Dear Editor,

We now provide a revised version of our manuscript, after some unfortunate delay whom we wish to apologise for. We hope the revisions will bring more clarity to our paper, as requested by the reviewers, and we also managed to expand our ensemble of PMIP4 models, which, we think, makes our paper more suitable for the PMIP4 special issue.
Additionally, we implemented in Appendix A a conjugate prior approach. Reviewer #2 has criticised the use of the complex Hamiltonian Monte Carlo method in our study, and although we do think that the flexibility given by the Markov Chain Monte Carlo methods is essential for our framework, we show that using and Inverse Gamma on sigma (and making the whole problem conjugate) leads to similar results for the posterior sensitivity.
In the following, the original comments from the reviewers (R) and short comments (C) are written in blue, while our replies (A) are in black.

Sincerely,
Martin Renoult

– – – – – – – – – – – – – – – – – – – –

Reviewer #1

R: In this paper the authors develop a novel technique to combine emergent constraints. Their main step forward is reconsidering the emergent constraint regression as a likelihood model so that it can be combined with a prior, allowing for Bayesian updating. This is particularly important for estimates of climate sensitivity, whose IPCC range has barely changed since 1990, even though independent lines of evidence have strengthened. The technique is elegant, transparent and I wish I'd come up with it. The accompanying code is also clear.
I suggest the authors clarify some of their text and if available include more PMIP4 models.

A: During the revision process, 4 models (INM-CM4-8 and AWI-ESM for the LGM, EC-EARTH3.3 and CESM2 for the mPWP) shared their outputs either on ESGF or directly with us, and therefore have been added to our study. Consequently, numbers, figures and list of authors have been changed. There were some interesting consequences in adding these models, as it reduced the correlation in the case of the LGM and increased it in the case of the mPWP. The overall storyline remains similar as during the first version of this paper. However, some details needed to be changed, in particular regarding the inclusion of PMIP4 / PlioMIP2 with the previous ensembles. These additions made us realise some typos in the Tables and Figures regarding some models: Mostly, some temperatures were erroneous (by only a few decimals, but the correct values were used in the code so there is no impact on the computations) and 2 models were wrongly plotted on the zonal mean plot (Fig. 1). This has been corrected in the revised version.

R:
Minor comments:
11: it's not a 100% clear whether this is a combination of the restricted ensemble of the nonrestricted ensemble. Either clarify, or remove the unrestricted estimate altogether.

A: The words "using the restricted ensemble" have been added, as we considered this estimate rather important in the paper.

R: 16: I don't quite understand the last half of the sentence: "higher bound by construction"

A: The whole sentence has been changed to "An interesting implication of this work is that OLS-based emergent constraints on ECS generate tighter uncertainty estimates, in particular at the lower end, an artifact due to a flatter regression line in case of lack of correlation."

R: 104: I didn't quite understand what "percentage of intervals to contain .." means. Please clarify.

A: Changed to "The former is the representation of the number of random intervals to contain the true interval bounds (at 90% confidence, this would lead to 90 out of 100 random intervals to contain the true bounds), while Bayesian credible interval is an interval which we believe (with the given probability) to contain the truth."

R: 126: typo: roles
139: “observation operator”. Operator is unnecessary jargon.

A: Both removed.

R: 169: A two line explanation of a (one step) Karman filter might benefit readers.

A: More details on the one-step process has been written, and the Kalman Filter section has been expanded.

R: 182: Phase 4 of PMIP are used in the study. Please replace explanation by saying not much data is available instead of none.

A: Added.

R: 236-237: I don’t think it’s necessary to include this test any more.

A: The other MCMC tests have been removed.

R: 291: I’m quite surprised that OLS is more tight. Could you check code or provide an explanation?

A: The following explanation has been added: “As previously argued for the combination of PMIP2 and PMIP3, the OLS produces a tighter posterior range. In the absence of a correlation, the Bayesian method relaxes to the prior, whereas the OLS method is heavily influenced by the range of the ensemble. However, we emphasise that this does not suggest that either range is closer to reality.”

Additionally, the number coming from OLS-based approaches have been updated to their predictive intervals values (minor changes which make them more consistent with previous studies).

R: 349-350: a logical extension of the methodology is to apply it to CMIP, where we find many emergent constraint on the same models. It would be nice if the authors could comment on whether they see this as a problem, given that these models may have similar systematic biases.

A: This is an interesting point which could lead to promising research in the future! The following text has been added in “Combining multiple constraints” : “A logical extension of the approach would be to apply it to the ensemble of models of CMIP, where multiple emergent constraints exist for the same models. In theory, this should be possible as long as the investigated relationships are physically plausible. This goes beyond the scope of our study, which uses the paleoclimates as an example for the method, and is left for future research.”

R: 337: merely → nearly or almost.

A: Changed to “nearly”

R: 374: add ‘in a systematic way’ or something similar. The principle behind emergent constraints relies on the fact that models deviate from reality, so that’s not the problem.

A: Added.

R: 386: pertinent → why not use simpler word such as relevant.

A: Changed to “relevant”.

R: 406: ordinary least squares doesn’t require capitalization

A: Changed in the whole manuscript.

R: Fig1 caption: what is a ‘wide’ ensemble proxy?

A: Changed to “multi-proxy ensemble”

R: Fig2 – Fig9: in the pdf, the colour orange might imply to a tired reader that only PMIP3 is used from the figure on the left. Purple or other dark colour might be more clear. I’m not convinced that all figures are necessary for the paper. The summary in the table may suffice for more regressions, such as the one in Fig 9.

A: The colours were updated in the corresponding figures to have a more distinct differences. Fig. 7 (“Latest version approach”) and Fig. 9 (“Model inadequacy”) were removed as it is summarised already in Table 1.

– – – – – – – – – – – – – – – – – – – –

Reviewer #2

R: The paper by Renoult et al presents a new Bayesian method for dealing with emergent constraints for estimating climate sensitivity from palaeoclimate model simulations. I have little expertise in the use of emergent constraints so I will concentrate my comments on the statistical methodology used. For such a simple approach they have made their technique remarkably opaque. For this reason it is hard to recommend an editorial decision for this paper - I will leave it up to the other reviewers to determine novelty and suitability for this journal. However I think there needs to be a considerable improvement in the explanation of the mathematical approaches.
If we start with the OLS method, we have a data set Si,Ti, for i = 1,...,n simulators where we use the model:
$Si = \alpha * Ti + \beta + \varepsilon$
And obtain estimates of alpha, beta, and the residual standard deviation sigma. A user comes along and provides us with a new value $T*$ and we obtain $S*$ from the fitted model. Uncertainty arises from the potential uncertainties in the estimates of the parameters, and the choice of whether prediction or confidence intervals are used.
So far so good. The authors point out that the Bayesian approach is often superior to these traditional models because of its more sensible handling of uncertainty and the allowance of brining in external information in the form of prior distributions. I agree totally.
Unfortunately here is where things get a little more confusing. The authors then state that the model they want to fit is:
$p(S|T) = p(T|S)p(S)/p(T)$,
i.e. a standard application of Bayes’ theorem which provides us with a posterior distribution of S given T. This is where the notation starts to get into a bit of a mess, because now we’re not told where the observations fit in to the model. My guess is that what the authors mean in the above equation (using my notation) is actually:
$p(S*|T*) = p(T*|S*)p(S*)/p(T*)$
Where the likelihood p(T*|S*) is actually integrated over the posterior set of parameters from a new linear regression model
$Ti = \gamma * Si + \delta + \gamma$
Where I’ve named these new slopes/intercepts differently to highlight the different from the previous OLS approach.
This is a more complicated model, and most of has come from guesswork because the authors haven’t provided enough information for me to work out exactly what is happening. I’d really appreciate the authors doing (the quite large job) of either clearing up their maths or making sure that my incorrect assumptions are not made by others.

A: First, as suggested in the last equation, the parameters of the OLS-based approach have been given different names, γ, δ and ζ, to emphasise the differences with the Bayesian framework, and also because their values are, indeed, different from α, β and ε.

A source of confusion, as pointed by Reviewer #2, were the annotations and the model not being explicit enough. We decided to change the name of the section “Bayesian Linear Regression” to “Bayesian Framework” to emphasise the 2-step Bayesian process we are using in the paper. Specifically, the Bayesian updating process which aims at computing the final posterior S is

$p(S|T^o_{tropical}) = p(T^o_{tropical}|S)p(S)/p(T^o_{tropical})$

The value $T^o_{tropical}$ is then a reconstruction from proxy data of the tropical temperature of either the LGM or the mPWP and allows to compute a posterior S in the emergent constraints framework. As noted by Reviewer #2, here, the likelihood $p(T^o_{tropical}|S)$ (noted $P(T^* \mid S^*)$ by Reviewer #2) was an integration over the posterior S of the set of parameters α, β and ε, coming from the second Bayesian process, the Bayesian Linear Regression. This BLR process is constrained on the model ensemble and in particular their values S and $T_{tropical}$. Thus, the BLR model is defined as:

$T_{i,tropical} = \alpha * S_i + \beta + \varepsilon, \varepsilon \sim N(0,\sigma^2)$

Where the "i" refers to the index of one of the climate model used. In this way, we create a likelihood (the posterior of the BLR) $P(T_{tropical} \mid S)$ by integrating over S, for the Bayesian updating process. We explicitly stated that $P(T_{tropical} \mid S)$ is approximated as $P(T_{tropical} = T^o_{tropical} \mid S)$ through an importance sampling method, allowing us to insert the proxy reconstruction and create a series of weights to update the prior P(S) into the posterior of interest, $P(S \mid T^o_{tropical})$.

We hope this clarification will make the comprehension of our method easier than before. We decided to explicitly split the 2 Bayesian processes, as the BLR is only used to compute the likelihood necessary for the Bayesian updating. The Bayesian updating, in fact, is the only way of computing the climate sensitivity; Our implementation of emergent constraints in a Bayesian framework, however, requires both processes.

R: The paragraph in the intro which starts "Two recent papers have also addressed. . ." makes some odd statements about KFs. It points out that everything is Gaussian then states that "it is fairly difficult to generate posterior values which are outside of the prior range". This seems surprising if everything is Gaussian. I haven't read the other paper so perhaps explain more clearly?

A: It has been changed to: "In particular, most of the posterior values would lie in the range covered by the ensemble of models if the observed value is either uncertain and/or close to the prior mean. This is a direct consequence of the joint probability distribution produced by the Kalman filter, which in the case of joint Gaussian distributions, will produce a tighter posterior Gaussian distribution. Because of that, it does not appear to correspond to the choice which is usually made, albeit implicitly."

R: The first sentence in the methods section involves, a load of unnecessary commas, which, in my view, makes the sentence, and hence, the definition, of the key concept, of emergent constraints, very hard to understand. There must be a simpler way of writing it.

A: Changed to: "The general method of "emergent constraints" seeks a physically plausible relationship in the climate system between two model variables in an ensemble of results from different climate models. Consequently, an observation of one measurable variable (such as past tropical temperatures) could be used to better constrain the other investigated variable, usually unobserved and difficult to measure (such as climate sensitivity)."

Additionally, the following has been added later on in the same paragraph, as we think it is a source of confusion for our approach: "Although the unobserved variable is usually taken as a future variable, the emergent constraints theory can be used with two variables within the same time frame, as long as the relationship is plausible."

R: Also in that sentence it says '. . .'then an observation . . .'. An observation of what?

A: Changed to "an observation of one measurable variable (such as past tropical temperatures) could be used to better constrain the other investigated variable, usually unobserved and difficult to measure (such as climate sensitivity)."

R: L95 should be N(0, σ2) to match standard notation. This mistake is made throughout. There's similarly a bizarre use of ∈ from set theory to write ε ∈ N(0,σ) which I think should be ε ~ N(0,σ2) everywhere.

A: We matched standard notation by writing N(0, σ^2) in the text. We also wrote ε ~ N(0,σ2) in the text.

R: There seems to be a kind of deeper issue that perhaps should be mentioned somewhere that these regression approaches really should involve measurement error (separate from model error as in the Williamson/Samson) paper. The literature on this is well-developed and is pretty easy to include in Bayesian models.

A: The following text has been added regarding observational errors. In particular, observational errors have been shown to have insignificant impact on the results in the case of this emergent constraints in a previous study on the mPWP (Hargreaves and Annan, 2016):

"It is not clear if observational errors have always been adequately accounted for in previous emergent constraints research. Our approach provides a natural framework for this, as the likelihood can include the uncertainty of the observational process as we have done. However, we have ignored uncertainties in the calculation of the model values of S and T_tropical as, while they are poorly quantified, we believe them to be too small to materially affect our result. In fact, it has been argued for the case of the mPWP that observational errors on S and T_tropical are small compared to the structural differences responsible of the dispersion of the points around the regression line and thus can be neglected (Hargreaves and Annan, 2016)"

R: L158 the use of sequential updates appears for the first time but I can't really see why this is relevant or used elsewhere?

A: Removed.

R: The Kalman filter method seems like a really important rival approach but despite being given a full subsection 2.3 this only has one paragraph and no mathematical definition. It would be nice to be able to compare the approaches more clearly

A: The mathematical definitions were added. Additionally, we added more references to the fact that we do not think the Kalman Filter is a relevant rival approach in that particular case, as it is too restrictive on its prior, and consequently, on its posterior S.

R: PlioMIP appears in L200 without being mentioned before

A: Corrected to "Pliocene Modelling Intercomparison Project (PlioMIP)".

R: There really is very little need to use a complex Hamiltonian Monte Carlo method like NUTS on a simple linear regression problem. With a small change in prior from Cauchy to Inv-Gamma the whole problem would become conjugate and could be done exactly on a pocket calculator.

A: Adding more mathematical restrictions on the form of prior would make the problem less relatable to reality where prior choices may be physically motivated. However, in Appendix A, we have illustrated the result obtained by the reviewer's suggestion. The differences are minimal.

R: L270 the posterior distributions of what?

A: Posterior distributions of S. This has been corrected in the text.

R: L273 and elsewhere. There are some weird mentions about Cauchy distributions having a finite integral whilst Uniform distributions do not. This makes no sense to me (the Uniform is only an improper prior if it has infinite limits). None of this is referenced so needs clarifying.

A: This has been removed/corrected. We still mention this once, but have explicitly stated that in our case, the prior can not be considered as improper since it is bounded. However, one could indeed choose a non-bounded Uniform prior to infer S, in which case the prior would be improper.

———————————————————

Short Comment #1

No changes were applied in the manuscript in response to SC#1. As we argued in the original answer, we do not think that the Kalman filter should be mentioned in the abstract to avoid an overload of information. Additionally, we did not wish to add other Bayesian approaches as it is beyond the scope of this paper. The initial online response was:

C: The paper's criticism of the Kalman filter method (section 2.3), as implying – very likely unrealistically – that the model ensemble is a credible predictor before consideration of the observational constraint, almost ruling out posterior estimates outside the model range, is valid and in my view sufficiently important to warrant mentioning in the Abstract.

A: This, indeed, is a relevant point. As pointed out by Reviewer #2, the Kalman Filter could appear as a significant rival approach. However, we consider it too restrictive, for the reasons reminded here by the author of the comment. Nevertheless, the main scope of the study is not to criticise the Kalman filter method, but to present a Bayesian method which we think is more appropriate to the question of emergent constraints. Therefore, we do not consider adding references to the Kalman Filter in the abstract, mainly to avoid an overload of information that could mislead the reader.

C: However, a major weakness of the paper is that it fails to investigate, or even acknowledge the existence of, an objective Bayesian method that has been applied for a very similar purpose, or of the frequentist likelihood ratio method that has also been so applied (Lewis and Grunwald 2018). Objective Bayesian methods use a 'noninformative prior' that reflects how the expected informativeness of the data about the parameter(s), derived from the likelihood function, varies over the parameter space, and where not all parameters are of interest may also reflect the targeted parameter(s). There is a huge statistical literature on objective Bayesian methods, as there also is on likelihood ratio methods.
Both the aforementioned objective Bayesian and likelihood ratio methods generate un- certainty distributions and ranges that that have been shown, in a perfect model test, to be well calibrated for combining, as well as evaluating separately, independent evidence (Lewis 2018). That is, the uncertainty ranges output by these two methods, although different in statistical nature, are both close to exact confidence intervals. Accordingly, in the long run probabilistic conclusions by an investigator employing either of these methods will on average be true statements, which is surely highly desirable for scientific investigations. That is not in general the case for subjective Bayesian methods (Fraser 2011, Lewis 2014).
Moreover, Bayesian updating does not in general produce satisfactorily calibrated inference when combining evidence, even if the related Bayesian inference from the separate pieces of evidence is well calibrated (Lewis 2013, Lewis 2018). Nor is Bayesian updating satisfactory as a method of incorporating probabilistic prior information, which can however be incorporated under the aforementioned objective Bayesian method. The appropriate way to do so is by treating the prior information not as a prior density to be used in Bayesian updating, but as equivalent to a notional observation with a certain probability density, from which a posterior density has been calculated using Bayes' theorem with a noninformative prior (Hartigan 1965).
In order to achieve satisfactory inference about climate sensitivity when combining evidence, climate scientists need to move on from fundamentally flawed subjective Bayesian methods, and to cease ignoring the existence of objective Bayesian and frequentist (profile) likelihood ratio based methods that are both demonstrably superior.

A: Besides the title of "objective" Bayesian method, which we find confusing and misleading, there are several reasons why such methods are not investigated nor acknowledged in this study. One of the first reasons would be that this paper does not aim at being a summary of every possible Bayesian method, but solely introduces one method for the question of emergent constraints in comparison to other (non-Bayesian) methods used in the past.

