RESPONSE TO REVIEWER 1

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.

This is a valid point that can be further underlined. We did so by further emphasizing it in the discussion paragraph regarding the PFT sub-grid heterogeneity.

However, to a certain point, the local aspect of topographical effect is mitigated by the masking out of areas with high topographical variability from the RS dataset (see Duveiller et al. 2018, SDATA). As a result, the values to which models are compared to at 1dd come from a sub-selection of 0.05dd values from quite homogeneously flat areas.

The delta Rn product actually has the same fine spatial resolution as the other RS products (that is 0.05dd), as the driving variable behind is the local spatial differences in albedo at 0.05 dd. The coarser data from the CERES instrument are used mostly to scale these differences from albedo to Rn, and to close the surface energy balance to retrieve H+G. Therefore, scale should not be a reason for higher model skill in Rn. Rather, this is because Rn is probably easier to parametrize in the models.

(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.

This is correct: models do not explicitly incorporate management in these runs. While some models arguably can include some management practices, it was decided not to turn them on to ensure comparability amongst models within this exercise. However, nothing prevents modelers to run their models with management under the presented framework and analyse themselves the improvement with respect to a baseline (e.g. the model runs currently in the dataset) and the RS product. We have added this in a final discussion paragraph that offers perspectives to extend the present work, and mentioned how this is a limitation in the current set-up.

(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.

We have added this as an extra limitation of the current set-up and included a suggestion in the final perspectives paragraph towards exploring the effect of using different forcing datasets on the resulting simulated biophysical impacts of LULCC.

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.

Done

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?

We agree that this is a promising avenue, but as it is, the current dataset is limited because it is based on MODIS and CERES instruments that are on-board of polar-orbiting satellite, which can only sample the earth at a fixed time. The RS dataset does have values for daytime and nighttime land surface temperature, but not for ET, H+G or Rn. We have added a suggestion at the end of the discussion that the use of geostationary satellites could be used in this sense.

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.

Done

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

Although this reference does apply as an example for space for time substitution, we have not included it here as the paragraph only deals with the context of satellite remote sensing, and Lee et al. is a flux-tower study.

3. Figure 1: The grey box is almost invisible in the pdf version of the manuscript.
We have changed the figure to make the grey box darker

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.

There seems to have been some confusion from our phrasing of the text. We have clarified it. We have also added a phrase specifying the sign convention for each flux in the place suggested by the reviewer.

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

Indeed, we use a generic C3 crop without irrigation for the CLM runs. We amended the text to clarify this.

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).

A phrase has been added in section 2.1 (Remote sensing estimations) to underline the fact that the RS dataset cannot provide values where vegetation types do not co-exist. A second phrase at the

end of section 2.3 (Harmonizing vegetation classes) was added to specify that model simulations are only retained where the RS data is available.

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

We assumed this was understood in the next phrase, but we made it more explicit by writing it down.

Textual comments

All textual comments have been addressed according to the suggestions of the reviewer.