Review of "Biophysics and vegetation cover change: a process-based evaluation framework for confronting land surface models with satellite observations", by Duveiller et al.

In this study, the authors analyze the skill of four global Land Surface models (LSMs) regarding their ability to represent land use and land cover change (LULCC) effects on the surface energy balance (SEB). Using an innovative simulation protocol and newly developed remote sensing data set, the authors uncover that the overall LSM skill is low, but that important variations in skill exist depending on the model, considered region, time of the year and SEB component.

This is an excellent manuscript. By performing a thorough quality assessment of LSM skill regarding biogeophysical impacts of LULCC, it prepares the way for improving LSMs. With the quality of future climate projections – especially regarding near-surface climate in populated regions – depending on the quality of their land surface component, this is indeed a much-needed work. Moreover, the performed analyses are comprehensive and go beyond the state-of-the-art. Finally, the paper is well structured and very well written, while reproducibility is ensured thanks to the analyses being well documented and the data set being publicly available. This paper thus demonstrates the potential to make a substantial contribution to the scientific literature. I only have some small concerns which ask for minor revisions of the manuscript. In general, the study may be considered for publication if the comments specified here below are sufficiently addressed.

General comments

1. The manuscript discusses several caveats related to the model set-up (P17L25). I would argue that three additional effects may affect model performance here:
   (i) The LSMs operate at much coarser resolution (factor 20 in either direction) than the native resolution of most of the RS products. As a consequence, the LSMs inevitably miss local differences in climate, e.g. induced by topography, and their potential effects on biogeophysical effects of LULCC. Interestingly (but likely by coincidence), the model skill is highest for the Rn product, which has a native resolution similar to the LSMs.
   (ii) Judging from the simulation protocol, the LSMs do not explicitly account for the effects of land management. Yet observational evidence indicates that the biogeophysical impacts of land management changes may be as important as the effects of land cover change (Luyssaert et al., 2014 NCC), which is now being confirmed by a growing number of observational as well as regional and global climate modelling studies. Omitting these effects likely negatively affects LSMs compared the RS product which does incorporate land management effects.
   (iii) Emergent scientific evidence highlights that the skill of the LSMs substantially depends on the quality of the meteorological forcing data set. For instance, CLM performance increases when switching from CRU-NCEP to GSWP3 forcing (

http://www.cesm.ucar.edu/events/wg-meetings/2018/presentations/lmwg/lawrence.pdf), and the skill of ISIMIP2a models regarding ET is predominantly determined by the atmospheric forcing data set (unpublished). While the current study moves beyond the state-of-the-art by considering multiple LSMs, uncertainty due to meteorological forcing is not considered.

Overall, I would suggest adding a short discussion of these additional limitations, thereby optionally also highlighting the potential of LUMIP and ISIMIP to address such issues.

2. In addition to background climate, season and region, the magnitude and even sign of the biogeophysical effect also depends on the time of the day. An increasing body of literature is now investigating these diurnal dynamics using models and observations (e.g. Lee et al., 2011 Nature; Vanden Broucke et al., 2015 JGR; Li et al., 2015 Nature Comm.). To what extent can the presented data set be used to study diurnal patterns?

Specific comments

1. P1L30: I would suggest specifying that this data set encompasses both the model and remote sensing-derived data, since the remote sensing data is already available on figshare.

2. P2L31: Also Lee et al. (2011 Nature)?

3. Figure 1: The grey box is almost invisible in the pdf version of the manuscript.

4. P4L13-17: Please mention the sign convention for the different SEB fluxes here (e.g. upward directed is positive). Note that I was expecting a different sign in figure 4 based on the text and the common sign convention.

5. P5L28: CLM4.5 can represent various crop types, so I presume the crop module – and thus irrigation – was switched off for these runs?

6. P9L5-7: This is the info I was searching for when looking at figure 2, perhaps it would be useful to mention this earlier in the manuscript (e.g. method section).

7. P18L4: Agree, but besides non-local effects, also local atmospheric feedbacks are not considered in this offline LSM set-up.

Textual comments

1. P3L11: Please introduce the acronym CMIP6. Also, some other acronyms appear to be not introduced?

2. P3L29: Please replace "consist" by "consists".

3. P5L29: Please split up this sentence to enhance readability: ". The *EvgTr* and *DecTr* simulations will therefore be identical in JULES".

4. P7L6 and elsewhere: I would recommend replacing "annually cumulated" to "annual accumulated" or simply "annual".

5. P7L18-20: Please check this sentence.

6. P7L23: Please add ", respectively". Also, mean and standard deviation should be both either singular or plural.

7. P9L21: Please specify to "shortwave albedo".

8. P16L9: Please replace "this" by "there".

9. P17L13: Please replace "and" by ", thereby".