Review of “Implementing a dynamic representation of fire and harvest including subgrid-scale heterogeneity in the tile-based land surface model CLASSIC v1.45”

by Curasi et al.
EGUsphere 2023-2003 for GMD

Curasi and colleagues present an implementation of sub-gridcell heterogeneity in the CLASSIC LSM, where the subgridcell tiles are dynamically created when the model simulates fire or harvest events. They test the model over Canada (offline with no coupling to an atmospheric model) with a custom set of PFTs, and find that the sub-gridcell heterogeneity makes as much, in fact often larger difference, to certain important C pool and energy flux variables than disturbance.

The presented advance is appropriate for publication in GMD. The topic is relevant for land surface modelling and therefore earth system modelling in general. The details of the implementation seem sensible, and the critical examination of the number of tiles with respect to balancing “accuracy” (the quotes because they don’t compare to data, just to the “full” model) and run time is very welcome. The effect on the fluxes and pools of the disturbances and the subgridcell heterogeneity are not negligible (so probably worthwhile enabling in many simulation setups), but not huge either (it will be interesting to see what affect they have in global, coupled simulations). I do have a few points of criticism which I believe should be addressed. Some of there are rather important to my mind, so I am going to recommend “Major Revisions” – even though the criticism may be considered comparatively mild for such a recommendation.

Key issues (in rough order of importance):

1. The authors made a commendable effort in setting up pre-satellite era disturbance scenarios for their model simulations to ensure a reasonable initial state before the evaluation period. But the time series for harvested area (Fig 2b) shows a huge discontinuity at the transition from bias corrected to observed. This is very probably not correct unless by some massive coincidence there was a change in legislation around that time or a bunch of logging companies simultaneously went bust. More likely the bias-correction procedure has gone wrong. Looking as the plot, the underlying data appear fine as the rate of increase is broadly similar across both periods. So if the earlier period was simply shifted downwards, the trajectory would be eminently reasonable! This should be investigated and fixed. The good news is that it will likely might make the simulated effects stronger because less logging beforehand will leave larger biomass pools at the start of the evaluation period, so the effects of disturbance and sub-gridcell heterogeneity will likely be stronger.

2. It is not made explicitly clear that when the number of tiles is greater than the number of disturbance events in a gridcell, the results for the gridcell will not improve (or change at all) with allowing more tiles (because they won’t be used). I assume the authors understand this, but the closest that the manuscript says to this is lines 426-433, although they don’t say this clearly. Instead the text talks about a “roughly exponential decline” and mentions “saturation”. This text is very wordy (it is hard to read) and misses the most important point: when you increase the number of dynamics tiles to 7, the disturbances in most gridcells are perfectly (or near perfectly) resolved and the results agree with the 32 tile run (for most variables). This should be clearly mentioned and its impact on the conclusions better discussed.

Relatedly, I need to flag the statement in the Discussion (line 528):

“Our results suggest that representing a relatively small number of heterogeneous tiles (e.g. < 12) may yield undesirable biases when compared to simulations using a larger number of tiles (32-tile; Figures 4a-h, 6a)”.

Based on Fig 4, the 7-tile simulation pretty much nails the 32-patch run and so doesn’t yield “undesirable biases”. The exception is the fire CO_2 emissions but these emissions are very insensitive to the tiling (Fig 6h). Also, Fig 6a) only shows comparisons of 1 vs 32 tiles, values like 7 and 12 are not shown, it is cannot support the statement.

Based on the two points above, I think the discussion of optimal number of tiles needs to be reconsidered.

3. The manuscript is rather long, but despite all the words, the wording is not always clear and can be difficult to follow. Here is an example starting at line 344:

“We utilized aspatial records of the total harvested and burned area within Canada to bias-correct inferred disturbance from 1920 - 1984 (Skakun et al., 2021; World Resources Institute, 2000). Before 1920, we utilized aspatial records of total disturbed area derived from 1920 stand age, with harvest held constant (0.3 Mha yr -1) (Chen et al., 2000; Kurz et al., 1995). First, for years in which inferred burned or harvested area (D inferred for i years, l grid cells; m 2) exceeded the aspatial records (d aspatial for i years; m 2) we correct the positive biases.”

