Response to Comments of Reviewer #1

Thank you for your constructive and insightful comments on our paper. These comments have helped us improve our manuscript. We have carefully reviewed the comments and revised our 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:

Liu et al. develop a new global GPP model called EGO using a CKML-XGBoost approach and compare it against FLUXCOM. The new model is interesting, but generally it's hard to make too much of a case for studying diurnal patterns of GPP given the massive uncertainties in partitioning GPP (e.g. the nighttime approach will give different patterns and new studies like those of Keenan et al. emphasize the importance of under appreciated processes like the Kok effect). But regardless it's an interesting challenge for models. The major shortcoming in my opinion is that the authors did not benchmark improvements of the CKML-XGBoost model against traditional approaches (e.g. LSTM) or standard XGBoost models. How are we to know if it's an improvement, and by how much? I don't doubt that it makes an improvement, but without explaining what is really new (and how much the novelty makes a difference), it's hard to know if CKML is really a future direction of an approach that might work in principle but may (or may not) be meaningfully better. Such an addition needn't extend to the FLUXCOM comparisons but it is important from a model development perspective. Minor comments follow.

Response: We sincerely thank you for the positive evaluation of our work and for recognizing EGO as an interesting and useful attempt for mapping global hourly photosynthesis. We also appreciate your thoughtful and constructive comments and we have addressed these concerns as follows.

1) We agree that uncertainties in eddy-covariance GPP partitioning, including differences between daytime- and nighttime-based approaches, can affect the estimated diurnal pattern of GPP. Our goal here is not to eliminate this source of uncertainty, but to provide an hourly GPP modelling framework whose performance is as robust as possible given the available observations. To examine whether our model performance is robust to different partitioning approaches, we added a sensitivity experiment in which CKML-GPP was applied to daytime-partitioned GPP (GPP_DT) and compared with the results based on nighttime-partitioned GPP (GPP_NT). This analysis is now

described in Lines 249–251, and the results are shown in Lines 339–342 and Fig. S4.

Lines 249–251:

"We also conducted an additional experiment by applying CKML-GPP to estimate daytime-partitioned GPP (GPP_DT) and comparing its performance with that obtained using GPP_NT (Fig. S4), to examine the robustness of model performance."

Lines 339–342:

"We further added a sensitivity experiment in which CKML-GPP was applied to daytime-partitioned GPP (GPP_DT) and compared with GPP_NT (Fig. S4). The results show that the model performance based on GPP_DT was generally consistent with that based on GPP_NT, suggesting the robustness of CKML-GPP across different partitioning methods."

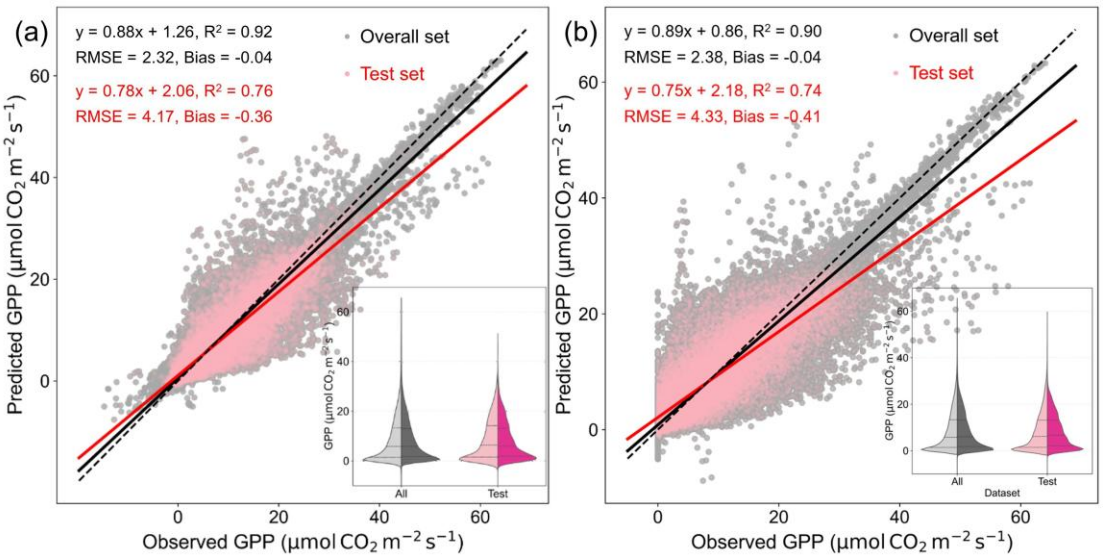

Figure S4. Performance of CKML-GPP in estimating (a) nighttime-partitioned GPP (GPP_NT) and (b) daytime-partitioned GPP (GPP_DT).

We also expanded the discussion of this uncertainty in Lines 615–618 (see below), including the potential influence of different respiration assumptions and processes such as the Kok effect.

"Another source of uncertainty arises from the fact that tower-based GPP is not directly measured but inferred from net ecosystem exchange through flux partitioning. Different partitioning approaches may introduce some uncertainty in estimated GPP and its diurnal patterns, as they rely on different assumptions about ecosystem respiration, including processes such as the Kok effect (Keenan et al., 2019; Ranjbar et al., 2026)."

