Response to Comments of Reviewer #2

Thank you for the positive evaluation and constructive suggestions on our manuscript. We have carefully considered all comments and revised the manuscript accordingly. In this response letter, your original comments are shown in **black**, our responses are given in **blue**, and the revised text quoted from the manuscript are given in **red**. The line numbers cited in this letter correspond to the tracking-change version of the manuscript.

General comments:

This study aimed to produce eddy covariance (EC) site-based Global Hourly GPP estimates (named EGO) using a causal inference approach (CKML-GPP) spanning the period from 2000 to 2022. While the model's performance in regions with limited EC flux sites may remain challenging, I believe this attempt to improve hourly GPP estimates by combining high-frequency on-site EC data with remote sensing data carries important insights for global carbon cycle studies. Overall, the manuscript reads well. However, I felt the explanations regarding how the authors compared EGO with earlier products (i.e., FLUXCOM and X-BASE) are currently lacking and require clarification. Below are my specific comments.

Response: Thank you for your positive evaluation of our study and for recognizing the potential value of EGO for global carbon-cycle studies. We agree that the comparison between EGO and existing products needed to be described more clearly. We have further added a detailed explanation in Lines 264–271 (see below). In addition, we have added Fig. S5 to more clearly illustrate the workflow used for product comparison.

"For product comparison, we used only the predictions from the combined test folds, ensuring that each record used for evaluation was predicted by a CKML-GPP model that had not seen that sites during training, thereby providing a strictly out-of-sample validation dataset (Fig. S5). The validation dataset was subsequently used as the basis for a fair comparison between CKML-GPP and existing hourly GPP products, including FLUXCOM and X-BASE. For each flux tower record in this dataset, we extracted the corresponding grid-cell GPP values from EGO, FLUXCOM, and X-BASE at the same site location, month, and local solar hour, and compared them against observations. Because these products have different native spatial resolutions, we further harmonized the spatial resolution for product-level comparisons. Specifically, EGO was aggregated from its native 0.05° grid to 0.25° and 0.5°, respectively, matching the native spatial resolutions of X-BASE and FLUXCOM (Fig. S6)."