Having said so, “non-informative” prior such as Jeffrey’s prior, are, in fact, very informative when dealing with a single problem which carry information by itself, such as the relationship between Sensitivity (S) and Temperature (T) in a defined ensemble of climate model. Actually, we do consider that informative priors are a valuable advantage of Bayesian methods (or updating), as it carries the knowledge, with a certain uncertainty, of the original problem (in that case, the plausible range of S). There is no reason for thinking that one prior would be more non-informative than others in every case - priors are more informative than others based on the problem. In the

case of this paper, we could consider that the uniform prior is more informative than the Cauchy prior towards high S. However, such affirmation could be completely different with a different set of climate system parameters. Thus, there is no reason for thinking that one specific Bayesian method would have more or less flaws than another.

– – – – – – – – – – – – – – – – – – –

Short Comment #2

A: The primary criticism of S&W is that our underlying model is different to that adopted in previous Emergent Constraint (EC) research. We agree that we are presenting a fundamental and perhaps radical point of departure from previous work and we will emphasise this more clearly in the text. In our view, it is entirely natural when attempting to estimate a parameter such as ECS, that we take a prior on ECS which is explicitly stated, and then ask how the evidence under consideration (in this example, a temperature observation and an ensemble of GCMs that exhibit a relationship between their ECS values and this temperature) can be used to update this prior. We believe that our approach, while unlike much previous work in EC, is actually much more closely aligned with the bulk of the research on Bayesian estimation of ECS, and our method allows for these data to be easily used to update any prior on ECS that might arise from a previous unrelated study. In contrast, as Sansom and Williamson showed so elegantly themselves, a Bayesian formulation of standard EC approach would be to take a prior over the measurand and infer ECS as a diagnostic of this. While this seems to work well for their “reference prior” case and they also discuss physically-motivated priors on the regression parameters, it does not seem so clear (and they do not discuss) alternative priors for the measurand and the implied predictive prior for ECS. Of course we do not insist that our view is the only correct one and must be adopted by others, but we believe strongly that it is reasonable and worthy of serious consideration.

C: The central feature of the proposed approach is to specify $p(X^\star \mid Y^\star)$ and $p(X \mid Y)$ rather than $p(Y^\star \mid X^\star)$ and $p(Y \mid X)$. It is important to realise that these are two fundamentally different statistical models, they cannot both be valid at the same time, and will inevitably result in different inferences for Y. Though both equations hold mathematically (as they are valid factorisations of the joint distribution), they imply different conditional independencies in the modelling that need to be physically interpreted and are certainly not interchangable. Therefore, one must consider carefully the underlying reasoning before adopting one or the other. For emergent constraints, $Y^\star$ is usually a measurable property of the future climate and $X^\star$ an observable property of the current or historical climate. Therefore it makes immediate sense to adopt the standard model for emergent constraints, i.e., the future depends upon the past via $p(Y^\star \mid X^\star)$ and $p(Y \mid X)$. In those cases the proposed approach would make no sense because it explicitly states that the past depends on the future via $p(X^\star \mid Y^\star)$ and $p(X \mid Y)$. Equilibrium Climate Sensitivity (ECS) is operationally defined as “the temperature anomaly reached at equilibrium following a instantaneous doubling of CO2”. To us, it seems natural to view this as a future climate quantity, and so it makes sense to adopt the standard model for ECS.
It might be possible to justify the proposed model for certain quantities of interest, but for an operationally defined future quantity, we would have to be willing to accept the reversal of time’s arrow, i.e., past depends on future. It may be that the authors have in mind some alternative definition or interpretation of ECS that renders time’s arrow an illusion for this quantity in particular.

A: We have to disagree with this comment for several reasons. In the definition of emergent constraints, there is strictly no assumption that one variable belongs to the future, nor that another variable belongs to the past. The theory of emergent constraints is solely based on a physically plausible relationship between two variables of the climate system. Having said so,

Sansom & Williamson argue that ECS should be viewed as a future climate quantity, which is not the view we share. We interpret ECS as a parameter of the climate system which is independent of time. In particular, it is usually accepted that small forcing, such as the traditional 4 times CO2 forcing used to compute ECS leads to weak non-linearity. Therefore, ECS can be approximated as a constant parameter of the climate system under a certain forcing, and has its own existence in the present time.

As written earlier for Reviewer #2, we added the following: “Although the unobserved variable is usually taken as a future variable, the emergent constraints theory can be used with two variables within the same time frame, as long as the relationship is plausible.”
We think this is a frequent source of confusion, when there is no restriction on the emergent constraint theory to be used with one past and one future variable. We also added physical arguments (see following comment) which explicitly states that ECS computed from 4xCO2 experiment usually leads to weak non-linearity, and thus can be considered as a constant parameter of the climate system under certain forcing and be used within the LGM and mPWP framework.

C: we must strongly insist that the physical argument behind the statistical model be made explicit, well defended, and open to the scrutiny of other researchers within the field.

A: It is rather straightforward to find arguments for the two models. One way (where S is the predicted and T the predictand) was already introduced by Annan and Hargreaves (2016). Here, we stipulate that, with a certain knowledge about S, one could estimate a value of tropical T, accounting for a certain uncertainty coming from temperature changes non-related to S. Physical arguments have been added in the revised paper, as the following:

“It implies that S is able to give a prediction of T_tropical, with a certain uncertainty. This is physically plausible, as S is considered as one of the best metric to represent temperature change. In particular, S is often diagnosed in climate models from abrupt and sustained quadrupling of CO_2 from pre-industrial conditions (4xCO_2), which usually leads to weak non-linearity similar to what shall be observed from LGM or mPWP climate dynamics. Therefore, it is possible to use 4xCO_2-computed S of climate models to predict T_tropical, assuming epsilon as a representation of all processes not related to S.”

C: However the statistical model is not stated explicitly in its entirety, nor are the models it is compared to. For the sake of transparency, we interpret the model used by the authors to be
$X_m \sim \text{Normal}(\beta_0 + \beta_1 Y_m, \sigma^2)$ form=1,...,M
$X^\star \sim \text{Normal}(\beta_0 + \beta_1 Y^\star, \sigma^2)$
$Z \sim \text{Normal}(X^\star, \sigma_{Z^2})$
Examination of those same Python scripts reveals that four chains of only 2 000 samples each with no warm-up were used in the production of the reported results (a total of only 8 000 samples). Both STAN and PyMC3 (used by the authors) implement the No U-Turn Sampler (NUTS) variant of Hamiltonian Monte Carlo for efficient mixing and fast convergence, so our results should be comparable. The lack of warm-up / burn-in period used by the authors is likely to lead to skewed estimates, even using NUTS. Further, the inefficient use of importance sampling to account for the observation uncertainty and the small total number of samples given notorious the difficulty of efficiently sampling from the Cauchy distribution are all likely to contribute to the differences we see when implementing their framework. Unless we have misunderstood their modelling (and by itself this would be an argument for making it explicit in the manuscript), we are not sure that the numbers given in the text actually represent the posteriors of their alternative model faithfully.

A: There is a difference between the model above (which is a simple linear regression) and the model used in the paper. This is mainly due to a lack of clarification of the model as already pointed out by Reviewer #2 which has been corrected. Therefore, the non-reproducibility of the results shown by Williamson and Sansom is not due the amount of samples as commented here. This is mainly due to a mistake in the code which was posted online. All computations were performed with 200 000 samples (and not 2000) for a total of 800 000 samples. By default, the code came with 2000 samples for faster simulation, but was not meant to be published with this number. This mistake has been corrected and we thank the authors of the comment to have

spotted it. Additionally, this comment made us think about the size of burn-in period, which has been increased. To check the convergence of the chains, each code comes with a Gelman-Rubin test, easily implemented by PyMC3.

Having said so, the model has been made more explicit as written above in the response to Reviewer #2. Additionally, the code provided has been changed to make them efficient but also slightly faster to run. By default, the BLR produces 100,000 samples (400,000 with 4 chains), and burn-in 10,000 samples. Importance sampling uses 800,000 samples. NUTS (PyMC3) comes by default with a substantial burn-in period, which is not originally written.

C: Bowman et al. (2018) include explicit expressions for projection using the Kalman filter model (Equations 17 & 23). However using these expressions we are also unable to reproduce the results for the Kalman filter quoted in Table 2 of the manuscript. Examining the Python script that accompanies this manuscript reveals an obvious error in the expression for the posterior variance on Line 80. The authors should therefore revisit the calculation of these credible intervals.

A: The code has been corrected, as a matrix multiplication was missing. The direct consequence of this is having a more constrained posterior range using the Kalman Filter (roughly 0.5 K more constrained on each bounds), which tend to show that the Kalman Filter is even more restrictive than expected.

C: We were able to reproduce the Ordinary Least Squares (OLS) estimates and intervals. However, in Section 2.1 the authors equate OLS with frequentist linear regression. This is incorrect. As discussed in detail in Williamson & Sansom (2019), OLS is a purely algorithmic method of parameter estimation in a mathematical model. The OLS estimates of the mean parameters in a linear regression model are equal to those obtained by frequentist maximum likelihood estimation, but OLS provides no estimate of uncertainty in either the parameters or the prediction. Frequentist regression is difficult to justify in a climate change context, but Bracegirdle & Stephenson (2011) presented emergent constraints within a frequentist linear regression framework and this approach has been adopted in many subsequent studies. However, the authors present heuristic uncertainty estimates based on mean-squared errors and on lines 287-288 claim that “OLS” underestimates uncertainty compared to their method. In fact, when standard frequentist regression is used, the inference for ECS in the LGM PMIP2 experiment is very similar to the proposed model with median 2.8K and 90% confidence interval 1.0–4.5K. As Williamson & Sansom (2019) point out, this is equal to an equal tailed 90% Bayesian credible interval under reference priors. Credible intervals from the reference model under the other experiments are less similar to the proposed model, but wider than either the “OLS” or Kalman filter estimates.

A: We might have taken a non-elegant shortcut here. The main idea was that OLS, at the opposite of Bayesian regression, does not require any prior knowledge, which makes it closer to frequentist philosophy. Nevertheless, this does not exclude OLS of being one of the most used method and therefore has a large interest for multi-study comparison. Our terminology may also have been a little sloppy. While we agree OLS per se does not provide uncertainty estimates, these are readily obtained from standard regression methods which are used for the OLS estimate. Finally, the uncertainty estimates were updated following the method of generating a predictive ensemble, as they are computed in the Bayesian methods and as they were computed in previous studies (Hargreaves et al., 2012; Schmidt et al., 2014; Hargreaves and Annan, 2016). As shown by the authors of the comment, the new intervals are slightly wider than the previous ones, but still much tighter than the Bayesian intervals in the case of low correlation, which was already the original argument of this method comparison.
We changed the terminology of OLS since it does not produce confidence intervals, but statistical models based on it will produce these intervals.

C: Treatment of model inadequacy
In the statistical reasoning above we omitted discussion of model inadequacy for brevity. In lines 130–134 and lines 376–378 the authors claim that in their approach model inadequacy can be entirely accounted for by specifying a larger residual variance for reality than the models and it is not necessary to consider differences in the regression parameters. With a minor modification to the proposed model, the intercept can be made independent of the slope, and therefore any additional uncertainty about the intercept in the real world can safely be pushed into the residual

since both sources⋆ of uncertainty are independent of Y . However, any uncertainty in the slope leads to uncertainty about X⋆ that is dependent on Y ⋆, i.e, the width of the predictive interval for X⋆ increases with the distance of Y ⋆ from the prior mean. Therefore any additional uncertainty in the slope in the real world is also conditional on the value of Y ⋆ and not accounted for by the residual variance. This is a simple matter of geometry and is therefore unavoidable. Williamson and Sansom (2019) developed a coherent elicitation of the regression parameters and structural error that was designed to account for these geometric considerations, and any emergent constraints framework that wishes to account for structural error, whether using the authors approach or the standard one, must grapple with the geometry.

A: We disagree with this comment. Although accounting for model inadequacy is relevant to the question of emergent constraints, accounting for it on every regression parameters leads to creation of regression lines based on an ensemble of non-existing models. We are aware of the Williamson and Sansom approach which postulates a different regression line based on some hypothetical ensemble of future models, and while we agree with them that treatment of model inadequacy is important we don't find their approach entirely compelling. In reality, there is one sensitivity value and one temperature observation, and it does not seem helpful to us to use 3 additional parameters including a new regression line to describe this. In the limit of improved models (W&S Section 4b), we would expect them all to converge to the truth and it is not clear that a regression line and error term is particularly meaningful in this case. Furthermore, it might be reasonable to expect that in this limiting ensemble sigma^* should be rather smaller, instead of larger, than sigma for the existing ensemble.
We emphasise that our point here is not to criticise W&S who have made a set of judgments that they consider reasonable, but merely to explain why ours differ.
We think this discussion is somewhat beyond the scope of our study, which is focused on the use of an existing ensemble of climate models.

A Bayesian framework for emergent constraints: case studies of climate sensitivity with PMIP

Martin Renoult[1], James Douglas Annan[2], Julia Catherine Hargreaves[2], Navjit Sagoo[1], Clare Flynn[1], Marie-Luise Kapsch[3], Qiang Li[4], Gerrit Lohmann[5], Uwe Mikolajewicz[3], Rumi Ohgaito[6], Xiaoxu Shi[5], Qiong Zhang[4], and Thorsten Mauritsen[1]

[1]Department of Meteorology, Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden
[2]Blue Skies Research Ltd, Settle, United Kingdom
[3]Max-Planck Institute for Meteorology, Hamburg, Germany
[4]Department of Physical Geography, Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden
[5]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
[6]Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan

Correspondence: Martin Renoult (martin.renoult@misu.su.se)

Abstract.

In this paper we introduce a Bayesian framework, which is ~~flexible and explicit about the~~explicit about prior assumptions, for using model ensembles and observations together to constrain future climate change. The emergent constraint approach has seen broad application in recent years, including studies constraining the equilibrium climate sensitivity (ECS) using the Last Glacial Maximum (LGM) and the mid-Pliocene Warm Period (mPWP). Most of these studies were based on ~~Ordinary Least Squares~~ ordinary least squares (OLS) fits between a variable of the climate state, such as tropical temperature, and climate sensitivity. Using our Bayesian method, and considering the LGM and mPWP separately, we obtain values of ECS of 2.7 K (~~1.1–4.8~~0.6–5.2, 5–95 percentiles) using the PMIP2, PMIP3 and PMIP4 data sets for the LGM, and ~~2.4 K (0.4–5.0~~2.3 K (0.5–4.4) with the PlioMIP1 and PlioMIP2 data sets for the mPWP. Restricting the ensembles to include only the most recent version of each model, we obtain 2.7 K (~~1.1–4.3~~0.7–5.2) using the LGM and ~~2.4~~ 2.3 K (0.4~~–5.1–~~4.5) using the mPWP. An advantage of the Bayesian framework is that it is possible to combine the two periods assuming they are independent, whereby we obtain a ~~slightly~~ tighter constraint of ~~2.6 K (1.1–3.9)~~ 2.5 K (0.8–4.0) using the restricted ensemble. We have explored the sensitivity to our assumptions in the method, including considering structural uncertainty, and in the choice of models, and this leads to 95% probability of climate sensitivity mostly below 5 ~~and never~~ K and only exceeding 6 K in a single and most uncertain case assuming a large structural uncertainty. The approach is compared with other approaches based on OLS, a Kalman filter method and an alternative Bayesian method. An interesting implication of this work is that OLS-based emergent constraints on ECS generate tighter uncertainty estimates, in particular at the lower end, ~~suggesting a higher bound by construction~~ an artifact due to a flatter regression line in case of ~~weaker~~ lack of correlation. Although some fundamental challenges related to the use of emergent constraints remain, this paper provides a step towards a better foundation of their potential use in future probabilistic estimation of climate sensitivity.

1 Introduction

In recent years, researchers have identified a number of relationships between observational properties and a future climate change, which was not immediately obvious a priori, but which exists across the ensemble of global climate models (GCMs) (Allen and Ingram, 2002; Hall and Qu, 2006; Boé et al., 2009; Cox et al., 2018) participating in the Climate Model Intercomparison Project (CMIP). These relationships are generally referred to as 'emergent constraints' as they emerge from the ensemble behaviour as a whole rather than from explicit physical analysis.

Such emergent constraints have been broadly used to constrain properties of the Earth's climate system which are not easily or directly observable. These are usually presented in probabilistic terms, mostly based on Ordinary Least Squares (OLS) methods. For example, studies have explored the constraint on equilibrium climate sensitivity (ECS), which is the global mean equilibrium temperature after a sustained doubling of CO_2 over pre-industrial levels, using model outputs from the Paleoclimate Model Intercomparison Project (PMIP) (Hargreaves et al., 2012; Schmidt et al., 2014; Hopcroft and Valdes, 2015; Hargreaves and Annan, 2016). Because of their relatively strong temperature signal, paleoclimate states like the Last Glacial Maximum (LGM) and the mid-Pliocene Warm Period (mPWP) are often considered as promising constraints for the ECS (Hargreaves et al., 2012; Hargreaves and Annan, 2016), in particular at the high end.

Almost all emergent constraint studies have used OLS to establish the link between variables in the model ensembles. However, whether ECS or another climate parameter was investigated, the theoretical foundations for the calculations have not previously been clearly presented. An additional problem arising from this is the resulting difficulty in synthesising estimates of climate system properties generated by different statistical methods with different, and often not explicitly introduced, assumptions. These methods include OLS but also alternative Bayesian approaches such as estimates of the climate sensitivity using energy-balance models (Annan et al., 2011; Aldrin et al., 2012; Bodman and Jones, 2016).