Reference:

Keenan, T. F., Migliavacca, M., Papale, D., Baldocchi, D., Reichstein, M., Torn, M., Wutzler, T.,

and Lawrence Berkeley National Laboratory Lbnl, B. C. U. S.: Widespread inhibition of daytime ecosystem respiration, Nat. Ecol. Evol., 3, 407-415, https://doi.org/10.1038/s41559-019-0809-2, 2019.
Ranjbar, S., Desai, A. R., Hoffman, S., Zahn, E., Bou Zeid, E., and Stoy, P. C.: Constrained carbon partitioning: a self-trained physics-informed machine learning model refines gpp estimates from eddy covariance measurements, Glob. Change Biol., 32, e70886, https://doi.org/10.1111/gcb.70886, 2026.

2) We acknowledge your point that, despite emphasizing the importance and novelty of causal knowledge in the Introduction (Lines 95–105, see below), a clearer quantification of its actual improvement is needed. Accordingly, we added a benchmark experiment comparing CKML-GPP with three baseline models (XGBoost, random forest, and LSTM). Detailed experimental design and results are provided in our response to Comment 9.

"Traditional machine learning models typically rely on correlations rather than directional causal relationships (Yuan et al., 2022). Without causal guidance, these models risk capturing spurious associations driven by confounding factors such as predictors sharing similar variations, which can render predictions unreliable (Galytska et al., 2023). 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). Another benefit is its ability to represent time-lagged dependencies, which is critical for estimating hourly GPP, since previous research has highlighted that vegetation photosynthesis exhibits certain lagged responses to environmental fluctuations in sub-daily (Krich et al., 2020). By embedding the causal structure into machine learning models, PCMCI acts as a form of biophysical constraint that guides model training (Runge et al., 2023; Yuan et al., 2024)."

Comments:

1. GPP is just canopy photosynthesis, the way it's being described is like you're trying to talk around something.

Response: We have revised the sentence in Lines 40–42 as:

"Gross primary productivity (GPP) is defined as the total carbon fixed by terrestrial ecosystems through photosynthesis (Beer et al., 2010)."

2. these early studies like Wofsy et al. and Grace et al. and more should be cited. In general, fundamental (or at least more fundamental) references are missing throughout the manuscript. For example, both Ruehr et al. references cited beforehand are great to note, for example, but these are not the first times that people realized that GPP is the largest flux in the global C cycle or that vegetation responds to environmental variability at multiple (including short) time scales.

Response: Thank you for this valuable suggestion. We have therefore revised the Introduction by replacing and supplementing the relevant citations with earlier studies, including Wofsy et al., 1993; Schimel et al., 2001, and other fundamental references. These changes have been made in Lines 39–51 and Lines 121–122.

Lines 39–40:

"Vegetation assimilates atmospheric CO_2 into organic carbon through photosynthesis, forming the largest carbon flux between the biosphere and the atmosphere (Schimel et al., 2001; Keenan and Williams, 2018)."

Lines 43–47:

"Analyses at longer timescales (e.g., daily, monthly and annual) provide valuable insights into photosynthetic phenology and interannual variability (Xiao et al., 2025). In contrast, shorter scales (e.g., sub-daily or hourly) focus on how vegetation photosynthesis responds to instantaneous changes in light, temperature, and water availability within a day (Urbanski et al., 2007)."

Lines 48–51:

"Early flux tower-based studies showed that short-term variations in environmental factors can induce pronounced diurnal fluctuations in vegetation photosynthesis, that are often obscured in longer-timescale analyses (Baldocchi et al., 2001; Goulden et al., 2004; Wofsy et al., 1993)."

Lines 121–122:

"Recent studies have revealed distinct diurnal features of vegetation photosynthesis, including global-scale morning–afternoon asymmetry (Khan et al., 2022; Lin et al., 2019; Liu et al., 2024), …"

Reference:

Baldocchi, D., Falge, E., Gu, L. H., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., Goldstein, A., Katul, G., Law, B., Lee, X. H., Malhi, Y., Meyers, T., Munger, W., Oechel, W., U, K., Pilegaard, K., Schmid, H. P., Valentini, R., Verma, S., Vesala, T., Wilson, K., and Wofsy, S.: Fluxnet : a new tool to study the temporal and spatial variability of

ecosystem-scale carbon dioxide , water vapor , and energy flux densities, Bull. Amer. Meteorol. Soc., 82, 2415-2434, https://doi.org/DOI 10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2, 2001.

Goulden, M. L., Miller, S. D., Rocha, H. D., Menton, M. C., de Freitas, H. C., Figueira, A., and de Sousa, C.: Diel and seasonal patterns of tropical forest co 2 exchange, Ecol. Appl., 14, S42-S54, https://doi.org/DOI 10.1890/02-6008, 2004.

Keenan, T. F. and Williams, C. A.: The terrestrial carbon sink, Annu. Rev. Environ. Resour, 43, 219-243, https://doi.org/10.1146/annurev-environ-102017-030204, 2018.

Khan, A. M., Stoy, P. C., Joiner, J., Baldocchi, D., Verfaillie, J., Chen, M., and Otkin, J. A.: The diurnal dynamics of gross primary productivity using observations from the advanced baseline imager on the geostationary operational environmental satellite-r series at an oak savanna ecosystem, J. Geophys. Res.-Biogeosci., 127, 28, https://doi.org/10.1029/2021JG006701, 2022.

