Response Letter

CEDAR-GPP: spatiotemporally upscaled estimates of gross primary productivity incorporating CO_2 fertilization

Referee #1

Dear reviewer,

We are grateful for your thorough and constructive feedback, which has helped us improve our manuscript. We have carefully addressed all concerns, including clarifying our methodology, providing additional context for the limitations of our dataset particularly in tropical regions in the revised manuscript. Below we provide our point-to-point response to reviewers' comments, with revisions highlighted in red. For revisions, we provide line numbers from the track-change revised manuscript.

[Reviewer Comment 1] The authors just added some results to make their results more convincing. However, I have pointed out the dataset itself is not robustness since the GPP trend at tropics is highly uncertain. Moreover, they didn't revise the dataset itself since the first time of submission. As a reviewer, my duty is to find out the potential shortcoming of this dataset and make this dataset reliable to the public. Till now, the dataset is not fully convincing. I am not doubting the innovation for this dataset, but the current version of dataset is not convincing.

Response: Thank you for your feedback. We understand your desire to ensure the dataset is as robust as possible, but respectfully disagree with the point that our dataset is "not convincing" due to high uncertainties in tropical GPP. Quantifying GPP trends, particularly in tropics, is a known challenge across the field. Current state-of-the-art datasets, such as FLUXCOM, do not capture these trends from site to global levels, partly because they overlook CO_2 fertilization. By incorporating this effect, our models (CFE-ML and CFE-Hybrid) significantly improved trend estimates, compared to the Baseline (which represents state-of-the-art approaches). This improvement demonstrates the robustness and significance of our work and how it advances over currently available approaches.

We acknowledge that estimation of GPP trends in tropics remains highly uncertain, a limit shared by all upscaling studies due to limitations in eddy covariance data availability and remote sensing data quality in tropical regions. We have quantified these uncertainties (Fig. 12) to ensure transparency. Additionally, we have expanded our discussion in the revised manuscript to emphasize the limitations and challenges

associated with validation. This aspect is highlighted in the abstract and conclusion (see text provided below).

We want to emphasize that the reviewer is correct, in that GPP trends in tropical regions are highly uncertain (as our results clearly show (Fig. 12)), but that this is an issue that affects our entire field, not just our paper, and the improvements we introduce greatly advance over previous efforts regardless.

Revisions: Line 857 – 865: Finally, direct validation of GPP trends is limited, particularly in tropical regions, constrained by the availability of long-term records. Detecting and evaluating trends is challenging and typically requires long monitoring records (e.g. over 10 to 15 years), since long-term changes, such as those induced by CO_2, are very small relative to large interannual variations. Evaluating aggregated GPP trends across multiple sites presents an alternative approach; however, there were still insufficient sites in tropical and evergreen broadleaf areas to robustly validate our estimates for those ecosystems (Figure 5). Partly due to data limitation, uncertainties in GPP estimated from bootstrapped samples are very high in tropical areas (Figure 12). Thus, trend estimates in these areas should be interpreted in the context of associated uncertainties and limitations.

Line 37 – 39: Estimating and validating GPP trends in data-scarce regions, such as the tropics, remains challenging, underscoring the importance of ongoing ground-based monitoring and advancements in modeling techniques.

Line 955 – 958: However, trend estimation and validation remain particularly challenging in data-scarce regions, such as the tropics, emphasizing the need for enhanced data availability and methodological advancements.

[Reviewer Comment 2] The CO2 fertilization effect indeed affects the global trend of GPP, but current dataset just explain 51% of the global trend, which is quite low.

Response: This comment appears to misinterpret our statement in the abstract: "*Incorporation of the direct CO_2 effects substantially enhanced the predicted long-term trend in GPP across global flux towers by up to 51%, aligning much closer to a strong positive trend from eddy covariance data*"

Our finding indicates that the CO_2 effects improved the model-predicted GPP trends at eddy covariance sites, when compared to models without considering CO_2. We have revised this sentence to improve clarity and avoid confusion.

Revisions: Line 28 – 32: After incorporating the direct CO_2 effects, predicted long-term GPP trend across global flux towers substantially increased from 3.1 $gCm^{-2}year^{-1}$ to 4.5 – 5.4 $gCm^{-2}year^{-1}$, which aligns more closely with the 7.7 $gCm^{-2}year^{-1}$ trend detected from eddy covariance data.

[Reviewer Comment 3] More importantly, the GPP trend in CEDAR-GPP at tropics is not convincing (based on the limited results in the revised manuscript), so I cannot recommend this study for publication. I need to remind the authors should revise and answer the questions point to point, but not omit some key questions in the last round of revision. Therefore, a rejection but an invitation for resubmission is appropriate.

Response: Thank you for your feedback. We agree that estimating and validating GPP trends in tropics remain a major challenge in the field, due to data availability. However, we conducted thorough validation in regions with sufficient data, demonstrating the robustness of our dataset (Figure 5b,c). Additionally, we provided detailed uncertainty quantification, transparently outlining the limitations of our estimates (Figure 12). In the revised manuscript, we have expanded our discussion to emphasize these limitations.

[Reviewer Comment 4] 1 Validation 1.1 '*Moreover, intercomparisons with other RSbased datasets, including FLUXCOM, FLUXSAT, and MODIS, confirm strong consistency inglobal patterns of annual mean GPP, interannual variability, and mean seasonal cycles*' .
Comment: I think this part of result is convincing, no need to clarify it again.

Response: Thanks for the feedback.

[Reviewer Comment 5] 1.2' *Our estimates of long-term trends agree with eddy covariance data across global sites. Notably,the Baseline model, which does not consider the direct CFE, underestimates GPP trends at flux towers. In contrast, the CFE-ML and CFE-Hybrid models show significant improvements, underscoring the need to consider both direct and indirect CFE. In our last revision, we performed additional validation of GPP trend by climate zones and plant functional types (Figure 5b, 5c). We also provided comparisons of estimated and observed trends at long-term sites (Figure S4).*'
Comment: I think this validation is weak, since in Figure 5b,c, the author still not show the results at tropics.