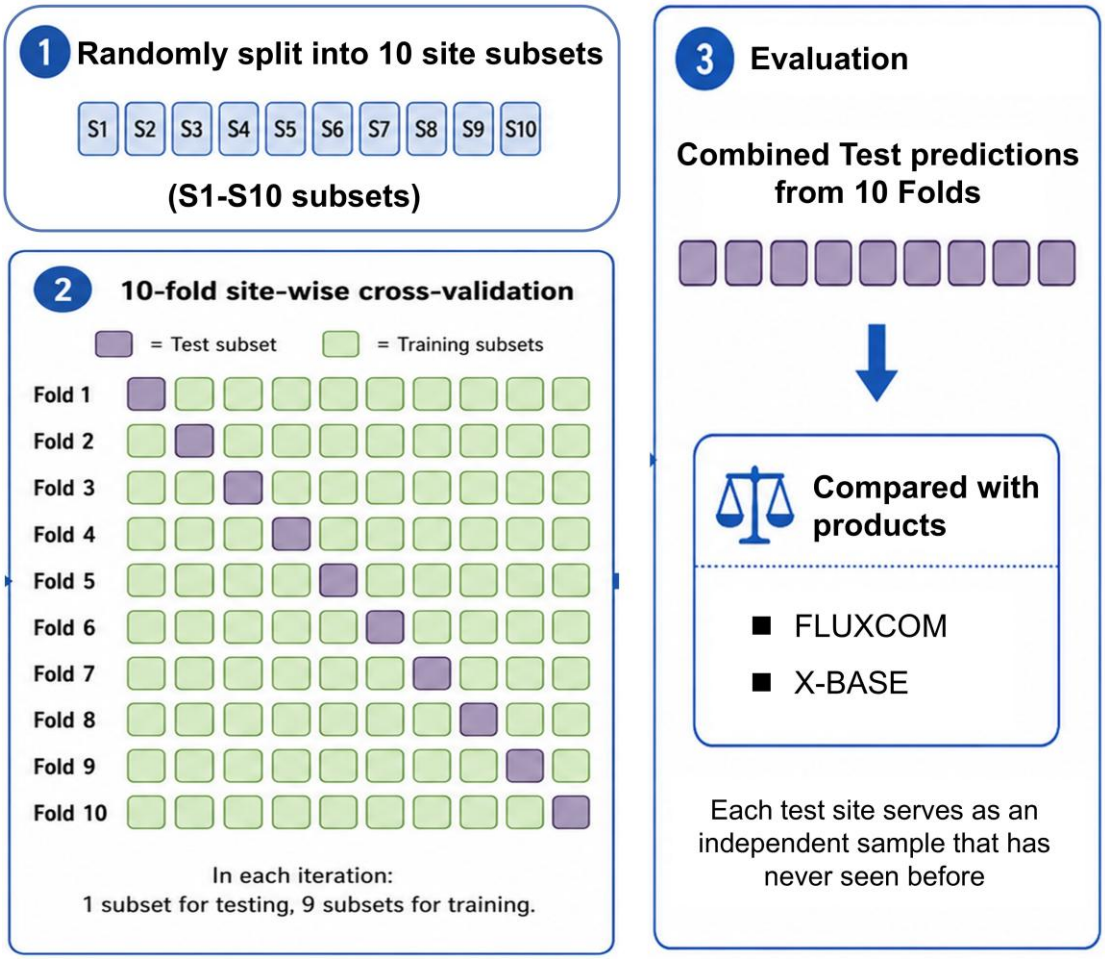

Figure S5. A visual explanation of the 10-fold site-wise cross-validation and product comparison.

Comments:

1. L65: Please spell out "OCO" at its first occurrence.

Response: We have spelled out "Orbiting Carbon Observatory (OCO)" at its first occurrence in Line 67:

"In parallel, the Orbiting Carbon Observatory-3 (OCO-3) instrument onboard the International Space Station measures solar-induced chlorophyll fluorescence (SIF), …"

2. L97: Lin et al. (2021) do not seem to use a causal inference approach in their analysis. Citing this paper here for the PCMCI method is somewhat confusing.

Response: Thank you for catching this. Lin et al. (2021) did not use a causal inference framework, so we have replaced it with Runge et al. (2019b) (Lines 100–101), which is more directly relevant to PCMCI.

"As one of the state-of-the-art causal inference methods, the PCMCI (Peter and Clark

Momentary Conditional Independence) offers a robust solution especially suitable for earth science applications, as it rigorously filters out confounding effects to identify true causal pathways (Runge et al., 2019b; Runge et al., 2023)"

Reference:

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D.: Detecting and quantifying causal associations in large nonlinear time series datasets, Sci. Adv., 5, 15, https://doi.org/10.1126/sciadv.aau4996, 2019b.

3. L165-166: Please subscript the 2 in "CO2".

Response: We have corrected it in the relevant text. The revisions have been made in Lines 179–181 and Lines 618–620.

Lines 179–181:

"To account for the influence of atmospheric CO_2 concentration on photosynthesis, we used global 3-hourly CO_2 concentration data at 3° × 2° from the NOAA CarbonTracker (http://carbontracker.noaa.gov) and applied time-weighted interpolation to generate continuous hourly CO_2 series (Chen et al., 2019)."

Lines 618–620:

"Moreover, we used MERRA-2 (0.5°× 0.625°) and CarbonTracker (3° × 2°) datasets to represent the effects of diffuse radiation fraction and atmospheric CO_2 concentration on diurnal GPP dynamics."

4. L207: Please add a space in "TableS2".

Response: We have corrected it in Line 222.

Lines 221–222:

"The final set of global gridded predictors used for modelling and upscaling is listed in Table S2."

