Biogeosciences Discuss., referee comment RC1
https://doi.org/10.5194/bg-2022-108-RC1, 2022

Comment on bg-2022-108

Anonymous Referee #1

Referee comment on "Evaluation of gradient boosting and random forest methods to model subdaily variability of the atmosphere–forest CO2 exchange" by Matti Kämäräinen et al., Biogeosciences Discuss., https://doi.org/10.5194/bg-2022-108-RC1, 2022

This study by Kämäräinen et al. compares two different machine learning algorithms for the prediction of CO_2 net ecosystem exchange of a boreal forest using ERA5 reanalysis data. Getting accurate estimates of CO_2 exchange outside of spatiotemporal domains covered by eddy covariance measurements is an important task and hence a relevant topic for Biogeosciences. The analysis of spatial and temporal neighborhood predictors and the emulation of less complete time series are interesting concepts and the study is overall well-written and structured. However, I have some general and specific concerns and questions listed below that should be addressed in a round of major revisions before publication.

General comments:

Introduction/Discussion: I think a more complete consideration of state-of-the-art literature could better present the novelty of this study which is not clear in the current form, and that literature should also be considered more for discussing the results. For example, please show what improvements can be expected from gradient boosting in view of previous research, e.g. Tramontana et al. 2016 (https://doi.org/10.5194/bg-13-4291-2016), who already compared various ML algorithms for NEE and GPP prediction. Please also cite literature regarding spatiotemporal neighborhood predictors, if there is any.

We thank the reviewer for the overall positive comments on our manuscript. We agree that both the Introduction and Discussion chapters could include more references to the previously published articles, and we are planning to include them in the corrected manuscript. However, as the time resolution and other details are quite different between different studies, the direct comparisons of the results between different studies might not be possible.

References to temporal lagging of the predictor data (temporal neighborhood) should be quite easy to find and we will include those: however, the utilization of the spatial neighborhood is perhaps a new invention and references for that approach are much harder to find for this reason.

As this study seems to be conducted in view of the general goal of estimating carbon fluxes for points in time and especially in space without direct flux observations, it would be more interesting to test the models also with spatially independent data, i.e., for a comparable boreal EC station (from Fluxnet for example) fully excluded from model training. Otherwise, it should be pointed out more clearly that the predictive error likely is much higher when the models are applied to new locations, see e.g., Roberts et al., 2017 https://doi.org/10.1111/ecog.02881

Our approach models the NEE only in time dimension: it does not receive any spatial information, and for this reason it can not be generalized outside the study site. To really make it applicable to other locations, a transformation of the model to three dimensions (time, latitude, longitude) would be necessary. For this an abundant set of NEE observation samples representing different (boreal) bioclimates would be necessary, and additionally, spatial information about the biology and geography (vegetation, land properties, orography, latitude, etc.) of those locations would be needed to allow the model learn the spatiotemporal relationships between the predictor variables and the NEE.

Applying the model as it is in different sites would implicitly contain an assumption that, for example, the vegetation and soil properties are the same everywhere, which of course is not a realistic assumption.

We will make this more clear in the text.

While I see that ML-models do not require causal relations, I'm still not convinced by the inclusion of negatively lagged, i.e. future, meteo-variables. Did the exclusion of negatively lagged variables actually deteriorate model accuracy or are they just redundant with spurious correlations? In any case, the explanation regarding advection requires references as a theoretical basis supporting it. To me, it makes sense only for grid cells downwind from cells representing the station well (e.g. the central cell). However, I still don't see what additional information can be gained as all the "advected" information already is contained in the non-time-lagged data of the more representative grid cells. Furthermore, it makes no sense for all meteo-variables, e.g., radiation and soil temperature, two of the most important variables, certainly are not advected directly. Hence, I think this explanation requires a more profound basis, i.e., by analyzing the importance of negatively lagged variables by grid cell in relation to wind direction, and by meteo variable. Otherwise, negatively lagged variables should rather be excluded from the analysis in my opinion.

The reason why they have been included in the first version of the manuscript are the spatiotemporal uncertainties of the ERA5 data. The reanalysis is a **modeled representation** of the observations of the meteorological variables, and for this reason it necessarily contains uncertainty, such as biases. Letting the gradient boosting learn the spatiotemporal neighborhood of the data makes the model able to actually learn – and take into account – the effect of the spatiotemporal biases.

We will consider excluding the negatively lagged time steps from the predictor data: that will not deteriorate the results too much, but as they seem to raise questions and confuse readers, it might be better to not use them. If we use the negative lags also in the next version of the manuscript, we will make sure the reasons why they were regarded as useful.

The Pearson correlation coefficient is insensitive to magnitude, so it does not tell how accurate the predicted values are and hence is not very meaningful for model evaluations. I recommend to focus on R2 instead.

We agree with Referee #1 that the Pearson CC does not take into account and measure the variance or bias of the data. However, using R2 instead of Pearson CC when evaluating the goodness-of-fit would not likely change the results, because the R2 measures the linear correlation quite similarly as the Pearson CC a) when bias is very small, b) when skill is high, and c) when the metrics are evaluated from large samples. All conditions, a), b), and c), are fulfilled in our modeling.

https://en.wikipedia.org/wiki/Coefficient_of_determination

See Figure 1 for comparisons of Pearson CC and R2 in a set of random simulations of correlated datasets. See also our response to general comments of Referee #2.

