Earth Syst. Dynam. Discuss.,
https://doi.org/10.5194/esd-2020-49-RC1, 2020

Earth System Dynamics
Discussions

Interactive comment on "Emergent constraints on Equilibrium Climate Sensitivity in CMIP5: do they hold for CMIP6?" *by* Manuel Schlund et al.

Peter Caldwell (Referee)

caldwell19@llnl.gov

Received and published: 19 August 2020

This study confronts 11 emergent constraints on ECS with CMIP6 data for the first time. The skill of most of these constraints collapses when faced with new data. All constraints predict higher ECS based on CMIP6 data relative to CMIP5 because many CMIP6 models have higher ECS yet similar constraint values. Overall I thought this paper was excellent, well-written, and very worthy of publication. I think the statistical significance methodology is inappropriate, however, which will require substantial revision. Otherwise I have a somewhat large number of fairly minor comments.

Major comments (in order of importance):

1. I'm uncomfortable with your bootstrap statistical significance testing method in sect

2.3.

1a. Your definition of statistical significance as "the sensitivity of the regression model to changes in the input data, i.e. the removal or addition of datasets" seems overly narrow. I think of statistical significance in this case as the probability of obtaining a correlation magnitude of at least r under the null hypothesis that no real correlation exists. By using such a narrow definition, I think you've avoided thinking about whether your methodology fully captures all sources of uncertainty. Your test is also weird in that it lacks any sense of a null hypothesis that there is no real correlation. Instead, you seem to just be taking the correlation obtained with all models and creating a histogram of possible values for it by recomputing correlations with models added or removed. I don't think this is appropriate.

1b. I suspect your bootstrapping results are strongly dependent on arbitrary sampling design choices: removing a model or adding multiple copies of a model makes a huge difference to your regression when you only have ∼30 models in the CMIP archive with data for a particular constraint. Thus I expect the number of random samples you draw to have a big impact on your histogram of bootstrapped correlations. To disprove my complaint, you could create histograms with M-2, M-1, M, M+1, and M+2 models. If these all look the same, then my criticism is misplaced.

1c. It seems more defensible - or at least complimentary - to use a T-test as described in https://atmos.uw.edu/∼dennis/552_Notes_3.pdf. If you did use the T-test, would you get similar results?

2. You don't mention any of the limitations of your methodology until the conclusions section. This left me reading through the methodology and results sections under the impression that you were unaware of the possible flaws with what you were doing. It would be helpful for readers if you describe potential problems with the methodology in the methodology section so readers won't traverse the paper thinking you don't know what you're doing and so they can interpret your results with an appropriate level of

skepticism.

2a. In addition to the problems with your methodology you currently mention in the conclusions, it is probably worth also mentioning that giving the models which agree worst with the observed constraint value equal weight in determining the regression is probably a bad idea. This issue is explained nicely in Brient (2020; https://link.springer.com/article/10.1007%2Fs00376-019-9140-8)

3. Caldwell et al (2018; https://journals.ametsoc.org/jcli/article/31/10/3921/94898/Evaluating-Emergent-Constraints-on-Equilibrium) tested 5 constraints trained on earlier CMIP ensembles on CMIP5 data and found that 4 of these constraints (Covey, Trenberth, and Fasullo D and M) also failed when confronted with out-of-sample data. In that context, I see your paper as a follow-up to the Caldwell paper. I think it would be worth mentioning this around L60. It is interesting that Volodin was trained on CMIP3 data but also holds up for CMIP5 and CMIP6 data.

4. On a related note, I felt you undersold the importance of Zhai failing when confronted with CMIP5 data which it wasn't originally trained on. A similar thing happened in Caldwell et al 2018 with the Qu constraint. Such sensitivity to sampling details seems to me an important indicator that the number of models we have in the CMIP archive is insufficient for making robust conclusions about the credibility - or lack of credibility - of the constraints we propose.

5. Your introduction argument that ECS hasn't changed in 40 yrs feels dated in light of Sherwood et al (2020; https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019RG000678). I know you didn't mention this study because it wasn't accepted when you submitted the paper, but should be cited in the revision.

6. L116: I'm pretty sure P(y|x)*P(x) can be written as a Gaussian function and therefore evaluated analytically rather than numerically integrated. You might have to use the fact that eˆ{x + C} = eˆx * eˆC for some constant C in conjunction with completing the

squares to manage this. This comment isn't a big deal - numerical integration is fine - but analytic integration is more elegant.

7. L118: I don't understand why you need to assume P(y|y0)=P(y0|y) in eq 6 and therefore that the prior is uniform. Perhaps you could explain this in more detail. As I see it, you are just assuming y has a Gaussian distribution with mean \hat{y} and variance sigma_{x_0}. These are definitely big assumptions, but don't imply a prior.

8. I got a bit lost regarding which of your results depend on the Gaussian approach of sect 2.2 for what results use the bootstrapping of sect 2.3. Am I correct that the left panels of Fig 2-5 use linear regression and the standard error, the middle panels of these same figures use the Gaussian approach and everything else is based on bootstrapping? It would be useful to mention at the end of sect 2.2 and 2.3 what figures use the methodology just described.

9. L156: how did you choose the 11 constraints you evaluate? Readers may think you cherry picked the constraints that behaved poorly if you don't say explicitly why you chose the ones you did.

10. L276 says Volodin was the first emergent constraint on ECS, which isn't true. Covey et al 2000 and Knutti et al 2006 provide earlier emergent constraints.

11. L433: Bretherton and Caldwell (2020; https://journals.ametsoc.org/jcli/article/33/17/7413/Emergent-Constraints-for-Climate) provide a multivariate technique for combining constraints on ECS. Doing so provided less conceptual insight than I expected - having most constraints predict high ECS led to the combined estimate also having high ECS with narrower uncertainty... which seems obvious in retrospect.

Minor Comments:

1. L18 "which stem the major source" is wrong. I think you mean "which is the major source"?

2. You often say things like "the emergent-constrained best estimate". "Emergent-

constrained" doesn't make sense. I think you mean the "emergent-constraint-constrained".

3. L66: you already gave the range of ECS in the previous line, so saying CMIP6 models exceed 5K is redundant/unnecessary.

4. Eq 3: x should either include or exclude "_m" on *both* sides of the equation.

5. Eq 8 uses P(y|x) from eq 6, which says it is an equation for P(y|x_0). I think eq 6 is really true for all x rather than just the observed value x_0. I suggest you remove mention of x_0 everywhere before eq 6.

6. Sect 2.3: does it really take 100,000 samples to characterize uncertainty in a correlation between the 20-50 samples you're getting from the CMIP archive? I would guess 1000 iterations would be sufficient.

7. L167: "Temperature (ERSST) is used": I'm confused because I thought you said you used HadISST on L164. Are you saying that the Brient + Schneider used ERSST?

8. L351: I've never seen the "(here: ...)" nomenclature you use. Do you mean "(e.g. ...)"?

Interactive comment on Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-49, 2020.