Interactive comment on "Global Evaluation of Gross Primary Productivity in the JULES Land Surface Model" *by* Darren Slevin et al.

Anonymous Referee #2

Received and published: 26 October 2016

This paper provides a useful evaluation of JULES GPP using three different climate data sets at three resolutions. The simulated GPP is compared to three global benchmarking datasets. The strength of the manuscript is the comparison to multiple datasets with different ways of running JULES. Sensitivity of model results to driving data (which itself is uncertain) is not well studied, so it's important to understand how robust properties like GPP are to the inputs for the model. However, the paper is not very well organized and there are many places where I feel more detail could be given. I think it needs some substantial revisions. I have two major comments, and several suggestions to help with the organization and readability of the manuscript.

Major Comments

1. Results: I think the discussion of results in Section 3 needs some improvement,

with more detail on the processes behind the modelled and observed patterns in GPP. The focus of the paper is on comparing JULES to these datasets, but it would be more interesting to first explain what the datasets show. Lead each section with a brief explanation of the observed pattern in GPP, and explain differences between the datasets. Then the results of JULES can be given within the context of the observations and CARDAMOM.

- For example, in Figure 2b, JULES does very well if you are only comparing to MODIS. But it overestimates the variability of GPP during winter months compared to the other two datasets. So does this mean that JULES captures the interannual variability, or not?

- Another example: Page 11, Lines 29-33 (Discussion of figure 6): Why are the results for the extratropics the only ones discussed? I think much more could be said here – instead of just listing the differences it would be better to provide some more evaluation. For example, it was already stated that JULES overestimates GPP in the tropics, and this analysis shows that the overestimation occurs in all tropical land areas. That is a useful thing to note. On the other hand, JULES does reasonable in the extratropics – but it is consistently lower than all three datasets in Northern Asia.

2. Robustness of results: A potential strength of this manuscript is the comparison of JULES using different datasets, however I found the discussion of this topic a bit thin. Could the authors provide some more detailed discussion and context of the results? Here are some examples where further information could be provided:

- It's interesting that the results were insensitive to the spatial resolution (Page 12, Lines 9-12). This is an important conclusion of the analysis, and as the authors point out, using courser resolutions can save computational resources. But – is the result surprising given that the same soil ancillary data was used for all experiments? The JULES parameterizations are not scale-dependent (for example, this isn't the same as comparing scales in a model that uses cloud microphysical processes). I think using a

different soil ancillary data set would have a larger impact than changing the resolution.

- Also the meteorological dataset did not strongly change the results. However this is dependent on two things: 1) Maybe there were not large differences in climate between the data sets? IE: Page 15, Lines 2-6: Why are these differences in GPP occurring? Is the temperature and precipitation (or other variables) very different between the datasets in these regions? Are there other regions where the climate is very different, but the JULES simulations do not show dramatically different GPP? It would be good to provide some more information on the climates from the different driving data sets. 2) Since JULES was run with prescribed PFTs, there was no feedback between NPP and the land cover. It's possible that the GPP would be much more sensitive to the meteorology if competition between PFTs were allowed. Could the authors provide two additional experiments where the competition is allowed (e.g. one with either WFDEI product and one with the PRINCETON dataset)? Or at least provide the caveat that these results are possibly only valid when TRIFFID is not turned on. Although it's more work, I do think the additional simulations with TRIFFID would make this paper more relevant to a larger audience, as it seems most investigations using JULES have TRIFFID predicting PFTs (for example in TRENDY, the HELIX project, ISIMIP, and most CMIP5 and upcoming CMIP6 experiments).

Other Comments

1. There are several places where the text is repetitive:

- GPP is important because errors in its calculation can propagate through the model and affect biomass and other flux calculations: Page 2, Lines 27-28; Page 3, Line 5; Page 4 Lines 31-33.

- JULES is compared against FLUXNET-MTE, MODIS GPP, and CARDAMOM: Page 4, Lines 1-2; Page 5, Lines 6-7; Page 5, Line 11.

- Simulations are 2001-2010 because of availability of data: Page 4, Lines 33-34; Page

5, Lines 6-7

- The list of driving meteorological variables is given three times on pages 5-6. Even though there are differences between what is available from WATCH vs PRINCETON, this information could be given in a more concise manner.

- The FLUXNET-MTE is described as being derived from a machine learning technique/model tree ensemble twice in lines 14-20 of Page 6.

- Section 2.4

- There are more examples of this, please proofread the text and remove all repetition.

2. Page 2, Lines 6-7: It would be incorrect to say the reduced ability of land to absorb CO2 in the future has been observed. Perhaps better to say "... has been shown by models and inferred from observations ..."

3. Page 3, Lines 5-9: This paragraph needs some revision. The comparison of JULES to these precise datasets is not an important part of model development in general. Would be better to say that evaluating the simulated GPP at a range of scales and its sensitivity to spatial resolution and meteorological data is essential for informing future model developments. The specific datasets can be mentioned next, ie "In this manuscript, we do this using the FLUXNET-MTE etc."

4. Page 3, Line 25: I suggest removing "In LSMs"

5. Page 5, Lines 11-12: Please specify what information is provided by the soil dataset.

6. Page 5, Lines 17-19: I don't see why the requirement for data at 6 hourly intervals or less leads to the need for a number of datasets. However, there is value in evaluating model response to a number of datasets – for example JULES is currently run with different datasets for a number of projects and MIPs, and it is not known to what extent these different datasets affect the results.

7. Page 7, Lines 10-11: What is meant by modelling "quality"?

8. All evaluation of GPP is based on area-weighted GPP, correct? I think this could be said once in Section 2.5 and then it does not need to be repeated throughout the remainder of the text.

9. I would lead the results with the evaluation of the global GPP, then examine seasonal and interannual variation (ie switch sections 3.2 and 3.1). The seasonal cycle discussion does not belong in the section on interannual variability. This section should be renamed "Seasonal and interannual variability." Each section in the results ends with a one sentence summary – consider moving this sentence to the beginning of each section instead.

10. Page 10, Lines 8-9: I would move the last sentence of this paragraph to earlier in the paragraph since it explains how the reader should interpret the CV plot.

11. Page 10, Lines 13-15: This sentence is unclear.

12. Page 10, Lines 21-23: These numbers are different from what's given in Figure 3.

13. In Section 3.3, it's a bit unusual to give total over the 10 year period, instead of annual fluxes, which is what is more usually reported in global-scale evaluations of GPP.

14. Throughout the results, it would be much easier to read through if a range of the results are given instead of listing each GPP value every time. For example, Page 11, Line 15: Replace with "JULES overestimates total annual GPP by 20-41%"

15. Page 12, Lines 22, 24: I think it would be more appropriate to refer to the "pattern" of zonal means rather than the "trend" in zonal means, as trends typically refer to change in time, rather than change in space.

16. It's difficult to distinguish between the reds and pinks in Figures 2, 3, and 5; and between the shades of blue/green in Figures 4 and 6. Could a different set of colors be used?

C5