Lin, C., Gentine, P., Frankenberg, C., Zhou, S., Kennedy, D., and Li, X.: Evaluation and mechanism exploration of the diurnal hysteresis of ecosystem fluxes, Agric. For. Meteorol., 278, 14, https://doi.org/10.1016/j.agrformet.2019.107642, 2019.

Schimel, D. S., House, J. I., Hibbard, K. A., Bousquet, P., Ciais, P., Peylin, P., Braswell, B. H., Apps, M. J., Baker, D., Bondeau, A., Canadell, J., Churkina, G., Cramer, W., Denning, A. S., Field, C. B., Friedlingstein, P., Goodale, C., Heimann, M., Houghton, R. A., Melillo, J. M., Moore, B., Murdiyarso, D., Noble, I., Pacala, S. W., Prentice, I. C., Raupach, M. R., Rayner, P. J., Scholes, R. J., Steffen, W. L., and Wirth, C.: Recent patterns and mechanisms of carbon exchange by terrestrial ecosystems, Nature, 414, 169-172, https://doi.org/10.1038/35102500, 2001.

Urbanski, S., Barford, C., Wofsy, S., Kucharik, C., Pyle, E., Budney, J., Mckain, K., Fitzjarrald, D., Czikowsky, M., and Munger, J. W.: Factors controlling co2 exchange on timescales from hourly to decadal at harvard forest, Journal of Geophysical Research: Biogeosciences, 112, n/a-n/a, https://doi.org/10.1029/2006JG000293, 2007.

3. The PCMCI is interesting but how does this compare to KGML?

Response: PCMCI can be viewed as related to the general concept of KGML, but it differs in how the "knowledge" is obtained. While KGML often incorporates prior process understanding or constraints, PCMCI identifies causal relationships directly from observational data and uses these data-driven causal constraints to guide the model. We have now clarified this relationship in the Introduction (Lines 105–108):

"This concept is closely related to knowledge-guided machine learning (KGML), which incorporates prior process understanding or physical constraints into data-driven models (Ranjbar et al., 2026). PCMCI identifies causal relationships directly from observational data without relying on predefined process equations, thereby providing data-driven causal constraints that can guide machine learning models and improve their robustness."

Reference:

Ranjbar, S., Desai, A. R., Hoffman, S., Zahn, E., Bou Zeid, E., and Stoy, P. C.: Constrained carbon partitioning: a self-trained physics-informed machine learning model refines gpp estimates from eddy covariance measurements, Glob. Change Biol., 32, e70886, https://doi.org/10.1111/gcb.70886, 2026.

4. this paragraph makes a great point, but what metric will be introduced instead? I've always wondered why people don't just use Nash-Sutcliffe modeling efficiency, and there would be a whole world of time and frequency (e.g. wavelet)-based approaches for choosing a superior metric.

Response: Thank you for recognizing this point. We agree that metrics such as Nash-Sutcliffe efficiency are useful for evaluating overall model performance. However, our objective here is to assess whether hourly GPP products can capture ecologically meaningful diurnal photosynthetic dynamics, rather than merely optimizing overall fit. To this end, we adopted diurnal metrics that are consistent with the existing literature on diurnal photosynthetic patterns, specifically the diurnal GPP centroid, which indicates the timing of peak photosynthesis and the morning/afternoon asymmetry, and the relative midday depression (RMD), which quantifies the intensity of midday photosynthetic depression. These metrics have been widely used in recent studies on diurnal GPP dynamics (Khan et al., 2022; Li et al., 2023; Lin et al., 2019; Liu et al., 2024; Wilson et al., 2003; Zhu and Zhu, 2025) and provide direct physiological interpretability. We have now explicitly introduced these metrics in the revised paragraph (Lines 114–121):

"Therefore, beyond conventional accuracy metrics, hourly GPP products should also be evaluated for their ability to reproduce key features of the diurnal photosynthetic cycle. Representative indicators include the diurnal GPP centroid, which reflects the temporal distribution of daily carbon uptake between morning and afternoon (Wilson et al., 2003), and relative midday depression (RMD), which captures the magnitude of midday photosynthetic suppression (Zhu and Zhu, 2025). Recent studies have revealed distinct diurnal features of vegetation photosynthesis, including global-scale morning–afternoon asymmetry (Khan et al., 2022; Lin et al., 2019; Liu et al., 2024), pronounced midday depression in mangrove ecosystems (Zhu and Zhu, 2025), and heatwave-induced advances of daily photosynthetic peak by 2–3 hours in dryland regions (Li et al., 2023)."

Reference:

Khan, A. M., Stoy, P. C., Joiner, J., Baldocchi, D., Verfaillie, J., Chen, M., and Otkin, J. A.: The diurnal dynamics of gross primary productivity using observations from the advanced baseline imager on the geostationary operational environmental satellite-r series at an oak savanna ecosystem,

J. Geophys. Res.-Biogeosci., 127, 28, https://doi.org/10.1029/2021JG006701, 2022.

Li, X., Ryu, Y., Xiao, J., Dechant, B., Liu, J., Li, B., Jeong, S., and Gentine, P.: New-generation geostationary satellite reveals widespread midday depression in dryland photosynthesis during 2020 western US heatwave, Sci. Adv., 9, 14, https://doi.org/10.1126/sciadv.adi0775, 2023.

