Responses to community comment #1

Gaber et al. present a comprehensive evaluation of using automated ML (AutoML) to estimate and upscale ecosystem GPP using four sets of remote sensing and reanalysis products. The comparative analysis of three AutoML frameworks reveals that AutoSklearn consistently outperforms the other frameworks and a baseline Random Forest model in reproducing spatial patterns, temporal variability, and trends in the observed GPP. Notably, the use of higher-resolution remote sensing products further enhances model performance, attributed to footprint matching. Additionally, the authors have produced a global wall-to-wall map of GPP (monthly, 0.05 deg) using AutoSklearn and a suite of remote sensing predictors, which agrees well with two other ML-based global GPP products.

The study highlights the potential of AutoML in quantifying global GPP, capturing its temporal and spatial variability and trend, and provides insights into feature selection for monthly GPP estimation. This topic matches the interests of the readers of Biogeosciences. While the manuscript is exceptionally well-written and the implementation of ML models is robust, several notable concerns, particularly regarding model interpretability, feature selection, and sources of uncertainty, warrant additional exploration and discussion.

Response: Thank you, Dr. Jiangong Liu, for your valuable and constructive comments on the manuscript. Your suggestions are very helpful for us in improving the manuscript.

Major comments:

1. When comparing estimations derived from "RS" and "RS + meteo", and observing no substantial improvement in model performance with additional meteorological predictors, the assertion that this is because meteorological data contains no additional information or the reanalysis data quality is not good might need further exploration (Lines 435-440). Given that several predictors from "RS + meteo" might contain overlapping information on a monthly scale (e.g., VIs, LAI, SIF, ET, and meteorological data), it might be premature to conclude that the inclusion of meteorological data yields marginal enhancement in modeling monthly GPP.

Response: You are right that the predictors are likely to contain overlapping information at a monthly scale, and thus, the apparent results from comparing “RS” and “RS+meteo” potentially undermine the actual contribution of meteorological factors to GPP prediction. We aimed to interpret this result in the context of the overall model predictive performance measured by goodness-of-fit metrics. Thus, we will adapt the corresponding text and emphasize that the reanalysis data does not additionally improve the predictive accuracy since meteorological data largely contains overlapping information with the RS variables. We will further underscore that metrological conditions are themselves important controls of GPP in the context of literature.

2. I am puzzled by the decision to leave out radiation (BESS_Rad) in the 'RS meteo' (Figure 3) and curious about the thinking behind splitting data sources into remote sensing and reanalysis, instead of classifying them into physical (BESS_Rad, ESA CCI, MODIS LST, and ERA5-Land) and biological (MODIS VI/LAI, CSIF, and ALEXI ET) controls. Also, I think it would be worthwhile to discuss whether SIF should be included as a predictor since it is commonly used as a GPP proxy.

Response: BESS_Rad is part of the RS meteo variable set, as stated in Figure 3 (“Features of RS + ERA-5 Land”). We will clarify this point in the text. Splitting the data into physical and biological controls is an interesting approach and would certainly give another valuable angle at variable importance. However, it isn't easy to draw the boundaries between these categories (for instance, LST and soil

moisture are significantly influenced by biological controls). In this regard, we will perform an additional analysis to assess the feature importance of individual variables based on a permutation approach. We expect the result to comprehensively quantify the importance of variables and the relative contribution of physical and biological controls.

3. While the Discussion does touch on various potential sources of uncertainties (e.g., section 4.2), it seems to overlook the potential for bias inherent in the eddy covariance GPP. The authors used night-time partitioned GPP, relying quite a bit on a temperature dependency function of night-time NEE. But there is still some debate about whether this dependency is exponential (Chen et al., 2023), if it can be extrapolated to the daytime (Keenan et al., 2019), and whether it should be referenced to air or soil temperature (Wohlfahrt & Galvagno, 2017). Given that AutoML isn't the easiest to interpret (Line 330), I am wondering if its top-notch performance is partly because it is picking up on some error structures during NEE partitioning.