Problem I encountered when reading this:

a. The order of the first two sentences should be swapped because they don’t match the conceptual temporal ordering (i.e. should be pre-1920 and *then* 1920-1984). In fact, both sentences could be combined into something much more succinct since they repeat a lot of words and deal very much with the same topic.
b It is not immediately clear if the bias correction was also applied to the pre-1920 data or not despite the data getting its own sentence. Likely not, but the sentence describing this data comes in between the first mention of the bias-correction and the description of it, so the implication is that it was in fact applied to pre-1920 data…? I don’t know what to make of it.
c. Getting into the description (third sentence), what is the “inferred” burned area or harvested area?
d. Why are positive and negative biases being corrected differently? No explanation is given.
e. In general, what follows this text is a (confusing) technical explanation of the bias correction procedure . At no point were the goals of the bias correction mentioned, potential pitfalls, the rationale for the choices made etc. And what is this “loop backward in time”? This is a completely new idea to me, why was this done?

As mentioned above, another example can be found in lines 426-432. Many words are used to basically say “When there are more tiles than disturbances, the results won’t change if you add more tiles.”

These are just two particular examples. The text is afflicted throughout with poor and non-logical flow, awkward wording and inappropriate levels of detail.

Because of the above, large parts of the manuscript should be critically reviewed and re-written.

Some general suggestions:

- Use more subheadings. This will force the writer to focus more precisely on what that text is supposed to say and will help the reader to understand exactly what they should be taking from the text.
- Describe the only methods briefly in the main text, but move the details to the appendix. This will improve the flow significantly.
- One specific point, please include more high-level details of the “normalized response metric”. How should the values be interpreted its their scale? And maybe give one sentence summarising the construction of the variable before jumping in the technical details.

4. The overall quality of the language is not too bad in terms of grammar and style, but it needs tightening up. Examples (non-exhaustive):

- Line 125 - “We use a domain, which encompasses all of Canada south of 76°N as our study area for demonstration.” Technically correct but very awkward .
- Line 191-195 – This single sentence sentence is huge (fully four lines with not a single comma) and completely cryptic. Reformulate.
- Line 236-238 – Not really a proper sentence.
- Line 417 – “Alternately” is not what the authors mean. I think “In contrast” or “Contrastingly”.
- Line 466 – Why are there citations behind an assertion about the paper’s own results?
- Line 524 – Use of the term “biases”. This is a matter of taste, but I dislike using “bias” when referring to model-to-model comparisons and prefer to reserve it for model-to-data comparisons. The authors may want to consider a different term.

5. There is very little comparison of the model improvements to data (only Fig 5b), but a large amount of model-to-model comparisons. The stated logic is that the 32-tile simulation with disturbance enabled is the best possible simulation, and therefore the standard to which the other simulations should be compared. But, is this 32-tile, disturbance-enabled simulation actually any better when compared to data than a single-tiled, non-disturbed simulation? Logically it might be better, but highly generalised but complex process-based models (such as LSMs) might not respond the way one expects. Some further model-to-data comparisons should be considered to make sure that changes are not actually making the model skill worse due to pre-existing cancellation of errors or some bias.

Relatedly, the explantation of the pre-existing “mosiac” tiling feature is not really clear (line 171). It is not specified how the tiles differ from one another – one must assume it varies across the cited studies. Giving more details about these studies and briefly summarising how mosaic tiling improved the model compared to the default “composite” approach would help justify the near exclusive use of model-to-model comparisons in the present study.

6. The authors made a point of describing the max of 66 disturbance events (due to the 33 years of the evaluation period) and relating that to the number of tiles. Simultaneously they spend a lot of effect deriving historical disturbance scenarios and also clearly state that disturbance is applied throughout. But they don’t make it clear how the disturbance is applied in the pre-evaluation period (since the can’t be contributing to the maximum 66 events). After carefully re-reading, it seems likely that the *composite* representation was used in the earlier period but I don’t think this is stated anywhere. If it was, will this transition from composite to mosaic change effect the results, particularly in the early phase of the evaluation period? Please make clear and discuss.

7. There is no mention of other disturbance agents such as biotic agents (importantly bark beetle), drought, wind throw and land slips. It is fine that they weren't included (not everything can be modelled at once) but some discussion is needed. Why where certain disturbances chosen and not others? Could this approach apply to other disturbance agents? What do the authors believe the consequences of leaving these out could be?