The authors have gone to great length to take into account many comments of the four reviewers and have in that process substantially improved the manuscript. However, despite this great work, I feel that some of my major points have not been sufficiently addressed and strong disagreements remain. Most importantly, it reads to me still more like an opinion piece, a very interesting opinion piece indeed, but with very little evidence to support the claims made in the manuscript. As already said before, it is very well written and a pleasure to read. However, I would have liked if the proposition by reviewer 2 to look at some ECS in much more detail would have been taken into consideration. Now it still remains quite speculative.

Thanks to the reviewer for taking the time to re-review the paper. We appreciate that this paper occupies an unusual niche for ESD – but the points made in the paper are not specific to climate sensitivity, they relate to the practise of making inference on an unknown climate parameter using correlations obtained from a structural model ensemble. Thus, although we refer in a number of places to examples which discuss climate sensitivity, this is not our focus.

That said, the reviewer raises a number of reasonable points which we have endeavoured to address better in the manuscript. The reviewer is correct that a clear line needs to be drawn between speculation and objective findings. For the simple model, the objective findings are very simple – that the presence and strength of a constraint are conditional on common structural model assumptions. Applying this logic to the complex models is more nuanced – and is dependent, as the reviewer suggests, on our absolute confidence in process representation in CMIP class models. The point is well taken that process biases ultimately exist in the models, and not in the constraints themselves. We attempt to clarify this in the revised version.

1) I feel that the authors judgement of the three kinds of emergent constraints is strongly biased in favour of type 1. The constraint of type 1 seems to be given more confidence than the ones of type 2 and 3. I find little to no justification why this should be the case.

For example, the authors mention that the EC that relates past warming to TCR is likely robust. So far, I do not see why this should necessarily be the case. What if a tipping point occurs: freshening of the Southern Ocean shuts down deep convection and no warm subsurface waters comes to the surface and heat uptake would be altered? The very different sea ice extent in the CMIP5 and CMIP6 models could move this moment to the early 21st century or to the end of the 21st century or even into the 22nd century. As such, it would also have strong consequences for the TCR and historic warming. historic warming due to a change in albedo and cannot act as a factor later on?

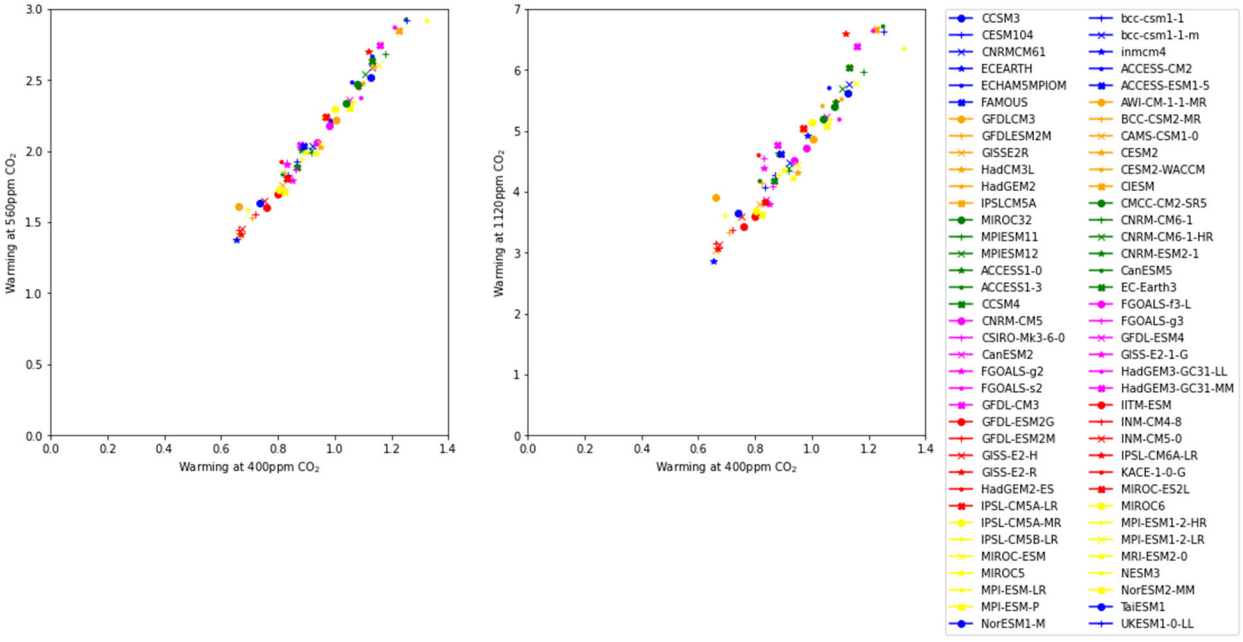
An example is the NorESM model that has a huge Southern Ocean Sea ice mass, much larger than the other models but a rather normal extent, suggesting that a very thick part of sea ice exist. Whereas the first sea ice (thin) disappears quickly, the thick part takes centuries to disappear. Or what if Arctic Sea Ice was melted early in the model and resulted in a strong albedo change but in others this comes later, maybe even after the 70 years of the 1% run that is used to quantify TCR? Without a mechanistic explanation, historic vs future trend relationship could potentially be pure 'luck'. I would argue that a type 1 constraint is, without a mechanistic explanation of the underlying processes, much less robust than an EC that identifies the driving process.

