Please find below our replies to both reviewers followed by a revised manuscript with tracked changes.

Reviewer #1

This paper describes statistical analysis of ISIMIP NPP dataset which weights models by their present day/historical performance in order to constrain the range of future estimates of NPP change. This is a well written and generally very clear disposition. I have a couple of small queries about the text, but no major issues.

Dear Reviewer

Many thanks for the review and the helpful comments on the approach on the manuscript. In the following, we provide an initial answer to your comments and will include additional text in the revised manuscript.

Comments:
Given the current popularity of the emergent constraint methodology, it would be useful to have a brief compare/contrast of how this method differs, as they seem superficially similar.

We agree with the reviewer that some aspects of the Reliability Ensemble Averaging (REA) are similar to multi-model averaging methods previously used in the context of terrestrial carbon cycle (e.g. Schwalm et al., 2015; Lovenduski and Bonan, 2017). Indeed, like in these recent studies, REA assigns more weight to simulations made by models that are more skilled to reproduce past observations. However, REA also considers how projections compare to each other by providing a measure of the convergence around the weighted average.

Beyond differences in the weighting schemes themselves, we also note discrepancies in the type and number of constraints, resolution and time period considered among studies. Lovenduski and Bonan (2017) consider a single value of cumulative terrestrial carbon uptake for 1959-2005 to derive one global coefficient per model. We apply the REA scheme on a pixel-by-pixel base using three different estimates of the same process while Schwalm et al. (2015) use multiple constraints on stocks and fluxes in each land pixel. However, we consider 21st century projections using a pixel-wise approach while Schwalm et al. (2015) focus on historical simulations.

We have added these aspects p2 l. 18-26:

> Until recently, applying these advanced multi-model averaging methods to simulations of the global carbon cycle has remained a challenge because of the lack of global observational datasets to constrain e.g. the REA weighting scheme. Schwalm et al. (2015) have presented results of skill-based model averaging applied to an ensemble of 10 models from the Multiscale synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger et al., 2013). This pixel-wise approach assigned weights to historical simulations based on their performance to simulate gross primary productivity and biomass stocks but did not consider future projections. Lovenduski and Bonan (2017) considered a single value of cumulative terrestrial carbon uptake for 1959-2005 to

derive one global coefficient per model to produce new projections. However, we are not aware of any studies using these methods in the context of spatially-explicit projections of the terrestrial carbon cycle under climate change.

The paper does an excellent job of explaining in appropriate detail the methods, but on page 6, line 1 three REAs are listed, but not explained what they are. It becomes clear in a figure caption later, but it would be good to explain in here too.

We refer to REA_C, REA_F and REA_M as the three REA cases driven by CARDAMOM, FLUXCOM and MODIS, respectively already on p. 5 l. 26-27. However, we take this comment as a necessity to remind the reader of the definition of each of the REA_C, REA_F and REA_M throughout the text and we do so p7 l. 7-8:

> The REA averaging method yields a global increase of NPP of 24.6 ± 8.5 Pg C y-1 (REA average ± RMSD) using CARDAMOM in REA_C, 24.8 ± 9.5 Pg C y-1 using FLUXCOM in REA_F and 25.0 ± 14.4 Pg C y-1 using MODIS NPP in REA_M.

I'd like to see a nod towards the uncertainties of the analysis in the abstract, particularly the lack of key processes (nitrogen, phosphorus, etc.) in the DGVMs. The discussion is good on this, but the abstract portrays a more uncritical acceptance of the reduction of uncertainty in the high latitudes, (especially boreal systems), which isn't completely supported by the data.

We agree that this is one of the major findings/limitations of our approach and needs to be highlighted in the abstract. We have added the following sentence to the abstract p1 l 21-22:

> This reduction in uncertainty is especially clear for boreal ecosystems although it may be an artefact due to the lack of representation of nutrient limitations on NPP in most models.

A brief discussion of the limits of this technique - especially regards whether we're increasing the precision but not the accuracy of the projections – would be useful. This is especially important given the issue about process representation, and the low weighting of the HYBRID model.

We agree that the low $R_{D,i}$ assigned to HYBRID at low and high latitudes may be due to its explicit representation of nitrogen limitations on NPP. These leads HYBRID to be the only model to project a possible decrease in global NPP by the end of the century and it becomes an outlier that is penalised by low values of $R_{D,i}$. However, HYBRID also performs less well than the other models as shown by low values of $R_{B,i}$ on Figure 4. Both these aspects play in the overall low R_i assigned to HYBRID as noted in the discussion p. 9 l 18-25:

> HYBRID is also the only model to predict a possible decrease in global NPP throughout the 21st century (Table 1 and Friend et al., 2014) because of a reduction at high latitudes and in tropical rainforests (Supplementary Figure S1). Thus, HYBRID is assigned low $R_{D,i}$ weights in these regions (Figure 4g-i and Supplementary Figures S4-12) and cannot influence the REA average and the calculation of its uncertainty (equation 4) despite integrating more detailed representation of ecosystem processes. However, HYBRID also exhibits stronger differences to the observational

datasets than the other models especially at high latitudes (Figure 4d-f) which may indicate a strong sensitivity of N limitations. Nevertheless, we note that all models' performances tend to decrease in regions north of 60°N where their ΔNPP projections also diverge (Figure 4d-f, Figure 5d-f)

Overall, the outcome of the REA approach cannot account for missing processes and remains conditional on the ensemble to which it is applied. This involves a risk to increase the precision around some inaccurate projections if treated like a black box. Following this comment and a similar comment from reviewer #2, we have added the following to the discussion p. 9 l 26-31:

> Overall, the promising REA results should be used carefully as they cannot correct for the omission of key processes by a large fraction of the ensemble members. Like in previous multi-model averaging studies focused on the carbon cycle (e.g. Schwalm et al., 2015; Lovenduski and Bonan, 2017) or climate (Krishnamurti et al., 1999; Giorgi and Mearns, 2002) we used already available simulations in a post-processing procedure. We note, however, that the ratio of two out of six models including carbon-nutrient interactions in the ISI-MIP ensemble is commensurate to other model inter-comparison projects: 3 out of 10 CMIP5 models (Exbrayat et al., 2014) or 2 out of 8 models in the new ISI-MIP experiments presented by Chen et al. (2017).

The map colour schemes are eye wateringly terrible, as well as not being colour blind friendly. The green in the middle makes it really difficult to read the plots accurately. The figure 4 plots would be enhanced by using different line patterns as well as colour, to help people read it in black and white print as well as colour blind readers. A cursory google or ask around the office should get the authors decent colour schemes. It's really not acceptable to use rainbow anymore.

We take this comment very seriously. Therefore, we have replaced Figures 2 to 5 using colour schemes that are compatible with colour-blindness (checked on http://www.vischeck.com) and have updated Figure 4 with line patterns. We have corrected the Supplementary Information accordingly.

There's a slightly higher than average number of words without spaces between them. This just needs checking.

We believe that this is an issue with the conversion of the original document into a pdf. We will double check upon submission of the revised manuscript.

Reviewer #2

The paper talks about a different approach of reliability ensemble averaging to calcu- late the average of multi-model estimates of global NPP for future scenario RCP 8.5. This new methodology takes into consideration 2 important aspects while allocating weights to different model estimates for calculating the ensemble mean: performance of the models as compared to the observations and convergence measure. Overall, introducing a new approach to calculate ensemble mean from different model estimates on a global scale is commendable and significant at this point in time when the world is focussing on quantifying the carbon fluxes for future and uncertainties in these estimates are large posing a challenge for scientists to come up with ways of reducing them. The analysis of the results obtained is extensive and comprehensive. However, there are some concerns that seem to be important.

Dear Reviewer,

Thank you for your insightful comments that will help improve the manuscript. We provide an initial answer to your comments in the following, and will include some additional text in a revised version of the manuscript.

Specific Comments:
In the discussion section, the major point that has been highlighted is the lack of representation of other elements, specifically N, in the GVMs used in this study and how their availability can limit carbon sequestration by vegetation in future. This has also been supported by multiple studies cited in the text. From the point of view of scientific knowledge and the focus on reduction in uncertainty from model estimates, the fact that of the 6 GVMs used in this study, only 2 (HYBRID and SDGVM) include the impact of N on model NPP estimation does not give a lot of reliability on results of this study. There should be some possible explanation for this difference in results of this study (increase in NPP) from other studies (reduction in NPP due to N limitation) to make the results more acceptable and reliable. In terms of introducing a new method for computing averages, the study has done a good job, but in terms of reliability and accuracy of the results of this study, it is questionable. This is a major concern.

Multi-model averaging is a post-processing procedure aiming at extracting knowledge from existing large ensemble of simulations. Like in previous multi-model averaging studies focused on the carbon cycle (e.g. Schwalm et al., 2015; Lovenduski and Bonan, 2017) or climate (Krishnamurti et al., 1999; Giorgi and Mearns, 2002) we used already available simulations in a "post-MIP" exercise. Overall, the outcome of the REA approach cannot account for missing processes and remains conditional on the ensemble to which it is applied. It is therefore beyond the scope of this paper to resolve the lack of process representation in some GVMs.

Nevertheless, we agree that the lack of representation of nutrient limitations on NPP in 4 out of 6 GVMs used here is a concern considering the possible implications for future productivity in response to increase CO2 concentrations (e.g. Wieder et al., 2015), a point we had already made in the discussion. We note, however, that this 1/3 ratio of models including carbon-nutrient interactions in the ISI-MIP ensemble is commensurate to other MIPs: 2 out of 10 CMIP5 models used by Exbrayat et al. (2014), 2 out of 8 models

in new ISI-MIP experiments presented by Chen et al. (2017). Furthermore, low weights Ri assigned to HYBRID (Figure 4a-c), which includes carbon-nutrient interactions, are not only due to a lack of convergence with the other models (Figure 4g-i) but also because of its poorer agreement with observational datasets (Figure 4d-f). SDGVM, the other model that includes carbon-nutrient interactions, is more similar to the carbon-only models in terms of historical performance and projected changes.

Overall, we accept this comment as a need to better explain the origin of the simulations and the post-processing nature of the averaging approach and we do so in the revised discussion p.9 l 26-31:

> Overall, the promising REA results should be used carefully as they cannot correct for the omission of key processes by a large fraction of the ensemble members. Like in previous multi-model averaging studies focused on the carbon cycle (e.g. Schwalm et al., 2015; Lovenduski and Bonan, 2017) or climate (Krishnamurti et al., 1999; Giorgi and Mearns, 2002) we used already available simulations in a post-processing procedure. We note, however, that the ratio of two out of six models including carbon-nutrient interactions in the ISI-MIP ensemble is commensurate to other model inter-comparison projects: 3 out of 10 CMIP5 models (Exbrayat et al., 2014) or 2 out of 8 models in the new ISI-MIP experiments presented by Chen et al. (2017).

