I was excited to receive the invitation to review this manuscript, as it is a much needed piece of work that contributes to improved understanding of Australia's terrestrial carbon cycle. It's great to see the use of the extensive OzFlux dataset for validation of the new AusEFlux product, as it is no trivial effort to keep sites running, collate and share eddy covariance data on a regular basis for use in these types of studies. The manuscript itself was very well written and easy to follow, with nicely designed figures and tables (I particularly loved figure 8!) - I applaud the authors on these aspects of their manuscript.

While I enjoyed reading this manuscript and feel it is an important contribution to the research community, I found the discussion section was very limited. First, I was looking for more critique on how the empirical upscaling approach has improved on previous methods, including a well-articulated argument for why regional empirical upscaling needs to be considered by the modelling community to ensure regions are correctly represented in global estimates. There is a heavy bias of flux sites and model parameterization from the temperate northern hemisphere, with regions such as Australia, South America and Africa lacking on-ground validation sites to adequately verify global models. What I really like about this study is that it has elegantly shown there are regional differences in Australia not being adequately captured by global models. If this is true for Australia, surely it could be true for other regions less represented too. I'd like to see more thought and critique on this point in the discussion.

In the introduction to our manuscript we argue that models built on global datasets and with a strong northern hemisphere bias may not accurately represent ecosystem dynamics in regions where ecosystem responses do not conform to the dominant dynamics in the global dataset. Perhaps owing to the unusual dynamics of the sclerophyllous, evergreen, woody species that dominate Australia's land mass we believe we have proven this hypothesis to be true in the case of Australia. Whether or not this is true for other under-represented regions such as Africa and South America will likely depend on the extent to which their dominant plant species and land cover types conform to the global datasets. Fortunately for Australia, we were able to test this hypothesis for a number of reasons. Firstly, the OzFlux network is reasonably comprehensive in its coverage of Australian ecosystems. Secondly, the EC sites have been 'harmonised' through the implementation of a single dataset standard (Isaac et al 2017) allowing them to be ingested into a single modelling framework. And lastly, the global upscaling product, FLUXCOM, had not included many sites from Australia so differences between FLUXCOM and AusEFlux are likely attributable to differences in the training data. For Africa and South America a set of high quality, harmonised EC sites covering the diversity of their ecosystems does not appear available, thus it may not be possible to test if these regions are being misrepresented by global models.

To highlight these points we have added a short paragraph to the discussion section that read as follows:

“Our study showed that increasing the diversity of flux tower sites beyond the small Australian set used in global products improved the quality of carbon flux estimates. We cannot predict whether the same might hold for other underrepresented regions, which mostly coincide with the global south, or whether the isolated evolution of Australia's ecosystems also plays a role.”

There's also a lack of discussion of how the limitations in this study could be overcome. For example, the method itself seems sound, but expanding on the points about more EC data being needed would be really helpful to the research community. Do the authors feel that just longer timeseries from the current network are needed, or are more sites required? If more sites, where are they needed? Figure 8 showed some interesting GPP and ER dynamics in certain areas where there is a distinct lack of EC observation sites, the WA Wheatbelt being one of them. The authors point to these areas in the results section, but do not address how these areas could be better understood by future research efforts. A better discussion around these points, in a nationally focused paper like this, would really help researchers on the ground level to make the case for the need to fill these missing gaps.

We agree with the reviewer that some discussion on where future OzFlux sites should be placed is worth adding into the manuscript, and to this end we have edited this discussion into a broader paragraph discussing some of the uncertainties with spatial sampling of OzFlux (the entire paragraph is quoted below).

“A limitation of the OzFlux network is the necessarily limited repeat spatial sampling of all main land cover types. Furthermore, not each bioclimatic region is equally well represented, leading to biases in the sampling. For example, desert and xeric ecosystems cover nearly half of the Australian land mass, but less than 10% of the sites are located in these regions (Beringer et al., 2016). Australia’s expansive cropping ecosystems are also under-represented. The limited representation of these systems in the training data is likely why we found comparatively high uncertainty in these regions (Fig. 6). Further uncertainty in the cropping regions may also be due to the heterogeneity of crop types and agricultural practices that may not be represented in our feature layers, and the potentially large carbon exports as agricultural commodities. Also, given the Australian Government's emphasis on emission offsetting through changes in agricultural practices and human-induced regeneration of native woody vegetation, especially in drier regions (DCEEW, 2023), new EC sites in cropping regions and in the (semi) arid rangelands areas of New South Wales, Queensland, and Western Australia might help reduce uncertainties in AusEFlux and expand the evidential basis for carbon sequestration through (re-)vegetation (Macintosh et al., 2022). Given the changing climate conditions of Australia, it is vital to maintain the current OzFlux infrastructure so that future changes to climate-carbon interactions are monitored at the continental level through iterative retraining of the AusEFlux model as new data is collected.”

