Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-29-RC2, 2020

Geoscientific Model Development Discussions

Interactive comment on "Using the International Tree-Ring Data Bank (ITRDB) records as century-long benchmarks for land-surface models" *by* Jina Jeong et al.

Anonymous Referee #2

Received and published: 29 May 2020

Jeong et al., proposed a new method to use ITRDB tree ring width to benchmark Land Surface Model (LSM) in the century-long period, to enable the benchmark could be extended back to those periods well before human-induced environmental changes. This creative way of using ITRDB for a longer-term benchmarking, transiting from pre-industrial to present-day environmental conditions over the past century, could be a very useful tool for model development and is very relevant for future predictions. Because it could potentially cover both the stable (pre-industrial) and fast (recent decades) climate change period. Four benchmarks, combined with the idea of the observed virtual tree are introduced to account for the sampling bias in ITRDB tree ring data and the fact that size-related growth exceeds the one caused by environmental changes. It

is very interesting to read and learn about how and why those metrics were chosen. The paper is well written and informative. I appreciate the huge and complicated work that the authors have done.

However, my biggest question or I hope to read from this paper is why and how this new approach works. For example, the data-based evidence is needed for why the size-related trend in diameter increment should be unique enough to be used as a character to distinguish different sites with different past century's climates. Why the diameter history, which contains not only the current year's growth signal but also carries previous years bias (possibly), was used to evaluate whether model performs well in diameter increment pattern in both young and mature trees? And the European regional case study didn't give a clear conclusion for the whole benchmarks.

There are a few minor queries, especially for the four benchmarks.

I am curious about whether the simulated ring width has been tuned before the final model run by adjusting some of the parameters. Could the authors be clear about whether there is the tuning process? And if so, the way of using RMSE or difference between observation and simulation can be tricky. Because those "artificial" bias could potentially have a big influence on such RMES-based benchmarks by simply changing/tuning the level of growth.

Figure 4: more details about what is compared are needed. Is y-axis the mean of ring width?

Figure 5: The exhibited slope estimation at Panel (d) looks not that convincing. The flat slope is heavily influenced by the big continuous underestimation for the young growth. And there is an obvious downward trend since the tree getting bigger. The slope estimation could make more sense (or be more robust) if data (difference) could be randomly arranged, not by age; or if it is not showing the consistent longer-term difference in either the positive or negative way for a certain period.

Figure 6: Details to explain how the “recent year” at Panel (d) was decided is needed? And would this “cut” of data scarify the length of data availability, considering this new methodology is targeting for “century-long” model-data comparison, and the mature tree is one of the more important benchmarks in the four?

Figure 7: Some logical reason why only the first few decades (30ish) years are chosen for this benchmarking is needed. The comparison was limited within the first few decades of the time series for young trees comparison. It was mentioned because the old fast-growing trees died well before sampling took place. But actually, those “young” fast-growing trees lived through a much longer period shown in Panel (a).

Figure 8: It looks like the extreme event benchmark is the most climate-sensitivity related benchmark. However, the period is limited for the most recent years when the most reliable observed climate data is available, which is not consistent with the other three benchmarks. This somehow downsized the importance of this new benchmarking method. (Because the longer-term benchmark is one of the major breakthroughs.) Does this mean the other three benchmarks are not that sensitive to the quality of the climate data, especially to the climate variations? The extreme value was extracted from the average of the observation, without any size related detrending. Would the size-related growth have any impact on the quantile statistic? Panel (d): “mm” in y-axis title should be “Normalized”. Panel (f): how different years’ value were matched if only the quantile was applied for both observation and simulation? Is there any explanation about why the model is always overestimating the growth for both the good and bad years. Is it because the original value of TRW (not the standardized one) was used. Again, I am wondering whether there is a modelling tuning process to adjust the simulated ring width closer to the observation. I understand Panel (e) and (f) is to test the ability to reproduce the amplitude of TRW, which has also been majorly targeted by the former three benchmarks. However, it might also logically make sense by simply using the normalized value if the above three benchmarks passed. Meanwhile, relative change can be more relative to climate sensitivity comparison, if the simulated growth

was tuned.

Table 1: Wider space between each row of the table could enhance the readability.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-29, 2020.