Response to Anonymous Referee #2 (https://doi.org/10.5194/gmd-2024-169-RC3)

The authors apply FLAML v2.3.3---an automated machine-learning toolkit---to predict gross primary productivity (GPP) across 20 eddy-covariance sites, which is a less interesting and less novel endeavor. The manuscript would benefit from a more sharply defined research question and a deeper interrogation of the ecological processes underlying the model's performance. In particular, the authors should clarify what novel scientific insight they seek --- beyond demonstrating sensitivities --- and explore how specific feature groups/selections inform mechanistic understanding rather than merely reflecting data redundancy and uncertainty. Given these substantive concerns about framing and ecological interpretation, I respectfully decline to continue with further review, if that is the case.

We sincerely appreciate your valuable time and insightful comments, which have significantly helped us improve the quality and clarity of our manuscript. In the revised version, we have carefully addressed all the issues you raised. Specifically, we have thoroughly revised the structure and content of the manuscript, resulting in substantial modifications—nearly a thousand changes were made throughout the document.

We believe that these revisions have greatly strengthened the overall presentation and scientific value of our work. Below, we provide a detailed point-by-point response to each of your comments.

General Comments

Q1. The whole work reads more like a sensitivity report than an ecological modeling study. What specific scientific insight are the authors seeking by comparing FLAML to not scientifically different feature groups?

Thank you for your thoughtful and constructive comment. We have further clarified the scientific rationale and objectives of our study in the Introduction section of the manuscript. Our study aims to bridge the gap between process-based ecological modeling and data-driven approaches by integrating domain-specific knowledge from LUE models with the automated and efficient learning capabilities of FLAML. The resulting FLAML-LUE framework is a knowledge-guided machine learning model designed to address key ecological questions related to the estimation of GPP.

Specifically, our scientific insights are centered on the following (**Line 122-131**):

- To evaluate the performance of models using different combinations of LUE-related variables, such as absorbed PAR (fPAR) and water stress factors, across multiple vegetation types and time scales.
- To investigate model robustness under extreme climatic conditions, including high temperatures, elevated vapor pressure deficits (VPD), and drought. By evaluating model stability under these stressors, we aim to assess the resilience and reliability of GPP estimation frameworks in the face of climate variability and change.

The ultimate objective is to identify optimal input combinations for FLAML-LUE

models tailored to different vegetation types and climate zones across China. This helps enhance regional-scale GPP estimation accuracy, which is crucial for carbon budget assessments and ecosystem management.

Q2. The main text suggests a "FLAML-LUE model", yet none of the analyses explicitly implement light-use-efficiency (LUE) theory. Instead, all results derive from various tree-based regressors. If the intent is to compare FLAML-derived machine-learning models against LUE theory, the authors should at least incorporate an explicit LUE model.

Thank you for your thoughtful comment. We have further clarified the structural framework of the FLAML-LUE model in Section 2.3.3 of the manuscript (**Lines 122 and 272**). In this study, the term "FLAML-LUE" does not refer to a direct implementation of a mechanistic light-use efficiency (LUE) model. Rather, it reflects a hybrid modeling strategy where we incorporate key explanatory variables that originate from LUE theory—such as absorbed photosynthetically active radiation (fPAR), light-use efficiency modifiers, and environmental stress indicators (e.g., VPD, temperature, and water stress indices) — into an automated machine learning framework (FLAML). These variables represent the core components influencing vegetation productivity in traditional LUE models.

$$GPP = f\ (PAR, T, fPAR, W_j, VT, Season) \quad (3)$$

where, the $fPAR$ include EVI, NDVI, and LAI; W_j denotes moisture factors including LSWI, EF, SW, PDSI, Pre, RH; VT represents vegetation types, in which forest ecosystems include: EBF, DBF, NF, MF, and SAV; grassland ecosystems include GRA, MEA, and SHR, and farmland ecosystems include SC and DC; $Season$ represents the season in which the original data were acquired.