Lin, C., Gentine, P., Frankenberg, C., Zhou, S., Kennedy, D., and Li, X.: Evaluation and mechanism exploration of the diurnal hysteresis of ecosystem fluxes, Agric. For. Meteorol., 278, 14, https://doi.org/10.1016/j.agrformet.2019.107642, 2019.

Liu, Y., Penuelas, J., Cescatti, A., Zhang, Y., and Zhang, Z.: Atmospheric dryness dominates afternoon depression of global terrestrial photosynthesis, Geophys. Res. Lett., 51, 12, https://doi.org/10.1029/2024GL110954, 2024.

Wilson, K. B., Baldocchi, D., Falge, E., Aubinet, M., Berbigier, P., Bernhofer, C., Dolman, H., Field, C., Goldstein, A., Granier, A., Hollinger, D., Katul, G., Law, B. E., Meyers, T., Moncrieff, J., Monson, R., Tenhunen, J., Valentini, R., Verma, S., and Wofsy, S.: Diurnal centroid of ecosystem energy and carbon fluxes at FLUXNET sites - art. No. 4664, J. Geophys. Res.-Atmos., 108, 13, https://doi.org/10.1029/2001JD001349, 2003.

Zhu, Z. and Zhu, X.: Increasing midday depression of mangrove photosynthesis with heat and drought stresses, Agric. For. Meteorol., 362, 9, https://doi.org/10.1016/j.agrformet.2024.110372, 2025.

5. Note Khan et al. (2022) here. The introduction is generally well reasoned but would benefit from a broader suite of references from the multiple groups who are working on these challenges, as well as the foundational papers only very briefly alluded to above.

Response: We have added Khan et al. (2022) and Lin et al. (2019) at this point in Line 123. In addition, as noted in our response to Comment 2, we have also revised the Introduction to include more earlier and foundational references.

Lines 121–122:

"Recent studies have revealed distinct diurnal features of vegetation photosynthesis, including global-scale morning–afternoon asymmetry (Khan et al., 2022; Lin et al., 2019; Liu et al., 2024), …"

6. this pixel/footprint matching requires more detail, especially given all the work that's gone into this topic by Chu et al. (2021, Ag. For. Met.) and others.

Response: We have added more details on the pixel/footprint matching procedure in Lines 147–154.

"Given that the footprint of EC observations differs from the 0.05° grid of the upscaled product, we applied additional filtering to reduce scale mismatch (Chu et al., 2021). We first used the MODIS land cover product (MCD12Q1 Version 6.1 with 500 m resolution,

covering 2001–2022) and reclassified its vegetation types following the scheme in Table S1 to ensure consistency with the flux-site classification. For each flux site, we identified the corresponding 0.05° grid cell containing the tower location and calculated the fractional coverage of each reclassified vegetation type from all 500 m MODIS pixels within that grid cell for each year. The vegetation type with the largest fractional coverage was defined as the dominant class of that 0.05° grid cell."

7. Figure 1 is nice, would benefit from larger text in some of the subplots.

Response: Thank you for the positive comment on Fig. 1. We have polished the figure by increasing the font size in the relevant subplots.