5. Fig 1 (c): The legends are a bit small. Please increase their font size.

Response: We have increased the font size of the legends in Fig. 1.

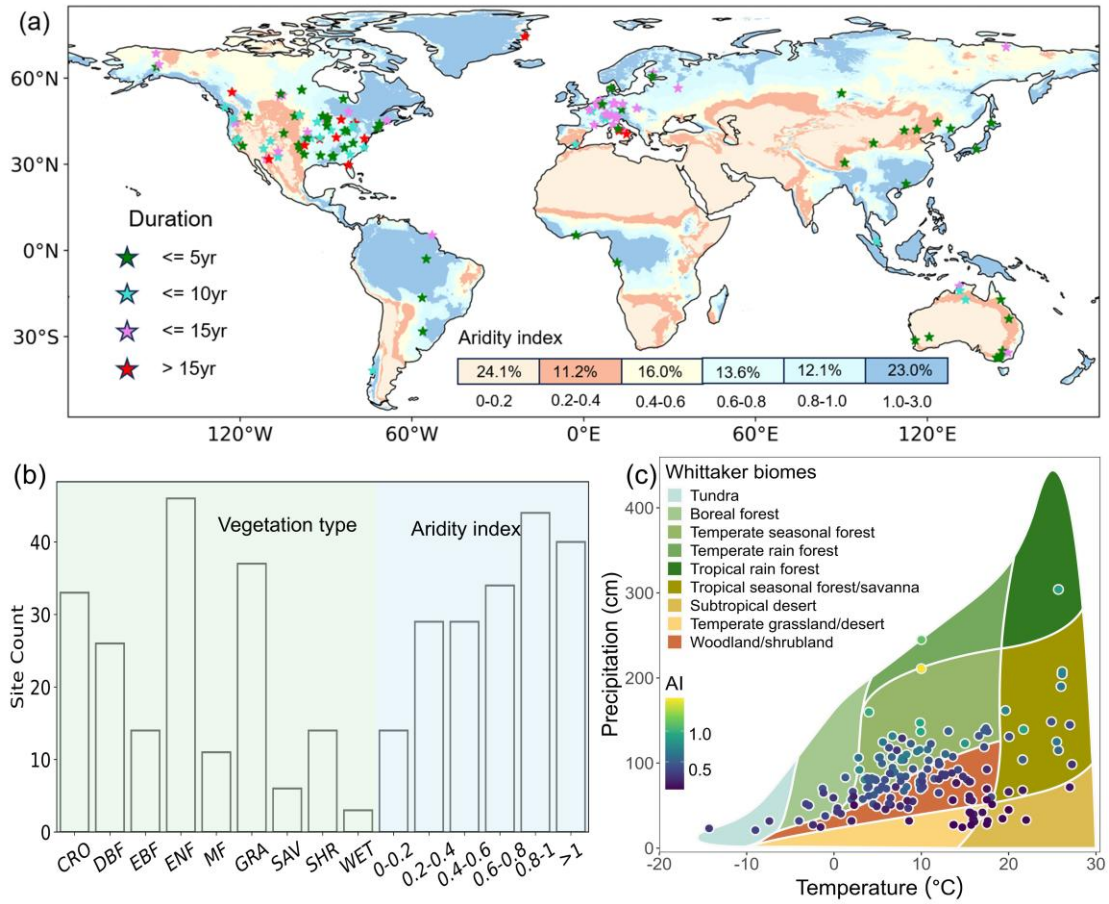

Figure 1. (a) Spatial distribution of the 190 eddy covariance flux sites used in this study. The background map shows the global aridity index, and the color bar indicates the proportion of land area within each aridity class. Sites with different observation durations are marked by star symbols in different colors. (b) Number of sites categorized by vegetation type and aridity gradient. (c) Distribution of sites across the Whittaker biome classification, where each site is positioned according to its mean annual temperature and precipitation. Vegetation type abbreviations are as follows: cropland (CRO), deciduous broadleaf forest (DBF), evergreen broadleaf forest (EBF), evergreen needleleaf forest (ENF), mixed forest (MF), grassland (GRA), savannas (SAV), shrubland (SHR) and wetland (WET).

