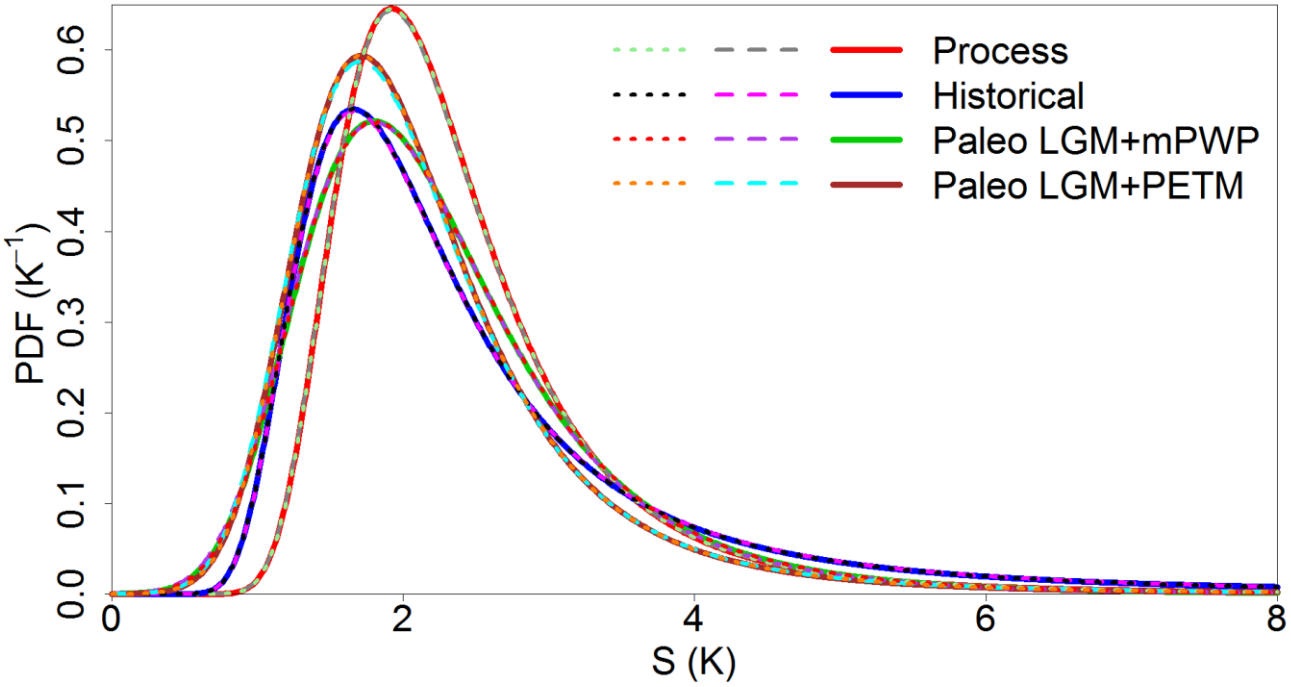

Figure S11 As for Figure 7(b), but showing also PDFs computed using the profile likelihood method and its data-space movement prior (dashed lines) and the data doubling method (dotted colored lines).