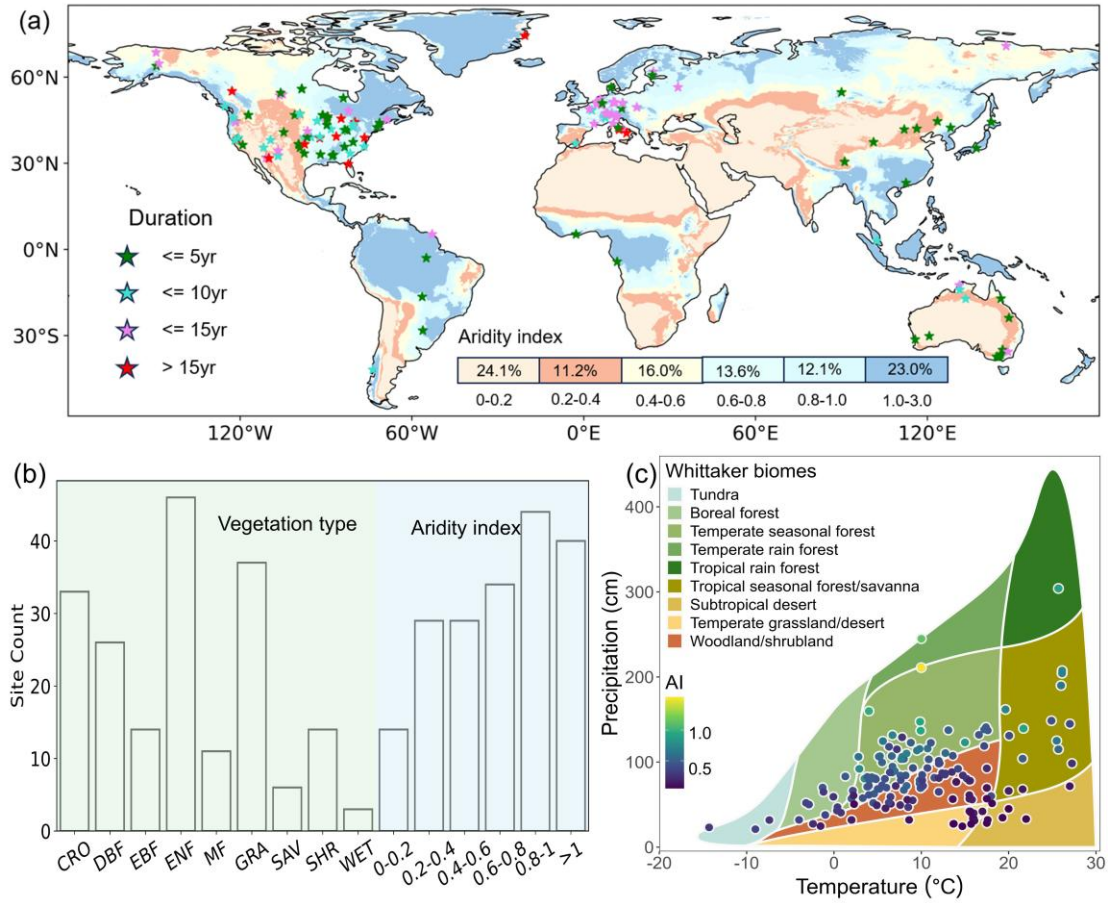

Figure 1. (a) Spatial distribution of the 190 eddy covariance flux sites used in this study. The background map shows the global aridity index, and the color bar indicates the proportion of land area within each aridity class. Sites with different observation durations are marked by star symbols in different 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).

8. just say 2 m. Probably too technical a point but using non-breaking spaces on 164,

165 & elsewhere between mathematical characters and values would be an improvement, also a subscript on the 2 in CO2 around here. Throughout, there are a number of minor usage issues that just require a careful read to check.

Response: Thank you for this suggestion. We have carefully revised these formatting issues throughout the relevant text. The revisions have been made as follows:

Lines 173–181:

"Based on 2 m air temperature and 2 m dew point temperature, we calculated hourly vapor pressure deficit (VPD) to quantify atmospheric dryness relevant for photosynthetic regulation (see Text S2 for details). The aridity index was obtained from the Global-AI_PET_v3 dataset and defined as mean annual precipitation divided by potential evapotranspiration over the climatological baseline, thereby characterizing long-term water availability across ecosystems (Zomer et al., 2022). Regions with Aridity < 0.65 were classified as drylands, and all others as non-drylands (Koppa et al., 2024). 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."

In addition, we further checked the manuscript and corrected several minor typographical issues, including changing "TableS2" to "Table S2" in Lines 221–222 and adding parentheses for panels (c) and (d) in the Fig. 11 caption in Lines 490–491.

Lines 221–222:

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

Lines 490–491:

"Figure 11. Global map of photosynthetic diurnal metrics derived from EGO GPP averaged over the growing seasons from 2000 to 2022. (a) RMD, (b) ΔGPP, (c) Centroid, and (d) Peak time."

9. I agree in principle with this approach but would be curious to know if a simple LSTM would fare much more poorly if the challenge is to incorporate causal inference,

or, alternately, if XGBoost alone gives a reasonable depiction of instantaneous flux as it often does. I don't question the CKML approach, rather I wonder if it's the only thing that might beat XGBoost here because a number of products including FLUXCOM-X, CASS, ALIVE, and many studies (that didn't give their model an acronym, e.g. https://doi.org/10.3390/land14010124) have arrived at XGBoost. But without benchmarking against XGBoost alone, the CKML-XGboost model's improvements can't be quantified.

Response: Thank you for this important comment. We agree that it is necessary to directly compare CKML-GPP with standard XGBoost and other commonly used models for flux upscaling. Therefore, we added a benchmark experiment using three conventional machine-learning models: XGBoost, Random Forest, and LSTM. The experimental design is described 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."

The results show that standard XGBoost performed best among the baseline models, confirming that XGBoost alone is already a strong approach for flux upscaling. CKML-GPP further improved the performance over XGBoost alone, increasing R^2 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."