6. L249: The authors use the phrase "following previous studies," but only a single study is cited. Please revise or add additional citations.

Response: We have added additional relevant citations (Kang et al., 2025; Zhao and Zhu, 2025) in Lines 277–278.

"This ensemble strategy, following previous studies, effectively reduces prediction uncertainty (Kang et al., 2025; Nathaniel et al., 2023; Zhao and Zhu, 2025)."

Reference:

Kang, Y., Bassiouni, M., Gaber, M., Lu, X., and Keenan, T. F.: CEDAR-GPP: spatiotemporally upscaled estimates of gross primary productivity incorporating CO 2 fertilization, Earth Syst. Sci. Data, 17, 3009-3046, https://doi.org/10.5194/essd-17-3009-2025, 2025.

Nathaniel, J., Liu, J., and Gentine, P.: MetaFlux: meta-learning global carbon fluxes from sparse spatiotemporal observations, Sci. Data, 10, 15, https://doi.org/10.1038/s41597-023-02349-y, 2023.
Zhao, C. and Zhu, W.: Vegetation structure and phenology primarily shape the spatiotemporal pattern of ecosystem respiration, Commun. Earth Environ., 6, 15, https://doi.org/10.1038/s43247-025-02240-1, 2025.

7. L250-252: Until reading this sentence, I expected the EGO outputs to be continuous hourly GPP estimates, similar to X-Base. It would be better to clarify this earlier in the manuscript, such as in the Abstract or the final paragraph of the Introduction.

Response: Thank you for this valuable suggestion. We have added a description of the EGO in the Abstract, specifying that EGO provides monthly-averaged hourly GPP estimates.

Lines 24–28:

"Based on eddy-covariance measurements and multi-source meteorological variables, vegetation properties, and land-cover fields, we generated a global 0.05° monthly-averaged hourly GPP product covering 06:00–18:00 local solar time from 2000 to 2022, named EGO (Eddy covariance site-based Global hOurly) GPP, and then evaluated how well EGO reproduces observed diurnal cycles and their responses to extreme events"

8. L304-305: What does "all products were compared under identical conditions" actually mean? Specifically: (1) Did you rerun the other models to apply the same cross-validation approach used for EGO?, (2) Were the comparisons conducted over the exact same time periods (both across years and hours of the day), and were they all evaluated using monthly-averaged hourly GPP outputs?, (3) How were the observed GPP data sampled? Were only the EC tower sites included in each respective model plotted? Adding a dedicated paragraph in the Methods section to clarify this comparison is necessary. Currently, the comparison feels like it abruptly begins in Results Section 4.1.

Response: Thank you for pointing these details out. We have added a dedicated paragraph in Lines 264–269 to clarify the comparison procedure.

Specifically, (1) we did not rerun FLUXCOM or X-BASE. (2) For each flux-tower record in this validation dataset, we extracted EGO, FLUXCOM, and X-BASE GPP at the same site location, month, and local solar hour, and compared all products against the same observed monthly-averaged hourly GPP values. (3) For EGO, only predictions from the combined test folds were used, ensuring that each evaluation record was predicted by a CKML-GPP model that had not seen that site during training. We have added Fig. S5 (please see our response for the overall comment) to more clearly show the comparison workflow.

Lines 264-269:

"For product comparison, we used only the predictions from the combined test folds, ensuring that each record used for evaluation was predicted by a CKML-GPP model that had not seen that sites during training, thereby providing a strictly out-of-sample validation dataset (Fig. S5). The validation dataset was subsequently used as the basis for a fair comparison between CKML-GPP and existing hourly GPP products, including FLUXCOM and X-BASE. For each flux tower record in this dataset, we extracted the corresponding grid-cell GPP values from EGO, FLUXCOM, and X-BASE at the same site location, month, and local solar hour, and compared them against observations."

