Answer to comments from the reviewer #2.

We thank reviewer for the constructive evaluation of the manuscript. Please find below our answers to questions/comments. Comments from the reviewer were left intentionally in this document and written in roman font. Our answers are written in italics.

Main comments

This article presents an evaluation of global-scale heterotrophic respiration (Rh) from CMIP6 output in comparison to three different observation-based data products.

The main conclusion presented by the authors is that even though the global aggregated Rh agrees well between models and the data products, the models 'fail' at reproducing spatial patterns. The authors also provide a list of well-known mechanisms that influence soil respiration and advocate for their inclusion in new versions of the models.

Although in general I agree with the importance of model evaluation studies, I find little incremental value in this analysis. Despite the author's claim of priority, other studies have already made comparisons between ESM output and Rh data products, pointing out disagreements (see comments and references from other reviewers). The list of potential mechanisms to be included in a new generation of models, presented in the Discussion, are well-known mechanisms that influence soil carbon dynamics and Rh, and this discussion is relatively shallow regarding more relevant modeling topics such as the type of functions that should be implemented and how to obtain parameters for those new functions at the global scale. The analysis of residuals and their relation to other variables is helpful in providing some clues about the importance of these different processes, but without a more clear and systematic analysis of different mathematical functions to be implemented in ESMs, there are no elements for modeling teams to make decisions about what new functions to implement and how to obtain their parameters. For instance, this analysis identified a major discrepancy between residuals of Rh and precipitation, and the authors advocate the inclusion of hump-shaped functions in models, which is something that has been previously said (e.g., Moyano et al. 2013, Davidson et al. 2014). There are a number of such functions proposed in the literature (Sierra et al. 2015), and a more relevant discussion would be which of those functions are more relevant at the grid-size level of an ESM, and what type of observations should be used to obtain parameter values for these functions, or whether one single set of parameters should be used at the global scale or whether they should change spatially and temporally. Although I am not trying to convince the authors that they should add this discussion here, I feel that without a more in depth analysis, there is little new value in the present study.

We thank the reviewer for the constructive comment. In the revised version we will carefully explain that indeed this is an incremental analysis but we still think our study is useful and worthy of publication for three main reasons:

1. *The analysis previously published was done on the CMIP5 models generation* (Shao et al., 2013) *and this paper focuses on the CMIP6 generation. Regarding the importance of the ESMs and their impact on others sciences, using results from the CMIP6 exercise is highly important to evaluate each model generations and share the results with the scientific community.*
2. *Our study is novel because we take advantage of the gridded products that were not available before to better understand the spatial pattern of the heterotrophic respiration flux and how it is represented in the new ESMs generation.*
3. *We used a model residue approach to disentangle the main effect and this was not used before. It helps to show that bias induced by the precipitation response is at least as important as those provide by temperature response.*

Regarding the existing hump-shaped functions, it has been suggested before indeed but never done in ESMs and we consider that suggesting to the ESM developer community that some solutions might exist to solve the bias we identified is useful. Nevertheless, we agree that a more in-depth discussion might be useful.

In the revised version, we will add : "Implementing this bell-shaped function approach is necessary to accurately represent the soil organic carbon stock of peatland in some land surface schemes used by ESMs (Qiu et al., 2019). The approach proposed by Moyano et al. (2012) seems well adapted to ESMs constraint since the author proposed several versions of the bell-shaped function and did the effort to define one function using drivers that are included in ESMs (the model 2 in Moyano et al., (2012)). The model including bulk density might perform better but bulk density is not calculated by ESMs and consequently such approach is hardly implementable in ESMs. Other approaches have been proposed in the literature (Davidson et al., 2014; Sierra et al., 2014) *but the solutions proposed are mostly based on Michaelis-Menten function whereas most of the ESMs used first order kinetics approach to describe SOM decomposition. Moreover, alternative solutions are based on O2 diffusion which is more mechanistic but more difficult to implement in an ESM compared to a more empirical solution as proposed by Moyano et al. (2012). Gas diffusion implementation at the spatial resolution of ESMs is quite challenging because it depends on drivers highly variables at small scales."*

