We sincerely thank the referee for her/his very constructive review and the important points raised. We think that we can improve the quality of our manuscript substantially due to the inputs of the referee. Hereafter the referee's comments are reported in black and our replies are highlighted in blue.

From the stylistic point of view, in my opinion the paper would read better if it is moderately restructured. The description of the sensitivity experiment (section 3.2) should be located in the methodology section before the result. This would avoid the feeling of jumping back and forth from results to methods, and would help justify the presence of CLM-PLUS results in the earlier figures (before its description). I know this may feel awkward as the CLM – PLUS simulation is seen as a response to the problems (i.e. results) identified in section 3.1, but with some effort I am confident the restructuring can be done. I would introduce this idea (that a CLM – PLUS simulation is done as a response to the first results) in the last paragraph of the introduction, and then describe it in the last section of the methods, saying you are anticipating (in the text) the results that will be presented thereafter.

Answer: We agree that the experiment description in section 3.2 could be placed elsewhere. We decided to add the methodological details to the appendix dedicated to the sensitivity experiments including the detailed description of the implementation. We will focus only on the results in section 3.2 and give a brief overview of the sensitivity experiments in the method section with a reference to the appendix for the more complete details. As the referee proposes we will describe the overall goal/motivation of the sensitivity experiment in the last part of the introduction.

Another restructuring point I would strongly recommend is to try to separate Results from Discussion. The combined section currently works quite well for 'Results', as such a section should not be just a description of results but also an interpretation of them. But some parts can be moved to a more general 'discussion' section in which the whole approach is discussed in a broader sense, providing more insight of the caveats and advantages of the whole experiment, and how it relates to the broader picture in Earth System Science

Answer: This is a good suggestion. We will add a separate Discussion section in the revised manuscript. Many of the points made below by the referees will effectively provide relevant material for the discussion section.

Deforestation is more complex than a simple transition from forest to open land described in the Li et al 2015 MODIS dataset, as different types of forest (e.g. evergreen or deciduous) would have different effects (on snow masking and albedo for instance), and different kind of open lands will also behave differently (management would arguably have a strong influence). With the GETA data, the authors do explore this variability for ET to some extent. In my opinion a more thorough discussion is warranted, even if further analyses are not required within this study. Could anything be said on PFT specific differences for albedo and LST? Are there other field-based datasets such as GETA that could be used for these variables? Could other datasets from remote sensing that differentiate amongst forest types be used? If not, mentioning this need could justify and stimulate the development of such products in the future.

Answer: We agree with this point and we will introduce a more complete discussion of this topic in the new discussion section (we already touched upon the issue by mentioning the need to distinguish between irrigated and rainfed crops in p. 7 ll. 21-24). In particular, we will add a reference to the new remote sensing-based dataset by Duveiller et al., (2018) which was released only after we submitted our manuscript. This dataset - distinguishing between different forest and open land types - is a promising element towards refining the type of evaluation strategy we presented here.

There are some doubts on how comparable the deltas that are extracted from GLEAM are with respect to the Li MODIS dataset and to the CLM sub-grid simulations. If I understand correctly, GLEAM provides separate values for tall canopies and low vegetation over the same 0.25 dd pixels, and to obtain a change between 'forest' to 'openland', one makes the difference a pixel level between the value for tall canopies and for low vegetation. However, to understand better the possible repercussions this may have on the analysis, it would be necessary to have more information on how the distinction between tall canopies and low vegetation is made in GLEAM. What land cover maps are used (if any)? How do these match with the CLM distribution of PFTs?

Answer: This is an important point which we will discuss in more detail in the revised manuscript. Indeed, GLEAM and Li et al. MODIS data are not based on the same land cover information (GLEAM uses MOD44B whereas Li et al. (2015) use MCD12C1). While MOD44B provides the fraction of a grid cell covered by trees, non-tree vegetation, and non-vegetated land surfaces, MCD12C1 provides the dominant IGBP land cover type in each pixel. The non-tree vegetation fraction of MOD44B incorporates shrubland which is excluded in MODIS and our CLM analysis. This might be the cause for the much lower value of the mean ET difference in the arid climate zone for GLEAM compared to the other two data sets (MODIS and GETA).

