Clim. Past Discuss., 9, C306–C311, 2013
www.clim-past-discuss.net/9/C306/2013/

Climate of the Past Discussions

Interactive comment on "Bayesian parameter estimation and interpretation for an intermediate model of tree-ring width" *by* S. E. Tolwinski-Ward et al.

Anonymous Referee #2

Received and published: 2 April 2013

I am happy to be part of a discussion around Bayesian model calibration, a process-based tree-ring model and the more global context of the use of mechanistic models for climate-proxy modelling and climate reconstruction. I am convinced that this avenue for new climate reconstructions should be more deeply considered.

This article presents the Bayesian calibration of a tree-ring growth model and a validation of the calibration method. The tree-ring growth model called VS-Lite has been extensively presented in another paper (Tolwinski et al. 2011). I think there are several things that could be developed and are of interest for the CPD audience but the article, in its present form, is unconvincing about the validation of the method

and the lack of an interesting discussion makes the article purely technical for a palaeoclimatology audience. In the following I explain the major comments I have and I propose ways to improve the article.

1. I disagree with the validation of the calibration method ; what the authors call a "pseudo-proxy experiment". In the validation experiment the authors omit to simulate the "model error" when simulating the "pseudo-proxy". This hampers the validation of the full calibration process. Let recall the way validation is performed:

- Four VS-Lite parameter values are selected, say $(\hat{T}_1, \hat{T}_2, \hat{M}_1, \hat{M}_2)$.
- A deterministic dendro signal $(\hat{D}_{t_1}, ..., \hat{D}_{t_n})$ is simulated using $(\hat{T}_1, \hat{T}_2, \hat{M}_1, \hat{M}_2)$ + a climate chronology.
- A statistical model made of VS-Lite + independent Gaussian error with a variance σ_W^2 is estimated.
- The hope is to recover the four selected parameters $(\hat{T}_1, \hat{T}_2, \hat{M}_1, \hat{M}_2)$.

In a pure optimisation framework this method is acceptable : authors find the 4 parameters allowing to best match the simulated chronolgy $(\hat{D}_{t_1}, ..., \hat{D}_{t_n})$ and the discrepancy measuring the "match" is Gaussian. But doing so, authors are not convincing about the behaviour of their method in situation where there is model error. In other words, we are not convinced that the value of σ_W^2 can be recovered and, perhaps more importantly, how can we interpret the σ_W^2 value obtained in validation when we know that it has been simulated equal to zero ? From a theoretical point of view, the validation does not fit within the Bayesian framework authors claim to use. In such a Bayesian framework the error has a meaning and is not a trick to "match" simulations.

In the discussion, section 4, the authors propose to interpret the value of σ_W^2 in the validation as coming from "parameter uncertainty" without explaining the meaning of "parameter uncertainty". I do not think the concept of "parameter uncertainty" is either relevant or proper in a situation where "data" are controlled simulations *without* any uncertainty.

By using experiments with added error (for a pre-selected valued $\hat{\sigma}_W^2$) the validation would

- convince us that it is able to recover the $\hat{\sigma}_W^2$ value,
- allow to test the effect of the noise amplitude on recovering the values $(\hat{T}_1, \hat{T}_2, \hat{M}_1, \hat{M}_2)$
- perhaps be more convincing than what is presented in the Figure 6. Indeed, plots (b) and (c) of Figure 6 show results with a systematic positive (b) and negative (c) bias. How do the authors explain this bias? This is not explained nor discussed in the manuscript.

Reconstructions could also be done with the whole available chronology instead of cutting the chronology in two pieces for making the validation at the same time as calibration (see my last remark on mixing calibration and validation).

2. The article focuses on the calibration method without explaining the context of the calibration. This hampers a fruitfull discussion on the modelling choices that have been made and this makes the article technical and not attractive for palaeoclimatologists. The calibration using Bayesian statistics is not new, even for tree-ring growth model (see Gaucherel et al, 2008). What would make the article interesting (especially for palaeoclimatologists) would be to describe calibration as a piece in a bigger picture (e.g palaeoclimate reconstruction or proxy modelling but at a specific scale and with a

specific goal in mind). This would allow to justify the modelling choices. Indeed, I have two comments/questions on the modelling choices.

