Reviewer 1

Teckentrup et al. made a lot of efforts to improve this research. Super nice! But I still have some minor comments at the current stage.

Abstract

Ln 8-9: The sentence "Carbon pools are insensitive to the type of bias correction method" is unclear to me. Do you mean "model carbon outputs before and after bias correction are similar" or "bias correction can improve model carbon output and different methods show similar outputs"?

We thank the reviewer for the comment and apologise for the confusion. The latter is correct; we have updated the abstract accordingly.

Ln11-12: "Some bias correction methods reduce the ensemble uncertainty more than others." I cannot extract any useful information from this sentence. Please describe some more clear findings. Which method can reduce more uncertainty? How much uncertainty can be reduced?

We apologise for the lack of clarity and now include more details in the abstract.

> *Multivariate bias correction methods tend to reduce the uncertainty more than univariate approaches, although the overall magnitude was similar.*

Ln16: I would suggest to re-write the result part of the abstract.

We thank the reviewer for the suggestion and revised the abstract as noted below.

> *Even after correcting the bias in the meteorological forcing dataset, the simulated vegetation distribution shows different patterns when different GCMs are used to drive LPJ-GUESS. […] This highlights that where possible, an arithmetic ensemble average should be avoided. However, potential target datasets that would facilitate the application of machine learning approaches, i.e., that cover both the spatial and temporal domain required to derive a robust informed ensemble average are sparse for ecosystem variables.*

Main text

Ln 74: Weighting methods should be the third strategy to reduce uncertainty, right? The hypothesis of this method is that the uncertainty of different models can cancel each other out

Weighting approaches are based on similar principles to GCM subselection approaches, i.e., the goal is to derive 'representative' ensemble statistics given model ensembles such as CMIP are ensembles of opportunity. Weighting methods **can** be the third strategy and have been shown to outperform simple ensemble averages (even after a subselection is chosen). However, they depend on the availability of suitable target datasets to derive weights from.

Ln 187: Confusing. Here, the authors said, "The bias correction was then applied to each calendar month". But in Ln 180, they said, "We show the correction based on daily timescales"

We apologise for the confusion and will clarify in the revised manuscript. For all bias correction approaches shown in the main manuscript, a correction term is derived and applied on a daily timestep for each month separately. We also tested the impact of the temporal resolution during the bias correction where in a sensitivity experiment, we derived and applied the bias correction on a monthly timestep covering all timesteps (instead of splitting the timeseries up into smaller slices). We revised the methods section accordingly

> *The projection period was split into ten 25-year slices. The bias correction was then derived and applied to each calendar month on a daily timestep within each time slice separately.*

Ln 276: "The error correlation coefficient is used as …" or "We use the error correlation coefficient…", "We derive the linear combination…" rather than "This method derives…"

We thank the reviewer for the suggestion and updated the revised manuscript accordingly.

Ln300: The sentence "While not all possible combinations of approaches were examined, we employed a wide range of methods" can be removed.

We thank the reviewer for the suggestion and updated the revised manuscript accordingly.

Figure 3: In panel (a). the biases of raw individual model data show most models having positive bias (except for MPI-ESM1-2-HR), so the arithmetic average value should be positive. But why the skill arithmetic average of precipitation bias in panel (a) is negative? Does this mean that the selected models are not representative?

Ln348: The arithmetic ensemble average of the biases in C_total, right?

In line 348 we refer to all variables presented in figure 3, i. e. average precipitation, temperature, and C_{Total}.

Ln391: why are no random forest and weighted ensemble results in Figure 6? These two ensembles with the lowest climate bias should have a good performance in predicting C_total.

We did not include the coefficient of variation for the independence or random forest weighted estimates because they produce a single estimate rather than an ensemble estimate. As such, there is no coefficient of variation across the ensemble, just a 'best estimate'. In the revised paper, we add this into the figure caption:

> *Note that we do not show a coefficient of variation for the weighted ensemble averages. Given they produce a single estimate rather than an ensemble estimate, a coefficient of variation does not exist for these methods.*

Figure 8: GPP seasonality forced by the R2D2 method corrected climate data always have the lowest bias. This would be useful information for readers
We thank the reviewer for the suggestion, and included in the revised manuscript

> *Notably, the R2D2 method always achieves the lowest bias to the target dataset compared to the remaining bias correction methods.*

Reviewer 2

Thank you for inviting me to review paper: "Opening Pandora's box: How to constrain regional simulations of the carbon cycle" by Teckentrup et al.

First, can I apologise for taking three weeks to return this review. It is me who has held things up – the start to 2023 has involved more meetings than is ideal.

I really like the messages of manuscript, and they are important – anything that provides concise information on ESM biases is important. And to my knowledge, there are few studies that frame near-surface meteorology uncertainty in the context of its impact on the terrestrial carbon cycle. However, I am sorry to say that I think the current write-up of the results needs substantial attention. There is little correlation in places between what the paper achieves, and how its findings are described in the text.