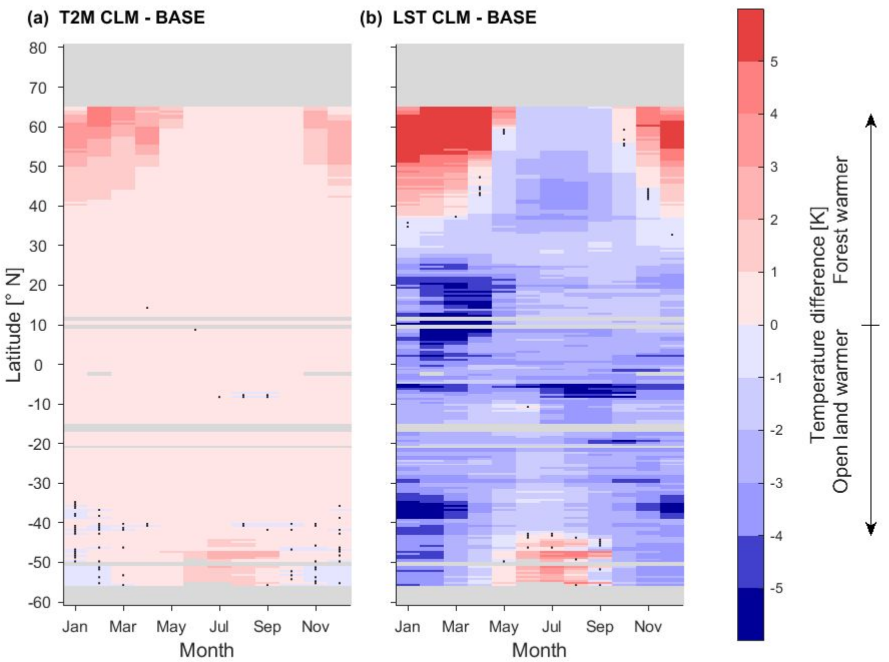
The MODIS Li et al. 2015 dataset depends on setting a threshold on the percentage of forest/trees that there are in a pixel so as to consider it 'forest' or 'openland'. They also show in their supplementary material that the choice of the threshold does have some effect of the results. How does this affect the comparison with CLM sub-grid results, for which the signal is fully 'un-mixed'? In my understanding this has the effect that the MODIS delta will often relate to a comparison from a 'not-so-full-forest' to a 'not-so-treeless-openland', while the simulations are from a 'full-closed forest' to a 'treeless openland'. How does this impact the results? Can something be done about it?

Answer: This reasoning is indeed correct. This issue was already mentioned in previous studies comparing satellite-derived albedo products with in situ measurements, as we discuss on page 6 lines 26-30 of the initial manuscript. However, this of course also applies to the other variables and we will therefore add this point in the discussion section.
Recently, Duveiller et al. (2018) established a relationship between vegetation cover fractions and different surface variables in similar moving windows as Li et al. (2015) to estimate the signal of vegetation cover changes to avoid this issue. As mentioned in our response to the 3rd comment, we encourage using this data set for future studies of a similar kind as ours and will make this clear in the revised manuscript.

Regarding the discussion on T2M vs LST in both the models and observations, an

Important point that is not completely clear is whether T2M is considered as 2m above the canopy or above the soil (i.e. within the forest). Note that in studies like Alkama & Cescatti (2016), the techniques to obtain T2M from satellite LST require weather stations, which typically use WMO definitions by which temperature is measured above a standard grass canopy, even if it is surrounded by forest. This means that the T2M obtained is not that which is observed within the canopy (i.e. under the trees) nor the one above the trees. In the model, and hence in this analysis, what temperature are we speaking about and how can the comparability between observations and models be ensured?

Answer: T2M in CLM4.5 is defined as the temperature 2 m above the apparent sink for sensible heat (Oleson et al., 2013; Eq. 5.58) which lies within the canopy air space. In the manuscript we argue that T2M is not the right temperature diagnostic to compare to LST observations (this is why we recalculated a radiative temperature (TRAD) based on the outgoing longwave radiation). We nevertheless show a comparison with daily maximum T2M difference in the appendix (Fig. A6) to highlight the different sign of the response in T2M compared to TRAD in CLM4.5. This result is surprising and is worth noting since it implies that modelling studies looking at land use effects might be affected by the choice of temperature diagnostics, which is an issue that has been overlooked in our community. That said, evaluating this T2M temperature signal in CLM4.5 against observations is very challenging since, as the referee rightfully points out, the WMO T2M concept is by definition not applicable to forest and therefore “T2M” in forest is ill-defined. For instance, the measurements of Lee et al. (2011) report “T2M” above the canopy and “T2M” in Alkama & Cescatti (2016) is indeed a mixed concept derived empirically, which is yet another definition compared to the CLM4.5 definition above. We will add these clarifications to the discussion. Also, we will replace Fig. A6 in the manuscript with the figure below to make it clearer for the reader that we do not intend to compare the T2M signal in CLM with the MODIS LST observations but emphasize that the T2M and the TRAD signal in CLM look very different.