Response: Thanks for the feedback! Tropics were not included because there were insufficient sites to support a reliable analysis, as noted in the manuscript (line 354 – 355). Specifically, at least five sites with five years of continuous observations (with no gaps more than one month per year) were required. For upscaling purposes, we applied strict quality control, discarding records with more than 20% missing or low-quality gap-filled data – a standard procedure in upscaling studies - which further limited data availability.
In the revised manuscript, we expanded our discussion to emphasize the limitations of our dataset in tropical and data scarce conditions, and we noted that the estimated uncertainty could be used as a reference of data quality.
Line 354 – 355: Categories with less than six long-term sites available were excluded from the analysis, which includes

355 EBF and Tropics.

In figure S4, what I can see is, half of the CFE-hybrid and CFE-ML cannot capture the trend in EC sites. So this validation is not convincing.

Response: We respectfully disagree with the statement that our "validation is not convincing". CEDAR-GPP greatly improved the quantification of trends when compared to existing state-of-the-art approaches. We established the Baseline model as a reference to the current methods ensuring a fair comparison. In Figure S4 (now S7), the CFE-Hybrid and CFE-ML models showed significantly better performance than the Baseline (current state-of-the-art) in most sites, demonstrating the effectiveness of our methods.

Note that directly benchmarking existing datasets (such as FLUXCOM) against eddy covariance data is not an apples-to-apples comparison, as the datasets were not developed using the same sites. To ensure fairness, we established the Baseline model to represent current methods used by datasets like FLUXCOM and FluxSat.

We agree that substantial differences exist between estimated (even with CO_2 effects) and EC-based GPP trends in Figure S4. It is important to consider factors other than CO_2 that can cause long-term GPP changes, such as nitrogen deposition, forest aging, succession, changes in surface roughness, or natural and manmade disturbances. These factors may be underrepresented in our models, contributing to the underestimation of trends. Robustly reconstructing trends in individual sites across the globe remains challenging, given the current limitations in eddy covariance and remote sensing inputs. Despite the challenges, our incorporation of the CO2 effects marks a significant improvement over current approaches and a meaningful advancement to the field.

In the revised manuscript, we have expanded our discussion on the limitations of our datasets in representing CO_2 fertilization and trends and highlighted this aspect in the abstract and conclusion. We have introduced the following changes to highlight these points:

Revisions: Line 830 – 850: Several limitations should be noted regarding GPP trend estimation and validation. First, the CFE-ML model may not fully capture the intricate mechanisms of plant physiological responses to CO_2. For example, eddy covariance towers, especially long-term sites, are typically located in homogeneous and undisturbed ecosystems, not representative of the full diversity of ecosystems globally. Thus, interactions between CO_2 and natural or human-induced disturbance, as well as many other stresses, are likely underrepresented in the models. Ultimately, the model's capacity to robustly quantify CO_2 fertilization is constrained by the scope and diversity of the eddy covariance data. Additionally, the use of spatially invariant CO_2 data may not fully represent the actual CO_2 variations that plants experience across different environments.

Secondly, CO_2 effects inferred by the CFE-ML models may be confounded by other factors that correlate with CO_2 over time. Industrialization-induced nitrogen

deposition could synergistically boost GPP alongside CO_2 (O'Sullivan et al., 2019). Technological and management improvements in agriculture that contribute to a global enhancement of crop photosynthesis (Zeng et al., 2014), might also be indirectly reflected in the model estimates. Moreover, interactions with the other input features that exhibit long-term trends, such as those induced by non-biological factors (e.g. sensor orbital drifts), also affect the CO_2 effects inference. Additionally, other factors that could lead to long-term GPP trends (e.g. forest aging, disturbances) might also be underrepresented in our models.

Line 37 – 39: However, estimating and validating GPP trends in data-scarce regions, such as the tropics, remains challenging, underscoring the importance of ongoing ground-based monitoring and advancements in modeling techniques.

Line 955 – 958: However, trend estimation and validation remain particularly challenging in data-scarce regions, such as the tropics, emphasizing the need for enhanced data availability and methodological advancements.

[Reviewer Comment 6] '*we show that CEDARGPP exhibits higher consistency with TRENDY models after incorporating the direct CFE*'

Comment: The TRENDY models cannot be the validation! The TRENDY models are results from simulations but not the ground measurement. So this could just be cross-validation at the model world but the observations at the real world.

Response: We fully agree with the reviewer that the TRENDY model comparison should not be considered validation. The TRENDY intercomparison was used to illustrate that previous inconsistencies between satellite-based GPP and TRENDY may be induced by an omission of the direct CO_2 effect in satellite estimates.

[Reviewer Comment 7] I think the authors didn't pay enough attention to my comments. Table S1 should provide the validated and training year for each flux site as I mentioned in the first time.

Response: We appreciate the feedback but noted a misinterpretation of our approach. We split data by sites rather than years. In our five-fold cross-validation, each site (including all data years) was randomly assigned to one of five groups, with each iteration training on four groups and testing on one. This scheme assesses model performance on unseen sites, which is more applicable for upscaling. Splitting sites by years would risk overfitting and deflating error metrics.

In the revised manuscript, we have added each site's time span and IGBP class to Table S1 and relocated it to the Appendix, so individual site citations can be integrated to the main text.

[Reviewer Comment 8] To my knowledge, there are at least 7 EC sites with long term GPP observation are at tropics:

(1) AU-Rob
https://ozflux.org.au/monitoringsites/robsoncreek/index.html
(2) GF-Guy
https://fluxnet.org/sites/siteinfo/GF-Guy
(3) BR-Sa1
https://ameriflux.lbl.gov/sites/siteinfo/BR-Sa1
(4) AU-ASM
https://ozflux.org.au/monitoringsites/alicesprings/index.html
(5) AU-LIF
https://ozflux.org.au/monitoringsites/litchfield/index.html
(6) CN-Xishuangbanna
https://www.scidb.cn/en/detail?dataSetId=db9bf2dde00746f7a40cfc3dbad324b2
(7) CG-YPS
https://www.congo-biogeochem.com/congoflux
All of these sites contain more than 8 years of observation.

Response: Thank you for sharing the list of EC sites with long-term observations. While increased data availability could benefit model robustness, adding a few sites is likely still insufficient for comprehensive inference and validation. As we noted earlier, evaluating CO_2 effects and benchmarking trends in a single site can be problematic, due to the impacts of other factors that may not be fully represented in the models.

