We thank the referee for her/his very useful comments which will help improve the manuscript severely. It appears that the referee is highly experienced in this field of research. In the following, we list the referee's comments in black and our replies in blue. Note that the order of the comments was changed to facilitate our response to related comments.

1) Complete the analysis of the sources of error. You test only things related to vegetation transpiration and not soil evaporation. Your data do not clearly indicate that vegetation is the main driver, and in fact show that soil evaporation could also be a dominant source of error. Just because your modifications for transpiration show some improvement does not mean that they are correct, because you could be over-compensating or over-fitting these parameter values.

page 8 lines 21-22 and 31-32: this statement is not supported by your data or the rest of this paragraph. while the visual pattern between the VTR and total is similar, the soil evap effects are compensated for by the interception effects, thus leaving VTR to dominate the pattern. but this doesn't mean that the soil evap is not a main contributor, especially in the tropics. and you mention the biases in the non-forest that contribute to this discrepancy as well. Figure 4 also indicates that the soil evaporation dominates the total ET pattern in the higher latitudes, which is where your modifications show little improvement.

page 12 lines 11-22: this indicates that your hypothesis regarding VTR as the main driver of discrepancies may not be correct. while you get improvements, soil evap remains a problem, and you may even be overcompensating with the VTR related modifications

Answer: We agree that the soil and canopy evaporation are both important for the creation of the total ET signal. In the manuscript this is not clear enough and the importance of VTR is emphasized too much. Thus, we will rephrase some of the statements in the manuscript (e.g., page 8 lines 21-22 and 31-32 and page 12 lines 11-22).
We tried to assess the importance of the individual components more objectively by calculating the pearson correlation and the index of agreement (Duveiller et al., 2016) between the monthly difference of forest minus open land of these individual components and the monthly difference in total ET. The difference in vegetation transpiration in CLM4.5 exhibits a stronger correlation over a given latitudinal band with the difference in total ET (r of 0.72) than the other two components (vegetation evaporation r of 0.33 and soil evaporation r of 0.19). Similarly, the transpiration difference exhibits a much higher index of agreement with the ET difference than the other two components of ET in the model (0.61, 0.22, and 0.23 for vegetation transpiration, vegetation evaporation, and soil evaporation, respectively; description of index of agreement in Duveiller et al., 2016). We therefore think that focusing on transpiration in a first sensitivity experiment is justified. We encourage however further investigations on the sensitivity of soil and canopy evaporation to land cover and will clarify this in the revised manuscript.

2) Please provide a metric for quantifying the effects of the modifications. Figure 7 (and the aggregate climate zones) is not adequate for demonstrating significant improvement of the results due to the modifications.
Also, while the pft level comparison with GETA looks good, the climate zone compar-

ison is more difficult to evaluate. Aggregating to these climate zones smooths out a lot of spatial variability, and may be too coarse to adequately evaluate the modifications. can you calculate a metric to quantify the effects of the modifications? what do pixel-level correlations between the model and the obs look like? are these correlations improved by the modifications? would zonal grouping make more sense than the climate zones?

Answer: We completely agree that introducing a metric will help assessing the performance of different model configurations more objectively. We tested two additional metrics in response to the referee's concern: The root-mean-squared deviation (RMSD) and the index of agreement (IA, as described in Duveiller et al., 2016). For some of the variables with relatively poor agreement (e.g. daily minimum LST difference) the IA tends to be zero or very close to zero in all climate zones. We therefore propose to add a table with the RMSD over the Köppen-Geiger climate zones to the main section of the manuscript and add a table with the IA to the Appendix.
We agree that aggregating over the climate zones smooths out some of the signal. We argue that zonal grouping will create more heterogeneous groups since there can be strong climatic variations at a given latitude as the referee mentions later in his review. We therefore propose to use the following refined climate zones: Equatorial humid (Ef and Em), equatorial seasonally dry (Es and Ew), arid, warm temperate fully humid (Cfa, Cfb, and Cfc), warm temperate summer dry (Csa, Csb, and Csc), warm temperate winter dry (Cwa, Cwb, and Cwc), snow warm summer (Dfa, Dfb, Dsa, Dsb, Dwa, and Dwb), and snow cold summer (Dfc, Dfd, Dsc, Dsd, Dwc, and Dwd). A figure illustrating the global distribution of these climate zones will be added to the Appendix.

page 6 lines 7-11: I think CLM also outputs a surface radiative temperature. Why didn't you use this?