9. L320: Please spell out the "SHAP" method at its first mention.

Response: We have spelled out the SHAP method at its first mention in Line 251.

"Finally, we used the SHAP (SHapley Additive exPlanations) method to interpret the feature contributions of CKML-GPP, quantifying both the magnitude and direction of each environmental variable's influence on hourly GPP variation (Zhang et al., 2024)"

10. Figs 6–8: "Local solar time" should include the unit ("Hour") in the axis titles.

Response: We have added the unit "hour" to the "Local solar time" axis titles in Figs. 6–8.

11. Fig 6: Please spell out the six vegetation types for each panel in the figure caption.

Response: We have revised the Fig. 6 caption as requested and spelled out the six vegetation types for each panel in Lines 408–409.

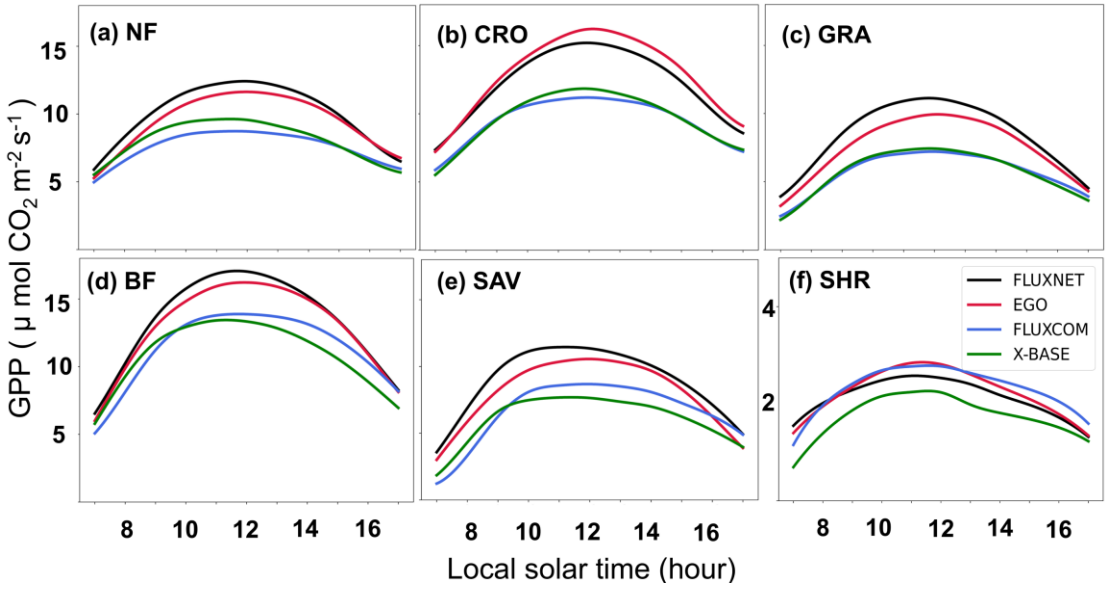

Figure 6. Mean diurnal GPP cycles for different vegetation types. Each curve represents the averaged diurnal GPP variation during growing seasons across all test sites within the same

vegetation category. Observations and product estimates are shown in colors consistent with panel (f). For clarity, vegetation types were moderately aggregated: NF includes evergreen needleleaf forests (ENF) and deciduous needleleaf forests (DNF); BF includes evergreen broadleaf forests (EBF), deciduous broadleaf forests (DBF), and mixed forests (MF). CRO, GRA, SAV, and SHR was the abbreviation of Cropland, Grassland, Savannas, and Shrubland, respectively.

12. Fig 8: The subpanel title (a) overlaps with the y-axis values. Also, the statistics table in panel (a) overlaps the Case 3 line. Please relocate them for better visualizations.