We have reviewed these data sources and found that they have limited years of publicly available data insufficient for robust evaluation (e.g. AU-Rob has only 2014 available in FLUXNET2015; CN-Xishuangbana, only has data from 2011 to 2015). Some sites were included in our analysis but lacked long records due to quality issues (e.g. GF-Guy and BR-Sa1).

Given that adding a few sites would not significantly alter our results, we did not include them in this manuscript. But we are committed to continue improving our methods and datasets as more data becomes available, such as the upcoming FLUXNET dataset.

[Reviewer Comment 9] 2 Model
As the second reviewer also concerned using a constant χ value for the generation of long-term GPP. So if my understanding is right, the constant χ value can only reproduce the spatial difference of different PFT, but it cannot reproduce the trend of VPD and soil moisture control to χ.
Following the assumption of FvCB model, a constant χ value indicated that the GPP is driven by atmospheric CO2 concertation and air temperature but not related to the water

stress such as VPD and soil moisture control. So this also lead to the CEDAR-GPP is leak of robustness.

Response: There appears to be a misinterpretation of our methods here. All our models account for GPP responses to VPD, soil moisture, and other environmental factors. The optimality theory that the reviewer refers to, which assumes a constant χ value, only simulates the response of LUE/GPP to CO_2 concentration. The machine learning models were designed to capture the effects of all other factors on χ/LUE/GPP. The CFE-Hybrid model combines the two components (machine learning and optimality theory) to fully represent GPP dynamics.

Our choice of a constant χ value is based on rigorous sensitivity analysis at eddy covariance towers, where we found that the CO_2 fertilization effects were highly consistent between the constant and dynamic χ approaches with $r^2 > 0.99$ (Figure S14). Given that the dynamic approach did not result in meaningful differences, and that quantification of VPD effects on stomatal conductance remains an active research area with high uncertainties, we decide to use the constant method for the main results. We therefore respectfully disagree with the reviewer's statements here, and hope these clarifications highlight why they are a misinterpretation of our methods and results.

[Reviewer Comment 10] Once again, it is not doubt that adding the direct CO2 fertilization effect to machine learning based GPP modelling is important. But I think the high uncertainties in the trend of GPP especially at tropics make this dataset untrusted. The seasonal trend, IAV, and mean spatial distribution of annual GPP are good, but the GPP trend is still not convincing and it is the most important outcome for direct CFE.

Providing a robust dataset for the science community is important, since sometime the wrong results caused the misleading to climate change feedback evaluation.

Response: We thank the reviewer for highlighting the importance of incorporating direct CO_2 effect and agree on the need for robust datasets. However, we respectfully disagree on the statement that "the high uncertainties in the trend of GPP especially at tropics make this dataset untrusted."

We have demonstrated significant improvements in trend estimates in eddy covariance sites against existing state-of-the-art datasets, particularly in arid, temperate, and cold regions. We acknowledge that estimation and validation of trends in tropics remain highly uncertain due to data limitation. To ensure transparency, we provided detailed uncertainty quantification and emphasized in the manuscript that our datasets should be used and interpreted in the context of uncertainties and limitations.

We concur that robust estimation and uncertainty quantification are important for accurate assessment of ecosystem changes and climate feedback. To further emphasize this, we have added a clarifying note in the revised manuscript.

Revisions: Line 929 – 936: Finally, **like other upscaled or remote sensing-based GPP datasets, CEDAR-GPP should not be regarded as "observations" but rather as model estimates informed by remote sensing and ground-based data**. The extent of assumptions or structural constraints varies across such datasets. CEDAR-GPP, particularly in its CFE-Baseline and CFE-ML configurations, is entirely data-driven and incorporates no explicit assumptions regarding the biological and environmental processes underlying photosynthesis, apart from the generic assumptions inherent in machine learning models. **Consequently, the usage and interpretation of this dataset should be carefully framed within the context of the input eddy covariance and environmental data as well as their limitations.**

Referee #4

Dear Dr. Nelson,

Thank you for your thorough evaluation and thoughtful suggestions for our manuscript. We have carefully revised the manuscript in response to your feedback, and these revisions have significantly enhanced its quality. Below is a summary of the major changes:

1. We expanded the discussion on limitations related to the quantification of CO_2 fertilization.

2. We ensured that all ten CEDAR-GPP model results were included in the cross-validation and intercomparison analyses.

3. We revised the cross-validation sampling protocol to account for co-located sites and updated the associated figures, tables, and text.

We also observed that your review was likely based on an earlier version of our manuscript (the preprint available in ESSD discussion). Our most recently submitted version differed from this earlier version in that we had implemented revisions in response to previous reviewers' comments, which addressed some of your points. This is noted in our point-by-point response below.

We hope our responses satisfactorily address your concerns and suggestions. We are looking forward to any further feedback or questions that you may have.

Below we provide our point-to-point response, highlighting revisions in red and line numbers from the track-change revised manuscript.

[Reviewer Comment 1] The manuscript "CEDAR-GPP: spatiotemporally up-scaled estimates of gross primary productivity incorporating CO2 fertilization" outlines an up-scaling approached based on eddy covariance estimates, which differentiates itself from other similar exercises by incorporating CO2 effects on global GPP. The methodology is benchmarked both in cross-validation and comparison to existing products, and shows similar performances. Overall, the evaluation is well done as a benchmark to introduce the new dataset, with a good framing of the motivation and discussion on the general limitations to such up-scaling exercises. The inclusion of an ensemble evaluation with some metrics of uncertainty is also well outlined and a welcome addition.

Response: Thank you for your positive feedback on our approach and evaluation of the CEDAR-GPP product.

[Reviewer Comment 2] One key aspect that I think needs to be improved is, given that the key addition compared to existing products relates to the CO2 fertilization effects, the manuscript needs further discussion of the potential limitations and implications of how these effects are introduced to the method and how users should interpret them. Some aspects, such as the fact that what is interpreted here as impacts of CO2 fertilization could include any factor with temporal trends, are mentioned only briefly in the discussion. Many other factors, including non-biological

factors such as developments in eddy covariance techniques, the bias in the fact that long term eddy covariance towers are generally placed in locations that are relatively undisturbed and protected, and potential trends in the feature sets such as due to sensor drifts, could be interpreted as CO2 fertilization effects. I do not think these represent a fundamental issue with the CEDAR-GPP products, as I would say they represent a very interesting hypothesis about how CO2 effects can be incorporated into a data driven product, and the set-up described is well evaluated. However, as data driven products tend to be seen as an "observation", it is especially important to highlight these issues to advise on the potential limitations to their use and interpretation.