Two recent papers have also addressed the question of emergent constraints in different ways. Bowman et al. (2018) presented a hierarchical statistical framework which went a long way to closing the gap in theoretical understanding of emergent constraints. Conceptually, it is very similar to a single step Kalman filter, where the iteration process is avoided to only keep a single updating of a prior into a posterior. Specifically, it uses the model distribution ~~(~~approximated as a Gaussian ~~)~~ as a prior, which is then updated using the observation to a posterior. However, such prior and the underlying assumptions attached to it could be seen as a restrictive choice to impose on the climate sensitivity area of research. In particular, ~~it is fairly difficult to generate posterior values which are outside of the prior range , even when the observation is outside the range covered by models~~most of the posterior values would lie in the range covered by the ensemble of models if the observed value is either uncertain and/or close to the prior mean. This is a direct consequence of the joint probability distribution produced by the Kalman filter, which in the case of joint Gaussian distributions, will produce a tighter posterior Gaussian distribution. Because of that, it does not appear to correspond to the choice which is usually made, albeit implicitly.

Another Bayesian statistical interpretation of emergent constraints has been recently presented by Williamson and Sansom (2019) who extended the standard approach to account for more general sources of uncertainty including model inadequacy. A key aspect of their approach is that they set a prior on the observational constraint rather than the climate system parameter(s)

that we are primarily interested in this study, i.e. the climate sensitivity. Thus, their prior predictive distribution for the climate system parameter is not immediately clear and may not be so easily specified as in the approach we explore ~~and which is described below~~here.

We present an alternative Bayesian linear regression approach in which the regression relationship is used as a likelihood model for the problem. This allows for the prior over the predictand to be defined separately from and entirely independently of the model ensemble and emergent constraint analysis. Thus the likelihood arising from the emergent constraint could be used to update a prior estimate of the predictand that arose from a different source.

In Section 2 we provide an overview of the concept of emergent constraints, the previous methods used for these analyses, introduce the Bayesian ~~linear regression method~~ framework as well as the models and data employed in the paper. Section 3 describes the results, starting with analysis of models and data from the Paleoclimate Intercomparison Project (PMIP) Phases 2 and 3 for the LGM and mPWP, that have previously been analysed for an emergent constraint on climate sensitivity (Schmidt et al., 2014; Hargreaves and Annan, 2016). We then incorporate some CMIP6/PMIP4 model outputs that have been made available to us for these periods to illustrate how these outputs fit into the same analysis. We also use the LGM and mPWP outputs to demonstrate how the method allows independent ~~lines of evidence~~ emergent constraints to be combined. Finally, we discuss the influences of the prior and of model inadequacy on climate sensitivity in Sections 3.5 and 3.6, respectively.

2 Methods

The general method ~~, referred to as "emergent constraints", suggests that if a relationship exists, in~~ of "emergent constraints" seeks a physically plausible relationship in the climate system between two model variables in an ensemble of results from different climate models~~, between an observable model variable and another model variable that we seek to constrain~~. Consequently, an observation of one measurable variable (such as ~~climate sensitivity) then an observation~~ past tropical temperatures) could be used to better constrain ~~that variable~~the other investigated variable, usually unobserved and difficult to measure (such as climate sensitivity). This idea has been used in climate science to ~~forecast variables~~ estimate quantities of interest such as snow albedo feedback (Hall and Qu, 2006), future sea ice extent (Boé et al., 2009; Notz, 2015), low-level cloud feedback (Brient et al., 2016), and ~~to estimate~~ the equilibrium climate sensitivity (Hargreaves et al., 2012; Schmidt et al., 2014; Cox et al., 2018). Although the unobserved variable is usually taken as a future variable, the emergent constraints theory can be used with two variables within the same time frame, as long as the relationship is plausible. A summary of several different emergent constraints on climate sensitivity was made by Caldwell et al. (2018). This approach using emergent constraints is meaningful only if we believe that reality satisfies the same relationship, and it was not observed purely by chance in the model ensemble. There is a risk in searching for such relationships in a small ensemble that we may find examples which are coincidental, with no real predictive value (Caldwell et al., 2014). Spurious relationships could also be found because of model limitations (Fasullo and Trenberth, 2012; Grise et al., 2015; Notz, 2015).

In this study, we focus on the relationship between equilibrium climate sensitivity, defined here as S, and the temperature change in the tropics which is observed at the Last Glacial Maximum (LGM) and the mid-Pliocene Warm Period (mPWP),

defined as T_{tropical}. We posit that a relationship between climate sensitivity and temperature change is physically plausible, as we expect the long-term quasi-equilibrium temperature to be mainly influenced by radiative forcing, and ~~past model ensembles~~in many model ensembles, variations in climate sensitivity have been dominated by tropical feedbacks, mostly arising from low-level clouds (Bony et al., 2006; Vial et al., 2013). ~~Furthermore, since mixing with the deep oceans happens mostly at high latitudes and equilibrating the deep ocean temperatures in a climate model takes several thousand years, there is the risk that the temperature in these regions have not yet equilibrated to the same extent in all models. This can cause high latitude temperature variations that are not related to climate sensitivity, but to how modelling centres conducted their experiments.~~

2.1 Ordinary Least Squares

The most widely-used approach to emergent constraint analysis is to find an observable phenomenon that exhibits some relationship to the parameter of interest, and use this as a predictor in a linear regression framework. The ~~Ordinary Least Squares~~ ordinary least squares (OLS) method has been widely used because of its simplicity, and so we also use it here as a ~~benchmark~~ starting point for comparison with alternative statistical methods. In the context of constraining climate sensitivity, the parameter of interest (i.e. the ECS) is considered as a predicted variable (Hargreaves et al., 2012; Schmidt et al., 2014; Hargreaves and Annan, 2016). This may be written as

$$S = \alpha \times T_{\text{tropical}} + \beta + \epsilon \tag{1}$$

where S is the climate sensitivity, ~~α and β two unknown constants,~~ T_{tropical} γ and δ two unknown parameters, T_{tropical} the temperature anomaly averaged over the tropical region for the given paleo-time interval, and ~~ϵ~~ ζ the residual term which is drawn from a Gaussian distribution ~~$N(0,\sigma)$~~ $N(0,\sigma^2)$ and which accounts for deviations from the linear fit. When we use this approach, the unknown constants of the linear fit are estimated via ordinary least squares (OLS) using the $(T^i_{\text{tropical}}, S^i)$ pairs representing the model ensemble (here i indexes the models) and then the equation is used to predict the true value of S for the climate system, based on the observed value T^o_{tropical}. A confidence interval for the predictor variable can be generated by accounting for uncertainties in the fit and in the observed value through a simulation of an ensemble of prediction as was demonstrated by Hargreaves et al. (2012). This procedure makes the assumption that reality satisfies the same regression relationship as the models, i.e. is likely to be at a similar distance from the line as the model points are.

Integrating the intrinsically frequentist ~~OLS-based confidence intervals~~ confidence intervals obtained from regression methods used for OLS estimates into a Bayesian framework is ~~somewhat unclear~~challenging. One issue is the misinterpretation of frequentist confidence intervals as Bayesian posterior credible intervals~~, where the~~ . The former is the representation of ~~a percentage of~~ the number of random intervals to contain the ~~fixed true value of the parameter~~true interval bounds (at 90% confidence, this would lead to 90 out of 100 random intervals to contain the true bounds), while Bayesian credible interval is ~~the probability of the true value to be included within a given interval~~an interval which we believe (with the given probability) to contain the truth. For instance, if there is an observed T_{tropical} = 1 K, with an assumed Gaussian observational uncertainty of σ = 0.25 at one standard deviation, then stating that there is a close-to-95% probability of having the true value of the

parameter within the interval 0.5–1.5 K is a Bayesian credible interval interpretation. However, the latter is a common interpretation of frequentist-based studies. This confusion has inherent drawbacks on the analysis of posterior outputs, as shown in various fields of science (Hoekstra et al., 2014) and more recently, for climate sensitivity computations (Annan and Hargreaves, 2019). Williamson and Sansom (2019) have presented a Bayesian interpretation of this approach using reference priors on ψ, as defined by Cox et al. (2018) as a metric of global mean temperature variability, and the regression coefficients. However, this approach does not appear to readily allow for the use of any arbitrary prior distribution for S which may either be desired for comparison with other research, or else have arisen through a previous unrelated analysis. The Bayesian linear regression approach that we introduce in the next section avoids these problems.

2.2 Bayesian ~~Linear Regression~~Framework

The (subjective) Bayesian paradigm is based on the premise that we use probability distributions to describe our uncertain beliefs concerning unknown parameters. We use Bayes' Theorem to update a prior probability distribution function (pdf) for the equilibrium climate sensitivity via

$$P(S|T^o_{\text{tropical}}) = \frac{P(T^o_{\text{tropical}}|S)P(S)}{P(T^o_{\text{tropical}})} \tag{2}$$

where ~~$P(S|T_{\text{tropical}})$~~ $P(S|T^o_{\text{tropical}})$ is the posterior estimate of S after conditioning on the ~~data~~geological proxy data T^o_{tropical}, $P(S)$ is the prior and ~~$P(T_{\text{tropical}})$~~ $P(T^o_{\text{tropical}})$ is a normalisation constant. The likelihood ~~$P(T_{\text{tropical}}|S)$~~ $P(T^o_{\text{tropical}}|S)$ is a function that takes any value of S and generates a probabilistic prediction of what we would expect to observe as ~~T_{tropical}~~ T^o_{tropical} if that value was correct. The use of the Bayesian paradigm requires us to create such a function. Using the principles of emergent constraint analyses in which a linear relationship between these two parameters, which was seen in the GCM ensemble, is believed to apply also to reality, it is natural to use the regression relationship

$$T_{\text{tropical}} = \alpha \times S + \beta + \epsilon \tag{3}$$

Where here, α, β and σ, the standard deviation of ϵ as $\epsilon \sim N(0, \sigma^2)$, are three a priori unknown parameters. Note that this reverses the ~~rôles~~ roles of predictor and predictand compared to the OLS-based approach (Eq. 1). ~~The values of~~ It implies that S is able to give a prediction of T_{tropical}, with a given uncertainty. This is physically plausible, as S is considered as one of the best metrics to represent temperature change. In particular, S is often diagnosed in climate models from abrupt and sustained quadrupling of CO_2 from pre-industrial conditions ($4xCO_2$), which usually leads to weak non-linearity similar to what shall be observed from LGM or mPWP climate dynamics. Therefore, it is possible to use $4xCO_2$-computed S of climate models to predict T_{tropical}, assuming ϵ as a representation of all processes not related to S.

The three parameters α, β and ~~σ, where σ represents the standard deviation of ϵ as $\epsilon \in N(0, \sigma)$, are estimated from~~ σ are conditioned on the model ensemble defined by its pairs of $(T^i_{\text{tropical}}, S^i)$ (with i indexing the models). We estimate them via a Bayesian linear regression (BLR) procedure, which requires priors to be defined over these parameters. Consequently, the likelihood $P(T_{\text{tropical}}|S)$ for a given S (as required by Eq. 2) is an integration over the posterior distribution of T_{tropical}

predicted by the regression relation (convolved with observational uncertainty where appropriate) and conditioned on the model ensemble through α, β and σ.

In this way, we create a statistical model ~~, conditioned on the model ensemble,~~ that can generate a predictive pdf for the tropical temperature change at the LGM or at the mPWP $P(T_{\text{tropical}}|S)$, for any given sensitivity. There is a structural difference between this approach and that of Eq. 1, in that here the residual uncertainties ~~$\epsilon \in N(0,\sigma)$~~ $\epsilon \sim N(0,\sigma^2)$ represent our inability to perfectly predict the tropical temperature anomaly arising from a given sensitivity, and are probabilistically independent of the latter rather than the former variable. The issue here is not a matter of which regression line is 'correct', but rather how, given the model ensemble, we can create a plausible likelihood model for $P(T_{\text{tropical}}|S)$.

~~The regression prediction for the temperature change as a function of sensitivity, together with the observed tropical temperature change as estimated through analysis of proxy data T^o_{tropical}, naturally leads to a likelihood function for the sensitivity of the climate system~~ It is important to note that Eq. 3 and the conditioning of parameters on the model ensemble only relates to the generation of the likelihood. The emergent constraint calculation itself is then a second step that uses this likelihood to calculate the posterior of interest $P(S|T^o_{\text{tropical}})$ (Eq. 2). To apply the emergent constraints theory, it is required to insert a geological observation T^o_{tropical} estimated through proxy data, and obtain the likelihood $P(T^o_{\text{tropical}}|S)$ which leads to the posterior $P(S|T^o_{\text{tropical}})$ by Bayesian updating. We perform this step through a simple importance sampling algorithm by approximating $P(T_{\text{tropical}} = T^o_{\text{tropical}}|S)$. That is, for ~~a~~ any given sensitivity S, we can calculate the probability of the observation of tropical temperature that we have, as the composition of the predictive pdf for actual tropical temperature, together with the ~~uncertain observation operator. In practice this is performed by a simple sampling algorithm~~uncertainty associated with the observation itself. The emergent constraint theory is thus applied with a 2-stage Bayesian process, including in first the BLR and in second, a Bayesian updating.

A prior belief ~~both~~ on climate sensitivity ($P(S)$) in the Bayesian updating process, and on the parameters of the regression model in the BLR process, has to be assumed. There is no clearly uncontested choice for prior distribution for climate sensitivity. However, Annan and Hargreaves (2011) argued that a Cauchy distribution has a reasonable behaviour with a long tail to high values, but unlike the uniform prior, does not assign high probability to these values. Thus we adopt this prior for our main analyses. In section 3.5 we test the sensitivity of the results to this choice and compare the results obtained using Gamma and uniform prior distributions. Priors for the parameters of the regression model are chosen with reference to the specific experiment and are intended to represent our reasonable ~~(albeit uncertain)~~ expectation that models do indeed generate a regression relationship as described.

An additional issue, that was briefly mentioned above, is that we may like to consider the probability that reality is qualitatively and quantitatively ~~distinct~~ distinguishable from all models. This issue, which was explicitly argued in the context of emergent constraint analysis by Williamson and Sansom (2019), seems reasonable since all models do share a theoretical heritage and certain limitations. However, this issue remains challenging to quantify. It has not been considered in most previous studies which also makes it difficult to compare. We investigate this issue in Section 3.6. Whilst the proposed resolution remains preliminary and although the concept is promising for understanding emergent constraints, we decide to omit it for the ~~rest of the~~ bulk of our analysis to enable more direct comparisons with previous studies.

The Bayesian ~~Linear Regression (BLR)~~ method is more explicit than the standard OLS approach, as the prior assumptions have to be given by the user. This transparency leads to more freedom and control of the statistical model. Moreover, it ~~is less sensitive~~ has a reduced sensitivity to outliers as the prior on the regression coefficients provides a form of regularisation. This should ~~result in~~ lead to lower variance in the results ~~, particularly when, as in the examples studied here, we have~~ compared to results with wider priors on the parameters, particularly with small model ensembles.

Additionally, the Bayesian method allows the user to add multiple lines of evidence by ~~sequentially~~ updating the chosen prior for S. The method for combining independent constraints is reasonably simple, as it only requires us to calculate and store the posterior of the first emergent constraint analysed, and use this distribution as the prior for the second emergent constraint. Thus it is a direct form of sequential Bayesian updating. This process results in a posterior distribution which will generally be narrower than either of the two posteriors that would have been generated from either of the emergent constraints separately. Although it may be tempting to simply combine all emergent constraints in this way, it is necessary to also consider possible dependencies between the uncertainties in the different emergent constraints before this can be done with confidence (Annan and Hargreaves, 2017).

It is not clear if observational errors have always been adequately accounted for in previous emergent constraints research. Our approach provides a natural framework for this, as the likelihood can include the uncertainty of the observational process as we have done. However, we have ignored uncertainties in the calculation of the model values of S and T_{tropical} as, while they are poorly quantified, we believe them to be too small to materially affect our result. In fact, it has been argued for the case of the mPWP that observational errors on S and T_{tropical} are small compared to the structural differences responsible of the dispersion of the points around the regression line and thus can be neglected (Hargreaves and Annan, 2016) .

2.3 Kalman Filter

Bowman et al. (2018) recently presented a new interpretation of emergent constraint analysis. Their framework is essentially a two-dimensional ensemble Kalman Filtering approach in which the prior, represented by the model ensemble, is updated according to the observation, using the Kalman ~~equation~~ equations which approximates all distributions by a multivariate Gaussian ~~.~~ Kalman (1960) . The Kalman equations are given by

$$K = P^{\mathrm{f}} H^{\mathrm{T}} \left(H P^{\mathrm{f}} H^{\mathrm{T}} + R \right)^{-1} \tag{4}$$

$$x^{\mathrm{a}} = x^{\mathrm{f}} + K \left(z - H x^{\mathrm{f}} \right) \tag{5}$$

$$P^{\mathrm{a}} = (I - KH) P^{\mathrm{f}} \tag{6}$$

where x is the mean, P its covariance, z the observations with associated observational uncertainty covariance matrix, R and H the operator that maps the model state onto observations. The superscripts f and a by convention refer to the forecast (i.e. the prior in this work) and analysis (the posterior) respectively. While in many applications, such as numerical weather prediction, this method is applied in an iterative fashion with the analysis being used as the starting point of the next forecast, here it is only applied once as a way of implementing Bayesian updating to our prior in order to generate the posterior.

