Earth Syst. Dynam., 16, 317–332, 2025
https://doi.org/10.5194/esd-16-317-2025
© Author(s) 2025. This work is distributed under
the Creative Commons Attribution 4.0 License.

Earth System
Dynamics
Open Access

Towards robust community assessments of the Earth's climate sensitivity

Kate Marvel[1] **and Mark Webb**[2]

[1]NASA Goddard Institute for Space Studies, New York, NY, USA
[2]Met Office Hadley Centre, Exeter, UK

Correspondence: Kate Marvel (kate.marvel@nasa.gov)

Received: 9 February 2024 – Discussion started: 22 February 2024
Accepted: 2 October 2024 – Published: 26 February 2025

Abstract. The eventual planetary warming in response to elevated atmospheric carbon dioxide concentrations is not precisely known. The uncertainty in climate sensitivity (S) primarily results from uncertainties in net physical climate feedback, usually denoted as λ. Multiple lines of evidence can constrain this feedback parameter: proxy-based and model evidence from past equilibrium climates; process-based understanding of the physics underlying changes; and recent observations of temperature change, top-of-the-atmosphere energy imbalance, and ocean heat content. However, despite recent advances in combining these lines of evidence, the estimated range of S remains large. Here, using a Bayesian framework, we discuss three sources of uncertainty – uncertainty in the evidence, structural uncertainty in the model used to interpret this evidence, and differing prior knowledge and/or beliefs – and show how these affect the conclusions we may draw from a single line of evidence. We then propose strategies to combine multiple lines of evidence. We end with three recommendations. First, we suggest that a Bayesian random-effects meta-analysis be used to estimate the evidence and its uncertainty from the published literature. Second, we advocate that the organizers of future assessments clearly specify an interpretive model or a group of candidate models and, in the latter case, use Bayesian model averaging to more heavily weight models that best fit the evidence. Third, we recommend that expert judgment be incorporated via solicitations of priors on model parameters.

1 Introduction

When radiative forcing (ΔF) is applied to the climate system, it induces a radiative imbalance (ΔN) at the top of the atmosphere and a response (ΔR) of the system itself. To first order, $\Delta R = \lambda \Delta T$, where ΔT is the change in global mean surface temperature. The feedback parameter (λ) thus measures the additional radiative flux density exported to space per unit of temperature change. On sufficiently long timescales, the climate comes into equilibrium ($\Delta N = 0$), internal variability is negligible, and we can write a simple energy balance model (denoted M_0) for the climate system. M_0 is expressed as follows:

$$\Delta N = \Delta F + \lambda \Delta T. \tag{1}$$

In the special case where radiative forcing results from a doubling of atmospheric CO_2 relative to its preindustrial concentration of 280 ppm ($\Delta F = F_{2\times CO_2}$), the resulting temperature change defines the equilibrium climate sensitivity (S), expressed as

$$S \equiv -\frac{F_{2\times CO_2}}{\lambda}. \tag{2}$$

S is often used as a metric to quantify expected warming in response to radiative forcing but has remained stubbornly uncertain, even as climate models have improved and become more sophisticated. A 2020 community assessment (Sherwood et al., 2020, hereafter referred to as S20) reduced this range using multiple lines of evidence, but the recent Intergovernmental Panel on Climate Change (IPCC) report (Forster, 2021) assessed only "medium confidence" in the upper bound. Is it possible to further narrow the estimated range of S, and can we increase our confidence in this result?

Published by Copernicus Publications on behalf of the European Geosciences Union.

S is determined by the net feedback (λ) at equilibrium and in response to doubled CO_2. While this is unobservable in the current system, in which CO_2 has not yet doubled and is out of equilibrium, several lines of evidence exist that might constrain λ. We have some process-based understanding of individual feedback processes and their correlations, derived from observations and basic physics. We also have evidence from the planet itself, which has been steadily warming in response to net anthropogenic forcing, including emissions of not just CO_2 but also other greenhouse gases and aerosols. Finally, we have proxies that provide evidence about equilibrium climates of the past. S20 attempted to synthesize these three lines of evidence, arriving at constraints on climate sensitivity that narrowed the former range.

In S20, the spread in S arose from reported and assessed uncertainty in historical observations and paleoclimate reconstructions, expert judgment about the uncertainty in physical processes, and the use of different priors on λ and/or S. The IPCC's Sixth Assessment Report (AR6) assessed confidence in the range of S based on support from individual lines of evidence, and the medium confidence assessed was in large part due to the fact that not all lines of evidence supported the same upper bound. By contrast, S20 sought to provide a robust estimate by combining lines of evidence in a coherent Bayesian framework. However, S20 used baseline priors and estimates of the evidence and investigated the impact of alternate choices as sensitivity tests, rather than attempting to combine multiple priors, estimates, and expert judgments into a single posterior probability distribution. In both the IPCC's AR6 and S20, as in almost all previous assessments, the means by which disagreements among experts were resolved or handled were not necessarily made transparent. This paper presents some lessons learned by two authors of S20 and attempts to chart a way forward.

Our goal is to understand where unavoidable subjective decisions enter into the analysis and to present a framework for systematically and fairly incorporating the subjective judgments of multiple experts. Ultimately, we seek to create a framework in which expert judgment is incorporated in the form of clearly specified priors.

The paper is organized as follows. In Sect. 2, we review the basic Bayesian analysis framework. Sections 3, 4, and 5 discuss uncertainty in the evidence, structural uncertainty, and prior uncertainty, respectively. In these sections, we use a single line of evidence – paleoclimate estimates from the Last Glacial Maximum – to illustrate how these sources of uncertainty shape estimates of climate feedbacks and sensitivity. In Sect. 6, we show how these sources of uncertainty affect constraints derived from multiple lines of evidence. In Sect. 7, we propose a new method for combining multiple published studies and multiple models, which may be used in the future to arrive at a robust community assessment of climate sensitivity. Finally, we discuss possible generalizations and extensions.

2 Analysis framework

Bayes' theorem can be written as

$$P(\Theta|Y,M)=\frac{P(Y|\Theta,M)P(\Theta|M)}{P(Y|M)}. \tag{3}$$

Here, we will define these terms as they apply to the problem of estimating climate sensitivity.

Evidence. The evidence (Y) used to constrain climate sensitivity consists of the global mean temperature change (ΔT) in response to forcing (ΔF), as well as, in non-equilibrium states, the net energy imbalance (ΔN). We have estimates of these quantities for the historical period (derived from observations and models) and for past climate states (derived from paleoclimate proxies and models); therefore, Y consists of multiple lines ($Y_1 \ldots Y_n$). For example, S20 used process-based understanding of underlying physics, recent observations, and proxy-based reconstructions of past climates to assess S.

Model. The model (M) codifies how we interpret the evidence (Y). It specifies the set of parameters (Θ) whose posterior distributions we estimate. For example, in the simple energy balance model, denoted by M_0, there is only one parameter, $\Theta=\lambda$. The model determines the *likelihood* $P(Y|\Theta,M)$ of observing the data given a particular set of parameter values (Θ). We discuss methods for calculating this likelihood in Sect. 3.1.

Prior. The *prior* probability distribution $P(\Theta|M)$ reflects prior beliefs or knowledge about the set of model parameters (Θ). For example, in the simple model M_0, the S20 community assessment adopted a uniform prior on λ as a baseline choice, choosing not to rule out net positive feedbacks (and, therefore, an unstable climate) a priori. Both the geological evidence and the process understanding presented in Sect. 3 of S20 effectively rule out both positive and extremely negative feedbacks, and thus an alternate prior reflecting this physical knowledge might be a normal distribution, denoted $N(\mu,\sigma)$, with a mean of $\mu=-1.30$ and a standard deviation of $\sigma=0.44$.

This framework allows us to use our prior understanding of the parameter values to calculate the posterior probability $P(\Theta|Y,M)$ of the model parameters given the evidence. This posterior can be updated as new evidence becomes available.

Bayesian statistics are both praised and criticized for their inherent subjectivity (see, e.g., Gelman et al., 1995). But *all* statistical analyses depend on prior knowledge and interpretive models, whether they are implicit or explicit. The Bayesian framework merely makes clear where unavoidable subjective decisions enter the analysis.

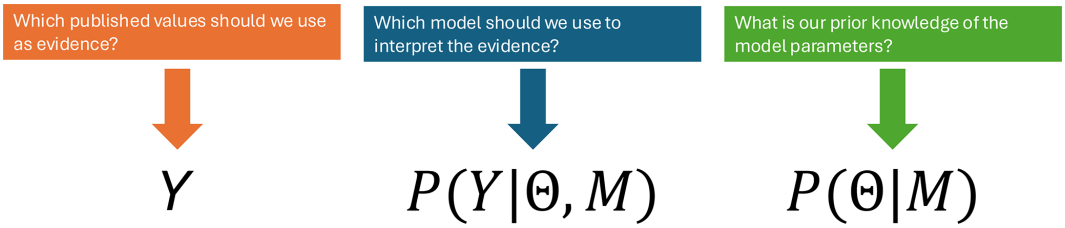
Which published values should we use as evidence?
Which model should we use to interpret the evidence?
What is our prior knowledge of the model parameters?
Y
$P(Y|\Theta, M)$
$P(\Theta|M)$

Figure 1. Schematic of unavoidable subjective decisions in an analysis of climate feedbacks.

Figure 1 summarizes the decisions that must be made in any Bayesian analysis of climate feedbacks. First, the analyst must decide what constitutes evidence. This requires an assessment of the literature evaluating ΔT, ΔF, and ΔN for each line of evidence. Second, the analyst must specify a model (and its set of parameters, Θ) in order to interpret that evidence. For example, the model M_0 assumes that the feedback parameter is time- and state-independent, and thus estimating the parameter from the past provides a reliable guide to the hypothetical future under doubled CO_2. Finally, the analyst must clearly specify their priors on the model parameters.

In the following sections, we show how different reasonable choices about evidence, models, and priors can lead to very different posterior distributions for λ (and hence for climate sensitivity, S) given a single line of evidence.