2.1 The authors independently calibrate VS-Lite using each of the 277 sites accross US. Doing so, the authors implicitly accept that the model parameters vary in space. I wonder how the authors would use VS-Lite to reconstruct past climate if it is what they have in mind? Which value of the parameter would be used for the past if they had to reconstruct climate through the inversion of VS-Lite? Do the authors think about using VS-Lite for that purpose (seemingly yes, see Tolwinski et al, 2011)? Personnaly I would have used the same set of parameter for the whole US and I would have spent more time on the modelling of the spatio-temporal error of the model. This is an example of the kind of discussion that would make the paper more interesting for palaeoclimatologists and earth-system modellers.

2.2 The authors use the simplest error model (independent Gaussian) without providing sound scientific evidences. In Section 2.2 the authors mention that studies in Tolwinski et al (2011) suggest the iid Gaussian assumption is appropriate. I do not see these evidences in Tolwinski et al (2011). Using an independent Gaussian error model implicitly assume that VS-Lite is able to simulate all the variability at a frequency lower than the year (data frequency). VS-Lite is presented as an "intermediate" complexity model and thus is expected to miss a few second-order processes. I then expect some sort of autocorrelation in the errors. Testing the presence of autocorrelation in the residuals is quite simple. I would like to find it in the manuscript (e.g a plot of temporal correlation of the difference between simulated and real dengro signal at the 277 site).

3. Two points lack comprehensive descriptions and scientific evidences that would greatly help in trusting what is stated.

- The number of runs (5000) for the Monte Carlo Markov Chain (MCMC) algorithm is low compared to what is classically used (tens or hundreds of thousands). This raises questions about (i) the convergence of the algorithm and (ii) the representativeness of the posterior estimates because only 240 points are retained to summarise the posterior distribution. It is a good point to run 3 MCMC chains and compute the R-hat statistic but It would have been convincing to also show a few MCMC chain plots (or to add them as supplementary material). Moreover, the algorithm is not completely described and this could help in evaluating the results. The authors give a code and an explanation on the website (www.ncdc.noaa.gov.siftlib/softlib.html) but they are designed for a simpler model where uniform priors are used.
- I did not find convincing proofs that, after calibration, VS-Lite fits the real data at least in a qualitative sense. The only measure of fit I see is the magnitude of the σ_W^2 which does not tell us if the model fits or not since it is not compared to the variation of the data themselves. I would really like to see a few plots of the fitted versus real data and a few computations of a R^2 or a comparable measure of fit. Note that during all the manuscript authors use and discuss the value of medians computed independently for each of the parameter. These marginal medians (for each parameter), when joined, do not form a global median! Then, when computing a fit the authors may use randomly picked MCMC outputs.

4. I appreciated the effort of the authors to make Bayesian statistics understandable for a wide audience. There are just a few things the authors could make clearer

- In the validation authors seem to use half of the interval for calibration of the parameter and the other half for validation (computing σ_W^2). If the authors choose to keep it in this form, this needs more clarification (which model is used, etc). But my advice is to separate calibration of the parameter from validation of the

variance estimate. Doing both in the same step reduces (to 29) the number of data available to estimate 5 parameter. 29 data for 5 parameters is low. Validation could also be done by modifying the model to allow two different variance parameters.

- Sometimes the authors seem to mix the Bayesian framework with the Markov Chain Monte Carlo algorithms which is not the only one method that could be used to find the posterior distribution. The authors compare numerical procedures (varying locally or globally the parameters) with a Bayesian scheme and say that the Bayesian scheme is computationaly efficient. A scientific comparison would consist in comparing the numerical procedures with MCMC directly. Moreover, MCMC may be considered efficient, but in a sense that has to be correctly explained because it is very computationally expensive.

Bibliography

Gaucherel, Campillo, Misson, Guiot, Boreux (2008) Parameterization of a process-based tree-growth model: Comparison of optimization, MCMC and Particle Filtering algorithms. *Environmental Modelling & Software* 23, 1280–1288.

Tolwinski-Ward, Evans, Hughes and Anchukaitis (2011) An efficient forward model of the climate controls on interannual variation in tree-ring width. *Climate Dynamics*, 36, 2419–2439.

Interactive comment on Clim. Past Discuss., 9, 615, 2013.