We thank the referee for providing a review of the manuscript and agree that the suggested changes and clarifications improve it. We have made the changes outlined below in the revised manuscript. Each item starts with the reviewer's comment followed by the changes to the manuscript. The text in blue is a re-written or new paragraph/sentence which has been added to the manuscript. The page and line numbers of where changes have been made to the updated manuscript are included at the end of each reply.

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.**
   As suggested, in the discussion section, We have started each subsection with a brief explanation of the pattern in observed and CARDAMOM GPP, followed by differences between the datasets. Finally, JULES GPP is given within the context of the observations and CARDAMOM. This has been done for subsections 4.1 and 4.2 (Pages 11–14). This has not been done for subsections 4.3 and 4.4 since the performance of JULES is being evaluated against itself. An extra paragraph was added to subsection 4.3 regarding the effect of spatial resolution on GPP simulations (Page 14, lines 22–26).

   Using a different soil ancillary dataset or land cover map (which specifies the PFT fractions) may have a larger impact than changing the spatial resolution. The regridding method used in this study was the conservative method, which preserves the same information when interpolating from 0.5° × 0.5° to 1° × 1° and 2° × 2° spatial resolutions, and results in only small differences in global GPP between the model simulations with varying spatial resolution. These small differences are due to differences in the PFT fractions of the land cover map after regridding.

2. **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?**
   In Figure 3b, JULES does very well if it is only compared to MODIS and overestimates the variability of GPP during winter months compared to the other two datasets (FLUXNET-MTE and CARDAMOM). We would say that JULES captures interannual variability since the coefficient of variation (CV) expressed as percentages of the mean monthly GPP for JULES lies between the CV values for the three observation-based estimates (Page 12, lines 3–5).

3. **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.
The results for the tropics has been added including some suggestions for improving simulated GPP. The following paragraphs were added to sections 3.3 and 4.2, respectively (Page 10, lines 15–21; Page 13, lines 17–19).

JULES overestimates GPP in all three tropical land areas compared to MODIS, FLUXNET-MTE and CARDAMOM (Figures 6c, e and f). Differences between JULES, MODIS, FLUXNET-MTE and CARDAMOM GPP with average annual GPP range from 7.4–12.1 Pg C year^{-1}, 7.7–13 Pg C year^{-1} and 1–1.3 Pg C year^{-1} for forests, grasslands and shrubs, respectively, in South and South-East Asia, 9.5–13.7 Pg C year^{-1}, 8.4–12.3 Pg C year^{-1} and 1.7–2.1 Pg C year^{-1} for forests, grasslands and shrubs, respectively, in Africa and 18-23.2 Pg C year^{-1}, 9–12.9 Pg C year^{-1} and 1.4–1.8 Pg C year^{-1} for forests, grasslands and shrubs, respectively, in Central and South America (Figures 6c, e and f, respectively).

JULES simulated average annual GPP to be 61, 54 and 7 Pg C year^{-1} for forests, grasslands and shrubs, respectively. JULES (JULES-WFDEI-GPCC) simulates higher GPP than MODIS, FLUXNET-MTE and CARDAMOM at global scales and this was found to be due to higher GPP simulated by JULES for forests and grasslands in the tropics (Figure 4b).

Yes we found that JULES performs reasonably well in the extratropics (Europe, Northern Asia, North America and Greenland and the Extratropical Southern Hemisphere), with the exception of Northern Asia and North America and Greenland, where the model is either equal to or lower than all three datasets. This may be due to the inability of this version of JULES to accurately simulate GPP in boreal regions where permafrost exists. It may also due to a different land cover map being used by JULES, MODIS and FLUXNET-MTE. The following paragraph were added to section 4.2 (Page 14, lines 4–8).

In the four extratropical regions (Europe, Northern Asia, Extratropical Southern Hemisphere and North America and Greenland), JULES simulated similar GPP to MODIS, FLUXNET-MTE and CARDAMOM for the three biomes in Europe and the Extratropical Southern Hemisphere (Figures 6a and d), with the exception of Northern Asia and North America and Greenland, where the model is either equal to or lower than all three datasets (Figures 6b and g). This is due to the inability of this version of JULES to accurately simulate GPP in boreal regions where permafrost exists.