Here we only have two dimensions for the Gaussian, these being the scalar predictor (e.g. sensitivity) and predictand (e.g. tropical temperature change). While this approach is a natural and attractive option in many respects, it has the specific drawback ~~of~~ (in the context of this work) of using the distribution of model samples as a prior (for both mean and covariance). Existing literature on emergent constraints does not make this assumption and this could be seen as a limiting aspect of the method, as it implies that the model ensemble is already a credible predictor even before consideration of the observational constraint. Some implications of this approach are that the posterior estimate ~~is~~ will be equal to the model distribution in the case that no observational constraint exists, either because there is in fact no relationship between observation and predictand, or else when the observational uncertainty is excessively large. ~~It is also~~ The use of a Gaussian prior based on the ensemble range also means that it is difficult for the method to generate posterior estimates that include values significantly outside the model range, even in the case where the observed value is outside the model spread. We present results generated with a Kalman filter in Section 3.1 for comparison with our main analysis.

2.4 Climate Models and Data

The ~~BLR~~ Bayesian method may be applied to any emergent constraint. In this study, we use the model outputs and data syntheses arisen from phases 2 and 3 of PMIP (Braconnot et al., 2007; Haywood et al., 2011; Harrison et al., 2014), as well as the few available models of the phase 4 (Haywood et al., 2016; Kageyama et al., 2017), summarised in Table 1. The Last Glacial Maximum (19–23 ka) corresponds to the period of the last ice age where ice sheets and sea ice had their maximum extent. Due to its temporal proximity, relative abundance of proxy data, and substantial radiative forcing anomaly, the LGM is widely considered one of the best paleoclimate intervals for testing global climate models and has been featured in all of the PMIP consortium experiments. A representation of several model LGM simulations compared to the SAT reconstruction of Annan and Hargreaves (2013) is shown in Fig. 1–(a).

Previous results from PMIP2 showed a significant correlation between LGM tropical temperatures and climate sensitivity in the models (Hargreaves et al., 2012), although the equivalent calculation for the PMIP3 models found no significant correlation (Schmidt et al., 2014; Hopcroft and Valdes, 2015). These two similar sized ensembles with contrasting characteristics are a good test-bed for exploring the properties of the different methods. For the tropical temperature anomaly relative to pre-industrial we use a value from Annan and Hargreaves (2013), for 20° S to 30° N, a T^o_{tropical} of -2.2 K with a Gaussian observational uncertainty of $\pm$ 0.7 K (5–95% confidence interval). Several data compilations are presently in development as part of PMIP4, but these have yet to be integrated into a global temperature field so revising the temperature estimate from Annan and Hargreaves (2013) is a topic for future work.

Interest in the mPWP (2.97–3.29 million years ago) as a more direct analogy for future climate change, has grown during the past years. This is the most recent period with a sustained high level of greenhouse gases and concomitant warmth relative to the pre-industrial period, however, the data are more sparse and uncertain. In Fig. 1–(b), the sea-surface temperature anomaly of different climate models which performed a mPWP simulation is displayed, as well as the PRISM3 SST reconstruction (Dowsett et al., 2009). Previous results for this period from the ~~PlioMIP~~ Pliocene Model Intercomparison Project (PlioMIP) experiment, which was part of PMIP3, indicated a fairly strong correlation between tropical temperature and climate sensitivity

in the models, but the confidence with which this can be used to constrain climate sensitivity was low due to high uncertainty in various observationally derived components as well as various compromises in the way the protocol was formulated (Hargreaves and Annan, 2016). For the mPWP, a tropical temperature anomaly of 0.8 $\pm$ 1.6 K (5–95% interval) is taken from Hargreaves and Annan (2016) for 30° S to 30° N, assuming the largest 5–95% uncertainty showed in that work. The reconstruction used here is the PRISM3 (Pliocene Research, Interpretation and Synoptic Mapping) SST anomaly field as described in Dowsett et al. (2009).

The Last Interglacial (127 ka, referred as lig127k in CMIP6) and the mid-Holocene (6 ka) are part of the PMIP simulations and also relatively warm climates. The forcings are, however, seasonal and regional in nature, mostly influencing the patterns of climate change. The global change in temperature and the global climate forcing are both very small, and this coupled with the large uncertainty in paleoclimate data makes these intervals poor candidates for constraining climate sensitivity. We do not explore these intervals further here.

Climate sensitivity has various definitions and there are also a number of different ways of approximating the value in climate models that have not been run to equilibrium. For PMIP3 LGM the model values are mostly based on the regression method of Gregory et al. (2004), but for the models which contributed to PMIP2 LGM and PlioMIP the exact definition and derivation used in each case is not always clear in the literature. In order to make comparisons with previous work, here we use the same values as those used in Hargreaves et al. (2012), Schmidt et al. (2014) and Hargreaves and Annan (2016) with two exceptions to ensure that only one value of sensitivity is used for identical versions of the same model across different experiments. Specifically, for FGOALS-g2 we use the value of 3.37 K (Yoshimori, pers. comm.) for both PMIP3 LGM and PMIP3 PlioMIP, and for HadCM3 we use 3.3 K (Randall et al., 2007) for both PMIP2 LGM and PMIP3 PlioMIP. Previous values used by Hargreaves and Annan (2016) for PMIP3 PlioMIP were 3.7 K for FGOALS-g2 (Zheng et al., 2013) and 3.1 K for HadCM3 (Haywood et al., 2013). These changes are minor compared to the ensemble range of climate sensitivity and thus, they have no significant effect on the posterior outputs.

In addition to the already published results from PMIP2 and PMIP3 we add to our ensembles the results that are currently available from PMIP4 in section 3.3. While the LGM protocol (Kageyama et al., 2017) remains very similar to that in previous iterations of PMIP, the mPWP protocol (Haywood et al., 2016) has more significant differences which address several of the limitations of the previous version. Most importantly, PlioMIP2 seeks to represent a specific quasi-equilibrium climate state in the past rather than representing an amalgamation of different warm peak climates as had been the case for PlioMIP1. A priori we are therefore less confident about combining the results from PlioMIP1 and PlioMIP2 and do so mostly to indicate where the new models lie in the ensemble and to highlight the potential for future research in this area once more model results based on the PlioMIP2 protocol become available.

3 Applications and Results

In order to apply the Bayesian Linear Regression and compute the likelihood ~~$P(T_{\text{tropical}}|S)$~~$P(T_{\text{tropical}}|S^*)$, several priors have to be established as initial conditions. Specifically, for both the LGM and the mPWP we use Eq. 3 as the basis for our

likelihood function. The prior expectations of the three unknown parameters α, β and the standard deviation of the residual ϵ, referred to as σ, need to be defined. The relative complexity of the likelihood function with three a priori unknown parameters requires the use of a sampling method for computational efficiency. In this study, we use the Markov Chain Monte Carlo (MCMC) method NUTS as described by Hoffman and Gelman (2014). The NUTS method is also included in the MCMC python package PyMC3 (Salvatier et al., 2016) which is applied here. ~~Other MCMC methods which have been tested, such as Metropolis sampling or Hamiltonian Monte Carlo, give equivalent~~ The approach is alternatively described as a conjugate priors problem using the R package spBayes (Finley et al., 2013, 2014), described in Appendix A, and leads to similar results.

Depending on the strength of the correlation among the dataset, one could expect a sensitivity of the regression to the choice of prior parameters. In the following sections, we first describe the physical arguments behind the choice of priors over α, β and σ, and then present the outputs of the BLR for both the PMIP2 and PMIP3 dataset of the LGM and the PlioMIP1 dataset of the mPWP. Then, we include the CMIP6 data in the ~~BLR~~ Bayesian framework for both paleo intervals, and present an approach of combining the two emergent constraints. Finally, we explore the sensitivity of the ~~BLR~~ Bayesian approach to the choice of priors over the climate parameter of choice (i.e. the climate sensitivity) and to the hypothetical inadequacy of climate models.

3.1 The Last Glacial Maximum

From consideration of energy balance arguments and fundamental physical properties, such as the response of Earth to an increase of CO_2, we have a prior expectation of a relationship between sensitivity and global LGM temperature anomaly (e.g. Lorius et al., 1990), and model experiments of Hargreaves et al. (2007) as well as simple physical arguments about the spatial distribution of forcing suggest that this relationship may be most clearly visible when we focus on the tropical region. While the total negative forcing at the LGM is roughly twice as large as the positive forcing that would be caused by a doubling of CO_2, the temperature response at low latitudes is generally expected to be lower than the global mean, due to polar amplification and the related presence of high latitude ice sheets. Thus we might reasonably expect the tropical temperature change at the LGM to be roughly equal to the global temperature rise under a doubling of CO_2. It would also be unexpected if the correlation had the opposite sign to that based on simple energy balance arguments, such that a more sensitive model had a lower temperature change at the LGM. However we cannot justify imposing a precise constraint on the slope and therefore our choice of prior for α is ~~$N(-1,1)$~~$N(-1,1^2)$. As for β, we expect the regression line to pass close to the origin, as a model with no sensitivity to CO_2 would probably have little response to any other forcing changes, especially in the tropical region where the influence of ice sheets is remote. However, we do not expect a precise fit to the origin and therefore, the prior chosen for β is ~~$N(0,1)$~~$N(0,1^2)$. Finally, we chose a wide prior for σ, a Half Cauchy with a scale parameter of 5. The Cauchy is fairly close to uniform for values smaller than the scale parameter, decaying gradually for higher values.

Deviations from the regression line may be due to different efficacies of other forcing components, especially ice sheets or dust. To take into account the uncertainty on the strength of the response, we performed two additional analyses where the prior response was smaller (α defined as ~~$N(-0.5,1)$~~$N(-0.5,1^2)$) and larger (α defined as ~~$N(-2,1)$~~$N(-2,1^2)$). We do not see much difference in the results using the three priors over α: the difference is approximately 0.2 K of climate sensitivity for

both the upper and lower percentiles quoted, giving us confidence in our choice of ~~$N(-1,1)$~~$N(-1,1^2)$. The computed 5–95% posterior climate sensitivity ranges for different values of α are summarised in Table 2.

The MCMC algorithm samples the posterior distribution of regression parameters which is represented by the ensemble of predictive regression lines in Fig. 2. This ensemble is used to infer the climate sensitivity following the Bayesian inference approach using the geological reconstruction of the LGM tropical temperature. The posterior distributions of S are computed using a truncated-at-zero Cauchy prior with a peak of 2.5 and a scale of 3, which corresponds to a wide 5–95% prior interval of 0.5–28.7 K. Such a prior was used previously by Annan et al. (2011) because it has a long tail, allowing for a substantial probability of having high climate sensitivity while still maintaining some preference for more moderate values. ~~Additionally, the Cauchy prior has a finite integral, unlike the uniform distribution (which is sometimes referred as an "improper prior" for this reason).~~ However, the sensitivity of Bayesian statistics to the choice of prior has often been noted. Thus, two alternative priors, including the widely used uniform prior, and their corresponding posterior distributions, are investigated in Section 3.5.

To test the robustness of the method and also to compare it with the statistical methods used in previous studies, three cases are investigated in which we use different combinations of the available model ensembles. The results are shown in Fig. 2 and Table 2.

For the PMIP2 ensemble, the correlation between tropical temperature and climate sensitivity was found to be reasonably strong and in this study the resulting ~~90~~5–95% range for inferred climate sensitivity is 1.0–4.5 K (Fig. 2–(b)). The range is slightly better constrained at the lower end than the 0.5–4 K from Hargreaves et al. (2012), however we have used the revised value for the LGM tropical anomaly of -2.2 $\pm$ 0.7 K rather than the value of -1.8 $\pm$ 0.7 K that was used by Hargreaves et al. (2012). The Bayesian-inferred value is similar to the OLS-inferred method with the revised version (Table 2), giving confidence on the proximity of both methods in case of high correlation.

When all the models of PMIP2 and PMIP3 (see Table 1) were considered jointly the correlation became weaker and the corresponding 5–95% range generated by the ~~BLR~~ Bayesian method is 0.7–4.8 K (Fig. 2–(d)). Schmidt et al. (2014) obtained 1.6–4.5 K using a similar ensemble although in that case multiple results obtained from the same modelling centre were combined by averaging. Using the OLS method on our ensemble we obtain ~~1.8~~1.4–~~4.3~~.6 K. The ~~BLR~~ Bayesian method generates a wider range here, particularly at the lower end, as the correlation is weaker and the prior starts to influence the posterior.

Finally, we consider the PMIP3 models in isolation. For this ensemble no correlation is found so for the ~~BLR~~ Bayesian method the result is heavily dependent on our prior assumptions. We obtain a 5–95% range here of 0.7–5.5 K (Fig. 2–(f)). Applying the OLS method on the PMIP3 dataset gives a 5–95% range of ~~2.2–4.7~~1.3–5.6 K. As previously argued for the combination of PMIP2 and PMIP3, the OLS produces a tighter posterior range ~~. It is also close to the~~ at the lower end. In the absence of a correlation, the Bayesian method relaxes to the prior, whereas the OLS method is heavily influenced by the range of the ~~initial ensemblebecause of the lack of correlation for this dataset~~ensemble. However, we emphasise that this does not suggest that either range is closer to reality.

The Kalman filtering approach presented by Bowman et al. (2018) has not previously been used for emergent constraint analyses in paleoclimate research. Thus, we also use this method to explore both PMIP2 and the combination of PMIP2 and PMIP3 (Fig. 3). With the same geological reconstruction value, and a prior 5–95% range (based on the PMIP2 GCM ensemble)

of ~~1.8~~1.7–~~4.6~~.5 K, a posterior range of ~~1.3~~S of 1.8–~~4.6~~.1 K is inferred. By combining the PMIP2 and PMIP3 models, the prior 5–95% range becomes 2.0–4.5 K and the posterior range is ~~1.6~~2.2–~~4.5~~.2 K. The increase in lower bound in these calculations is the largest change compared to our Bayesian linear regression method. However, this is strongly forced by the underlying assumptions of a Kalman filter (Section 2.3) which uses the model ensemble as a prior, making it difficult to compute a posterior range outside of the model range, in particular when the observed value is considered as excessively uncertain. Thus, although the Kalman filtering method could be interesting, we do not consider it further, as we stipulate its assumptions are too restrictive for the question of emergent constraints and therefore can not be a relevant method in its current form to efficiently assess S and, in particular, its uncertainty.

3.2 The mid-Pliocene Warm Period

As for the LGM, priors parameters have to be defined to perform the BLR with the mPWP data. In principle these may be different to those used for the LGM experiment, since the total positive forcing of the mPWP is not as large as the negative forcing of the LGM, but in practice we have adopted the same priors for our base case, apart from the obvious sign change for α. ~~Regarding the slope term α, the total positive forcing of the mPWP is not as large as the negative forcing of the LGM. Therefore, it seems reasonable to expect a roughly similar slope in the regression.~~ We performed the same sensitivity experiments as for the LGM, with three different priors over alpha: ~~$N(1,1)$, $N(0.5,1)$, $N(2,1)$~~$N(1,1^2)$, $N(0.5,1^2)$, $N(2,1^2)$. There was only a small difference between the results using the three priors: the differences at the 5th percentile being less than ~~than~~ 0.1 K and the differences at the 95th percentile being approximately 0.3 K (see Table 2). Regarding β and σ, there is no physical reason for ~~them~~ their priors to be substantially different than the ones chosen for the LGM. ~~Indeed, although the mPWP is a warm climate, it should also be expected that there is little temperature change to other forcing if the climate sensitivity is null.~~ Thus, a ~~$N(0,1)$~~ $N(0,1^2)$ prior for β is selected and the same prior for σ as for the LGM analysis is chosen.

The Bayesian inference method applied above for the LGM model outputs is now applied on the mPWP model outputs (Fig. 4). With less abundant models and less well-constrained temperature data, we prefer to assume large uncertainties in the mPWP SST reconstruction (0.8 ± 1.6 K, 5–95% confidence). We adopt the Cauchy prior on climate sensitivity as for the LGM analysis (5–95% interval of 0.5–28.7 K) and compute a 5–95% interval for the ECS of 0.5–5.0 K for the PlioMIP1 dataset. Similar to the results for the LGM, the OLS method (Hargreaves and Annan, 2016) resulted in a slightly narrower 5–95% range than the ~~BLR~~ Bayesian method (1.3–4.2 K, assuming 1.6 K of uncertainty on the data).

3.3 Inclusion of CMIP6 / PMIP4 data

The ongoing PMIP4 experiments have produced LGM and mPWP (PlioMIP2) simulations. Here we add those results to our ensembles. There are two model runs available for the LGM and three for the mPWP (see Table 2) on 1 May 2020.

For the LGM we have previously combined the PMIP2 and PMIP3 results, and the protocol for PMIP4 is not very different. If we combine all three ensembles we obtain a 5–95% range for the ECS of ~~0.8–4.7~~0.6–5.2 K using the ~~BLR~~ Bayesian method (Fig. 5–(b)). The ensemble size is now ~~16~~18, but we note that this includes several models coming from the same modelling centres. Past studies have investigated the proximity of models with hierarchical trees (Masson and Knutti, 2011; Knutti et al.,

2013) and the influence of their dependency on statistical methods (Annan and Hargreaves, 2017). Thus, although we believe such dependencies exist in the ensemble, it is in reality difficult to quantify and correct for this. How to deal with this possible duplication of information is therefore a subjective decision. In Schmidt et al. (2014) it was taken into account by averaging the results from models from the same modelling centre. Here we take an alternative approach of including only the latest version of each model. This gives an ensemble size of ~~9 models~~ (Fig. **??**–(a)) ~~and a rather well-constrained~~ 11 models (Table 2) and a 5–95% climate sensitivity range of ~~1.1–4.3~~0.7–5.2K with the ~~BLR method(Fig. **??**–(b)).~~ Bayesian method. The range here is relatively wide and close to the range computed with the ensemble of PMIP2, PMIP3 and PMIP4. This is due to the removal of almost all PMIP2 models in this restricted ensemble, which leaves mainly the poorly correlated PMIP3 ensemble and the ensemble of PMIP4 together.