3 Evidence uncertainty

The strongest constraints on equilibrium climate sensitivity in S20 were derived from paleoclimate evidence, with the closest equilibrium climate to that of the present being the Last Glacial Maximum (LGM), approximately 21 000 years ago. Reconstructions (Annan and Hargreaves, 2013; Bereiter et al., 2018; Friedrich et al., 2016; Holden et al., 2009; von Deimling et al., 2006; Shakun et al., 2012; Snyder, 2016) or model-based estimates (Braconnot et al., 2012; Kageyama et al., 2021) of the global mean temperature change (ΔT) and radiative forcing (ΔF) have been used to calculate the global mean feedback (λ) inferred from this period. Neither of these observed quantities is precisely known. For example, multiple seemingly incompatible estimates of the LGM global mean cooling (ΔT) are available in the published literature (Annan and Hargreaves, 2013; Holden et al., 2009; Shakun et al., 2012; von Deimling et al., 2006; Friedrich et al., 2016; Hansen et al., 2023; Annan et al., 2022; Bereiter et al., 2018; Snyder, 2016; Kageyama et al., 2021). These estimates are derived from climate models participating in the Paleoclimate Modelling Intercomparison Project (PMIP; Kageyama et al., 2021), as well as from combinations of models and various proxies, and are often in conflict with one another.

We will illustrate the impact of this uncertainty by comparing the evidence used in two recent studies. S20 used expert judgment applied to a literature review to estimate $\Delta T = -5$ K, with a 95 % confidence interval of -3.0 to -7.0 K. However, a contemporaneous study using a new temperature reconstruction (Tierney et al., 2020, hereafter referred to as T20) estimated both colder values (with a mean of -6.1 K) and less uncertain values (with a 95 % highest posterior density interval of -6.5 to -5.7 K) for LGM cooling. We note that the two studies are not exactly comparable: S20 represents a community assessment of evidence that took into account a broad range of evidence and uncertainties, whereas T20 was a single study. The temperature estimates in T20 may also be cold biased and overconfident due to the reliance on a prior derived from a single climate model (Annan et al., 2022). However, in order to illustrate evidence uncertainty, we here treat S20 and T20 as different reasonable estimates of ΔT and ΔF over the LGM. We discuss methods for incorporating estimates, such as those from T20, into expert assessments in Sect. 7.1.

The two studies, S20 and T20, also differ in their estimates of the radiative forcing that led to this temperature change. Both agree that it was colder 21 000 years ago because a change in orbital forcing, while negligible with respect to the global mean, led to the development of large, reflective ice sheets in the Northern Hemisphere and lower levels of atmospheric greenhouse gases. The forcings associated with orbital changes (Kageyama et al., 2021) and CO_2 (Siegenthaler et al., 2005) are relatively well constrained; the forcings from other well-mixed greenhouse gases (Loulergue et al., 2008) and ice sheets are less so, but they are still informed by proxy and model evidence. The forcings from dust (Mahowald et al., 2006; Albani and Mahowald, 2019), other aerosols, and vegetation (Köhler et al., 2010) are highly uncertain. While S20 estimated total radiative forcing at the LGM to be $N(-8.43, 2)\,\mathrm{W\,m^{-2}}$, T20 uses a best estimate of $-46.8\,\mathrm{W\,m^{-2}}$, with a 95 % confidence interval of -9.6 to $-5.2\,\mathrm{W\,m^{-2}}$.

The contour lines in Fig. 2a show the joint probability distribution $\rho(\Delta T, \Delta F)$ (assuming uncorrelated errors) as reported by S20 (black) and T20 (red). Rather than comprising exact measurements of the temperature change and radiative forcing, our evidence (Y) consists of estimates of the joint probability density $\rho(\Delta T, \Delta F)$.

3.1 Calculating the likelihood

The likelihood of observing this probability density for any given value of the feedback parameter (λ) is determined by the model, which dictates the relationship between λ, ΔT, and ΔF. For example, the simple energy balance model M_0 constrains all possible pairs of ΔT and ΔF to lie along a line with a slope of $-\lambda$. Intuitively, the value of $-\lambda$ that maximizes the likelihood is the slope of the line that passes through the greatest probability density. These maximum likelihood estimates are indicated by straight lines in Fig. 2a.

We therefore define the likelihood of $\rho(\Delta T, \Delta F)$ for any λ value as the probability mass along the curve (C), as described by the energy balance model with a fixed λ value. This definition is written as follows:

$$P(Y|\lambda) \propto \int_C \rho(\Delta T, \Delta F)\mathrm{d}s,$$

where C is the curve defined by $\Delta F + \lambda\Delta T = 0$. If the joint evidence is a multivariate normal distribution (as it is in S20), this leads to an exact analytic expression for $P(Y|\lambda)$ (Appendix A). Otherwise, the integral can be computed numerically. The resulting likelihood functions are indicated by thick lines in Fig. 2b.

3.2 Climate sensitivity estimates depend on the evidence

Clearly, the constraints placed on the climate feedback by the Last Glacial Maximum depend on our estimates of the temperature difference and radiative forcing that caused it. Using S20 evidence, this energy balance model, and a uniform prior ($P(\lambda) = U(-10, 10)$), we find that the most likely value of the feedback parameter is $\lambda = -1.7\,\mathrm{W\,m^{-2}\,K^{-1}}$ (thick black line in Fig. 2b), with a 5 %–95 % range of -3.37 to $-1.09\,\mathrm{W\,m^{-2}\,K^{-1}}$. Using T20 evidence, the most likely value is $\lambda = -1.1\,\mathrm{W\,m^{-2}\,K^{-1}}$ (thick red line in Fig. 2b), where the 5 %–95 % range is -1.49 to $-0.87\,\mathrm{W\,m^{-2}\,K^{-1}}$.

For simplicity, here we calculate the likelihood $P(Y|\lambda)$ and use the resulting posterior, $P(\lambda|Y) \propto P(Y|\lambda)P(\lambda)$, to calculate S (Appendix B). This neglects the small correlation between ΔF and the forcing with doubled CO_2, but this simplification does not substantially affect our results (Appendix C).

Using S20 evidence from the LGM, we find a 5 %–95 % range of 1.17 to 3.69 K for the climate sensitivity (S), assuming, as in S20, that $F_{2\times CO_2} \sim N(4.0, 0.3)$. Using T20 evidence, the 5 %–95 % range for S is 2.61 to 4.72 K.

4 Structural uncertainty

Thus far, we have relied on the simple energy balance model to interpret the LGM evidence. However, many recent studies (e.g., Rohling et al., 2018; Stap et al., 2019; Friedrich et al., 2016; Renoult et al., 2023) have suggested that M_0 might not be appropriate for past climates due to the dependence of the feedbacks on the background climate state. If the relationship between temperature change and radiative forcing is nonlinear, then the feedbacks in a past cold climate should not be treated as identical to those in a future warm climate. To model this background temperature dependence, we might use an alternate model that includes a second-order term in the radiative response. This model (M_α) is given by

$$0 = \Delta F + \lambda\Delta T + \frac{\alpha}{2}\Delta T^2, \tag{4}$$

where $\alpha = \partial\lambda/\partial(\Delta T)$ is an additional parameter reflecting the background state dependence (Sellers, 1969; Caballero and Huber, 2013; Budyko, 1969; Sherwood et al., 2020). Intuitively, nonzero values of α change the relationship between the paleoclimate evidence and the feedback parameter (λ). This, in turn, makes the evidence more or less likely given a value of λ. For example, if $\alpha = +0.1$ (which translates to a change in feedback of $-0.5\,\mathrm{W\,m^{-2}\,K^{-1}}$ at a cooling of -5 K), the most likely value of λ is not the same as the most likely value of λ when assuming $\alpha = 0$ (dotted and solid lines in Fig. 2a). In this case, the likelihoods (Fig. 2b) are calculated by integrating the joint probability distribution for ΔT and ΔF along the curve defined by Eq. (4), and they depend on the value of the state dependence parameter (α).

If α is not a fixed value but an unknown parameter, then the evidence can constrain only the joint distribution of $\Theta = (\lambda, \alpha)$. Obviously, in order for the climate of the past to tell us anything about the climate of the future, we must have some information about how they relate to one another.

There is no limit to the complexity of models we might use to interpret the evidence from the LGM. We might allow for both non-unit forcing efficacy and state dependence. We might assign different efficacies to different forcing agents or allow the parameter α to bifurcate at lower temperatures. We might also include an additive pattern effect ($\Delta\lambda$) that reflects differences in the spatial pattern of temperature change at the LGM and the pattern of warming expected at elevated CO_2 concentrations (e.g., Cooper et al., 2024).

Regardless of the interpretive model used, the model is both required for analysis and subjectively chosen by the analyst. Different reasonable analysts might make different choices about the model that should be used. This means that the choice of model is an important source of uncertainty that must be clearly specified or quantified. There is, however, one more source of uncertainty to discuss. Even with a single model, such as M_α, our degree of confidence in the constraints placed by paleoclimate evidence on the feedback parameter (λ) necessarily reflects our prior knowledge of the state dependence of climate feedbacks. It is to this prior uncertainty that we turn in Sect. 5.