4. **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.**
Yes, we found it interesting that the results were insensitive to spatial resolution. This is a useful since lower resolution global simulations can be performed to save computational resources. Using a different soil ancillary dataset or land cover map would have a larger

impact than changing the resolution. A paragraph discussing these points has been added to section 4.3 (Page 14, line 22–26).

Using a different soil ancillary dataset or land cover map (which specifies the PFT fractions) may have a larger impact than changing the spatial resolution. The regridding method used in this study was the conservative method, which preserves the same information when interpolating from 0.5° × 0.5° to 1° × 1° and 2° × 2° spatial resolutions, and results in only small differences in global GPP between the model simulations with varying spatial resolution. These small differences are due to differences in the PFT fractions of the land cover map after regridding.

5. **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).**
   1) When JULES was driven with different meteorological datasets, differences in simulated GPP occurred mostly in the tropics (between 5°N-5°S) with JULES driven with WFDEI-GPCC-1degree simulating higher GPP than JULES driven with PRINCETON and slightly higher GPP in the extratropics was simulated by JULES was driven with PRINCETON (Figure 5). There are differences in climate between the two datasets. Positive biases in the downward longwave radiation fluxes and surface air temperatures in the meteorological datasets are the reason for these differences (Figures G.5 and G.6 in Slevin (2016)). In general, precipitation in the WFDEI-GPCC dataset is higher than that of PRINCETON (Figures G.6b and d in Slevin (2016)) with surface air temperatures higher in PRINCETON (Figures G.6a and c in Slevin (2016)). However, since JULES is more sensitive to downward longwave radiation and surface air temperature than precipitation when simulating GPP, the main reason for differences in simulated GPP when JULES was driven with two different meteorological datasets is due to differences in downward longwave radiation fluxes and surface air temperatures. There are differences in northern Eurasia (above 60°N) in the meteorological datasets with slightly higher radiation fluxes (downward shortwave and longwave) and surface air temperatures in the PRINCETON dataset with little difference between the JULES simulations driven with WFDEI-GPCC and PRINCETON in this region (Figure 5). Information on differences in the meteorological dataset (WFDEI-GPCC and PRINCETON) led to differences in simulated GPP has been added to section 4.4 on Page 15, lines 1–16.

The higher simulated GPP in the tropics when JULES was driven with WFDEI-GPCC is due to positive biases in downward longwave radiation fluxes in WFDEI-GPCC in the Amazonian, African and South-East Asian tropics (Figures G.5b and d in Slevin (2016)) and the higher GPP simulated by JULES (driven with PRINCETON) in the extratropics are a result of positive biases in downward longwave radiation in the PRINCETON dataset in North America and Northern Asia (Figure G.5b in Slevin (2016)) and positive biases in surface air temperature in the PRINCETON dataset in the Northern Hemisphere (Figures G.6a and c in Slevin (2016)). As with the JULES-WFDEI-GPCC simulations, there are also differences in GPP between the PRINCETON driven JULES simulation and the observation-based and CARDAMOM estimates at latitudes 15°N-30°N (Figure 5). There was no improvement in simulated GPP when a different meteorological dataset was used.

In general, precipitation in the WFDEI-GPCC dataset is higher than that of PRINCETON (Figures G.6b and d in Slevin (2016)) with surface air temperatures higher in PRINCETON (Figures G.6a and c in Slevin (2016)). However, since JULES is more sensitive to downward longwave radiation and surface air temperature than precipitation when simulating GPP(Alton et al., 2007), the main reason for differences in simulated GPP when JULES was driven with two different meteorological datasets is due to differences in downward longwave radiation fluxes and surface air temperatures. There are differences in northern Eurasia (above 60°N) in the meteorological datasets with slightly higher radiation fluxes (downward shortwave and longwave) and surface air temperatures in the PRINCETON dataset with little difference between the JULES simulations driven with WFDEI-GPCC and PRINCETON in this region (Figure 5).

and on Page 15, line 33–Page 16, line 9.