For ~~PlioMIP~~ PlioMIP1 and PlioMIP2 the situation is a little more complex as the protocol has been redesigned to represent a specific interglacial state rather than a generic warm climate, referred to as a "time slab" in the PlioMIP protocol. Thus there could be a different regression relationship for these two ensembles. However, when we plot the PlioMIP1 ensemble members (Fig. 5–(d)) we see that they do not look different to the PlioMIP2 ensemble members. The straight combination of PlioMIP1 and PlioMIP2 gives an ensemble range of ~~12~~ 14 models and we computed a 5–95% range of ~~0.4–5.0 K(Fig. 5–(d)).~~ 0.5–4.4 K. Including only the most recent versions of models results in an ensemble size of ~~9 models (Fig. **??**–(c))~~ 11 models (Table 2) and generates a ~~merely~~ nearly identical 5–95% climate sensitivity range of 0.4~~–5.1~~–4.5 K with the mPWP simulation~~(Fig. **??**–(d))~~. Thus, for this period the inclusion of the PlioMIP2 models allows for a tighter constraint at the upper bound, much aided by the larger spread of S in these new models.

3.4 Combining multiple constraints

As described in section 2.4, the mPWP and the LGM are very different climates. If the observational data are generated by unrelated analyses, we may be able to consider the two lines of evidence to be independent, and combine them using Bayes theorem to create a new posterior which is likely to be narrower than that arising from either analysis alone. Assuming that the uncertainties arising from the mPWP and the LGM analyses are independent of each other may be plausible as the proxy reconstructions use different observations and analyses to estimate both the tropical temperatures and the other variables that act as boundary conditions for the model experiments. Moreover, modelling uncertainties that influence the regression analysis are expected to arise from rather different sources, such as the response to ice sheets and a cold climate in one case, versus the influence of a warmer climate in the other. Having said that, model biases influencing the simulation of one climate change may also influence the other, which means that if similar models occur in both ensembles, this could lead to dependencies. Using Bayes theorem to combine the constraints means that it is not necessary for the same set of models to be used for each ensemble but, as we can see from Table 1, a few models do occur in both ensembles.

It is straightforward to first compute the posterior estimate of S from the LGM analysis as previously described, and then use this as a prior for the mPWP analysis. Priors over the regression coefficients are considered independent between the two analyses. Because of the issues discussed above, we perform an analysis using both ensembles of latest model versions in the

LGM and the mPWP as described in section 3.3. The posterior of the LGM is used as the prior for the mPWP analysis and the resulting posterior from this process has a narrower 5–95% interval for S of ~~1.1–3.9~~0.8–4.0 K (Fig. 6).

A logical extension of the approach would be to apply it to the ensemble of models of CMIP, where multiple emergent constraints exist for the same models. In theory, this should be possible as long as the investigated relationships are physically plausible. This goes beyond the scope of our study, which uses the paleoclimates as an example for the method, and is left for future research.

3.5 Alternative Priors on sensitivity

A major strength of the Bayesian analysis developed here is the way that the prior on the parameter of interest, here climate sensitivity, can easily be specified independently of all other aspects of the analysis. A uniform prior for S has been widely used (e.g. Tomassini et al., 2007; Aldrin et al., 2012). However, it has also been argued that such prior could give an unrealistically high ~~probability~~ weight to high climate sensitivity (Annan and Hargreaves, 2011). Here we test our method with the commonly-used uniform prior U[0;10] which has a 5–95% range of 0.5–9.5 K. The resulting posterior 5–95% range for climate sensitivity is 0.8–5.0 K when analysing the LGM PMIP2 models only, and 0.6–5.4 K with the LGM PMIP2 and PMIP3 models together. These posteriors are wider than the ranges previously computed with a Cauchy prior, particularly for the case of combining PMIP2 and PMIP3 where the correlation is rather weak, ~~meaning that~~ in which case the prior has a ~~relatively~~ higher influence. These results are shown in Fig. 7. Due to the questions which have arisen over the use of a uniform prior and the fact that it has an infinite integral, unless bounded arbitrarily as done here, we also perform a comparison with an alternative prior which features a decaying tail and a finite integral. For this purpose, a Gamma prior is chosen with a shape parameter of 2 and a scale of 2, which corresponds to a similar 5–95% prior range of 0.7–9.5 K. The posterior computed 5–95% range is 1.0–4.5 K for LGM PMIP2 models and 0.9–4.8 K for the combination of PMIP2 and PMIP3, which is very close to the one computed with the Cauchy prior. Although the Bayesian paradigm will inevitably involve such subjective choices, the sensitivity of the results to a sensible choice of prior appears to be low as long as a reasonable correlation exists in the ensemble.

3.6 Model Inadequacy

As previously explored and described by Williamson and Sansom (2019), we investigate the probability that all models deviate in a systematic way from reality to a certain extent, mainly because of computational limitations and their shared technical heritage. Statistically, this issue is best described by the terminology that while the models are considered 'exchangeable' with each other, they are not exchangeable with reality. Williamson and Sansom (2019) provide further discussion on this point. In our methodology, this can simply be accounted for by considering that the regression prediction of S for reality has a larger residual than that arising for the models themselves:

$$T^t_{\text{tropical}} = \alpha \times S^t + \beta + \epsilon^*, \tag{7}$$

where the superscript t indicates here that we are referring to the truth (i.e. the real climate system) and ϵ^* has the distribution N(0,~~σ^*) for some $\sigma^* > \sigma$~~σ^{*2}) for some $\sigma^{*2} > \sigma^2$. There can be various reasons why such an inadequacy, represented as ϵ^* in

Eq. 7, may be thought to exist. Models all share a common heritage and theoretical basis, which is certainly incomplete even if not substantially wrong, and computational constraints limit their performance. Particularly in the paleoclimate context, there may be biases in the experimental protocol and differences in number of feedbacks included in the different model systems, e.g. interactive vegetation and prognostic dust. Such errors would lead to reality being some distance from the model regression line, even if the models were otherwise perfect. Such issues are ~~pertinent~~ relevant to both the LGM, where there are significant uncertainties relating to dust and vegetation effects, and the mPWP where even the GHG forcing is somewhat uncertain, and furthermore where the older simulations are designed as a general representation of interglacial warm periods rather than a specific quasi-equilibrium climate state.

However, while we may anticipate reality deviating further from the regression line, it is difficult to quantify such deviation. Here, we perform two sensitivity tests where we define ~~$\sigma^* = 2\sigma$~~ $\sigma^{*2} = (2\sigma)^2$, that is to say the distribution for the residual term ϵ^* is defined as N(0,~~2σ~~) $(2\sigma)^2$) for our predictions. We consider that this corresponds to a rather large inadequacy term. To compare with our previous analysis, we investigate the effect of the model inadequacy using the data set of PMIP2 and PMIP3 combined for the case of the LGM, and the data set of PlioMIP1 for the case of the mPWP~~, and present them in Fig.~~ **??**.. For the LGM, the 5–95% posterior range computed after doubling σ is 0.5–5.8 K~~(Fig.~~ **??** ~~(b))~~, while the 5–95% posterior range for the mPWP is 0.5–5.4 K~~(Fig.~~ **??** ~~(d)).~~. When we consider the 'latest model version' approach outlined in Section 3.3 and take the same approach of doubling the estimated residual, the 5–95% posterior ranges increase to ~~0.7–5.1~~0.5–6.3 K for the LGM and a 5–95% posterior range of 0.4–5~~.7~~.0 K for the mPWP. Thus these sensitivity tests typically ~~add~~ involve a change of around half a degree to the upper bound obtained, while having much less influence on the lower bounds in these examples.

4 Conclusions

Past climates are relevant sources of information on the properties of the climate system, specifically the equilibrium climate sensitivity, due to the quasi-equilibrium changes in response to external forcing, which are of similar magnitude to the projected future climate changes. In this study, we have described a new statistical method based on Bayesian inference to approach the question of emergent constraints. We believe this method provides a reasonable representation within the Bayesian paradigm of the underlying structure of emergent constraint principles. This Bayesian method is designed to be as explicit and flexible as possible. Previous work using ~~Ordinary Least Squares~~ ordinary least squares usually applied implicit assumptions. Because of these assumptions, OLS tends to generate tight posterior ranges, particularly on the lower end and when the correlation is rather weak; something that may well be regarded an artifact of using OLS.

By applying the method to the LGM tropical temperature model ensemble used in Schmidt et al. (2014), which included 14 models from the PMIP2 and PMIP3 generations, we estimate the climate sensitivity to be 2.6 K (0.7–4.8, 5–95 percentiles). Similarly, applying the method to the mPWP tropical temperature data set of Hargreaves and Annan (2016) gives a climate sensitivity of 2.4 K (0.5–5.0), but with the more uncertain ensemble of models which contributed to PlioMIP1.

With the new generation of climate models, the LGM and mPWP analyses have been widened by the addition of several CMIP6 model outputs. By adding the PMIP4 LGM simulations, we computed a 5–95% interval for climate sensitivity of ~~0.8–~~

4.70.6–5.2 K. We performed the same analysis by combining PlioMIP1 and PlioMIP2 models and obtained a 5–95% interval of 0.4–5.00.5–4.4 K. However, these results come with some caveats attached. In particular, combining the two model generations of the mPWP could lead to biased results, since the experimental protocol substantially changed in PlioMIP2. An alternative approach is to consider solely the latest version of each model. By doing this we reduce expected redundancy in the ensemble, and so improve our confidence in the result, despite the smaller ensemble sizes. This leads to a more tightly constrained similarly constrained climate sensitivity of 2.7 (1.1–4.30.6–5.2 5–95%) for the LGM simulations, and a less well-constrained sensitivity 2.4 2.3 (0.4–5.1–4.5, 5–95%) for the mPWP simulations. Our experiment considering a substantial model inadequacy term resulted in an increase of up to a degree in the upper bounds presented here, though this aspect is as yet poorly understood and quantifiedAlthough most of the computed ranges are wider than the ranges obtained with both OLS or Kalman filtering, the Bayesian framework avoids the underlying assumptions of both methods and in particular, makes us regard the Kalman filtering approach in its current form as too restrictive for the question of emergent constraints.

Our Nevertheless, our results obtained by analysing the LGM or the mPWP in isolation are broadly consistent with results obtained by other statistical methods used in previous studies. The differences between the way the information is obtained from the paleo record for the mPWP and the LGM and the different dominant climate features of the intervals suggest it may be reasonable to consider these estimates to be statistically independent, given climate sensitivity. It is then possible to combine them within the same Bayesian framework to compute a narrower range of climate sensitivity. By doing so, we evaluated the climate sensitivity to be 2.6 K (1.1–3.92.5 K (0.8–4.0, 5–95%). NeverthelessHowever,this approach requires independence between the different combined emergent constraints.

It is, in principle, straightforward to include other independent emergent constraints into our Bayesian framework. As well as evidence from historical or present day analyses, other past climates are starting to be explored by modellers and may be potential candidates for future analyses, such as the Eocene, the Miocene and the last deglaciation. Over the next couple of years we expect new outputs for models from CMIP6 and new data analyses to become available, which will enable these preliminary analyses to be compared with results from expanded LGM and mPWP ensembles and improved data estimates.

Code and data availability. The Python codes used for the different statistical methods are available from the Bolin Centre Code Repository at https://git.bolin.su.se/bolin/renoult-2020 (https://doi.org/10.5281/zenodo.3838204). The data of the PMIP2 models can be obtained by asking the corresponding modelling groups. The data of the PMIP3 and CMIP6 models can be downloaded from the ESGF Portal at CEDA, located at https://esgf-index1.ceda.ac.uk/. The data of the PlioMIP1 models can be downloaded from Redmine at the School of Earth and Environment of the University of Leeds, located at https://www.see.leeds.ac.uk/redmine/public/. For username and password, email Alan Haywood (a.m.Haywood@leeds.ac.uk). The PRISM3 SST reconstruction can be downloaded from the PRISM/PlioMIP web page, located under "Experiment 1 AGCM version 1.0, Preferred Data" at http://geology.er.usgs.gov/egpsc/prism/prism_1.23/prism_pliomip_data.html, files PRISM3_SST_v1.1.nc and PRISM3_modern_SST.nc. The LGM SAT geological reconstruction can be downloaded from the Supplementary material of Annan and Hargreaves (2013), currently located at http://www.clim-past.net/9/367/2013/cp-9-367-2013-supplement.zip. For data of CMIP6 models which are not yet published on ESGF, please refer to the corresponding modelling groups.

Author contributions. The BLR method was conceived by JDA and JCH. TM put the project together. The code for the Bayesian framework and for the OLS was written by MR. The code for the Kalman Filter was written by JDA and translated to Python by MR.The statistical analysis were performed by MR. The climate sensitivities of the CMIP6 models were computed by CF. The manuscript was written by MR, JDA, JCH, NS and TM. RO provided the LGM outputs of MIROC-ES2L. UM and MLK provided the LGM outputs of MPI-ESM1.2-LR. GL and XS provided the LGM outputs of AWI-ESM-1-1-LR. QZ and QL provided the mPWP outputs of EC-EARTH3.3.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We thank the two anonymous reviewers for their insightful comments. We are very grateful for the extensive work of the scientists involved in the different PMIPs and their efforts to publish and share their data with us. We acknowledge the Bolin Centre for Climate Research at Stockholm University for giving us access to the code repository which allows to freely share our codes. We acknowledge the STFC Centre for Environmental Data Analysis (CEDA) in collaboration with the InfraStructure for the European Network for Earth System Modelling (IS-ENES) and the Natural Environment Research Council via the National Centre for Atmospheric Science (NCAS) for making available on the ESGF portal the CMIP5 and CMIP6 dataset. We acknowledge Alan Haywood and Richard Rigby for giving us access to the PlioMIP1 dataset. We acknowledge the Statistical Research Group at the Department of Mathematics at Stockholm University and its director Jan-Olov Persson for his useful comments. Rumi Ohgaito acknowledges support from the Integrated Research Program for Advancing Climate Models (TOUGOU programme) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan. The simulations using MIROC models were conducted on the Earth Simulator of JAMSTEC. Gerrit Lohmann and Xiaoxu Shi acknowledge support from the PACMEDY project of the Belmont Forum and the PalMod project through BMBF. The simulations using AWI-ESM-1-1-LR were conducted on the German Climate Computing Centre (DKRZ). Qiong Zhang acknowledges the support from Swedish Research Council VR grants 2013-06476 and 2017-04232. The MPI-M contribution was supported by the German Federal Ministry of Education and Research (BMBF) as a Research for Sustainability initiative (FONA) through the project PalMod (FKZ: 01LP1504C). This result is part of ~~a~~ the highECS project that has received funding from the European Research Council (ERC) (Grant agreement No.770765) ~~under~~ and the CONSTRAIN project, the European Union's Horizon 2020 research and innovation program (Grant agreement No.820829). The analysis and storage of data were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at the National Centre at Linköping University (NSC).

References

Aldrin, M., Holden, M., Guttorp, P., Skeie, R. B., Myhre, G., and Berntsen, T. K.: Bayesian estimation of climate sensitivity based on a simple climate model fitted to observations of hemispheric temperatures and global ocean heat content, Environmetrics, 23, 253–271, 2012.

Allen, M. and Ingram, W.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419, 224–232, 2002.

Andrews, T., Gregory, J. M., Webb, M. J., and Taylor, K. E.: Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models, Geophysical Research Letters, 39, 2012.

Annan, J. D. and Hargreaves, J. C.: On the generation and interpretation of probabilistic estimates of climate sensitivity, Climatic Change, 104, 423–436, 2011.

Annan, J. D. and Hargreaves, J. C.: A new global reconstruction of temperature changes at the Last Glacial Maximum, Climate of the Past, 9, 367–376, 2013.

Annan, J. D. and Hargreaves, J. C.: On the meaning of independence in climate science, Earth System Dynamics, 8, 211–224, https://doi.org/10.5194/esd-8-211-2017, 2017.

Annan, J. D. and Hargreaves, J. C.: Bayesian deconstruction of climate sensitivity estimates using simple models: implicit priors, and the confusion of the inverse, Earth System Dynamics Discussions, 2019, 1–18, https://doi.org/10.5194/esd-2019-33, 2019.

Annan, J. D., Hargreaves, J. C., and Tachiiri, K.: On the observational assessment of climate model performance, Geophysical Research Letters, 38, 2011.

Bodman, R. W. and Jones, R. N.: Bayesian estimation of climate sensitivity using observationally constrained simple climate models, Wiley Interdisciplinary Reviews: Climate Change, 7, 461–473, 2016.

Boé, J., Hall, A., and Qu, X.: September sea-ice cover in the Arctic Ocean projected to vanish by 2100, Nature Geoscience, 2009.

Bony, S., Colman, R., Kattsov, V. M., Allan, R. P., Bretherton, C. S., Dufresne, J.-L., Hall, A., Hallegatte, S., Holland, M. M., Ingram, W., et al.: How well do we understand and evaluate climate change feedback processes?, Journal of Climate, 19, 3445–3482, 2006.

Bowman, K. W., Cressie, N., Qu, X., and Hall, A.: A Hierarchical Statistical Framework for Emergent Constraints: Application to Snow-Albedo Feedback, Geophysical Research Letters, 45, 13,050–13,059, https://doi.org/10.1029/2018GL080082, 2018.