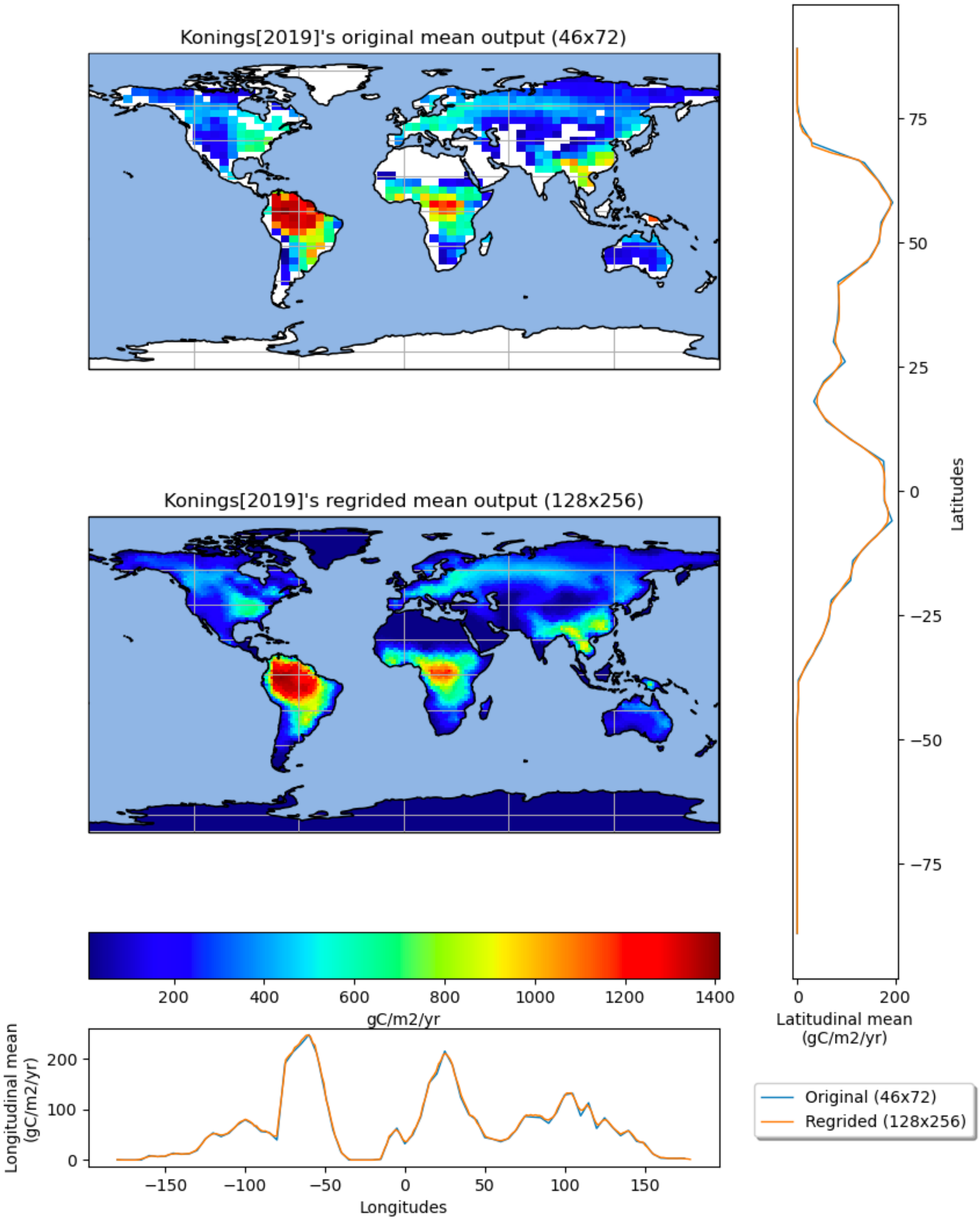
In addition, there are other topics of model evaluation that are very relevant for this study that are not discussed at all. One topic is the use of objective metrics to characterize distance between model output and data products. The authors claim that the models 'fail' to reproduce spatial patterns, but a definition of 'failure' is not provided, nor a measure of distance or probability of model output to lay in some rejection zone. A more formal analysis would be required to assess how far the model output is with respect to data-products, which are also uncertain. Throughout the manuscript the authors use the three data products as error free, but it is well-known that these products are also subjected to biases and errors. Despite their growing size, Rh databases still lack comprehensive coverage in some key regions such as the tropics. If all the models would agree well with a biased data-product, we would be very misled in our carbon-climate projections!

We obviously agree that all the observation-based products are somehow uncertain and this is why we decided to use several of them in this study. We do not have all the information to *calculate an in-depth error propagation but in the revised version, we will better quantify the uncertainties by:*

1. *Calculating the median absolute deviation (MAD) of the median of the three products*
2. *Ranking the ESMs by the number of pixels that are within the MAD*
3. *Calculate RMSE for each ESM.*

Another topic of relevance is the issue of spatial aggregation in soil respiration estimates. Since the 1990s, there has been a discussion on how to deal with aggregation errors in estimates of Rh at ecosystem and global scales (Kicklighter et al. 1994, Rastteter et al. 1992). The authors downscaled the CMIP6 output to a common spatial resolution, but it is not clear how this 'dis-aggregation' would affect uncertainties and biases.