There are different time periods that are included in the text. For instance, data from the 3 datasets used (CARDAMOM, FLUXCOM, MODIS) are from 2001-2010. While calculating Bi in equation (2), the difference between model predictions during last 10 years of historical simulations (1996-2005) and NPP from observations (2001-2010) is considered, or so it seems. It would be good to clarify why 2 different time periods are considered for calculating the performance measure (Bi) of models with observed values. Ideally, a comparison should be done for the same time period.

We agree that the benchmarking period should be the same. Therefore, we have redone the experiments using the time period 2001-2005 to evaluate Bi. As a result, we now compare the 2001-2005 reference period to the last five years of the projections for 2095-2099. Results are similar and numbers have been updated throughout the manuscript. For example, the first paragraph of the results section now reads p.7 l. 7-13:

> The REA averaging method yields a global increase of NPP of 24.6 ± 8.5 Pg C y-1 (REA average ± RMSD) using CARDAMOM in REAC, 24.8 ± 9.5 Pg C y-1 using FLUXCOM in REAF and 25.0 ± 14.4 Pg C y-1 using MODIS NPP in REAM. As the ISI-MIP ensemble mean indicated a ΔNPP of 24.2 Pg C y-1, these results represent a ~2% increase of the mean for both REAC and REAF and 3% for REAM. The pixel-wise one standard deviation uncertainty in the ISI-MIP ensemble was 26.3 Pg C y-1 and the REA results indicate strong reduction of 68% for REAC, 64% for REAF and 45% for REAM. These results further indicate that in all three cases the REA averaging method reduces the uncertainty of the ensemble spread toward an agreement on a future increase in the global land carbon uptake.

Captions of figures should be improved to include details like time period for which the given figure represents mean. For instance, in the caption of figure 1, what years comprise the historical simulation can be added. Captions should be as complete in themselves as possible.

We have improved the figure captions to include more detailed descriptions. For example, the caption of figure 1 now reads (p. 19):

> Figure 1: Zonal mean ΔNPP by the end of the 21st century (averaged over 2095-2099) under RCP8.5 compared to the end of the historical simulations (averaged over 2001-2005). Shading represents the uncertainty around the zonal mean across the ISI-MIP ensemble, taken as one standard deviation for ISI-MIP, and calculated following equation (4) for REA. REAC, REAF and REAM, refer to REA values calculated based on observationally-constrained CARDAMOM, FLUXCOM and MODIS NPP respectively.

Title of section 2.2 on page 3 'Estimates of current NPP' is confusing since the ISI-MIP model simulations also include the current period.

We have replaced with "Benchmark datasets of modern NPP" (p. 3 l. 27)

In the manuscript, appropriate spaces have been missed between 2 words or a word and a full stop. Like in page 5 line 17, the word 'integratealso'. The authors are advised to go through the text and revise these typographical mistakes.

We note that this comment is similar to reviewer #1's and will make sure that these typos will disappear in the revised manuscript.

In section 2.3 on Reliability Ensemble Averaging, before the actual method has been described there is a lot of description of the other methods used for calculating mean. This part from line 10 to 16 on page 5 can be a part of the introduction, where it identifies why these other methods are not serving the purpose and there is a need for a better strategy. Since REA is the method finally adopted in this study, the description of only this method used should be a part of this section 2.3.

We agree that this section of the text is misplaced, and actually redundant with the text page 2 l. 7 to 18. Therefore, we will remove it from the method section that now starts with the following (p5 l. 14-17):

> The Reliability Ensemble Averaging method (REA; Giorgi and Mearns, 2002) was developed to assign coefficients to models in the context of future projections. Additionally to using a measure of model performance to reproduce historical conditions, the REA weighting scheme implements measure of model convergence to penalize models that do not predict the same response to changes (Exbrayat et al., 2013b).

Since REA is a new approach introduced for calculating NPP in this study, it would be good if the terms in equation (1) and (5) are described in terms of their maximum and minimum possible values, and their significance to give a more meaningful perspective of this approach.

Terms R_i, $R_{B,i}$ and $R_{D,i}$ are model weights and range from 0, for a poorly performing model, to 1. We have improved the explanation of the possible range taken by these three terms p.5 l 23-26:

> The performance coefficient $R_{B,i}$ ranges from 0, for a poorly performing model, to 1 if the absolute value of B_i is smaller than the variability ε. Similarly, the convergence coefficient $R_{D,i}$

ranges from 0 for outlier projections to 1 if the absolute value of D_i, the difference between the projection and the REA mean, is smaller than ε. As a result, the final model weight R_i also takes values ranging from 0 to 1.

References

Chen, M., Rafique, R., Asrar, G. R., Bond-Lamberty, B., Ciais, P., Zhao, F., Reyer, C. P. O., Ostberg, S., Chang, J., Ito, A., Yang, J., Zeng, N., Kalnay, E., West, T., Leng, G., Francois, L., Munhoven, G., Henrot, A., Tian, H., Pan, S., Nishina, K., Viovy, N., Morfopoulos, C., Betts, R., Schaphoff, S., Steinkamp, J. and Hickler, T.: Regional contribution to variability and trends of global gross primary productivity, Environ. Res. Lett., 12(10), doi:10.1088/1748-9326/aa8978, 2017.

Exbrayat, J.-F., Pitman, A. J., and Abramowitz, G.: Response of microbial decomposition to spin-up explains CMIP5 soil carbon range until 2100, Geosci. Model Dev., 7, 2683-2692, https://doi.org/10.5194/gmd-7-2683-2014, 2014.

Giorgi, F. and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, J. Clim., 15, 1141–1158, doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2, 2002.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S. and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science (80-.)., 285(5433), 1548–1550, doi:10.1126/science.285.5433.1548, 1999.

Lovenduski, N. S. and Bonan, G. B.: Reducing uncertainty in projections of terrestrial carbon uptake, Env. Res. Lett., 12(4), 044020, 2017.

Schwalm, C. R., Huntzinger, D. N., Fisher, J. B., Michalak, A. M., Bowman, K., Ciais, P., Cook, R., El-Masri, B., Hayes, D., Huang, M., Ito, A., Jain, A., King, A. W., Lei, H., Liu, J., Lu, C., Mao, J., Peng, S., Poulter, B., Ricciuto, D., Schaefer, K., Shi, X., Tao, B., Tian, H., Wang, W., Wei, Y., Yang, J. and Zeng, N.: Toward "optimal" integration of terrestrial biosphere models, Geophys. Res. Lett., 42(11), 4418–4428, doi:10.1002/2015GL064002, 2015.

Wieder, W. R., Cleveland, C. C., Smith, W. K. and Todd-Brown, K.: Future productivity and carbon storage limited by terrestrial nutrient availability, Nat. Geosci., 8(6), 441–444, doi:10.1038/NGEO2413, 2015.

Reliability Ensemble Averaging of 21st century projections of terrestrial net primary productivity reduces global and regional uncertainties

Jean-François Exbrayat[1], A. Anthony Bloom[2], Pete Falloon[3], Akihiko Ito[4], T. Luke Smallman[1], Mathew Williams[1]

[1]School of GeoSciences and National Centre for Earth Observation, University of Edinburgh, Edinburgh, EH9 3FF, UK
[2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, US
[3]Met Office Hadley Centre, Fitzroy Road, Exeter, EX1 3PB, UK
[4]National Institute for Environmental Studies, Tsukuba, Japan

Correspondence to: Jean-François Exbrayat (j.exbrayat@ed.ac.uk)

Abstract. Multi-model averaging techniques provide opportunities to extract additional information from large ensembles of simulations. In particular, present-day model skill can be used to evaluate their potential performance in future climate simulations. Multi-model averaging methods have been used extensively in climate and hydrological sciences, but they have not been used to constrain projected plant productivity responses to climate change, which is a major uncertainty in earth system modelling. Here, we use three global observation-orientated estimates of current net primary productivity (NPP) to perform a reliability ensemble averaging (REA) using 30 global simulations of the 21st century change in NPP based on the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP) 'business as usual' emissions scenario. We find that the three REAs support an increase in global NPP by the end of the 21st century (2095-2099) compared to 2001-2005, which is 2 – 3% stronger than the ensemble ISIMIP mean value of 2.4 Pg C y^{-1}. Using REA also leads to a 45 – 68% reduction in the global uncertainty of 21st century NPP projection, which strengthens confidence in the resilience of the CO_2-fertilization effect to climate change. This reduction in uncertainty is especially clear for boreal ecosystems although it may be an artefact due to the lack of representation of nutrient limitations on NPP in most models. Conversely, the large uncertainty that remains on the sign of the response of NPP in semi-arid regions points to the need for better observations and model development in these regions.

Deleted: 2090s
Deleted: the
Deleted: 2000s
Deleted: 4
Deleted: 6
Deleted: 3
Deleted: 7
Deleted: 43
Deleted: 67
Deleted: .

1 Introduction

Anthropogenic emissions of carbon dioxide (CO_2) enhance the uptake of atmospheric carbon by terrestrial ecosystems through net primary productivity (NPP). This so-called CO_2-fertilization effect has helped offset 25-30% of CO_2 emissions responsible for climate change in recent decades (Canadell et al., 2007; Le Quéré et al., 2009). There exists a large uncertainty as to whether this positive effect of CO_2-fertilization will be resilient to climate change, as shown by the spread between model projections from various intercomparison projects (Friedlingstein et al., 2006; Arora et al., 2013; Friend et al., 2014; Nishina et al., 2014, 2015), especially in highly productive tropical regions (Rammig et al., 2010; Cox et al., 2013). However, large ensembles of

projections are challenging to interpret as they may include models with an opposite response to the same change in boundary conditions (Friedlingstein et al., 2006). Simulations from the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP, Warszawski et al., 2014) have shown that most of the uncertainty in 21st century projections of the terrestrial carbon cycle can be attributed to differences between global vegetation models (GVMs; Friend et al., 2014; Nishina et al., 2014, 2015), although a non-negligible part of the uncertainty arises from differences in climate projections themselves (Ahlström et al., 2012).

In recent years multi-model averaging has been widely used to extract information from large ensembles of simulations in studies targeting climate change (Bishop and Abramowitz, 2012; Krishnamurti et al., 1999), rainfall-runoff processes (Georgakakos et al., 2004; Huisman et al., 2009; Shamseldin et al., 1997; Viney et al., 2009) and catchment-scale nutrient exports (Exbrayat et al., 2010, 2013b). These methods range from simple arithmetic means of model ensembles to more elaborate weighting schemes that take model performance into account such as Bayesian Model Averaging (Raftery et al., 2005). The underlying assumption is that a model that is better able to reproduce current conditions should be given more weight in the final projection than a poorly performing model. The more complex Reliability Ensemble Averaging (REA; Giorgi and Mearns, 2002) approach takes into account a measure of convergence between projections to identify the most likely change: this way, the REA method avoids giving too much weight to an over-fitted model which may accurately represent current conditions for the wrong reasons but predicts vastly different change than other ensemble members (Exbrayat et al., 2013b). Metrics measuring model independence (Bishop and Abramowitz, 2012) have also been introduced in weighting schemes to avoid pseudo-replication.

