Dear Editor,

Please find our reply below. We copy your comments in full and our replies are in **bold blue**.

I thank the authors for their detailed response to the comments raised by the reviewers to their paper. The reviewers raised a few important scientific points that I think were addressed well, although I feel like some additional analysis is merited in one case. Due to this additional analysis, as well as a desire to confirm with reviewers that the revised text (combined) alleviates their concerns, I will ask one reviewer to look over the next version, and have thus requested "major revisions." However, I don't believe too much additional work is required.

One common substantive question raised by the reviewers has to do with whether the use of tiling (and choice of tiling setup) actually improves model accuracy relative to the real world. This is a very important consideration for many modeling papers, but it's less critical here, since this is a development & technical paper. As described on the GMD manuscript types webpage (https://www.geoscientific-model-development.net/about/manuscript_types.html#item2), such papers "usually include a significant amount of evaluation against standard benchmarks, observations, and/or other model output as appropriate." Here, the authors' model-model comparisons fall under the "other model output" category. They do also compare to observations, though, as illustrated in Fig. 5b.

I'm thus satisfied with the existing analyses given the authors' explanation. To clarify their goals, the authors have added and amended text in various places, removed the term "bias" in a few lines (see response to Reviewer 1, comment 4, last bullet point) and replaced "error" with "deviation" in, e.g., Fig. 4.

Another common and important point the reviewers mentioned is their concern about the abrupt transitions in disturbance regimes between the three periods in Fig. 2a-b, especially around 1985 for harvest (Fig. 2b). The authors explain in their response to Reviewer 1's comment 1 that (a) they chose to use distinct sets of derived records for these periods, as data were available; (b) that harvest is a relatively minor contribution to disturbance compared to fire; and (c) that all evaluated setups used the same disturbance history, so to some extent the possible worrying effects of abrupt regime transitions cancel out in model-model comparisons. Reviewer 1 also notes that the possible too-high harvest rate in the middle period means that the experimental results are conservative, if anything. I'm more or less convinced by the authors' arguments, but the fact that both reviewers raised the issue suggests that the authors could do more to convince readers.

To that end, I think it would be valuable to do some sort of additional experiment with at least a subset of the model setups (e.g., 1-tile/disturbed vs. 3 tiles vs. optimal) with an alternative disturbance scenario. Reviewer 2's suggestion of simply using the first observed value throughout the pre-observation period seems as good a choice as any. I suspect that we would still see strong effects of tile count, which would bolster the conclusion that the tiling scheme is important.

Thank you for your work with our article and follow-up comments. We appreciate your evaluation of our responses. We have provided a revised version of the manuscript including the additional analyses requested. A description of our further edits to the manuscript and relevant explanation follows.

We have revised the manuscript to include Figure S4 (copied below). This figure details a set of model runs using alternative pre-1985 fire and harvest scenarios. In these scenarios (figure S4a-b) pre-1985 fire and harvest are bias-corrected to match their respective means from the observed period. Based upon these model runs we can demonstrate that by 2017 the pre-1985 disturbance has a limited impact on the distribution of aboveground tree biomass in the model domain across our various simulation setups (figure S4c). We are also able to demonstrate that due to the construction of our model-on-model comparisons, the differences in the normalized response metric ($\Delta\bar{X}_{norm}$) between the runs using the alternative disturbance scenario ($\Delta\bar{X}_{norm,alternative}$; figure S4a-b) and the original disturbance scenario ($\Delta\bar{X}_{norm,original}$; Figure 2) are an order of magnitude lower than those shown in figure 6a/S3. The alternative disturbance scenario has a similar impact on all the simulations utilized in calculating $\Delta\bar{X}_{norm}$, which is then removed via the normalization (see eqn 12). These follow-up analyses demonstrate that our conclusions regarding the dynamic tiling parameters and the relative impacts of subgrid-scale heterogeneity are robust.

We have revised the main text to refer to these additional analyses and our conclusions therein. We have added a paragraph to the results near line 530 detailing the implications of this further analysis:

> **"The relative impacts of subgrid-scale heterogeneity demonstrated herein are robust given our model-on-model approach, which cancels out pre-existing biases. This is evident in model runs using alternative pre-1985 fire and harvest scenarios that are bias-corrected to match their respective means from the observed period (Figure S4a-b). These alternative scenarios have a limited impact on the modeled statistical distributions of aboveground tree biomass (figure 5a, S4c). Moreover, because our statistical analysis focuses on the period during which disturbance observations are available (1985-2017) and because of the statistical metrics utilized in our model-on-model comparisons (eqn 12), the differences in $\Delta\bar{X}_{norm}$ with these alternative scenarios (Figure S4d) are an order of magnitude smaller than shown in figure 6a. Our evaluations provide insights into the impacts of subgrid-scale heterogeneity alone, further reinforcing its importance and the value of representing this heterogeneity within models."**

We have also added a reference to figure S4 in the conclusion near line 562:

> **"These results are strengthened by our model-on-model approach, which acts to cancel out pre-existing biases, to demonstrate the impacts of subgrid-scale heterogeneity, and discretization error alone (Figure S4)(Torres-Rojas et al., 2022;**

Curasi et al., 2022; Melton et al., 2017; Melton & Arora 2014; Moorcroft et al., 2001)."

