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 will review these aspects in relevant parts of the introduction, methods and discussion of the revised manuscript.

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. 24-25. 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 for improved clarity and we will do so in a revised version of the manuscript.

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 will add the following sentence to the abstract:

> 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 $R_{D,i}$ (p.9 l. 20-23).

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, the revised manuscript will emphasise that REA outcomes should be cautiously interpreted with respect to the ensemble members.

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 provide new Figures 2 to 5 using a colour scheme that is compatible with colour-blindness (checked on http://www.vischeck.com) and renders well on black and white printers. We attach these updated figures at the end of this document and will correct the Supplementary Information accordingly upon submission of a revised manuscript.

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.

References

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.

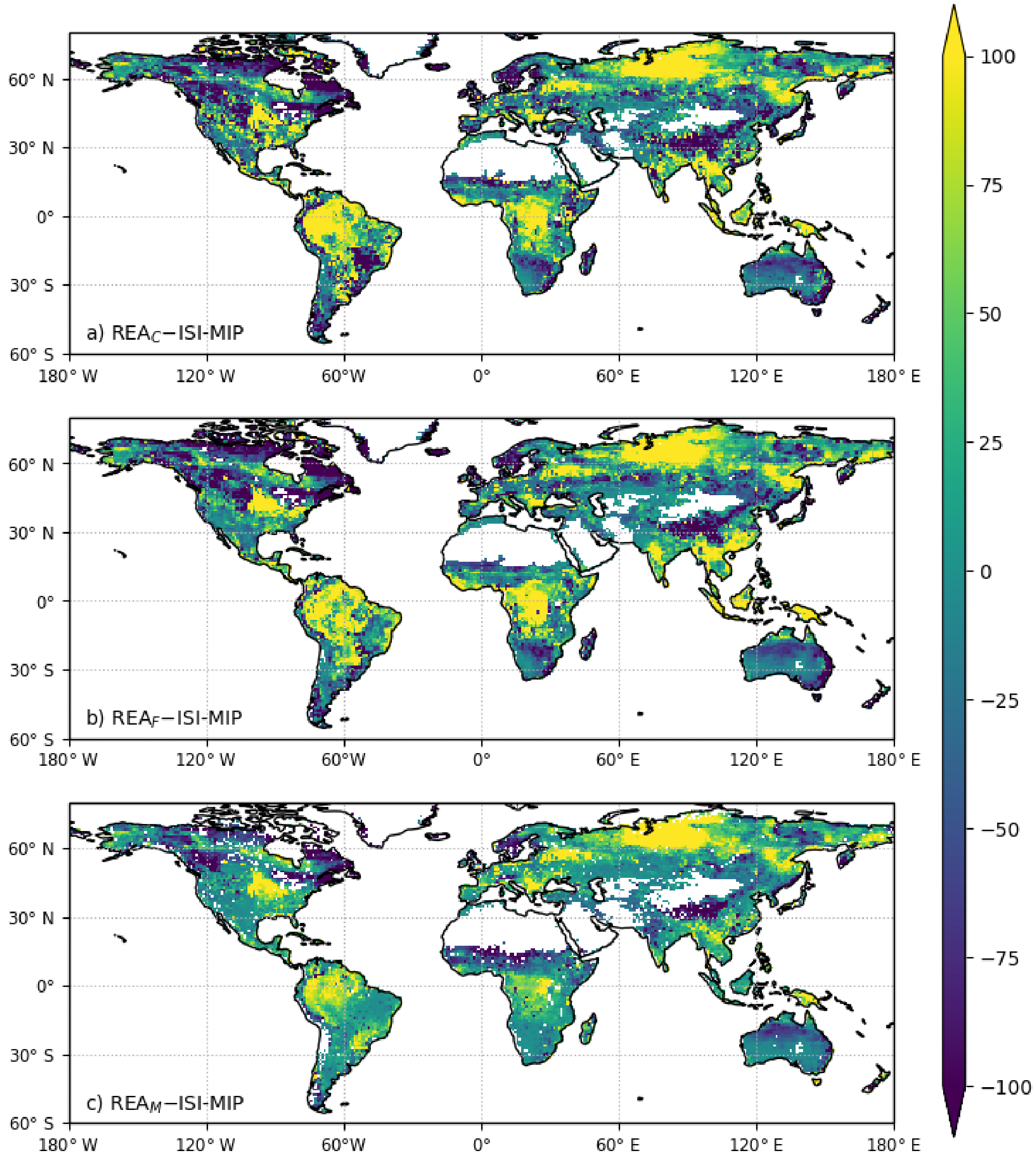

Figure 2: Differences between ΔNPP from REA average and ISI-MIP ensemble mean (in g C m^{-2} y^{-1}). Red indicates where the REA averages predict ΔNPP greater than the ISI-MIP ensemble mean. Blue indicates where the REA averages predict ΔNPP less than the ISI-MIP ensemble mean.