Answer: To our knowledge, CLM4.5 does not provide a radiative temperature output. Therefore, we added a runtime calculation of radiative temperature in the code which however is calculated according to the CLM4.5 Tech-note (Eq. 4.10 of Oleson et al., 2013).

page 7 line 21: confidence in which observations? the non-outliers I assume.

Answer: The statement we try to make is that the agreement across different independent data sources gives more confidence in the fact that ET is generally higher over forest. We will reformulate this sentence as: "The considered global ET data sets however consistently exhibit higher ET over forests in most regions (Fig. 2). This agreement across the different independent global data sources gives some confidence in the fact that ET is generally higher over forests."

page 10 lines 5-6: if comparing for lee et al, why reference alkama and cescatti for the amplification? you should include the delta LST per degree from lee et al for a consistent comparison, and to show that these observations also show this amplification

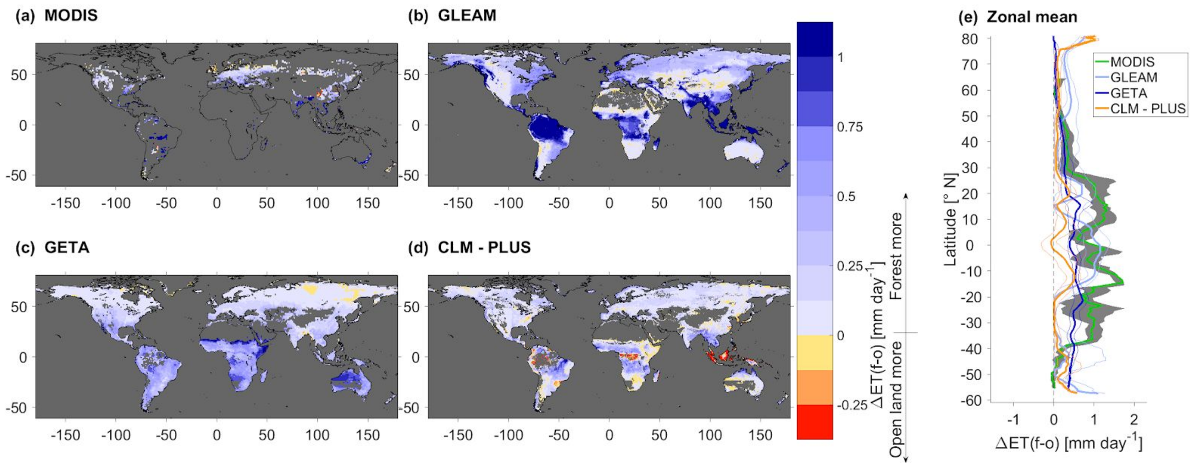
Answer: It appears that the current formulation of this section creates confusion. Lee et al. (2011) compared temperature measurements 2 m above the canopy to standard station data. To our knowledge for most of the paired sites there was no direct observations of LST in this study (all but 4). We try to emphasize here that the latitudinal dependence of the T2M difference in CLM4.5 is much weaker than in the observations of Lee et al. (2011) while the latitudinal dependence of LST is slightly stronger than the T2M dependence of Lee et al. (2011). The second part states that we expect a stronger latitudinal dependence of the LST difference compared with the T2M difference since Alkama and Cescatti (2015) observe stronger LST effects of forest in general. We will re-formulate this section to make it more understandable for the reader.

page 13 line 18: is this because you used prescribed atmospheric forcing?

Answer: We indeed considered this hypothesis (page 9 lines 27-28), but the fact that Lejeune et al. (2017) observed the same for coupled simulations with CLM tends to rule it out.

Generally, why show a CI for only the modis zonal average? What about the other data and the model outputs? And is CI the best metric to depict variability here? There are many reasons for variability around the globe at a given latitude (e.g., different weather patterns, continental vs maritime), and we should not expect a zonal mean to behave like a population mean estimate that supposedly characterizes a more homogenous group.

Confidence intervals become very narrow for CLM, GLEAM, and GETA because the zonal sample sizes are much larger. Therefore, they can hardly be seen in Figs. 1 and 5 and were not plotted in Figs. 2 and A9. We agree it is not the ideal metric to display variability here. We will therefore plot the median and the interquartile range instead of the confidence interval in Figs. 1, 2, 5, and A9. An example for such a figure can be seen below for Fig. A9.

