Point-to-point reply

Referee #1

Referee comment	**Authors' response**	**Authors' changes**
Define 'feature importance'. Clearly state what you refer to with 'the importance of features' or 'the contribution of variables'. It seems you refer to the contribution to/ the importance for model accuracy, but this only becomes clear after having read the complete manuscript. In several instances it is not clear what is meant, it could also be the contribution to/the importance for the predicted GPP value itself, as opposed to its accuracy. Examples of instances with need for clarification are in paragraph l.49-64, l.210, 415.	Thank you for this suggestion. Yes, we refer indeed to the importance for model accuracy. We will include a clear definition of variable importance in the introduction of the manuscript. Furthermore, we will make sure that the terminology is consistent throughout the manuscript. We assume you are referring to l.241, l.491 and hope that our changes will resolve these clarity issues.	**l.50-53:** "The contribution of different explanatory variables, such as greenness measures, photosynthetically active radiation (PAR), land surface temperature (LST), soil moisture (SM), and meteorological variables (vapor pressure deficit, temperature, precipitation) to the accuracy of the GPP predictions (hereafter referred to as variable importance) has not been conclusively clarified." **l.245-247:** "The explanatory variable sets can provide information about the importance of the input features on the performance of the upscaling frameworks." **l.502-503:** "AutoML is a powerful approach for assessing the importance of the variables on model performance since it selects the optimal base models and constructs optimal pipelines independently for each feature set under consideration."
Do you perform any quality checks on the MODIS products before aggregating to monthly and 0.05deg? Data quality is another very important factor determining the accuracy of model results, so information whether and if so, how this has been handled in the work is necessary to report in the manuscript. How do you handle data gaps or low sample availability within a month?	We will reformulate the paragraph on quality checks and gap filling (l.134-142) to make it clearer and more understandable.	**l.142-152:** "We filtered the data for poor-quality pixels, performed gap-filling, and matched spatial and temporal resolutions. We used NBAR (MCD43C4 v006), where more than 75 % of high-resolution NBAR pixels were available from the full BRDF inversion. We selected LST data by applying the quality control mask and where the average emissivity error was less than 0.02. LAI and FPAR were used when retrieved using the main algorithm with or without saturation. Data gaps were filled at the native resolution, similar to the procedure of Walther et al. (2022). We filled gaps of less or equal five days (8 days for four-day resolution datasets) with the average of a fifteen-days moving window for high-frequency datasets (NBAR, LAI, FPAR, BESS_Rad, CSIF). We gap-filled LST with a 9-day moving window because we observed

		higher variations. For SM, we used the moving window median for short gaps and the mean seasonal cycle for long gaps. Finally, we resampled all datasets to 0.05 ° spatial resolution and monthly temporal resolution. Coarser-resolution datasets were resampled using a nearest neighbor approach, while high-resolution data was down-sampled using the conservative remapping method (Jones, 1999)."
I still wonder what potential consequences it has that many of the features included are model output themselves, partly driven by very similar remotely sensed data sets like in the feature set. Any speculations or justification?	The features used for the modeled datasets, such as CSIF and ET, largely overlap with the features used to model GPP in this study. CSIF draws from the MODIS NBAR product whereas ET is using MODIS LST, LAI, albedo, ASTER surface emissivity, land cover (Hansen), and a GTOPO DEM. Despite these overlaps, they incorporate information that is not provided to our model. The CSIF dataset is trained on OCO-2 SIF data, allowing their model to establish a functional relationship specifically between NBAR and SIF. The ET data, on the other hand, is modeled from a process-based model, which includes domain knowledge that may be challenging for our ML approach to learn. With ideal model capabilities and scalable training data, we would expect a redundancy of these input data since these relationships would be learned inherently by the ML models just from the MODIS input features. However, given that the ML models might not capture all relationships and that they are trained on	**l.134-140:** "Many of the explanatory variables are themselves datasets that have been modeled from MODIS data. For instance, SIF was predicted from MODIS NBAR using a feed-forward neural network, trained on OCO-2 SIF retrievals (Zhang et al., 2018). ET estimates were modeled by a coupled land-surface and atmospheric boundary layer model (Atmosphere Land Exchange Inverse, ALEXI), which used MODIS LST and LAI as inputs, among others (Hain and Anderson, 2017). Although their input data largely overlap with the inputs to our model, we expected additional improvements from including these datasets due to the domain knowledge of their models, which would otherwise be difficult to replicate in this study by solely relying on MODIS data and limited GPP measurements."