Until recently, applying these advanced multi-model averaging methods to simulations of the global carbon cycle has remained a challenge because of the lack of global observational datasets to constrain e.g. the REA weighting scheme. Schwalm et al. (2015) have presented results of skill-based model averaging applied to an ensemble of 10 models from the Multiscale synthesis and Terrestrial Model Intercomparison Project (MsTMIP; Huntzinger et al., 2013). This pixel-wise approach assigned weights to historical simulations based on their performance to simulate gross primary productivity and biomass stocks but did not consider future projections. Lovenduski and Bonan (2017) considered a single value of cumulative terrestrial carbon uptake for 1959-2005 to derive one global coefficient per model to produce new projections. However, we are not aware of any studies using these methods in the context of spatially-explicit projections of the terrestrial carbon cycle under climate change.

Here, we present the first example of a pixel-wise application of the REA approach to extract a best estimate of NPP change (ΔNPP) during the 21st century under a business-as-usual scenario of emissions from a large ensemble of projections. We perform the REA procedure three times using different observation-constrained estimates of current NPP: retrievals of the terrestrial carbon cycle with the CARDAMOM model-data fusion approach (Bloom and Williams, 2015; Bloom et al., 2016), an approximation of NPP based on the up-scaled FLUXCOM GPP datasets (Jung et al., 2009, 2011, 2017; Tramontana et al., 2016), and the MOD17A3 MODIS NPP product (Running et al., 2004; Zhao et al., 2005; Zhao and Running, 2010). Based on optimally-weighted model averages, we evaluate the impact of the REA method on 21st century projections of ΔNPP but also on the uncertainty in the future resilience of the CO_2-fertilization that exist among the models. We show that the REA procedure

Deleted: terrestrial

Deleted: To our knowledge only

Deleted: to historical simulations of the terrestrial carbon cycle

Deleted: .

Deleted: ¶

Deleted: spatially-explicit

Deleted:

can help identify regions where uncertainties remain large and thereby inform the future development of models and observational networks needed to improve climate change projections.

2 Materials & methods

2.1 The ISI-MIP ensemble

We used an ensemble of simulations of net primary production (NPP) from the Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP; Warszawski et al., 2014). The ISI-MIP simulations included here consist of 6 global vegetation models: HYBRID (Friend and White, 2000), JeDi (Pavlick et al., 2013), JULES (Clark et al., 2011), LPJmL (Sitch et al., 2003), SDGVM (Woodward et al., 1995) and VISIT (Ito and Inatomi, 2012). Each of these 6 GVMs was driven by bias-corrected output (Hempel et al., 2013) from 5 general circulation models (GCMs): GFDL-ESM2M (Dunne et al., 2012), HadGEM2-ES (Collins et al., 2011), IPSL-CM5A-LR (Dufresne et al., 2013), MIROC-ESM-CHEM (Watanabe et al., 2011) and NorESM1-M (Bentsen et al., 2013), generating a total of 30 global simulations of NPP for the historical period and under the representative concentration pathway 8.5 (RCP8.5). We chose the ISI-MIP ensemble over other initiatives like C4MIP (Friedlingstein et al., 2006) or CMIP5 (Taylor et al., 2012) because the combination of multiple GVMs with multiple GCMs in ISI-MIP allows a more comprehensive coverage of the uncertainty in the terrestrial carbon cycle and attribution of dominant factor in the uncertainty of the future (Friend et al., 2014; Nishina et al., 2014, 2015) although we note that these simulations omit feedbacks from the biosphere on weather and atmospheric CO_2 concentrations. As the ensemble integrates 5 representations of the same GVM, and 6 representations of the same GCM, we avoid issues related to model genealogy (Knutti et al., 2013) that could lead similar models to bias results of the averaging because of intrinsic lack of independence between the different ensemble members (Bishop and Abramowitz, 2012). We focus our approach on NPP projections under the RCP8.5 scenario of emissions for which more simulations were available (Nishina et al., 2015). Mean annual current NPP and projected changes are summarised in Table 1 and Supplementary Figure S1. We note a large spread in current global NPP simulated by the models from 51.7 Pg C y^{-1} to 77.8 Pg C y^{-1} during 2001-2005, the last five years of the historical simulations, as well as ΔNPP in the last five years of the projections (2095-2099), ranging from -3.7 to 41.6 Pg C y^{-1}. Further information on the models and the ISI-MIP protocol are to be found in the Supplementary Information of Friend et al. (2014) and the respective model description papers listed in Table 1.

2.2 Benchmark datasets of modern NPP

We use three different estimates of current NPP: (a) an observation-bound terrestrial carbon cycle analysis estimate, (b) an estimate based on up-scaled eddy-covariance CO_2 flux measurements, and (c) an estimate based on satellite measurements of absorbed photosynthetically active radiation. To harmonize the approach, we re-gridded all observationally-constrained NPP datasets to the lowest dataset resolution (1°×1°), and confined our analysis to the overlap period 2001-2005, for which

Deleted: 6
Deleted: 5
Deleted: 10
Deleted: 2090s
Deleted: 17
Deleted: 0
Deleted: 4
Deleted: Estimates of current
Deleted: NPP dataset
Deleted: (
Deleted: -2010
Deleted:)

benchmark datasets and ISI-MIP models were available. Mean annual NPP and variability for each dataset is presented in Figures S2 and S3.

2.2.1 CARDAMOM retrievals

The CARbon DAta Model fraMework (Bloom et al., 2016) produces spatially explicit retrievals of the global terrestrial carbon cycle following a model-data fusion procedure. In each 1°×1° pixel, the Data-Assimilation Linked Ecosystem Carbon version 2 (DALEC2; Bloom and Williams, 2015; Williams et al., 2005) is driven by ERA-Interim reanalysis climate data (Dee et al., 2011) and burned area from the Global Fire Emission Database version 4 (Giglio et al., 2013). A Bayesian Markov Chain Monte-Carlo approach is implemented to constrain DALEC2 according to observations of MODIS leaf area index (Myneni et al., 2002), tropical biomass (Saatchi et al., 2011), soil carbon content from the Harmonized World Soil Database (HWSD; FAO, 2012) and a set of Ecological and Dynamic Constraints (Bloom and Williams, 2015). Through this Bayesian procedure, CARDAMOM provides an explicit estimation of the uncertainty in model parameters and hence in land-atmosphere carbon fluxes such as net primary production (NPP) from site to global-scale (Bloom et al., 2016; Smallman et al., 2017). However, as not all the other datasets (see sections 2.2.2 and 2.2.3) provide a measure of the parametric uncertainty, in this study we rely on CARDAMOM's highest confidence estimates of a mean annual NPP of 49.9 Pg C y^{-1}. More details on the framework can be found in the supplementary information of Bloom et al. (2016).

Deleted: 50

Deleted: .1

2.2.2 FLUXCOM

The FLUXCOM project uses machine-learning methods (Tramontana et al., 2016) to up-scale global datasets from eddy-covariance measurements of CO_2 and energy fluxes from the FLUXNET network (Baldocchi et al., 2001). In a first step, a machine-learning algorithm is used to extract a relationship between local environmental drivers and ecosystem fluxes (Jung et al., 2009). Then, the trained algorithm is used in combination with gridded climate data and remote sensing observations to produce a global estimate of monthly ecosystem fluxes at a 0.5°×0.5° spatial resolution. In its first instance, FLUXCOM products relied on a random forest method (Breiman, 2001) but newly available datasets have been produced using additional machine learning methods (Tramontana et al., 2016; Jung et al., 2017).

Here, we use the average of an ensemble of six FLUXCOM GPP datasets to derive an estimate of annual NPP for 2001-2005. These datasets were created using three machine-learning methods: random forest, artificial neural networks and multivariate regression splines. Each machine-learning method was used to produce two GPP datasets corresponding to two partitioning methods of net ecosystem exchange (see Reichstein et al. (2005) and Lasslop et al. (2009)). Then, we used CARDAMOM's retrievals of carbon use efficiency (Bloom et al., 2016), the ratio of NPP to GPP, to derive a current value of NPP of 52.7 Pg C y^{-1} for the first five years of the 21st century from the 126.9 Pg C y^{-1} FLUXCOM estimated GPP.

Deleted: 2010

Deleted: 8

Deleted: ten

Deleted: 127

Deleted: 1

2.2.3 MODIS NPP

The MOD17 MODIS GPP/NPP dataset provides 8-day estimates of GPP and annual NPP at a 1-km spatial resolution since the year 2000. Therefore, GPP is calculated as the product of the amount of absorbed photosynthetically active radiation (estimated from the MOD15 MODIS LAI/FPAR product, Myneni et al., 2002) and a biome-specific radiation use efficiency that is adjusted as a function of air temperature and vapour pressure deficit. Land cover classification is derived from MODIS using the MCD12Q1 product (Friedl et al., 2002) while meteorological data are taken from the National Centers for Environmental Prediction (NCEP)/ Department of Energy (DOE) Reanalyses II. Then, annual maintenance respiration is estimated using a temperature-acclimated Q_{10} relationship (Tjoelker et al., 2001) while growth respiration is assumed to be a fixed fraction of NPP. The MODIS NPP dataset has been used to quantify the impact of droughts (Zhao and Running et al., 2010) and the El Niño/Southern Oscillation on global terrestrial ecosystem productivity (Bastos et al., 2013). We re-gridded the annual NPP data to a 1°×1° spatial resolution for the reference years 2001-2005 from which we derived a 53.4 Pg C y^{-1} mean annual value.

2.3 Reliability Ensemble Averaging

The Reliability Ensemble Averaging method (REA; Giorgi and Mearns, 2002) was developed to assign coefficients to models in the context of future projections. Additionally to using a measure of model performance to reproduce historical conditions, the REA weighting scheme implements measure of model convergence to penalize models that do not predict the same response to changes (Exbrayat et al., 2013b).

In each 1°×1° pixel, each model projection i of the 30 GVM-GCM ensemble is assigned a reliability factor R_i that is calculated such as

$$R_i = R_{B,i} \times R_{D,i} = \left(\frac{\varepsilon}{|B_i|}\right) \times \left(\frac{\varepsilon}{|D_i|}\right) \quad (1)$$

where ε represents the variability in observations expressed as the difference between the largest and smallest values of annual NPP in each pixel (Figure S3; Giorgi and Mearns, 2002), while B_i and D_i correspond to a measure of model i's performance and convergence, respectively. The performance coefficient $R_{B,i}$ ranges from 0, for a poorly performing model, to 1 if the absolute value of B_i is smaller than the variability ε. Similarly, the convergence coefficient $R_{D,i}$ ranges from 0 for outlier projections to 1 if the absolute value of D_i, the difference between the projection and the REA mean, is smaller than ε. As a result, the final model weight R_i also takes values ranging from 0 to 1. We produce three REA estimates based on CARDAMOM, FLUXCOM and MODIS NPP, further referred to as REA_C, REA_F and REA_M, respectively. For each REA application, terms ε, B_i, D_i and hence R_i (equation 1) are recalculated based on the particular observational dataset to produce three independent sets of model coefficients.

