Answer to Short Comment #2 by Philip Sansom and Daniel Williamson

We thank the authors of the comment for their perspective and constructive criticism and we answer the different points where C: refers to the original comment, and A:, our answer.

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.

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, such as pattern effects / time-dependant feedbacks. Physical arguments will be added in the revised paper, as the following:

"The model introduced here implies that S is able to give a prediction of T, with a certain uncertainty. This is physically plausible, as S is often 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 CO2 from pre-industrial conditions (4xCO2), which usually leads to weak non-linearity similar to what shall be observed from LGM or mid-Pliocene climate dynamics. Therefore, it is possible to use 4xCO2-computed S of climate models to predict T, assuming epsilon as a representation of all processes not related to S. These processes are eventually non-linear processes which are by definition not included in the linear expression of S, such as tropical time-dependant feedbacks."

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 will be 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 will be 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 will be increased in the revised paper and should lead to better constraint results. For the interest of the users, there is an easy-to-integrate Gelman-Rubin test coming with PyMC3 but which was later on removed from the shared codes. This will be added again.

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: We thank the authors of the comment for showing us the mistake which will be corrected. 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.

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 $X^\star$ that is dependent on $Y^\star$, i.e, the width of the predictive interval for $X^\star$ increases with the distance of $Y^\star$ from the prior mean. Therefore any additional uncertainty in the slope in the real world is also conditional on the value of $Y^\star$ 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.