When JULES was driven with the PRINCETON dataset, it was found that simulated photosynthesis was mostly Rubisco-limited (Figure 5.25 in Slevin (2016)). A similar trend was found when JULES was driven with the WFDEI-GPCC dataset (Figure 5.6 in Slevin (2016)). Similar trends in transport limitation were found with the JULES-PRINCETON model simulation, though the number of model gridboxes in which transport limitation dominated was less than that for the JULES-WFDEI-GPCC-1degree model simulation (Figures 5.25 and 5.28 in Slevin (2016)). When comparing the model gridbox fractions for the JULES-WFDEI-GPCC-1degree and JULES-PRINCETON model simulations, it was found that when JULES was driven with the PRINCETON dataset, simulated photosynthesis was more Rubisco-limited than when the model was driven with WFDEI-GPCC (Figure 5.26 in Slevin (2016)). Light-limitation was more important in simulating photosynthesis when JULES was driven with WFDEI-GPCC than PRINCETON (Figure 5.27 in Slevin (2016)). The percentage of model gridboxes which are transport-limited show a pronounced geographical variation with the WFDEI-GPCC driven simulation being more transport-limited in the Southern Hemisphere and the PRINCETON driven simulation being more transport-limited in the Northern Hemisphere (Figure 5.28 in Slevin (2016)).

2) Yes, since JULES was run with prescribed PFTs, there was no feedback between NPP and the land cover and there is a possibility that GPP could be more sensitive to the meteorology if competition between PFTs were allowed. These additional simulations with TRIFFID would make this paper more relevant to a larger audience. Therefore, two more model simulations were carried out where vegetation competition (and TRIFFID) were switched on. This was done with the WFDEI-GPCC and PRINCETON datasets (both at 1° × 1° spatial resolution). A new figure was added showing the results from these extra model simulations (Page 32, Figure 8). A paragraph describing the results

from these simulations was added to section 4.4 (Page 16, lines 10-24)

In this study, the model simulations were performed with prescribed PFTs (i.e. no vegetation competition). If competition between PFTs was allowed (i.e. vegetation competition), the annual average global GPP would be higher by 15 % and 17 %, for the WFDEI-GPCC and PRINCETON driven simulations, respectively (Figures 8b and e). In general, with vegetation competition switched on, higher GPP was simulated by JULES when driven with both datasets (Figures 8c and f). Higher GPP occurred mostly in Europe, south-eastern US, and in the tropical regions of Central and South America, Africa and South and South-East Asia (Figures 8c and f). This increased GPP in tropical regions is due to the tree-shrub-grass dominance heirachy in TRIFFID with dominant types (trees) limiting the expansion of subdominant types (shrubs and grasses). In savanna regions, such as the Sudanian Savanna, which stretches from the Atlantic Ocean in the west to the Ethiopian Highlands in the east of Africa, and northern Australia, there is higher GPP with prescribed PFTs (Figures 8c and f). These are also fire-prone regions. The version of JULES used in this study has no fire module and TRIFFID may overestimate woody cover and therefore GPP.

In terms of global GPP, the WFDEI-GPCC and PRINCETON driven simulations produce similar increases (Figures 8b and e). However, the spatial pattern is slightly different with higher GPP simulated in the Amazon region when JULES was driven with the WFDEI-GPCC dataset and higher GPP in southern Brazil and Argentina and Southeast Asia when JULES was driven with the PRINCETON dataset (Figures 8c and f). The spatial pattern of simulated GPP is more sensitive to the meteorological data than the annual average global GPP if competition between PFTs is allowed. This may be due to compensating differences in the sensitivity of the model to the two meteorological datasets.

and to the conclusions (Page 17, lines 26–29).

The model simulations in this study were largely performed with prescribed PFTs (i.e. no competition between PFTs was allowed). With competition between PFTs, the annual average global GPP was higher by 15 % and 17 %, for the WFDEI-GPCC and PRINCETON driven simulations, respectively, with the spatial pattern of simulated GPP more sensitive to the meteorological data used.