Braconnot, P. et al.: Results of PMIP2 coupled simulations of the mid-Holocene and Last Glacial Maximum, Part 1: experiments and large-scale features, Climate of the Past, 3, 261–277, 2007.

Brient, F., Schneider, T., Tan, Z., Bony, S., Qu, X., and Hall, A.: Shallowness of tropical low clouds as a predictor of climate models' response to warming, Climate Dynamics, 47, 433–449, https://doi.org/10.1007/s00382-015-2846-0, 2016.

Caldwell, P. M., Bretherton, C. S., Zelinka, M. D., Klein, S. A., Santer, B. D., and Sanderson, B. M.: Statistical significance of climate sensitivity predictors obtained by data mining, Geophysical Research Letters, 41, 1803–1808, 2014.

Caldwell, P. M., Zelinka, M. D., and Klein, S. A.: Evaluating Emergent Constraints on Equilibrium Climate Sensitivity, Journal of Climate, 31, 3921–3942, https://doi.org/10.1175/JCLI-D-17-0631.1, 2018.

Cox, P. M., Huntingford, C., and Williamson, M. S.: Emergent constraint on equilibrium climate sensitivity from global temperature variability, Nature Publishing Group, 553, 319–322, 2018.

Dowsett, H. J., Robinson, M. M., and Foley, K. M.: Pliocene three-dimensional global ocean temperature reconstruction, Climate of the Past, 5, 769–783, 2009.

Fasullo, J. T. and Trenberth, K. E.: A less cloudy future: The role of subtropical subsidence in climate sensitivity, Science, 338, 792–794, 2012.

Finley, A., Banerjee, S., and Gelfand, A.: spBayes for Large Univariate and Multivariate Point-Referenced Spatio-Temporal Data Models, Journal of statistical software, 63, https://doi.org/10.18637/jss.v063.i13, 2013.

Finley, A. O., Banerjee, S., and Cook, B. D.: Bayesian hierarchical models for spatially misaligned data in R, Methods in Ecology and Evolution, 5, 514–523, https://doi.org/10.1111/2041-210X.12189, https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12189, 2014.

Gettelman, A., Hannay, C., Bacmeister, J. T., Neale, R. B., Pendergrass, A. G., Danabasoglu, G., Lamarque, J.-F., Fasullo, J. T., Bailey, D. A., Lawrence, D. M., and Mills, M. J.: High Climate Sensitivity in the Community Earth System Model Version 2 (CESM2), Geophysical Research Letters, 46, 8329–8337, https://doi.org/10.1029/2019GL083978, https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019GL083978, 2019.

Goosse, H., Crowley, T., Zorita, E., Ammann, C., Renssen, H., and Driesschaert, E.: Modelling the climate of the last millennium: What causes the differences between simulations?, Geophysical Research Letters, 32, 2005.

Gregory, J. M., Ingram, W., Palmer, M., Jones, G., Stott, P., Thorpe, R., Lowe, J., Johns, T., and Williams, K.: A new method for diagnosing radiative forcing and climate sensitivity, Geophysical Research Letters, 31, 2004.

Grise, K. M., Polvani, L. M., and Fasullo, J. T.: Re-examining the relationship between climate sensitivity and the Southern Hemisphere radiation budget in CMIP models, Journal of Climate, 2015.

Guo, C., Bentsen, M., Bethke, I., Ilicak, M., Tjiputra, J., Toniazzo, T., Schwinger, J., and Otterå, O. H.: Description and evaluation of NorESM1-F: a fast version of the Norwegian Earth System Model (NorESM), Geoscientific Model Development, 12, 343–362, 2019.

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Ohgaito, R., Ito, A., Yamazaki, D., Okajima, H., Ito, A., Takata, K., Ogochi, K., Watanabe, S., and Kawamiya, M.: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks, Geoscientific Model Development, 13, 2197–2244, https://doi.org/10.5194/gmd-13-2197-2020, https://www.geosci-model-dev.net/13/2197/2020/, 2020.

Hall, A. and Qu, X.: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, Geophysical Research Letters, 33, https://doi.org/10.1029/2005GL025127, 2006.

Hargreaves, J. C. and Annan, J. D.: Could the Pliocene constrain the equilibrium climate sensitivity?, Climate of the Past, 12, 1591–1599, 2016.

Hargreaves, J. C., Abe-Ouchi, A., and Annan, J. D.: Linking glacial and future climates through an ensemble of GCM simulations, Climate of the Past, 3, 77–87, 2007.

Hargreaves, J. C., Annan, J. D., Yoshimori, M., and Abe-Ouchi, A.: Can the Last Glacial Maximum constrain climate sensitivity?, Geophysical Research Letters, 39, 2012.

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., Holmgren, K., Izumi, K., and Willis, K.: Climate model benchmarking with glacial and mid-Holocene climates, Climate Dynamics, 43, 671–688, https://doi.org/10.1007/s00382-013-1922-6, https://doi.org/10.1007/s00382-013-1922-6, 2014.

Haywood, A., Hill, D., Dolan, A., Otto-Bliesner, B., Bragg, F., Chan, W.-L., Chandler, M., Contoux, C., Dowsett, H., Jost, A., Kamae, Y., Lohmann, G., Lunt, D., Abe-Ouchi, A., Pickering, S., Ramstein, G., Rosenbloom, N., Salzmann, U., Sohl, L., Stepanek, C., Ueda, H., Yan, Q., and Zhang, Z.: Large-scale features of Pliocene climate: results from the Pliocene Model Intercomparison Project, Climate of

the Past, 9, 191–209, © Author(s) 2013. This is an open access article under the terms of the Creative Commons Attribution CC BY 3.0 License., 2013.

Haywood, A. M., Dowsett, H. J., Robinson, M. M., Stoll, D. K., Dolan, A. M., Lunt, D. J., Otto-Bliesner, B., and Chandler, M. A.: Pliocene Model Intercomparison Project (PlioMIP): experimental design and boundary conditions (Experiment 2), Geoscientific Model Development, 4, 571–577, https://doi.org/10.5194/gmd-4-571-2011, 2011.

Haywood, A. M., Dowsett, H. J., Dolan, A. M., Rowley, D., Abe-Ouchi, A., Otto-Bliesner, B., Chandler, M. A., Hunter, S. J., Lunt, D. J., Pound, M., and Salzmann, U.: The Pliocene Model Intercomparison Project (PlioMIP) Phase 2: scientific objectives and experimental design, Climate of the Past, 12, 663–675, https://doi.org/10.5194/cp-12-663-2016, 2016.

Hoekstra, R., Morey, R. D., Rouder, J. N., and Wagenmakers, E.-J.: Robust misinterpretation of confidence intervals, Psychonomic bulletin & review, 21, 1157–1164, 2014.

Hoffman, M. D. and Gelman, A.: The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo., Journal of Machine Learning Research, 15, 1593–1623, 2014.

Hopcroft, P. O. and Valdes, P. J.: How well do simulated last glacial maximum tropical temperatures constrain equilibrium climate sensitivity?, Geophysical Research Letters, 42, 5533–5539, 2015.

K-1 Model Developers, .: K-1 Coupled Model (MIROC) Description (K-1 Technical Report 1), 2004.

Kageyama, M., Albani, S., Braconnot, P., Harrison, S. P., Hopcroft, P. O., Ivanovic, R. F., Lambert, F., Marti, O., Peltier, W. R., Peterschmitt, J.-Y., Roche, D. M., Tarasov, L., Zhang, X., Brady, E. C., Haywood, A. M., LeGrande, A. N., Lunt, D. J., Mahowald, N. M., Mikolajewicz, U., Nisancioglu, K. H., Otto-Bliesner, B. L., Renssen, H., Tomas, R. A., Zhang, Q., Abe-Ouchi, A., Bartlein, P. J., Cao, J., Li, Q., Lohmann, G., Ohgaito, R., Shi, X., Volodin, E., Yoshida, K., Zhang, X., and Zheng, W.: The PMIP4 contribution to CMIP6 ? Part 4: scientific objectives and experimental design of the PMIP4-CMIP6 Last Glacial Maximum experiments and PMIP4 sensitivity experiments, Geoscientific Model Development, 10, 4035–4055, https://doi.org/10.5194/gmd-10-4035-2017, 2017.

Kalman, R. E.: A new approach to linear filtering and prediction problems, J. Basic Engineering, 82D, 33–45, 1960.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophysical Research Letters, 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.

Lorius, C., Jouzel, J., Raynaud, D., Hansen, J., and Le Treut, H.: The ice-core record: climate sensitivity and future greenhouse warming, Nature, 347, 139, 1990.

Masson, D. and Knutti, R.: Climate model genealogy, Geophysical Research Letters, 38, https://doi.org/10.1029/2011GL046864, 2011.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Brovkin, V., Claussen, M., Crueger, T., Esch, M., et al.: Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1. 2) and Its Response to Increasing CO2, Journal of Advances in Modeling Earth Systems, 11, 998–1038, 2019.

Notz, D.: How well must climate models agree with observations?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 373, 20140 164, 2015.

Randall, D. A., Wood, R. A., Bony, S., Colman, R., Fichefet, T., Fyfe, J., Kattsov, V., Pitman, A., Shukla, J., Srinivasan, J., et al.: Climate models and their evaluation, in: Climate change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR), pp. 589–662, Cambridge University Press, 2007.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C.: Probabilistic programming in Python using PyMC3, PeerJ Computer Science, 2, e55, 2016.

Schmidt, G., Annan, J., Bartlein, P., Cook, B., Guilyardi, É., Hargreaves, J., Harrison, S., Kageyama, M., LeGrande, A., Konecky, B., et al.: Using palaeo-climate comparisons to constrain future projections in CMIP5, Climate of the Past, 10, 221–250, https://www.clim-past.net/10/221/2014/, 2014.

Sueyoshi, T., Ohgaito, R., Yamamoto, A., Chikamoto, M. O., Hajima, T., Okajima, H., Yoshimori, M., Abe, M., O'ishi, R., Saito, F., Watanabe, S., Kawamiya, M., and Abe-Ouchi, A.: Set-up of the PMIP3 paleoclimate experiments conducted using an Earth system model, MIROC-ESM, Geoscientific Model Development, 6, 819–836, https://doi.org/10.5194/gmd-6-819-2013, 2013.

Tomassini, L., Reichert, P., Knutti, R., Stocker, T. F., and Borsuk, M. E.: Robust Bayesian Uncertainty Analysis of Climate System Properties Using Markov Chain Monte Carlo Methods, Journal of Climate, 20, 1239–1254, 2007.

Vial, J., Dufresne, J.-L., and Bony, S.: On the interpretation of inter-model spread in CMIP5 climate sensitivity estimates, Climate Dynamics, 41, 3339–3362, 2013.

Williamson, D. B. and Sansom, P. G.: How are emergent constraints quantifying uncertainty and what do they leave behind?, Bulletin of the American Meteorological Society, 0, null, https://doi.org/10.1175/BAMS-D-19-0131.1, 2019.

Wyser, K., van Noije, T., Yang, S., von Hardenberg, J., O'Donnell, D., and Döscher, R.: On the increased climate sensitivity in the EC-Earth model from CMIP5 to CMIP6, Geoscientific Model Development Discussions, 2019, 1–13, https://doi.org/10.5194/gmd-2019-282, https://www.geosci-model-dev-discuss.net/gmd-2019-282/, 2019.

Zheng, W., Zhang, Z., Chen, L., and Yu, Y.: The mid-Pliocene climate simulated by FGOALS-g2, Geoscientific Model Development, 6, 1127–1135, 2013.

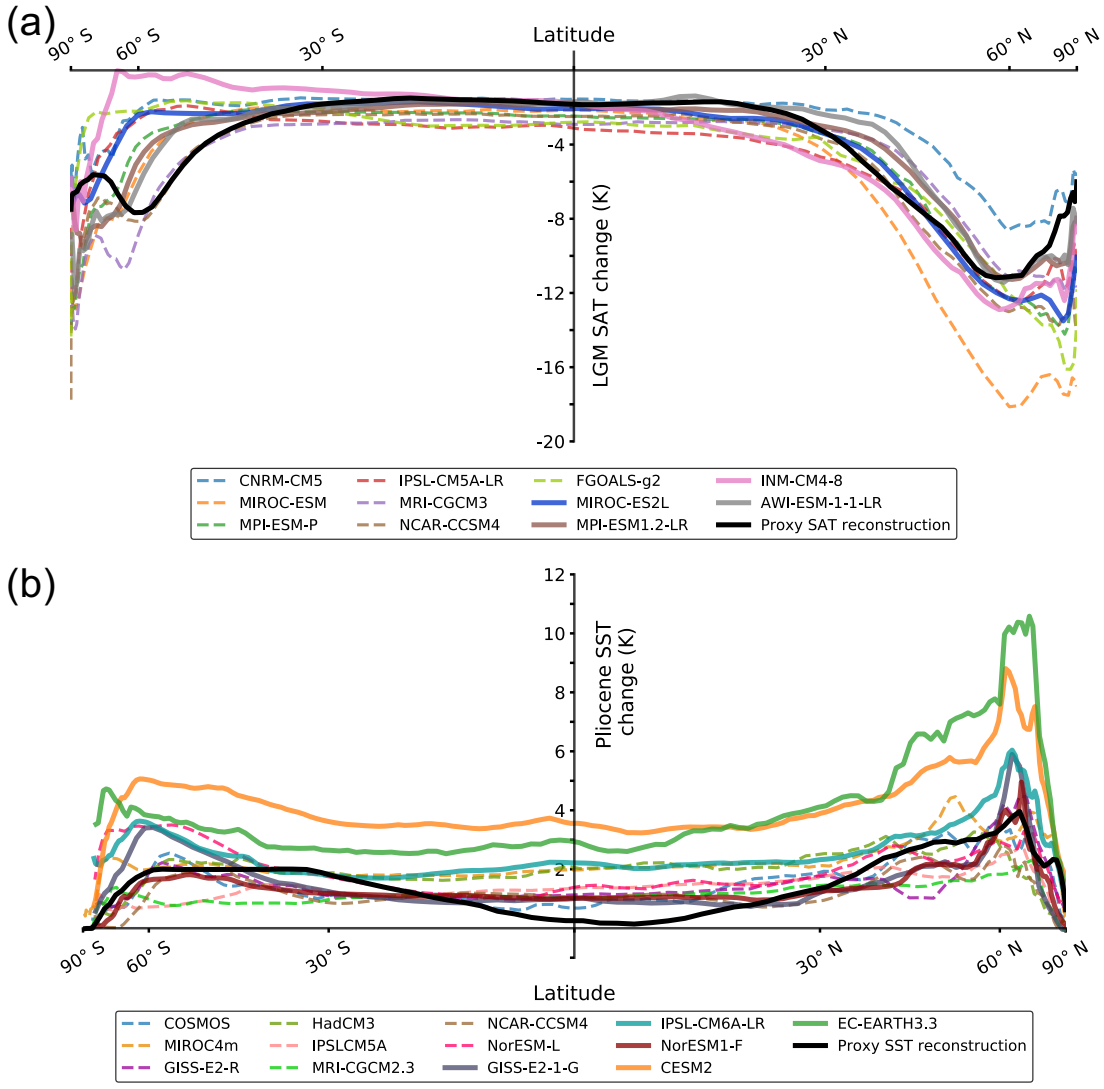

Figure 1. Latitudinal distribution of temperature changes relative to pre-industrial for both simulated climates for various climate models and a proxy reconstruction. Dashed lines are models of the CMIP5 generation, while ~~thick~~ solid lines are models from the CMIP6 generation. All model distributions correspond to 100-year zonal averages when possible; certain CMIP5 PlioMIP1 models were averaged over 30 years. (a), SAT change of the LGM. The ~~thick~~solid black line is a ~~wide~~ multi-proxy ensemble ~~proxy~~ reconstruction taken from Annan and Hargreaves (2013). (b), SST change of the mPWP. The ~~thick~~solid black line is the ~~wide~~ multi-proxy ensemble ~~proxy~~ reconstruction PRISM3 described by Dowsett et al. (2009).

