Responses to referee #1

Max Gaber and colleagues investigate the effect of several technical choices in the process of predicting GPP from eddy-covariance measurements and satellite(-derived) data sets using machine learning. The focus is on novel methods in the field of automated machine learning applied to predict monthly GPP at site level from different sets of predictor variables, as well as on the effect of their spatial resolution. The authors demonstrate the applicability of AutoML, and of AutoSklearn in particular, and show that in the global upscaled product, spatiotemporal patterns reasonably compare to other products. They also illustrate the importance of adequate spatial resolution of the predictor variables by increased model performance at site level, when part of the predictor variables are fed into the machine learning at 500m instead of at 0.05deg resolution.

Given the growing number of research studies that implement such data-driven approaches at global and regional scales (large part of whom are cited in the paper) and the still unquantified importance of several technical choices in the set-up, this study is timely and definitely of relevance. The fact that the overall R2 at site level is similar to or slightly higher than from a plane random forest or the results in comparable upscaling exercises at monthly scale (Jung et al. 2011) is interesting, and highlights that tuning the machine-learning set-up may not be the most promising way forward to improving the performance of data-driven models, but rather more informative predictor variables (at least Fig.5 may be interpreted in this way) . I find this a very valuable finding which may also deserve to be communicated/ highlighted more clearly (like you did for example in l. 384-389, but not in the abstract or elsewhere). At the moment the differences in significance are stressed more than the very similar magnitude of performance between the different AutoML methods. Also, your finding that the AutoML does not help to reproduce interannual changes (l. 267) is an important finding, because it is a common problem in data-driven upscaling and very relevant question in the carbon cycle community, and therefore in my opinion deserves to be stressed more.

Response: Thank you for your valuable feedback! We appreciate your constructive comments, which are very helpful in improving the manuscript. Indeed, tuning the machine learning setup might only produce marginal improvements compared to considering more informative predictor variables. We will highlight this finding more prominently in the text. We will, furthermore, stress the problem of AutoML to reproduce interannual variabilities, as suggested.

I suggest publication of the paper after addressing the following major questions/ comments:

1. What is the reason for doing this analysis at a monthly temporal scale when structural vegetation changes dominate rather than finer temporal resolution? I would expect higher gains from AutoML and also more differentiated contributions between predictor variables (especially meteorological features) at higher temporal resolution. This is also the time scale which is more relevant to be able to properly represent seasonal and anomalous trajectories. I would expect large potential from automated model tuning especially for short extreme events, which are relevant for the carbon uptake and hard to represent in a data-driven model set-up, but clearly smeared out at a monthly time step. Much of the discussion in section 4.2 does neglect the coarse time step when for example LUE changes are not expected to play major role.

Response: The temporal resolution is indeed an important factor in the contribution of the different predictor variables. A higher temporal resolution could enable the models to represent better anomalies, extreme events, and their impact on GPP (see, f.x. Bodesheim et al. (2018)). Since many previous upscaling works focus on monthly scales, and these data have been instrumental in

informing global long-term dynamics of GPP across different regions in many studies, we have chosen to perform this evaluation at monthly scales as an initial step. Our team has follow-up studies that examined more advanced machine learning algorithms, such as the temporal fusion transformer (TFT), in modeling the dynamics of GPP at hourly scales across space (Rumi Nakagawa et al., 2023). More assessments are necessary to quantify machine learning performance under different time scales. We will make sure to highlight better the consideration of temporal scales on the upscaling framework and model choice in our discussion in the revised manuscript.

References:

Bodesheim, P., Jung, M., Gans, F., Mahecha, M. D., and Reichstein, M.: Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product, Earth System Science Data, 10, 1327–1365, https://doi.org/10.5194/essd-10-1327-2018, 2018.

Nakagawa, R., Chau, M., Calzaretta, J., Keenan, T., Vahabi, P., Todeschini, A., Bassiouni, M., Kang, Y., 2023. Upscaling Global Hourly GPP with Temporal Fusion Transformer (TFT). https://doi.org/10.48550/arXiv.2306.13815

2. Data sets:

A number of predictor variables are model outputs themselves, relying on input data and model assumptions. This is not discussed at all.

Response: We will include further discussion about the sources of the variable input with a focus on introduced uncertainty from the modeling process.

What is the reason for ingesting both SIF and instantaneous SIF, or both PAR and RSDN?

Response: Our approach was to include as many predictor variables as possible and let the AutoML frameworks decide themselves what variables are necessary for a good prediction. This includes variables showing a high intercorrelation and potentially small differences in predictive capacity. We will provide clarification in the revised manuscript.

How is the temporal aggregation done?

Response: We aggregated with a simple average within the respective period after filling the data gaps (see below). We will clarify this in the text.

How do you handle data gaps?

Response: We filled gaps at native temporal resolution. For high-resolution data products (frequency <=4 days), such as NBAR, LAI/FPAR, BESS, CSIF, and CCI, we filled gaps less or equal to 5 days (8 days for products with a 4 day resolution) with the average of a 15-day moving window. We gap-filled LST with a 9-day moving window since we observed higher variations. Soil moisture was filled after Walther et al. (2021) with moving window medians for short gaps and mean seasonal cycle for long gaps. We will clarify this in the text.

