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

Abstract. The eventual planetary warming in response to elevated atmospheric carbon dioxide concentrations is not precisely known. The uncertainty in 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-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 that 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 a Bayesian random effects meta-analysis be used to estimate the evidence and its uncertainty from published literature. Second, we advocate that the organizers of future assessments clearly specify an interpretive model or group of candidate models,in the latter case using 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 a 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 warming. 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:

$$M_0 : \Delta N = \Delta F + \lambda \Delta T. \tag{1}$$

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

$$S \equiv -\frac{F_{2\times CO2}}{\lambda}. \tag{2}$$

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 (**?**, hereafter S20) reduced this range using multiple lines of evidence, but the recent IPCC report **?** 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?

S is determined by the net feedbacks λ at equilibrium and in response to doubled $\mathrm{CO_2}$. While these are unobservable in the current system, in which $\mathrm{CO_2}$ has not yet doubled and which is out of equilibrium, there exist several lines of evidence 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 the evidence of the planet itself, which has been steadily warming in response to net anthropogenic forcing, which includes not just emissions of $\mathrm{CO_2}$ but of other greenhouse gases and aerosols as well. 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 judgement about the uncertainty of physical processes, and the use of different priors on λ and/or S. IPCC 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 attempt to combine multiple priors, estimates, and expert judgements into a single posterior probability distribution. In both IPCC AR6 and S20, as in almost all previous assessments, the means by which disagreements among experts were resolved or handled was 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 in to the analysis and to present a framework for systematically and fairly incorporating the subjective judgements of multiple experts. Ultimately, we seek to create a framework in which expert judgement is incorporated in the form of clearly specified priors.

The paper is organized as follows. In section 2, we review the basic Bayesian analysis framework. Sections 3, 4, and 5 discuss evidence, structural, and prior uncertainty, respectively. In these sections, we use a single line of evidence– paleoclimate data from the Last Glacial Maximum– to illustrate how these sources of uncertainty shape estimates of climate feedbacks and sensitivity. In Section 6 we show how these sources of uncertainty affect constraints derived from multiple lines of evidence. In Section 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, in section **??** we discuss possible generalizations and extensions.

2 Analysis framework

Bayes’ Theorem can be written as

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

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 a forcing ΔF as well as, in non-equilibrium states, the net energy inbalance 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), and Y therefore 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 parameters Θ whose posterior distributions we estimate. For example, in the simple energy balance model denoted M_0, there is only one parameter and $\Theta = \lambda$. The model determines the **likelihood** $P(Y|\Theta,M)$ of observing the data given particular values of the parameters Θ. We discuss methods for calculating this likelihood in Section 3.1.

Prior The **prior** probability distribution $P(\Theta|M)$ reflects prior beliefs or knowledge about the model parameters Θ. For example, in the simple model M_0, the community assessment S20 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 process understanding presented in Section 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 $N(\mu,\sigma)$ with mean $\mu =$-1.30 and standard deviation $\sigma = 0.44$.

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

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

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 assessing ΔT, ΔF, and ΔN for each line of evidence. Second, the analyst must specify a model (and its parameters Θ) in order to interpret that evidence. For example, the model M_0 assumes the feedback parameter is time- and state-independent, and thus estimating it from the past is a reliable guide to the hypothetical future under doubled CO_2. Finally, the analyst must clearly specify her or his 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 climate sensitivity S) given a single line of evidence.

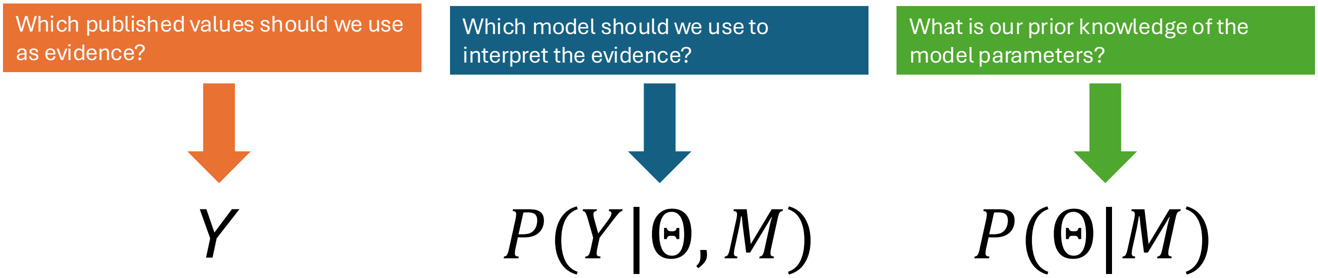
Which published values should we use as evidence?
Which model should we use to interpret the evidence?
What is our prior knowledge of the model parameters?
Y
$P(Y|\Theta, M)$
$P(\Theta|M)$