Response: We have adjusted Fig. 8 to avoid these overlaps. Specifically, we relocated the subpanel title in panel (a) and repositioned the statistics table to improve the clarity of the visualization.

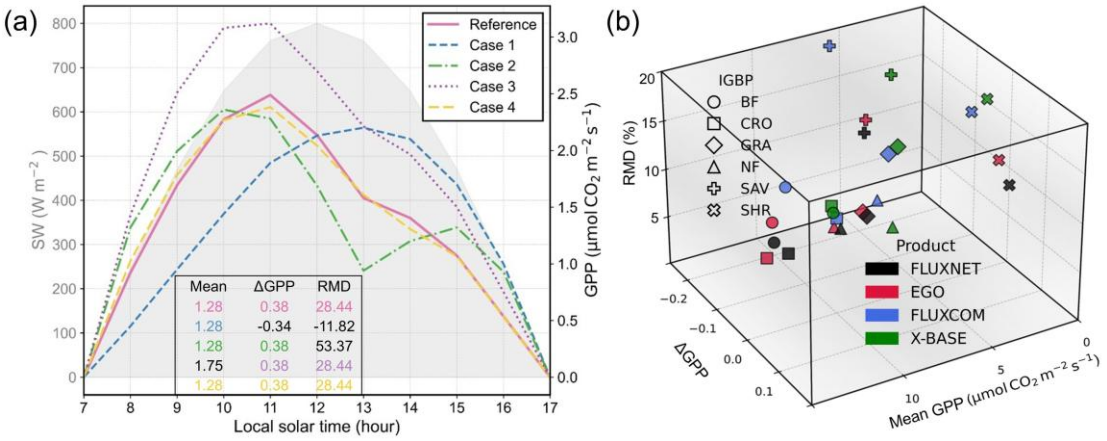

Figure 8. Diurnal GPP pattern evaluation based on diurnal metrics. (a) shows reference GPP and four controlled cases: (1) mean magnitude only, (2) magnitude and asymmetry, (3) asymmetry and midday depression, and (4) all three factors jointly controlled. For each case, a Gaussian-function-based curve generation method were used to produce 100 random diurnal GPP curves satisfying the control conditions, and their mean profiles were shown. (b) presents the performance of products in a three-dimensional metric space, where the X, Y, and Z axes represent Mean GPP (absolute magnitude), ΔGPP (asymmetry), and RMD (midday depression), respectively. Different vegetation types are marked by distinct symbols, and products are distinguished by colour. A smaller distance between a product and the observed point in this space indicates a closer match to the observed diurnal GPP dynamics.

13. Fig 9: The subplots span from 8:00 to 16:00, but the figure caption states they run from 8:00 to 18:00.

Response: Sorry for this typo. The figure caption for Fig. 9 has now been corrected to match the time range shown in the subplots (8:00–16:00).