Lastly, there's a lack of discussion around future directions for this work. There's momentum building and wider interest in understanding carbon fluxes from landscapes in real time and in collating annual budgets at national scale more frequently. There's an opportunity for the work presented in this manuscript to be incorporated into a regularly produced national annual estimate of Australia's terrestrial carbon accounting, but there's no mention of this in the discussion. I suggest the authors consider adding text along these lines, perhaps pointing to the vision outlined in Papale 2020 and efforts already underway to deliver national carbon observing infrastructure, such as TERN, NEON (USA) and ICOS (EU), that could be input into approaches like the one presented by the authors to help realise these goals.

We agree with the reviewer on this point and thank them for prompting further discussion on the future directions of the work. Please see the comment below in dot point **"Lines 467-469: …"** *for our response to the argument on this work being incorporated into a regularly produced national estimate of Australia's terrestrial carbon budget.*

We have also added further information on future directions for this work at the end of the discussion section. The last paragraph of the discussion now reads:

"While our estimate provides a step-forward in our means for assessing the complex, seasonal, and interannual dynamics of Australia's carbon cycle, future work can improve upon this current effort. Firstly, we aim to extend AusEFlux further back in time through the inclusion of satellite observations from the AVHRR and Landsat missions. However, this effort will inform a separate study as it will require solving cross-sensor calibration issues. A longer record of empirically derived terrestrial carbon fluxes will assist in defining robust environmental baselines from which future changes to the carbon cycle can be assessed. Secondly, new or improved feature layers can be incorporated as they become available (e.g., time-varying estimates of the percentages of trees, grass and bare). And lastly, we aim to explore the prospects of ecological forecasting (Dietze et al., 2018) of the terrestrial carbon cycle as seasonal forecasts may be possible where forecasts of the climate are sufficiently detailed."

There are a few other specific items I feel need to be addressed before the manuscript is ready for publication. I've identified these as follows and believe that if the authors can address them, their manuscript will be more widely cited as a result:

- Lines 37-45: This is quite a difference between the two studies, but then at line 55 it's revealed that the Villalobos et al. 2022 study was from years 2015-2019, while the Friedlingstein et al. 2022 study was from 2003-2021. Looking up both studies reveals these time frames to be accurate. While I completely agree that regionally forced studies usually provide more accurate estimates of carbon cycling in Australia, I think it is misleading not to mention the temporal mismatch between these studies. I suspect the temporal mismatch could be the primary cause of the difference of >50 % between studies, as the millennium drought

(2001-2009) would be captured in the Friedlingstein et al. anaylsis but not in the Villalobos et al analysis. Please amend the text to take this into consideration.

The references for this section has led to some confusion, for which we apologise. Here we are comparing CABLE-POP (extracted from TRENDY v10) with CABLE-BIOS3, provided by Villalobos (2022). Both datasets were clipped to the 2003-2019 range (we incorrectly stated the range as 2003-2021 in the manuscript and have corrected this) for the comparison between long-term mean GPP, thus there is no temporal mismatch. In the Villalobos reference their atmospheric inversion only ran from 2015-2019, but their run of CABLE extends much longer. We have amended the text slightly to make this more clear. We have also included a table in the appendix that outlines the main features of each comparison dataset, which should also help reduce confusion.

- Line 57; Please add spatial resolution for better comparison with OCO-2, i.e. as at lines 54-55.

The Metz (2023) dataset is averaged over the entire Australian TRANSCOM region (including NZ) and is provided as a zonally summarised time-series so there is no spatial resolution given. The initial CO_2 fluxes were resampled to 1 x1 degree grid before applying their TRANSCOM region mask, but it is not clear to us from their supplementary material if the atmospheric inversion results were first predicted on a grid and then summarised, or summarised and then predicted.