Thanks for this point and its illustration with a detailed example. We have reassessed our statements about type 1 constraints. We fully agree that the presence of a tipping point or mechanism for the transient trend to alter would weaken a first order constraint. Indeed, we already note this when we introduce the concept.

We do agree, however, that our level of confidence on the TCR constraint may be too high in places, and we've edited section 5.4 to convey this. However, we do not see evidence of deviation in either CMIP ensemble of near-linear warming response to linearly-increasing forcing. See Gregory (2015), and the following plot, which shows in CMIP5 and CMIP6 there is a very strong relationship in both ensembles between warming at 400ppm (approx. present day) and warming and TCR at 560ppm (2xCO2) in the 1pctCO2 simulations, with a slightly weaker relationship for warming at 4xCO2:

This points to at least the potential for extrapolation of a forced transient trend in CMIP ensembles in the idealised case of the 4xCO2 simulation, potentially complicated by forcing uncertainties in a historical simulation. However, the point is well taken that this is an empirical observation (we believe the constraint because it's there) – and there is strength in your argument that a process-based constraint allows for a deeper investigation for responses like TCR and ECS, which are driven by multiple mechanisms. We have updated the text to reflect this better.

What intrigues me in this example (Tokarska et al., 2020) even more is that the slope of the EC changes from CMIP5 to CMIP6 by around 100%. So, an overestimation of the historic warming by 0.1°C would have a twice as large effect on TCR. This suggest that other processes are in place and merits more discussion and that two relationships are found in both model ensembles but apparently not the same. This seems to remain undiscussed in that paper and in this review, which assesses other types of constraints much more critical. Having said all this, I find the claim that type 1 constraints are more robust unfounded.



This was an interesting observation, and we briefly followed it up. The plot above shows that the relationship between transient response on different timescales is strong in both CMIP5 and CMIP6 in the 1pctCO2 simulations.

More important than the slope itself is perhaps the fact that the relationship between past warming and TCR in CMIP5 showing in Tokarska 2020, is much weaker – with a 0.52 correlation – so confidence in the slope is lower than the CMIP6 case (corelated at 0.74). We agree that the lack of strong relationship in CMIP5 is a relevant issue, especially given the addition of the CMIP5 to the CMIP6 ensemble weakens the constraint seen in CMIP6 alone. We've adjusted the text to reflect this.

We would expect there to be two major factors in the difference between the two ensembles: CMIP6 containing some models with higher TCR values than CMIP5, and there are potentially differences in historical forcing trends between the two studies. However, following this up further is a study in itself.

2) On the other hand, type 2 constraints are more criticized although they identified the leading process. It states, "A plausible, robust, process-based EC is still conditional on the plausibility of the relevant process as it is represented in the class of models used in the ensemble."

We stand by this sentence, but we take the reviewers point – that the plausibility of process representation can be reinforced by additional observations. We've added the following qualifier:

"However, confidence in process representation can be assessed and potentially increased through consideration of the plausibility of common model assumptions (Klein and Hall, 2015) or identification of independent observables which can be used to assess the degree to which models represent relevant processes (Terhaar et al., 2020)."

In many cases, these processes are demonstrated in observational studies. For example, Terhaar et al. (2020) have found their EC after observational studies indicated that deep water formation in the Barents Sea is responsible for most of the anthropogenic carbon inventory change in the Arctic Ocean. From my perspective, this is more robust: Identifying with observations the dominant process for a projection, see how this process is represented in models and see if the observational-based hypothesis holds in models. If this is the case, the EC should be considered very robust.