Revised Figure A6. Seasonal and latitudinal variations of (a) the daily maximum T2M difference of forest minus open land and (b) LSTmax(f-o) in CLM- BASE.

Page 9 Line 31, Could you speculate on why the model would have this behaviour?

Answer: We did not further investigate why the model exhibits this behavior. We think that we can exclude the fact that our simulations were made in offline mode as the cause, since online simulations using CLM exhibit similar daily maximum T2M signals, as mentioned in the original manuscript (p. 8 ll. 28-31).

Could you add some info on whether this shallower root distribution is closer to what is observed, perhaps based on information from the references cited (Fan2017 & Canadell1996)? Ideally it would be good to have a line in Fig 6 for the observations over 'openland'.

Answer: We will visualize the rooting depths reported in Fan et al. (2017) in Fig. 6. Note that when creating this new root distribution, the aim was not to fit the new distribution to a particular root distribution found in the literature. It was rather a test on how the model would react to a shallower distribution for open land. In fact, the resulting root profile for open land PFTs is rather shallow compared to what is observed but still within the observational range. We will mention this in the discussion of the sensitivity experiment.

How do you calculate the confidence interval in MODIS? Do they come from the original product of Li et al. 2015? If so, do explain a bit more how they are calculated and how should the reader interpret it?

Answer: The confidence interval is the original one from Li et al. (2015). It is the two-sided 95% confidence interval estimated by a t-test. As the 2nd referee mentions, the data we plot are not normally distributed. We will thus replace the confidence intervals with the interquartile range which is more suitable to visualize the variability of not-normally-distributed data.

For all plots like that of figure 1, I am not too sure how much we gain in insight by having the fine 1dd resolution. I would recommend using broader latitudinal bins (e.g. 2.5 or perhaps even 5 dd) so as to have larger boxes in which the points of the t-test are larger and clearer.

Answer: At 0.5° resolution the points displaying the results of the t-test are indeed hard to see. We will therefore average the CLM and GLEAM data to latitudinal bands of 1° for those plots.

References:

Bright, R. M., Davin, E. L., O'Halloran, T., Pongratz, J., Zhao, K., and Cescatti, A.: Local temperature response to land cover and management change driven by non-radiative processes, Nat. Clim. Change, 7, 296–302, https://doi.org/10.1038/nclimate3250, 2017.

Duveiller, G., Forzieri, G., Robertson, E., Li, W., Georgievski, G., Lawrence, P., Wiltshire, A., Ciais, P., Pongratz, J., Sitch, S., Arneth, A., and Cescatti, A. (2018). Biophysics and vegetation cover change: a process-based evaluation framework for confronting land surface models with satellite observations.Earth System Science Data Discussions, 2018:1-24.

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., and Otero-Casal, C.: Hydrologic regulation of plant rooting depth, P. Natl. Acad. 10 Sci. USA, 114, 10 572–10 577, https://doi.org/10.1073/pnas.1712381114, 2017.

Lee, X., Goulden, M. L., Hollinger, D. Y., Barr, A., Black, T. A., Bohrer, G., Bracho, R., Drake, B., Goldstein, A., Gu, L., Katul, G., Kolb, T., Law, B. E., Margolis, L. H., Meyers, T., Monson, R., Munger, W., Oren, R., Paw U, K. T., Richardson, A. D., Schmid, H. P. Staebler, R.,Wofsy, S., and Zhao, L.: Observed increase in local cooling effect of deforestation at higher latitude, Nature, 479, 384–387, https://doi.org/10.1038/nature10588, 2011.

Oleson, K. W., Lawrence, D., B. Bonan, G., Drewniak, B., Huang, M., Koven, C., Levis, S., Li, F., Riley, W., M. Subin, Z., C. Swenson, S., E. Thornton, P., Bozbiyik, A., Fisher, R., Heald, C., Kluzek, E., Lamarque, J.-F., Lawrence, P., Leung, L., and Yang, Z.-L.: Technical Description of version 4.5 fo the Community Land Model (CLM), https://doi.org/10.5065/D6RR1W7M, 2013.