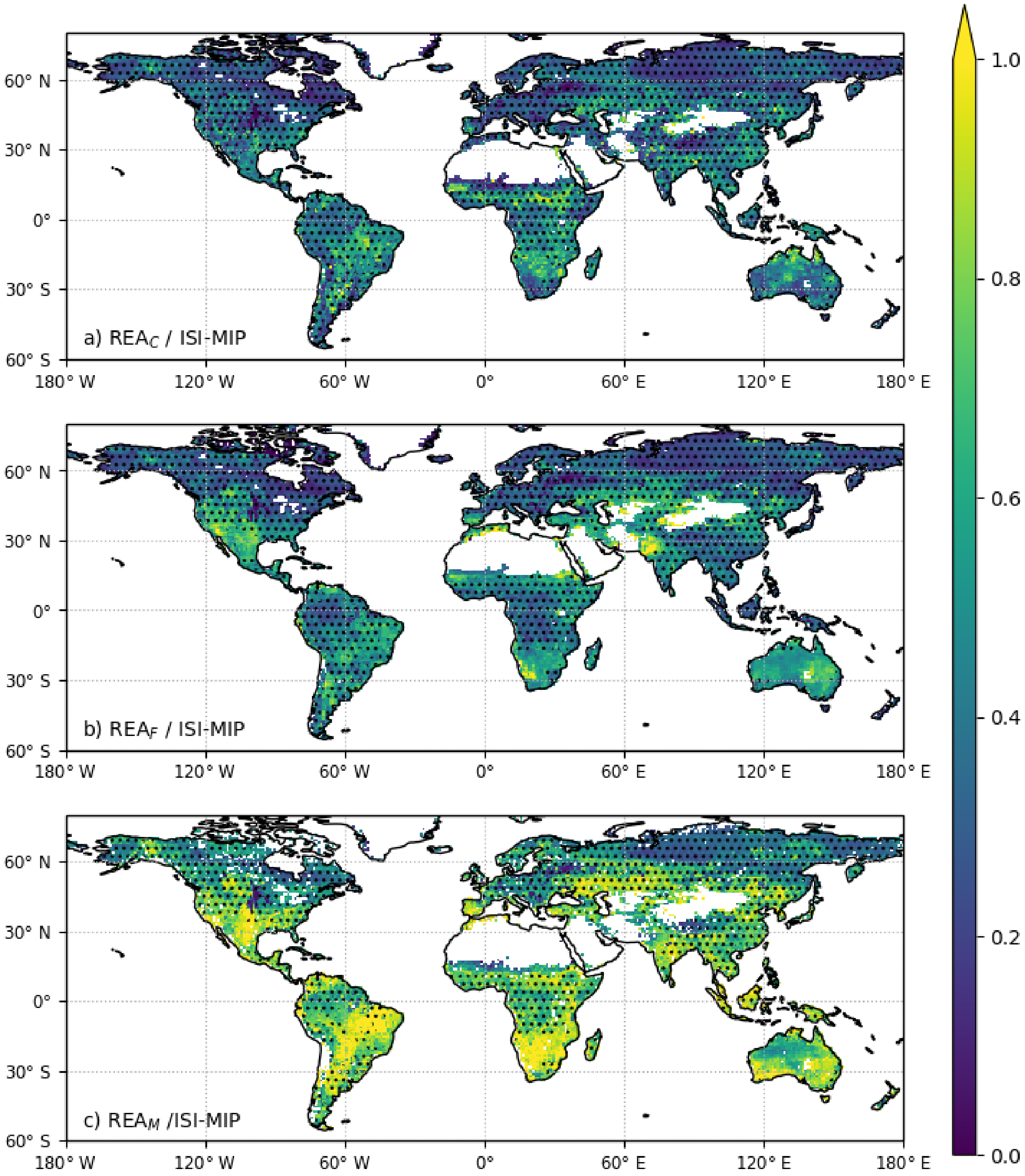

Figure 3: Ratio of the uncertainty from each REA to the uncertainty in the ISI-MIP ensemble. For ISI-MIP, the uncertainty is calculated as the standard deviation across the ensemble while the uncertainty around the REA averages is calculated following equation 4. Stippling indicates regions where there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