We certainly agree that process understanding based on observations is highly desirable. However, process-based emergent constraints are still fundamentally based on the differences among model simulations, so in this sense models remain fundamental to the interpretation of the constraints. Furthermore:

(1) this sequence of inquiry is not universal –it remains possible to retrofit a plausible process hypothesis upon the discovery of a constraint, and it is difficult to objectively assess from the published literature whether this has been done, given the primary quantitative evidence for an emergent constraint is generally presented as the constraint itself.
(2) Though the example you present of Terhaar (2020) is compelling (that the base-state observationally derived understanding of Arctic water transport provides a good conceptual model for an emergent constraint based on the persistence of ocean circulation), in this case, a simple model expectation (expected deep ocean carbon transport given persistence of circulation biases) is supported by the ESM relationships. In this case, confidence in the EC is boosted by the persistence-based hypothesis, and the presence of the EC in the ESMs is

confirmation that nothing unexpected is happening in this class of model. We would still argue that confidence in the EC is ultimately conditional on the ESM process representations being complete and accurate.

That said, we do agree that a strong hypothesis, with supporting observational evidence, allows the confidence to be built in process-based ECs through supporting, potentially independent means.

3) This leads me to my main criticism, which was in my opinion not sufficiently addressed in the responses. The authors 'unload' model shortcomings almost entirely on the ECs and argue that model projections do not have uncertainties.

However, the IPCC report and multiple studies use the standard deviation across a model ensemble as uncertainty and the mean of the entire ensemble (or a subsample after excluding physically wrong models) as the best estimate. I hence strongly disagree with the response that MME do not make a statement about uncertainties. Like all scientific studies, the method of EC is not perfect and never claims. It, however, can help to analyse a model ensemble and to learn about its strengths and weaknesses.

This point is well taken. The use of the MME distribution mean and standard deviation in assessment is indeed making implicit assumptions about how the model distribution relates to uncertainties. We agree that treating the CMIP distribution as a proxy for uncertainty is problematic – and this is also appreciated in the IPCC assessment process. AR5 was quite specific with uncertainty language while regarding to the MME - and referred to the CMIP distribution as a range or spread, but generally not as an uncertainty. AR6 is also not going to take CMIP6 as an uncertainty because the distribution of CMIP6 ECS contains a significant fraction of models above the assessed likely range.

We also agree that ECs can be a powerful tool in identifying model feedback processes and relevant observables, and potentially for understanding ensemble limitations – and we have revised our discussion to highlight that these uses of ECs can be a powerful way to understand the ensemble. Our critique is the use of ECs in isolation as a tool to confidently narrow projections while potentially ignoring other aspects of model performance.

I think a fundamental misunderstanding between the authors and me is the way we interpret emergent constraints. To me emergent constraints help to analyse existing model outputs. To that extent they cannot erase strong shortcomings like missing processes (in most cases). They can however reduce uncertainties in the model ensemble because of how the existing knowledge is numerically represented (see lambda example). They hence reduce the uncertainties in existing projections and can account for an identified bias, but they may miss biases if all models do.

Thanks for this point. Firstly, we are in agreement with the reviewer that the biases and approximations ultimately exist in the models, and not in the constraints themselves. We also agree that emergent constraints are potentially a powerful way to understand diversity in model results.

However, our primary point is that the use of emergent constraints to reduce model projection spread must be treated with caution because *if* models make common simplistic assumptions (as in some cases, we know they do – e.g. eddy diffusion parameterisations, soil temperature respiration relationships), then (1) an EC may emerge because there are few degrees of freedom in the common model structure repeated in the ensemble, (2) calibrating a projection using this EC is then conditional on those common

structural assumptions and (3) the EC framework disregards other observable quantities which might highlight the deficiencies of the model parameterisations.

We fully agree that ECs are useful to isolate potentially relevant observable processes for feedback processes in MMEs, but we differ on whether this information should then be used in isolation to constrain the MME distribution of projections. As we argue in the discussion – multi-metric skill scores and model weights represent one extreme (many metrics, no consideration of relationship to response), while emergent constraints represent the other (one metric chosen because of its correlation to response).