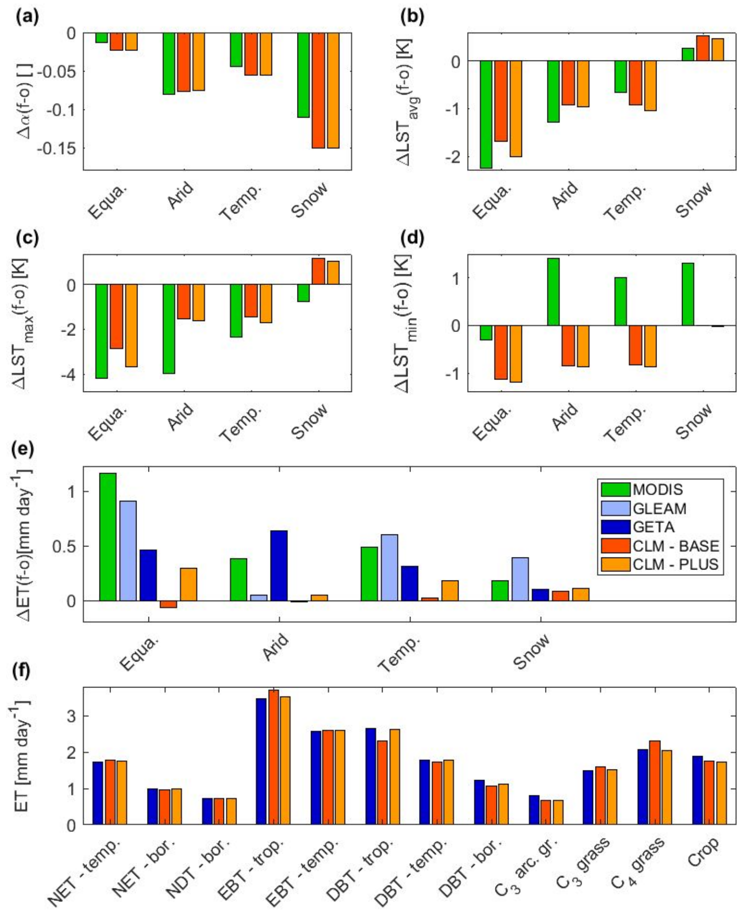


Revised Fig. A9: Annual mean ET(f-o) in (a) MODIS, (b) GLEAM, (c) GETA, and (d) CLM- BASE. Panel (e) shows the zonal median and the interquartile range of MODIS (in green along with its interquartile range in grey), GLEAM (blue), GETA (orange), and CLM-BASE (red). Note that on this subfigure results have been smoothed with a 4° latitudinally-running mean.

Figure 2a: this does not appear to be the correct figure. it does not match with the averages in panel 2c, nor table 2

Answer: This seems to be a graphical issue. For some reason there is a thin red line on the margin between data and NaN's which looks dominant when not zooming into the picture. When zoomed in the graph starts to look more blueish. We remade Figs. 2 and A9 using a different format to resolve this (See revised Fig. A9 as an example above).

Additional remark: We had to make a correction in the calculation of the mean values over the climate zones for the MODIS data. The resulting changes are small for all climate zones except for the arid zone and do therefore not affect the conclusions of the manuscript (See new figure below).



Revised Figure 7. Area-weighted annual mean over Köppen-Geiger climate zones (Kottek et al., 2006) of (a) (f-o), (b) LSTavg(f-o), (c) LSTmax(f-o), and (d) LSTmin(f-o) in MODIS (green), CLM- BASE (red), and CLM-PLUS (orange). Only grid cells containing valid data in the MODIS observations were considered for analysis of CLM4.5. Panel (e) shows the area weighted mean over the Köppen-Geiger climate zone of ET(f-o) in MODIS (green), GLEAM (light blue), GETA (dark blue), CLM- BASE (red), and CLM- PLUS (orange) and panel (f) the area weighted mean ET for each PFT analyzed in this study according to the GETA (dark blue), CLM-BASE (red), and CLM- PLUS (orange). The acronyms of the PFTs are defined in Table 3.

References:

Alkama, R. and Cescatti, A.: Biophysical climate impacts of recent changes in global forest cover, Science, 351, 600–604, 2016.

Duveiller, G., Fasbender, D., and Meroni, M. (2016). Revisiting the concept of a symmetric index of agreement for continuous datasets. Sci. Rep.-UK, 6:19401.

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.

Lejeune, Q., Seneviratne, S. I., and Davin, E. L.: Historical Land-Cover Change Impacts on Climate: Comparative Assessment of LUCID and CMIP5 Multimodel Experiments, J. Climate, pp. 1439–1459, https://doi.org/10.1175/JCLI-D-16-0213.1, 2017.

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.