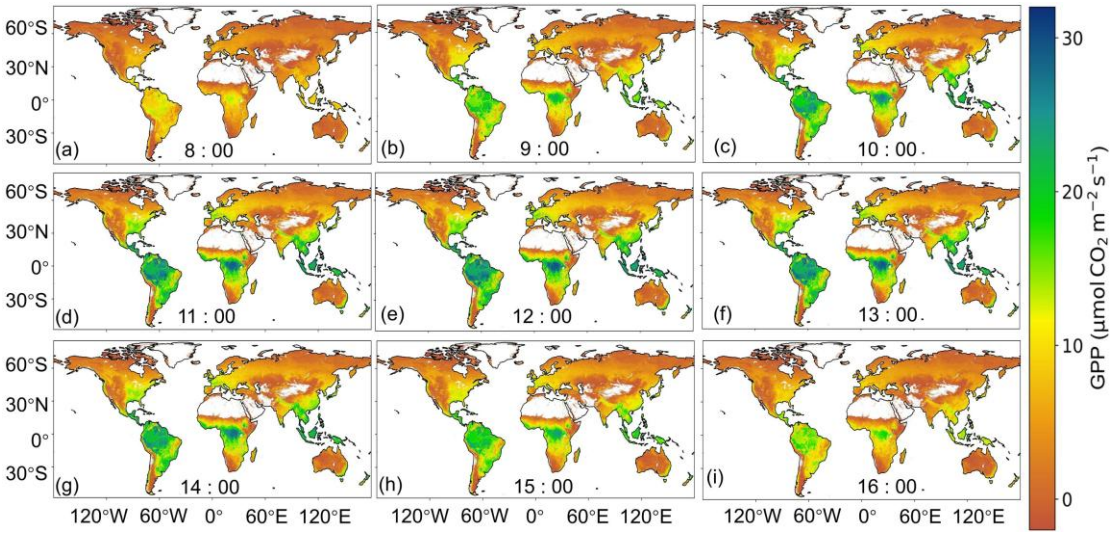

Figure 9. Global diurnal pattern of EGO GPP averaged over the growing seasons from 2000 to 2022, a-i represents the spatial distribution of hourly GPP at 1-hour intervals spanning from 08:00 to 16:00.

14. Fig 11: Please add parentheses for panels (c) and (d) in the figure caption.

Response: We have added parentheses for panels (c) and (d) in the Fig. 11 caption.

15. Figs 11 & 14: Please add the appropriate units for each variable in the color scale bars.

Response: We have revised Figs. 11 and 14 by adding the appropriate units for each variable.

16. 5.1: The authors suggest that EGO performs better than past models (especially X-Base, given its use of hourly data inputs) mainly due to the inclusion of the aridity index and the implementation of PCMCI. While the authors have compared the model with and without AI in Fig. S6, proving that the inclusion of AI improves performance, how does the comparison look between PCMCI + XGBoost vs XGBoost alone?

Response: Thank you for this valuable comment. We agree that it is necessary to directly compare PCMCI + XGBoost (CKML-GPP) with XGBoost alone. We therefore added a comparison between CKML-GPP and baseline models, including XGBoost, Random Forest, and Long Short-Term Memory (LSTM). The experimental design is described in the Methods in Lines 244–246.

"We also established three conventional machine learning models as baselines for performance comparison: EXtreme Gradient Boosting (XGBoost), Random Forest (RF), and Long Short-Term Memory (LSTM) (Fig. S3). These models all have strong capabilities for nonlinear regression, and have been widely applied in flux upscaling studies."

This comparison directly evaluates PCMCI + XGBoost against XGBoost alone, and the results show that PCMCI + XGBoost further further improved the performance over XGBoost alone, increasing R2 by 0.03 for the training set and by 0.05 for the test set (see below: Lines 335–339 and Fig. S3). This indicates that incorporating causal knowledge into XGBoost not only improves the model's ability to capture complex driver–GPP relationships during training, but also enhances its generalization to unseen test samples. Furthermore, we emphasize that the value of the CKML framework is not limited to accuracy improvement. Its PCMCI-derived causal constraints help reduce reliance on spurious correlations and provide a more robust basis for modelling and interpreting large-scale diurnal photosynthetic dynamics.

"Comparative experiments among the three baseline models (Fig. S3) showed that XGBoost achieved the best performance (test set: $R^2 = 0.71$, RMSE = 4.48 μmol CO_2 m^{-2} s^{-1}), followed by Random Forest, whereas LSTM performed relatively poorly. The proposed CKML-GPP model further improved prediction accuracy over XGBoost, increasing R^2 by 0.03 for the training set and by 0.05 for the testing set. These improvements indicate that the PCMCI-guided model not only enhances predictive skill to capture complex driving relationships, but also improved its generalization to unseen test samples."