Response: Thank you for this insightful suggestion. In the revised manuscript, we have expanded the discussion on the limitations of ML-based CO_2 fertilization quantification, addressing potential biases stemming from eddy covariance tower representation and the influence of non-biological trends. We have also emphasized the challenges in isolating CO_2 fertilization effects, noting the constraints due to the limited availability of long-term observations, as well as the confounding interactions with other environmental factors.

In addition, we have previously introduced a new section offering detailed guidance on data usage, specifically discussing considerations related to CO_2 fertilization, during previous reviews. In this revision, we have further emphasized the "modeled" nature of CEDAR-GPP products, providing additional cautionary notes regarding their interpretation and application.

Revisions: Line 830 – 872: Several limitations should be noted regarding GPP trend estimation and validation. First, the CFE-ML model may not fully capture the intricate mechanisms of plant physiological responses to CO_2. For example, eddy covariance towers, especially long-term sites, are typically located in homogeneous and undisturbed ecosystems, not representative of the full diversity of ecosystems globally. Thus, interactions between CO_2 and natural or human-induced disturbance, as well as many other stresses, are likely underrepresented in the models. Ultimately, the model's capacity to robustly quantify CO_2 fertilization is constrained by the scope and diversity of the eddy covariance data. Additionally, the use of spatially invariant CO_2 data may not fully represent the actual CO_2 variations that plants experience across different environments.

Secondly, CO_2 effects inferred by the CFE-ML models may be confounded by other factors that correlate with CO_2 over time. Industrialization-induced nitrogen deposition could synergistically boost GPP alongside CO_2 (O'Sullivan et al., 2019). Technological and management improvements in agriculture that contribute to a global enhancement of crop photosynthesis (Zeng et al., 2014), might also be indirectly reflected in the model estimates. Moreover, interactions with the other input features that exhibit long-term trends, such as those induced by non-biological factors (e.g. sensor orbital drifts), also affect the CO_2 effects inference. Additionally, other factors that could lead to long-term GPP trends (e.g. forest aging, disturbances) might also be underrepresented in our models.

Finally, direct validation of GPP trends is limited, particularly in tropical regions, constrained by the availability of long-term records. Detecting and evaluating trends is challenging and typically requires long monitoring records (e.g. over 10 to 15 years), since long-term changes, such as those induced by CO_2, are very small relative to large interannual variations. Evaluating aggregated GPP trends across multiple sites presents an alternative approach; however, there were still insufficient sites in tropical and evergreen broadleaf areas to

robustly validate our estimates for those ecosystems (Figure 5). Partly due to data limitations, uncertainties in GPP estimated from bootstrapped samples are very high in tropical areas (Figure 12). Thus, trend estimates in these areas should be interpreted in the context of associated uncertainties and limitations.

Line 886 – 894: Lastly, quantifications of GPP trends and their causes remain highly uncertain from site to global scales. Trend detection is often complicated by data noises and interannual variabilities, thus requiring long-term records which are limited in certain areas, biomes, and environmental conditions, such as tropics, polar regions, wetlands, as well as ecosystems with regular or anthropogenic disturbances (Baldocchi et al., 2018; Zhan et al., 2022). Moreover, isolating the effect of CO_2 is challenging, as it is confounded by other factors, such as forest regrowth, land cover change, and disturbances, which also significantly impacts long-term GPP variations. To this end, continued efforts in expanding ecosystem flux measurements and standardizing data processing present new opportunities to assess ecosystem productivity responses to changing climate conditions (Delwiche et al., 2024; Pastorello et al., 2020). Future research could also leverage novel machine learning techniques, such as knowledge-guided machine learning (Liu et al., 2024) and hybrid modeling that combines process-based and machine learning approaches (Kraft et al., 2022; Reichstein et al., 2019).

Line 915 – 924: CO_2 Fertilization Effect (CFE) configurations: the CFE-Hybrid and CFE-ML setups are preferable when assessing temporal GPP dynamics, especially long-term trends. The CFE-Hybrid setup includes a hypothetical trend from the direct CO_2 effect, while CFE-ML is purely data-driven and does not make any specific assumption about the sensitivity of photosynthesis to CO_2. Averaging the CFE-Hybrid and CFE-ML estimates is acceptable, with the difference between them reflecting the uncertainty surrounding the direct CO_2 effect. Note that the Baseline setup should not be used to study long-term GPP dynamics, especially those induced by elevated CO_2. The Baseline setup may be useful to compare with other remote sensing-derived GPP datasets that do not consider the direct CO_2 effect. Differences between these setups regarding mean GPP spatial patterns, seasonal and interannual variations are considered to be minor.

Line 929 – 936: Finally, like other upscaled or remote sensing-based GPP datasets, CEDAR-GPP should not be regarded as “observations” but rather as model estimates informed by remote sensing and ground-based data. The extent of assumptions or structural constraints varies across such datasets. CEDAR-GPP, particularly in its CFE-Baseline and CFE-ML configurations, is entirely data-driven and incorporates no explicit assumptions regarding the biological and environmental processes underlying photosynthesis, apart from the generic assumptions inherent in machine learning models. Consequently, the usage and interpretation of this dataset should be carefully framed within the context of the input eddy covariance and environmental data as well as their limitations.

[Reviewer Comment 3] Furthermore, in the comparisons between CEDAR GPP and the reference datasets, it would be important to always include the three variants (baseline and ML), at least in the supplement. Each up-scaling product has many differences due to small choices (e.g. QC, version of eddy covariance data, feature set and processing, etc...), so referencing how the three variants of CEDAR GPP differ, where the underlying methodology is most similar, will help differentiate the effects of CO2 and other methodological choices.

Response: We completely agree! During previous rounds of reviews, we have included all the CEDAR-GPP model variants in the cross-validation and intercomparison analyses, supporting a more thorough interpretation of the results. In the main text, we noted the major differences and general consistencies between the models. Please refer to our responses to your specific comments for detailed changes.