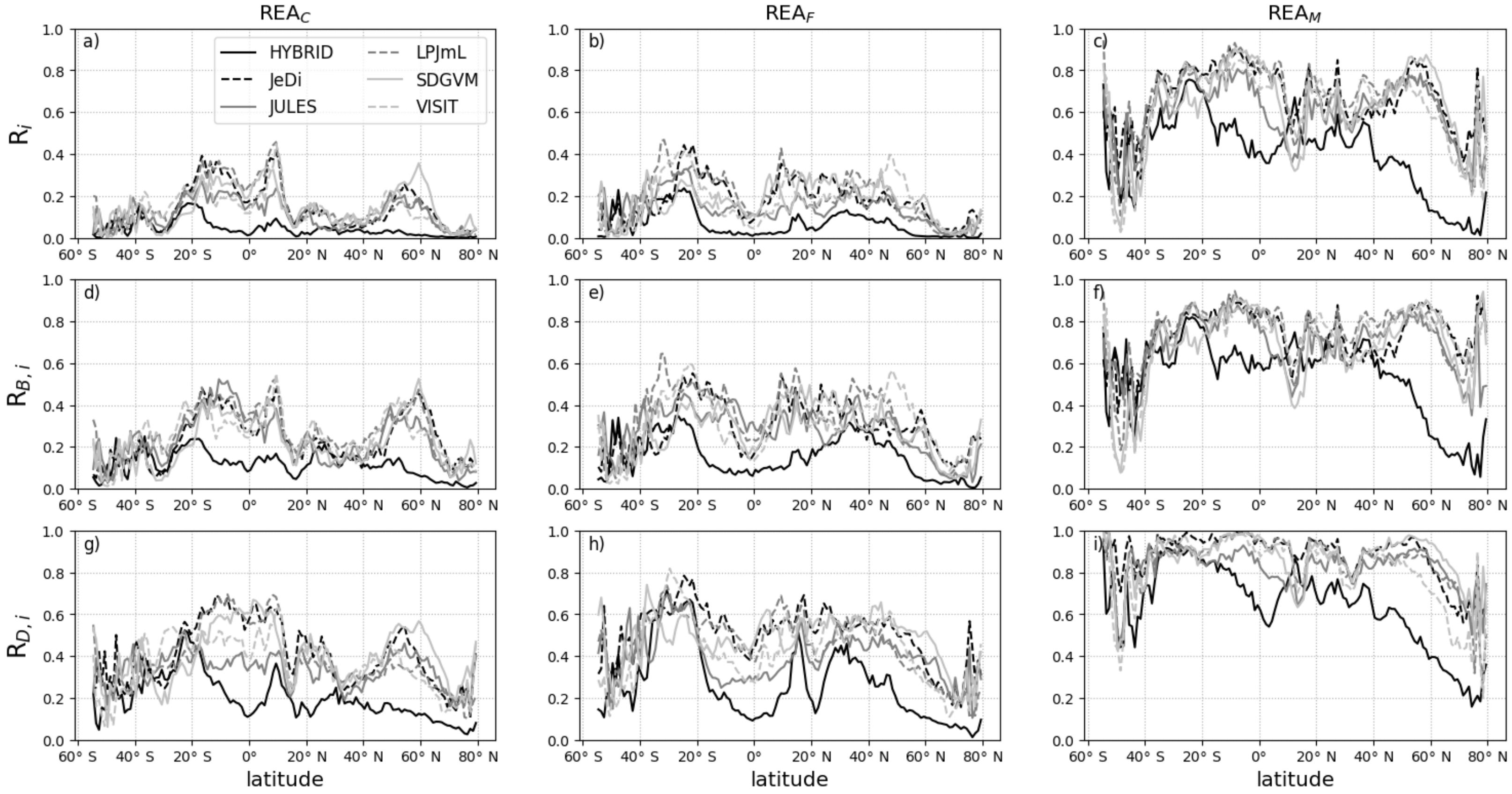

Figure 4: Zonal mean R_i, $R_{B,i}$ and $R_{D,i}$ (row-wise) in each REA$_C$, REA$_F$ and REA$_M$ (column-wise). Each line represents the average value obtained across the five simulations of each GVM.

