Dear editor and reviewers, we would like first to thank you for your useful feedbacks and comments on our manuscript. You can find here below the Referee's comments in *italics* and our answer in blue. In **bold**, you can find the modifications that will be made to the manuscript.

Referee#2

This manuscript presents a useful analysis of the use of the model MAIDEN as a PSM for potential paleoclimatic reconstructions. I have some minor comments, corrections, and requests for clarification.

We would like to deeply thank the Referee for the positive evaluation of our manuscript and the interesting comments. They will all be accounted for in the revised manuscript, as described here below.

I think it would be important to state more prominently that the results here come with the caveat that they are done over a limited range of climate regimes. In my experience using VS-lite, I have found large differences for Eastern North America vs. Western North America, where Eastern North America (the primary region used here) did clearly worse than Western North America. It's therefore possible that MAIDEN will be less clearly the winner in certain climate regimes.

We totally agree with the reviewer that the results come with the caveat that they are done over a limited range of climate regimes and that an analysis on a broader scale is needed to have a complete view on the performance of both models under various climate conditions. The objective here is clearly not to present an exhaustive evaluation of the two models or of our calibration method but to test our methodology on a few sets of tree-ring sites with different configurations (a network and few individual sites in Europe), so as to present our methodology. We are currently testing the methodology exemplified in the manuscript to a wider range of environmentally different sites to test the applicability of our calibration method for the MAIDEN model.

Therefore, we will state this again on lines 367-370 (p. 18), as follows: "**As our objective is to provide a first test of our calibration methodology using only a few sets of tree-ring sites, the obtained results only give an incomplete view of the MAIDEN model performance and its comparison with VS-Lite, focussing over a limited range of climate regimes. More experiments in different conditions are required in the future to exhaustively evaluate and compare the performance of both models.**"

All of the validations are done with only the correlation metric. Correlation will miss potentially important differences like a variance bias. Is this not a concern here because the time series being compared are all standardized to have no mean and unit variance?

We agree with the reviewer that our analysis do not allow estimating the variance bias. Ideally, an exhaustive quantitative evaluation of MAIDEN would require a comparison of the variable simulated by MAIDEN to represent tree-growth (which is the annual quantity of carbon allocated to the stem in gC.m^2 of forest per year) directly with observations. In this case, all biases (including on the variance) can be estimated. Unfortunately, this would, for example, imply to have observations such as tree-ring density measurements, which are less widely distributed than tree-ring widths, and to account for biases in tree-ring observations due to the chronology building process. Those biases may indeed deteriorate the comparison with what MAIDEN simulates, i.e forest carbon accumulation and not tree-ring indexes. In specific cases, we are able to compare outputs variables from MAIDEN with observations, as it is the case for example for simulated gross primary

production with eddy covariance stations measurements of gross ecosystem production (Gennaretti et al., 2017) but this is not possible in most paleoclimate applications.
Consequently, such as VS-Lite which produces a unitless tree-growth index, we have to use a simple normalization procedure, assuming that annual quantity of carbon allocated to the stem is proportional to tree-ring width observations, as stated in our original manuscript on lines 106-108 (p.4). The disadvantage is that this normalization forbids us to assess error in the variance. This is why we only analyse the correlations for simplicity as using other metrics like the RMSE would not help us in this aspect. Similarly, studies on VS-Lite such as Breitenmoser et al. (2014) or Tolwinski-Ward et al. (2011) have used correlation as a unique statistical indicator.

This will be mentioned more explicitly in the revised version of the manuscript on lines 216-221 (p.9-10):”To compare observed and simulated tree-ring growth data after the optimization of the model parameters, both observed tree-ring width series and simulated time series have been normalized to unitless indexes. **Ideally, an exhaustive quantitative evaluation of MAIDEN would require a comparison of the variable simulated by MAIDEN to represent tree-growth directly with observations. However, this would imply the use of other tree-growth observations such as tree-rings density measurements, while tree-ring width represents the most widely available tree-growth observations which makes it a relevant candidate given our global scale goals. The disadvantage is that this normalization forbids us to assess error in the variance. This is why we only analyse the correlations for simplicity as using other metrics like the RMSE would not help us in this aspect.**”.