Revisions: Figures S1, S4, S5, S6, S8, S9, S11, S12

Besides the main point of discussing the limitations to interpreting the CO2 fertilization effects, there are a number of clarifications to the methodology that are described in the specific points below that should be addressed before publication.

[Reviewer Comment 4] L120 - I guess here high-quality would be based on the ONEFLUX flags as measured or high quality gap filling? Best to be explicit.

Response: Thank you for pointing this out. We have clarified in the revised manuscript

Revisions: Line 133 – 135: High-quality data refers to GPP derived from measured or high-quality gap-filled Net Ecosystem Exchange (NEE) data.

[Reviewer Comment 5] L123 - The classification of C3/C4 at all eddy covariance sites can be a difficult task, especially at crop sites with rotations. It could be useful to add this information somewhere, either as a supplement or as a referenced dataset.

Response: Thank you for the suggestion. We have added the data in the supplement.

Revisions: Line 139: This classification information is included in Supplementary Text S1.

[Reviewer Comment 6] L125 - A list of all sites used should be included, ideally with the corresponding DOI where available.

Response: Thanks for the suggestion. We included this table in the supplement during our first revision. In this round of revision, we have moved the table to Appendix A to ensure the individual site DOIs are cited in the main text references.

Revisions: Appendix A.

[Reviewer Comment 7] Table 1 - Having the information on which datasets were used in which model would be useful to show here (e.g. combining with Table S1), and would make a nice overall summary of the set-up, particularly as there is a lot of empty space in this table.

Response: Thanks for the suggestion. We have revised Table 1 to incorporate specific usage of each dataset in the model setups. Table S1 was kept to list original and extracted variables from each dataset.

Revisions: Table 1.

[Reviewer Comment 8] L213 - Here it says all data was aggregated to 0.05° resolution, but many datasets are at courser resolutions. How was the resampling done? linear interpolation? Also, I did not see a description of how the gridded data was matched to the eddy covariance towers, I guess the nearest 0.05° pixel was used?

Response: Following your suggestion, we have included a description of the resampling and the nearest pixel approach in the revised manuscript.

Revisions: Line 243 – 247: Finally, all the datasets were aggregated to a monthly time step and 0.05-degree spatial resolution. We employed the conservative resampling approach using the xESMF python package (Zhuang et al., 2023). To generate the machine learning model training data, we extracted values from the nearest 0.05 degree pixel relative to the site locations within the gridded dataset.

[Reviewer Comment 8] L270 - Was there a significant difference between the baseline and the reference models?

Response: Great question! The Baseline and CFE-REF models show no difference in long-term trends and their spatial patterns (Figure R1), as indicated by a close to 1 slope and near-zero bias term in the regression line (Figure R3a, c). This consistency suggests that the CO_2 effect has been effectively removed from the CFE-ML models.

The key difference between the Baseline and CFE-REF models lies in the magnitude of predicted GPP (Figure R2, R3). Global annual GPP from the CFE-REF models is systematically lower by 1.2 – 1.4% compared to the Baseline models (Figure R2, R3). Despite the bias, GPP predictions from both models agree well with an R^2 over 99% and a regression slope of 1. The difference in GPP magnitudes stems from the models' assumptions. The Baseline model, without accounting for CO_2 changes over time, essentially assumes a fixed CO_2 level, corresponding to an "average" level from 2001 to 2020. However, the CFE-REF models remove CO_2 effects by fixing it to the minimum level of year 2001. Therefore, the CFE-REF predicts lower GPP values, though GPP temporal dynamics are consistent between the models.

Revisions: Line 308 – 309: Long-term trends from the reference and the Baseline models are consistent.

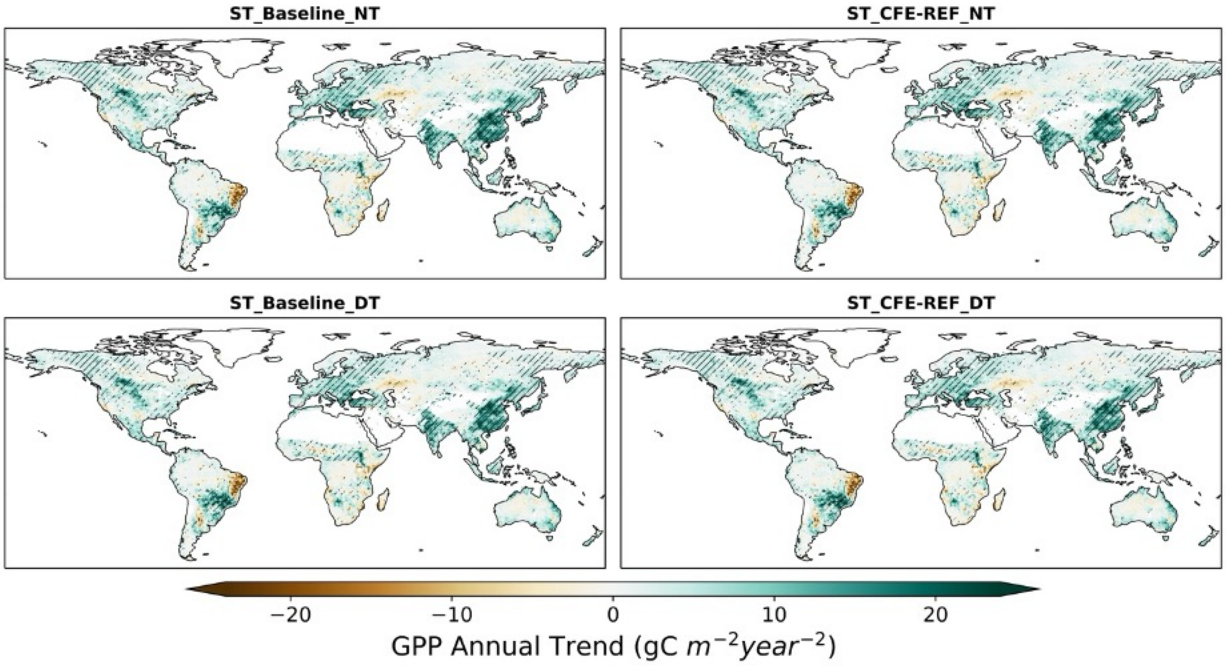

Figure R1. Comparison of GPP trend spatial patterns between the Baseline and CFE-REF setups for NT and DT models.

