Review of the Paper: "Empirical upscaling of OzFlux eddy covariance for high-resolution monitoring of terrestrial carbon uptake in Australia"

The paper develops high-resolution estimates of GPP, ER, and NEE in Australia using empirical upscaling of flux tower measurements. Comparisons with other products show regional empirical upscaling outperforms global upscaling and process-based models. Rainfall deficits and surpluses drive NEE anomalies, with GPP responding more than ER. The paper introduces "AusEFlux" as a benchmark for high-resolution monitoring of Australia's carbon cycle.

Upon careful evaluation and analysis of the manuscript, it is evident that a major revision is necessary in light of the following critical comments. Addressing these concerns will enhance the overall quality and impact of the paper, ensuring its suitability for publication in our esteemed journal.

- **The paper highlights the performance of regional empirical upscaling in improving global upscaling products and outperforming existing LSMs, but could you provide more insight into the specific limitations of Australia's comparatively sparse network of EC towers and their potential impact on the accuracy of the derived estimates?**

  We thank the reviewer for the opportunity to provide some more detail on the impact of Australia's sparse EC network on the uncertainty of the AusEFlux estimates. We have edited the paragraph below into the Discussion section of the manuscript.

  The principal limitation of the OzFlux EC network is its relatively limited spatial sampling of all the landcover types in Australia. Furthermore, each bioclimatic region is not equally represented, leading to biases in the sampling. For example, desert and xeric ecosystems cover nearly half of the Australian land mass, yet less than 10 % of the network is in these regions (Beringer et al. 2016). Agricultural cropping ecosystems are also under-represented (this is probably why the uncertainty plots of Figure 6 show the greatest variance between predictions is in these regions). Owing to the limitations of the OzFlux network in spatial representation of ecosystems, it is a challenge to confidently claim that cross-validation alone provides an accurate estimate of terrestrial carbon fluxes. As such, we rely heavily on an intercomparison between products as being more indicative (albeit qualitatively) of spatial uncertainty, as we believe the convergence of results from multiple, independent lines of evidence tells us more about the true nature of Australia's terrestrial carbon cycle than any given cross-validation method. We are encouraged by the convergence of our results with the GPP estimates from MODIS, GOSIF, and CABLE-BIOS as each of these datasets applies a different method to quantify GPP. Ecosystem respiration (ER) is harder to effectively validate as only FLUXCOM (similar method to ours) and CABLE provide estimates of ER. Net ecosystem exchange (NEE) offers the prospect of independent validation as

the satellite assimilated atmospheric inversions are a wholly independent measurement of NEE (though they still contain significant uncertainties owing to the uncertainties in the satellite CO_2 measurements themselves, along with the atmospheric transport model used). This is why we include the two most recent regional-scale inversions in our intercomparisons. Although mean NEE varied between our estimate and those of the two atmospheric inversions, anomalies and the seasonal cycle show better agreement than other methods. We take as evidence that our empirical upscaling of the OzFlux network provides a better estimate of Australia's terrestrial carbon cycle than the global empirical upscaling product, FLUXCOM, which to-date has been the only product available of its type for Australia.

- **How were datasets resampled to monthly resolution and reprojected to 1*1 km in section "2.1.2 Gridded explanatory variables"? what was the raw data specifications?**

  Datasets such as MODIS LST, kNDVI, and NDWI were aggregated (simple averaging) to monthly scale using the mean of all available observations within the given month. Spatial reprojection onto a 1 km x1 km geographic grid (EPSG:4326) using either 'bilinear' or 'averaging' resampling methods depending on the native resolution of the product (bilinear for LST, averaging for the others). We have updated the text with this information in section 2.1.2. These datasets were all downloaded from Google Earth Engine, as described in Table 1. The specific MODIS products used were the MCD43A4 v6.1, and MOD11A1 v6.1

  The climate datasets (from ANUClimate) are provided at 1-km spatial resolution and monthly temporal resolution so no resampling or reprojection was required.

  The static variables such as landcover fractions and vegetation height were resampled from their native resolutions using the 'average' of all pixels within a 1 km grid cell. The vegetation height product is provided at native 25-m resolution, and the landcover fractions are provided at 250-m resolution. This information has also been added to the manuscript in section 2.1.2.