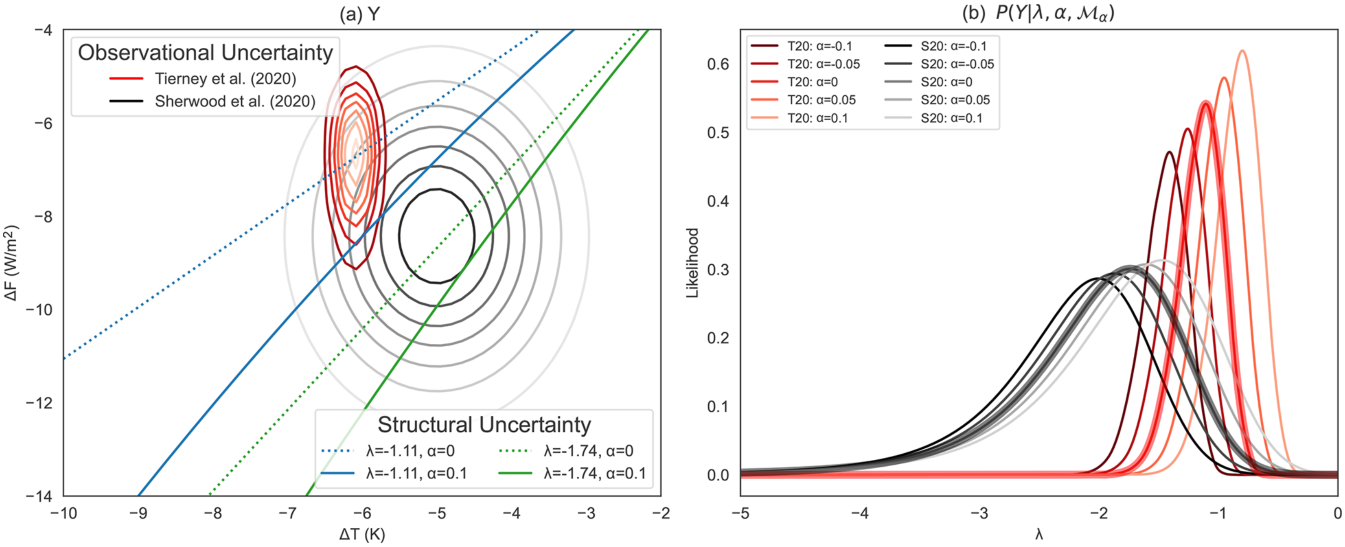

Figure 2. (a) Joint-evidence distributions for ΔT and ΔF used in Sherwood et al. (2020) (black contours) and Tierney et al. (2020) (red contours). Structural uncertainty is illustrated using solid lines (corresponding to fixed values of λ using the model M_0) and dashed lines (corresponding to fixed values of λ and α using the model M_α). **(b)** Likelihoods as a function of λ given evidence from S20 (black lines) or T20 (red lines) and different values of the state dependence parameter (α). **(b)** Resulting likelihoods for λ given evidence from S20 (black) or T20 (red) and different values of the state dependence parameter (α). Likelihoods derived using the simple energy balance model ($\alpha = 0$) are highlighted by thick lines.

5 Prior uncertainty

Once a model is specified, we aim to use the evidence to gain insights into its set of parameters (Θ). Bayes' theorem states that the posterior distributions of the parameters are simply obtained by multiplying the likelihood by the prior probability distributions, reflecting our preexisting beliefs and/or knowledge. These priors incorporate expert judgment, the results of other analyses, and knowledge of physical processes. Posterior distributions of individual parameters can depend strongly on prior knowledge of all parameters. For example, Fig. 3a shows the joint posteriors for the feedback parameter (λ) and the state dependence parameter (α), assuming the model M_α, the temperature and radiative-forcing values reported in S20, and uniform priors on both parameters.

In the absence of any physical knowledge about these parameters, the joint posterior is not very informative. In fact, considerable posterior weight is placed on extremely large positive values of α and positive values of λ, which would make negative climate sensitivity appear more likely than most scientists would consider credible. A well-informed scientist, however, is unlikely to think that $\alpha = 1$ (which implies an enormous mean change in feedback of $-5\,\mathrm{W\,m^{-2}\,K^{-1}}$ for 5 K of glacial cooling) is just as likely as $\alpha = 0$ (implying no change in feedback). In S20, a prior of $N(+0.1, 0.1)$ was assigned to the state dependence parameter (α), reflecting the current state of the literature. This prior substantially constrains the resulting joint posterior distribution (Fig. 3b). Conversely, imposing a more informative prior on the feedback parameter (λ) – for example, by using the process constraints in S20 that result in $\lambda \sim N(-1.30, 0.44)$ – also constrains the joint distribution: positive values of α (i.e., values which imply a lower sensitivity at the LGM than for doubled CO_2) receive more posterior weight. Combining the informative priors on both λ and α further constrains the joint posterior (Fig. 3d).

6 Combining multiple lines of evidence

The examples we have presented thus far have all used a single line of evidence – paleoclimate reconstructions of the Last Glacial Maximum – to constrain λ. However, it is not necessary to look back over 20 000 years to gauge the planet's response to external influences. More recently, a large increase in radiative forcing has resulted in significant global warming and a large radiative imbalance at the top of the atmosphere. To constrain λ with transient historical observations, we use the evidence $Y = (\Delta T, \Delta F, \Delta N)$, where ΔN is estimated from observed changes in ocean heat uptake and/or satellite observations constrained by ocean heat content (Forster, 2016).

6.1 Historical likelihood

In this three-dimensional joint probability space, the simple energy balance model M_0 defines a plane, rather than a line, in the evidence space (Fig. 4), and the likelihood of the evidence given λ is proportional to the integral over this surface.

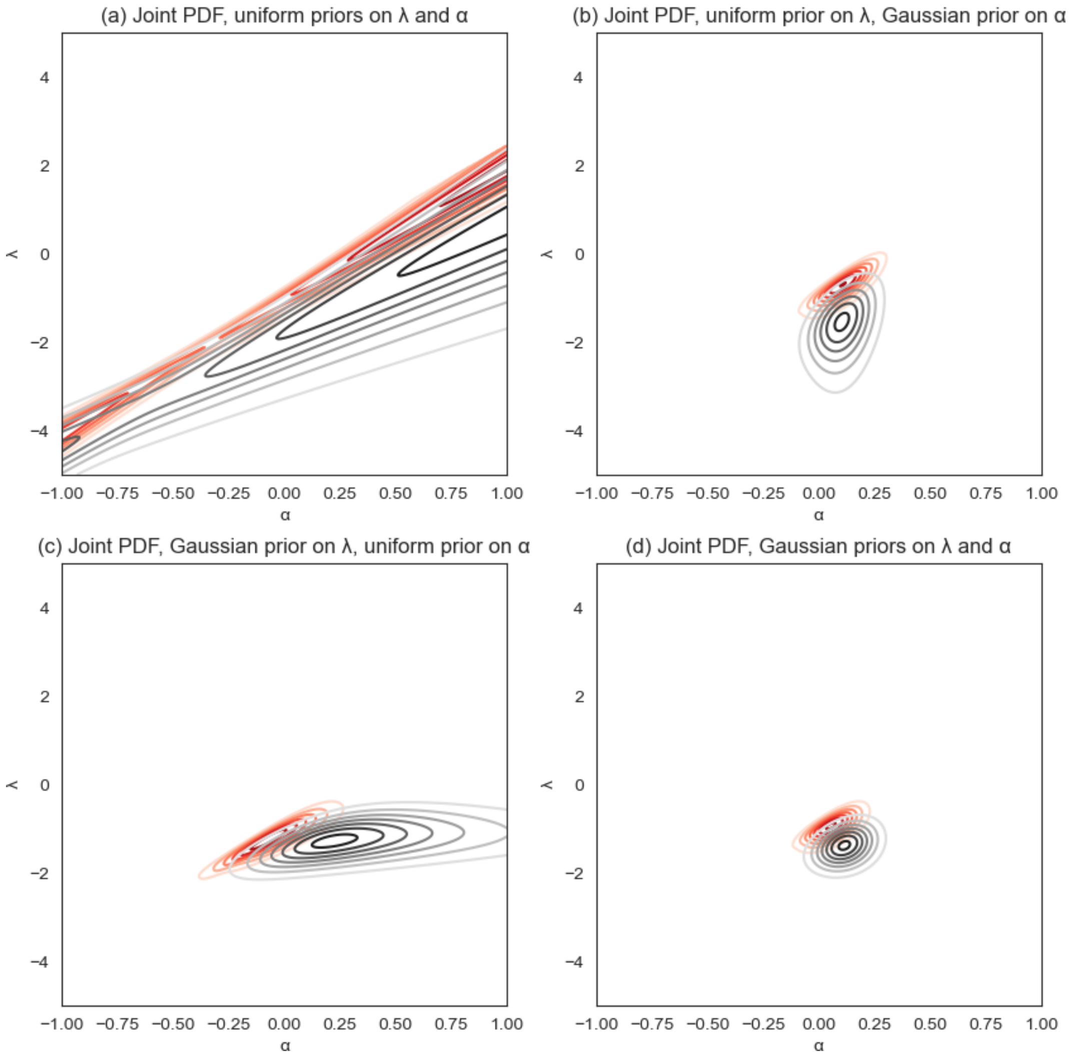

Figure 3. Joint posteriors for the feedback parameter (λ) and the state dependence parameter (α) under the influence of different priors: **(a)** uniform priors on both parameters; **(b)** a uniform prior on λ and a Gaussian prior, based on expert judgment from the published literature (used in S20), on α; **(c)** a Gaussian prior, based on process evidence (used in S20), on λ and a uniform prior on α; **(d)** and Gaussian priors (from S20) on both parameters. PDF: probability density function.

Figure 4 shows the historical evidence reported in S20, in which

$$\Delta T \sim N(1.03, 0.085), \tag{5}$$

$$\Delta N \sim N(0.6, 0.18), \tag{6}$$

and ΔF is calculated using unconstrained aerosol effective radiative forcings (ERFs) from Bellouin et al. (2020), with a median of $1.83\,\mathrm{W\,m^{-2}}$ and a 5 %–95 % range of -0.03 to $2.71\,\mathrm{W\,m^{-2}}$. The gray line in Fig. 4 shows the resulting likelihood as a function of λ. The maximum likelihood value is $\lambda = -1.53\,\mathrm{W\,m^{-2}\,K^{-1}}$.