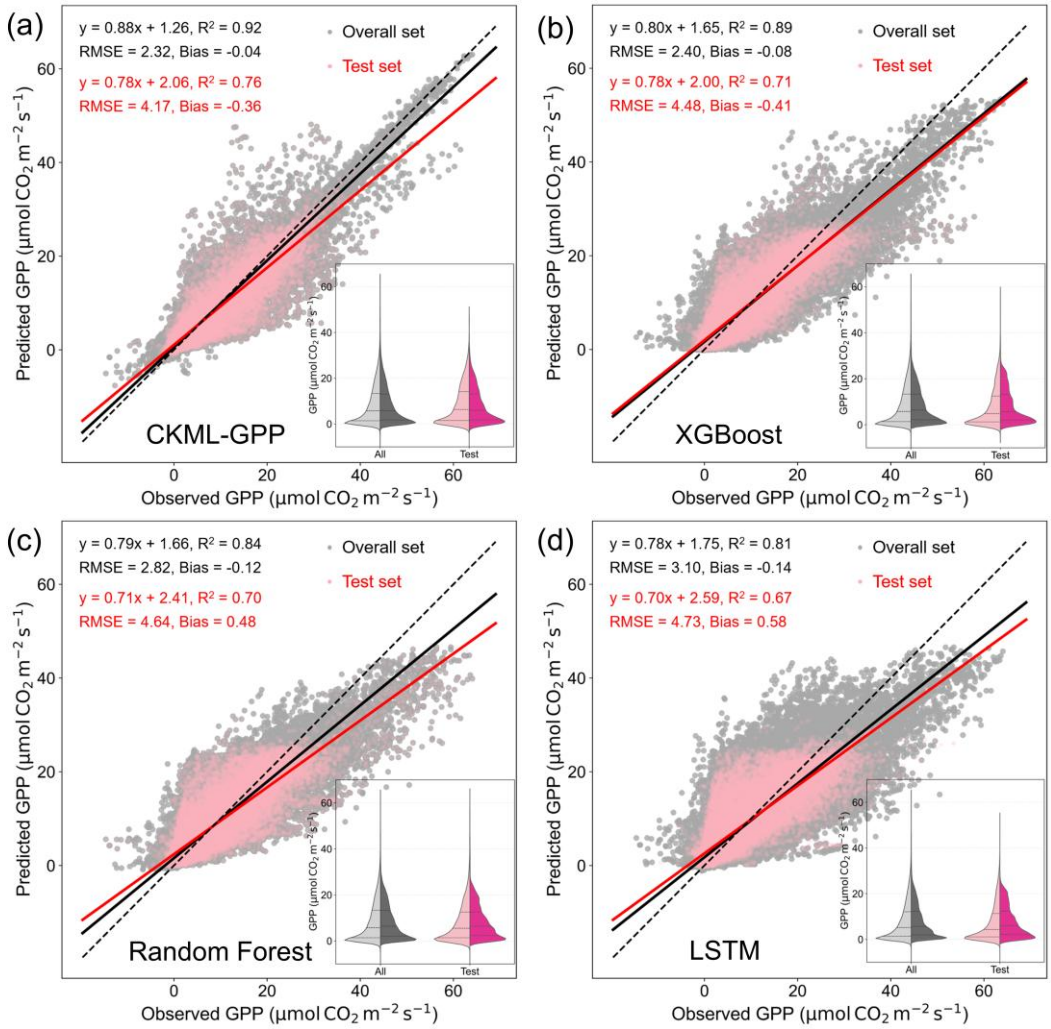

Figure S3. Comparison of the accuracy of CKML-GPP and other baseline models. (a-d) represent the predictive performance of CKML-GPP, XGBoost, RF and LSTM, respectively. In each panel, the black solid line and black text indicate the regression line and corresponding performance metrics for all samples, 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.

10. were any sites fully held out for the training testing split (commonly 70:30 or 80:20 or similar) or was the fold approach adopted alone? I'm wondering if this isn't a full fair comparison against FLUXCOM; how does the model perform against data from sites that it has never seen before?

Response: Thank you for raising this point. As noted in the revised Methods (Lines 257–263, see below), we did not use a random record-level split. Instead, we adopted a 10-fold site-wise cross-validation, where entire sites, not individual records were held out as independent test sets. In each fold, nine site subsets were used for training and

the remaining one was fully excluded from training. This was rotated so that each site served as a test site exactly once (Fig. S5). Thus, the CKML GPP model was tested on sites it had never seen during training. We have added Fig. S5 to illustrate the workflow.

"We adopted a 10-fold site-wise cross-validation strategy to train and evaluate the CKML-GPP model. Specifically, the available sites were randomly partitioned into 10 distinct subsets. In each iteration, nine subsets served as the training set, while the remaining one was used for testing. This procedure was rotated 10 times to ensure that every site functioned as an independent test sample exactly once, allowing us to compile a comprehensive validation dataset covering all sites (Fig. S5). Furthermore, to mitigate the impact of random initialization and ensure result stability, we performed 20 independent runs with different random seeds for each fold. The final prediction for each test sample was derived from the ensemble average of these 20 runs."