Our goal was to combine domain knowledge from LUE theory with the flexibility and efficiency of data-driven models. While we do not simulate GPP using a process-based LUE equation, the LUE-related predictors guide the learning process of the machine learning models, enabling a knowledge-informed estimation of GPP across different vegetation types and environmental conditions.

Q3. The model groups differ mainly in dryness index definition, data source or temporal averaging (e.g., PDSI vs. evaporative fraction, flux - tower vs. ERA5-Land temperature, actually Ta_flux is typically gapfilled by ERA5). These inputs often carry overlapping information, so comparisons may reflect data uncertainty or scale mismatches rather than mechanistic differences. Exploring a truly critical predictor --- such as soil moisture --- could strengthen the ecological relevance and offer interesting insights. A basic clarification to mention here is that ERA5-Land is a reanalysis dataset rather than a remote sensing product, and it should not be confused with ERA5. ERA5Land provides hourly rather than daily data.

Thank you for your valuable suggestion. We have addressed both issues you raised with corresponding revisions.

First, based on your comments, we have revised the selection of input variables used

in the model construction process (see **Table 1**). Following this adjustment, we re-train the models and re-evaluated the results accordingly. Specifically, to ensure consistency and reliability across all 18 variable combinations, we standardized the sources of temperature and PAR data by uniformly adopting ERA5-Land products. Additionally, we removed the PDSI dataset from our analysis because it is only available at a monthly temporal resolution, which is inconsistent with the finer time scales of other datasets used in this study. Instead, we carefully selected variables that more accurately capture vegetation moisture constraints from multiple ecological perspectives: atmospheric moisture stress (e.g., relative humidity and precipitation), vegetation-level moisture stress (e.g., LSWI and EF), and soil moisture limitations (e.g., SW). These choices are grounded in ecological theory and supported by previous research (Chang et al., 2023)。

Table 1

Input variable combinations of fPAR and water stress indicators.

Group	Input variables	Group	Input variables	Group	Input variables
FLAML00	NDVI, LSWI	FLAML10	EVI, LSWI	FLAML20	LAI, LSWI
FLAML01	NDVI, EF	FLAML11	EVI, EF	FLAML21	LAI, EF
FLAML02	NDVI, SW	FLAML12	EVI, SW	FLAML22	LAI, SW
FLAML03	NDVI, VPD	FLAML13	EVI, VPD	FLAML23	LAI, VPD
FLAML04	NDVI, Pre	FLAML14	EVI, Pre	FLAML24	LAI, Pre
FLAML05	NDVI, RH	FLAML15	EVI, RH	FLAML25	LAI, RH

Regarding the second issue you mentioned about the description of the ERA5-Land dataset, we have made corresponding revisions in the updated manuscript. Specifically, **Section 2.2.3** now reads as follows: "ERA5-Land (Hersbach et al., 2020) is a global high-resolution reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) under the Copernicus Climate Change Service (C3S). It provides hourly land surface variables at a spatial resolution of 0.1°, generated using a dedicated land surface model driven by the ERA5 climate reanalysis. The dataset integrates advanced land surface modeling and data assimilation techniques, offering a wide range of variables such as air temperature, soil moisture, precipitation, and snow depth. In this study, site-specific variables including air temperature (T), soil water content (SW), precipitation (Pre), and leaf area index (LAI) were extracted from ERA5-Land. In addition, photosynthetically active radiation (PAR), evapotranspiration fraction (EF), VPD and relative humidity (RH) were calculated and derived from available ERA5-Land variables using GEE."

Once again, thank you for your insightful feedback. Your suggestions have significantly contributed to improving the depth and rigor of our study. We will continue to build on this work and aim to present our findings more comprehensively in future research.