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

3 Evidence uncertainty

The strongest constraints on equilibrium climate sensitivity in S20 were derived from paleoclimate evidence, and the closest equilibrium climate to the present is the Last Glacial Maximum (LGM) approximately 21,000 years ago. Reconstructions **???????** or model-based estimates **??** of the global mean temperature change ΔT and the radiative forcing ΔF have been used to calculate the global mean feedbacks 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 **??????????**. These are derived from climate models participating in the Paleoclimate Model Intercomparison Project (PMIP **?**) and 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 judgement applied to a literature review to estimate $\Delta T = -5$K with a 95$\%$ confidence interval of (-3.0K, -7.0K). However, a contemporaneous study using a new temperature reconstruction (Tierney et al 2020 **?**, hereafter T20) estimated both colder (mean -6.1K) and less uncertain (with a 95 % highest posterior density interval of -6.5 to -5.7 K) values 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 reliance on a prior derived from a single climate model **?**. 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 T20 in expert assessments in Section 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 in 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 **?** and CO_2 **?** are relatively well-constrained; the forcings from other well-mixed greenhouse gases **?** and ice sheets less so but still informed by proxy and model evidence (Section **??**), and those from dust **??**, other aerosols, and vegetation **?** highly uncertain. While S20 estimated total radiative forcing in the LGM to be N(-8.43, 2) W m^{-2}, T20 use a best estimate of -6.8 W m^{-2} with a 95$\%$ confidence interval of -9.6 to -5.2 W m^{-2}.

Contour lines in Figure 2a show the joint probability distribution (assuming uncorrelated errors) $\rho(\Delta T, \Delta F)$ as reported by S20 (black) and T20 (red). Rather than 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 $(\Delta T, \Delta F)$ to line on a line with slope $-\lambda$. Intuitively, the value of $-\lambda$ that maximizes the likelihood is the slope of the line that passes through through the greatest probability density. These maximum likelihood estimates are shown as straight lines in Figure 2a.

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

$$P(Y|\lambda) \propto \int\limits_C \rho(\Delta T, \Delta F) ds$$

$$C : 0 = \Delta F + \lambda \Delta T$$

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 1). Otherwise, the integral can be computed numerically. The resulting likelihood functions are shown as thick lines in Figure 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 the 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 $\lambda = -1.7$ $\mathrm{Wm^{-2}K^{-1}}$ (thick black line, Figure 2b) with a 5-95% range of (-3.37, -1.09) $\mathrm{Wm^{-2}K^{-1}}$. Using T20 evidence, the most likely value is $\lambda = -1.1\,\mathrm{Wm^{-2}K^{-1}}$ (thick red line, Figure 2b). The 5-95% range is (-1.49, -0.87) $\mathrm{Wm^{-2}K^{-1}}$.

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 2). This neglects the small correlation between ΔF and the forcing at doubled $\mathrm{CO_2}$, but this simplification does not substantially affect our results (Appendix 3).

Using S20 evidence from the LGM, we find a 5-95% range of (1.17K, 3.69K) 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 (2.61K, 4.72K).

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.**????**) suggest that M_0 might not be appropriate for past climates due to the dependence of the feedbacks on the background

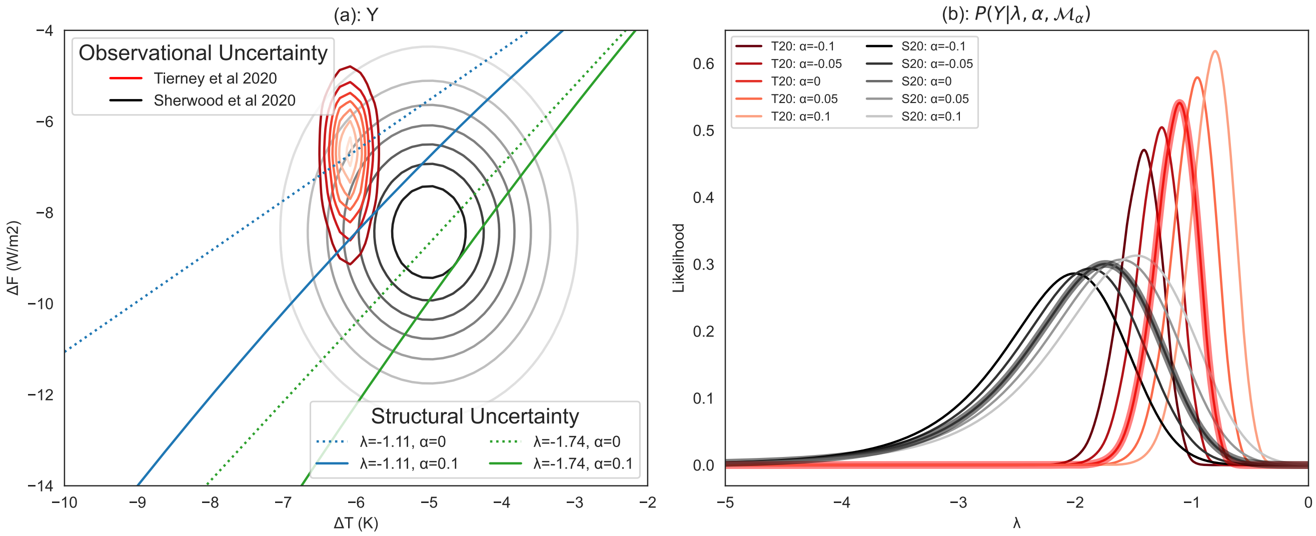

Figure 2. Panel a: joint evidence distributions for ΔT and ΔF used in Sherwood et al (black contours) and Tierney et al (red contours). Structural uncertainty is illustrated using solid lines (corresponding to fixed values of λ using the model M_0) and dashed lines lines (corresponding to fixed values of λ and α using the model M_α). b: Likelihoods as a function of λ and given S20 (black lines) or T20 (red lines) evidence and different values of the state dependence α. b: Resulting likelihoods for λ given the evidence from S20 (black) and 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.