	limited GPP measurements (difference in scale), the datasets might provide additional information that are useful for the GPP modeling. The variable importance analysis provides further insights. We will also include a statement highlighting this matter.	
What is the reason for leaving out the vegetation indices from the RS and RSmeteo feature sets?	We will provide more clarity about the use of the VIs in the manuscript. While VIs slightly improved the performance of the RS minimal datasets, we could not detect any performance improvements in the other datasets. However, this is not one of our main findings in this study and might confuse the reader more than provide insights. For better understanding, we will not consider the VI variables as a separate variable set anymore and instead just mention the finding concerning the VIs directly in the text and in an additional table in the appendix.	**Tab. 1,2:** Removed the VIs **Fig. 3-6:** Removed RS minimal +VI **l.232-241:** ” We organized the explanatory variables into three sets to determine their impact on GPP predictions within different AutoML frameworks (Tramontana et al., 2016; Joiner and Yoshida, 2020). Each set consisted of different features that could explain the variation in GPP. The minimal set of remotely sensed variables (RS minimal) included surface reflectance from seven MODIS visible to infrared bands and PAR, which largely reflect the ability of the vegetation canopy to intercept solar radiation for photosynthesis. The "RS" set included all remotely sensed variables and their products. Notably, compared to the "RS minimal" set, the "RS" set also included land surface temperature, evapotranspiration, and soil moisture, which provide an additional link to vegetation heat and water stress (Green et al., 2022; Stocker et al., 2018). Finally, the "RS meteo" set included all remotely sensed variables and, in addition, meteorological variables from the ERA5-Land reanalysis (see Table 2). Additionally, we replaced the MODIS reflectance bands, LAI, FPAR, and land cover products with their native 500 m resolution data in the "RS" set to evaluate the impact of satellite data spatial resolution on GPP estimation.” **l.346-347: “**In addition, we evaluated whether vegetation indices (VI) could improve the performance of the variable sets, but no

		improvements were found beyond the "RS minimal" dataset (Tab. A1)." **Tab. A1:** Added an additional table of mean r2 values to the appendix for the repeated cross-validations, including the results with VIs.
Please clarify which data product versions were used for the ESA CCI soil moisture and for the Fluxcom (l.250 states Fluxcom v6, so does this refer to the Fluxcom RSonly data from MODIS c006?).	We used ESA CCI v06.1 (see Table 1) and FluxCom v6, RS only (see l.284). This refers to RS only from MODIS collection 006. We will highlight this in the text.	**l.287-289:** "We produced global GPP and standard error maps at a resolution of 0.05 ° in monthly frequency from 2001 to 2020, which we compared with the two ML-based reference datasets FluxCom v6 (RS only, based on data from the MODIS collection 6) (Jung et al., 2020) and FluxSat (Joiner and Yoshida, 2020)."
How was the R2 computed (http://www.jstor.org/stable/2683704)?	Thank you for raising this matter. Our R2 definition aligns with the Nash-Sutcliffe model efficiency (l.258), which corresponds to Eq. 1 in Kvalseth (1985). We will include an equation in the appendix.	**Eq. A1:** $r^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$
Figure 4 and 5: Do these distributions represent 30 (cross-validation rounds) R2 values computed across all sites, or 30 (cross-validation rounds) x 245 (sites) R2 values computed for each site and shown all together? For spatial variability I understand it is always 30, right? Similarly, the question on the grouping for the R2 in Fig.7.	Thank you for raising this question. The graphs show the distribution of the R2-values from the 30 repeated cross-validations (hence, a distribution of 30 R2-values, the first part of your question). Within each cross-validation, the R2 is calculated over the entire prediction, in which all sites are merged. This applies to total, trend, seasonality, anomalies, and also spatial variability. The only difference for spatial variability is, that we compare averages instead of temporal time series components. We will formulate the corresponding text passages more clear.	**l.302-304 (Caption fig. 4): "**Overall framework performance, expressed as the coefficient of determination (r^2) for the candidate frameworks and the three different explanatory variable sets. Each distribution belongs to one framework and one set of explanatory variables and results from the repeated cross-validations, for each of which one r^2 value is calculated over the predictions at all sites." **l.313-317 (Caption fig. 5): "**Evaluation of the temporally and spatially decomposed time series expressed as the coefficient of determination (r^2). Each distribution belongs to one framework and one set of explanatory variables and results from the repeated cross-validations, for each of which one r^2 value is calculated over the predictions at all sites. The r^2 values for seasonality and anomalies were calculated from seasonal cycles