Here, we apply the REA method to the ensemble of 30 ISI-MIP simulations of 21st century ΔNPP under RCP8.5 emission scenario. We first re-gridded the ISI-MIP data using the *remapcon* function of the Climate Data Operators version 1.6.9 to

Deleted: 2010

Deleted: 6

Deleted: Multi-model averaging techniques have been developed to extract information and quantify the uncertainty from large ensembles of simulations (e.g. Krishnamurti et al., 1999). These methods range from simple arithmetic mean to more complex statistical methods (Viney et al., 2009) such as Bayesian Model Averaging (Raftery et al., 2005). A common assumption is that models which better reproduce available observations should be given more weight in a final prediction than poorly performing models. However, models may be over-fitted to match observations, providing the good answers for the wrong reasons (Exbrayat et al., 2013b). These models are likely to represent improperly, or even omit, processes which may become key under changed conditions, and this challenges their reliability. Therefore, the

Deleted: integrate

Deleted: also

Deleted: in the weighting scheme and

Deleted: which

Formatted: Font: Italic

Formatted: Font: Italic, Subscript

Formatted: Subscript

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic, Subscript

Formatted: Font: Italic

Formatted: Font: Italic, Subscript

Formatted: Font: Italic

Formatted: Subscript

match the 1°×1° spatial resolution of the observationally constrained datasets (see section 2.2) and performed the procedure in each land pixel to create maps of REA averages. We then apply the REA method three times (REA_C, REA_F and REA_M) to evaluate their current performance.

For each 30 simulations of the ISI-MIP ensemble we calculated B_i in each pixel such as

$$B_i = NPP_i - NPP_{obs} \quad (2)$$

where NPP_i is the mean annual NPP predicted by model i during the 10 last years of the historical simulations and NPP_{obs} corresponds to either of the observational datasets mean annual NPP. Then for each model the value of D_i was calculated in each pixel as the difference between the change predicted by model i and the REA average such as

$$D_i = \Delta NPP_i - \frac{\sum_{i=1}^{N} R_i \cdot \Delta NPP_i}{\sum_{i=1}^{N} R_i} \quad (3)$$

where ΔNPP_i is the change in mean NPP in the last five years of the RCP8.5 simulation (2095-2099) compared to the last five years of the historical simulations (2001-2005) predicted by the ensemble member i and N is the total number of ensemble members. The REA average is not known beforehand and weights $R_{D,i}$ are calculated iteratively (Giorgi and Mearns, 2002).

The uncertainty around the REA average change is calculated as the weighted root-mean square difference (RMSD) calculated following

$$RMSD = \left(\frac{\sum_{i=1}^{N} R_i \cdot \left(\Delta NPP_i - \Delta NPP_{REA} \right)^2}{\sum_{i=1}^{N} R_i} \right)^{1/2} \quad (4)$$

where ΔNPP_{REA} is the REA average change. Assuming that the error distribution is somewhere between uniform and Gaussian, the 60-70% confidence interval of the REA is represented by $\Delta NPP_{REA} \pm RMSD$ (Giorgi and Mearns, 2002).

Giorgi and Mearns (2002) further introduced a quantitative measure of the collective model reliability ρ, based on R_i, where

$$\rho = \frac{\sum_{i=1}^{N} R_i^2}{\sum_{i=1}^{N} R_i} \quad (5)$$

which will vary pixel-wise based on each model's performance with respect to the mean and variability represented in each observational dataset as well as the convergence to the REA average. The reliability measure ρ can be further decomposed in ρ_B and ρ_D, such as

Deleted: 10

Deleted: 2090

Deleted: 10

Deleted: 1996

Deleted: Finally, weights $R_{B,i}$ and $R_{D,i}$ are assigned a maximum value of 1 if the absolute value of B_i and D_i are smaller than ε, the measure of variability in the observations.

$$\rho_B = \frac{\sum_{i=1}^{N} R_{B,i}}{N} \quad (6)$$

$$\rho_D = \frac{\sum_{i=1}^{N} R_{D,i}}{N} \quad (7)$$

where ρ_B and ρ_D correspond to the ensemble reliability with respect to model biases and model convergence respectively. All ρ, ρ_B and ρ_D take values ranging from 0, indicating a lack of agreement between models, to 1, indicating a consensus between models in terms of performance and projected changes.

3 Results

The REA averaging method yields a global increase of NPP of 24.6 ± 8.5 Pg C y^{-1} (REA average ± RMSD) using CARDAMOM in REA_C, 24.8 ± 9.5 Pg C y^{-1} using FLUXCOM in REA_F and 25.0 ± 14.4 Pg C y^{-1} using MODIS NPP in REA_M. As the ISI-MIP ensemble mean indicated a ΔNPP of 24.2 Pg C y^{-1}, these results represent a ~2% increase of the mean for both REA_C and REA_F and 3% for REA_M. The pixel-wise one standard deviation uncertainty in the ISI-MIP ensemble was 26.3 Pg C y^{-1} and the REA results indicate strong reduction of 68% for REA_C, 64% for REA_F and 45% for REA_M. These results further indicate that in all three cases the REA averaging method reduces the uncertainty of the ensemble spread toward an agreement on a future increase in the global land carbon uptake.

Zonal means (Figure 1) indicate that the ISI-MIP ensemble mean and all three REA_C, REA_F and REA_M averages estimate an increase in NPP across all latitudes. All three REA averages predict a weaker increase in NPP at high latitudes of the northern and southern hemispheres. They also agree on a stronger increase in NPP than the ISI-MIP ensemble mean for tropical regions between 15°S and 10°N but also between 20°N and 25°N and temperate regions around 55°N to 65°N. REA_C and REA_F indicate a weaker increase in NPP than ISI-MIP around 20°S while the REA_M average is similar to the ISI-MIP ensemble mean in these regions. The uncertainty around each of the REA averages is smaller than the uncertainty around the ISI-MIP ensemble mean across all latitudinal zones. Furthermore, while the very large uncertainty around the ISI-MIP ensemble mean does not provide confidence on the sign of ΔNPP across most regions, the uncertainty around all three REA averages is constrained toward an increase in NPP across all regions, except around 20°S.

The spatial distribution of the ISI-MIP ensemble mean ΔNPP contrasts with that of the three REA averages with noticeable differences across all regions of the globe (Figure 2). All three REA averages predict a weaker increase in NPP than the ISI-MIP ensemble in Canada and Scandinavia, while they predict a stronger increase in NPP in Eurasia. Similarly, all three REA averages predict a stronger increase in NPP than the ISI-MIP ensemble in tropical rainforest of South America, Africa and south-east Asia. The REA averages agree on a weaker ΔNPP in semi-arid regions of the Sahel, southern Africa, Australia and the Tibetan Plateau. Overall, all REA_C, REA_F and REA_M exhibit broadly similar patterns in the spatial distribution of ΔNPP

Formatted: Subscript
Formatted: Subscript
Deleted: 7
Deleted: 9
Deleted: 8
Deleted: for
Deleted: 5
Deleted: 0
Deleted: 10
Deleted: 0
Deleted: for
Deleted: 3
Formatted: Subscript
Formatted: Subscript
Deleted: 9
Deleted: 6
Deleted: 2
Deleted: for
Formatted: Subscript
Deleted: 3
Deleted: 7
Deleted: 4
Deleted: ,
Deleted: 5% for
Deleted: 6
Deleted: 29
Deleted: 6
Deleted: 7
Deleted: 6
Deleted: 3
Deleted: 4
Formatted: Subscript
Deleted: s
Deleted: REA_F and
Deleted: s
Deleted: are

differences with the ISI-MIP ensemble mean that is confirmed by Pearson's correlation coefficient of 0.63 between REA_C and REA_F, 0.61 between REA_C and REA_M and 0.68 between REA_F and REA_M.

The uncertainty in ΔNPP is reduced across most regions of the globe for all three REA_C, REA_F and REA_M (Figure 1 and Figure 3). This reduction of uncertainty leads to a confidence on the sign estimation of ΔNPP in 86%, 80% and 76% of all the land pixels for REA_C, REA_F and REA_M respectively, against 43% for the ISI-MIP ensemble. The average reduction in uncertainty is large in regions north of 40°N (Figure 1), mostly corresponding to a reduction in uncertainty in boreal Eurasia (Figure 3) that provides better confidence in an increase in NPP (Figure 2). We note that the uncertainty in the REA_M remains similar to the uncertainty around the ISI-MIP ensemble mean for large portions of the southern hemisphere such as southern Africa. However, all three REA_C, REA_F and REA_M cannot provide confidence on the sign of ΔNPP for southern Africa and Australia. The zonal means of the mean values of the three coefficients R_i, $R_{B,i}$ and $R_{D,i}$ (Figure 4) show that MODIS-based REA_M yields larger values of all coefficients compared to REA_C and REA_F. We note strong inter-model similarities in the spatial distribution of model weights (R_i; Figure 4a-c), biases ($R_{B,i}$; Figure 4d-f) and convergence of the projected ΔNPP ($R_{D,i}$; Figure 4g-i). Only the HYBRID models are almost systematically assigned lower weight R_i as a result of lower values for both $R_{B,i}$ (i.e. a larger bias than the other models) and $R_{D,i}$ (i.e. a divergence in projected ΔNPP). This is especially obvious in boreal regions north of 60°N where HYBRID is assigned values significantly closer to 0 in all REA_C, REA_F and REA_M.

The collective model reliability measure ρ provides a quantification of the spread of model weights determined through the REA method (Figure 5). Regions where ρ is close to 1 indicates places where there is a strong consensus between models on the current NPP but also on 21st century ΔNPP. There are large differences in ρ depending on the NPP observational datasets using to constrain the REA (Figure 5). Indeed, while the average value of ρ is 0.29 for REA_C and 0.32 for REA_F, it is 0.62 for REA_M. REA_C and REA_F yields very low values of ρ in boreal regions (Figure 5) while REA_M leads to values of ρ close to 1 in many regions south of 60°S. The measure of reliability ρ can be further decomposed in two components ρ_B and ρ_D (Figure 5d-i, equations 6 and 7). Results indicate that ρ_D is consistently greater than ρ_B for all REA_C, REA_F and REA_M. This result means that model convergence in the simulation of ΔNPP is greater than the model performance to reproduce current NPP. In other words, the model performance evaluated against the three current NPP datasets contributes the most to decreasing the ensemble reliability ρ. Values of ρ_B are lower than 0.10 in boreal regions for REA_C and REA_F, indicating that model bias is greater than the variability of NPP ε estimated from the CARDAMOM retrievals and the FLUXCOM based NPP by a factor 10. Conversely, regions where ρ_B is close to 1 for REA_M indicate that the variability in the MODIS NPP observations is larger than model biases.