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 one. To model this background temperature dependence, we might use an alternate model that includes a second-order term in the radiative response

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

where $\alpha = \partial\lambda/\partial(\Delta T)$ is an additional parameter reflecting the background state dependence **????**. 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 $\mathrm{Wm^{-2}K^{-1}}$ at a cooling of -5 K), the most likely value of λ is not the same as the most likely value of λ assuming $\alpha = 0$ (dotted and solid lines, Figure 2a). In this case, the likelihoods (Figure 2b) are calculated by integrating the joint probability distribution for ΔT and ΔF along the curve defined by Eq. 4, and 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 of 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 in the LGM and the pattern of warming expected at elevated CO_2 concentrations (e.g. **?**).

Regardless of the interpretive model used, it is both required for analysis and subjectively chosen by the analyst. Different reasonable analysts might make different choices about the model to use. 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 given a single model, for example 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 Section 5.

5 Prior uncertainty

Once a model is specified, we would like to use the evidence to tell us something about its parameters Θ. Bayes' theorem says that the posterior distributions of the parameters are simply obtained by multiplying the likelihood by prior probability distributions reflecting our pre-existing beliefs and/or knowledge. These priors incorporate expert judgement, 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, Figure 3a shows the joint posteriors for the feedbacks λ and the state dependence 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 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 W m^{-2} K^{-1} for 5K 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 α, reflecting the current state of the literature. This prior substantially constrains the resulting joint posterior distribution (Figure 3 b). 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., which imply a lower sensitivity in the LGM than at doubled CO_2) receive more posterior weight. Combining the informative priors on both λ and α further constrains the joint posterior (Figure 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 twenty thousand years to gauge the planet's