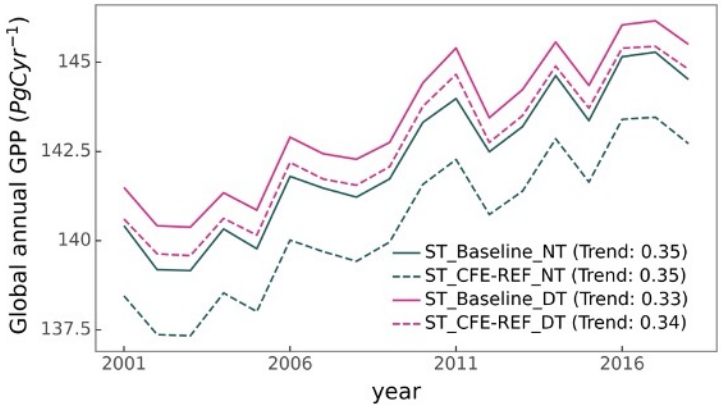

Figure R2. Global annual GPP from 2001 to 2018 for Baseline and CFE-REF models.

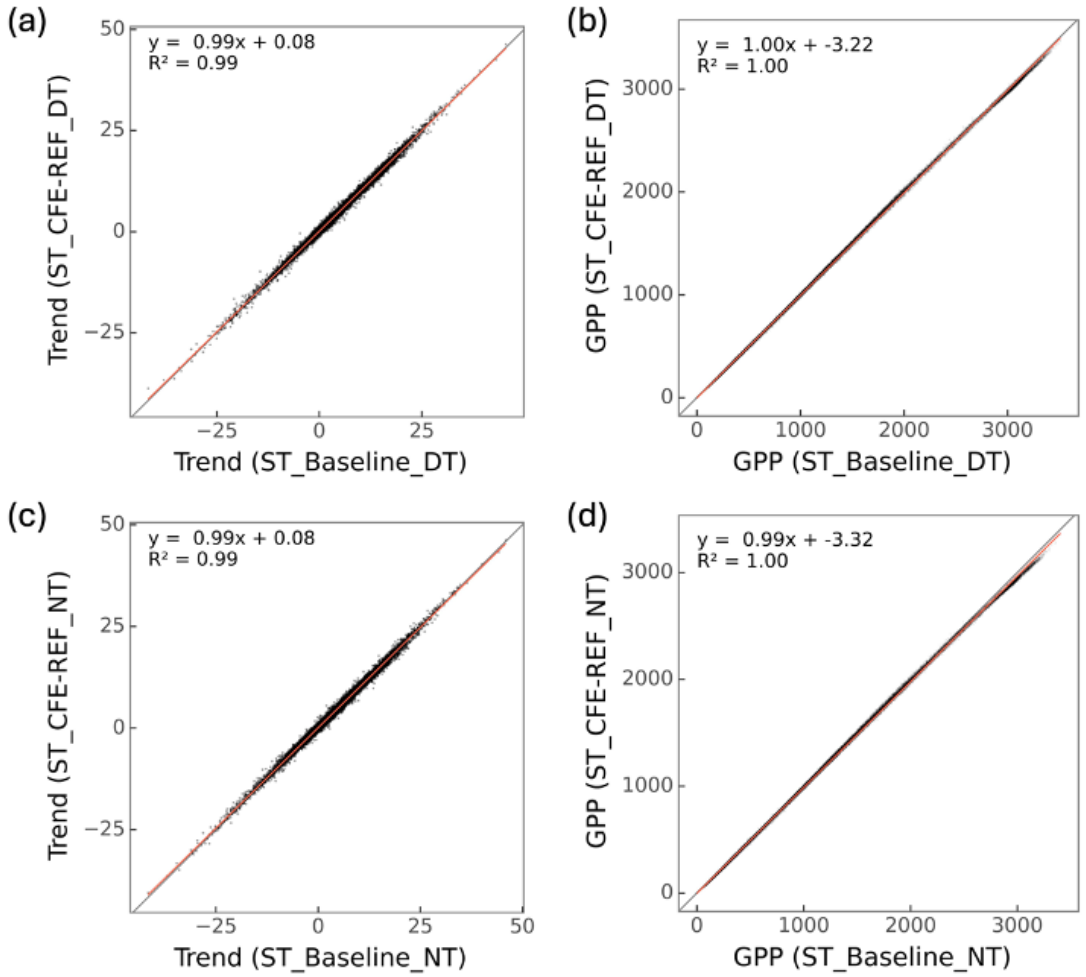

Figure R3. Comparison of Baseline and CFE-REF setups in terms of long-term trends (a,c) and annual GPP magnitudes (b,d) for DT models (a,b) and NT models (c,d). Panels (a) and (c) show long-term trends, represented by Sen slopes, calculated for each grid using annual GPP data from 2001 to 2018. Panels (b) and (d) display the annual GPP values for 2010.

[Reviewer Comment 9] L290 - When doing the split, did you account for co-located eddy covariance towers, as a number of sites are replicates and should be treated as effectively one site in the context of cross validation.

Response: We appreciate this valuable comment. We did not account for co-located sites and we agree that co-located towers should be treated as a single site in the context of cross-validation to avoid inflating model performance of generalization. We have updated our cross-validation approach to group co-located towers, i.e. towers that are no more than 0.05-degree apart, ensuring that each group is treated as a single entity during train/test splitting. The updated cross-validation results in slightly lower model accuracy (overall R^2 = 0.72 vs 0.74), but the comparative

patterns across model setups are consistent with the previous approach. We have revised all the figures and text in the cross-validation result section.

Revisions: Figure 3 – 5; Table S2; Figure S1 – S7; Text Section 3.1

L303 - When aggregating to annual values, this would mean you only took sites with at least 5 years of data that also passed the high QC threshold (>80% high-quality) for all 12 months? How many sites ended up meeting this threshold to be included in the evaluations? Also, please briefly describe Chen et al. 2022 here.

