Clim. Past Discuss.,
https://doi.org/10.5194/cp-2019-162-RC2, 2020

Climate of the Past Discussions

Interactive comment on "A Bayesian framework for emergent constraints: case studies of climate sensitivity with PMIP" *by* Martin Renoult et al.

Anonymous Referee #2

Received and published: 12 March 2020

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 S_i, T_i, for $i = 1, \ldots, n$ simulators where we use the model:

$$S_i = \alpha * T_i + \beta + \epsilon$$

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

$$T_i = \gamma * S_i + \delta + \gamma$$

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.

Apart from this major clanger (on my behalf or theirs), there are a load of other minor points: - 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? - 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. - Also in that sentence it says '. . .'then an observation . . .'. An observation of what? - L95 should be $N(0, \sigma^2)$ to match standard notation. This mistake is made throughout. There's similarly a bizarre use of $\in fromsettheorytowrite \epsilon \in N(0, \sigma)$ which I think should be $\epsilon \sim N(0, \sigma^2)$ everywhere. - 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 - L158 the use of sequential updates appears for the first time but I can't really see why this is relevant or used elsewhere? - 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 - PlioMIP appears in L200 without being mentioned before - 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. - L270 the posterior distributions of what? - 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.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2019-162, 2020.