4 Discussion

The globally integrated values of the REA average change (24.6 to 25.0 Pg C y-1) and the ISI-MIP ensemble mean (24.2 Pg C y^{-1}) are similar. This is in agreement with a previous multi-model approach that only found a 0.01 Pg C y-1 difference in historical mean annual net ecosystem exchange between a simple mean and a weighted average based on model performance

Deleted: R^2
Deleted: 74
Deleted: 66
Deleted: 70
Deleted: 4
Deleted: 3
Deleted: 35
Deleted: and
Formatted: Subscript
Deleted: .
Deleted: 35
Deleted: 38
Deleted: 75
Deleted: most
Deleted: 23
Deleted: 9
Deleted: 3
Deleted: 7

(Schwalm et al., 2015). However, by contrast with this previous study, we find that in all three REA_C, REA_F and REA_M a large spatial variability in grid cell differences (Figure 2) that compensate each other to yield a relatively small global difference with the ISI-MIP ensemble mean. The three REA averages indicate a stronger positive ΔNPP than the ISI-MIP ensemble mean for boreal Eurasia and tropical rainforests (Figures 1 and 2), and a weaker but still positive ΔNPP in northern Canada and semi-arid regions like the Sahel, the Tibetan plateau, southern Africa and Australia.

The reduction in uncertainty arising from the REA method helps putting a greater confidence in a sustained CO_2-fertilization effect throughout the 21st century although these results may be influenced by model-wise differences in process representation. In both the ISI-MIP ensemble mean and the three REA averages, the sustained increase of NPP at high latitudes, where nitrogen (N) limitation on NPP dominates (Zhang et al., 2011; Exbrayat et al., 2013a) but is only represented in the HYBRID and SDGVM models (Table 1; Nishina et al., 2014). The increase in NPP in these N-limited regions is in contrast with observations at Free-Air CO2 Enrichment experiments that indicate a quick weakening of the CO2-fertilization effect as soil N stores deplete (Norby et al., 2010). Models which integrate coupled C-N cycles generally predict the historical land carbon sink in good agreement with estimates from the Global Carbon Budget (Huntzinger et al., 2017) and project a decrease in NPP throughout the 21st century (Thornton et al., 2009; Goll et al., 2012; Zhang et al., 2013; Wieder et al., 2015).

Similarly, recent observations have concluded a total absence of CO_2-fertilization effect under phosphorus-limited conditions (Ellsworth et al., 2017) which dominates in the tropics and leads to an additional reduction of NPP in model projections (Goll et al., 2012; Zhang et al., 2013; Wieder et al., 2015). Here, only the HYBRID and SDGVM models integrate the representation of N limitations on NPP (Nishina et al., 2014) and none of them represent phosphorous limitations. HYBRID is also the only model to predict a possible decrease in global NPP throughout the 21st century (Table 1 and Friend et al., 2014) because of a reduction at high latitudes and in tropical rainforests (Supplementary Figure S1). Thus, HYBRID is assigned low $R_{D,i}$ weights in these regions (Figure 4g-i and Supplementary Figures S4-12) and cannot influence the REA average and the calculation of its uncertainty (equation 4) despite integrating more detailed representation of ecosystem processes. However, HYBRID also exhibits stronger differences to the observational datasets than the other models especially at high latitudes (Figure 4d-f) which may indicate a strong sensitivity of N limitations. Nevertheless, we note that all models' performances tend to decrease in regions north of 60°N where their ΔNPP projections also diverge (Figure 4d-f, Figure 5d-f).

Overall, the promising REA results should be used carefully as they cannot correct for the omission of key processes by a large fraction of the ensemble members. Like in previous multi-model averaging studies focused on the carbon cycle (e.g. Schwalm et al., 2015; Lovenduski and Bonan, 2017) or climate (Krishnamurti et al., 1999; Giorgi and Mearns, 2002) we used already available simulations in a post-processing procedure. We note, however, that the ratio of two out of six models including carbon-nutrient interactions in the ISI-MIP ensemble is commensurate to other model inter-comparison projects: 3 out of 10 CMIP5 models (Exbrayat et al., 2014) or 2 out of 8 models in the new ISI-MIP experiments presented by Chen et al. (2017).

There is also considerable debate on how good large-scale NPP observational products are (Kolby-Smith et al., 2015; de Kauwe et al., 2016), a problem that we address by performing the REA approach three times.

Formatted: Subscript

Formatted: Subscript

Formatted: Subscript

Deleted: g-i

Deleted: 5g

Deleted: i

Deleted: Overall, we note that the promising REA results should be used carefully as they cannot correct for the omissions of key processes by a large fraction of the ensemble members.

Deleted:

In all three REA_C, REA_F and REA_M cases, the global uncertainty around the REA average is reduced compared to the uncertainty within the ISI-MIP ensemble which provides a higher degree of confidence in the resilience of the global CO2-fertilization effect to warming. The reduction in uncertainty, and the gain in confidence on the sign of ΔNPP, is especially obvious in boreal regions for all three REA (Figure 3). Conversely, uncertainties on the sign of ΔNPP remain large for all REA in semi-arid regions of Southern Africa and Australia. It is a non-trivial result as the response of these ecosystems to climate events like El Niño and La Niña drives the inter-annual variability and the trend of the global terrestrial carbon sink (Bastos et al., 2013; Poulter et al., 2014; Ahlström et al., 2015), while projections indicate a gain of forest ecosystems over savannahs in the future (Moncrieff et al., 2016).

Because of the way the REA method assigns coefficients to ensemble members with respect to the annual variability in the data ε_A (equation 1), the final REA average and uncertainty are conditional on the variability represented in current estimate of NPP. Figure 5a-c shows that the reliability of the ensemble measured by ρ varies depending on which observational dataset is used, although generally lower values of ρ_B and ρ_D at high latitudes indicate that models disagree on the current NPP and future ΔNPP in these regions. Furthermore, high values of ρ for REAM indicate a larger variability ε in the MODIS dataset compare to CARDAMOM and the FLUXCOM based NPP data (Figure S3). This larger variability leads to more models being given a weight close to 1 in the averaging scheme because the variability is larger than their bias (Figure 5f) or the predicted change (Figure 5i). Conversely, the relatively smaller variability in CARDAMOM retrievals leads more models to be weighted poorly according to both their performance (Figure 5d) and their convergence with other models (Figure 5g). The variability ε influences the final uncertainty and as a result the REA_C has a smaller uncertainty because it is more penalizing on models, and vice-versa with MODIS NPP.

5 Conclusion

We applied the REA method on a pixel-by-pixel base to an ensemble of 30 simulations of historical and 21st century NPP from the ISI-MIP project. Our results indicate that using either CARDAMOM retrievals, a FLUXCOM based estimate of current NPP or data from MODIS to constrain the REA scheme helps at least halving the uncertainty in 21st century global ΔNPP. This process leads to a higher confidence in a sustained CO_2-fertilization effect. We nevertheless note that a large uncertainty remains in semi-arid regions that is mostly attributable to differences in process representation in global vegetation models. Furthermore, most models used here do not account for N limitations on NPP and this may have altered the outcome of the convergence coefficient used in REA.

Acknowledgements

This work was supported by the Natural Environment Research Council through the National Centre for Earth Observation. Part of this work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with

Deleted:
Formatted: Font: Italic
Formatted: Subscript
Formatted: Subscript
Formatted: Superscript
Formatted: Subscript

the National Aeronautics and Space Administration. PF was supported by the Joint UK DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). For their roles in producing, coordinating, and making available the ISI-MIP model output, we acknowledge the modelling groups and the ISI-MIP coordination team.

References

Ahlström, A., Raupach, M. R., Schurgers, G., Smith, B., Arneth, A., Jung, M., Reichstein, M., Canadell, J. G., Friedlingstein, P., Jain, A. K., Kato, E., Poulter, B., Sitch, S., Stocker, B. D., Viovy, N., Wang, Y. P., Wiltshire, A., Zaehle, S. and Zeng, N.: Carbon cycle. The dominant role of semi-arid ecosystems in the trend and variability of the land CO_2 sink, Science, 348(6237), 895–9, doi:10.1126/science.aaa1668, 2015.

Ahlström, A., Schurgers, G., Arneth, A., and Smith, B.: Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections, Env. Res. Lett., 7, 044008, doi:10.1088/1748-9326/7/4/044008, 2012.

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., Bonan, G., Bopp, L., Brovkin, V., Cadule, P., Hajima, T., Ilyina, T., Lindsay, K., Tjiputra, J. F. and Wu, T.: Carbon–Concentration and Carbon–Climate Feedbacks in CMIP5 Earth System Models, J. Clim., 26(15), 5289–5314, doi:10.1175/JCLI-D-12-00494.1, 2013.

Baldocchi, D., Falge, E., Gu, L. H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X. H., Malhi, Y., Meyers, T., Munger, W., Oechel, W., U, K. T. P., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K. and Wofsy, S.: FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities, Bull. Am. Meteorol. Soc., 82(11), 2415–2434, doi:10.1175/1520-0477(2001)082<2415:fantts>2.3.co;2, 2001.

Bastos, A., Running, S. W., Gouveia, C. and Trigo, R. M.: The global NPP dependence on ENSO: La Niña and the extraordinary year of 2011, J. Geophys. Res. Biogeosciences, 118(3), 1247-1255, doi:10.1002/jgrg.20100, 2013.

Beer, C., Reichstein, M., Tomelleri, E., Ciais, P., Jung, M., Carvalhais, N., Rödenbeck, C., Arain, A. M., Baldocchi, D., Bonan, B. G., Bondeau, A., Cescatti, A., Lasslop, G., Lindroth, A., Lomas, M., Luyssaert, S., Margolis, H., Oleson, W. K., Roupsard, O., Veenendaal, E., Viovy, N., Woodward, I. F., and Papale, D.: Terrestrial Gross Carbon Dioxide Uptake: Global Distribution and Covariation with Climate, Science, 329, 834–838, doi:10.1126/science.1184984, 2010.

Bentsen, M., Bethke, I., Debernard, J. B., Iversen, T., Kirkevåg, A., Seland, Ø., Drange, H., Roelandt, C., Seierstad, I. A., Hoose, C., and Kristjánsson, J. E.: The Norwegian Earth System Model, NorESM1-M – Part 1: Description and basic evaluation of the physical climate, Geosci. Model Dev., 6, 687-720, doi:10.5194/gmd-6-687-2013, 2013.

Bishop, C. H. and Abramowitz, G.: Climate model dependence and the replicate Earth paradigm, Clim. Dyn., 41(3–4), 885–900, doi:10.1007/s00382-012-1610-y, 2012.

Bloom, A. A. and Williams, M.: Constraining ecosystem carbon dynamics in a data-limited world: integrating ecological "common sense" in a model–data fusion framework, Biogeosciences, 12(5), 1299–1315, doi:10.5194/bg-12-1299-2015, 2015.

Bloom, A. A., Exbrayat, J.-F., van der Velde, I. R., Feng, L. and Williams, M.: The decadal state of the terrestrial carbon cycle: Global retrievals of terrestrial carbon allocation, pools, and residence times., Proc. Natl. Acad. Sci. U. S. A., 113(5), 1285–1290, doi:10.1073/pnas.1515160113, 2016.

Breiman, L.: Random forests, Mach. Learn., 45(1), 5–32, 2001.