Other Comments

1. **There are several places where the text is repetitive:**
   The text has been updated to avoid repetition (see below).

   - **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.**
     Done (Page 2, lines 19–24).
   - **JULES is compared against FLUXNET-MTE, MODIS GPP, and CARDAMOM: Page 5, Lines 1–2; Page 5, Lines 6–7; Page 5, Line 11.**
     Done (Page 4, lines 27–28).
   - **Simulations are 2001–2010 because of availability of data: Page 4, Lines 33–34; Page 5, Lines 6–7.**
     Done (Page 4, lines 26–27).

- **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.**
  Done (Page 5, lines 26–28 and 32–33). The driving meteorological variables is also listed in the model description section since it is required when explaining the connection between the meteorological variables and GPP (Page 3, lines 25–26).
- **The FLUXNET-MTE is described as being derived from a machine learning technique/ model tree ensemble twice in lines 14–20 of Page 6.**
  Done (Page 6, lines 2–6).
- **Section 2.4 - There are more examples of this, please proofread the text and remove all repetition.**
  This section was re-written in order to avoid repetition (Page 7, lines 5–29).

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 ..."**
   This sentence has been changed to (Page 2, lines 3–5)

   The reduced ability of the land surface to absorb increased anthropogenic CO_2 emissions in the future has been shown by models and inferred from observations (Friedlingstein et al., 2006; Canadell et al., 2007; Friedlingstein et al., 2014; Sitch et al., 2015).

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."**
   This paragraph has been re-written (Page 2, line 33–Page 3, line 2).

   JULES has been evaluated at various scales: point (Blyth et al., 2010, 2011; Slevin et al., 2015; Ménard et al., 2015), regional (Galbraith et al., 2010; Burke et al., 2013; Chadburn et al., 2015) and globally as part of model-intercomparison studies (Anav et al., 2015; Sitch et al., 2015). Evaluating simulated GPP at a range of scales and its sensitivity to spatial resolution and meteorological data is essential for informing future model developments. In this manuscript, we do this using two observation-based datasets (FLUXNET-MTE and MODIS) and the CARbon DAta MOdel fraMework (Bloom et al., 2016, CARDAMOM).

4. **Page 3, Line 25: I suggest removing "In LSMs"**
   Done (Page 3, lines 18–19).

5. **Page 5, Lines 11–12: Please specify what information is provided by the soil dataset.**
   The following information is provided by the soil dataset (Page 5, lines 3–6).

   The soil dataset used was the Harmonized World Soil Database version 1.2 (Nachtergaele et al., 2012, HWSD) and contains soil property data such as soil texture fractions, water storage capacity, soil depth and pH (Nachtergaele et al., 2012). In this study, the soil texture fractions (% of sand, silt and clay) were used to calculate the soil thermal and hydraulic conductivity parameters listed in Table 3 of Best et al. (2011).

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.**
   Yes, I agree that evaluating JULES' response to various datasets (soil, vegetation and meteorological) can help to explain its behaviour when used as part of a multi-model inter-comparison project. This sentence has been changed to (Page 5, lines 8–10)

   Two meteorological datasets were used to drive the model offline (i.e. run separately from its host Earth System Model) at global scales; WFDEI (Weedon et al., 2014) and PRINCETON (Sheffield et al., 2006).

7. **Page 7, Lines 10-11: What is meant by modelling "quality"?**
   The sentence is missing the word "improve". It now reads (Page 6, lines 27–29)

   The CARbon DAta MOdel fraMework (CARDAMOM) is a model-data fusion approach which consists of merging observational data with models in order to improve model quality and characterise its uncertainty.

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.**
   Yes, all evaluation of GPP is based on area-weighted GPP. Since it is mentioned in Section 2.5, it has been removed from 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.**
   Sections 3.2 and 3.1 were switched (Pages 8–9). This also required that the abstract (Page 1, lines 5–8), the study questions (Page 3, lines 6–10), the list of experiments (Section 2.4; Page 7) and parts of Section 4.1 (Discussion; Page 12–14) and the Conclusions (Section 5; Page 16–18) be slightly re-written. Section 3.2 has been renamed to "Seasonal and interannual variability of GPP." (Page 9). The one sentence summary at the end of each section in the results section has been moved to the beginning.