Response: We included a site-year with at least 11 months of high-quality data, and filled the remaining gap with a temporal linear interpolation. There were 81 sites used in this analysis. We have included these details in the revised manuscript.

Revisions: Line 343 – 350: To evaluate the models' ability to capture long-term GPP trends, we aggregated the monthly GPP to annual values following Chen et al. (2022), which detected the CO_2 fertilization effect across global eddy covariance sites. For sites with at least five years of observations, GPP anomalies were computed by subtracting the multi-year mean GPP from the annual GPP for each site. Anomalies were aggregated across sites to achieve a single multi-site GPP anomaly per year. We excluded a site-year if less than 11 months of data was available and used linear interpolation to fill the remaining temporal gaps. This resulted in 81 sites used in the GPP trend evaluation.

L315 - Please include an overview of the hyperparameters used in for XGBoost, including model sizes and stopping criteria. Best would be in the supplement.

Response: Thanks! We've included a description of the XGBoost hyperparameters used in the final product generation in the Supplementary. These parameters were determined based on the results of the nest cross-validation for each model setup.

Revisions: Line 364 – 365: Hyperparameters of the XGBoost models used in the final product generation were described in Supplementary Text S2.

Text S2: During the nested cross-validation (Main text Section 2.3.3), XGBoost model hyperparameters were determined using a randomized search based on 3-fold cross-validation within each training set. This process generated a best-fit parameter set for each of the five folds. When generating the global product, the final hyperparameters were determined based on a majority vote from the five best-fit parameter sets. For the short-term model setups, the XGBoost models were trained with 500 estimators (parameter "n_estimator" in the XGBoost python API), a learning rate ("learning_rate" of 0.01, and a subsample ratio of columns/features ("colsample_bytree") of 0.3 for each tree. For the long-term model setups, the XGBoost models used 300 estimators, a learning rate of 0.05, and a subsample ratio of columns of 0.3. Note that adding the CO_2 features to the models or using NT versus DT GPP did not change the selected best-fit parameter sets.

L317 - Was it possible to sample all PFTs using only 150 sites? Some PFTs have relatively few sites, and I guess each would need something like 10 sites to be represented? Here again is where a site list would be useful.

Response: Good point. Since we combined some similar IGBP classes, such as WSA with SAV, and OSH with CSH, to create our PFT categories, all PFTs have at least a few sites in all bootstrapped samples. However, as you pointed out, this does not guarantee that the samples are fully representative. For example, there were only 8 EBF sites in total. In fact, it is difficult to determine the number of sites needed to sufficiently represent each PFT. Despite the reduced sample size, bootstrapping introduces variability in representativeness across sub-samples. This helps capture at least a portion of uncertainties associated with under-sampling, which remains a major challenge in eddy covariance upscaling.

L341 - Just as a check, but was the global GPP average calculated using area weighting to account for latitudinal differences in pixel size? How were coastal pixels handled?

Response: Yes, that is an important aspect to be careful. We calculated the pixel-area-weighted GPP global average and now specified it in the revised text.

Table 3 - Here an overall performance is reported. A more comprehensive view would be to characterize the distribution of performance across each individual site, such as the median and interquartile range of each metric.

Response: Great suggestion! We have included a figure showing the distribution of R2 across sites in monthly GPP, mean seasonal cycle, and monthly anomalies (Figure S2). Please also note that Table 3 had been moved to Supplementary Table S2 following suggestions in the previous rounds of review.

Revisions: Line 412- 413: Model performance also varied across sites, and models were more advantageous in explaining mean seasonal cycles than monthly anomalies in most sites (Figure S2).

Figs. 3/4 - Include a version of these plots for all model variants in the supplement.

Response: We have included Figure S1 with all the DT models supplementing Figure 3 and Figures S3 including all the DT short-term and long-term CFE-Hybrid models supplementing Figure 4. Figure S5 provided all the model performance for different PFT and Koppen zones. Figure S6 was provided for all the DT models supplementing Figure 5.

Note that following previous reviewers' requests, we have revised Figure 5 to include trend evaluation for individual PFT and Koppen zones with at least six long-term sites.

L415 - Is this comparing trends of EC and the model output for the EC sites, or the global trends from the models?

Response: Thanks for the note. This section compares trends of EC and model predictions at the EC sites. We have revised the text for improved clarity.

Revisions: Line 470 – 473: The eddy covariance GPP from the night-time partitioning approach indicated an overall trend of 7.7 $gCm^{-2}year^{-2}$. In contrast, the ST_ Baseline_NT model predicted a more modest overall trend of 3.1 $gCm^{-2}year^{-2}$ across the flux sites, primarily reflecting the indirect CO_2 effect manifested through the growth of LAI.

Fig. 7 - Here more regions would give a clearer view of the differences in mean seasonal cycles. Also, please include all CEDAR variants.

Response: Thanks for the suggestion! In the revised manuscript, we summarized mean seasonal cycles of the 11 TransCom regions (Figure 7) and have provided all CEDAR-GPP model results in Figure S10.

Revisions: Line 542 – 563: CEDAR-GPP agreed with other GPP datasets on seasonal variabilities (average between 2001 and 2018) at the global scale, characterized by a peak in GPP in July and a nadir between December and January (Figure 7, Figure S9). At the global scale, CEDAR-GPP was most closely aligned with FLUXSAT in GPP seasonal magnitude and amplitude, while both FLUXCOM and MODIS displayed a relatively less pronounced magnitude.

In boreal and temperate regions of the Northern Hemisphere, all datasets agreed on seasonal GPP variation, with only minor variances in the magnitude of peak GPP. In Southern Hemisphere temperate regions, datasets demonstrated similar seasonality, though with greater variability in peak amplitudes compared to the Northern Hemisphere. The largest disparities were found in the South American tropical areas, where seasonal variation is less prominent. Here, FLUXSAT showed a distinct bi-modal pattern with peaks in March-April and September-October. CEDAR-GPP and FLUXCOM-ERA5 aligned with the second peak, but exhibited a less pronounced first peak. Interestingly, the DT setups of CEDAR-GPP showed slightly higher peaks in March-April in this region (Figure S10). MODIS, in contrast, indicated an inverse seasonal pattern with a small peak from June to August. Across all regions, CEDAR-GPP's seasonality aligned more closely with FLUXSAT and FLUXCOM-ERA5 than with other datasets. Differences among the ten CDEAR-GPP model setups were minimal, except for small variations in GPP magnitude in some tropical areas between NT and DT setups (Figure S10).

