Responses to referee #2

Gaber et al. test the ability of multiple automated machine learning (AutoML) approaches, each based on multiple individual machine learning methods, to upscale gross primary production (GPP) with remote sensing. They specifically test three different AutoML methods (as well as a random forest model as a baseline) with different subsets of remote sensing and meteorological data, finding that they provide very similar performance, with r^2 ranging from ~0.7-0.75 at monthly scale. They also find similar abilities to capture trends, spatial variation, and seasonality across most approaches but that none of them is particularly effective at capture monthly GPP anomalies. The best models were typically based on a combination of MODIS surface reflectance with additional remote sensing-based estimates of LAI/FPAR, land surface temperature, soil moisture, evapotranspiration, and solar-induced fluorescence (SIF); adding meteorological reanalyses of precipitation, temperature, and vapor pressure deficit did not notably improve model performance.

Overall, the manuscript presents an interesting comparison of some cutting edge approaches to automated machine learning and adds a new dimension to ongoing discussions of flux upscaling. It's also a well written and well-constructed study. The fact that the approaches achieve similar results to each other and to other upscaled products is itself interesting and perhaps suggests that further improvement in upscaled GPP estimates may come from avenues aside from just algorithmic optimization (e.g., better and more extensive ground data, improved remotely sensed data streams). I have a few suggestions for improved presentation and additional analysis, but overall I think this is likely to be a high quality contribution.

Response: Thank you for your constructive feedback! Your comments are very valuable to us and will considerably contribute to improving the manuscript. We will highlight the similar results of the different frameworks and other upscaled products more in the text.

1) My main suggestion for the analysis would be provide, if possible, a more refined and specific assessment of the importance of individual variables. The analysis of the different subsets is interesting, but I think the impact of the study could be enhanced by assessing specifically which variables within those subsets are giving the most "bang for the buck." I know random forests, for example, provide variable importance metrics and perhaps those are doable from the AutoML approaches as well? I'm curious, for example, in the RS subsets, which variables added the most predictive skill beyond what was achieved with RSmin? How important were LST and soil moisture? Did the ET and SIF data, which are themselves modeled from remote sensing data, add any additional independent information? The CSIF product, for example, is itself an upscaled SIF product based on machine learning of MODIS NBAR data, so it seems like it wouldn't necessarily add anything beyond what the methods were able to get directly from the NBAR data.

Response: Thank you for raising this interesting point. We agree that the importance of the individual predictor variables would, indeed, add value to the study. We will include an assessment of the importance of individual variables in the form of an ablation study for the best-performing model-variable combination, AutoSklearn-RS. This can be done by calculating the permutation importance, which would indicate the model's sensitivity towards individual features. That technique takes a fitted model and has it predict on data, where one feature is recursively replaced by random noise, resulting in a potential decrease in the performance metric. The magnitude of the decrease indicates the importance of that feature to the particular model. While this technique allows us to assess the model-specific sensitivity, it can only provide a limited insight into the intrinsic information content of the input variables.

2) I think the Discussion could use a little improvement in places. I think it would be especially helpful to improve how the findings are contextualized in light of previous literature. I'll provide more specific suggestions below.

Response: Thank you for the suggestions. See below for the responses.

3) I find Fig. 6 very difficult to interpret. Is it possible to present those results in a more intuitive form?

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

Specific comments:

L12: should that be "scale" instead of "scales"?

Response: We will adapt the text.

L14: parameterization is misspelled (missing an "e")

Response: We will adapt the text.

Fig. 2: Just to clarify, this is showing number of sites, not site-years, correct? If so, I wonder if it would be more relevant to show site-years since that's a better representation of how much training data is available in each biome?

Response: We will include this information in the figure.

L122: I think it would be worth expanding more on these different sources, including references. Especially since some of these (ET and SIF) are themselves modeled based on remote sensing. Given that, what would you expect them to add beyond what would be coming from the NBAR data itself? Would they actually be providing independent information?

Response: We will include further discussion about the sources of the variable input. We don't expect additional information from SIF, as mentioned in your comment. ET, i.e., the ALEXI model, is derived based on energy balance and surface temperature, which is highly coupled with GPP due to stomatal control. Therefore, we hypothesize that the physical mechanisms inherent in the ET data may contribute additional information on GPP beyond remote sensing signals. We expect the feature importance analysis to shed light on the unique contribution of these variables. The revised manuscript will also discuss the impacts of modeled vs. observational variables.

L274-275: These may be "statistically different," but to me, it seems like an r^2 of say 0.74 is not particularly different from an r^2 of 0.75 in any meaningful sense. The authors do a good job stating this later in the paper, but I do think it's worth not overinterpreting small differences even if they are "statistically significant." Any difference, however small, could be "significant" given a large enough sample size, but that doesn't necessarily make it a meaningful difference.

Response: Thanks for raising this point. We will adapt the corresponding text passages.