First, what does the paper actually do? If I have understood correctly, then a single land surface model (LPJ-GUESS) is forced with many models from the CMIP6 climate model ensemble. Then various methods are used to remove biases in the CMIP6 models' meteorological output, by relaxing back (via different ways) to the CRUJRA climate data. The resultant scaled climatologies are then used to force, again, the LPG-GUESS DGVM. Changes in predicted features of the terrestrial carbon cycle are analyzed, with an emphasis on Australia.

In light of the above, my starting point was the title and Abstract, but a vagueness meant a continuous jumping forward and backward within the paper to make sure my interpretation was valid. Hence, the following changes would help at the paper start:

(1) Title. Avoid an odd title ("Pandora's box"), and state clearly what the paper is about. First, the paper does not cover multiple "regional" locations – it is almost exclusively about Australia, and second and more importantly, I initially thought the paper was about constraining models of the terrestrial carbon cycle. (I was anticipating some sort of analysis of the TRENDY ensemble of DGVMs, given mention of "the carbon cycle"). In fact, the paper is almost exclusively about bias-correcting ESM outputs and assessing its impact on a single uncorrected land model.
We thank the reviewer for the feedback. We here chose Australia as a test bed for other water-limited regions but appreciate the title may be misleading and accordingly changed it to reflect the focus on Australia. We have opted to retain other elements of the title. Given the strong focus on the simulated carbon cycle, and the fact that referee 1 explicitly stated they appreciated the title, we updated the title to:

Opening Pandora's box: Reducing GCM uncertainty in Australian simulations of the carbon cycle

While we agree that assessing the uncertainty linked to different terrestrial biosphere modelling framework is crucial, the impact of the climate forcing on the emerging carbon cycle response has been proven to contribute strongly to carbon cycle uncertainty in previous research (i.e. Ahlström et al., 2012 and 2017; Wu et al., 2017). Indeed, Fig. 2 in the manuscript presents an uncertainty in simulated C_{Total} of 30-70 PgC which is comparable to the uncertainty in C_{Veg}+C_{Soil} across the TRENDY ensemble of roughly 12-90 PgC (see Teckentrup et al., 2021). This emphasizes the need to constrain the impact of climate uncertainty on the simulated carbon cycle. Choosing a single model, in this case LPJ-GUESS, as a representative DGVM of the TRENDY ensemble, allows this manuscript to specifically focus on the impact of climate biases on the carbon cycle/ vegetation response. It is not possible to achieve this using the TRENDY ensemble which was forced using a single meteorological dataset. In the updated manuscript, we updated the introduction to make this point clear:

Using a single model forced with multiple realisations of climate us to separate climate-driven uncertainties from those arising from model.

(2) In the Abstract, as LPG-GUESS itself is not corrected in any way, then this does not provide an overall constraint on land carbon cycle projections. That's fine, as ESM correcting is important, but this needs to be made clear. The confusion occurs with wordings such as: "None of the bias correction methods consistently improve the change in carbon over time". At first reading, I was expecting a comparison again carbon pool datasets – possible some sort of EO product. But the "data" is really LPG-GUESS projections?

Yes, we did not correct LPJ-GUESS. A detailed parametrisation and correction of process representations in any DGVM are not trivial and this was not the focus of our manuscript. This is explicitly mentioned in the methods ('We adopted the global configuration of the model'). Instead, our aim was to get a sense of the impact of uncertainty in the climate datasets used to drive any DGVM in Australia, or, more generally, water limited regions. Therefore, a tuned DGVM is not necessary to understand the effect of climate uncertainty on carbon cycle simulations.

We also note that we are aware that the target datasets chosen in this study, i.e. LPJ-GUESS simulations driven by reanalysis, are not equivalent to observation datasets of any kind. Since we wanted to analyse the impact of the methods chosen to deal with biases rather than achieving a constrained estimate of the simulated carbon cycle (which indeed would need a 'corrected' version of LPJ-GUESS), a somewhat synthetic experiment set up is suitable for this manuscript.

However, we appreciate the reviewer's concern and we have updated ambiguous sections in the revised manuscript. We also revised the abstract

These biases have been identified as a major source of uncertainty in carbon cycle projections, hampering predictive capacity. In this study we examine different methods to reduce uncertainty in simulations of the carbon cycle in Australia arising from biases in climate projections.

(3) At multiple locations, the beginning of the paper talks about geographical variation. For instance, "especially at regional scales, climate projections display large biases…". Again, I think the authors need to be clear that the focus of this paper is almost exclusively on Australia.
We thank the reviewer for their comment and note that we chose Australia as a test bed for water-limited regions. To address this concern, we updated the revised manuscript accordingly and clarify we are focussing on Australia as opposed to several different regions globally. We also now mention Australia in the title as noted previously. We still use the term 'regional' in the paper as Australia is a huge continent with many different climate zones, vegetation types etc. and there is a clear need to better understand carbon cycle uncertainties at regional scales within Australia.

The above points only refer to the start of the paper. But this consistent need for clarity as to what the paper achieves need to follow through the entire manuscript.
We have made changes throughout the paper to clarify the text as the reviewer can hopefully see from the tracked changes document.

