Review: Efficient Reconstructions of Common Era Climate via Integrated Nested Laplace Approximations

The authors present a nice application of statistical methods to paleoclimate. They explore different models for reducing dimension of the proxy dataset and explore the impact of different inference algorithms and the use of forcings to inform the reconstruction. The paper is interesting and important, but could use some more focus. For example, it is unclear what the primary thesis of the manuscript is. Is it the dimension reduction? INLA? or the forcings? Throughout the manuscript, it seems like each of the above is the **primary** focus and then later the topic is dropped. I think a revision that provides more focus and clarity about these points is important.

1) Describe the proxy data better. Is it annually resolved tree ring data, pollen data at irregular resolutions, etc? Has the data been standardized? Are the response of the proxies to climate linear? time uncertain?

2) On lines 44-45 of page 3, you say "to ensure reproducible workflows." I suggest you do the same and post the code as an appendix or on a online repository.

3) By doing "screening" of the data, the inference is no longer truly "Bayesian". I'm always skeptical of picking and choosing proxy data. Perhaps you are missing a signal by throwing out the baby with the bathwater. Perhaps you could assimilate the proxy data with each record having a different variance. E.g. for proxy j at time t with proxy data value $y_j(t)$ and climate variable $X(t)$, $y_j(t) = \beta_0 + \beta_1 X(t) + \varepsilon_j(t)$ with $\varepsilon_j(t) \sim \mathrm{N}(0, \sigma_j^2)$. Then, important proxies will have a smaller σ_j^2 and unimportant proxies will have their influence reduced due to a larger σ_j^2.

4) Why not just go fully Bayesian and reconstruct the proxy data that is unobserved given the observed proxies and the model? Why try to do the "RegEM" type of methods (nested reconstructions, etc.) and not take advantage of the full power of the Bayesian framework? The dimensions don't seem too large to me - especially because you aren't doing any Gaussian Process type estimation.

5) I have a hard time using Figure 4. Please rescale the correlations to only positive values and add numeric values to the correlations.

6) The model validation and comparison is interesting. What is the point of using a proper scoring rule if you are using an estimated reconstruction? You are just calibrating to that reconstruction regardless of whether the previous reconstruction is reasonable. For example, sPCR and PCR-mixed predict well for MSE using the borehole data but seems to do poorly in predicting the instrumental data - this seems concerning. In addition, are you calibrating to a smoothed version of your reconstruction? Then the scoring rule is likely not proper even when you are using CRPS. In addition, make sure you state that the scoring rules are negatively oriented - lower scores are better. A pseudo-proxy experiment can be useful in validating reconstructions. See Smerdon (2012) and Tipton et. al. (2016) for examples.

7) What is the computational time for the MCMC model? Was it coded in R, Python, or c++? Was the code profiled and optimized? This model seems like it should be easy to fit in Stan Nimble, or JAGS as there is nothing non-standard about the model. What about effective sample size per second? It seems to me that the model is relatively simple - Is INLA really even needed for this model? If you can do a full MCMC what is the need for INLA - especially when the INLA approximation is not performing as well (e.g. large dips in climate state in $\approx$ 1450, 1250, etc)? Is 8 hours really that much of a burden to offset the improved inference from MCMC? If this were a spatio-temporal problem with large data I would agree about INLA, but the dataset and the model are actually small.

8) page 5 line 27 - what type of splines? In the abstract you say Bsplines - is this what you use here?

9) page 11 lines 48-50 - You state the central hypothesis of the article is to test "no-forcing" to "forcing" models. Why are you just now introducing this here? If this is the **central thesis** it needs to be front and center.

10) There are many undefined acronyms in the abstract. (INLA, MCMC, HadCRUT4, etc.). Make sure you define all acronyms in place.

References

Smerdon, J. E. (2012). Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. Wiley Interdisciplinary Reviews: Climate Change, 3(1), 63-77.

Tipton, J., Hooten, M., Pederson, N., Tingley, M., & Bishop, D. (2016). Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models. Environmetrics, 27(1), 42-54.