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 CO_2 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: 3 out of 12 CMIP5 models used by Todd-Brown et al. (2014), 2 out of 8 models in new ISI-MIP experiments presented by Chen et al. (2017). Furthermore, low weights R_i 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 in the revised manuscript.

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 will be updated throughout the manuscript. For example, the first paragraph of the results section will now read (updated numbers in red):

> The REA averaging method yields a global increase of NPP of 24.6 ± 8.5 Pg C y^{-1} (REA average ± RMSD) for CARDAMOM, 24.8 ± 9.5 Pg C y^{-1} for FLUXCOM and 25.0 ± 14.5 Pg C y^{-1} for MODIS NPP. 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.

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 will improve figure captions to include more detailed descriptions. For example, the caption of figure 1 will now read (updated text in red):

> Figure 1: Zonal mean ΔNPP by the end of the 21st century (averaged over 2095 to 2099) under RCP8.5 compared to the end of the historical simulations (averaged over 2001 to 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.

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 will replace with "Benchmark datasets of modern NPP".

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 l. 7 to 17. Therefore, we will remove it from the method section.

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. As noted p 6 l 12-13:

> 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.

We will move the above closer to equation 1 and will include a better description of the range in the revised manuscript.

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.

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.

Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., Shevliakova, E., Tjiputra, J., Volodin, E., Wu, T., Zhang, Q. and Allison, S. D.: Changes in soil organic carbon storage predicted by Earth system models during the 21st century, Biogeosciences, 11(8), 2341–2356, doi:10.5194/bg-11-2341-2014, 2014.

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.