- Line 70: I think its important to identify here the unequal representation of EC sites across the globe, as some biomes (i.e. the tropics) contain a limited number of sites compared to the temperate northern hemisphere. This bias is also likely to be affecting ML empirical upscaling approaches. I see the authors allude to this at line 78, but I think it needs to be addressed here too. See Baldocchi et al. 2018 (https://doi.org/10.1016/j.agrformet.2017.05.015) for a good review of inter-annual variability in NEE from sites around the world, and where long-term monitoring sites are lacking.

We agree that the northern hemisphere bias in the Fluxnet dataset is an important attribute to highlight, and that mentioning this earlier in the manuscript during the discussion on the limitations of FLUXCOM is worthwhile. We have included the following sentence in that section:

"The global FLUXNET2015 dataset is also biased to the northern hemisphere, which may preclude global upscaling products from making quality predictions in regions that are both underrepresented in the training data, and do not conform to northern hemisphere climate dynamics."

- Line 77: A better introduction citation for the OzFlux network is Beringer et al. 2022 (https://doi.org/10.1111/gcb.16141) or Beringer et al. 2016 (https://doi.org/10.5194/bg-13-5895-2016). Isaac et al. 2017 is an excellent publication to cite for how the flux data were processed, which should be in the methods.

We have included the Beringer 2016 & 2022 references in this location.

- Line 108: Please add the following text here to clarify how the data were processed "using PyFluxPro vXXX (Isaac et al. 2017),..." The authors may need to check with TERN regarding the PyFluxPro version used.

We argue that providing the reader with the specific PyFluxPro versions used to process the Level-6 data will not be valuable to the reader unless they are already aware of what the software iterations mean. However, we accept the more general point underlying this comment about reproducibility of AusEFlux, and to this end we have included specific reference to the versions of the datasets used. We have also included URL paths to find the data, and Table A1 has been updated to include start and end dates of the datasets used. The beginning of section 2.1.1 now reads:

"We used monthly fluxes of NEE, GPP, and ER produced by the OzFlux (https://ozflux.org.au/) regional network of eddy covariance flux towers. These data are processed to Level 6 and are freely accessible through the Terrestrial Ecosystem Research Network THREDDS portal (https://dap.tern.org.au/thredds/catalog/ecosystem_process/ozflux/catalog.html (TERN, 2023). All site data used in this study were version "2022_v2", and in instances where both "site-pi" and "default" versions of the datasets were available, we utilised the "default" datasets."

- Lines 189-191: Can the authors comment on this more specifically? Are there any biomes or land uses missing that in their opinion would make the analysis more robust? Perhaps this could come in the discussion instead...?

We have included in the discussion section more comments on the limitations of the training data, including where we believe further EC sites would help reduce uncertainties that are derived from the training data.

- Lines 321-325: I agree with this statement, but it should appear in the discussion, not results. Please move to discussion, a good place would be the final discussion paragraph.

We agree and have moved these statements to a new section at the beginning of the discussion section.

- Lines 406-422: This paragraph is mixing results and discussion a bit, i.e. lines 408-410 and lines 415-416, Please consider moving these points to the discussion, which would help beef up the section.

While in general we agree with the reviewer that mixing results and discussion is an issue, in this case we argue that there is merit in including these brief explanations for the results as it assists the reader in understanding discrepancies between products at the point of encounter. As the 'discussion' parts of this section only amount to two *sentences, we argue there is utility in keeping these comments where they currently are in the manuscript because it improves clarity.*

- Line 438: Remind readers of this study here, it's the Villalobos et al. 2022 study, correct? In fact, it would be useful to include a small table that includes information about each of the models used in this study, who published them, their general characteristics (temporal and spatial resolution), etc... That way the authors can refer the reader here to table X for a refresh and avoid re-citing each study, that would add clutter to the text below. The table could be a brief summary of information presented in section 2.1 and included at the end of that section.

We appreciate the suggestion of adding a table to summarise the comparison datasets as there are quite a few products and it does get convoluted at times. We have created a table that includes the dataset name, dataset type (process model, inversion etc.), the spatio-temporal extents, and the key reference for each study. We have added this table (Table A1) to the appendix to avoid cluttering the manuscript with too many tables and figures. We've included a copy of the table below.