However, the simple energy balance model M_0 assumes that the feedback parameter is the same for climate changes in the deep past, the transient historical period, and the future. Many studies (e.g., Marvel et al., 2016; Andrews et al., 2018; Dong et al., 2020; Rose et al., 2014; Armour et al., 2013; Gregory and Andrews, 2016; Marvel et al., 2018; Modak and Mauritsen, 2023) now argue that a more appropriate model should include a pattern effect ($\Delta\lambda$) that reflects the differences between feedbacks triggered by the observed spatial pattern of transient warming and feedbacks expected in response to the long-term equilibrium warming pattern. This model ($M_{\Delta\lambda}$) is given by

$$\Delta N = (\lambda - \Delta\lambda)\Delta T + \Delta F.$$

S20 placed a Gaussian prior on this pattern effect, with $\Delta\lambda = N(0.5, 0.3)\,\mathrm{W\,m^{-2}\,K^{-1}}$. This corresponds to a modification of the tilt of the plane shown in Fig. 4a. Because this model assumes that the pattern effect is linearly additive, no further curvature is introduced. By multiplying the joint

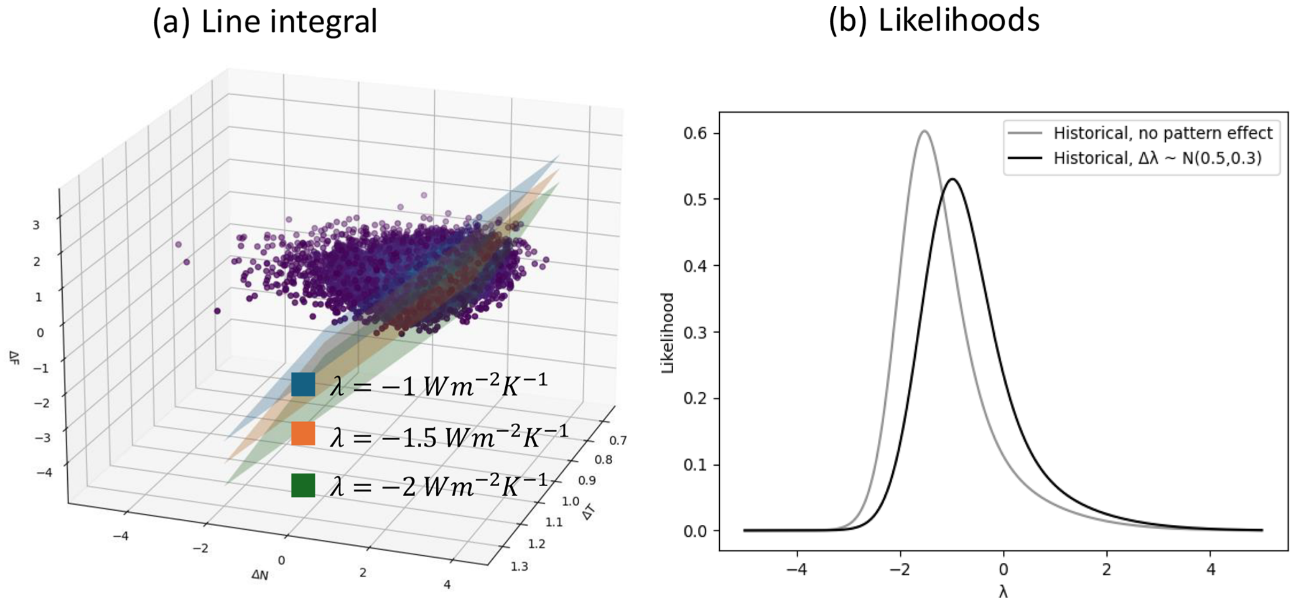

Figure 4. **(a)** Calculating the likelihood of observing the historical evidence used in S20 for a putative value of λ. Each value of λ defines a plane; shown are $\lambda = -1\,\mathrm{W\,m^{-2}\,K^{-1}}$ (blue), $\lambda = -1.5\,\mathrm{W\,m^{-2}\,K^{-1}}$ (orange), and $\lambda = -2\,\mathrm{W\,m^{-2}\,K^{-1}}$ (green). The likelihood is the surface integral of the joint PDF along the plane. **(b)** The likelihood of the feedback parameter (λ) given a simple energy balance model with no pattern effect (gray line) and the marginal likelihood of λ given an additive pattern effect with the prior $\Delta\lambda \sim N(0.5, 0.3)$.

likelihood ($P(\Delta Q, \Delta T, \Delta F|\lambda, \Delta\lambda)$) by the prior ($P(\Delta\lambda)$) and integrating over all values of $\Delta\lambda$, we obtain a marginal likelihood for the historical evidence as a function of the feedback parameter (λ). This is shown by the black line in Fig. 4b. The inclusion of the additive pattern effect and our physics-informed intuition that said effect is likely to be positive shifts the most likely value of the feedback parameter to $\lambda = -1.0\,\mathrm{W\,m^{-2}\,K^{-1}}$.

The pattern effect estimate used in S20 was based on the Atmospheric Model Intercomparison Project II (AMIP-II) dataset, which produces the largest estimate of the pattern effect (Modak and Mauritsen, 2023), and therefore the priors on $\Delta\lambda$ used in S20 may be both overconfident and weighted too strongly toward high values. However, while noting this important caveat, for illustrative purposes we will use the S20 historical likelihood marginalized over the pattern effect estimate as the historical likelihood for the rest of this paper.

6.2 The "Twin Peaks" problem

Assuming conditional independence between lines of evidence, the posterior distribution of the feedback parameter (λ) is expressed as

$$P(\lambda|Y) \propto P(Y_{\mathrm{hist}}|\lambda)\, P(Y_{\mathrm{LGM}}|\lambda)\, P(\lambda). \tag{7}$$

That is, the posterior estimate of λ given two lines of evidence is proportional to the product of the individual likelihoods. But what if the likelihoods have a small (or no) region of overlap? Can we really be confident that the posterior estimate is well constrained in this case? Figure 5a highlights this potential pitfall. The black line shows the marginal likelihood of the historical evidence as a function of λ. The light-blue line shows the likelihood of the S20 LGM evidence as a function of λ, assuming no state dependence ($\alpha = 0$). The product of these likelihoods is indicated by the dashed green line. The less the historical and paleoclimate likelihoods overlap, the narrower the posterior will be. We refer to this conundrum as the Twin Peaks problem: should greater incompatibility between multiple lines of evidence *really* reduce the uncertainty in λ? Or could it be that the two lines of evidence are not, in fact, measuring the same thing?

We can take the latter possibility into account using an alternate model for the paleo-evidence. Note that the posterior for λ shown in Fig. 5a is conditional on the model M_0 for the paleoclimate evidence, which contains only one parameter, λ. The model assumes that the equilibrium feedbacks in a warmer climate are exactly the same as those in a colder climate, that the response to pure $\mathrm{CO_2}$ forcing is equivalent to the response to LGM forcings, and that the pattern effect is zero over the LGM. An alternate model, say M_α, allows for state dependence via an additional parameter (α). The marginal likelihood for the paleodata, given M_α and Gaussian priors on α, is indicated by a dark-blue line in Fig. 5b. While the overlap between these two distributions is far from exact, it is substantially larger than that for the no-state-dependence case, illustrated in Fig. 5a. Simply put, the historical evidence and the LGM evidence appear to be more compatible when we correct for the state dependence of the past cold period.

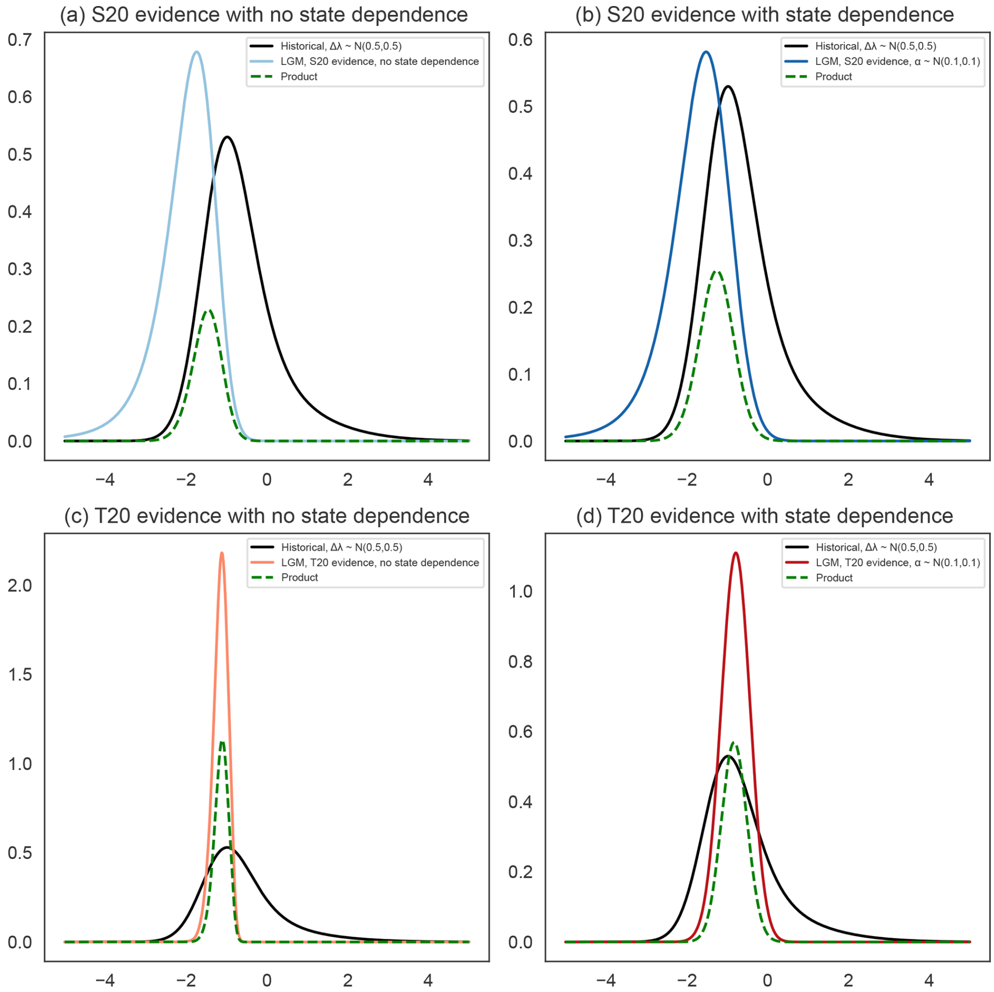

Figure 5. Likelihoods from multiple lines of evidence. In all four panels, the black line shows the likelihood of the historical evidence given λ, assuming the pattern effect $\Delta\lambda \sim N(0.5, 0.3)$. **(a)** The likelihood of S20 evidence given λ, assuming no state dependence during the LGM (light-blue line) and overlap (dashed green line). **(b)** The likelihood of S20 evidence given λ, assuming state dependence and $\alpha \sim N(0.1, 0.1)$ (dark-blue line), as well as overlap (dashed green line). **(c)** The likelihood of T20 evidence given λ, assuming no state dependence during the LGM (orange line) and overlap (dashed green line). **(d)** The likelihood of T20 evidence given λ, assuming state dependence and $\alpha \sim N(0.1, 0.1)$ (dark-red line), as well as overlap (dashed green line).

