Clim. Past Discuss.,
https://doi.org/10.5194/cp-2018-87-RC2, 2018

Climate
of the Past
Discussions

Interactive comment on "Combining a pollen synthesis and climate simulations for spatial reconstructions of European climate using Bayesian modelling" *by* Nils Weitzel et al.

Anonymous Referee #2

Received and published: 12 November 2018

This is an interesting manuscript which presents a sophisticated Bayesian approach to model-data synthesis. I have some minor criticism of the presentation which often appears poorly-organised to me. Specifically, many parameters and concepts are used prior to their definition and introduction which made the manuscript tough going for me. That is easily fixable of course.

A more fundamental concern I have is demonstrated in the results, which (unless I have misunderstood something) show a strong degree of instability and overconfidence. This is a common occurrence in situations where all members of a model ensemble are inadequate and a Bayesian framework is used which does not allow for

this - for instance, by the explicit use of model inadequacy or model error terms. The z weight of 0.98 when all data are used is itself likely to indicate a problem, and this feeling is only strengthened by the fact that the weight on this model was essentially 0 when half the data were used, in which event two different models are preferred. What seems to be happening here is that the method is homing in on the model(s) for which the data are most likely, while in fact these data are incompatible with any model.

I've seen this in a variety of contexts and am sure this sort of phenomenon will be familiar to the authors. A trivial demonstration of the phenomenon can be given by considering a situation in which we have two models which generate data according to N(0,1) and N(10,1) respectively, whereas the data are in fact generated by N(5,1). A naive filtering which ignores the possibility of model inadequacy will assign essentially all weight to whichever model happens to be closer to the sample mean of the data set (which is of course close to, but not precisely, 5). As more data are sampled, the weight will jump randomly from one model to the other with extremely high confidence, rather than converging to the answer that the models are equiprobable but poor, which might be found if a reasonable model error term was considered. This issue seems to me to undermine the results and details of the application in a fairly fundamental manner but I would hope that the authors could revise their method to adequately account for it, ie by accounting for model inadequacy in a formal manner.

Related to this, the Brier score analysis seems to indicate that the posterior is generally closer to the data than the prior, but it does not indicate whether the posterior is valid in the sense of having reasonably calibrated uncertainties. Also, while the data are in taxa-space, the aim of the paper is actually to recreate a climate, so it is important that the validity of this estimate is assessed. Perhaps a cross-validation (in which one model is used as a target) could be tried.

In summary, I don't want to reject the paper, it is interesting and useful and presents an attractive framework for paleoclimate state estimation. However, I am not convinced that the details of the application as presented here are adequate, nor, therefore, can

I believe that their results are credible. I urge them to modify and extend their work in order to address this.

Minor comments:

As I mentioned, the ordering of some of the paper made it hard work for me.

The prior is introduced in Sect 2.2 and Fig 1 but only explained in Sect 3.2. The variables w and z are first mentioned in 3.1, but for their definitions we are referred forward to 3.2. It is only after substantial usage of the variables, right at the end of 3.2, that we are informed that z was only introduced as a help parameter, with its definition and explanation again pushed forward another two sections to 3.4 with the largely unrelated discussion of the transfer function (two words for this please, and note also that burn-in should be hyphenated) intervening in section 3.3. I would have found it far more straightforward to have had the variables explained as they were introduced.

The multinomial with n=1 would I think be better described as the categorical distribution.

In 3.4, "alternately" usually refers to two alternates, not a sequence. "Sequentially" might be better.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2018-87, 2018.