I’m confused about the use of NRCAN data in the VS-lite model. If I’ve understood the manuscript correctly, the NRCAN data provides daily max-min temperature and precip data. But I believe that VS-lite is designed for monthly mean data. Is NRCAN (and daily max/min values) the right data to be using for VS-lite? I’m wondering if this might contribute to the reduced skill of VS-lite.

We agree with the reviewer that using daily maximum and minimum values could be a source of bias for VS-Lite. This problem has been highlighted in the PhD thesis of Alexandre Devers available online (https://www.theses.fr/2019GREAU029), on p.56 for example, where for France the average difference between daily average temperature and daily average temperature calculated from minimum and maximum temperature has been estimated to be around 0.5°C. The bias should be relatively weak and thus not impact so much the skill of VS-Lite.

The following information will be added in the revised manuscript on lines 165-166 (p.7), as follows:”**Note that monthly average temperature has been computed by averaging daily maximum and minimum temperature, which could lead to a small bias.**”

Alexandre Devers. Vers une réanalyse hydrométéorologique à l’échelle de la France sur les 150 dernières années par assimilation de données dans des reconstructions ensemblistes. Hydrologie. Université Grenoble Alpes, 2019. Français. NNT: 2019GREAU029. tel-02506254

Can the authors comment on the computation cost of running MAIDEN vs VS-lite? This is particularly relevant for paleoclimate DA where an expensive PSM might be justification enough for not using it if something else is much faster.

We agree that it is an important information to add in the manuscript. This information will be added on lines 232-235 (p.10), as follows: “**Running MAIDEN takes around 2.5 seconds on one CPU for a 50 years time span while running VS-Lite takes around 0.30 seconds. Currently, calibrating MAIDEN with our method takes around 18 hours on one CPU for a site due to the**

high number of iterations and calibrated parameters, while the calibration method used for VS-Lite and developed by Tolwinski-Ward et al. (2013) takes only a few seconds. ”

p2.l51-53 This isn’t actually true. Several reconstructions have assimilated proxy values directly using linear statistical "PSMs" (e.g., Hakim et al. 2016, Steiger et al. 2018, Tardif et al. 2018). While these are not physically-based, they still are a kind of PSM and the proxy values are not converted to temperature and then assimilated. Additionally there are reconstructions methods that have tested the direct assimilation of real isotope data using isotope GCMs (Steiger et al. 2017, Okazaki and Yoshimura 2019), and thus employed fully physically-based PSMs.

We agree with the reviewer that it has not been stated clearly in our manuscript. In the introduction, we are only talking about physically-based PSMs and this will be corrected in the revised manuscript accordingly. Also, there are indeed examples where physically-based GCMs have been used with direct assimilation but for other variables (isotopes) and not for tree-rings.

We will revise the manuscript on lines 51-53 (p.2) as follows: “**However, so far, as for physically-based tree-rings PSMs,** all tests using actual data have been based on temperature reconstructions derived from **tree-rings** proxies, not on proxies themselves.”

p3.l62-64 Is the inclusion of CO2 influences needed for Common Era paleoclimate though? Over most of the Common Era CO2 changes very little. Then when CO2 does start to matter, we have plentiful observations? Maybe there’s some other aspect of the MAIDEN model that would be more beneficial to highlight for paleoclimatic applications? It just seems like the use of MAIDEN might not be sufficiently motivated here.

We think that the inclusion of CO_2 influences is very important as models are calibrated over the recent period where CO_2 concentration has changed a lot. If we do not take the CO_2 effect into account, then it could potentially induce stationarity problems which can, ultimately, have an impact on other parameters, such as the ones related to temperature that can covariate with CO_2.

The following sentence will be added to the revised manuscript on lines 64-66 (p.3): “**As models are calibrated over this recent period, not taking into account CO_2 concentration could potentially induce stationarity problems which can, ultimately, have an impact on the calibration of parameters, such as the ones related to temperature that can covariate with CO_2.**”