The authors raise an important point in the Abstract that it may be necessary to "account for temporal properties in correction or ensemble averaging methods". Unfortunately, a lot of the time-evolving issues are lost in the presentation. Figure 2 makes it very clear that removal of an overall invariant bias (approximated by setting all changes to be such that they are zero in year 1920) fails to remove the subsequent large range of gradients in the years out to 2010. Yet after bias correction, we cannot see the individual time-evolving paths – because in Figure 4, if I understand things OK, the individual lines are more a form of ensemble mean. And this will, by definition, remove the spread of gradients.
We thank the reviewer for the feedback and apologise for the confusion. Figure 4 shows the change in CTotal over time for both ensemble averaging methods (Fig. 4f) and for individual simulations based on the five bounding GCMs (Fig. 4a-e). In the revised manuscript, we updated the figure caption

Figure 4. 30-year moving average of the change in CTotal. In each panel, the bold black line is the change in CTotal obtained using the CRUJRA reanalysis and the grey shaded area represents the full unconstrained CMIP6 model ensemble. Panel a–e show the CTotal change simulated using input from the five individual bounding models separately. The colors show the change in CTotal based on the different bias correction methods. Panel f shows the change in CTotal estimated by the ensemble averaging methods.

The issue with eventual removing trends (so not just offsets) is that this will in effect give each climate model the same transient climate sensitivity, or even the same equilibrium climate sensitivity; ECS. Hence this will make each ESM warm at roughly similarly rates to each other, for the same GHG pathway, as based on known historical warming. Of course, determining the

true ECS is the planet is the number one task of climate research. However, the reason this is not a trivial task, i.e. simply fitting to historical temperature trends, is because we do not know if we are living in a world with a high ECS value, but aerosols are currently offsetting much warming – or the opposite. The authors should be aware of these issues, because simply adding a correction to gradients based on the historical period could still cause major errors when estimating out to year 2100. This might be worth stating in the Discussion part.

We thank the reviewer for their comment. Firstly, to be clear, while we agree with this comment, it has no qualitative implications for what we present here. While this issue could affect some bias correction approaches, those presented here to do not explicitly correct trends. We have nevertheless noted this issue in the discussion:

> *Conversely, explicitly bias correcting trends based on historical data, when the spatiotemporal nature may not yet have clearly emerged, could equally be problematic for unbiased estimation of climate system properties like equilibrium climate sensitivity.*

I would link the Abstract two sentences "Some bias correction methods…." and "The vegetation distribution…." – make the same sentence? Because this is a key point of the manuscript, that when scanning across a range of bias-correction possibilities, major DGVM projection differences remain. Such differences are sufficiently large that even for a single land surface response, that DGVM can estimate alternative vegetation distributions. (This statement then encourages the reader to consider in detail Figure 7).

Following this Reviewer's suggestion and feedback from referee 1, the abstract now reads

> *Multivariate bias correction methods tend to reduce the uncertainty more than univariate approaches, although the overall magnitude is similar. Even after correcting the bias in the meteorological forcing dataset, the simulated vegetation distribution presents different patterns when different GCMs are used to drive LPJ-GUESS.*

In the maps, it is obvious that the analysis is applied to Australia, but for the spatially-averaged time-evolving diagrams, it is not always clear if they apply to Australia, globally, or some other area. Please make sure that that all captions are complete in stating what each diagram represents.

We thank the reviewer for their feedback and updated the figure captions accordingly.

Despite these quite severe caveats, I genuinely believe this can be turned in to a very useful manuscript and is appropriate to ESD. As I am not suggesting any additional analysis, then I am sure that a new version could be generated relatively quickly. There needs to be a rewrite that is much clearer as to what the analysis does (Australia only, multiple bias-correction methods for climate models, no temporal bias-correction and importantly, uncorrected LPJ-GUESS acting as "data"). And that, critically, will illustrate what the manuscript does not do (No corrections to LPJ-GUESS, no "data" as true vegetation carbon data e.g. from EO datasets, no

temporal bias-correction methods – and again, take note of the dangers of the later due to highly uncertain contemporary aerosol forcings).

We thank the reviewer for their detailed feedback and hope we addressed their concerns. We updated the title (see above) and figure captions to make clear that this analysis is focussed on Australia, clearly state that the bias correction methods applied do not correct temporal properties

> *and note that none of the correction methods used here are designed to correct temporal properties of the climate forcing.*

and explicitly mention the fact that simulated carbon variables (when LPJ-GUESS is forced with the CRUJRA reanalysis) are used as target datasets in this study in the introduction

> *We use a single dynamic global vegetation model, LPJ-GUESS (Smith et al., 2014), forced with different versions of CMIP6 climate forcing, as well as LPJ-GUESS forced with the CRUJRA reanalysis (Harris, 2019) as a target dataset for carbon variables, and focus on responses at seasonal to centennial timescales.*

and in the methods

> *In addition, we use LPJ-GUESS runs forced with the CRUJRA reanalysis as reference datasets for carbon variables.*

We further included a discussion point about the equilibrium climate sensitivity (see above).