Canadell, J. G., Le Quéré, C., Raupach, M. R., Field, C. B., Buitenhuis, E. T., Ciais, P., Conway, T. J., Gillett, N. P., Houghton, R. A., and Marland, G.: Contributions to accelerating atmospheric CO2 growth from economic activity, carbon intensity, and efficiency of natural sinks, Proc. Natl. Acad. Sci., 104, 18866–18870, doi:10.1073/pnas.0702737104, 2007.

Chen, M., Rafique, R., Asrar, G. R., Bond-Lamberty, B., Ciais, P., Zhao, F., Reyer, C. P. O., Ostberg, S., Chang, J., Ito, A., Yang, J., Zeng, N., Kalnay, E., West, T., Leng, G., Francois, L., Munhoven, G., Henrot, A., Tian, H., Pan, S., Nishina, K., Viovy, N., Morfopoulos, C., Betts, R., Schaphoff, S., Steinkamp, J. and Hickler, T.: Regional contribution to variability and trends of global gross primary productivity, Environ. Res. Lett., 12(10), doi:10.1088/1748-9326/aa8978, 2017.

Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., Pryor, M., Rooney, G. G., Essery, R. L. H., Blyth, E., Boucher, O., Harding, R. J., Huntingford, C. and Cox, P. M.: The Joint UK Land Environment Simulator (JULES), model description – Part 2: Carbon fluxes and vegetation dynamics, Geosci. Model Dev., 4(3), 701–722, doi:10.5194/gmd-4-701-2011, 2011.

Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C. D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A., and Woodward, S.: Development and evaluation of an Earth-System model – HadGEM2, Geosci. Model Dev., 4, 1051-1075, doi:10.5194/gmd-4-1051-2011, 2011.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Q. J. R. Meteorol. Soc., 137(656), 553–597, doi:10.1002/qj.828, 2011.

De Kauwe, M. G., Keenan, T. F., Medlyn, B. E., Prentice, I. C. and Terrer., C.: Satellite based estimates underestimate the effect of CO2 fertilization on net primary productivity, Nature Clim. Change, 6, 892–893, doi:10.1038/nclimate3105, 2016.

Dufresne, J.-L., et al.: Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5, Clim. Dyn., ., 40, 2123–2165, doi:10.1007/s00382-012-1636-1, 2013.

Dunne, J. P., et al.: GFDL's ESM2 Global Coupled Climate–Carbon Earth System Models. Part I: Physical Formulation and Baseline Simulation Characteristics, J. Clim., 25, 6646–6665, doi:10.1175/JCLI-D-11-00560.1, 2012.

Ellsworth, D. S., Anderson, I. C., Crous, K. Y., Cooke, J., Drake, J. E., Gherlenda, A. N., Gimeno, T. E., Macdonald C. A., Medlyn, B. E., Powell, J. R., Tjoelker, M. G. and Reich, P. B.: Elevated CO2 does not increase eucalypt forest productivity on a low-phosphorus soil, Nature Clim. Change, 7, 279-282, doi: 10.1038/nclimate3235, 2017.

Exbrayat, J.-F., Viney, N. R., Seibert, J., Wrede, S., Frede, H.-G. and Breuer, L.: Ensemble modelling of nitrogen fluxes: data fusion for a Swedish meso-scale catchment, Hydrol. Earth Syst. Sci., 14(12), 2383–2397, doi:10.5194/hess-14-2383-2010, 2010.

Exbrayat, J.-F., Pitman, A. J., Zhang, Q., Abramowitz, G. and Wang, Y.-P.: Examining soil carbon uncertainty in a global model: response of microbial decomposition to temperature, moisture and nutrient limitation, Biogeosciences, 10(11), 7095–7108, doi:10.5194/bg-10-7095-2013, 2013a.

Exbrayat, J.-F., Viney, N. R., Frede, H.-G. and Breuer, L.: Using multi-model averaging to improve the reliability of catchment scale nitrogen predictions, Geosci. Model Dev., 6(1), 117–125, doi:10.5194/gmd-6-117-2013, 2013b.

Exbrayat, J.-F., Pitman, A. J., and Abramowitz, G.: Response of microbial decomposition to spin-up explains CMIP5 soil carbon range until 2100, Geosci. Model Dev., 7, 2683-2692, https://doi.org/10.5194/gmd-7-2683-2014, 2014.

FAO/IIASA/ISRIC/ISSCAS/JRC: Harmonized World Soil Database (version 1.21), FAO, Rome, Italy and IIASA, Laxenburg, Austria, 2012.

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., Woodcock, C. E., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., and Schaaf, C.: Global land cover mapping from MODIS: algorithms and early results, Remote Sens. Environ., 83(1-2), 287-302, doi:10.1016/S0034-4257(02)00078-0 , 2002.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., Cadule, P., Doney, S., Eby, M., Fung, I., Bala, G., John, J., Jones, C., Joos, F., Kato, T., Kawamiya, M., Knorr, W., Lindsay, K., Matthews, H. D., Raddatz, T., Rayner, P., Reick, C., Roeckner, E., Schnitzler, K.-G., Schnur, R., Strassmann, K., Weaver, A. J., Yoshikawa, C., and Zeng, N.: Climate–Carbon Cycle Feedback Analysis: Results from the C4MIP Model Intercomparison, J. Clim., 19, 3337–3353, doi:10.1175/JCLI3800.1, 2006.

Friend, A. D. and White, A.: Evaluation and analysis of a dynamic terrestrial ecosystem model under preindustrial conditions at the global scale, Global Biogeochem. Cycles, 14(4), 1173–1190, doi:10.1029/1999GB900085, 2000.

Friend, A. D., Lucht, W., Rademacher, T. T., Keribin, R., Betts, R., Cadule, P., Ciais, P., Clark, D. B., Dankers, R., Falloon, P. D., Ito, A., Kahana, R., Kleidon, A., Lomas, M. R., Nishina, K., Ostberg, S., Pavlick, R., Peylin, P., Schaphoff, S., Vuichard, N., Warszawski, L., Wiltshire, A. and Woodward, F. I.: Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO2., Proc. Natl. Acad. Sci. U. S. A., 111(9), 3280–5, doi:10.1073/pnas.1222477110, 2014.

Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J. and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298(1–4), 222–241, doi:10.1016/j.jhydrol.2004.03.037, 2004.

Giglio, L., Randerson, J. T. and van der Werf, G. R.: Analysis of daily, monthly, and annual burned area using the fourth-generation global fire emissions database (GFED4), J. Geophys. Res. Biogeosciences, 118(1), 317–328, doi:10.1002/jgrg.20042, 2013.

Giorgi, F. and Mearns, L. O.: Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) Method, J. Clim., 15, 1141–1158, doi:10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2, 2002.

Hempel, S., Frieler, K., Warszawski, L., Schewe, J. and Piontek, F.: A trend-preserving bias correction – the ISI-MIP approach, Earth Syst. Dyn., 4, 219–236, doi:10.5194/esd-4-219-2013, 2013.

Huisman, J. A., Breuer, L., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., Viney, N. R. and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM) III: Scenario analysis, Adv. Water Resour., 32(2), 159–170, doi:10.1016/j.advwatres.2008.06.009, 2009.

Huntzinger, D. N., Schwalm, C., Michalak, A. M., Schaefer, K., King, A. W., Wei, Y., Jacobson, A., Liu, S., Cook, R. B., Post, W. M., Berthier, G., Hayes, D., Huang, M., Ito, A., Lei, H., Lu, C., Mao, J., Peng, C. H., Peng, S., Poulter, B., Riccuito, D., Shi, X., Tian, H., Wang, W., Zeng, N., Zhao, F., and Zhu, Q.: The North American Carbon Program Multi-Scale Synthesis and Terrestrial Model Intercomparison Project – Part 1: Overview and experimental design, Geosci. Model Dev., 6, 2121-2133, https://doi.org/10.5194/gmd-6-2121-2013, 2013.

Huntzinger, D. N., Michalak, A. M., Schwalm, C., Ciais, P., King, A. W., Fang, Y., Schaefer, K., Wei, Y., Cook, R. B., Fisher, J. B., Hayes, D., Huang, M., Ito, A., Jain, A. K., Lei, H., Lu, C., Maignan F., Mao J., Parazoo N., Peng S., Poulter B., Ricciuto D., Shi, X., Tian, H., Wang, W., Zeng, N., and Zhao, F.: Uncertainty in the response of terrestrial carbon sink to environmental drivers undermines carbon-climate feedback predictions, Scientific Reports 7, 4765, doi:10.1038/s41598-017-03818-2, 2017.

Ito, A. and Inatomi, M.: Water-Use Efficiency of the Terrestrial Biosphere: A Model Analysis Focusing on Interactions between the Global Carbon and Water Cycles, J. Hydrometeorol., 13(2), 681–694, doi:10.1175/JHM-D-10-05034.1, 2012.

Jung, M., Reichstein, M. and Bondeau, A.: Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model, Biogeosciences, 6(10), 2001–2013, doi:10.5194/bg-6-2001-2009, 2009.

Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlström, A., Arneth, A., Gustau Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Jain, A. K., Kato, E., Papale, D., Poulter, B., Raduly, B., Rödenbeck, C., Tramontana, G., Viovy, N., Wang, Y.-P., Weber, U., Zaehle, S. and Zeng, N.: Compensatory water effects link yearly global land CO2 sink changes to temperature, Nature, 541, 516-520, doi:10.1038/nature20780, 2017.

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., Arneth, A., Bernhofer, C., Bonal, D., Chen, J., Gianelle, D., Gobron, N., Kiely, G., Kutsch, W., Lasslop, G., Law, B. E., Lindroth, A., Merbold, L., Montagnani, L., Moors, E. J., Papale, D., Sottocornola, M., Vaccari, F., and Williams, C.: Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations, J. Geophys. Res.-Biogeo., 116, G00J07, doi:10.1029/2010JG001566, 2011

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett. 40(6), 1194-1199, doi:10.1002/grl.50256, 2013.

Kolby Smith, W., Reed, S. C.,Cleveland, C. C., Ballantyne, A. P., Anderegg, W. R. L., Wieder, W. R., Liu, Y. Y., and Running, S. W.: Large divergence of satellite and Earth system model estimates of global terrestrial CO2 fertilization, Nature Clim. Change, 6, 306–310, doi:10.1038/nclimate2879, 2015.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S. and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science (80-.)., 285(5433), 1548–1550, doi:10.1126/science.285.5433.1548, 1999.

Lasslop, G., Reichstein, M., Papale, D., Richardson, A. D., Arneth, A., Barr, A., Stoy, P., and Wohlfahrt, G.: Separation of net ecosystem exchange into assimilation and respiration using a light response curve approach: critical issues and global evaluation, Glob. Change Biol., 16(1), 187-208, doi:10.1111/j.1365-2486.2009.02041.x

Le Quéré, C., Raupach, M. R., Canadell, J. G., Marland, G., Bopp, L., Ciais, P., Conway, T. J., Doney, S. C., Feely, R. A., Foster, P., Friedlingstein, P., Gurney, K., Houghton, R. A., House, J. I., Huntingford, C., Levy, P. E., Lomas, M. R., Majkut, J., Metzl, N., Ometto, J. P., Peters, G. P., Prentice, I. C., Randerson, J. T., Running, S. W., Sarmiento, J. L., Schuster, U., Sitch, S., Takahashi, T., Viovy, N., van der Werf, G. R., and Woodward, F. I.: Trends in the sources and sinks of carbon dioxide, Nat. Geosci., 2, 831–836, doi:10.1038/ngeo689, 2009.