Dataset Name	*Dataset type*	*Spatial resolution*	*Temporal range*	*References*
CABLE-POP	*Process-model*	*1°*	*2003-2020*	*Friedlingstein et al. (2022)*
CABLE-BIOS3	*Process-model*	*0.25°*	*2003-2019*	*Villalobos et al. (2022)*
OCO-2 Inversion	*Atmos. inversion*	*0.8°*	*2015-2019*	*Villalobos et al. (2022)*
GOSAT Inversion	*Atmos. inversion*	*-*	*2009-2018*	*Metz et al. (2023)*
FLUXCOM-Met	*ML upscaling*	*0.5°*	*2003-2015*	*Jung et al. (2020)*
FLUXCOM-RS	*ML upscaling*	*0.083°*	*2003-2015*	*Jung et al. (2020)*
MODIS-GPP	*Obs. Based*	*0.01°*	*2003-2021*	*Running et al. (2015)*
GOSIF-GPP	*Obs. Based*	*0.01°*	*2003-2021*	*Li and Xiao (2019)*
DIFFUSE-GPP	*Obs. Based*	*0.01°*	*2003-2021*	*Donohue et al. (2014)*

- Lines 438-442 - this is all one sentence, which is long and rather confusing to follow. Please revise and more clearly articulate to the reader that this study was verified using OzFlux EC sites.

We agree that this sentence was unwieldy and have rephrased the paragraph.

"We found evidence that Australia is, on average, a stronger annual carbon sink than previous CABLE LSM and FLUXCOM estimates have concluded. Our estimate of the long-term annual mean carbon sink over Australia (-0.44 PgC/yr) is higher than those reported by any study besides the regional OCO-2 inversion (-0.47 PgC/yr). We take the consilience between our estimate and the OCO-2 inversion's; the fact that 25 out of the 29 OzFlux EC sites used here report strong annual mean carbon sinks (Figure A7), and the theoretical argument that ML predictions tend to produce good estimates of the mean as evidence that Australia's status as a comparatively strong net carbon sink is robust."

- Line 460: Table shouldn't appear in the discussion.

Agreed, it's been moved back to the results section below section 3.5.

- Lines 467-469: Can the authors expand on this point more? In an ideal world, how frequently do the authors think a product like this should be updated? Realistically, how frequently is this likely to be? I recommend reading Papale 2020 (https://doi.org/10.5194/bg-17-5587-2020) and publications from the global carbon project to tease this discussion point out further.

We are happy to expand on this point further, and it is a future aim of the authors to annually update and release this product. We have included in the discussion section the following paragraph:

"An advantage of this approach over other methods is its computational efficiency, and, owing to the mature architecture of the OzFlux infrastructure, the ability to programmatically ingest updated or new EC datasets to further refine models. Thus, there is an opportunity for AusEFlux to be incorporated into an annually produced national estimate of Australia's terrestrial carbon fluxes. Any annually produced 'bottom-up' estimate of Australia's terrestrial carbon fluxes could also serve as a compliment to the Global Carbon Project's aims of annually reporting the carbon balance of the world (Papale, 2020). The primary challenge with any operational reporting framework is ensuring the necessary feature layers are updated with a similar cadence, so future work will involve identifying reliable and regularly updated data streams to serve this end. Through regular updating of this dataset, the ecosystems that play an outsized role in controlling Australia's mean carbon sink and contribute substantially to its IAV can begin to be systematically monitored for change."

- Lines 477-478: What about the role of fire in consuming biomass in the dry season and how that might affect carbon emissions from savannas? Can the authors expand on this please. Beringer et al. 2015 (https://doi.org/10.1111/gcb.12686) might be a good place to start.

We thank the reviewer for prompting us to discuss the seasonal role of fire. To elucidate the seasonal role of fire in the savanna regions, we have amended Figure A5 to include fire emissions (copied below). The addition of fire emissions shows that the late dry-season (Aug-Oct) fires lead to an earlier net carbon pulse to the atmosphere and larger peak emissions than the out-of-phase ER-GPP effect alone. We have amended the discussion in the manuscript to reflect this.