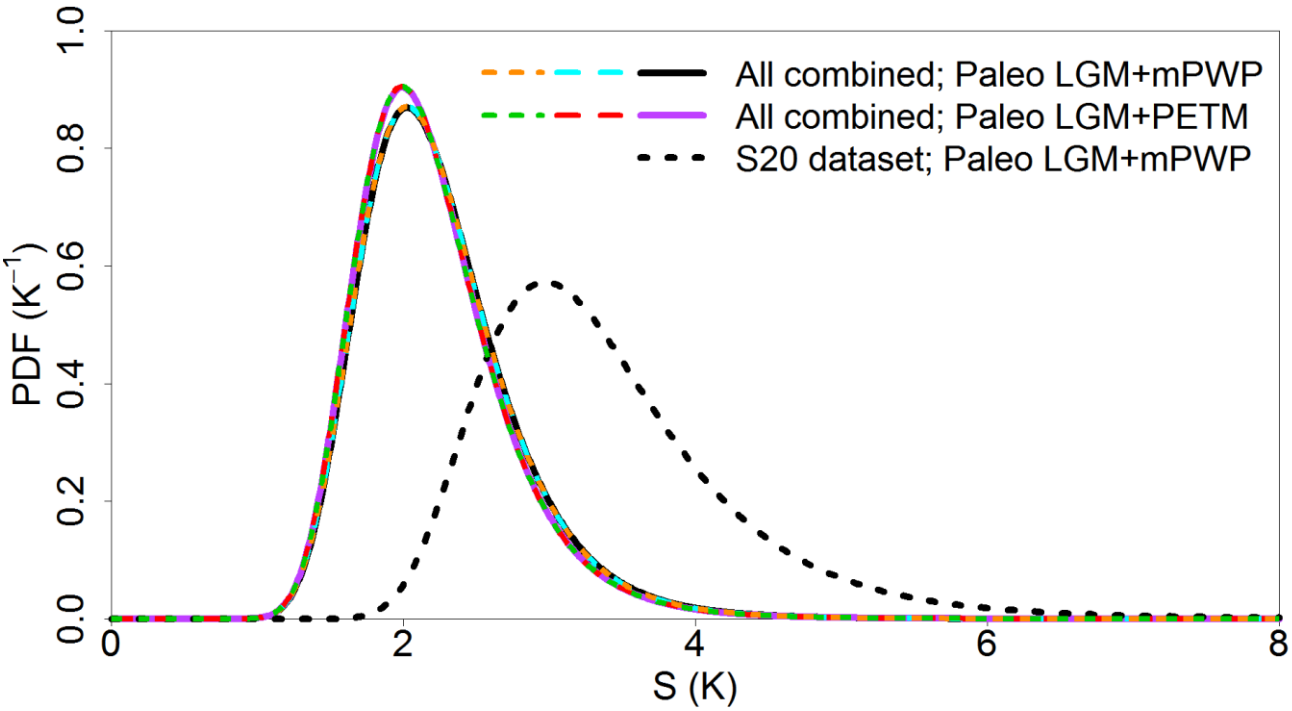

Figure S12 As for Figure 7(c), but showing also PDFs computed using the profile likelihood method and its data-space movement prior (dashed lines) and the data doubling method (dotted colored lines).

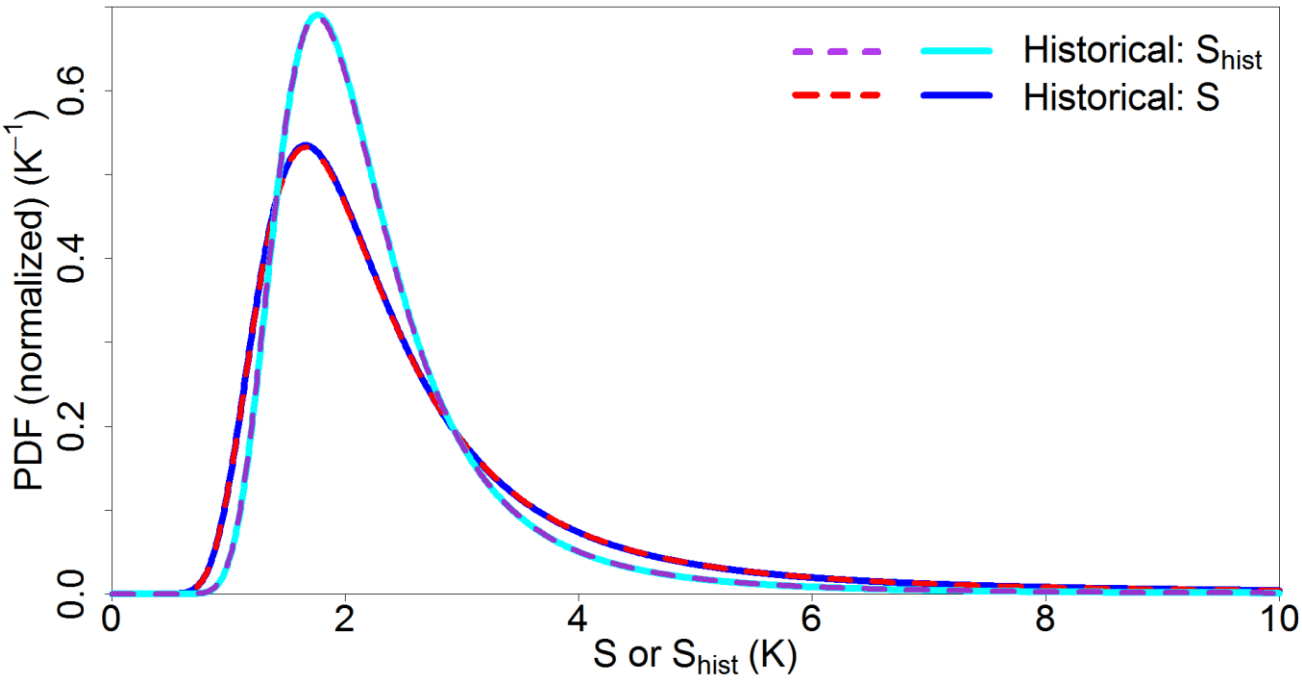

Figure S13 As for Figure 7(d), but showing also PDFs computed using the profile likelihood method and its data-space movement prior (dashed lines). As the latter method cannot derive probability lying outside the *S* range over which computations are carried out, these PDFs have been normalized over 0-20 K, unlike in Figure 9(d). No lines are shown using the doubled data method since for individual lines of evidence it uses the same sampling-derived PDF as does the primary integrated likelihood method.