When using T20 evidence, however, there is considerable overlap between the historical likelihood (with a pattern effect) and the paleoclimate likelihood (with no state dependence). As in Fig. 5a and b, the black lines in Fig. 5c and d show the historical likelihood. The likelihood for λ obtained from T20 evidence, assuming no state dependence (orange line in Fig. 5c), closely overlaps with the historical likelihood, as does the likelihood assuming state dependence with a prior on α, as in S20 (red line in Fig. 5d). The latter model, however, yields a broader likelihood for λ, and therefore the region of overlap with the historical evidence is smaller.

Combining multiple lines of evidence, therefore, introduces another source of unavoidable subjectivity: how can we be sure that, in doing so, we are comparing "apples to apples"?

6.3 Model odds

The question of how to compare separate lines of evidence is a question of models: namely, how do we interpret these separate lines? Fortunately, Bayesian methods allow us to compare and criticize models based on the evidence. Consider, for example, two models for the LGM: M_0 and M_α. The model odds are defined as

$$\begin{aligned}\text{odds} &= \frac{P\left(M_\alpha|Y_{\text{hist}}, Y_{\text{paleo}}\right)}{P\left(M_0|Y_{\text{hist}}, Y_{\text{paleo}}\right)}\\ &= \frac{P\left(Y_{\text{hist}}, Y_{\text{paleo}}|M_\alpha\right)P(M_\alpha)}{P\left(Y_{\text{hist}}, Y_{\text{paleo}}|M_0\right)P(M_0)}\\ &\equiv \text{BF} \times \frac{P\left(M_\alpha\right)}{P(M_0)},\end{aligned}$$

where the Bayes factor (BF) is the ratio of the evidence for each model.

The model evidence for any given model (M_ℓ) is defined as the integrated likelihood over the entire set of parameter values (Θ_ℓ), expressed as follows:

$$P(Y|M_\ell) = \int P(Y|\Theta, M_\ell)\, P(\Theta_\ell|M_\ell)\, \mathrm{d}\Theta_\ell. \tag{8}$$

This reflects the probability that the model M_ℓ could have generated the observed evidence under a given set of priors on its parameters (θ_ℓ).

For example, the model evidence for the model M_0 is given by

$$P\left(Y_{\text{hist}}, Y_{\text{paleo}}|M_0\right) \propto \int P\left(Y_{\text{paleo}}|\lambda\right) P_{\Delta\lambda}\left(Y_{\text{hist}}|\lambda\right) P(\lambda)\mathrm{d}\lambda,$$

where $P_{\Delta\lambda}(Y_{\text{hist}}|\lambda)$ is the marginal historical likelihood (black line in Fig. 5a). When combined with a uniform prior on λ, the model evidence for M_0 is therefore the area under the green curve in Fig. 5a.

By contrast, the model evidence for the model M_α is given by

$$P\left(Y_{\text{hist}}, Y_{\text{paleo}}|M_\alpha\right) \propto \int P\left(Y_{\text{paleo}}|\lambda, \alpha\right) P_{\Delta\lambda}\left(Y_{\text{hist}}|\lambda\right) P(\alpha) P(\lambda)\mathrm{d}\alpha\mathrm{d}\lambda.$$

When combined with a uniform prior on λ, the model evidence for M_α is the area under the green curve in Fig. 5b.

Using S20 evidence and these priors, we find that the Bayes factor is 1.33. This means that if our prior assumes both models are equally likely, the evidence shifts these odds: the model depicted in Fig. 5b is about 33 % more likely to have generated the observed paleo-evidence and historical evidence.

However, using T20 evidence, the Bayes factor is 0.93. This suggests that the better model to use, given T20 evidence, is one without state dependence. Clearly, the best model depends on the evidence used; the prior knowledge of whether we are comparing apples to apples; and the priors we place on λ, $\Delta\lambda$, and α.

We note that whether the Twin Peaks problem is indeed a problem is largely dependent on the prior odds ($P(M_\alpha)/P(M_0)$), which must be specified. If we have prior knowledge that the two lines of evidence are measuring the same thing, then we will give more prior weight to the simple model M_0, and the Bayes factor will do little to shift the odds. This will result in a narrower posterior estimate: if two lines of evidence are compatible only for a small range of values, and we are confident in what the evidence is telling us, then we may be more confident in its posterior value.

7 A way forward

Thus far, we have established that there are three points at which unavoidable subjective decisions must be made: when collecting evidence, when choosing the interpretive model, and when assessing prior knowledge of that model's parameters. We have also established that multiple lines of evidence appear more or less compatible, depending on the models used. Here, we present a suggested framework for making these decisions in a community assessment framework.

7.1 Handling evidence uncertainty

Whether and how much a newly published estimate of a particular quantity (for example, ΔT or ΔF from the Last Glacial Maximum) affects the evidence base depends on prior knowledge of that quantity. It also depends on expert assessment of how the new study relates to the existing literature. A single, highly certain, high-quality study can strongly shift previously uncertain estimates, while low-quality or uncertain published estimates may not change previously firm understandings.

We suggest formalizing these intuitions using a Bayesian random-effects meta-analysis (Smith et al., 1995), frequently used in fields as diverse as psychology (Gronau et al., 2021), medicine (Sutton and Abrams, 2001), and ecology (Koricheva et al., 2013). This model can be written as

$$\hat{y}_j \sim N(y_j, \sigma_j), \tag{9}$$

$$y_j \sim N(Y, \tau), \tag{10}$$

where $\hat{y}_j$ and σ_j are the reported mean and standard deviation of each study (j). We assume that the true (latent) mean (y_j) of each study is normally distributed about an overall mean (Y), with τ representing the expected inter-study standard deviation.

The priors we place on the quantities of interest – the overall mean (Y) and the between-study spread (τ) – quantify our previous knowledge of and views about the literature. A τ value very close to zero suggests homogeneity across the studies (and, in fact, choosing to set $\tau = 0$ reduces the random-effects model to the fixed-effects model). By contrast, if we have reason to believe that multiple studies should vary in their reported values due to structural and design factors, then we might place a broad prior on τ. For example, a fixed-effects model might be appropriate for calculating the ensemble mean of a quantity within a single Coupled Model Intercomparison Project (CMIP) model, whereas a random-effects model might be more appropriate for combining ensembles of multiple CMIP models, which we know differ structurally.

As a specific example relevant to calculating the feedback parameter (λ), we can consider the multiple published values of LGM global mean temperature change (ΔT) derived from proxies and models, as well as from PMIP3 and PMIP4 models (Table 1).

Figure 6 illustrates how the posterior distribution of ΔT depends on prior beliefs about the nature and quality of the published literature assessing it.

Table 1. Estimates of global cooling (ΔT) during the Last Glacial Maximum. SST: sea surface temperature. CESM: Community Earth System Model. CNRM: Centre National de Recherches Météorologiques. IPSL: Institute Pierre-Simon Laplace. MRI: Meteorological Research Institute.

Mean (K)	Standard deviation	Reference	Source	Generation
−4.00	0.41	Annan and Hargreaves (2013)	Proxies and models	–
−5.80	0.77	von Deimling et al. (2006)	Proxies and models	–
−6.20	0.46	Holden et al. (2009)	GENIE-1	–
−3.58	0.12	Shakun et al. (2012)	Proxies	–
−6.20	0.92	Snyder (2016)	Proxies and models	–
−6.30	0.61	Bereiter et al. (2018)	Proxies (e.g., ocean temperature) and models	–
−5.70	0.20	Friedrich and Timmermann (2020)	–	–
−5.75	0.38	Friedrich et al. (2016)	SST proxies and a model simulation	–
−6.10	0.20	Tierney et al. (2020)	Proxies and an isotope-enabled climate model	–
−5.00	1.00	Sherwood et al. (2020)	Synthesis	–
−4.85	–	Kageyama et al. (2021)	CESM	PMIP3
−2.70	–	Kageyama et al. (2021)	CNRM	PMIP3
−4.63	–	Kageyama et al. (2021)	FGOALS-g2	PMIP3
−4.92	–	Kageyama et al. (2021)	GISS-E2-p1	PMIP3
−5.19	–	Kageyama et al. (2021)	GISS-E2-p2	PMIP3
−4.64	–	Kageyama et al. (2021)	IPSL	PMIP3
−5.40	–	Kageyama et al. (2021)	MIROC	PMIP3
−4.41	–	Kageyama et al. (2021)	MPI-p1	PMIP3
−4.67	–	Kageyama et al. (2021)	MPI-p2	PMIP3
−4.71	–	Kageyama et al. (2021)	MRI	PMIP3
−3.75	–	Kageyama et al. (2021)	AWI-ESM-1	PMIP4
−3.81	–	Kageyama et al. (2021)	AWI-ESM-2	PMIP4
−6.80	–	Kageyama et al. (2021)	CESM1.2	PMIP4
−7.16	–	Kageyama et al. (2021)	HadCM3-PMIP3	PMIP4
−5.92	–	Kageyama et al. (2021)	HadCM3-ICE6GC	PMIP4
−6.46	–	Kageyama et al. (2021)	HadCM3-GLAC1D	PMIP4
−3.28	–	Kageyama et al. (2021)	iLOVECLIM-ICE-6G	PMIP4
−3.26	–	Kageyama et al. (2021)	iLOVECLIM-GLAC1D	PMIP4
−3.73	–	Kageyama et al. (2021)	INM-CM4-8	PMIP4
−4.63	–	Kageyama et al. (2021)	IPSLCM5A2	PMIP4
−4.02	–	Kageyama et al. (2021)	MIROC-ES2L	PMIP4
−3.90	–	Kageyama et al. (2021)	MPI-PMIP4	PMIP4
−5.27	–	Kageyama et al. (2021)	UT-CCSM4	PMIP4

Consider, for example, a random-effects model in which we place broad priors on the mean $\mu \sim N(0, 100)$ and the inter-study standard deviation $\tau \sim U(0, 100)$. With these prior assumptions, 90 % of the resulting posterior density for μ (the true value of ΔT) lies between −5.9 and −4.8 K. Assuming that there is no inter-study spread (i.e., assuming that τ is zero with zero uncertainty, corresponding to a fixed-effects model) would yield an estimate of ΔT that is 90 % likely to be between −4.8 and −4.5 K. This much narrower (and warmer) estimate results from the extremely restrictive prior belief that every study, regardless of method, targets the same underlying ΔT value and would yield the same results if performed perfectly and with adequate data. Similarly, we might set the prior on μ using the result of a single published study (for example, ΔT from T20). Combined with a broad uniform prior on the inter-study spread, this results in an 90 % posterior density estimate ranging from −6.2 to −5.6 K. If, however, we adopt the restrictive fixed-effects model, the T20 study is merely treated as an outlier and fails to substantially move the posterior distribution toward cooler values of ΔT (red line), even when using the T20 prior.

