

Interactive comment on “CLASSIC v1.0: the open-source community successor to the Canadian Land Surface Scheme (CLASS) and the Canadian Terrestrial Ecosystem Model (CTEM) – Part 2: Global Benchmarking” by Christian Seiler et al.

Anonymous Referee #1

Received and published: 16 December 2020

This manuscript describes the results of a global benchmarking exercise in which the CLASSIC land surface model is compared with observation-based estimates of various quantities. Uncertainty in driving meteorological data and in the observations is assessed using an ensemble of runs and in the analysis procedure.

I recommend that this manuscript be accepted for publication with minor corrections (it is bit of a toss up between no corrections or minor corrections, but I will go for minor as

C1

there is some small room for improvement).

This is clearly a long manuscript! Despite that I have few comments to make. The nature of the material (model benchmarking) almost inevitably leads to text that is descriptive and repetitive as different aspects/variables/datasets are considered in turn. This level of detail is mainly appreciated by people using the model, or by other modelling groups who wish to understand how models compare, or by someone with a particular interest in one sub-area (say energy fluxes) – but in general it does not make for a paper that is particularly easy to read or to review. Thus I am quite happy to admit that I have not studied every line in every table, nor every figure. However, I am reassured by the quality of what I have seen and am fairly confident that the many details provided would also prove satisfactory (if I had the time and inclination to study them). This might sound like slightly veiled criticism but is intended as an endorsement of the manuscript!

I particularly value the treatment of uncertainty, with three meteorological datasets used to drive the model and more than one "observation-based" dataset used to assess most of the variables. This should really be standard practice but often lack of time encourages modellers to consider a reduced set of possibilities. The value of the approach taken here is shown by the headline result that for 10 of 19 variables the sign of the bias depends on the datasets used. Narrowing these uncertainties is a key challenge for the land surface community, but beyond the scope of this paper.

In general I think this is an excellent attempt to benchmark (or evaluate; some people distinguish these terms) a land surface model, with a commendable level of detail presented (without being completely overwhelming). I think some other modelling centres will look at this with some envy!

Specific points

Is it possible to make more of the comparison with the results of Part 1 (Melton et al.)? At present I think this is limited to the brief observation in the Discussion that the "overall

C2

statistic" is similar. Given that a holy grail for this kind of work is to get a coherent picture of model performance across multiple variables/sites/scales/datasets, it might be good to pursue the comparison to the results of Melton et al. in slightly greater depth (though I appreciate that it might be difficult to draw insightful conclusions beyond the fact that the results are similar).

Similarly, there is little or no mention of benchmarking studies using other models (e.g. CLM) and any common conclusions. Again, this might be difficult, but some brief mention would be warranted.

Are all simulations carried out at the same spatial resolution and/or at the resolution of each meteorological dataset? I suspect all the datasets are 0.5deg.

Figures 1 & 2 (and some others, such as Fig.7) - the colour scale could be better, in particular to distinguish the darker tones (e.g. >5 and <-40 in Fig.1a). This is particularly important for Figs. 1&2 because of the small size of the symbols - they pretty much all look the same dark colour to me. Can the symbols be made larger and/or the panels larger? Without that these plots are of very limited value.

Line 214: "it's" should be "its"

There is repeated but not ubiquitous use of italics when numbering lists (e.g. *i*, *ii*)- which looks a bit odd.

Fig.A1 - Eyeballing this it looks like CRUNCEP is often rather anomalous compared to the other two datasets (e.g. SW, Tas, precip, huss, wind). Is that because the other products are based on more similar data sources, and/or might this suggest that CRUNCEP is an anomalous (possibly inferior) dataset?

Interactive comment on Geosci. Model Dev. Discuss., <https://doi.org/10.5194/gmd-2020-294>, 2020.