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.**
    The last sentence of this paragraph has been moved to earlier in the paragraph as suggested (Page 9, lines 13–15).

11. **Page 10, Lines 13-15: This sentence is unclear.**
    This sentence has been rewritten (Page 9, lines 22–23).

    The model is able to capture simulated monthly anomalies from 2001 to 2010 with the exception of those in 2002 (Figure 3c).

12. **Page 10, Lines 21-23: These numbers are different from what's given in Figure 3.**
    The numbers have been changed to reflect those given in Figure 2 (Page 8, lines 24–27).

    This value is greater than that estimated by MODIS, FLUXNET-MTE and CARDAMOM with annual average global GPP estimated to be 112, 130 and 114 Pg C year^{-1}, respectively, for the same period (Figures 2a, b and d). The higher global GPP simulated by the JULES-WFDEI-GPCC driven simulations is greater than the MODIS, FLUXNET-MTE and CARDAMOM estimates by 25 %, 8 % and 23 % on average, respectively.

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.**
    Annual fluxes have been provided for GPP in Section 3.3 (Pages 9–10).

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%"**
    The results section has now been changed so that a range of results are given instead of listing each GPP value every time (Pages 8–11).

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.**
    Yes, you are correct. This has been changed (Page 11, lines 15–18).

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?**
    A different set of colors has been used to distinguish between the various model simulations in order to make it easier for the reader (Pages 26–30).

Bibliography

P. Alton, L. Mercado, and P. North. A sensitivity analysis of the land-surface scheme JULES conducted for three forest biomes: Biophysical parameters, model processes, and meteorological driving data. *Global Biogeochemical Cycles*, 20:GB1008, 2007. doi: 10.1029/2005GB002653.

A. Anav, P. Friedlingstein, C. Beer, P. Ciais, A. Harper, C. Jones, G. Murray-Tortarolo, D. Papale, N. C. Parazoo, P. Peylin, S. Piao, S. Sitch, N. Viovy, A. Wiltshire, and M. Zhao. Spatio-temporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, 2015. doi: 10.1002/2015RG000483.

M. J. Best, M. Pryor, D. B. Clark, G. G. Rooney, R .L. H. Essery, C. B. Ménard, J. M. Edwards, M. A. Hendry, A. Porson, N. Gedney, L. M. Mercado, S. Sitch, E. Blyth, O. Boucher, P. M. Cox, C. S. B. Grimmond, and R. J. Harding. The Joint UK Land Environment Simulator (JULES), Model description–Part 1: Energy and water fluxes. *Geoscientific Model Development*, 4:677–699, 2011. doi: 10.5194/gmd-4-677-2011.

A. A. Bloom, J.-F. Exbrayat, I. R. van der Velde, L. Feng, and M. Williams. The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times. *Proceedings of the National Academy of Sciences*, 113:1285–1290, 2016. doi: 10.1073/pnas.1515160113.

E. Blyth, J. Gash, A. Lloyd, M. Pryor, G. P. Weedon, and J. Shuttleworth. Evaluating the JULES Land Surface Model Energy Fluxes Using FLUXNET Data. *Journal of Hydrometeorology*, 11:509–519, 2010. doi: 10.1175/2009JHM1183.1.

E. Blyth, D. B. Clark, R. Ellis, C. Huntingford, S. Los, M. Pryor, M. Best, and S. Sitch. A comprehensive set of benchmark tests for a land surface model of simultaneous fluxes of water and carbon at both the global and seasonal scale. *Geoscientific Model Development*, 4:255–269, 2011. doi: 10.5194/gmd-4-255-2011.

E. J. Burke, R. Dankers, C. D. Jones, and A. J. Wiltshire. A retrospective analysis of pan Arctic permafrost using the JULES land surface model. *Climate Dynamics*, 41:1025–1038, 2013. doi: 10.1007/s00382-012-1648-x.

