Any bias in climate forcing directly influences the model projection of carbon cycle dynamics. Teckentrup et al. quantify the impacts of different bias correction (including univariate correction, multivariate correction, model averages, and random forest method) methods on improving the outputs of carbon stock changes from a dynamic global vegetation model (LPJ-GUESS). This draft was well-written, but I still have some comments on the algorithms used in this analysis, and I think the novelty is insufficient for a paper in ESD.

Major comments:

1. The biggest concern is that after reading I still have no idea which bias correction method should be used to assess the spatial variability or short-term and long-term temporal variation in the total carbon stocks. The results are quite confusing. It would be good to evaluate the classifications of correction methods by function
2. The authors should perform a synthetic analysis and evaluation. The current results are very preliminary.

We thank the reviewer for their comments. The role of biases in climate forcing is particularly problematic when considering the future of ecosystems that experience the extremes of climate (i.e., very wet, very dry, very hot, very cold). To date, research has focussed on the impact of corrections at global scales, which makes it hard to be sure whether correction approaches affect the nuance of conclusions on regional scales. Our contribution is novel exactly because it tackles a suite of bias correction through the lens of a regional element of the terrestrial carbon cycle that is subject to climate extremes.

The reviewer is then making two major comments.

First, we interpret the reviewers comment to imply that they were after some sort of ranking of best practice. Our results do not definitively support a single "best" approach. Instead, we show that considering forcing biases is an integral step in carbon cycle projections and has the potential to considerably reduce biases in projections of carbon pools. This is an important step in reducing the large uncertainties in carbon cycle projections. Furthermore, we highlight that bias correction methods have a large potential to reduce carbon cycle uncertainty, namely the simulation of long-term average of carbon pools. Conversely, the bias-correction of forcing data was not able to significantly reduce biases in carbon cycle trends and interannual variability. These findings provide clearer guidance for selecting ensemble weighting and bias correction methods in future studies.

Although we did group our evaluation by function, in light of the reviewers comment we will carefully revise to ensure grouping by function is clearer. We will also revise the discussion to better highlight key findings and hope these aspects of the paper will be clearer in the revised version.

Second, the reviewer asks for a synthetic analysis. On this point we do not feel this is within the scope of aims of our manuscript. We feel that this type of approach would be more suited for evaluation of individual methods, whereas our paper is trying to synthesise the impact of different approaches through a common lens (LPJ-GUESS and the carbon cycle). We feel that *any additional synthetical analysis would simply overcomplicate the story by focussing on very specific details.*

Specific comments:

Fig 1: It would be good to differentiate steps and the name of methods in each step. Can use different icons or colors.

We will review this figure and the caption to see if we can make it easier to interpret. We will also expand the caption to better explain how steps/methods should be interpreted.

Table 2: Some of these selected metrics reflect the same (similar) property. For example, all the Root mean squared error, Normalised Mean Error, and Mean bias error indicates the bias in mean value. So the model with good skill in simulating mean value tends to have a higher rank. It is unfair.

We based the selection of error metrics based on Naughton et al, 2018, who choose these metrics. However, we appreciate the concern of reviewer 1 and will test the ranking using only one of these bias metrics to see how it influences the rankings and revise the manuscript accordingly.

Ln148: Why use the correlation of 0.3 as a threshold to select the models?

We agree with the reviewer that the choice of 0.3 as a threshold may seem arbitrary (as would any choice of threshold). Given the selection of independent GCMs is based on the assumption that simulations are independent when their biases are not correlated, we choose 0.3 as a threshold given it is relatively commonly interpreted as the threshold between weak and moderate correlations. We will clarify this in the revised manuscript.

Ln305-314: Please clarify which meteorological forcing influence the mean value of C_total, the short-term variability (i.e., inter-annual variability) in C_total, and the long-term variability in C-total.

We thank the reviewer for the suggestion and will adjust the manuscript accordingly. Given Australia is strongly water-limited, both total ecosystem carbon and interannual variability is mostly driven by precipitation (compare Haverd et al., 2013). We also tested this in an additional analysis (which is not shown in the paper) where we only bias-corrected the precipitation data, leaving temperature and radiation as the raw data values. This yielded very similar C_total to runs where all meteorological variables were bias-corrected, suggesting the influence of precipitation dominates in our study region (discussed in "General caveats"). However given bioclimatic limits are prescribed in the model used for this study, temperature will influence the vegetation distribution as well (see below).

Ln320: In Fig3, only squares and circles indicate a larger bias in mean PPT after multivariate bias correction. Why?

We agree with the reviewer that this is a surprising result, and note that in fact this behaviour is also apparent for the univariate CDF-t approach. This will require further investigation, and will be discussed in the revisions.

Ln350: The authors should give a summarized metric showing which bias correction method is better. It is difficult to find the best model by eyes.