		and anomalies at monthly granularity, while those for trend and across-site variability were calculated from one trend or mean value per site, respectively." **l.365-366 (Caption fig. 7): "**The distribution results from the repeated cross-validations, for each of which one r^2 value is calculated over the predictions at all sites."
l. 485-490: Would you expect more (dummy) training data or feature selection enabled to be more promising for higher model accuracy?	We expect higher robustness of the model predictions and the evaluation metrics from more and better-balanced training data.	**l.499-500:** "More training data with better geographic representation could help mitigate these shortcomings and could lead to more robust predictions, model evaluations, and potentially higher model performance."
l.14/15: These are some, but not all choices that affect the accuracy of the regression model.	We will highlight that these are just some aspects.	**l.13-15:** "However, the accuracy of the regression model can be affected by uncertainties introduced by model selection, parameterization, and choice of explanatory features, among others."
Discussion on light-use-efficiency is unclear to me given the monthly temporal scale of interest in this study. Line 428, 436 contrast instantaneous GPP reductions due to environmental stress. I suggest to rephrase this paragraph and give the (in my opinion) more reasonable explanation of the difference between site level and reanalysis meteorology more visibility.	Thank you for this suggestion. We are a bit unclear as of to what you refer to by l. 428 and l. 436. We assume you are referring to l. 507 and l. 515 and hope that our changes satisfactorily address your concerns. We will reformulate this paragraph to highlight the scale mismatch better as a possible explanation.	**l.522-532:** "Including the meteorological explanatory features (ERA5-Land) in the training data does not significantly improve the prediction quality for any of the frameworks. This implies that meteorological data may not contain additional information that the machine learning frameworks in this study can effectively use to predict GPP. A possible explanation could be the mismatch between reanalysis and site meteorology. The coarse resolution and large uncertainties of the reanalysis data may result in a poor representation of the flux tower footprints, which are often smaller than one pixel of the reanalysis data, leading to uncertainties in the modeling. For example, Joiner and Yoshida (2020) showed that using site-measured meteorological data instead of reanalyzed data significantly improved the performance of GPP predictions. At the monthly scale, the "RS" variable set may already encode information about the instantaneous environmental stress from adverse meteorological conditions through, for example, LST, ET, and soil moisture, which are important controls on GPP (Bloomfield et al., 2023). Further

		studies could potentially assess these uncertainties by comparing models trained with tower meteorological data to gridded reanalysis datasets.”
Discussion of spatial resolution l. 441-445: To me this paragraph suggests between the lines that training such models by pairing eddy-covariance data with 0.05 or 0.25 pixels as done in this work is the state-of-the-art. It is not. And this could become more clear. For some variables there is no other option because data are not available at finer spatial resolution, this is clear. But the authors chose to take the coarser pixels as the normal standard, and I suggest to make this difference between author choice and state-of-the-art clear.	We assume you refer to l.551-556. We will highlight that the spatial resolution in this study is a result of the authors’ choice and reformulate the paragraph less suggestive.	**l.563-565:** “These results underscore the importance of spatial resolution and suggest the use of data with a resolution that better represents smaller landscape features and flux tower footprints, in contrast to our initial choice of 0.05 ° resolution in this study. (Xiao et al., 2008; Yu et al., 2018; Chu et al., 2021).”