Response: Thank you for raising this relevant point. We will explain the origin of the GPP estimates and how they can affect prediction performance/uncertainty better in the text. Thank you also for providing the references, which we will consider in the text.

4. I am excited about a new global GPP product. Would the authors like to give it an official name, and give the name a spotlight in the Title or Abstract? Additionally, it is recommended that the authors articulate both the interannual variability and the annual magnitude of GPP relative to the new product, as such information would likely be invaluable to the flux community. I am also curious about why the authors did not use the high-resolution RS data (500 m) for the product, considering it seems to pull better performance.

Response: Our analysis focuses primarily on the benchmark of different AutoML frameworks. The upscaled maps were mainly used to verify the results of AutoML in comparison to benchmarking products. At this stage, the release of a new GPP dataset is not planned but could be considered in the future. 500m RS data improved the performance significantly; we used the 0.05 degree data for upscaling to compare with other upscaled datasets, which are typically at a similar or coarser resolution. In light of our result, the production of 500m-resolution data from upscaling is highly encouraged to improve accuracy and reduce uncertainties associated with scaling errors. We will clarify and highlight these aspects in the discussion section.

Minor comments

Line 90: Since negative outliers are in a unit of "gC m-2 d-1", did the authors aggregate daily values to monthly for both fluxes and their predictors? More details should be provided for the quality control.

Response: We used the monthly data provided by the original data sources, i.e., FLUXNET2015, AmeriFlux ONEFLUX, and ICOS. The monthly data has been aggregated from daily and half-hourly/hourly values. Outlier removal was performed on monthly data that corresponds to average daily NEE. We will provide clarification in the text.

Line 100: Add the source/reference for IGBP here, and also in Fig 2.

Response: We will include this in the text.

Line 115: It is a very minor point, but I think terminology for explanatory variables/predictor (e.g., Table 1)/feature (e.g., line 40) is used a bit random in the manuscript. Though they share the same meaning, readers might get confused.

Response: We will better align the terminology.

Line 130-140: It might be worthwhile to relocate this paragraph concerning the challenges with CASH to the Introduction to serve as an additional motivation statement. In the current Introduction, the authors highlighted the advantages of using AutoML, which are "... to overcome the challenges of algorithm selection, hyperparameter tuning, and pipeline creation through an automated approach". They introduced well the existing problem of feature selection. However, the knowledge gaps in the existing ML-based products of fluxes regarding algorithm selection and hyperparameter tuning should also be clarified.

Response: Thank you for raising this point. We will include the challenge of algorithm selection and hyperparameter tuning in current ML-based products in the introduction.

Line 255: Offering details about the calculation of trends, seasonality, across-site variability, and anomalies in the Methodology section, prior to Figure 10, might enhance comprehension. I am also unsure what R2 values mean for trend comparison, as trends are the fitted slopes.

Response: We will include a reference to 2.3.2 to clarify the calculation of trends, seasonality, across-site variability, and anomalies. The R2 for trends represents the spatial variability of their slopes. We will clarify that in the text.

Figure 7: what do R2 values smaller than -1 mean?

Response: We define R2 as the coefficient of determination that provides a measure of the proportion of variation that can be predicted from the predictor variables. Negative R2 values mean that the model performs worse than a simple model that just predicts the mean of the dependent variable. This definition of R2 aligns with the Nash- Nash–Sutcliffe model efficiency coefficient that is typically used in hydrological models, and it is commonly used as a metric for regression models in machine learning applications. We will provide more descriptions in the text to improve clarity.

Line 490: While the models also underestimate large GPP values (Line 305), further discussion on this aspect may provide additional insight.

Response: We will include further discussion/literature regarding this behavior.

Line 520: I appreciate the authors raising this point about the cautious use of AutoML. The inherently 'black-box' nature of AutoML, which presents challenges in interpretability as indicated (Line 330), is a notable issue.

Response: Thanks for this feedback!