7.1.1 Recommendations

Unavoidable subjective decisions about the evidence can be made explicit by adopting a random-effects meta-analysis. This requires the specification of priors on the inter-study spread (τ) and the overall mean (Y). Our recommendation is that the organizers of community assessments choose and clearly specify these priors, rather than allowing individual experts to choose their own.

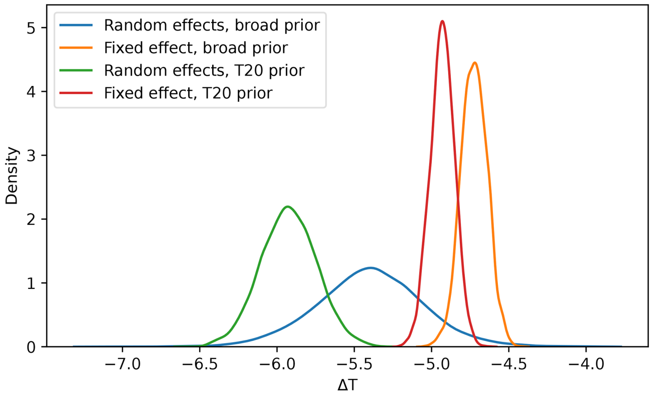

Figure 6. How cold was the Last Glacial Maximum? The answer depends on your prior beliefs about the cooling and the literature. Shown are posterior distributions for the LGM cooling (ΔT), assuming a random-effects model with a broad prior (blue line) or a T20 prior (green) on the mean or assuming a fixed-effects model with a broad prior (orange line) or a T20 prior (red line) on the mean.

7.2 Handling model uncertainty

As shown in Sect. 4, the constraints placed on climate sensitivity by multiple lines of evidence depend on the model(s) used to interpret that evidence. This means that the design of each expert assessment must be explicit regarding its interpretive models. As the assessment is planned, it is crucial to arrive at a consensus on credible interpretive models for the evidence. For example, one possible model for the Last Glacial Maximum might incorporate the parameters α (representing state dependence), ξ (representing the difference between long-term equilibrium feedbacks for the LGM and target quasi-equilibrium feedbacks for doubled CO_2), and $\Delta\lambda_{\mathrm{LGM}}$ (representing sea surface pattern differences between the LGM and doubled CO_2), resulting in the following equation:

$$\Delta T = \frac{-\Delta F}{\frac{\lambda+\Delta\lambda_{\mathrm{LGM}}}{1+\xi} + \frac{\alpha}{2}\Delta T}.$$

Given a model, experts may then be asked to specify their prior beliefs about each parameter. If an expert disagrees with the inclusion of a parameter in the model, they would be free to set a prior that is very narrowly clustered around zero.

If consensus cannot be reached on a particular model, then we suggest that the planners for any assessment arrive at a list of candidate models ($M_1 \ldots M_K$). The aggregate posterior can then be taken as a weighted average over different models as follows:

$$P(\Theta|Y) = \sum_{k=1}^{K} w_k P(\Theta|M_k, Y). \tag{11}$$

Here, $(\Theta|M_k, Y)$ is the posterior obtained using the model M_k to interpret the evidence (Y).

The weights reflect how well the model fits the data and are given by

$$w_k = P(M_k|Y) = \frac{P(Y|M_k)P(M_k)}{\sum_{k=1}^{K} P(Y|M_k)P(M_k)}. \tag{12}$$

The term $P(M_k|Y)$ is the model evidence (Eq. 8), as discussed in Sect. 6.2. These weights, and hence the combined posterior, depend on the set of priors ($P(M_k)$) we place on the correctness of each model. If an assessment allows for experts to use one of multiple models, it is imperative to specify assessment-wide priors for these models upfront.

7.2.1 Recommendations

We recommend that organizers of community assessments clearly specify a single interpretive model for the evidence used. If this is not possible, organizers should specify a list of possible candidate models (M_k) and a prior ($P(M_k)$) for each candidate model. The resulting estimate will then be a weighted average over the models.

7.3 Expert elicitation via priors

Finally, it is necessary to quantify the degree of preexisting knowledge and/or beliefs through the use of prior distributions. This is where a wide variety of expert opinions may be usefully incorporated into an assessment.

However, we require consistent ways to aggregate the judgments of multiple experts. In theory, sufficient evidence should lead to a high degree of agreement, even if different experts begin the analysis with very different priors. Figure 7a shows the priors placed on the parameter λ by two hypothetical experts. Expert A (solid red line) believes that the feedback parameter is less negative than Expert B (solid blue line) thinks it is and is even open to the idea that it might be positive. The dashed red and blue lines show both experts' posteriors, updated using the evidence presented in S20. While the experts began their analysis with differing opinions, the weight of the evidence has updated their understandings, and they now agree about the feedback parameter (λ).

However, some experts may not be as open-minded as Experts A and B. Expert C (blue line in Fig. 7b believes that the feedback parameter is strongly negative. Moreover, they are extremely confident in this: their prior distribution is very narrowly peaked around a value of $\lambda = -3\,\mathrm{W\,m^{-2}\,K^{-1}}$. Expert C's confidence remains unshaken by the evidence presented in S20, and their posterior remains nearly identical to their prior beliefs. How should an assessment handle such excessively confident experts, whose beliefs appear to be unshaken by any reasonable amount of evidence?

Consider an assessment in which N experts each specify a set of priors ($P_i(\theta)$), where $i = 1 \ldots N$. A reasonable aggregate prior might then be a linear combination of the individ-

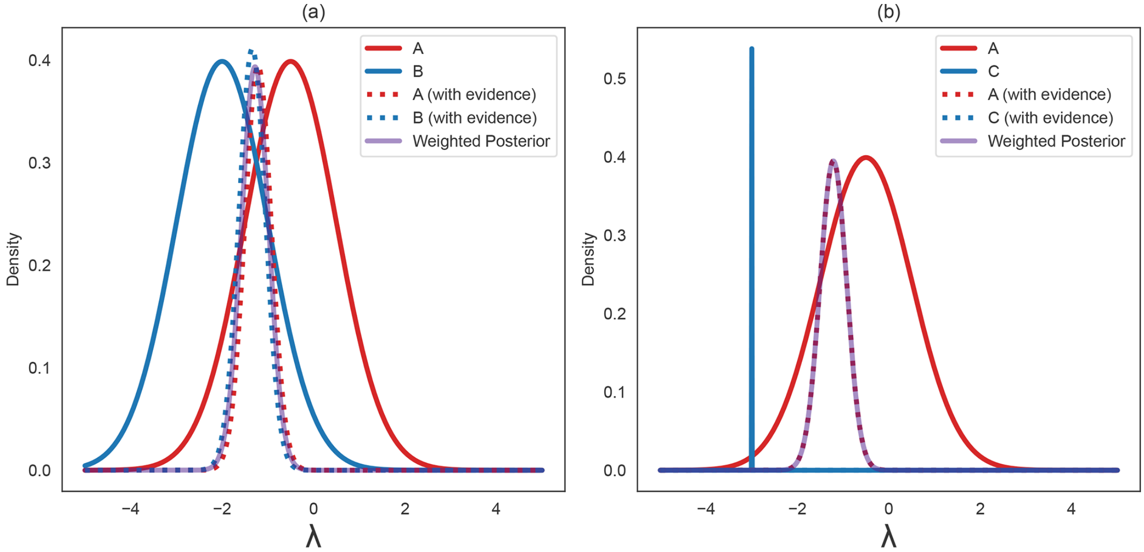

Figure 7. **(a)** Experts A (solid red line) and B (solid blue line) begin with different priors on λ. The evidence presented in S20 updates these priors, and the resulting posteriors are nearly identical (dotted red and blue lines). The purple line shows the weighted posterior. **(b)** Experts A (solid red line) and C (solid blue line) begin with different priors on λ, but Expert C's prior is very narrowly peaked. The evidence presented in S20 updates these priors, but the posteriors remain very different (dotted red and blue lines). The purple line shows the weighted posterior, which is almost identical to Expert A's posterior.

ual expert priors, expressed as

$$P(\theta) = \sum_{i=1}^{N} a_i P_i(\theta). \tag{13}$$

The aggregate posterior is therefore a weighted average of the individual expert posteriors, given by

$$P(\theta, Y) = \sum_i \tilde{a}_i P_i(\theta | i, Y), \tag{14}$$

where

$$\tilde{a}_i = \frac{a_i \int P(Y|\theta) P_i(\theta) \mathrm{d}\theta}{\sum_{i=1}^{N} a_i \int P(Y|\theta) P_i(\theta) \mathrm{d}\theta}. \tag{15}$$

This method introduces N new parameters, representing the prior weight (a_i) we assign to each expert's judgment. This is a far easier task than setting priors on models (as discussed in Sect. 7.2) because it requires no physical understanding but rather a belief about the quality of each expert's initial beliefs. We recommend weighting each expert equally by setting $a_1 = a_2 = \ldots a_N = \frac{1}{N}$, in which case the posterior weight becomes

$$\tilde{a}_i = \frac{\int P(Y|\theta) P_i(\theta) \mathrm{d}\theta}{\sum_{i=1}^{N} \int P(Y|\theta) P_i(\theta) \mathrm{d}\theta}. \tag{16}$$

The purple line in Fig. 7a shows the resulting aggregate posterior given Expert A and Expert B's priors. Because these experts are similarly able to update their priors, the weighting process has no effect on the outcome. However, the weighted average of Expert A and Expert C's posteriors, indicated by the purple line in Fig. 7b, is similar to Expert A's posterior distribution. The narrowness of Expert C's prior causes their posterior distribution to be down-weighted in the weighted average. We suggest this as an effective strategy for handling inflexible or extremely anomalous expert opinions.