- **Provide additional details or references to explain the 'SOLO' data version used for partitioning NEE into GPP and ER. This will aid readers in understanding the specific data processing steps and methods employed.**

  We understand and agree with the reviewer that the SOLO flux partitioning methods may be of interest to some readers. However, we do not believe the details of the methods are within scope of the study because as: (a) the cited work of Isaac et al. (2017) provided the full detail necessary and is available via the references; and (b) adding a description of the methods within this paper would add a lengthy section of text. Nevertheless, we provide a link to the datasets, which are freely available for download through the Terrestrial Ecosystem Research Network (TERN). And we have added to the text within section 2.1.1 a note that a full description of the partitioning method is available from the stated reference. Within the same paragraph we have also

provided links to both the OzFlux website, and the TERN website where datasets can be downloaded.

- **In Section 2.1.3, it is mentioned that the MODIS-GPP and DIFFUSE-GPP products were resampled to a 1 km resolution to match the resolutions of the ML upscaling product. Could you please provide more details regarding the specific method used for resampling these datasets? It would be beneficial to understand the resampling technique employed to ensure compatibility between different resolutions. Additionally, any information regarding potential implications or limitations of the resampling process would be valuable.**

  Both MODIS- and DIFFUSE-GPP datasets were resampled to 1-km resolution using the average of all pixels within the 1 km cell. MODIS-GPP is provided at a native 500-m, and DIFFUSE-GPP is provided at 250-m resolution. This information is provided in the text at lines 170-171. We do not believe there to be any significant implications of using cell averaging from the moderate resolution (250- and 500-m) to 1-km. Such resampling is common practice in remote sensing.

- **Elaborate on the resampling and reprojection of gridded explanatory variables. Specify the resampling resolution and provide a rationale for selecting a common 1-km x 1-km geographic grid. Discuss potential errors or limitations associated with spatial resampling and its impact on the accuracy or comparability of the datasets.**

  We understand the reviewer's point here and refer them to the previous comment above, as some information in response has been given. The rationale for using a 1 km grid is twofold: one, (and the primary reason) is it matches the coarsest native resolution explanatory variables, namely the climate datasets; and two, because the footprint of a flux tower is on the order of 1 x 1 km (very approximately), so it makes sense to extract training data at resolutions comparable to the flux footprint. We have added this short rationale to the 2.1.2 section.

- **In Section 2.1.3.3, it is mentioned that the regional inverse modeling product by Villalobos et al. (2022) provides a spatial resolution of approximately 81 km. Could you please provide details on how the other datasets with different resolutions were processed and plotted to ensure compatibility for comparison? Specifically, how were the ML results, MODIS-GPP and DIFFUSE-GPP products, which were resampled to 1 km resolution, handled in the analysis?**

  The ML results are predicted at 1 km spatial resolution, and monthly temporal resolution so were not post-processed (except for some masking of cities). The MODIS and DIFFUSE GPP products were resampled as per the text in section 2.1.3.4. and no further post-processing was done. For the scatter plots of Figure A3, a simple

extraction of the pixels over the EC tower locations was done for comparison of the different GPP products against flux tower estimates.

For the time-series plots in figures 9, the datasets are all converted to PgC/year from their respective native units (usually gC/m^2/month). This involved converting the datasets from their native grid to an equal-area grid of the same spatial resolution (we used EPSG:3577), then multiplying every pixel by its area, along with the conversion factor from grams to petagrams to arrive at PgC/year. This step is documented in the notebook 7_Compare_products.ipynb available on the github repository linked in the assets for this paper. This processing step is common in the literature, and since some effort has gone into creating well documented Jupyter notebooks, we feel there isn't a pressing need to add these steps to the methods section (at the end of section 2.2.2 we have added an explicit reference to the Jupyter Notebooks which describe all processing and analysis steps). After conversion of all datasets to PgC, we summed across all pixels at each time-step to produce 'zonal' time-series. We consider this step to be common enough to not require specific description in the text. The time-series are then smoothed using a three-month rolling mean, which is stated in the caption. For figure 10, seasonal climatologies are created by grouping common months together (i.e. all the Jan., all the Feb. etc.) in a time-series, and then finding the long-term average for each month. The result is then summed across the continent at each monthly time-step to produce a 'zonal' average seasonal cycle. Again, we argue these processing steps are standard practice and do not require a specific description in the text.

To assist the reader in understanding if AusEFlux performs better because of the finer spatial resolution, or because it is intrinsically better even when coarsened to the scale of competing products, we have amended Figure A3 in the manuscript. All products in the inter-comparison scatter-plots have now been reprojected to match the resolution of CABLE-BIOS3 (~25km). The figure as it is shown in the manuscript is reproduced below for convenience.