Lovenduski, N. S. and Bonan, G. B.: Reducing uncertainty in projections of terrestrial carbon uptake, Env. Res. Lett., 12(4), 044020, 2017.

Moncrieff, G. R., Scheiter, S., Langan, L., Trabucco, A., and Higgins, S. I.: The future distribution of the savannah biome: model-based and biogeographic contingency, Philos. Trans. R. Soc. B-Biol. Sci., 371(1703), 20150311, doi: 10.1098/rstb.2015.0311, 2016.

Myneni, R. B., Hoffman, S., Knyazikhin, Y., Privette, J. L., Glassy, J., Tian, Y., Wang, Y., Song, X., Zhang, Y., Smith, G. R., Lotsch, A., Friedl, M., Morisette, J. T., Votava, P., Nemani R. R., and Running, S. W.: Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data, Remote Sens. Environ., 83(1-2), 214-231, 2002.

Nishina, K., Ito, A., Beerling, D. J., Cadule, P., Ciais, P., Clark, D. B., Falloon, P., Friend, A. D., Kahana, R., Kato, E., Keribin, R., Lucht, W., Lomas, M., Rademacher, T. T., Pavlick, R., Schaphoff, S., Vuichard, N., Warszawaski, L. and Yokohata, T.: Quantifying uncertainties in soil carbon responses to changes in global mean temperature and precipitation, Earth Syst. Dyn., 5(1), 197–209, doi:10.5194/esd-5-197-2014, 2014.

Nishina, K., Ito, A., Falloon, P., Friend, A. D., Beerling, D. J., Ciais, P., Clark, D. B., Kahana, R., Kato, E., Lucht, W., Lomas, M., Pavlick, R., Schaphoff, S., Warszawaski, L. and Yokohata, T.: Decomposing uncertainties in the future terrestrial carbon budget associated with emission scenarios, climate projections, and ecosystem simulations using the ISI-MIP results, Earth Syst. Dyn., 6(2), 435–445, doi:10.5194/esd-6-435-2015, 2015.

Norby, R. J., Warren, J. M., Iversen, C. M., Medlyn, B. E., and McMurtrie, R. E.: CO2 enhancement of forest productivity constrained by limited nitrogen availability, Proc. Natl. Acad. Sci., 107, 19368–19373, doi:10.1073/pnas.1006463107, 2010.

Pavlick, R., Drewry, D. T., Bohn, K., Reu, B. and Kleidon, A.: The Jena Diversity-Dynamic Global Vegetation Model (JeDi-DGVM): a diverse approach to representing terrestrial biogeography and biogeochemistry based on plant functional trade-offs, Biogeosciences, 10(6), 4137–4177, doi:10.5194/bg-10-4137-2013, 2013.

Poulter, B., Frank, D., Ciais, P., Myneni, R. B., Andela, N., Bi, J., Broquet, G., Canadell, J. G., Chevallier, F., Liu, Y. Y., Running, S. W., Sitch, S. and van der Werf, G. R.: Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle, Nature, 509(7502), 600–603, doi:10.1038/nature13376, 2014.

Raftery, A. E., Gneiting, T., Balabdaoui, F. and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133, 1155–1174, doi:10.1175/MWR2906.1, 2005.

Rammig, A., Jupp, T., Thonicke, K., Tietjen, B., Heinke, J., Ostberg, S., Lucht, W., Cramer, W. and Cox, P.: Estimating the risk of Amazonian forest dieback, New Phytol., 187(3), 694–706, doi:10.1111/j.1469-8137.2010.03318.x, 2010.

Reichstein, M., Falge, E., Baldocchi, D., et al.: On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm, Glob. Change Biol., 11(9), 1424–1439, 2005.

Running, S. W., Nemani, R. R., Heinsch, F. A., Zhao, M., Reeves, M. and Hashimoto, H.: A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production, Bioscience, 54(6), 547, doi:10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2, 2004.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M. and Morel, A.: Benchmark map of forest carbon stocks in tropical regions across three continents., Proc. Natl. Acad. Sci. U. S. A., 108(24), 9899–9904, doi:10.1073/pnas.1019576108, 2011.

Schwalm, C. R., Huntzinger, D. N., Fisher, J. B., Michalak, A. M., Bowman, K., Ciais, P., Cook, R., El-Masri, B., Hayes, D., Huang, M., Ito, A., Jain, A., King, A. W., Lei, H., Liu, J., Lu, C., Mao, J., Peng, S., Poulter, B., Ricciuto, D., Schaefer, K., Shi, X., Tao, B., Tian, H., Wang, W., Wei, Y., Yang, J. and Zeng, N.: Toward "optimal" integration of terrestrial biosphere models, Geophys. Res. Lett., 42(11), 4418–4428, doi:10.1002/2015GL064002, 2015.

Shamseldin, A. Y., O'Connor, K. M. and Liang, G. C.: Methods for combining the outputs of different rainfall–runoff models, J. Hydrol., 197(1–4), 203–229, doi:10.1016/S0022-1694(96)03259-3, 1997.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J. O., Levis, S., Lucht, W., Sykes, M. T., Thonicke, K. and Venevsky, S.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Glob. Chang. Biol., 9(2), 161–185, doi:10.1046/j.1365-2486.2003.00569.x, 2003.

Smallman, T. L., Exbrayat, J.-F., Mencuccini, M., Bloom, A. A., and Williams, M.: Assimilation of repeated woody biomass observations constrains decadal ecosystem carbon cycle uncertainty in aggrading forests, J. Geophys. Res. Biogeosciences, 122(3), 528-545, doi:10.1002/2016JG003520, 2017.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bull. Am. Meteorol. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Thornton, P. E., Doney, S. C., Lindsay, K., Moore, J. K., Mahowald, N., Randerson, J. T., Fung, I., Lamarque, J.-F., Feddema, J. J., and Lee, Y.-H.: Carbon-nitrogen interactions regulate climate-carbon cycle feedbacks: results from an atmosphere-ocean general circulation model, Biogeosciences, 6, 2099-2120, https://doi.org/10.5194/bg-6-2099-2009, 2009.

Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, Biogeosciences, 13, 4291-4313, https://doi.org/10.5194/bg-13-4291-2016, 2016.

Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Gräff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G. W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M. and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32(2), 147–158, doi:10.1016/j.advwatres.2008.05.006, 2009.

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O. and Schewe, J.: The Inter-Sectoral Impact Model Intercomparison Project (ISI-MIP): project framework., Proc. Natl. Acad. Sci. U. S. A., 111(9), 3228–32, doi:10.1073/pnas.1312330110, 2014.

Watanabe, S., Hajima, T., Sudo, K., Nagashima, T., Takemura, T., Okajima, H., Nozawa, T., Kawase, H., Abe, M., Yokohata, T., Ise, T., Sato, H., Kato, E., Takata, K., Emori, S., and Kawamiya, M.: MIROC-ESM 2010: model description and basic results of CMIP5-20c3m experiments, Geosci. Model Dev., 4, 845-872, doi:10.5194/gmd-4-845-2011, 2011.

Wieder, W. R., Cleveland, C. C. Kolby Smith, W. and Todd-Brown, K. E. O.: Future productivity and carbon storage limited by terrestrial nutrient availability, Nature Geosci., 8, 441-444, doi: 10.1038/ngeo2413, 2015.

Williams, M., Schwarz, P. A., Law, B. E., Irvine, J. and Kurpius, M. R.: An improved analysis of forest carbon dynamics using data assimilation, Glob. Chang. Biol., 11(1), 89–105, doi:10.1111/j.1365-2486.2004.00891.x, 2005.

Woodward, F., Smith, T., and Emanuel, W.: A global land primary productivity and phytogeography model, Global Biogeochem. Cy., 9, 471–490, 1995.

Zhang, Q., Wang, Y. P., Pitman, A. J., and Dai, Y. J.: Limitations of nitrogen and phosphorous on the terrestrial carbon uptake in the 20th century, Geophys. Res. Lett., 38, L22701, doi:10.1029/2011GL049244, 2011.

Zhang, Q., Wang, Y. P., Matear, R. J., Pitman, A. J., and Dai, Y. J.: Nitrogen and phosphorous limitations significantly reduce future allowable CO2 emissions, Geophys. Res. Lett., 41, 632-637, doi: 10.1002/2013GL058352, 2013.

Zhao, M., Heinsch, F. A., Nemani, R. R., and Running, S. W.: Improvements of the MODIS terrestrial gross and net primary production global data set, Remote Sensing of Environment, 95, 164–176, doi:10.1016/j.rse.2004.12.011, 2005.

Zhao, M. and Running, S. W.: Drought-Induced Reduction in Global, Science (80-.)., 329(5994), 940–943, doi:10.1126/science.1192666, 2010.

Deleted: ¶

Tables

Table 1: Information about global vegetation models used here. For each GVMs we indicate the range of values obtained while driving it with 5 GCMs.

Model	NPP (2001-2005) Pg C y^{-1}	ΔNPP (2095-2099) Pg C y^{-1}	Nitrogen[a]	Reference
HYBRID	63.7 – 77.8	-3.7 – 26.2	Yes	Friend and White (2000)
JeDi	55.8 – 65.2	23.3 – 32.3	No	Pavlick et al. (2013)
JULES	65.6 – 72.3	34.6 – 41.6	No	Clark et al. (2011)
LPJ	70.4 – 76.8	25.2 – 35.9	No	Sitch et al. (2003)
SDGVM	72.0 – 76.0	30.2 – 37.1	Yes	(Woodward et al., 1995)
VISIT	51.7 – 60.7	28.4 – 32.6	No	Ito and Inatomi (2012)

[a] from Nishina et al. (2015)

Deleted: 1996
Deleted: 5
Deleted: 76
Deleted: 1
Deleted: 17
Deleted: 0
Deleted: 25
Deleted: 3
Deleted: 5
Deleted: 63
Deleted: 8
Deleted: 24.6
Deleted: 2
Deleted: 3
Deleted: 1
Deleted: 1
Deleted: 5
Deleted: 4.1
Deleted: 4
Deleted: 69.6
Deleted: 5.6
Deleted: 6.7
Deleted: 0
Deleted: 0.9
Deleted: 4
Deleted: 8
Deleted: 32.3
Deleted: 5
Deleted: 59.7
Deleted: 9.1
Deleted: 3