Figure:

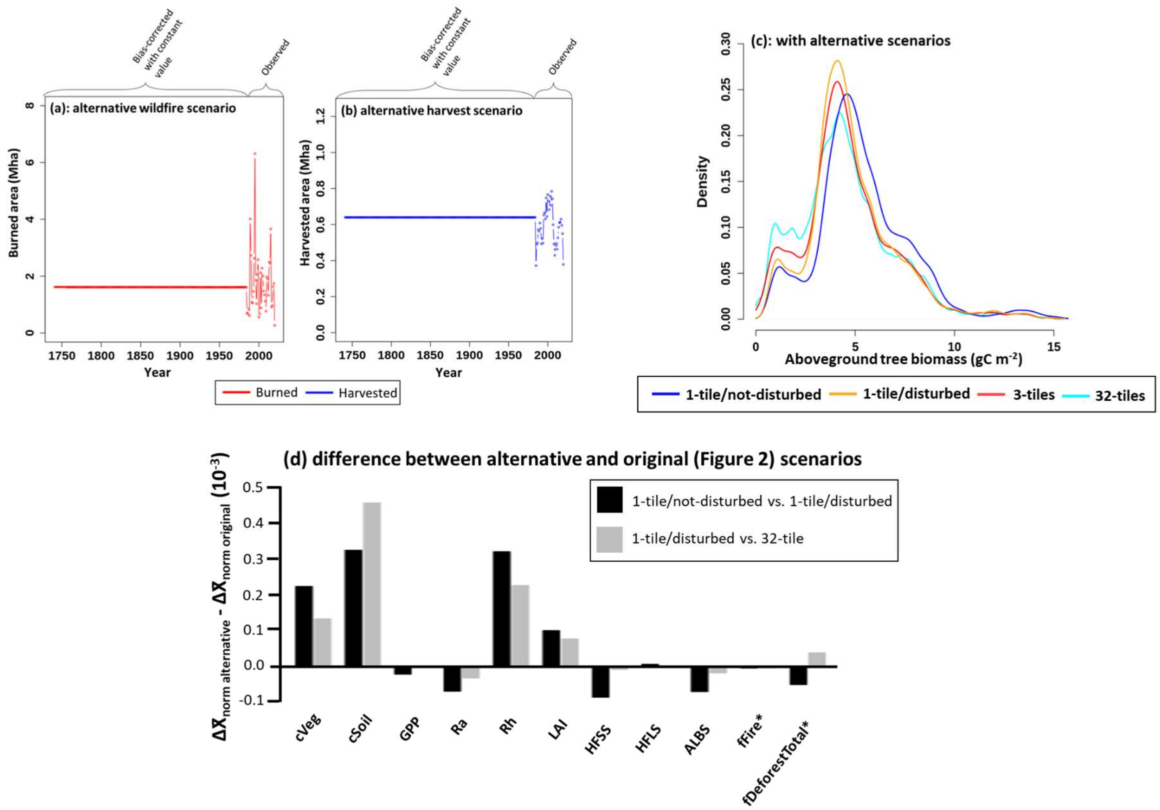

Figure S4: Model runs using alternative historical harvest and fire scenarios. Plots of the alternative disturbance drivers over time. a) annual total burned and b) harvested area from 1740 - 2020. Observed indicates the period that uses the Landsat fire and harvest observations (Hermosilla et al., 2016, 2015a, b). Bias-corrected with constant values refers to the period where the inferred disturbance was bias-corrected using the mean of the Landsat observations. c) Weighted histogram of aboveground tree biomass for forested areas of Canada at the end of a selection of model runs including the 1-tile/not-disturbed run, 1-tile/disturbed, 3-tiles, and 32-tiles. The contributions of all forested subgrid areas weighted by their fractional area within the modeled region are considered. d) Plot of the difference in the normalized response metric between model runs using the alternative disturbance scenario ($\Delta\bar{X}_{norm,alternative}$) and the original disturbance scenario ($\Delta\bar{X}_{norm,original}$; Figure 2) for 1-tile/not-disturbed versus 1-tile/disturbed, 1-tile/disturbed versus 32-tile, for vegetation carbon (cVeg), soil carbon (cSoil), gross primary productivity (GPP), autotrophic respiration (Ra), heterotrophic respiration (Rh), leaf area index (LAI), sensible heat flux (HFSS), latent heat flux (HFLS), albedo (ALBS), fire emissions (fFire) and total deforested carbon (fDeforestTotal). All runs using >1 tile include disturbance. *denotes disturbance-related fluxes omitted in the 1-tile/not-disturbed model run.