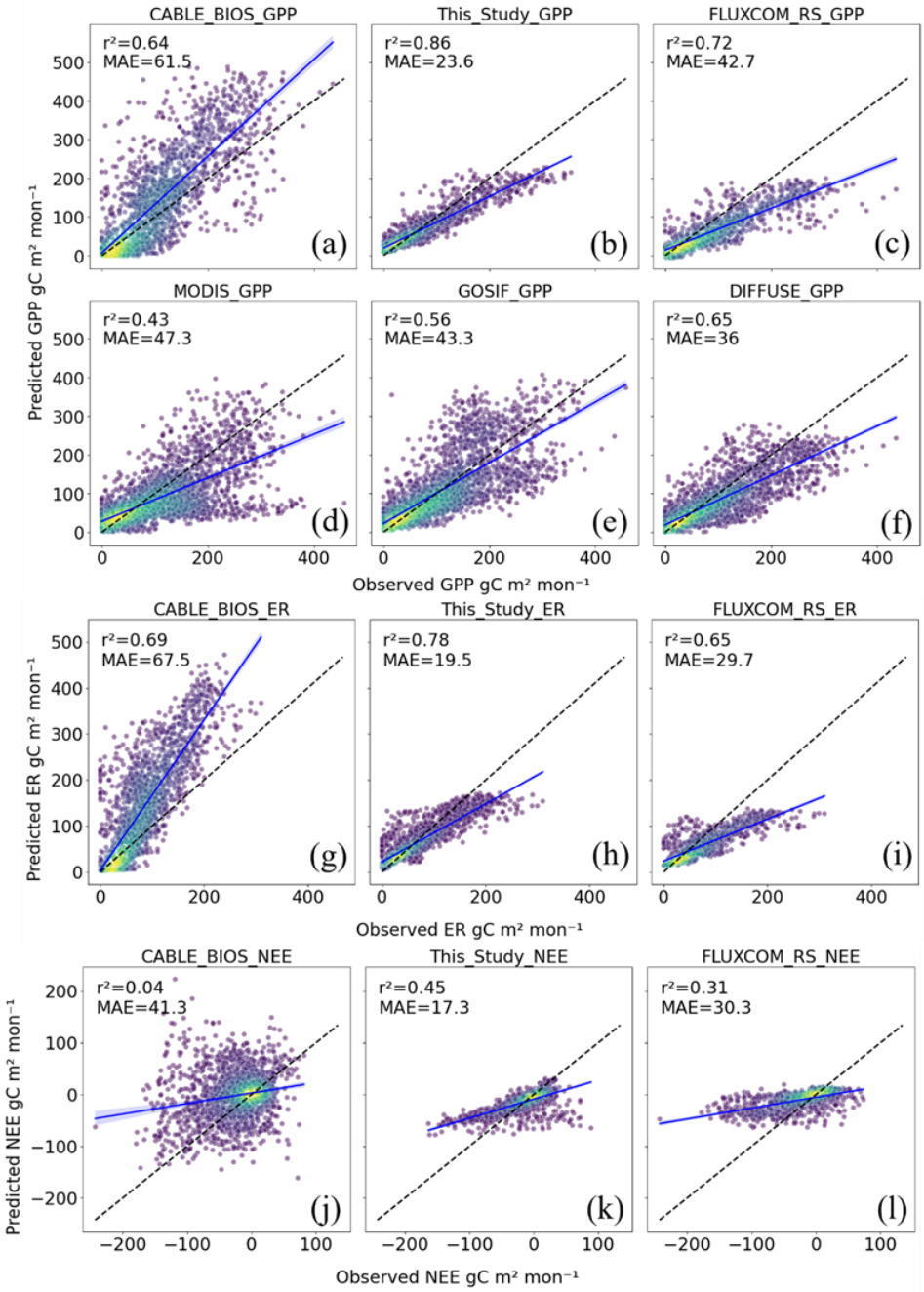

Figure A3. Scatter plots of modelled vs EC flux tower monthly carbon fluxes for a suite of products. The EC tower flux values are compared with the nearest pixel in each product, and the products have been reprojected to match the resolution of CABLE-BIOS3 (~25 km). Only those products with a reasonably high spatial resolution have been compared with the flux tower (i.e.,CABLE-POP, FLUXCOM-Met, and the OCO-2 Inversion have been excluded from these plots).