The method we used is conservative and the main problem we faced during the regriding step was to take into account that the land fraction changed because of the regriding. The difficulties were similar for model outputs and for observation-based products. Once we corrected for the land fraction change the regriding effect was quite limited. For instance, our regrided products estimate the total heterotrophic respiration to be 50, 43 and 51 PgC yr-1 for Warner et al, Konings et al and Hashimoto et al., products respectively whereas in the original publications they authors estimated 49.7, 43.6 and 51 Pg PgC yr-1 for Warner et al, Konings et al and Hashimoto et al., products respectively. The spatial distribution was also not hardly affected by the regriding (see the figure below using the Konings et al. product to illustrate).



Konings[2019]'s original mean output (46x72)
Konings[2019]'s regrided mean output (128x256)
200
400
600
800
1000
1200
1400
gC/m2/yr
Longitudinal mean (gC/m2/yr)
200
100
0
−150
−100
−50
0
50
100
150
Longitudes
75
50
25
0
−25
−50
−75
Latitudes
0
200
Latitudinal mean (gC/m2/yr)
Original (46x72)
Regrided (128x256)

Mean Rh spatial distribution over 2010-2012 from the Konings et al., (2019) product –original (46x72, top panel) vs regrided (128x256, bottom panel).

In summary, although the results presented here are interesting to explore differences between CMIP6 Rh output with respect to observation-based data products, the authors make claims about scientific priority/novelty and 'failure' of the models that are poorly supported.

Minor comments

- L37-40. What do you mean by that these fluxes are not well characterized? Do you mean 'evaluated' instead of 'characterized'? What has been done with plant and ocean fluxes that has not been done with Rh?

We rephrase to clarify : "Despite the importance of heterotrophic respiration fluxes, the scheme representing this flux in ESMs, which aim to simulate the most important drivers of the earth's climate system, are currently challenged because important drivers are missing (Huang et al., 2021; Wieder et al., 2015) *but the proposed new schemes lacks of sufficient evaluation on long term time series (Le Noë et al., 2023). Thus, how accurate are the prediction of ESMs for heterotrophic respiration fluxes is a key question to well constraint the carbon climate feedbacks in ESMs."*

- L94. Please provide more details about 'cdo remapdis (nco module)'. What is this? A software, a package of a programing language? Can you provide a reference?

We added this information: "The Climate Data Operators (CDO) software is a collection of multiple operators for standard processing of climate and forecast model data. The operators include simple functions (statistical and arithmetic) to be used for data selection, subsampling, and spatial interpolation."

- Section 2.5. This paragraph is very difficult to understand. I get the general idea of the analysis, but I can't understand well the specific details. Please consider rewriting this section, adding more details for each step, adding some equations about how the medians and model differences were obtained, and maybe a figure describing the different steps.

We rewrote this section: "We defined here the ESM's model residuals as median of the difference between each single CMIP6's model output and the observation-based products median calculated for each grid cell. The ESM's model residuals were calculated in three steps: (i) we calculated first the median for each cell using the three observation-derived products. We consider this median as our best-estimate. (ii) Then, we calculated the difference between each CMIP6's model output and our best-estimate for each grid cell. (iii) Finally, we calculated the ESM's model residuals as the median of this difference.

Using the ESM's model residuals, we performed a statistical analysis to identify the main drivers. We proceed with a two-step methodology. First, we compared several linear generalized least square models with different spatial structures (gaussian, exponential, spherical, linear or rational (gls package, (Venables and Ripley, 2002))) and without spatial structures to estimate the effect of spatial correlation. Based on AIC values we selected the rational quadratic spatial correlation structure that had the smallest AIC values for the second step of the analysis. Then, we used generalized additive mixed model with ESM's model residuals as variable to explain and mean annual temperature (MAT), mean annual precipitation (MAP), observation derived SOC, ESM's model residuals on NPP and lithology as predictors variables. MAT and MAP are derived from the Global Soil Wetness Project Phase 3 (GSWP3) reanalysis (http://hydro.iis.u-tokyo.ac.jp/GSWP3/ last access: April 5 2022). SOC was taken from the Soilgrid250m product(Hengl et al., 2017). ESM's model residuals on NPP *are calculated as the median of the difference between ESM's NPP and NPP from the global inventory monitoring and modelling studies group (GIMMS). Lithology maps from the global lithological map (GLiM) (Hartmann and Moosdorf, 2012) was used but since lithology was not significant (p>0.05) and the model has a lower AIC without it was not included in the final generalized additive mixed model presented here. All statistical analysis were made using R v3.5 (R Core Team, 2018)."*

- L114. From what programing language is the gls package? Add a reference.