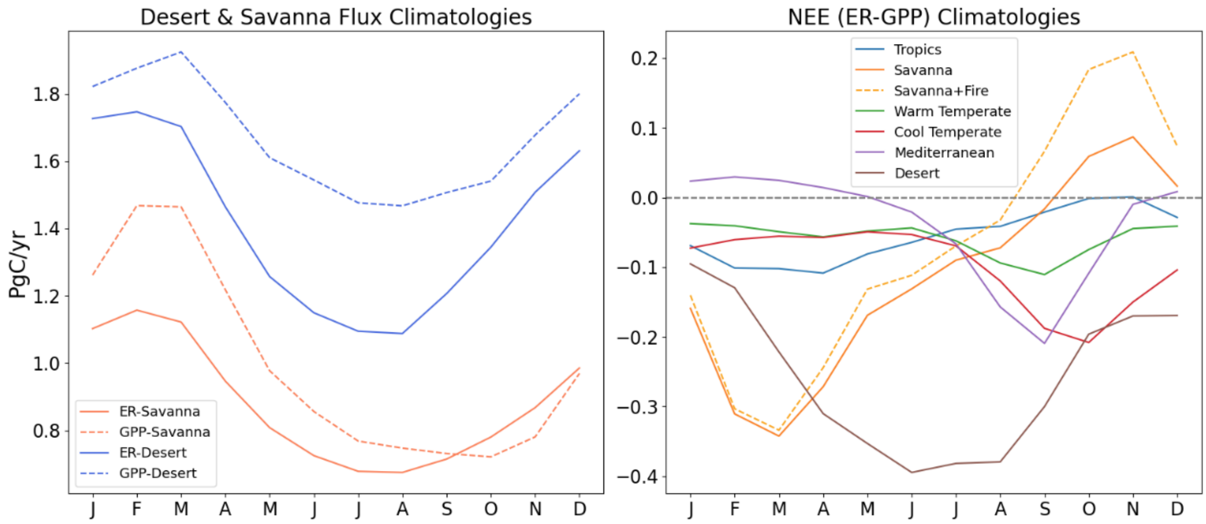

Figure A5. (a) Flux climatologies for the Savanna and Desert region, showing the same results as those in Figure 10, but shown on a single plot to enhance interpretability. (b) NEE per bioclimatic region calculated by subtracting GPP from ER (i.e., not directly modelled), presented here to show how the fluxes interact to produce NEE. Fire emissions from the GFAS product have been added to the Savanna fluxes in (b) to highlight how dry season fires interact with ER to enhance a seasonal pulse of carbon to the atmosphere.

- Lines 478-480: Here again, a missed opportunity to critique with site-based studies, such as Cleverly et al. 2013 (10.1002/jgrg.20101)

We have added reference to the Cleverly et al. (2013) paper in this section, which now reads:

"This finding agrees with Renchon et al. (2018) at the Cumberland Plains EC flux tower site, where the forest was a CO2 sink in winter and a source in summer due to larger seasonal amplitudes in ER. Similarly, Metz et al. (2023) found that seasonal rainfall in semi-arid regions after the dry season drives pulses of heterotrophic respiration that precede the GPP response, leading to net carbon uptake not beginning until March. Cleverly et al. (2013), in a site-based study of a semi-arid acacia woodland in central Australia, observed that the first large springtime storms following the dry season

resulted in pulses of ecosystem respiration owing to an uptick in moisture limited microbial decomposition of photodegraded litter and flushing of CO_2 from soil pore spaces through infiltration. Our results confirm that ER over the savanna region responds quickly to seasonal rainfall events at the end of the dry-season, while GPP responds more slowly resulting in carbon pulses to the atmosphere during the Oct-Dec period."

- Lines 489-492: This is a rather subjective and negative way to begin a conclusion. One could argue that the OzFlux network already captures a diverse range of Australian ecosystems, and it is certainly the largest network in the underrepresented southern hemisphere. However, one could also argue that there are key systems missing, which can bias any upscaling approaches that use OzFlux data. How many flux towers are needed for a network like OzFlux to have "good" coverage? My point being, given this paper did not assess whether the quantity of sites in OzFlux was adequate for upscaling (in fact it used correlation with OzFlux sites as an indicator that the results were robust), I suggest rethinking the opening sentence of the conclusion to be more focused on the key result/finding and less about the limitations of OzFlux.

We thank the reviewer for this suggestion and agree that the opening sentence was needlessly negative. We have deleted the opening line of the conclusion and the conclusion now has a better focus on the key findings.