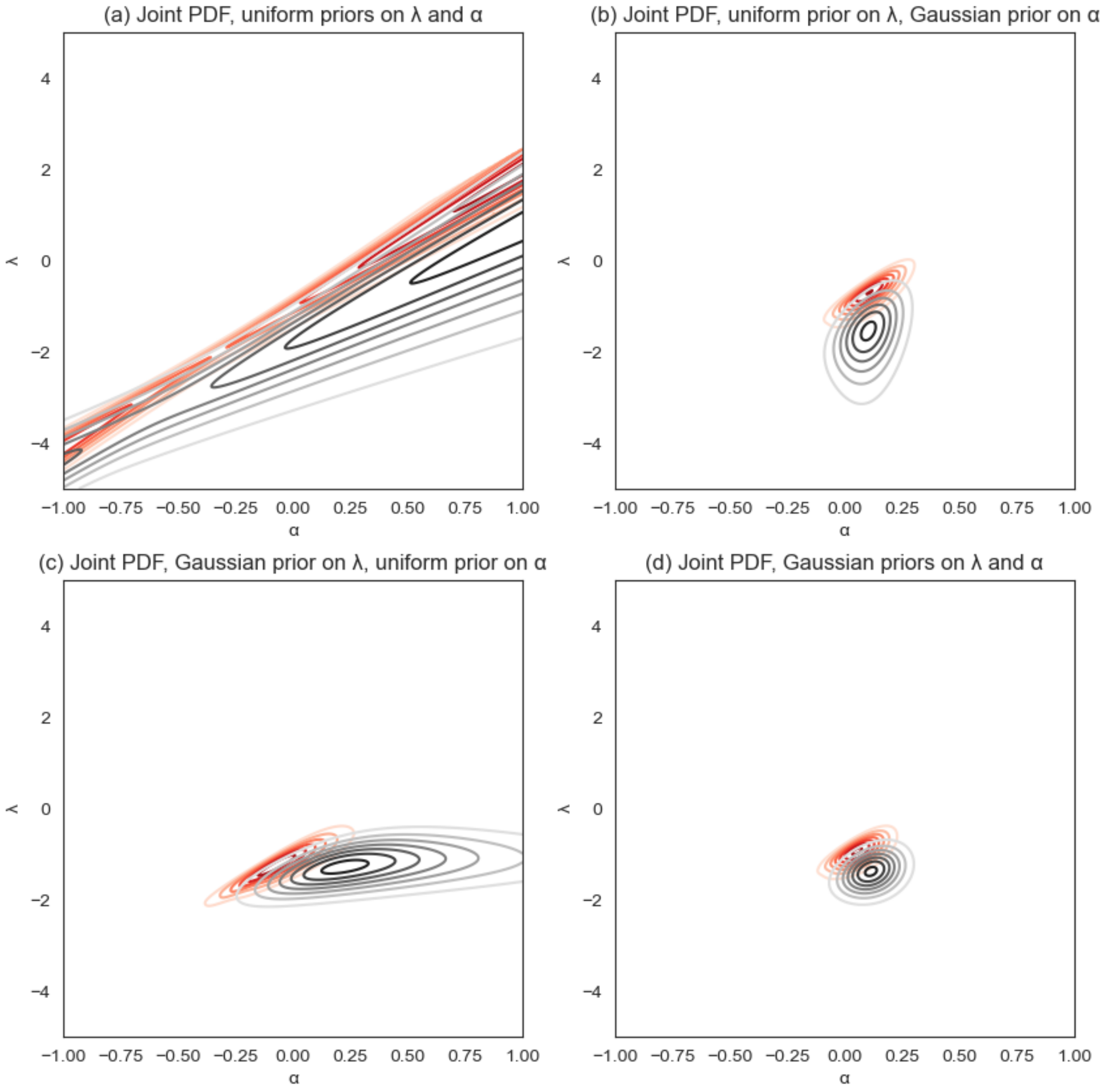

Figure 3. Joint posteriors for the feedback parameter λ and the state dependence α under different priors: a Uniform priors on both parameters b Uniform prior on λ, Gaussian prior from expert judgement of published literature (used in S20) on α (c) Gaussian prior from process evidence (used in S20) on λ, uniform prior on α (d) Gaussian priors (from S20) on both.

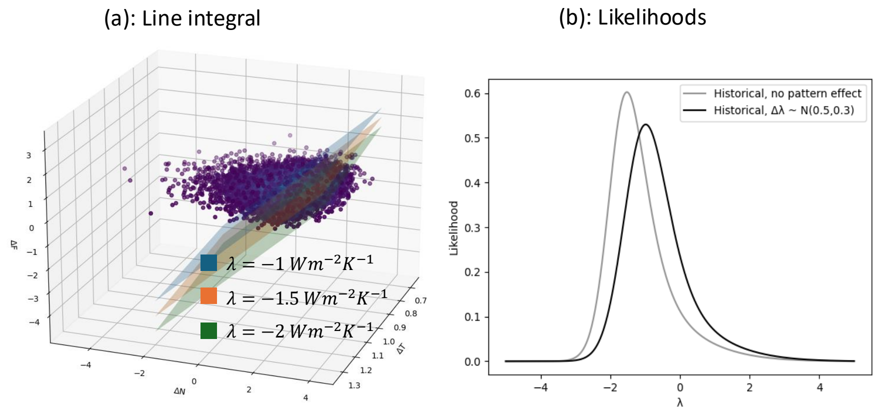

Figure 4. a: Calculating the likelihood of observing the historical evidence used in S20 for a putative value of λ. Each value of λ defines a plane; shown are $\lambda = -1\text{Wm}^{-2}\text{K}^{-1}$ (blue),$\lambda = -1.5\text{Wm}^{-2}\text{K}^{-1}$ (orange) and $\lambda = -2\text{Wm}^{-2}\text{K}^{-1}$(green).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

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 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 **?**.

6.1 Historical likelihood

In this three-dimensional joint probability space, the simplest energy balance model M_0 defines a plane rather than a line in evidence space (Figure 4), and the likelihood of the evidence given λ is proportional to the integral over this surface. Figure 4 shows the historical evidence reported in S20, in which

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

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

and ΔF is calculated using unconstrained aerosol ERFs from **?** with median 1.83 W m^{-2} and 5-95% range (-0.03, 2.71) W m^{-2}. The gray line in Figure 4 shows the resulting likelihood as a function of λ. The maximum likelihood value is $\lambda = -1.53\text{Wm}^{-2}\text{K}^{-1}$.

However, the simplest energy balance model M_0 assumes the feedback parameter is the same for climate changes in the deep past, the transient historical period, and the future. Many studies (e.g. **????????**) 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 the feedbacks expected in response to the long-term equilibrium warming pattern:

$$M_{\Delta\lambda} : \Delta N = (\lambda - \Delta\lambda)\Delta T + \Delta F$$

S20 placed a Gaussian prior on this pattern effect $\Delta\lambda$ =N(0.5 , 0.3) W $\mathrm{m}^{-2}\mathrm{K}^{-1}$. This corresponds to a modification of the tilt of the plane in Figure 4a. Because this model assumes the pattern effect is linearly additive, no further curvature is introduced. By multiplying the joint likelihood $P(\Delta Q, \Delta T, \Delta F|\lambda, \Delta\lambda)$ by this 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 Figure 4b. The inclusion of the additive pattern effect and our physics-informed intution that it is likely to be positive shift the most likely value of the feedback parameter to $\lambda = -1.0\mathrm{Wm}^{-2}\mathrm{K}^{-1}$.

The pattern effect estimate used in S20 was based on the Atmospheric Model Intercomparison Project II (AMIPII) dataset, which produces the largest estimate of the pattern effect **?**, and therefore the priors on $\Delta\lambda$ used there may be both overconfident and too strongly weighted 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

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

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 for the historical evidence as a function of λ. The light blue line shows the likelihood for the S20 LGM evidence as a function of λ, assuming no state dependence ($\alpha = 0$). The product of these likelihoods is shown as a green dashed line. The less the historical and paleo likelihoods overlap, the narrower the posterior will be. We refer to this conundrum as the "Twin Peaks" problem: should larger 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 by using an alternate model for the paleo evidence. Note that the posterior for λ shown in Figure 5a is conditional on a model M_0 for the paleoclimate evidence that 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 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 paleo data given M_α and Gaussian priors on α is shown as a dark blue line in Figure 5b. While the overlap between these two distributions is far from exact, it is substantially larger than for the no-state-dependence case illustrated in Figure 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. When using T20 evidence, however, there is considerable overlap between the historical (with pattern effect) and paleo (with no state dependence) likelihoods. As in the top two panels, the black lines in Figure 5c and d show the historical likelihood. The likelihood for λ obtained from T20 evidence and assuming no state dependence (orange line, Figure