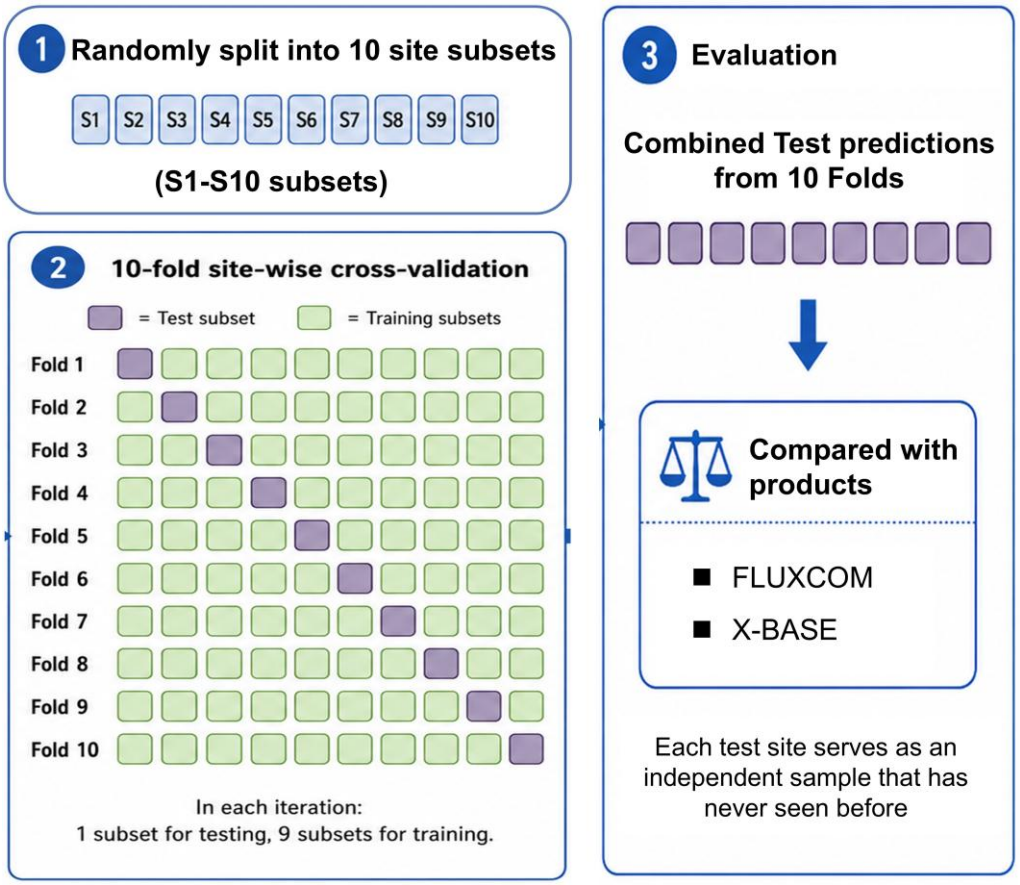

Figure S5. A visual explanation of the 10-fold site-wise cross-validation and product comparison.

To ensure a rigorous comparison, we clarified (Lines 264–269, see below) that all evaluations were conducted strictly on out-of-sample predictions from the combined test folds, where each site was excluded from training before being predicted (Fig. S5). This validation dataset was then used for a fair comparison between CKML-GPP and existing hourly GPP products, including FLUXCOM and X-BASE.

"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."

11. 3.2.3 (really before): how was flux data quality considered?

Response: We have described the flux data quality control procedure in the Methods section (Lines 165–168).

"To ensure data quality for model training and evaluation, we applied the following procedures to the tower-based GPP (GPP_NT_VUT): (1) removed negative values that are physiologically unrealistic; (2) retained only records with quality-control flags equal to 0 or 1; and (3) converted all timestamps to local solar time (LST) to remove phase shifts related to solar position and ensure physical consistency of diurnal variations (see Text S1 for details)"

12. the relationship between GPP and radiation isn't linear, it saturates.

Response: We agree that the GPP–radiation relationship is not linear but saturates. We have therefore replaced our original metric (MDI, midday depression intensity) with the published RMD (relative midday depression) metric (Zhu and Zhu, 2025) to quantify the intensity of midday photosynthetic depression.

Our primary focus in this study is on developing hourly GPP products and rigorously validating their accuracy at the hourly timescale. RMD (or MDI) is simply one of four metrics we examine for diurnal behavior, along with ΔGPP, centroid and peak time. Therefore, we think it is more appropriate to adopt an existing metric rather than proposing a new one. The switch to RMD did not change the main evaluation results or conclusions, and we have updated the relevant text and figures accordingly. As shown in the Figure.R1 below, EGO (red) remains the closest to FLUXNET observations (black) across vegetation types when either MDI or RMD is used, indicating that EGO provides the most consistent representation of observed diurnal GPP dynamics among the evaluated products.

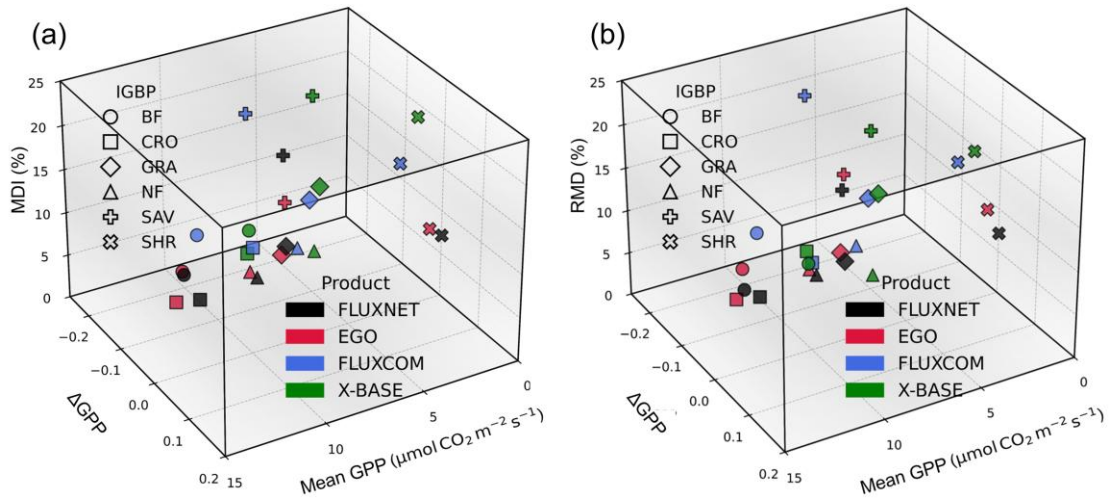

Figure R1. Similar with Fig. S8b, it presents the performance of products in a three-dimensional metric space, where the X, Y, and Z axes represent Mean GPP, ΔGPP, and (a) RMD / (b) MDI, respectively. Different vegetation types are marked by distinct symbols, and products are distinguished by colour. A smaller distance between a product and the observed point in this space indicates a closer match to the observed diurnal GPP dynamics.