L286-297 (but also in other places throughout the results): There are places here that could use references to specific figures or panels within figures. Sometimes it's hard to tell where the results as described are shown in the figures.

Response: We will adapt the text.

L304-305: The overestimation of low values and underestimation of high values is interesting and consistent (I think) with some of the early studies of MODIS GPP (perhaps from David Turner and/or Faith Ann Heinsch, if I'm remembering correctly?). Some reference to those earlier works here would provide valuable context. The fact that we're still trying to solve long-standing problems is itself interesting!

Response: Thanks for providing these insights and references. We will consider them in the text.

L390-399: This paragraph (about differences among approaches) seems to slightly contradict the previous one (about how there aren't really major differences). I'm not suggesting that the authors do a complete rewrite of the paragraph or anything, but I do think it might be worth making sure that they are sending a consistent message: that the differences are generally pretty slight.

Response: We will adapt the text passage.

L401-407: It could also be that the quality of the eddy covariance data itself is a limiting factor. EC GPP is used as the ground truth in this case, but it's not a perfect representation of GPP: EC data has sources of noise and EC GPP is a modeled quantity from the more directly measured NEE. I imagine there may therefore be upper limits to the performance metrics that we can expect when upscaling EC GPP just because of uncertainties in what we're using as "truth."

Response: We agree and will include discussions about the uncertainties and modeling background of GPP in the text.

Section 4.2: I think this section would definitely benefit from a more thorough dive into the variable importance, as suggested in general comments. Also, I don't think there's any mention of SIF in this section while other variables composing the RS subset are discussed?

Response: We will include an assessment of variable importance (see above) and consider the results in this paragraph.

L433-439: The authors mention this at the end of the paragraph, but I think it could be more up front: reanalysis data (especially for precip) can be very flawed. So maybe temperature and VPD do matter (precipitation probably less so since soil moisture is already included in the model and ultimately it's soil moisture, not precipitation, that gets directly used by plants) but the reanalysis data just doesn't do a good job capturing it. Could also be worth a citation to previous literature that has assessed reanalysis data.

Response: This is a good point. We will discuss the impact of reanalysis data with reference to previous studies, e.g., Tramontana et al. (2016). Additionally, microwave soil moisture retrievals are noisy with limitations, which may undermine their contributions to the model. Thus, the lagged precipitation may still provide useful information. Our feature importance analysis will provide further information in this respect.

Reference:

Tramontana, G., Ichii, K., Camps-Valls, G., Tomelleri, E., Papale, D., 2015. Uncertainty analysis of gross primary production upscaling using Random Forests, remote sensing and eddy covariance data. Remote Sensing of Environment 168, 360–373. https://doi.org/10.1016/j.rse.2015.07.015

L444: I'd suggest rephrasing "It is to be explored." That's somewhat awkward, passive phrasing.

Response: We will adapt the text.

L463-466: This paragraph is kind of light on citations and the final sentence feels out of place and incomplete, like there's something more that should be coming that connects the first part of the paragraph to this final thought.

Response: We will provide more references for this paragraph and embed the last sentence better in the paragraph.

L477-484: This paragraph is also pretty light on citations. A couple suggestions: Smith et al. 2019 (Remote Sensing of Environment) on challenges specifically in dry regions and the early MODIS papers by Turner that assessed biome differences in MODIS GPP performance. It'd be interested to see the results here contextualized with the challenges that have faced remote sensing of productivity for a long time!

Response: Thanks for suggesting these references! We will provide more references in this paragraph.

L481: It's unclear what's meant by "high proportion of biomass" or how that would affect productivity estimation. To me, it seems like it's not high biomass that would lead to good performance but rather high seasonal variation in leaf area (which both DBF and MF have).

Response: We will rephrase this paragraph.

L484: A little unclear what's meant by "complex biophysical and environmental characteristics." I think it'd be worth expanding on this and being more specific.

Response: We will rephrase this paragraph.

L487: I think "It is to further research to…" is also somewhat awkward and passive phrasing and would suggest rewording.

Response: We will adapt the text.

L490: This is another good place to cite Smith et al. 2019, which also shows that drylands are underrepresented in flux networks relative to their global proportion. Haughton et al. 2018 (Biogeosciences) could be a good one too since they showed that drylands are more "unique" (meaning less easy to apply a globally-trained model to an unseen site) than most other systems, which may be partly why the underrepresentation of dryland sites in flux networks can be such a problem for upscaling in those regions.

Response: Thank you for providing these references. We will consider them in the text.

L504: For the Conclusions section, it might be worth expanding on what's meant by "RS" here. That's referring to a specific subset of the variables but for readers who are skimming and skip to the conclusions section, they might miss what that subset refers to.

Response: We will adapt the text.

L519-520: Maybe to some extent, but it's interesting to note that RF (not automated and with, I think, some amount of subjectivity in choices) performed nearly as well as the AutoML methods.

Response: It is an interesting point. We will include this in the text.