The gls package is from R as explained at the end of the paragraph. We added the Venables, W.N. and Ripley, B.D. (2002) "Modern Applied Statistics with S", 4th Edition, Springer-Verlag. Reference which is cited in the documentation of the function.

- L140-141. The median of the mean across products? or the median of the residuals after fitting a statistical model? Legend of Fig 3 says that each map is a residual. Be more specific.

We first calculated the median for each observation-based product for each grid cells. We this obtained our best-estimate spatially distributed. Then we calculated the residual for each model at each grid cell. We modified the text to clarify:

"To generate our best-estimate of heterotrophic respiration fluxes from the three observation-derived products we calculated the median for each cell. Thus, we obtained the spatially distributed best-estimate. At each grid cell, we then compared each ESM with the observation-derived products median (Fig. 3)."

- L142. I'm not sure if 'overestimate' is the right word to use here. The comparison is not directly with measured data, but with the output of a model that was informed by data. The data-products may also include biases.

We rephrase to clarify: "Compared to observation-based products, ESMs tend to overestimate heterotrophic respiration flux in tropical regions…"

- L158-159. I still don't understand how the use of first-order rates in models is connected to the need to use the median of the residuals in this comparison. Can you explain this better?

We modified in the revised version to clarify: "In order to improve predictions of heterotrophic respiration fluxes in future ESMs we need to understand the spatial biases we observed and determine their causes. To explore these biases, we performed a statistical analysis based on a generalized additive mixed model of the ESMs residuals defined as the median of the difference between each CMIP6's model output and the median of the observation-based products calculated in each grid cell (see online methods). ESMs share a very common approach based on first order kinetics with soil organic decomposition driven by soil moisture and temperature (Ito et al., 2020). This approach is derived from the very first attempts to describe soil organic decomposition with mathematical equations (Henin and Dupuis, 1945) and is still the most used to describe this process (Manzoni and Porporato, 2009; Wutzler et al., 2008). Since SOM decomposition schemes in ESMs are very similar, comparing each model individually can be redundant and not very informative and less generalizable. To allow broader conclusions and suggestions to improve ESMs performances, *we decided to perform the residual analysis on the ESMs median rather on each individual model."*

- L160-161. This set of drivers of Rh is well-know, even before Swift et al. (1979). I'm not sure why this single recent reference is relevant here.

We used the Doetterl et al. (2015) study to support this claim because it was done at global scale with a very large dataset.

- L160-163. The entire sentence is difficult to understand. Consider rewriting.

We modified the sentence in the revised version: "The main drivers of heterotrophic respiration are soil carbon availability, soil moisture and temperature, carbon inputs and mineralogy (Doetterl et al., 2015). To explain our model residues we used soil organic carbon, net primary production residuals calculated using similar methods to heterotrophic respiration flux residuals, mean annual precipitation, mean annual temperature and lithology.

References

E. A. Davidson, K. E. Savage, and A. C. Finzi. A big- microsite framework for soil carbon modeling. Global Change Biology, 20(12):3610–3620, 2014.

D. W. Kicklighter, J. M. Melillo, W. T. Peterjohn, E. B. Rastetter, A. D. McGuire, P. A. Steudler, and J. D. Aber. Aspects of spatial and temporal aggregation in estimating regional carbon dioxide fluxes from temperate forest soils. J. Geophys. Res., 99(D1):1303–1315, 1994.

F. E. Moyano, S. Manzoni, and C. Chenu. Responses of soil heterotrophic respiration to moisture availability: An exploration of processes and models. Soil Biology and Biochemistry, 59(0):72 – 85, 2013.

Rastetter, King, Cosby, Hornberger, O'Neill, and Hobbie] E. B. Rastetter, A. W. King, B. J. Cosby, G. M. Hornberger, R. V. O'Neill, and J. E. Hobbie. Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems. Ecological Applications, 2(1):55–70, 1992.

C. A. Sierra, S. E. Trumbore, E. A. Davidson, S. Vicca, and I. Janssens. Sensitivity of decomposition rates of soil organic matter with respect to simultaneous changes in temperature and moisture. Journal of Advances in Modeling Earth Systems, 7(1):335–356, 2015.

M. J. Swift, O. W. Heal, and J. M. Anderson. Decomposition in terrestrial ecosystems. University of California Press, Berkeley, 1979.