- **Specify a specific website or source where readers can access the CO2 flux tower data used in the study. This will facilitate replication and further exploration of the data.**

We thank the reviewer for noting this and have added to the text at line 109. It now reads: "These data are processed to Level 6 and are freely accessible through the Terrestrial Ecosystem Research Network (https://portal.tern.org.au/)"

- **Provide more details on the implementation of random forest regression and gradient-boosting decision tree algorithms, including parameter settings, and more importantly elaborate on how predictions from the ensemble of random forest and GBDT models are combined or weighted as an ensemble learning.**

  We appreciate the reviewer's suggestion of providing more information on the ML methods. As such we have provided sufficient detail in the appendix and via the python notebooks should a researcher wish to know more about the hyperparameters used or to reproduce the results. This information has been added to the text in section 2.2.1. We also note that this information is provided in the notebook 3_Generate_ensemble_of_models.ipynb contained on the github page for this study.

  However, as *we fit 30 distinct models per flux, for a total of 90 unique models, it would be impractical to provide hyperparameter settings for each model configuration in the main text. Instead, in the appendix we have provided the parameter grids over which a random grid search was conducted during hyperparameter optimization (Table A1). In section 2.2.1 we provide details on how the model ensembles were combined. A simple per pixel median is conducted across the 30 gridded predictions for a given flux. The interquartile range (25^{th} and 75^{th} percentile) are taken as the uncertainty envelope. There are no weightings, each model and gridded prediction is conducted independently and then combined through the calculations of medians/percentiles. We argue this is quite clearly outlined in section 2.2.1. The notebook 5_Combine_ensembles.ipynb provides the documented code on how this was run.*

- **Clarify the rationale and details behind the iterative training procedure with randomly selected EC sites for uncertainty estimation. How can you ensure that all sites were removed in the 30 repeats? Provide details on how the randomness is controlled to achieve this objective.**

  The reason we selected two sites to remove per iteration was because it balanced the need to significantly alter the training dataset per iteration, while not overly degrading the quality of the model by removing too much data. As some of the site's time-series are relatively short, removing only one site could result in removing only 20-30 samples from the training data (from a total of ~2800), and thus the model may only be marginally different from the full-dataset model. Removing more than two sites could result in some iterations where so much data is removed from the training dataset that the quality of the predictions is severely degraded. We have added a statement to section 2.2.1 to clarify this point.

It was not our objective to ensure all sites were removed during iteration of the models. Rather, we elected for a uniform random approach where sampling two sites, fifteen times, was merely likely to remove every site. To ensure every possible combination of sites is removed would require 29^2=841 permutations per flux, which would be impractically time consuming to run, and would be unlikely to tell us much more about the uncertainty than the approach already described. We concede that fifteen iterations are an arbitrary number, but it provides a balance between allowing for a reasonable chance for all sites being removed once, while also keeping the computation time to within reason. The other important consideration is the difficulty in calculating per-pixel percentiles across more than 30 predictions, as each gridded prediction is equivalent to 12.5GiB of data (so already we are summarising close to 400 GiB of data to calculate the ensemble medians).

- **Provide a detailed description of the data split methodology used in the nested, time-series-split cross-validation approach. Did you consider the aspect of time when splitting and testing the methods? (e.g. did you allocate 5 years for training and 1 or 2 years for testing?)**

  We understand the reviewer's suggestion of providing more information on the temporal cross-validation methods. The time-series split method blocks the testing samples by continuous lengths of time. The exact length of time tested depends on the length of the overall time-series as we allocated 20 % of a timeseries to testing, and 80 % to training. For example, if a dataset is 10 years long, then 8 years is used for training, while a two-year continuous block is used for testing. As we conducted five-fold cross-validation, this procedure was repeated five times and at each iteration the two-year testing 'block' is moved forward in time, such that over the 5-folds, the entire time-series is tested. We have added this example to the text in section 2.2.2. to clarify the method. Also, note that every flux tower record is included in a k-fold, so 20 % of every flux record is tested per fold.

  The 'nested' part of the CV procedure refers to using a separate, internal split on an outer k-fold to conduct hyperparameter optimization. Using a nested approach to CV prevents testing on the same data used to tune model parameters, and thus prevents creating overly optimistic CV scores. We have included in the text of section 2.2.2 a sentence discussing this, along with a reference that outlines the benefits of using a nested approach.