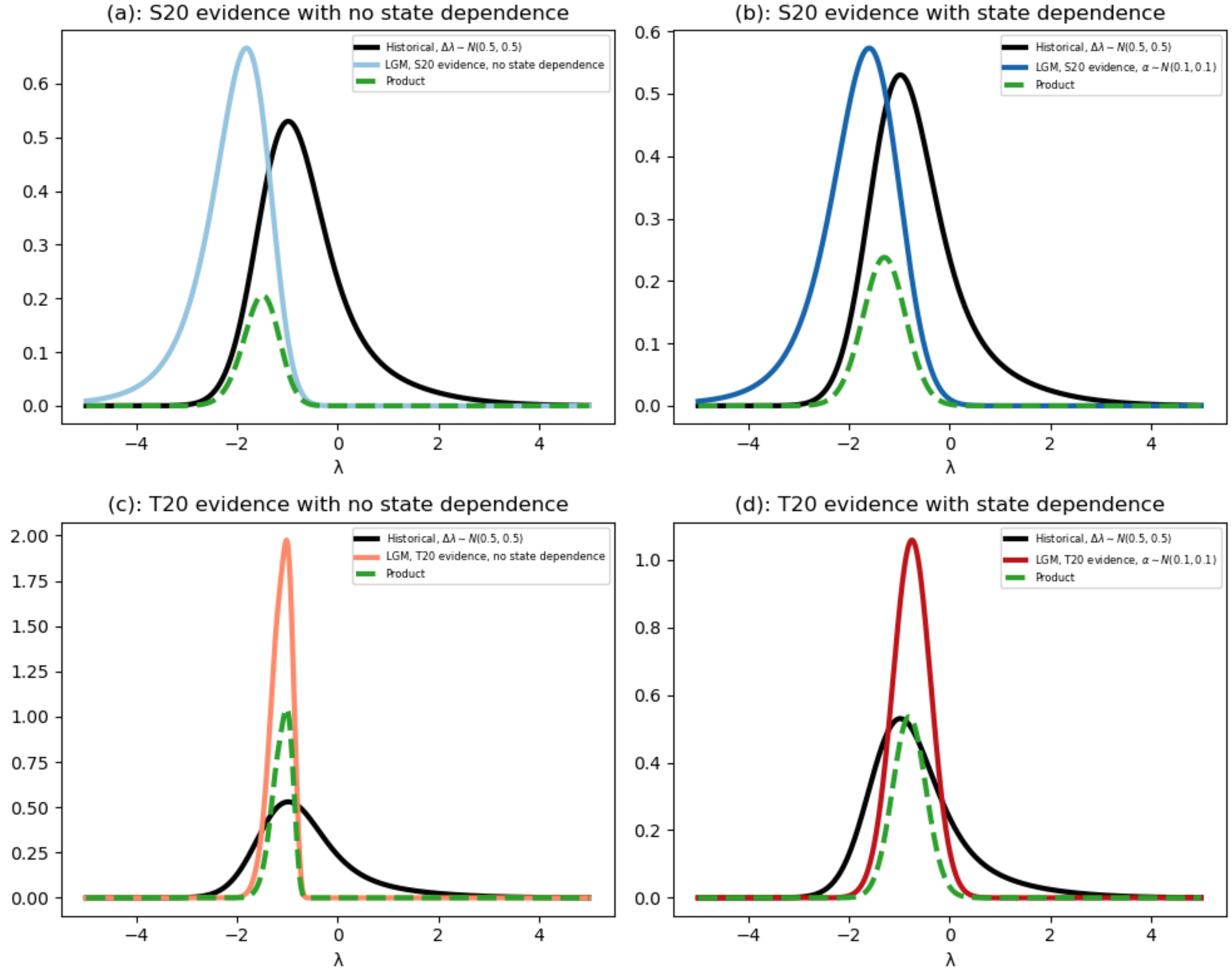

Figure 5. Likelihoods from multiple lines of evidence. In all four panels, the black line shows the likelihood for the historical evidence given λ and assuming a pattern effect $\Delta\lambda \sim N(0.5, 0.3)$. a: Likelihood of S20 evidence given λ assuming no state dependence in the LGM (light blue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

5c closely overlaps the historical likelihood, as does the likelihood assuming state dependence with a prior on α as in S20 (red line, Figure 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 the 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

$$\mathrm{odds} = \frac{\mathrm{P(M_\alpha|Y_{hist}, Y_{paleo})}}{\mathrm{P(M_0|Y_{hist}, Y_{paleo})}}$$

$$= \frac{P(Y_{hist}, Y_{paleo}|M_\alpha)P(M_\alpha)}{P(Y_{hist}, Y_{paleo}|M_0)P(M_0)}$$

$$\equiv BF \times \frac{P(M_\alpha)}{P(M_0)}$$

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 all values of its parameters Θ_ℓ:

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

This reflects the probability that 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 model M_0 is

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

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

By contrast, the model evidence for the model M_α is

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

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

Using S20 evidence and these priors, we find that the Bayes Factor is 1.33. This means that if our prior is that both models are equally likely, the evidence shifts those odds: the model depicted in panel b is about 33% more likely to have generated the observed paleo 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 knowlege 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 places where unavoidable subjective decisions must be made: collecting evidence, choosing the interpretive model, and 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 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 **?**, frequently used in fields as diverse as psychology **?**, medicine **?**, and ecology**?**. 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 the true (latent) mean y_j of each study is normally distributed about an overall mean Y, with τ the expected inter-study standard deviation.