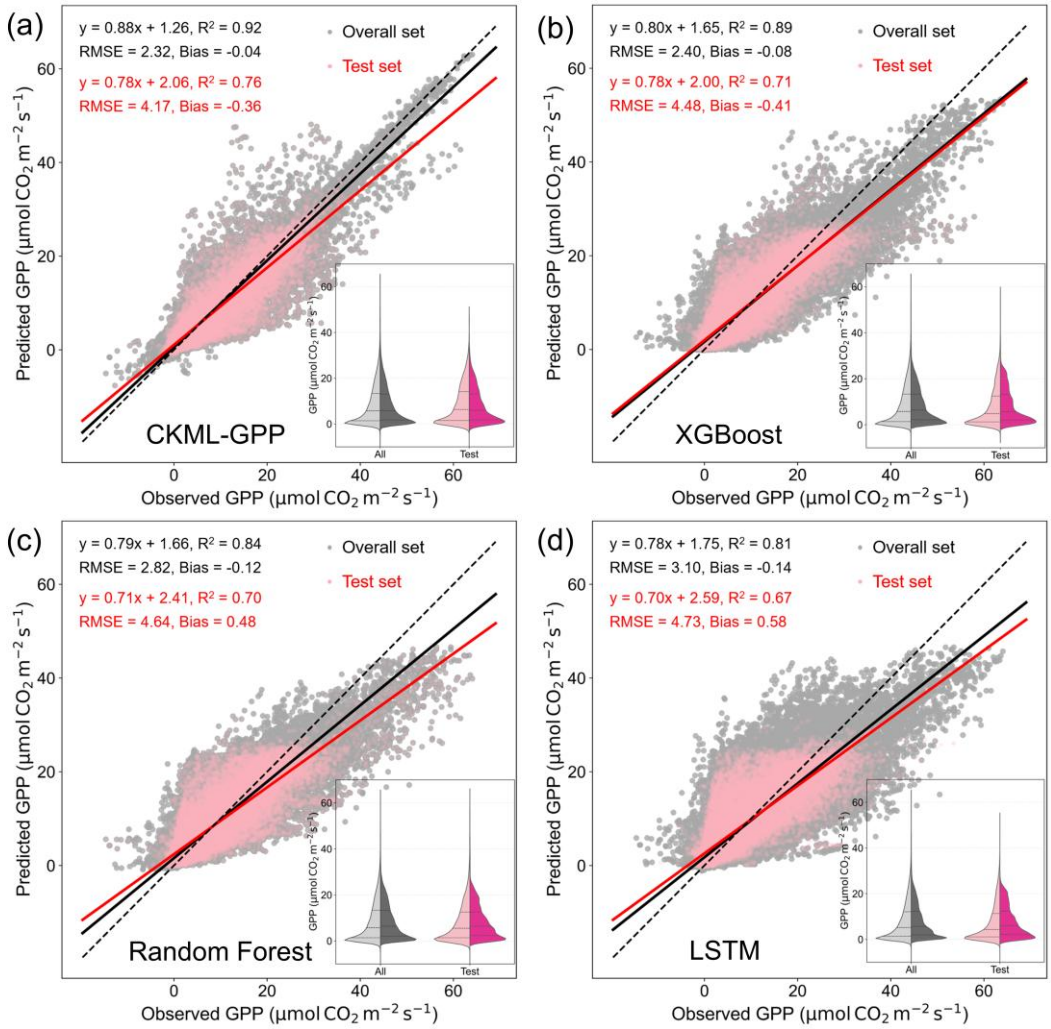

Figure S3. Comparison of the accuracy of CKML-GPP and other baseline models. (a-d) represent the predictive performance of CKML-GPP, XGBoost, Random Forest (RF) and LSTM, respectively. In each panel, the black solid line and black text indicate the regression line and corresponding performance metrics for all samples, while the red one represents test samples. The black dashed line represents the 1:1 reference line. The violin plots in the lower right corner of each panel illustrate the distributions of observed and predicted GPP values—left for the overall dataset and right for the test set. Light colors indicate observed GPP, and dark colors indicate predicted GPP. The numerical values above each violin represent the corresponding mean, with labels matching the violin colors.