Q4. The rationale for analyzing 8-day, 16-day vs. monthly statistics is not fully

developed. Because GPP seasonality dominates many signals, the differences in model performance may simply reflect sample size (it is unsurprising that monthly R2 exceed those at the 8-day scale, and this comparison offers no insight).

Thank you for your insightful comment. We agree that the seasonal dynamics of GPP and the differences in sample sizes across temporal scales (e.g., 8-day, 16-day, monthly) can inherently influence model performance metrics such as R^2. However, our rationale for analyzing multiple temporal resolutions goes beyond statistical comparisons.

The primary objective of incorporating different temporal scales is to evaluate the robustness and generalizability of the FLAML-LUE model across varying degrees of temporal aggregation. As indicated in the revised manuscript (**Line 464 - 467**), compared to the daily scale, the nuRMSE decreases by 12.97%, 16.52%, and 25.92% at the 8-day, 16-day, and monthly scales, respectively. This highlights that the uncertainty of the FLAML-LUE model is significantly reduced at coarser temporal resolutions.

Furthermore, from an application perspective, transitioning from site-level to regional-scale GPP estimation across China requires temporal resolutions that align with commonly used satellite products. In this context, 8-day or monthly models are more practical, as they not only reduce noise through temporal aggregation but also ensure greater consistency with large-scale remote sensing data. These coarser time scales offer a more effective trade-off between capturing ecological dynamics and enabling broader spatial applicability.

Q5. Presenting each PFT group in separate sections can make cross‐comparison cumbersome. I suggest grouping figures by PFT (forest, grassland, cropland) with sub-panels for each site or model variant.

Thank you for your helpful suggestion.

We agree that organizing the figures by plant functional type (PFT)—such as forest, grassland, and cropland—can improve clarity and facilitate more effective cross-comparisons. In response to your comment, we have revised the relevant figures accordingly, grouping them by PFT with sub-panels representing individual sites or model variants. This required a substantial amount of work, as it involved reprocessing the results and essentially rewriting this section of the manuscript. Nonetheless, we believe this reorganization enhances both the readability and interpretability of the results. We sincerely appreciate your constructive feedback.

Q6. The manuscript contains kind of repeated descriptions across all sessions. I recommend restructuring the whole manuscript thoroughly to avoid duplication. Meanwhile, the authors claim that all results are from validation, but without describing the split strategies.

Thank you for your valuable comment. In response, we have thoroughly restructured the manuscript to reduce redundancy and improve overall clarity. Repetitive descriptions across sections have been removed or streamlined to avoid duplication and enhance readability.

Additionally, we have now clearly described the dataset split strategy in **Section 2.3.1** of the revised manuscript. Specifically, the pre-processed dataset was divided into training and testing sets using the Blocked Time Series Split strategy. Given the temporal dependency of the data, standard cross-validation is not suitable for time series analysis (Reichstein et al., 2019). Instead, a block-based and non-continuous split is applied to preserve the temporal structure. In this approach, the time series is partitioned into several non-overlapping continuous training blocks (e.g., 2003-2005, 2007-2009, 2011-2013, 2015-2017, 2019-2021), with independent years reserved as the validation set following each training block (e.g., 2006, 2010, 2014, 2018, 2022). This strategy ensures that the temporal order is maintained, preventing future data from leaking into the training process and thus avoiding invalid predictions. Additionally, the method incorporates validation over multiple periods, enabling the assessment of model generalization across different climate conditions, which is crucial for evaluating the model's robustness under varying environmental scenarios.

Reference

Chang, X., Xing, Y., Gong, W., Yang, C., Guo, Z., Wang, D., Wang, J., Yang, H., Xue, G., Yang, S., 2023. Evaluating gross primary productivity over 9 ChinaFlux sites based on random forest regression models, remote sensing, and eddy covariance data. Sci. Total Environ. 875, 162601. https://doi.org/10.1016/j.scitotenv.2023.162601

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204. https://doi.org/10.1038/s41586-019-0912-1