Table 1. Models, tropical temperature (T_{tropical}) outputs and Climate Sensitivity (S) used in this study

Experiment	Figure reference	Model	T^{*}_{tropical}	S	S Reference
PMIP2 LGM	1	MIROC	~~-2.70~~ -2.75	4.0	K-1 Model Developers (2004)
PMIP2 LGM	2	IPSL	~~-2.73~~ -2.83	4.4	Randall et al. (2007)
PMIP2 LGM	3	CCSM	~~-2.16~~ -2.12	2.7	Randall et al. (2007)
PMIP2 LGM	4	ECHAM	~~-3.18~~ -3.16	3.4	Randall et al. (2007)
PMIP2 LGM	5	FGOALS	~~-2.42~~ -2.36	2.3	Randall et al. (2007)
PMIP2 LGM	6	HadCM3**	~~-2.73~~ -2.77	3.3	Randall et al. (2007)
PMIP2 LGM	7	ECBILT**	~~-1.37~~ -1.34	1.8	Goosse et al. (2005)
PMIP3/CMIP5 LGM	8	CCSM4**	~~-2.56~~ -2.6	3.2	Andrews et al. (2012)
PMIP3/CMIP5 LGM	9	IPSL-CM5A-LR**	~~-3.46~~ -3.38	4.13	Andrews et al. (2012)
PMIP3/CMIP5 LGM	10	MIROC-ESM	~~-2.41~~ -2.52	4.67	Sueyoshi et al. (2013)
PMIP3/CMIP5 LGM	11	MPI-ESM-P	~~-2.58~~ -2.56	3.45	Andrews et al. (2012)
PMIP3/CMIP5 LGM	12	CNRM-CM5**	~~-1.68~~ -1.67	3.25	Andrews et al. (2012)
PMIP3/CMIP5 LGM	13	MRI-CGCM3**	~~-2.80~~ -2.82	2.6	Andrews et al. (2012)
PMIP3/CMIP5 LGM	14	FGOALS-g2**	-3.15	3.37	Yoshimori, pers. comm.[1]
PMIP4/CMIP6 LGM	24	MPI-ESM1.2-LR**	-2.06	3.01	Mauritsen et al. (2019)
PMIP4/CMIP6 LGM	25	MIROC-E2L**	-2.23	2.66	Hajima et al. (2020)
PMIP4/CMIP6 LGM	26	INM-CM4-8**	-2.43	1.81	This study
PMIP4/CMIP6 LGM	27	AWI-ESM-1-1-LR**	-1.75	3.61	This study
PMIP3/CMIP5 PlioMIP1	15	CCSM4**	1.03	3.2	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	16	IPSLCM5**	1.33	3.4	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	17	MIROC4m**	1.99	4.05	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	18	GISS ModelE2-R	1.16	2.8	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	19	COSMOS**	2.18	4.1	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	20	MRI-CGCM2.3**	1.15	3.2	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	21	HadCM3**	1.93	3.3	Randall et al. (2007)
PMIP3/CMIP5 PlioMIP1	22	NorESM-L	1.45	2.1	Haywood et al. (2013)
PMIP3/CMIP5 PlioMIP1	23	FGOALS-g2**	2.14	3.37	Yoshimori, pers. comm.[1]
PMIP4/CMIP6 PlioMIP2	~~26~~ 28	GISS-E2-1-G**	0.92	2.6	This study
PMIP4/CMIP6 PlioMIP2	~~27~~ 29	IPSL-CM6A-LR**	2.12	4.50	This study
PMIP4/CMIP6 PlioMIP2	~~28~~ 30	NorESM1-F**	1.37	2.29	Guo et al. (2019)
PMIP4/CMIP6 PlioMIP2	31	CESM2**	3.5	5.3	Gettelman et al. (2019)
PMIP4/CMIP6 PlioMIP2	32	EC-EARTH3.3**	2.94	4.3	Wyser et al. (2019)

*For the LGM simulations (generations PMIP2, PMIP3 and PMIP4), the tropical average was defined between 20° S and 30° N, where the correlation was computed as the highest with climate sensitivity (Hargreaves et al., 2012). For the mPWP simulations (generations PlioMIP1 and PlioMIP2), the tropical average was defined between 30° S and 30° N (Hargreaves and Annan, 2016). All temperature values are defined as changes compared to pre-industrial. **Latest version of a model that was kept for the approach described in Section 3.3.

[1]Calculated using the Gregory method on 150 years of output making it consistent with the values of Andrews et al. (2012).

Table 2. Summary of the methods and computed posterior sensitivities.

Experiment	Method*	5–95% prior (K)	5–95% T^{o}_{tropical} (K)	Median (K)	5–95% posterior (K)
LGM PMIP2	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	2.7	1.0–4.5
LGM PMIP2	~~BLR~~ BF Gamma prior	0.7–9.5	-2.9 – -1.5	2.6	1.0–4.5
LGM PMIP2	~~BLR~~ BF Uniform prior	0.5–9.5	-2.9 – -1.5	2.7	0.8–5.0
LGM PMIP2	OLS predicted CS	n/a	-2.9 – -1.5	2.8	~~1.5~~1.0–~~4.1~~.5
LGM PMIP2	Kalman filter	~~1.8~~1.7–~~4.6~~.5	-2.9 – -1.5	2.9	~~1.3~~1.8–~~4.6~~.1
LGM PMIP2	~~BLR~~ BF α prior mean=-2	0.5–28.7	-2.9 – -1.5	2.7	1.0–4.4
LGM PMIP2	~~BLR~~ BF α prior mean=-0.5	0.5–28.7	-2.9 – -1.5	2.7	0.9–4.6
LGM PMIP2+PMIP3	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	2.6	0.7–4.8
LGM PMIP2+PMIP3	~~BLR~~ BF Gamma prior	0.7–9.5	-2.9 – -1.5	2.6	0.9–4.8
LGM PMIP2+PMIP3	~~BLR~~ BF Uniform prior	0.5–9.5	-2.9 – -1.5	2.7	0.6–5.4
LGM PMIP2+PMIP3	OLS predicted CS	n/a	-2.9 – -1.5	3.0	~~1.8~~1.4–~~4.3~~.6
LGM PMIP2+PMIP3	Kalman filter	2.0–4.5	-2.9 – -1.5	~~3.1~~3.2	~~1.6~~2.2–~~4.5~~.2
LGM PMIP2+PMIP3	~~BLR~~ BF α prior mean=-2	0.5–28.7	-2.9 – -1.5	2.6	0.8–4.7
LGM PMIP2+PMIP3	~~BLR~~ BF α prior mean=-0.5	0.5–28.7	-2.9 – -1.5	2.6	0.7–4.8
LGM PMIP2+PMIP3	~~BLR~~ BF Model inadequacy	0.5–28.7	-2.9 – -1.5	2.8	0.5–5.8
LGM PMIP3	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	2.8	0.7–5.5
LGM PMIP3	OLS predicted CS	n/a	-2.9 – -1.5	3.4	~~2.2~~–~~4.7~~1.3–5.6
LGM PMIP2+PMIP3+PMIP4	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	2.7	~~0.8~~–~~4.7~~0.6–5.2
LGM "Latest" models	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	2.7	~~1.1~~–~~4.3~~0.7–5.2
LGM "Latest" models	~~BLR~~ BF Model inadequacy	0.5–28.7	-2.9 – -1.5	~~2.7~~2.8	~~0.7~~–~~5.1~~0.5–6.3
mPWP PlioMIP1	~~BLR~~ BF Cauchy prior	0.5–28.7	-0.8 – 2.4	2.4	0.5–5.0
mPWP PlioMIP1	~~BLR~~ BF α prior mean=2	0.5–28.7	-0.8 – 2.4	2.4	0.5–4.8
mPWP PlioMIP1	~~BLR~~ BF α prior mean=0.5	0.5–28.7	-0.8 – 2.4	2.4	0.5–5.1
mPWP PlioMIP1	~~BLR~~ BF Model inadequacy	0.5–28.7	-0.8 – 2.4	2.5	0.5–5.4
mPWP PlioMIP1+PlioMIP2	~~BLR~~ BF Cauchy prior	0.5–28.7	-0.8 – 2.4	~~2.4~~2.3	~~0.4~~–~~5.0~~0.5–4.4
mPWP "Latest" models	~~BLR~~ BF Cauchy prior	0.5–28.7	-0.8 – 2.4	~~2.4~~2.3	0.4–~~5.1~~–4.5
mPWP "Latest " models	~~BLR~~ BF Model inadequacy	0.5–28.7	-0.8 – 2.4	~~2.5~~2.4	0.4–5.~~7~~.0
mPWP and LGM, "Latest" models	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	~~2.6~~2.5	~~1.1~~–~~3.9~~0.8–4.0
mPWP and LGM, with CMIP6	~~BLR~~ BF Cauchy prior	0.5–28.7	-2.9 – -1.5	~~2.5~~2.4	~~0.8~~0.7–4.1

*BLR: Bayesian ~~Linear Regression~~framework . OLS: Ordinary ~~L~~least ~~S~~squares. Truncated-at-zero Cauchy prior: peak=2.5, scale=3. Gamma prior: peak=2, scale=2. Uniform prior: bounded 0–10. The "Latest" models ensembles are those created from the most recent versions of each model (see Section 3.3).

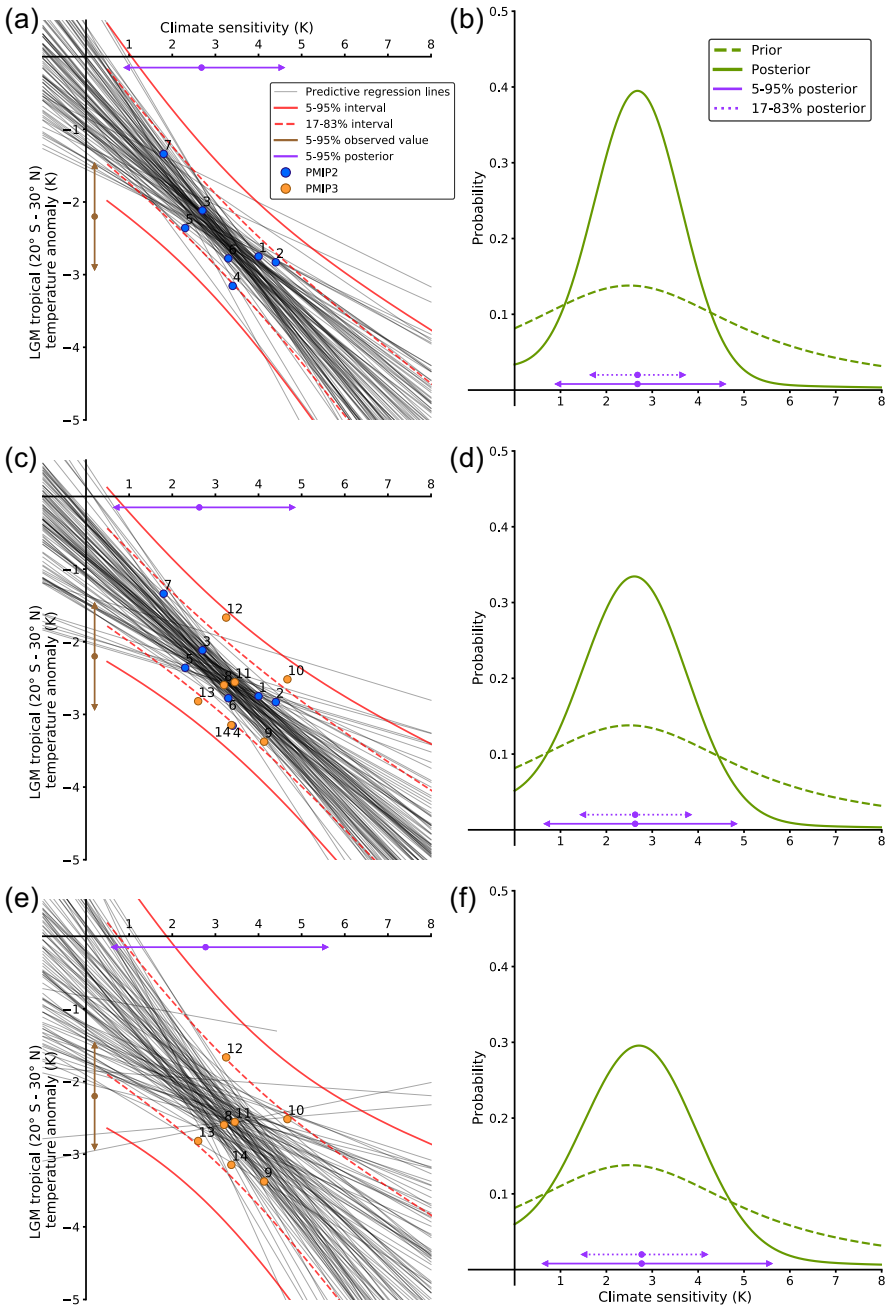

Figure 2. LGM northern tropical (20° S–30° N) temperature versus climate sensitivity for the PMIP2 and PMIP3 models. On the left, predictive regression lines sampled with the MCMC method. On the right, corresponding posterior climate sensitivity computed with a Cauchy prior and inferred from a geological reconstruction taken from Hargreaves et al. (2012). (a) and (b), analysis done on the PMIP2 dataset; (c) and (d), analysis done on the PMIP2 and PMIP3 combined dataset; (e) and (f), analysis done on the PMIP3 dataset. The numbers on each point refer to the models used as listed in Table 1.

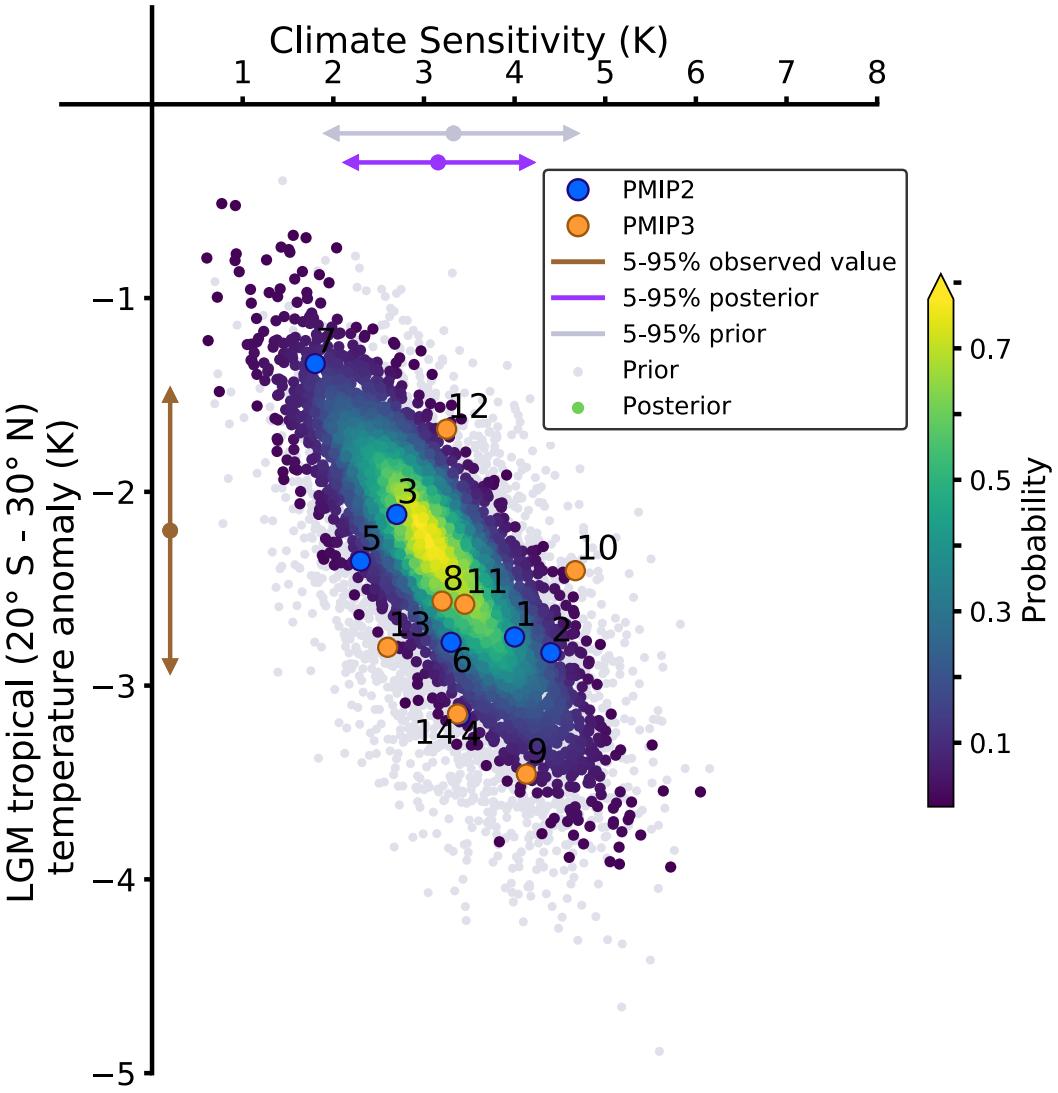

Figure 3. LGM northern tropical (20° S–30° N) temperature versus climate sensitivity of the PMIP2 and PMIP3 models. The Kalman filtering is applied on the ensemble of both PMIP2 and PMIP3. The numbers on each point refer to the models used as listed in Table 1.

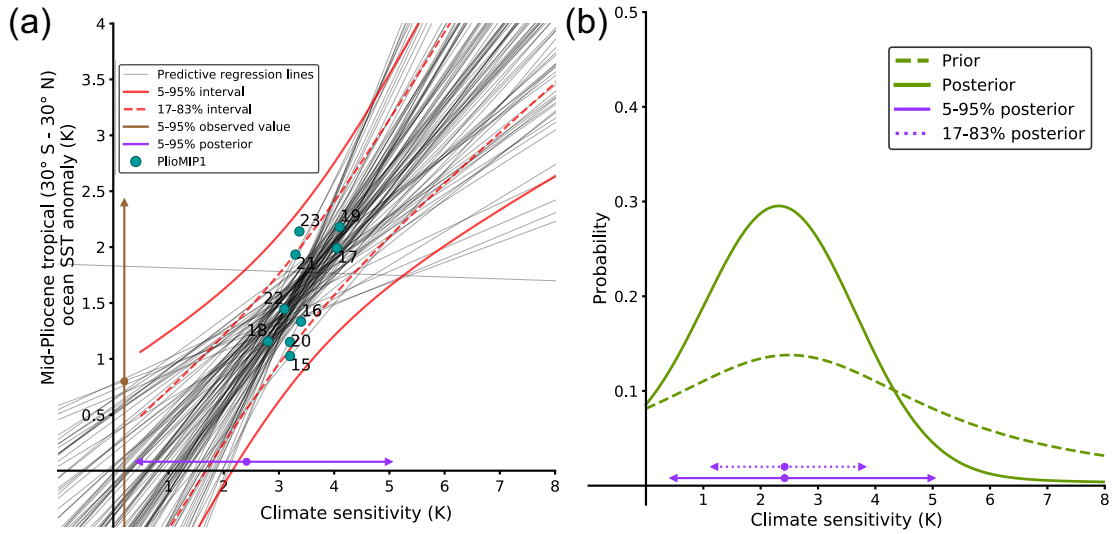

Figure 4. mPWP tropical (30° S–30° N) temperature versus climate sensitivity of the PlioMIP1 models. (a), predictive regression lines sampled with a MCMC method. (b), corresponding posterior climate sensitivity computed with a Cauchy prior and inferred from a geological reconstruction taken from Dowsett et al. (2009). The numbers on each point refer to the models used as listed in Table 1.