Handling of bad data quality is only mentioned for the site-level fluxes, what about the explanatory variables?

Response: We used NBAR, where >75% high-resolution NBAR pixels were available from full BRDF inversion. We applied the quality control mask for LST where the average emissivity error is < 0.02. LAI/FPAR was used with and without saturation. We used all data for soil moisture. We will include this in the text.

Specify more clearly the data sources, e.g. for the CCI soil moisture, which version did you use? Presumably, FluxCom v6 refers to the FluxCom set up with RSonly (only satellite-based predictors using MODIS collection 6), which is 8-daily and at high spatial resolution?

Response: We used CCI Soil moisture v.06.1 and FluxCom v6 RS only. We will include this in the text.

3. Spatial resolution: Why not also ingest tower meteorology instead of the coarser ERA5-Land? The scale mismatch could be further discussed, especially between a 0.05deg pixel and the tower footprint. The way the authors approach the analysis suggests using the 0.05deg pixel is the generally accepted default, which is not the case.

Response: Thank you for raising this interesting point. The spatial mismatch is a large uncertainty factor in the prediction, as outlined in the manuscript (l.309-314). Tower meteorology is expected to increase predictive performance substantially compared to the coarse-resolution ERA-5 product. Regarding using meteorological variables as predictor variables for global upscaling, however, tower meteorology poses a limitation due to its spatially constrained availability. It cannot be used as a predictor for regions where no flux tower data exists. For this reason, we chose ERA-5 land, since it is globally available and, hence, can be used for global predictions. It would be interesting to evaluate uncertainties in reanalysis data using tower meteorology and understand the potential impacts on upscaling uncertainties. We will clarify and discuss this aspect in the text.

4. In parts the manuscript uses very technical language and describes key concepts only in a very short manner. I suggest to rephrase certain passages to make the manuscript better accessible to a wider audience which may also not be very familiar with the newest developments in the machine learning world – or at least expand more in the supporting information. Examples of very technical sentences in my opinion are l.160-161, l.165-166, l.170-172, l.177-181, l.243-246

Response: We will make the text more accessible and reformulate the mentioned sentences.

5. I am afraid, but I cannot follow the meaning of Fig.6.

Response: We will include a better explanation in the caption and make the figure more understandable.

Minor comments for clarification:

Throughout the manuscript: The analysis is not done on climatological time scales, so VPD, precipitation and temperature are meteorological variables, it's not climate data.

Response: We will change the corresponding text passages.

l.22: I suggest to stress in the abstract already the small differences between the AutoML frameworks, eg. by writing '...AutoSklearn consistently but marginally outperformed other AutoML frameworks…'

Response: We will change the corresponding text passages.

l.49 and later in the manuscript: In the literature the term 'variable importance' is used with very different meanings. Please clearly state that for your work, importance refers to the contribution of a variable to model accuracy.

Response: We will provide clarification for the use of "variable importance" in the text.

l.49-56: I am not convinced that the conclusions of the different cited papers are strictly comparable because the analyses have been done at different temporal scales, from daily to monthly, and using different feature sets. Although the machine learning results are analysed which do not necessarily need to obey conceptual understanding, the contributions of different features are expected to differ between time scales.

Response: We will more explicitly mention the different time scales of these studies and the limitation in comparing them.

l.66 (and later as well, eg l.146, 149, 319, 325): Could you clarify/ give examples of what is meant by 'pipeline creation' and 'data processing steps'? The legend of Fig.A2 is hardly understandable for the non-expert without any further context or info.

Response: The term 'pipeline' refers to the entire process of developing and training a machine learning (ML) model. A pipeline typically consists of several tasks, such as preprocessing, feature engineering, model training, hyperparameter tuning, and model deployment. Preprocessing involves various tasks to convert raw input data into a shape accessible for ML training. It typically includes steps such as data cleaning, transformation, integration, or reduction with the goal of improving the quality, accuracy, and reliability of ML models. We will provide further clarification in the corresponding text passages.

l.81: 'predictive contribution' to what? To prediction accuracy?

Response: We will include further clarification in the text passages.

l. 202: Is there a reason for leaving out the VIs?

Response: Including the VIs in the RS minimal set did not improve the prediction. Hence, we did not include them in the other feature sets. We will clarify this in the text.

l. 232: So you compute a linear trend also for time series of just 2 years?

Response: We will change the threshold to a longer period (5 years) and update the corresponding figures and text passages to ensure a more robust trend estimation.

l. 241: What value does the critical difference take?

Response: The critical difference is calculated with

$$\mathrm{CD} = q_\alpha \sqrt{\frac{k(k+1)}{6N}}$$

(CD: critical difference, q: critical values, k: number of algorithms, N: number of datasets). For more information, see Demšar (2006). We will include more clarifying information in the text.

Reference:

Demšar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 7, 1–30, 2006.

Section 3.4: So the main take-away is that the patterns from AutoML in general make sense when compared to other upscaling products? Or do you want to convey another message?

Response: We will include a concluding sentence to highlight this finding.

l.465: the deforestation is mentioned the first time here and I cannot follow what is meant.

Response: We will leave this part out since it is confusing and not connected to the main message of the manuscript.

l.519-525: This last part may be slightly overstating, I do not see very clear indications of more robust and accurate GPP predictions yet.

Response: We will adapt this part.