The priors we put 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 very close to zero suggests homogeneity across 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 CMIP model, whereas a random-effects model might be more appropriate for combining ensembles of multiple CMIP models, which we know to differ structurally.

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

Mean (K)	Standard Deviation	Reference	Derived From	Generation
-4.00	0.41	**?**	Proxies and models	N/A
-5.80	0.77	**?**	Proxies and models	N/A
-6.20	0.46	**?**	GENIE-1	N/A
-3.58	0.12	**?**	Proxies	N/A
-6.20	0.92	**?**	Proxies and models	N/A
-6.30	0.61	**?**	Proxies (ocean temperature) and models	N/A
-5.70	0.20	**?**	N/A	N/A
-5.75	0.38	**?**	SST proxies and a model simulation	N/A
-6.10	0.20	**?**	proxies and isotope-enabled climate model	N/A
-5.00	1.00	**?**	Synthesis	N/A
-4.85	N/A	**?**	CESM	PMIP3
-2.70	N/A	**?**	CNRM	PMIP3
-4.63	N/A	**?**	FGOALS-g2	PMIP3
-4.92	N/A	**?**	GISSE2-p1	PMIP3
-5.19	N/A	**?**	GISSE2-p2	PMIP3
-4.64	N/A	**?**	IPSL	PMIP3
-5.40	N/A	**?**	MIROC	PMIP3
-4.41	N/A	**?**	MPI-p1	PMIP3
-4.67	N/A	**?**	MPI-p2	PMIP3
-4.71	N/A	**?**	MRI	PMIP3
-3.75	N/A	**?**	AWIESM1	PMIP4
-3.81	N/A	**?**	AWIESM2	PMIP4
-6.80	N/A	**?**	CESM1-2	PMIP4
-7.16	N/A	**?**	HadCM3-PMIP3	PMIP4
-5.92	N/A	**?**	HadCM3-ICE6GC	PMIP4
-6.46	N/A	**?**	HadCM3-GLAC1D	PMIP4
-3.28	N/A	**?**	iLOVECLIM-ICE-6G	PMIP4
-3.26	N/A	**?**	iLOVECLIM-GLAC1D	PMIP4
-3.73	N/A	**?**	INM-CM4-8	PMIP4
-4.63	N/A	**?**	IPSLCM5A2	PMIP4
-4.02	N/A	**?**	MIROC-ES2L	PMIP4
-3.90	N/A	**?**	MPI-PMIP4	PMIP4
-5.27	N/A	**?**	UT-CCSM4	PMIP4

Table 1. Estimates of global cooling ΔT during the Last Glacial Maximum

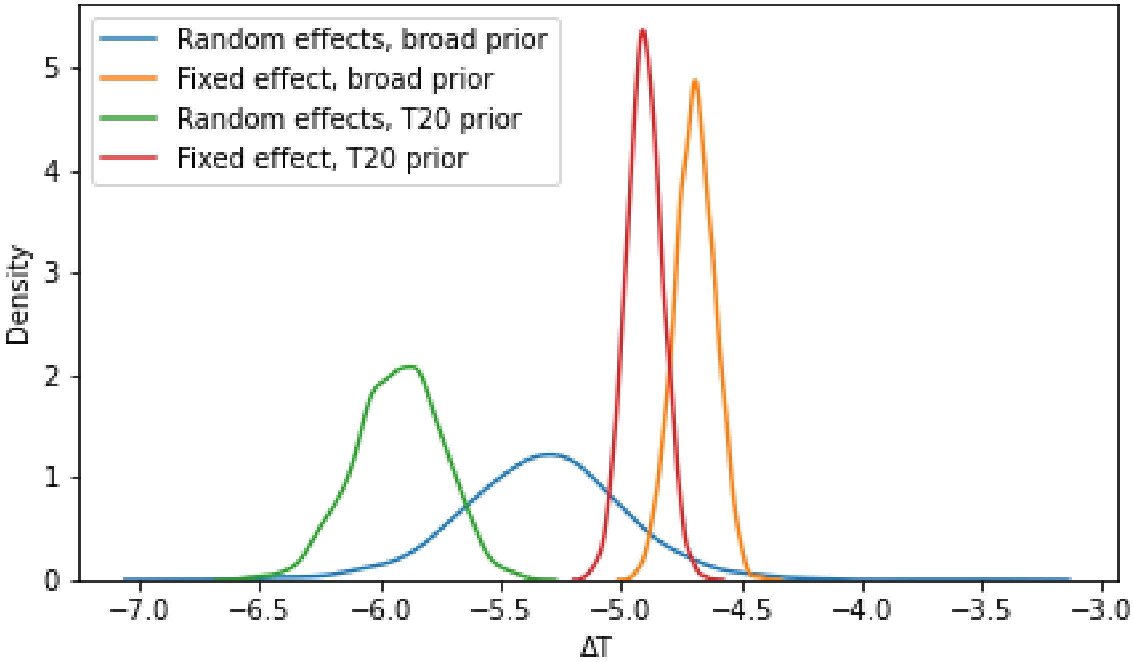

Figure 6. How cold was the Last Glacial Maximum? The answer depends on your prior beliefs about the cooling and about the literature. Shown are posterior distributions for the LGM cooling ΔT assuming a random effects model and broad (blue line) or T20 (green) priors on the mean or a fixed effects model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