7.3.1 Recommendations

We recommend eliciting expert judgment in a systematic way by allowing experts to specify priors on predetermined model parameters. The analysis can then be performed using a single aggregate posterior, calculated as the weighted average of individual expert posteriors.

8 Conclusions

Here, we have presented three sources of uncertainty that enter into estimates of climate sensitivity, which can be summarized as three sets of questions as follows. First, what evidence should we use to constrain climate sensitivity, how do we decide what counts as evidence, and how should we handle estimates that disagree or conflict with each other? Second, what interpretive model should we use to relate the evidence to climate sensitivity, and what parameters are required? Third, what prior knowledge of these parameters is appropriate and should be included? We then propose a strategy to make the role of expert judgment in subsequent assessments fairer and more transparent. The advantage of this strategy, combining Bayesian meta-analysis and Bayesian model averaging, is that it can incorporate newly published data and is easily expanded to handle uncertainties at multiple levels.

There is no limit to the number of nested levels we could theoretically use within a Bayesian hierarchical model: the prior for radiative forcing from ice sheets, for example, can be updated using a global ice sheet reconstruction, which itself is constrained by individual geological measurements. Similarly, a prior on ocean heat uptake (ΔN) or historical warming (ΔT) can be updated as new measurements become available. However, to remain tractable, every project must truncate the hierarchy at some finite level. In practice, this means treating the posteriors that arise from observational, general-circulation-model (GCM), or paleoclimate studies as evidence; where we draw the line between evidence and parameters sets the bounds of our analysis.

As a result, we propose a framework in which experts are required to specify their choices at clearly defined decision points. Once priors are specified, the model and evidence will update them accordingly, arriving at a new, aggregate consensus posterior. We review this framework here.

Somewhat obviously, experts' beliefs about the data are based on their prior beliefs, updated by the evidence. But how experts interpret and use that evidence depends on the subjective choices they make – i.e., what counts as a study or evidence? How should we best compare estimates derived from proxies or observations with estimates from GCMs? Should some studies receive more weight than others? In our framework, experts must make judgments about the evidence by asking the following questions:

1. What is your informed belief about the evidence? (For example, what is your prior on μ?)

2. What is your belief about the published literature? (What is your prior on τ?)

Second, we suggest taking the choice of model out of individual participants' hands to the greatest extent possible. Ideally, assessment planners should arrive at a single model and set of parameters on which experts may specify their priors. If not, they should arrive at a list of candidate models, specify firm prior beliefs about these models, and perform Bayesian model averaging over the posteriors of individual experts, which will depend on the model they use.

Third, once a model is specified, experts should specify their prior beliefs about the parameters of that model.

The results presented here are meant to begin, rather than end, a conversation. The beauty of Bayesian methods is that we can allow new evidence to update our existing knowledge. As climate researchers gear up for the next generation of model intercomparison projects and assessments, it is important to consider how these new results will be integrated with existing knowledge. Our methods presented here allow for new discoveries to advance our understanding, ultimately narrowing the bounds of climate sensitivity and informing future research and decision-making.

Appendix A: Exact forms of integrals

To estimate the likelihood of the evidence (ΔT and ΔF) given the simple energy balance model, we integrate the joint probability distribution ($\mathcal{J}(\Delta T, \Delta F)$) over the curve ($C$) using the following model:

$$P(Y|\lambda, M_0) = \int\limits_C \mathcal{J}(\Delta T, \Delta F)\mathrm{d}s. \tag{A1}$$

C can be parameterized as

$$\mathbf{r}(t) = t\hat{i} + -\lambda t\hat{j}, \tag{A2}$$

and the integral is then given by

$$\begin{aligned} P(Y|\lambda, M_0) &= \int\limits_{-\infty}^{\infty} \mathcal{J}(\mathbf{r}(t))\,||\mathbf{r}'(t)||\,\mathrm{d}t \\ &= \int\limits_{-\infty}^{\infty} \mathcal{J}(t, -\lambda t)\sqrt{1+\lambda^2}\,\mathrm{d}t. \end{aligned} \tag{A3}$$

In the case where ΔT and ΔF are Gaussian and independent (with the means μ_T and μ_F and the standard deviations σ_T and σ_F, respectively), the likelihood has an exact analytic form, substantially speeding up its computation. This form is expressed as

$$P(Y|\lambda, M_0) = C\left(\frac{2\pi}{A}\right)^{1/2} \exp\left(\frac{B^2}{2A}\right), \tag{A4}$$

where

$$C = \frac{\sqrt{1+\lambda^2}}{2\pi\sigma_T\sigma_F}\exp\left(\frac{\mu_T^2}{\sigma_T^2} + \frac{\mu_F^2}{\sigma_F^2}\right),$$

$$A = \frac{1}{\sigma_T^2} + \frac{\lambda^2}{\sigma_F^2},$$

$$B = \frac{\mu_T}{\sigma_T^2} - \frac{\lambda\mu_F}{\sigma_F^2}.$$

In the case of a three-dimensional space (as with the historical evidence), the curve (C) defines a plane rather than a line, and we have

$$\begin{aligned} P(Y|\lambda) &\propto \int\limits_C \mathcal{J}(\Delta T, \Delta F, \Delta N)\mathrm{d}S \\ &= \int\int \mathcal{J}\left(\boldsymbol{r}(u,v)\right)||r_u \times r_v||\mathrm{d}u\,\mathrm{d}v, \end{aligned} \tag{A5}$$

where

$$\boldsymbol{r} = u\hat{i} + v\hat{j} + (\lambda u + v)\hat{k}. \tag{A6}$$

Appendix B: Likelihood vs. probability

We note that this method is distinct from estimating λ as the ratio of the distributions ΔF and ΔT. This is due to a conceptual difference between probability and likelihood. Constructing the likelihood answers the following question: (1) how likely is a particular hypothesis (in this simple case, a particular value of λ) given the evidence? This is a fundamentally different question from the following: (2) what is the probability density function of the ratio $-\Delta F/\Delta T$? The first question involves fixing a putative value of λ, which is *not* treated as a random variable. The second question treats λ as a random variable. Mathematically, this is reflected in the difference between (1) a line integral over the curve $y = -\lambda x$,

$$P(x,y|\lambda) = \int_C P_{xy}(x,y)\mathrm{d}s = \int_{-\infty}^{\infty} P_{xy}(x,-\lambda x)\sqrt{1+\lambda^2}\,\mathrm{d}x,$$

and (2) the ratio distribution of the random variable $\lambda = -y/x$,

$$P_\lambda(\lambda) = \int_{-\infty}^{\infty} P_{xy}(x,-\lambda x)|x|\,\mathrm{d}x.$$

We use the ratio distribution (b) to estimate S once we have the posterior probability density function (PDF) for λ. This is because we treat S as the ratio of two random variables, $F_{2\times\mathrm{CO_2}}$ and λ.

Appendix C: Correlations between $F_{2\times\mathrm{CO_2}}$ and ΔF

$\mathrm{CO_2}$ emissions are the primary contributor to present-day radiative-forcing change relative to preindustrial concentrations. Atmospheric concentrations of $\mathrm{CO_2}$ were lower in the Last Glacial Maximum. This means that the forcing term (ΔF) used as evidence in the LGM and historical periods is correlated with the forcing corresponding to doubled $\mathrm{CO_2}$. For visual clarity, we neglect this correlation in this paper. To take it into account, we can write the simple energy balance model as

$$\Delta N = \Delta F' + \beta F_{2\times\mathrm{CO_2}} + \lambda \Delta T.$$

In this case, the likelihood $P(E|\lambda, F_{2\times\mathrm{CO_2}})$ is defined as the integral of the joint probability distribution of the evidence (E) over the curve defined by the model. Following S20, we can then calculate S by changing the variables and marginalizing over $F_{2\times\mathrm{CO_2}}$ as follows:

$$P(S|E) = \int P\left(\lambda', F'_{2\times\mathrm{CO_2}}|E\right) \delta\left(S - F'_{2\times\mathrm{CO_2}}/\lambda'\right)\left(\partial S/\partial\lambda'\right)^{-1} \left(\partial S/\partial F'_{2\times\mathrm{CO_2}}\right)^{-1}\mathrm{d}F'{}_{2\times\mathrm{CO_2}}\mathrm{d}\lambda'.$$

Practically, we can draw samples of λ and $F'{}_{2\times\mathrm{CO_2}}$ from the joint posterior distribution and use these to calculate a posterior distribution for S. This correlation contributes very little to the results; when taking it into account, we obtain similar ranges for S as when we neglect it.

Code availability. The code for this project is available at https://github.com/netzeroasap/LambdaBayes/ (https://doi.org/10.5281/zenodo.13905523, Marvel and Webb, 2024).

Data availability. No new data sets were generated for this article; references to existing data sets are provided throughout the text.

Author contributions. KM and MW carried out the analysis jointly. KM wrote the analysis code and performed the mathematical calculations. Both authors contributed to writing the paper.

Competing interests. The contact author has declared that neither of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. We are grateful to John Eyre and to three anonymous reviewers, whose comments on the paper led to substantial improvements. Additionally, we acknowledge James Webb for suggesting the name for the Twin Peaks problem.

Financial support. This research has been supported by the NASA Modeling, Analysis, and Prediction Program and the Met Office Hadley Centre Climate Programme, funded by DSIT.