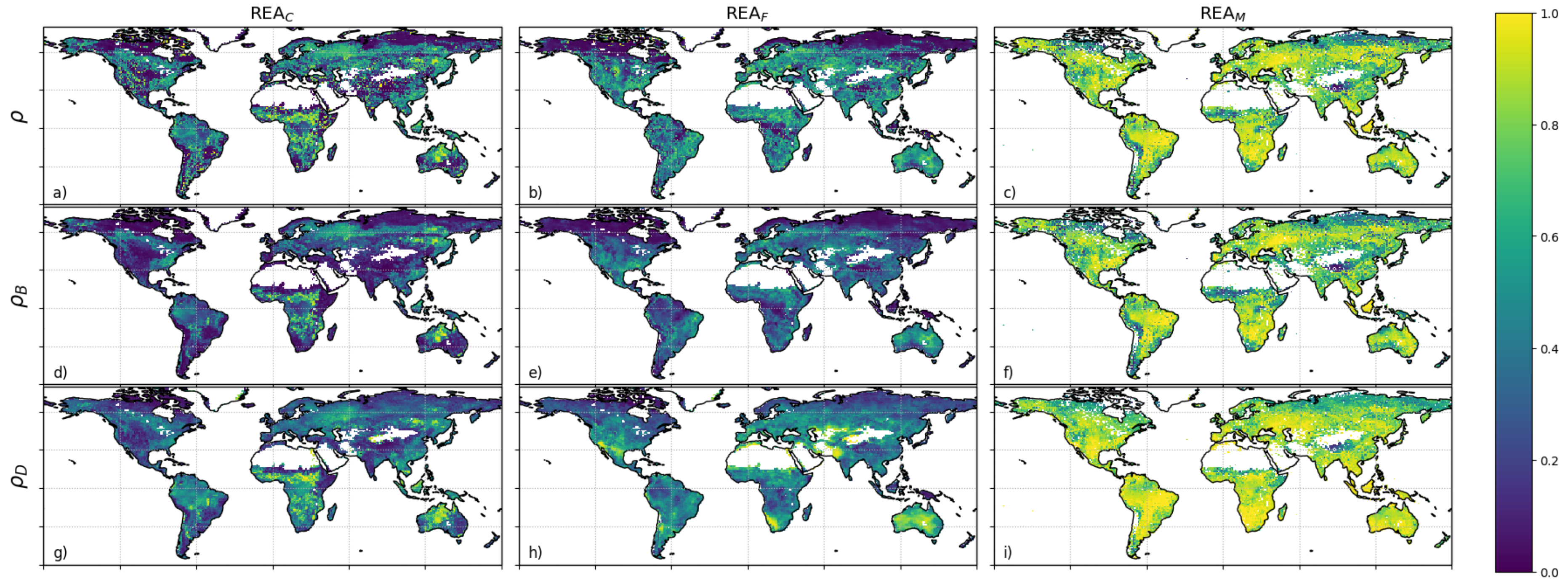

Figure 5: Collective model reliability ρ, model performance ρ_B and model convergence ρ_D (row-wise) for each REA$_C$, REA$_F$ and REA$_M$ (column-wise).