Referee #2

Referee comment	Authors’ response	Authors’ changes
Line 129: I think these variables could use a little more explanation plus citations. My reasoning here is, as I mentioned in my previous comments, that several of these variables (SIF and ET) are themselves modeled from remote sensing data, and I think that is important context for how they are interpreted. The SIF product used here, for example, is not truly a measured SIF signal but a modeled SIF based solely on surface reflectance (the NBAR product). In my opinion, it’s important to	We will include citations and refer to the variables’ modeled background. SIF is modeled from MODIS NBAR and OCO-2 SIF retrievals, the former of which is also an input to our GPP models. ET is modeled from MODIS day and nighttime LST, MODIS LAI, MODIS albedo, ASTER surface emissivity, land cover (Hansen), and a GTOPO DEM, hence overlapping in terms of LAI and LST. We will provide more context on these variables.	**l.134-140:** “Many of the explanatory variables are themselves datasets that have been modeled from MODIS data. For instance, SIF was predicted from MODIS NBAR using a feed-forward neural network, trained on OCO-2 SIF retrievals (Zhang et al., 2018). ET estimates were modeled by a coupled land-surface and atmospheric boundary layer model (Atmosphere Land Exchange Inverse, ALEXI), which used MODIS LST and LAI as inputs, among others (Hain and Anderson, 2017). Although their input data largely overlaps with the inputs to our model, we expected additional improvements from including these datasets due to the domain knowledge of their models, which would otherwise be difficult to

make that clearer here since I definitely think it affects the interpretation of its importance as a variable.		replicate in this study by solely relying on MODIS data and limited GPP measurements."
Line 137: I think just a little more detail about the resampling could be helpful. For products with finer resolutions than 0.05 degrees, were all pixels within the 0.05 degree cell averaged together? For those with coarser resolutions, how were they down-scaled to 0.05 degrees?	We performed the down sampling using the conservative remapping method, and the up sampling using nearest neighbor. We will include this in the manuscript.	**l.149-152:** "Finally, we resampled all datasets to 0.05 ° spatial resolution and monthly temporal resolution. Coarser-resolution datasets were resampled using a nearest neighbor approach, while high-resolution data was down-sampled using the conservative remapping method (Jones, 1999)."
Lines 331-332, 501-502, and 626-628: I think this is slightly overstating the improvement of the RS set over the RS-minimal and RS-minimal+VI sets. Per 331-332, the full RS set only added 2% variance explained, so I think it's too strong to say that the NBAR + PAR "did not provide the models with sufficient information." I think it would be more accurate to say that NBAR + PAR is responsible for the vast majority of model skill but the remaining variables can add some additional information on the margins.	Thank you for this valid point. We will rephrase the corresponding paragraph and make it less overstating.	**l.511-520:** "The frameworks' performance depends significantly on the choice of predictive features on which they are trained. The results show that while the seven NBAR bands and PAR from the "RS minimal" variable set provide the model with sufficient information for a GPP prediction, the full set of "RS" variables adds additional information that all the frameworks can exploit. The additional variables in the "RS" variable set, such as SIF, LAI, FPAR, ET, LST, SM, and plant function type, appear to include important environmental forcings and structural variables that provide a marginal advantage over the variables on only vegetation structure and radiation in "RS minimal" (Green et al., 2019; Stocker et al., 2019; Xu et al., 2020)." **l.633-637:** "We found that remotely sensed (RS) explanatory variables provided the best results in combination with the investigated frameworks. While only relying on the MODIS NBAR reflectance bands and PAR ("RS minimal") provided the models with sufficient information for GPP prediction, considering other proxies of photosynthetic activity and canopy structure, such as solar-induced fluorescence, leaf area index, and fraction of absorbed photosynthetic activity, increased the performance of all models."

Community comment #1

Referee comment	Authors’ response	Authors’ changes
I have carefully examined the authors' responses to the previous comments and the changes made in the revised manuscript. The authors have thoughtfully addressed the concerns raised, enhancing the quality of the work. I congratulate the authors on their diligent efforts and appreciate the opportunity to review this manuscript. Based on the revisions made, I recommend the manuscript be accepted for publication.	Thank you for your review. We appreciated your suggestions, which have greatly improved this manuscript.	