We appreciate the reviewer's suggestion, and will include new text to help guide the reader better by discussing the advantages and disadvantages of each approach.

The spatial patterns of bias and CV of C_total simulated by the model in Fig 5 and Fig 6 have a clear and strange strip with extreme values. This is not reasonable. Could you please explain why this strip exists?

In this study, we employed the dynamic global vegetation model LPJ-GUESS. Like many DGVMs (see Fisher et al., 2015), it prescribes bioclimatic limits that define the geographic location growth based on temperature. For example, vegetation growth of C_4 grasses and tropical trees is restricted by a lower temperature boundary such that these vegetation types cannot establish or survive when the 20-year average minimum temperature falls below 15.5 degrees Celsius. Therefore, C_4 grasses and tropical trees only grow north of the Tropic of Capricorn, while south of it only temperate trees and C_3 grasses are simulated. The strong variation across GCMs in simulated temperature thus leads to very different simulated vegetation cover in LPJ-GUESS. Given the boundary between C_3 and C_4 grasses will depend on the GCM used to force LPJ-GUESS, the type of vegetation in this area varies by the GCM forcing, leading to the large error and uncertainty in simulated carbon. We will discuss this in the revised manuscript.

Ln374-375: It is not clear why C4 grasses would have a higher CV. The authors did not convince me that this is the real reason.

We explained the pattern in the comment above, and apologise that this was not made clearer, and will revise the manuscript accordingly.

Ln379-380: The authors did not explain why the bias in C_total relates to foliar projective cover/ Could you please show the relation between C_total and foliar projective cover? Which factors or processes can influence foliar projective cover in the LPJ-GUESS model?

We apologise for the lack of clarity in this point and will update the manuscript accordingly. Foliar projective cover can be seen as in indicator for the vegetation growth, which ultimately defines the ecosystem carbon. For example, areas with high foliar projective cover (i.e., trees) will tend to intercept greater PAR and thus assimilate more carbon

Simulated foliar projective cover also results from vegetation competition in LPJ-GUESS. This in turn is influenced by the climate and other input datasets: For example, the arid areas of Australia will show strong water limitation (i.e. no precipitation, high temperatures) which create unfavourable conditions for tree growth, so that grasses become more competitive. In the coastal areas, and in the tropics of Australia, water is abundant and allows tree growth at the cost of grass expansion. There are also competitive processes amongst tree species, and C3 and C4 grasses, that are driven by temperature (either dynamically or prescribed), or based *on incoming short-wave radiation, i.e., vegetation can be shade-tolerant or shade-intolerant. However, incoming shortwave radiation is not a limiting factor in Australia and can therefore be largely neglected for this study. We will clarify this in the revised manuscript.*

Ln415-417: The peak of seasonal GPP was underestimated a lot. Is this because the peak of meteorological variables (like precipitation or temperature) was underestimated and uncorrected?

We assume that indeed the lower peak in seasonal GPP is driven by the overall bias in both temperature, and precipitation. However, removing the bias in the input forcing does not achieve a perfect match with the target dataset. This is likely because the bias correction did not consider the spatial domain, i.e., the spatial pattern of vegetation distribution remains similar, only the magnitude in bias is reduced. We will make this clearer in the revised manuscript.

Ln421-422: The bias in the dry season seems very small. So the effects of correcting data in the dry season may not be very useful?

We thank the reviewer for the question and will clarify this point in the manuscript. The bias for the individual GCMs may appear relatively low, however the ensemble spread across the full CMIP6 ensemble (shaded grey area) indicates significant uncertainty, and depending on the GCM chosen, dry season GPP can either be zero or reach values over 0.2 PgC which is roughly two third of peak wet season GPP in the target dataset.

Ln442-445 The introduction of the importance of Australia for the estimation of global land carbon sink should not be in the Discussion. Can put it into the Introduction.

Yes, we will move the paragraph to the introduction.

Ln469-470: Don't repeat the results of the analysis in the Discussion.

We thank the reviewer for the suggestion, and will remove this paragraph.

References

Fisher, R. A., Muszala, S., Verteinstein, M., Lawrence, P., Xu, C., McDowell, N. G., Knox, R. G., Koven, C., Holm, J., Rogers, B. M., Spessa, A., Lawrence, D., and Bonan, G.: Taking off the training wheels: the properties of a dynamic vegetation model without climate envelopes, CLM4.5(ED), Geosci. Model Dev., 8, 3593–3619, https://doi.org/10.5194/gmd-8-3593-2015, 2015.

Haverd, V., Raupach, M. R., Briggs, P. R., J. G. Canadell., Davis, S. J., Law, R. M., Meyer, C. P., Peters, G. P., Pickett-Heaps, C., and Sherman, B.: The Australian terrestrial carbon budget, Biogeosciences, 10, 851–869, https://doi.org/10.5194/bg-10-851-2013, 2013.