  Overall, we have provided a detailed description of the cross-validation procedure, including figure 2 which presents a schematic of the procedure.

  We would like to also note that the primary focus of this paper is not on providing novel methods for cross-validation (CV), nor on exploring/testing various approaches to ML upscaling. Rather, we have implemented common methodological procedures for empirical upscaling to produce a higher-quality estimate of Australia's terrestrial

carbon cycle than already exists (taking advantage of the expansion in the OzFlux tower network and regional feature layers), so that it can be considered alongside other approaches to quantifying the carbon cycle. This is why most of the discussion in the paper is devoted to the intercomparison between products as we feel consiliences between lines of evidence are more important than the results of any given CV method.

- **Consider incorporating any additional limitations or uncertainties associated with the data sources, processing steps, or comparison datasets. This will provide a more comprehensive understanding of the potential impacts on the study's results and conclusions.**

  We agree with the reviewer that all the products used as explanatory variables in the model are subject to uncertainties. However, the datasets employed by this study are widely used and accepted in the literature. The remote sensing explanatory variables (either MODIS or MODIS-derived) have been widely used at continental scales and their errors have been well documented elsewhere. The interpolated climate data is subject to uncertainties due to the distribution of the measurement network, which in Australia is skewed towards the coast. However, ANUClimate is a well-regarded dataset (Hutchinson et al. 2004, 2014 & 2015) and the density of weather records over Australia is very good during the modern era.

  We consider describing the uncertainties associated with the inter-comparison datasets as beyond the remit of this paper as it would take considerable time and effort to summarise uncertainties from nine other products. We argue the interested reader can follow-up with the citations provided.

- **Why was the model not tested on individual sites after training? It is crucial to determine whether the model can perform effectively at a single location.**

  We agree with the reviewer that it is important to assess the model's ability to predict at each flux tower location. However, we believe the manuscript has amply described the model's performance in this respect.

  Figure A2 compares the predicted seasonal cycles with the observed seasonal cycle for every site in the training dataset (only for NEE as NEE seasonality was a key focus of the paper).

  In addition to the cross-validation plots of figure 3a-c, we have also included scatter plots of the predicted annual means (Figure 3d-f) and the observed annual means from every site. The colour coding does not show individual sites because distinguishing between 29 unique colours is very difficult, instead we grouped them by bioclimatic regions to make the analysis more legible.

  We do not see a need to include a figure showing the full time-series predictions of every site as it would make for a very large and unwieldy figure and we feel would not

provide the reader with any more useful information than the results shown in Figure 3, and Figure A2.

References

Beringer, J., Hutley, L. B., McHugh, I., Arndt, S. K., Campbell, D., Cleugh, H. A., ... & Wardlaw, T. (2016). An introduction to the Australian and New Zealand flux tower network–OzFlux. *Biogeosciences*, *13*(21), 5895-5916.

Hutchinson, M. F., & Xu, T. (2004). ANUSPLIN version 4.4 user guide. *Centre for Resource and Environmental Studies, The Australian National University, Canberra*, *54*.

Hutchinson, M., & Xu, T. (2014). Methodology for generating Australia-wide surfaces and associated grids for monthly mean daily maximum and minimum temperature, rainfall, pan evaporation and solar radiation for the periods 1990–2009, 2020–2039 and 2060–2079. *NARCliM Report to the NSW Office of Environment and Heritage*.

Hutchinson, M. F., Kesteven, J. L., Xu, T., Evans, B. J., Togashi, H. F., & Stein, J. L. (2015, December). Fine Scale ANUClimate Data for Ecosystem Modeling and Assessment of Plant Functional Types. In *AGU Fall Meeting Abstracts* (Vol. 2015, pp. B43G-0631).

Isaac, P., Cleverly, J., McHugh, I., Van Gorsel, E., Ewenz, C., & Beringer, J. (2017). OzFlux data: network integration from collection to curation. *Biogeosciences*, *14*(12), 2903-2928.

Villalobos, Y., Rayner, P. J., Silver, J. D., Thomas, S., Haverd, V., Knauer, J., ... & Pollard, D. F. (2022). Interannual variability in the Australian carbon cycle over 2015–2019, based on assimilation of Orbiting Carbon Observatory-2 (OCO-2) satellite data. *Atmospheric Chemistry and Physics*, *22*(13), 8897-893