Figures

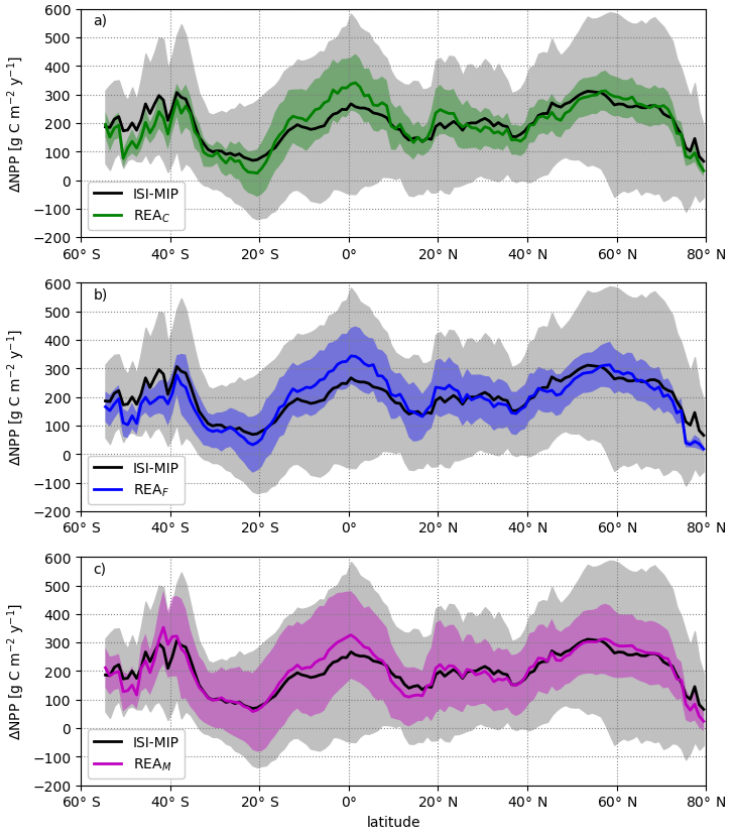

Figure 1: Zonal mean ΔNPP by the end of the 21st century (averaged over 2095-2099) under RCP8.5 compared to the end of the historical simulations (averaged over 2001-2005). Shading represents the uncertainty around the zonal mean across the ISI-MIP ensemble, taken as one standard deviation for ISI-MIP, and calculated following equation (4) for REA. REA$_C$, REA$_F$ and REA$_M$, refer to REA values calculated based on observationally-constrained CARDAMOM, FLUXCOM and MODIS NPP respectively.

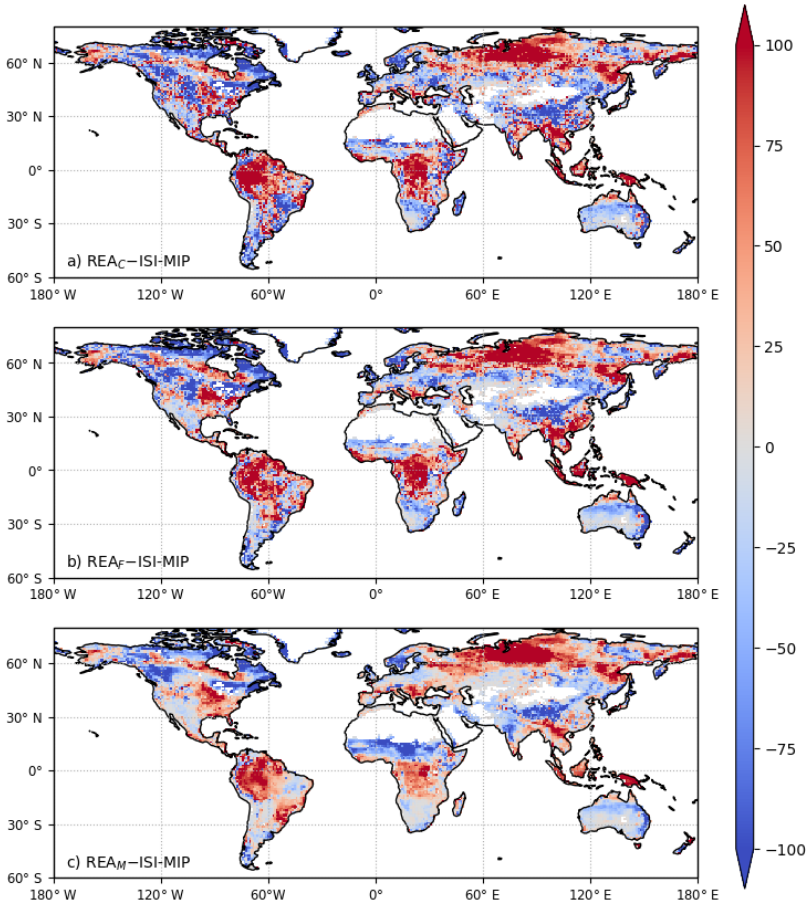

Figure 2: Differences between ΔNPP in 2095-2099 compared to 2001-2005 from the REA average and ISI-MIP ensemble mean (in g C m^{-2} y^{-1}). Red indicates where the REA averages predict ΔNPP greater than the ISI-MIP ensemble mean. Blue indicates where the REA averages predict ΔNPP less than the ISI-MIP ensemble mean.

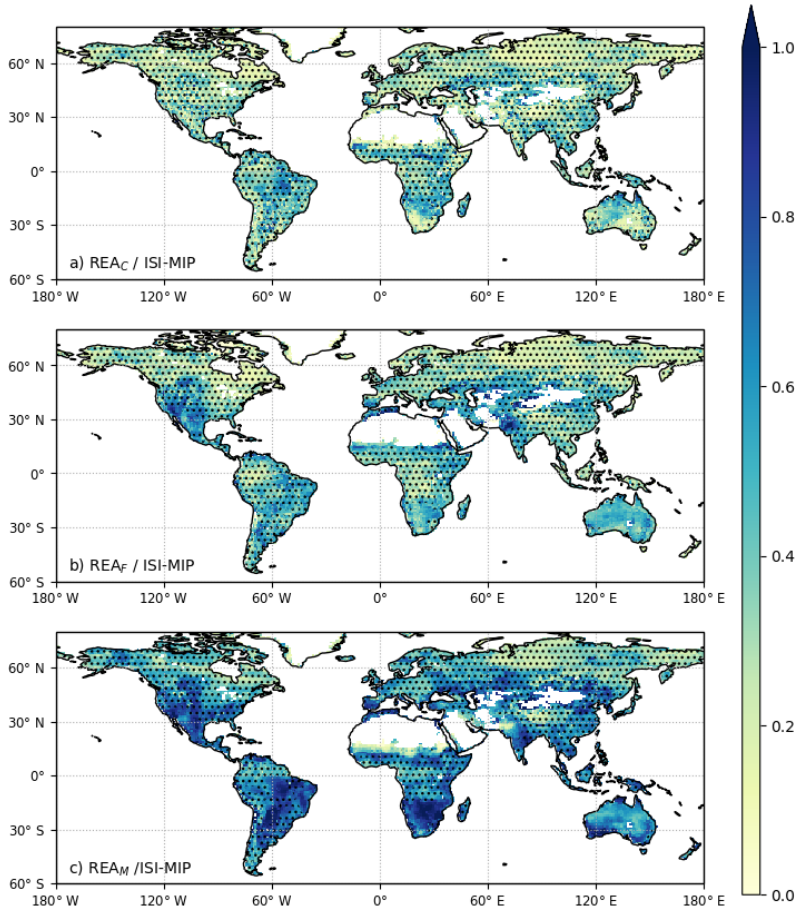

Figure 3: Ratio of the uncertainty in 21st century ΔNPP from each REA to the uncertainty in the ISI-MIP ensemble. For ISI-MIP, the uncertainty is calculated as the standard deviation across the ensemble while the uncertainty around the REA averages is calculated following equation 4. Stippling indicates regions where there is an agreement on the sign of ΔNPP through the uncertainty.

Formatted: Superscript

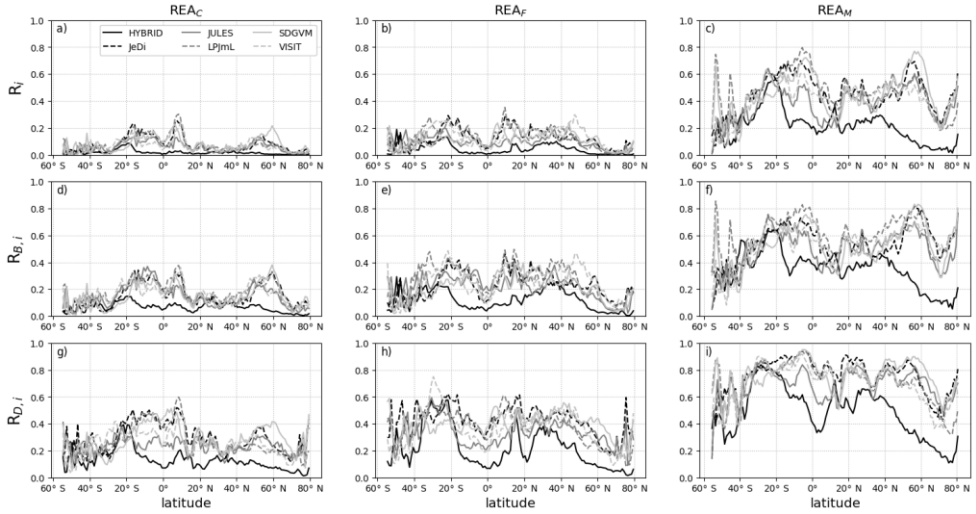

Figure 4: Zonal mean R_i, $R_{B,i}$ and $R_{D,i}$ (row-wise) in each REA$_C$, REA$_F$ and REA$_M$ (column-wise). Each line represents the average value obtained across the five simulations of each GVM.

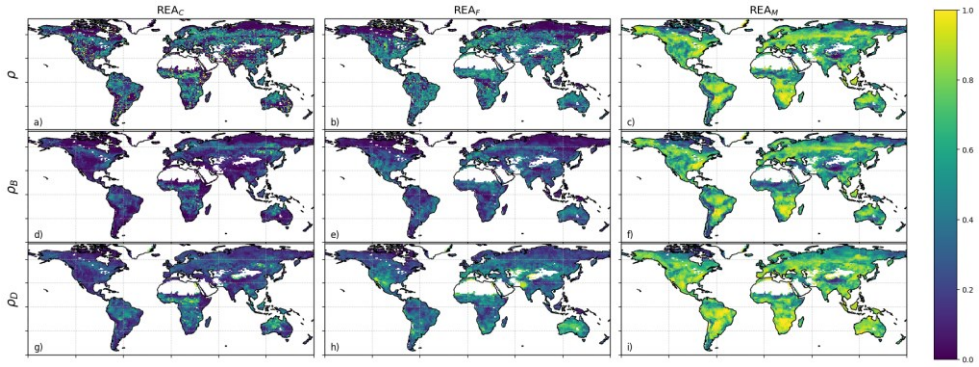

Figure 5: Collective model reliability ρ, model performance ρ_B and model convergence ρ_D (row-wise) for each REA$_C$, REA$_F$ and REA$_M$ (column-wise).

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic, Subscript

Formatted: Font: Italic

Formatted: Font: Italic, Subscript