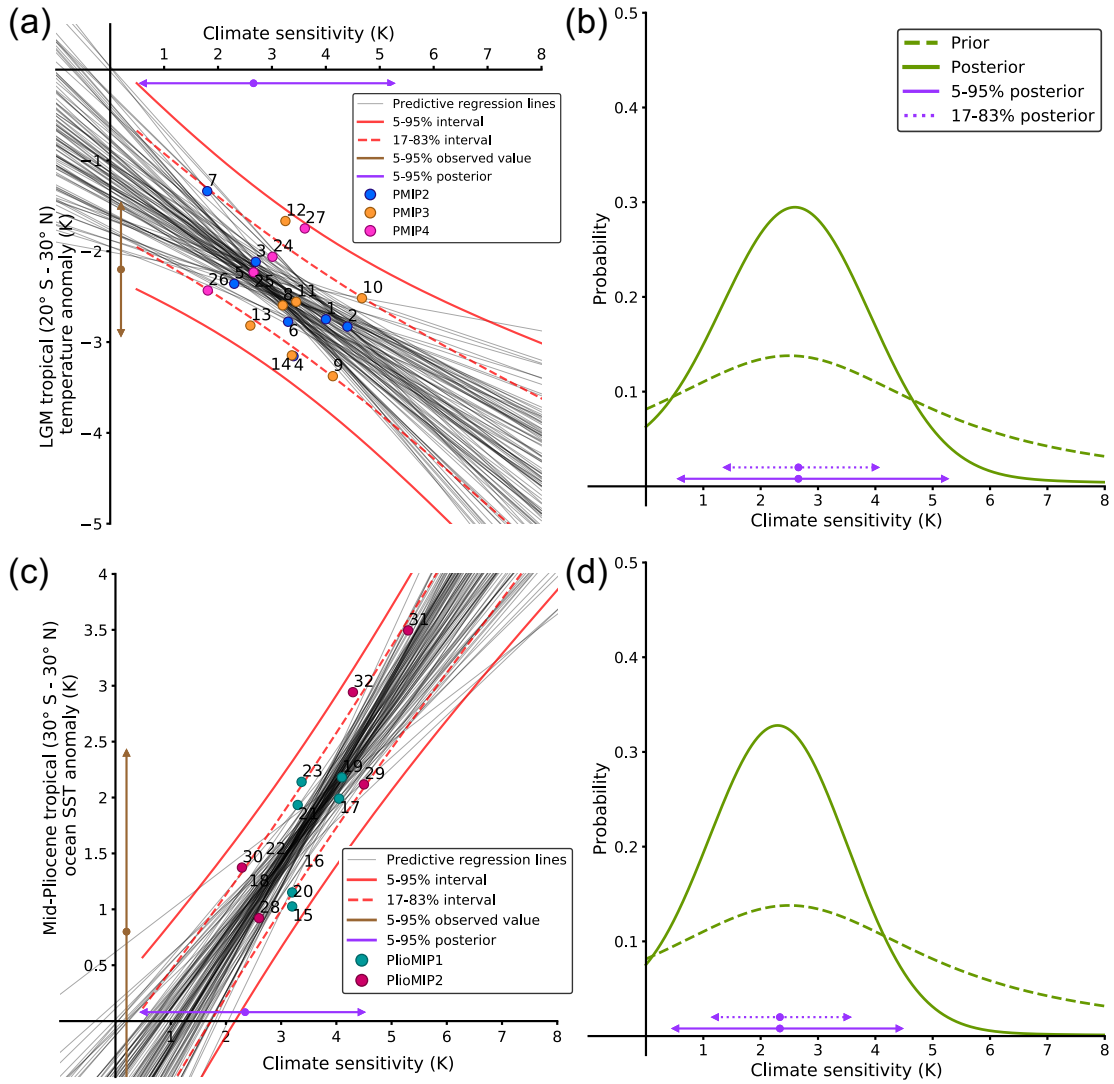

Figure 5. Inclusion of the CMIP6 models into the Bayesian method for the LGM and the mPWP. (a), LGM northern tropical (20° S–30° N) temperature versus climate sensitivity of the PMIP2, PMIP3 and PMIP4 models and (b), inferred climate sensitivity. (c), mPWP tropical (~~20~~30° S–30° N) temperature versus climate sensitivity of the PlioMIP1 and PlioMIP2 models and (d), inferred climate sensitivity. For both inferences, the prior used is a Cauchy distribution defined with a peak of 2.5 and a scale of 3. The numbers on each point refer to the models used as listed in Table 1.

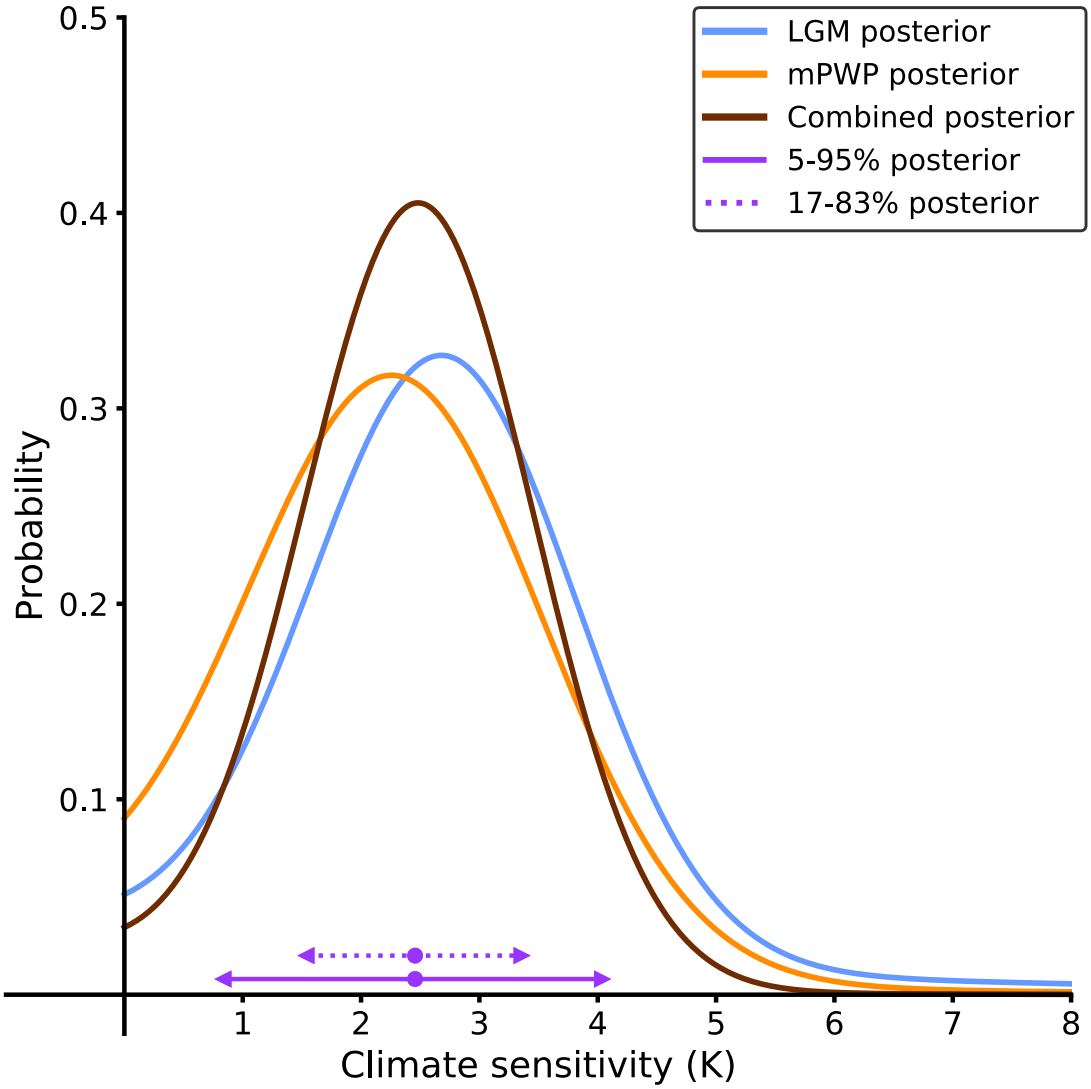

Figure 6. Posterior distribution of climate sensitivity computed with a Cauchy prior by combining two assumed independent emergent constraints. The method does not explicitly use both posteriors of the LGM and the mPWP, but use the LGM posterior as the mPWP prior. However, the resulting combined posterior will usually be narrower than the two independent posteriors. For the LGM, the posterior is computed by using the latest model versions of PMIP, including PMIP4. For the mPWP, the posterior is computed by using the latest model versions of PlioMIP, including PlioMIP2.

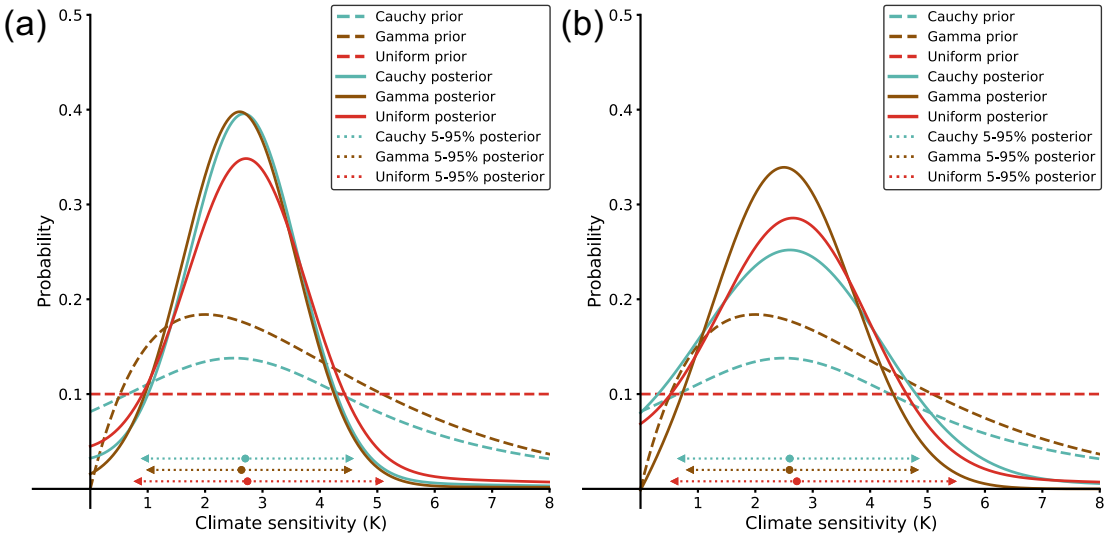

Figure 7. Posterior distributions computed with different priors and dataset. (a), posteriors computed with the PMIP2 dataset (strong correlation). (b), posteriors computed with the PMIP2 and PMIP3 dataset combined (weak correlation). The Cauchy prior is defined with a peak of 2.5 and a scale of 3; The Gamma prior is defined with a peak of 2 and a scale of 2; The Uniform prior is bounded between 0 and 10.

Appendix A: Conjugate priors approach

In section 3, we introduce the NUTS Markov Chain Monte Carlo method, used with the Python package PyMC3 (Salvatier et al., 2016) to compute the posterior distributions of α, β and σ and obtain the likelihood $P(T_{\text{tropical}}|S)$. However, the likelihood model of the Bayesian Linear Regression is defined as $T_i \sim N(\alpha \times S_i + \beta, \sigma^2)$, where (T_i, S_i) is the T_{tropical} and S of the i models. Thus, it is possible to choose conjugate priors in this specific case of emergent constraints to avoid using the complex Hamiltonian-based NUTS method. We show here that both approaches lead to similar results.

For the case of the mPWP, we defined the priors $\alpha \sim N(1, 1^2)$, $\beta \sim N(0, 1^2)$ and $\sigma \sim HalfCauchy(Scale = 5)$. Changing the prior σ to another family of distribution, such as $\sigma \sim InvGamma(Shape = 1.5, Rate = 1)$ leads to a conjugate problem and allows to generate a well-defined form for the posterior distributions of these parameters, while giving a relatively good approximate of the original Half Cauchy prior. To illustrate this approach, we use the R implementation bayesLMConjugate (where LM stands for Linear Model) of the package spBayes. Running the code is significantly faster than the use of a MCMC method, and both posterior outputs are compared in Fig. A1. The differences between both methods is minimal with respect to the range of the posterior parameters and can be attributed to natural variability and differences between the Half Cauchy and Inverse Gamma prior. Thus, both approaches lead to similar estimates of posterior S when they are inserted in the likelihood of the Bayesian updating process. If we take the full ensemble of PlioMIP1 and PlioMIP2 models simulating the mPWP, and using the posterior distributions of α, β and σ from the conjugate prior method, we estimate a 95% S of 2.3 K (0.4–4.6), compared to a value of 2.3 K (0.5–4.4) obtained with NUTS. For the case of the LGM PMIP2, PMIP3 and PMIP4, the conjugate prior generates a 95% estimate of 2.8 K (0.9–4.9) while we obtain 2.7 K (0.6–5.2) with NUTS.

Although the choice of conjugate priors would simplify the computation, NUTS (or MCMC methods in general) have the advantage of allowing an explicit and flexible choice of priors for the users. Having such flexibility is a vital element for the analysis presented in this paper. The example taken here to illustrate the Bayesian framework, i.e. the relationship of T_{tropical} and S is a simple linear regression problem. However, we stipulate that such framework could be used in more complex cases, such as higher complexity emergent constraints relationships, where the use of MCMC methods would become essential.

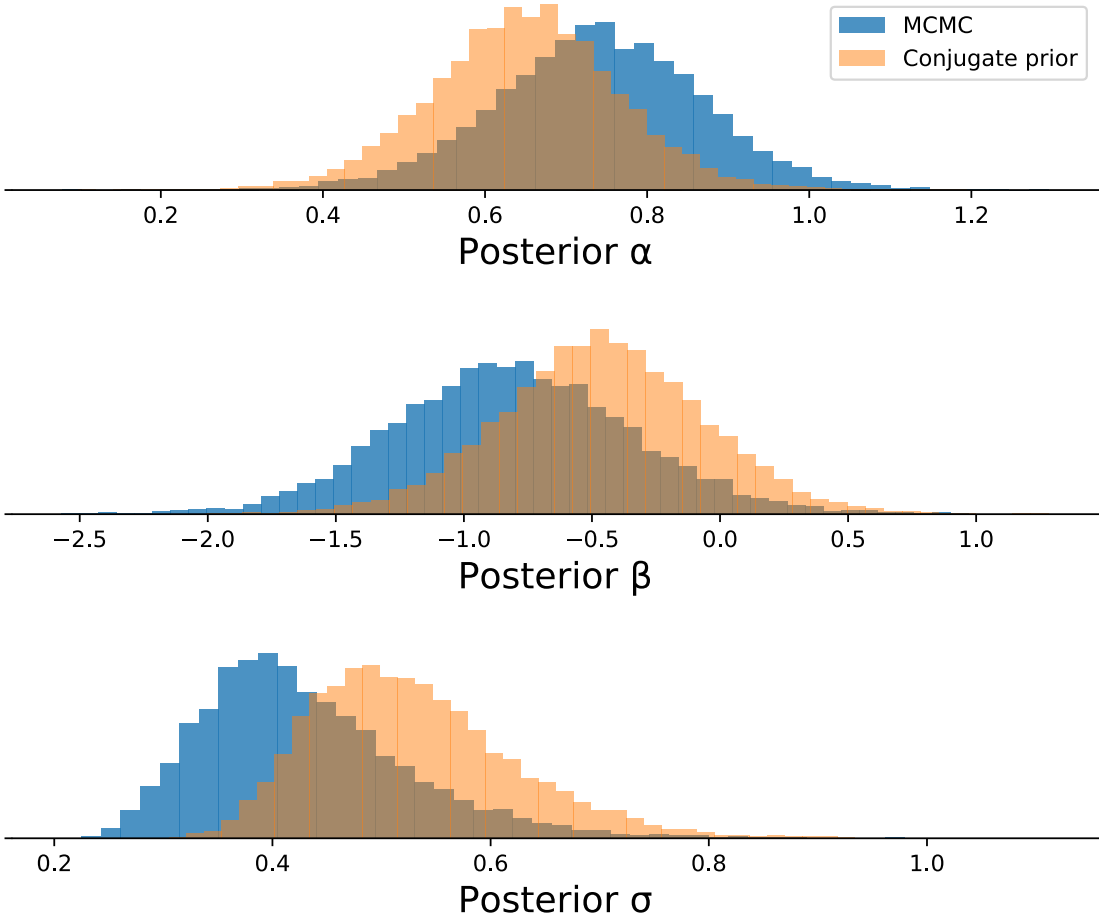

Figure A1. Posterior distributions of the three parameters α, β and σ for the case of the combined PlioMIP1 and PlioMIP2 simulating the mPWP. A re-sample of 4 chains of the MCMC method NUTS (in blue) is compared to the conjugate priors approach (in orange). Some differences can be seen due to natural variability and differences on the prior of σ, but they result in only small differences on the computed posterior S^*