Figure 6 illustrates how the posterior distribution of ΔT depends on prior beliefs about the nature and quality of the published literature assessing it. Consider, for example, a random-effects model in which we place broad priors on the mean $\mu \sim N(0,100)$ and 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 (-5.9K, -4.8K). Assuming that there is *no* inter-study spread (i.e, τ is assumed to be zero with zero uncertainty: a fixed effect model) would yield an estimate of ΔT 90% likely to be between -4.8 and -4.5K. This much narrower (and warmer) estimate results from the extremely restrictive prior belief that every study, regardless of method, targets the same underlying ΔT 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 (say, for example, ΔT from T20). Combined with a broad uniform prior on the inter-study spread, this results in an 90% posterior density estimate of (-6.2K, -5.6K). 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 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.

7.2 Handling model uncertainty

As shown in Section 4, the constraints placed on climate sensitivity by multiple lines of evidence evidence depend on the model(s) used to interpret that evidence. This means that the design of every expert assessment must be explicit about its interpretive models. As the assessment is planned, it is crucial to arrive at consensus on credible interpretive models for the evidence. For example, one possible model for the Last Glacial Maximum might incorporate parameters α (representing state dependence), ξ (representing the difference between long-term equilibrium LGM feedbacks and the target quasi-equilibrium feedbacks to doubled CO2) and $\Delta\lambda_{LGM}$ (representing radiatively important sea-surface pattern differences between the LGM and doubled CO2):

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

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 a model, s/he would be free to set a prior very narrowly clustered around 0 on that prior.

If consensus cannot be reached on a particular model, then we suggest that the planning team 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:

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

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

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

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

7.2.1 Recommendations

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

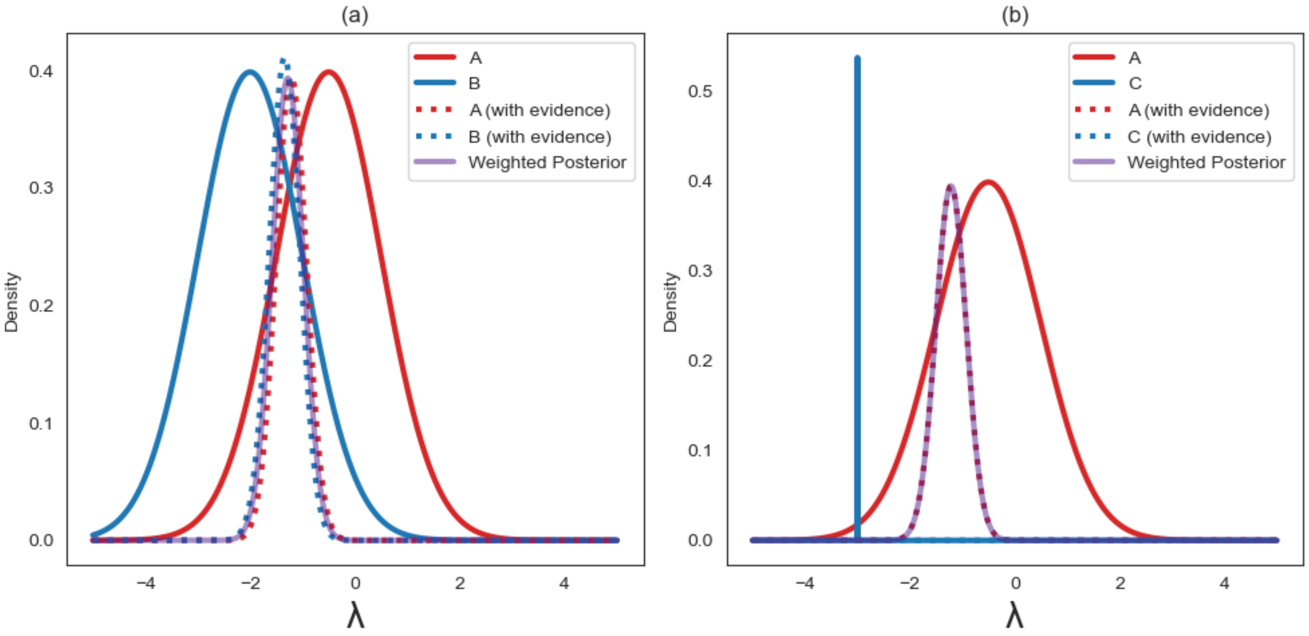

Figure 7. a: Experts A (solid red line) and B (solid blue line) begin with different priors on λ. The evidence presented in S20 updates these priors, and the resulting posteriors are nearly identical (dotted red and blue lines). The purple line shows the weighted posterior. a: Experts A (solid red line) and C (solid blue line) begin with different priors on λ, but 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 A's posterior.