Fig. 8 - Does the IAV calculation include the trend? As much of the discussion revolves around the trend, it would be good to separate the IAV and the trend in the analysis. For example, are the differences between FLUXCOM-RS006 due more to a lack of variability or a lack of trend. Furthermore, including the CEDAR baseline and ML versions would help understand what are the effects of the added CO2 effect and what is the underlying variability due to the feature set.

Response: IAV was calculated based on the detrended time series. We have included Figure S11 showing the IAV spatial patterns for the ten CEDAR-GPP models in response to previous review comments. In general, there were no differences between the Baseline and CFE models. The LT models showed slightly lower IAV than the ST models, likely due to differences in the input RS data.

L619 - Drylands is another area that is difficult to represent, which is reflected in the highest uncertainties in Figure 12. This is likely due to both the tower representation, and the difficulty in capturing drought responses.

Response: This is a good point! We have highlighted this aspect in the revised manuscript.

Revisions: Line 709 – 711: However, data availability remains limited in critical carbon exchange hotspots such as tropics, subtropics, drylands, and boreal regions, as well as in mountainous areas (Figure 1).

Line 771 – 774: High prediction uncertainties (Figure 12) in drylands also suggest the machine learning models did not sufficiently represent the mechanisms of water stress and drought responses.

L621 - While the uncertainty estimates are an important and welcomed inclusion, it is important to mention that tree methods tend to deflate uncertainties in extrapolation, as they tend to stick to the edges of the training distribution. This can lead to what looks like high certainty when extrapolating, when in fact there may be strong biases.

Response: Good point! We have discussed this aspect in the revised manuscript.

Revisions: Line 799 – 802: Furthermore, tree-based models do not generalize well to unseen conditions, and the uncertainty estimates derived from bootstrapping of XGBoost models may underrepresent actual biases stemming from limitations in training data representation.

L640 - Here it would be very relevant to compare the new FLUXCOM X-BASE products, as they are based on similar underlying datasets. However, as it seems both products were developed in parallel, it does not seem fair to hold this as a requirement. The data is available here: https://gitlab.gwdg.de/fluxcom/fluxcomxdata

Response: Thanks for pointing this out and sharing the link to the FLUXCOM X-BASE product. We will definitely use it as a reference for our future analysis. We have incorporated the X-Base preprint as a major reference for upscaling approaches in the Introduction section.

L689 - See comments on L621.

Response: Thanks! We've highlighted the limitations of using tree-based models to quantify uncertainty in this section.

L716 - Here there are no mentions of the effects of water limitations, which show the highest uncertainty in the model spread, and is a key limitation on GPP.

Response: Thank you for highlighting this point. The underrepresentation of water stress in the models likely constitutes a key limitation, as evidenced by the high uncertainties in GPP estimates within dryland regions. We have addressed this aspect in previous sections in response to your earlier comments (e.g., Lines 771 - 773).

Regarding CO_2 fertilization under water stress, it is anticipated that elevated CO_2 could improve water use efficiency, thereby enhancing plant productivity under moderate water stress by reducing stomatal conductance. While the CFE-Hybrid setup does not incorporate this factor—yielding a conservative estimate of CO_2 effects—the CFE-ML model may capture some aspects of water use efficiency as represented in the eddy covariance measurements. We have noted this in the text (see below). However, due to the complex nature of CO_2 effects, it remains challenging to isolate individual pathways through which CO_2 influences plant physiological processes.

Lines 819–821: This strategy allowed the model to potentially capture multiple physiological pathways of the CO_2 impact evidenced in the eddy covariance measurements, including the increases of the biochemical rates and enhancements in the water use efficiency (Keenan et al., 2013).

L722 - While the CFE-ML product is learning long term trends from the eddy covariance data, as I understand it, the CO2 data is from Mauna Loa. I think it is important to be very careful to distinguish that fact so as to not conflate the models here with experiments which are actually looking at the impacts of CO2 as a plant sees it (such as lab or FACE experiments).

Response: This is a good point. In the revised manuscript, we discussed the limitations of using a spatially invariant CO_2 in the model. We have also revised the sentence in question to enhance clarity.

Revisions: Line 819 – 821: This strategy allowed the model to potentially capture multiple physiological pathways of the CO_2 impact evidenced in the eddy covariance measurements, including the increases of the biochemical rates and enhancements in the water use efficiency (Keenan et al., 2013).

Line 837 – 839: Additionally, the use of spatially invariant CO_2 data may not fully represent the actual CO_2 variations that plants experience across different environments.

L731 - In addition to other factors that have long term trends which might coincide with CO2 increases, it is important to mention that eddy covariance measurements tend to be made in specific ecosystems, namely relatively undisturbed and curated. So the long term trends of FLUXNET, ICOS, and AmeriFLUX, particularly the sites with long term monitoring of trends, are not representative of "real-world" ecosystems. This point should be discussed as a parallel to the spatial sampling bias, particularly when talking about trends.

Response: This is an important point! We have expanded the discussion on the limitations of the machine learning approach in quantifying CO_2 effects, specifically addressing the sampling bias inherent in eddy covariance data. Additionally, we have highlighted the need for long-term eddy covariance observations across a broader range of ecosystems to improve representativeness and better capture real-world trends.

Revisions: Line 830 – 837: First, the CFE-ML model may not fully capture the intricate mechanisms of plant physiological responses to CO_2. For example, eddy covariance towers, especially long-term sites, are typically located in homogeneous and undisturbed ecosystems, not representative of the full diversity of ecosystems globally. Thus, interactions between CO_2

and natural or human-induced disturbance, as well as many other stresses, are likely underrepresented in the models. Ultimately, the model's capacity to robustly quantify CO_2 fertilization is constrained by the scope and diversity of the eddy covariance data.

Line 887 – 890: Trend detection is often complicated by data noises and interannual variabilities, thus requiring long-term records which are limited in certain areas, biomes, and environmental conditions, such as tropics, polar regions, wetlands, as well as ecosystems with regular or anthropogenic disturbances (Baldocchi et al., 2018; Zhan et al., 2022).