Review statement. This paper was edited by Ben Kravitz and reviewed by three anonymous referees.

References

Albani, S. and Mahowald, N. M.: Paleodust insights into dust impacts on climate, J. Clim., 32, 7897–7913, 2019.

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., Webb, M. J., Armour, K. C., Forster, P. M., and Titchner, H.: Accounting for changing temperature patterns increases historical estimates of climate sensitivity, Geophys. Res. Lett., 45, 8490–8499, 2018.

Annan, J. D. and Hargreaves, J. C.: A new global reconstruction of temperature changes at the Last Glacial Maximum, Clim. Past, 9, 367–376, https://doi.org/10.5194/cp-9-367-2013, 2013.

Annan, J. D., Hargreaves, J. C., and Mauritsen, T.: A new global surface temperature reconstruction for the Last Glacial Maximum, Clim. Past, 18, 1883–1896, https://doi.org/10.5194/cp-18-1883-2022, 2022.

Armour, K. C., Bitz, C. M., and Roe, G. H.: Time-varying climate sensitivity from regional feedbacks, J. Clim., 26, 4518–4534, 2013.

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A.-L., Dufresne, J.-L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J. M., Lohmann, U., Malavelle, F., Mauritsen, T., McCoy, D. T., Myhre, G., Mülmenstädt, J., Neubauer, D.,, Possner, Rugenstein, M., Sato, Y., Schulz, M., Schwartz, S. E., Sourdeval, O., Storelvmo, T., Toll, V., Winker, D., and Stevens, B.: Bounding global aerosol radiative forcing of climate change, Rev. Geophys., 58, e2019RG000660, https://doi.org/10.1029/2019RG000660, 2020.

Bereiter, B., Shackleton, S., Baggenstos, D., Kawamura, K., and Severinghaus, J.: Mean global ocean temperatures during the last glacial transition, Nature, 553, 39–44, https://doi.org/10.1038/nature25152, 2018.

Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, Nat. Clim. Change, 2, 417–424, https://doi.org/10.1038/nclimate1456, 2012.

Budyko, M. I.: The effect of solar radiation variations on the climate of the Earth, Tellus, 21, 611–619, 1969.

Caballero, R. and Huber, M.: State-dependent climate sensitivity in past warm climates and its implications for future climate projections, P. Natl. Acad. Sci. USA, 110, 14162–14167, https://doi.org/10.1073/pnas.1303365110, 2013.

Cooper, V. T., Armour, K. C., Hakim, G. J., Tierney, J. E., Osman, M. B., Proistosescu, C., Dong, Y., Burls, N. J., Andrews, T., Amrhein, D. E., Zhu, J., Dong, W., Ming, Y., and Chmielowiec, P.: Last Glacial Maximum pattern effects reduce climate sensitivity estimates, Sci. Adv., 10, eadk9461, https://doi.org/10.1126/sciadv.adk9461, 2024.

Dong, Y., Armour, K. C., Zelinka, M. D., Proistosescu, C., Battisti, D. S., Zhou, C., and Andrews, T.: Intermodel spread in the pattern effect and its contribution to climate sensitivity in CMIP5 and CMIP6 models, J. Clim., 33, 7755–7775, 2020.

Forster, P. M.: Inference of climate sensitivity from analysis of Earth's energy budget, Annu. Rev. Earth Pl. Sc., 44, 85–106, 2016.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D. J., Mauritsen, T., Palmer, M. D., Watanabe, M., Wild, M., and Zhang, H.: The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity, in: Climate Change 2021: The Physical Science Basis, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Chap. 7, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 923–1054, https://doi.org/10.1017/9781009157896.009, 2021.

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

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

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian data analysis, Chapman and Hall/CRC, ISBN 9781439840955, 1995.

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

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., and Wagenmakers, E.-J.: A primer on Bayesian model-averaged meta-analysis, Adv. Method. Pract. Psychol. Sci., 4, 25152459211031256, https://doi.org/10.1177/25152459211031256, 2021.

Hansen, J. E., Sato, M., Simons, L., Nazarenko, L. S., Sangha, I., Kharecha, P., Zachos, J. C., von Schuckmann, K., Loeb, N. G., Osman, M. B., Jin, Q., Tselioudis, G., Jeong, E., Lacis, A., Ruedy, R., Russell, G., Cao, J., and Li, J.: Global warming in the pipeline, Oxford Open Climate Change, 3, kgad008, https://doi.org/10.1093/oxfclm/kgad008, 2023.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., and Wilkinson, R. D.: A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Clim. Dynam., 35, 785–806, https://doi.org/10.1007/s00382-009-0630-8, 2009.

Kageyama, M., Harrison, S. P., Kapsch, M.-L., Lofverstrom, M., Lora, J. M., Mikolajewicz, U., Sherriff-Tadano, S., Vadsaria, T., Abe-Ouchi, A., Bouttes, N., Chandan, D., Gregoire, L. J., Ivanovic, R. F., Izumi, K., LeGrande, A. N., Lhardy, F., Lohmann, G., Morozova, P. A., Ohgaito, R., Paul, A., Peltier, W. R., Poulsen, C. J., Quiquet, A., Roche, D. M., Shi, X., Tierney, J. E., Valdes, P. J., Volodin, E., and Zhu, J.: The PMIP4 Last Glacial Maximum experiments: preliminary results and comparison with the PMIP3 simulations, Clim. Past, 17, 1065–1089, https://doi.org/10.5194/cp-17-1065-2021, 2021.

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

Koricheva, J., Gurevitch, J., and Mengersen, K.: Handbook of meta-analysis in ecology and evolution, Princeton University Press, ISBN 9780691137285, 2013.

Loulergue, L., Schilt, A., Spahni, R., Masson-Delmotte, V., Blunier, T., Lemieux, B., Barnola, J.-M., Raynaud, D., Stocker, T. F., and Chappellaz, J.: Orbital and millennial-scale features of atmospheric CH_4 over the past 800,000 years, Nature, 453, 383–386, 2008.

Mahowald, N. M., Yoshioka, M., Collins, W. D., Conley, A. J., Fillmore, D. W., and Coleman, D. B.: Climate response and radiative forcing from mineral aerosols during the last glacial maximum, pre-industrial, current and doubled-carbon dioxide climates, Geophys. Res. Lett., 33, L20705, https://doi.org/10.1029/2006GL026126, 2006.

Marvel, K. and Webb, M.: netzeroasap/LambdaBayes: Code for "Towards robust community assessments of the Earth's climate sensitivity" (v1.0.0), Zenodo [code], https://doi.org/10.5281/zenodo.13905523, 2024.

Marvel, K., Schmidt, G. A., Miller, R. L., and Nazarenko, L. S.: Implications for climate sensitivity from the response to individual forcings, Nat. Clim. Change, 6, 386–389, https://doi.org/10.1038/nclimate2888, 2016.

Marvel, K., Pincus, R., Schmidt, G. A., and Miller, R. L.: Internal Variability and Disequilibrium Confound Estimates of Climate Sensitivity From Observations, Geophys. Res. Lett., 45, 1595–1601, 2018.

Modak, A. and Mauritsen, T.: Better-constrained climate sensitivity when accounting for dataset dependency on pattern effect estimates, Atmos. Chem. Phys., 23, 7535–7549, https://doi.org/10.5194/acp-23-7535-2023, 2023.

Renoult, M., Sagoo, N., Zhu, J., and Mauritsen, T.: Causes of the weak emergent constraint on climate sensitivity at the Last Glacial Maximum, Clim. Past, 19, 323–356, https://doi.org/10.5194/cp-19-323-2023, 2023.

Rohling, E. J., Marino, G., Foster, G. L., Goodwin, P. A., Von der Heydt, A. S., and Köhler, P.: Comparing climate sensitivity, past and present, Annu. Rev. Mar. Sci., 10, 261–288, 2018.

Rose, B. E., Armour, K. C., Battisti, D. S., Feldl, N., and Koll, D. D.: The dependence of transient climate sensitivity and radiative feedbacks on the spatial pattern of ocean heat uptake, Geophys. Res. Lett., 41, 1071–1078, https://doi.org/10.1002/2013GL058955, 2014.

Sellers, W. D.: A global climatic model based on the energy balance of the earth-atmosphere system, J. Appl. Meteorol. Climatol., 8, 392–400, 1969.

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

Sherwood, S., Webb, M. J., Annan, J. D., Armour, K., Forster, P. M., Hargreaves, J. C., Hegerl, G., Klein, S. A., Marvel, K. D., Rohling, E. J., and Watanabe, M.: An assessment of Earth's climate sensitivity using multiple lines of evidence, Rev. Geophys., 58, e2019RG000678, https://doi.org/10.1029/2019RG000678, 2020.

Siegenthaler, U., Stocker, T. F., Monnin, E., Luthi, D., Schwander, J., Stauffer, B., Raynaud, D., Barnola, J.-M., Fischer, H., Masson-Delmotte, V., and Jouzel, J.: Stable carbon cycle climate relationship during the Late Pleistocene, Science, 310, 1313–1317, 2005.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A.: Bayesian approaches to random-effects meta-analysis: a comparative study, Stat. Med., 14, 2685–2699, 1995.

Snyder, C. W.: Evolution of global temperature over the past two million years, Nature, 538, 226–228, https://doi.org/10.1038/nature19798, 2016.

Stap, L. B., Köhler, P., and Lohmann, G.: Including the efficacy of land ice changes in deriving climate sensitivity from paleodata, Earth Syst. Dynam., 10, 333–345, https://doi.org/10.5194/esd-10-333-2019, 2019.

Sutton, A. J. and Abrams, K. R.: Bayesian methods in meta-analysis and evidence synthesis, Stat. Method. Med. Res., 10, 277–303, 2001.

Tierney, J. E., Zhu, J., King, J., Malevich, S. B., Hakim, G. J., and Poulsen, C. J.: Glacial cooling and climate sensitivity revisited, Nature, 584, 569–573, 2020.

von Deimling, T. S., Ganopolski, A., Held, H., and Rahmstorf, S.: How cold was the Last Glacial Maximum?, Geophys. Res. Lett., 33, L14709, https://doi.org/10.1029/2006gl026484, 2006.