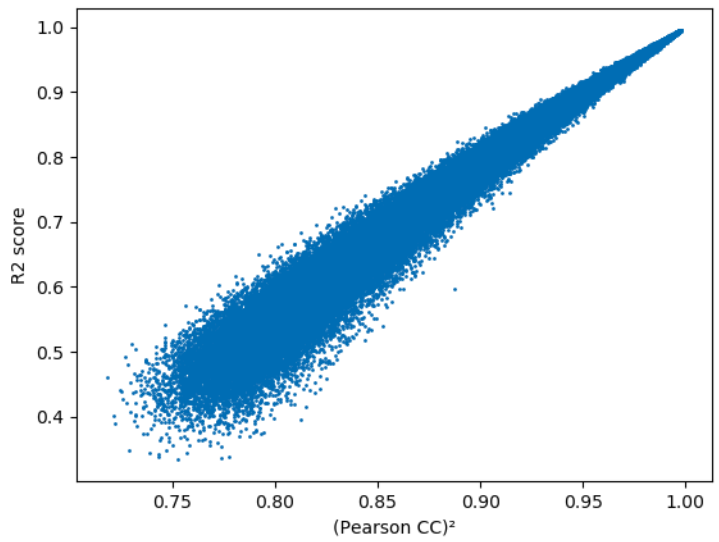

Figure 1. Relation of the square of the Pearson correlation coefficient and the R2 score in unbiased random (but correlated) datasets. Each point represents one dataset of size (250,2): the Pearson correlation and R2 score were calculated between the two columns.

Using larger sample sizes than 250 (our CO2 data contains 10000 samples) would make the relationship even stronger.

Because Pearson CC does not measure the variance or bias, we have used the RMS as an alternative measure of the skill, as RMS is affected by the variance and bias of the modeled data. When we draw conclusions about the goodness-of-fit, we take into account both measures: the Pearson CC, and the RMS.

We can change the analysis such that R2 (or NSE as suggested by Referee #1) will be used instead of the Pearson CC, as it better takes into account the bias and variance in a single metric, but very likely it will not have major effects on the conclusions.

Specific comments:

L24-25: This recommendation is too general, as the models have been evaluated just for one specific ecosystem.

We will modify this recommendation such that it better takes into account the limitations of the data.

L45-48: Please add a reference here.

We will consider adding a reference. However, this piece of general knowledge might not necessarily need one in our opinion.

L65: Is the reference to kaggle really necessary? This rather comes across as an advertisement for a company, so please remove it.

We will remove the reference.

L80-81: Please make clear that EddyUH (and REddyProc?) processing was not done within this study but before data was acquired.

We will modify the sentence, for example like this: "Flux processing for the NEE was done previously by Mammarella et al. (2016) using the EddyUH software (a summary of the data is shown in Fig. 1, presented as multi-year mean values)."

L81-83: Rather explain NEE when the term is first introduced in Section 1.

We agree that this sentence would be better to locate earlier in the text: we will move it to Section 1.

L85: What constitutes a missing value? Were flux data filtered according to a certain quality control strategy, e.g. a test on stationarity, well-developed turbulence, footprint etc.?

Because the flux processing was done earlier, as mentioned by Referee #1, we do not take a closer look into the constitution of the missing data in this article. In general, the very raw data contains missing values due to technical faults in the instruments, power outages, and so on. Additionally, the flux processing is an additional filter, which causes some other data to be discarded, as suggested by the Referee. And finally, the averaging process discards a major part of the data as explained in the text.

We can extend the sentence to make it more clear, for example: "...windows. Only complete 6 hourly aggregates, i.e. those with no missing values arising from flux processing and instrument faults, were accepted for the averaging process."

L90 (Fig. 1): Why is 1998 written below Jan? The title seems superfluous, rather add NEE to the y-axis.

The year 1998 was accidentally left in the figure: it will be removed in the next version of the draft.

We will shorten the title and add NEE to the y-axis.

L95-96: Were missing values gap-filled or just omitted for the calculation of the weekly means? The latter would likely introduce a bias towards more negative NEE values as likely more nighttime data are missing compared to daytime data. This could at least be mentioned.

The missing values were omitted from the calculation of the weekly means. However, the same missing steps were also removed from the modeled data: this makes quality-of-fit measures fair.

We can mention the diurnal distribution of the missing data in the text, even though it won't affect the quality-of-fit/skill estimation of the model.

L107: Are you sure 1° is the spatial resolution? I think it's 0.1°, otherwise it would be really coarse.

The 1° resolution data is what we have been using here. The original data is 0.25°, which is of course denser. We can download the data in this denser resolution and re-calculate the results, and if the new results are significantly different, we will change the text and figures accordingly.

L110: Some of the abbreviations appear quite bulky, rather use more common ones like H, LE and rH. Also, is diffuse or total shortwave radiation not available in the ERA5 product?