7.3 Expert elicitation via priors

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

However, we require consistent ways to aggregate the judgement 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 the feedback parameter to be less negative than Expert B (solid blue line) and is even open to the idea that it might be positive. Dashed red and blue lines show both experts' posteriors, when 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 our researchers A and B. Expert C (blue line, Figure 7b believes the feedback parameter to be strongly negative. Moreover, s/he is extremely confident in this: his/her prior distribution is very narrowly peaked around a value of $\lambda = -3Wm^{-2}K^{-1}$. Expert C's confidence remains unshaken by the evidence presented in S20, and his/her posterior remains nearly identical to his/her prior beliefs. How should an assessment handle such excessively confident experts, whose beliefs appear to be unshakeable by any reasonable amount of evidence?

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

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

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

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

where

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

This method introduces N new parameters: the prior weight a_i we assign to each expert's judgement. This is a far easier task than setting priors on models (as discussed in Section 7.2) because it requires no physical understanding, only 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 weights become

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

The purple line in Figure 7a shows the resulting aggregate posterior given A and B's priors. Because both these experts are similarly able to update their priors, the weighting process has no effect on the outcome. However, the weighted average of A and C's posteriors, shown as a purple line in 7b, is similar to A's posterior distribution. The narrowness of C's prior causes his/her 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 judgement in a systematic way by allowing experts to specify priors on pre-determined 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 in to estimates of climate sensitivity. First, what evidence are we using to constrain climate sensitivity, how do we decide what counts as “evidence", and how should we handle estimates that disagree or conflict? Second, what interpretive model should we be using to relate the evidence to the climate sensitivity, and what parameters are required? Third, what prior knowledge of these parameters is it appropriate to include? We then propose a strategy to make the role of expert judgement 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, GCM, or paleoclimate studies as evidence; where we draw the line between evidence and parameter 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 they interpret and use that evidence depends on the subjective choices they make: what counts as a "study" or "evidence"? How should we best compare estimates derived from proxies or observations and estimates from GCMs? Should some studies receive more weight than others? In our framework, experts must make the following judements about the evidence:

1. What is your informed belief about the evidence? (E.g. 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 would 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, not 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.

Code availability. The code for this project is available at https://github.com/netzeroasap/LambdaBayes/

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 defined by the model :

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

C can be parameterized as

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

and the integral is then

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

In the case where ΔT and ΔF are Gaussian and independent with means μ_T, μ_F and standard deviations σ_T, σ_F respectively, the likelihood has an exact analytic form, substantially speeding up its computation:

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

where

$$\begin{aligned} C &= \frac{\sqrt{1+\lambda^2}}{2\pi\sigma_T\sigma_F} \exp\left(\frac{\mu_T^2}{\sigma_T^2} + \frac{\mu_F^2}{\sigma_F^2}\right) \\ A &= \frac{1}{\sigma_T^2} + \frac{\lambda^2}{\sigma_F^2} \\ B &= \frac{\mu_T}{\sigma_T^2} - \frac{\lambda\mu_F}{\sigma_F^2} \end{aligned}$$

In the case of a three-dimensional space (as for the historical evidence), the curve C defines a plane, not a line, and we have

$$P(Y|\lambda) \propto \int\limits_C \mathcal{J}(\Delta T, \Delta F, \Delta N) dS = \int\int \mathcal{J}\left(\boldsymbol{r}(u,v)\right)||r_u \times r_v||du\,dv \tag{A5}$$

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 question, "a: 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 "b: 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 a line integral over the curve $y = -\lambda x$:

$$a : P(x,y|\lambda) = \int\limits_C P_{xy}(x,y)ds = \int\limits_{-\infty}^{\infty} P_{xy}(x, -\lambda x)\sqrt{1+\lambda^2}\,dx$$

and the ratio distribution of the random variable $\lambda = -y/x$

$$b : P_\lambda(\lambda) = \int\limits_{-\infty}^{\infty} P_{xy}(x, -\lambda x)|x|\,dx$$

We use the ratio distribution b to estimate S once we have the posterior PDF for λ. This is because we treat S as the ratio of two random variables F_{2xCO_2} and λ.

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

CO_2 emissions are the primary contributor to present-day radiative forcing change relative to preindustrial. Atmospheric concentrations of CO_2 were lower in the Last Glacial Maximum. This means that the forcing terms ΔF used as evidence in the LGM and historical periods are correlated with the forcing corresponding to doubled 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

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

In this case, the likelihood $P(E|\lambda, F_{2\times CO_2})$ 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 variables and marginalizing over $F_{2\times CO_2}$

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

Practically, we can draw samples of λ and $F'_{2\times CO_2}$ 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.

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 authors declare no competing interests.

Acknowledgements. We are grateful to John Eyre and to three anonymous reviewers whose comments on the manuscript led to substantial improvements. Additionally we acknowledge James Webb for suggesting the name for the "Twin Peaks problem". Mark Webb was supported by the Met Office Hadley Centre Climate Programme funded by DSIT. Kate Marvel was supported by the NASA Modeling, Analysis, and Prediction program.