We argue that both approaches are non-ideal, and that the more defensible middle ground has been underexplored. Constraining projections based on an EC is a very strong statement that only the EC variable is relevant in our assessment of the plausibility of different values of the response variable, and all other model performance metrics can be ignored. However, it is only by consideration of multiple model metrics, and trade-offs between different calibration targets, that model structural errors become apparent (see e.g Hourdin 2017 or McNeall 2016). This multi-metric perspective highlights that complex models cannot be tuned to match all observable targets simultaneously, and by restricting our consideration to only one variable, we would get an overly confident projection of the future. Clearly – a simple skill score/bias weighting also has disadvantages, with no focus on aspects of model response which are relevant to the projected quantity.

As such, we agree with the reviewer that emergent constraints can help us analyse ensembles, and that they should inform which variables should be included in model evaluation. What we argue against is the use of emergent constraints as a direct means to constrain model projections, implicitly ignoring all other possible evaluation metrics as well as the known process assumptions in the component models.

I will try to make my point clear with the simple example in the paper. First, the authors show that not assuming the deep ocean can lead to a wrong constraint on T280 warming by using the T70 warming. This is indeed the case; however, it is not an issue with the emergent constraint. The problem lies in the models that do not consider a deep ocean. If our knowledge does not include the deep ocean, we would expect the warming as simulated by model 1. The difference lies thus only in the difference of lambda, which itself would depend on how the feedback mechanisms are calculated. Within that model ensemble with very different lambdas, knowledge of the 'real' lambda or T70 would indeed improve the projection of that model ensemble. The EC would give the most likely projection assuming no deep ocean exist and hence improve the projection of such a model ensemble. This is, however, not reality, but that is due to the models and not the way these models are analysed. The example in this manuscript is hence rather an example why model shortcomings are a problem and not emergent constraints.

In the case of the simple model example, we agree that the bias exists in the shallow ocean model. However, in this simple case, we disagree that the EC derived in the simple model relating T70 to T280 would improve the projection. The use of that EC in that model would result in a constrained projection which excluded the real value of T280 (where 'real' is in this case the two-layer model). Use of the EC alone to constrain projections would therefore result in a failed forecast where the truth lies outside the constrained distribution. This failure happens because the EC does not factor in uncertainty due to the model structural errors.

In this case, the identification of the EC is still useful because it helps us understand the degrees of freedom in the model and the processes which govern its long-term response. It even helps us identify the model's structural limitations, given it illustrates that the ensemble does not represent a plausible diversity of equilibration responses. However, the example underlines that the use of the EC alone in this ensemble to constrain projections of long-term warming would result in a confident wrong answer – and our argument highlights the risk of this type of error.

4) In the Conclusions, the authors argue that EC is effectively model weighting. I disagree. No model is weighted when using emergent constraints. An important mechanism is identified for a projection and that mechanism relates the predictor and predictand in the same way across all models, they are equally weighted.

This is a misunderstanding - apologies, we have clarified our position. Our argument is not that weighting is used in the derivation of ECs, rather that their application to constrain projections is a form of weighting. We agree that almost all published ECs weight each model member equally in the derivation of the relationship between predictor and predictand. But ECs usually present a calibrated projection conditional on the observed value and uncertainty range of the predictor, usually using the ensemble as a transfer function. This is effectively weighting the projected values of the response according to modelled skill in the predictor.

5) In general, I have the feeling, that the authors and me agree but that that the assessment of EC depends on the variable that is constrained. A local advection driven process, such as C uptake in the Southern Ocean, can by observations and models be linked to the formation of mode and intermediate waters. ECS or TCR are, however, depended on many different variables and unlikely to be constrained by one single process. I think this difference should be emphasized more strongly, especially given the Conclusions about multi-variable metrics. Overall, I feel that the ECs that constrain a local process are being taken prisoners by the often-spurious ECS constraints.

Completely agreed – the composite response of a complex system is subject to a different set of considerations than a single process component, the latter enabling a clearer assessment of model assumptions and performance. We already discussed the difference between 'top-down' and 'bottom-up' emergent constraints in the discussion – we agree that Terhaar 2021 is an excellent example of the latter, and have included the citation.

6) Could you add prediction intervals and r2 values on figure 1 please. That would help a lot.

Agreed, done.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J. C., Balaji, V., Duan, Q., ... & Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, *98*(3), 589-602.

McNeall, D., Williams, J., Booth, B., Betts, R., Challenor, P., Wiltshire, A., & Sexton, D. (2016). The impact of structural error on parameter constraint in a climate model. *Earth System Dynamics*, *7*(4), 917-935.