These abbreviations were inherited from the ERA5 data. We are not completely sure whether more common abbreviations actually exist for all quantities which all readers could accept, but we will consider whether we can improve the readability of the presentation of the variables.

Different shortwave variables are available in the ERA5. We will consider if adding one or more of them in the list of predictors would offer significant added value for the study.

L123-127 (Fig. 2): in the caption, temporal lags from -2 to +2 are stated, though in the figure lags from +3 to -1 are visualized. Please also make the numbering uniform between Fig. 2 and A2, i.e. that the central cell is number 13 in both figures and so on. As some of the most important grid cells are outside the inner circle (10, 11 ,21), I think it's necessary and less confusing to show them all.

We will modify the figure according to the suggestions of the Referee.

L129-135: Were all 23752 operations carried out in the end (as PCA was not used) despite technically being too laborious? Please clarify.

Unfortunately, we could not carry it out, as the computation would have taken too much time.

L165-168: Please add a reference.

We will add a reference.

L171-172: How many data points were included in each of the 10^3 bootstrap samples?

We have used standard bootstrapping. In it the same number of data points is sampled as in the original data: 10500. Variation between the samples is caused by sampling **with replacement**. See https://en.wikipedia.org/wiki/Bootstrapping_(statistics)

L201-202: Are 00 and 06 UTC the start or the end of the averaging period? (also relevant for Fig. 4)

It is the beginning of the period. We will mention this in the corrected text.

L242: Rather write "cope better" as this is not a yes-no question. To evaluate which one copes better, wouldn't it also be necessary to compare the decrease in prediction skill of each model to its own 100% CORR and RMSE values?

We will add the word "better" to that sentence.

We are not quite sure what Referee #1 means with the latter comment. At 100% there is no difference between the sampling procedures – the data are literally the same. This is also visible in Figure 5: the lines of the same color (dashing for non-random and solid line for random sampling) merge after 90%, and at 100%, the values are the same.

L258: Are highly correlated variables an issue for gain? I know they induce a bias for permutation importance, so can this be excluded for gain?

This is an interesting question, but we do not have a direct answer for it.

If two highly correlated variables are present in the predictor data, both will get high gain values if they are relevant for the prediction, but only if random forest style subsampling is used when building the trees. Without subsampling, probably only the better one of those two would get a high gain score (as it would be selected for all trees), the other getting near zero gain (as it would be probably rejected from most trees).

In most cases, and especially with uncertain and noisy atmospheric data, using subsampling is recommended, not only as it yields more accurate models, but also because it likely makes the gain values more stable.

We have used subsampling both for the GB and RF models in this study, which makes their results comparable (even though we did not present the gain from the RF models).

Whether the differences in gain results between the different approaches (subsampling vs. no subsampling) is a problem depends on the desired outcome. Subsampling "softens" the differences between gain values of predictors, which can reveal potentially interesting results.

See also the answers to the question of SHAP values by Referee #2.

L266: direct or total SW radiation?

We had only one SW radiation parameter in this study: "Mean surface direct short-wave radiation flux", see Table 1.

L269 (& L312-313): I think this could be worth some more detailed analysis. To what percentage was (pine) forest the dominant land cover in each grid cell? Does this correlate with the importance statistics? Is there any spatial pattern? (Figure A2 could be visualized as a map for this).

Figure 2 shows the typical, scattered landscape of the nearest grid cell of Hyytiälä. The surrounding cells are quite similar. Mostly forest covered, lots of lakes, and when zoomed closer, agricultural land is quite dominant as well. Unfortunately, telling the requested percentages, or examining why this area was not giving the highest gain value in the analysis, might be too laborious tasks.

However, recalculating the results in the denser spatial resolution might change the result – we will see.

Figure 2. A satellite image of the area corresponding roughly to the location of the nearest ERA5 grid cell of the study site.

L275 (Fig. 6): The figure would be more readable if the importance results were averaged over the five models with a measure of variation. Alternatively, swap the figure with one or all figures of the Appendix, as they are more easily recognizable and hence more valuable to the reader. Also, an x-axis label is missing.

We will swap Figure 6 and A1.

L276-279: the same results for grid cells and time lags would also be interesting.

We will consider, but not promise, testing this experiment for the requested dimensions as well.

L302: Are there any papers investigating the model accuracy for out-of-range data?

We can try to find examples. However, in general, tree-based methods do not extrapolate well outside the range.

L309-310: “more of a proxy-like” sounds clumsy. Suggestion: “represented by proxy”.

We will change the text as suggested.

L339: I think the quoted FLUXCOM approach by Jung et al. already is a global flux model for this very purpose. Hence, the formulation “could act as a first step” sounds rather misleading to me.

Please note that in that sentence we do not refer to creation of **the first** global flux model: we refer to creation of **a** global flux model, and to make one, it has to be started from the first step.

Technical comments:

Articles are sometimes used excessively for generic nouns, e.g. L12-16, L.78, L.242-243
Write CO_2 with a subscript 2

We will remove articles from generic nouns and formulate “CO2” with a subscript.