13. 320 and beyond: I'd be curious about the time dependencies of these variables as soil moisture should increase in importance at longer time scales and at annual timescales temperature will likely be more important following the findings of Jung et al. This is perhaps a different topic though.

Response: Thank you for this insightful comment. We agree that the importance of environmental drivers may depend strongly on the temporal scale considered. Here, our SHAP analysis was mainly used to show the relative contribution of the input features during model development. Following your suggestion, we have added a discussion in Lines 600–604 to clarify this point. We also highlight that EGO provides a useful data basis for exploring these time-scale-dependent controls in future applications.

"It should be noted that the dominant controls on GPP may vary with temporal aggregation. For example, soil moisture effects can become more important at longer time scales because of ecosystem water-memory effects, whereas temperature may play a stronger role in regulating annual-scale carbon uptake (Green et al., 2019; Jung et al., 2017). Here, EGO provides a useful opportunity for future studies to examine how the relative controls of radiation, temperature, and soil moisture shift from hourly to seasonal and annual time scales."

Reference:

Green, J. K., Seneviratne, S. I., Berg, A. M., Findell, K. L., Hagemann, S., Lawrence, D. M., and Gentine, P.: Large influence of soil moisture on long-term terrestrial carbon uptake, Nature (London), 565, 476-479, https://doi.org/10.1038/s41586-018-0848-x, 2019.

Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlstrom, A., Arneth, A.,

Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Ain, A. K. J., Kato, E., Papale, D., Poulter, B., Raduly, B., Rodenbeck, C., Tramontana, G., Viovy, N., Wang, Y., Weber, U., Zaehle, S., and Zeng, N.: Compensatory water effects link yearly global land co 2 sink changes to temperature, Nature, 541, 516-520, https://doi.org/10.1038/nature20780, 2017.

14. Fig. 6 and related comparisons: as noted above can we be sure that the FLUXCOM products are treated to a fair comparison in these analyses?

Response: Thank you for this point. As we outlined in our response to Comment 10, the comparison with existing products was based exclusively on predictions from the combined test folds of CKML GPP (Lines 264–269). In this setup, every evaluation record came from a site that was completely withheld during training. We hope this clarifies that the comparison with FLUXCOM and X-BASE is indeed a fair one.

15. For Figure 12, what time period is the reference? Also, this is more than the US, also a good chunk of Canada and Mexico.

Response: The reference period is the multi-year June mean for all available years (2000–2022) excluding 2021, and we have now clarified this in the figure caption. We also note that the mapped region extends beyond the CONUS and includes parts of Canada and Mexico; we have therefore revised the wording from "the U.S." to "central North America" accordingly.

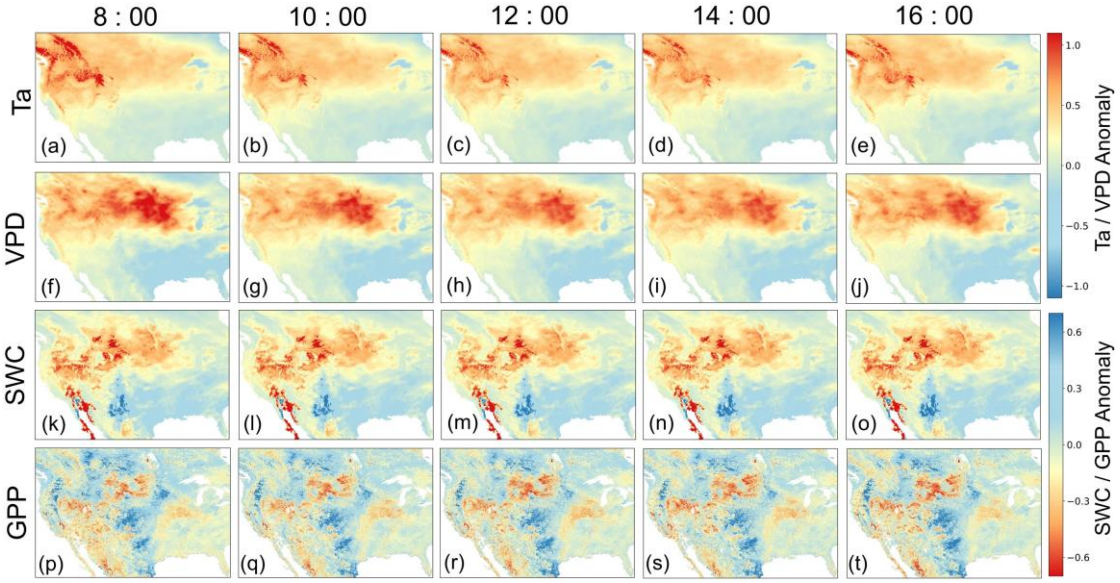

Figure 12. Spatial distribution of (a-e) Ta, (f-j) VPD, (k-o) SWC, and (p-t) GPP anomalies in the central North America during June 2021. For instance, GPP Anomaly = (GPP_{2021} – $GPP_{reference}$) / $GPP_{reference}$. $GPP_{reference}$ denotes the multi-year average for June excluding 2021.