J. G. Canadell, C. Le Quéré, M. R. Raupach, C. B. Field, E. T. Buitenhuis, P. Ciais, T. J. Conway, N. P. Gillett, R. A. Houghton, and G. Marland. Contributions to accelerating atmospheric CO_2 growth from economic activity, carbon intensity, and efficiency of natural sinks. *Proceedings of the National Academy of Sciences of the United States of America*, 104:18866–18870, 2007. doi: 10.1073/pnas.0702737104.

S. Chadburn, E. Burke, R. Essery, J. Boike, M. Langer, M. Heikenfeld, P. Cox, and P. Friedlingstein. An improved representation of physical permafrost dynamics in the JULES land-surface model. *Geoscientific Model Development*, 8:1493–1508, 2015. doi: 10.5194/gmd-8-1493-2015.

P. Friedlingstein, P. Cox, R. Betts, L. Bopp, W. Von Bloh, V. Brovkin, P. Cadule, S. Doney, M. Eby, I. Fung, et al. Climate-Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison. *Journal of Climate*, 19:3337–3353, 2006. doi: 10.1175/JCLI3800.1.

P. Friedlingstein, M. Meinshausen, V. K. Arora, C. D. Jones, A. Anav, S. K. Liddicoat, and R. Knutti. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, 27:511–526, 2014. doi: 10.1175/JCLI-D-12-00579.1.

D. Galbraith, P. E. Levy, S. Sitch, C. Huntingford, P. Cox, M. Williams, and P. Meir. Multiple mechanisms of Amazonian forest biomass losses in three dynamic global vegetation models under climate change. *New Phytologist*, 187:647–665, 2010. doi: 10.1111/j.1469-8137.2010.03350.x.

C. B. Ménard, J. Ikonen, K. Rautiainen, M. Aurela, A. N. Arslan, and J. Pulliainen. Effects of Meteorological and Ancillary Data, Temporal Averaging, and Evaluation Methods on Model Performance and Uncertainty in a Land Surface Model. *Journal of Hydrometeorology*, 16: 2559–2576, 2015. doi: 10.1175/JHM-D-15-0013.1.

F. Nachtergaele, H. van Velthuizen, L. Verelst, D. Wiberg, N. Batjes, K. Dijkshoorn, V. van Engelen, G. Fischer, A. Jones, L. Montanarella, M. Petri, S. Prieler, E. Teixeira, and X. Shi. Harmonized World Soil Database v1.2. Technical report, International Institute for Applied Systems Analysis (IIASA), Food and Agriculture Organization of the United Nations (FAO), February 2012.

J. Sheffield, G. Goteti, and E. F. Wood. Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. *Journal of Climate*, 19:3088–3111, 2006. doi: 10.1175/JCLI3790.1.

S. Sitch, P. Friedlingstein, N. Gruber, S. D. Jones, G. Murray-Tortarolo, A. Ahlström, S. C. Doney, H. Graven, C. Heinze, C. Huntingford, S. Levis, P. E. Levy, M. Lomas, B. Poulter, N. Viovy, S. Zaehle, N. Zeng, A. Arneth, G. Bonan, L. Bopp, J. G. Canadell, F. Chevallier, P. Ciais, R. Ellis, M. Gloor, P. Peylin, S. L. Piao, C. Le Quéré, B. Smith, Z. Zhu, and R. Myneni. Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences*, 12:653–679, 2015. doi: 10.5194/bg-12-653-2015.

D. Slevin. *Investigating sources of uncertainty associated with the JULES land surface model.* PhD thesis, School of GeoSciences, University of Edinburgh, 2016. URL `http://hdl.handle.net/1842/18757`.

D. Slevin, S. F. B. Tett, and M. Williams. Multi-site evaluation of the JULES land surface model using global and local data. *Geoscientific Model Development*, 8:295–316, 2015. doi: 10.5194/gmd-8-295-2015.

G. P. Weedon, G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo. The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, 50:7505–7514, 2014. doi: 10.1002/2014WR015638.