Biogeosciences Discuss., referee comment RC2
https://doi.org/10.5194/bg-2022-108-RC2, 2022

Comment on bg-2022-108

Anonymous Referee #2

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-RC2, 2022

This paper (GCB-21-2684) evaluated the predictive skill of two machine learning models for estimating sub-daily net ecosystem exchange (NEE) in a long-term boreal forest site. Although using machine learning to model NEE is not a new topic, this study provides informative results on model choice (XGBoost vs commonly used random forest), the use of climatic data solely to estimate NEE, and the benefits of incorporating spatial and temporal autocorrelated information. These results are potentially helpful to carbon flux modeling with machine learning. I have several outstanding questions and suggestions that I hope the authors would consider.

Major comments:

1. The introduction should provide more background on the use of machine learning to model eddy covariance measured NEE and identify the knowledge gap that this paper tries to fill. Many studies have employed machine learning models to upscale eddy covariance NEE, and global products such as FLUXCOM are available. Therefore, what makes this study significant or informative when it models NEE with machine learning in a single site? This paper looks at novel aspects which were not discussed in the introduction, such as comparing GB vs. RF; incorporating spatial and temporal information.

It is true that spatially more comprehensive prior work exists. As mentioned by Referee #2, we have introduced some new ideas on how to model NEE perhaps better, or at least differently, than what has been achieved earlier.

For example, dozens of general circulation models (GCMs) worldwide contribute to the IPCC assessments of the ongoing global climate change. Each of those GCMs model the same common goal – the spatiotemporal variability of the key climate variables – using more or less different approaches. Together their results complete each other. Similarly, different impact models (such as the one presented in our work) could be used to 1) find new ways to achieve the common goal of modeling accurately the NEE, to 2) complete the estimations of the (spatio-) temporal variability of the NEE, to 3) help other modelers perhaps improve their own approaches, and so on.

We will improve the Introduction by better discussing the novelties of this study, as suggested by Referee #2.

2. A more rigorous model evaluation procedure would help improve the robustness of the model comparison results. This could include 1) using different types of goodness-of-fit metrics (e.g. NSE and bias), 2) estimating uncertainties of model performance from repeated cross-validation with random splitting and model initialization. Please see my specific comments.

We can improve the evaluation by including NSE (or R2 as suggested by Referee #1). However, at this point, we do not expect the conclusions to be different, as NSE, Pearson CC and R2 all produce comparable results a) when bias is 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/Nash%E2%80%93Sutcliffe_model_efficiency_coefficient

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

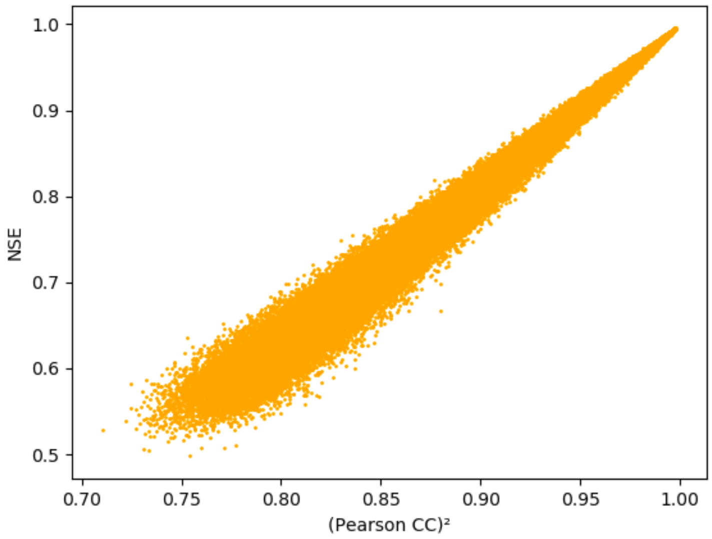

Figure 1. Relation of the square of the Pearson correlation coefficient and the NSE in unbiased random (but correlated) datasets. Each point represents one dataset of size (250,2): the Pearson correlation and NSE were calculated between the two columns.

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

We can change the analysis such that NSE (or R2 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.

See specific comments later for the question about cross-validation experiments.

3. It would be interesting to look at how incorporating neighboring temporal and spatial information affects the predictability of NEE by the machine learning models since previous studies usually only focus on concurrent and collocated measurements/inputs.

While the feature importance analysis shed light on the benefits of spatiotemporal information, the importance metrics are difficult to interpret for tree-based models, given that many features are highly correlated. A direct comparison between models with and without spatial/temporally neighboring information would be appreciated.

We can perform a control simulation without the spatiotemporal neighboring data. Depending on the results we will decide how to report them – either visually in the Figures or in the text only.

4. Global feature importance metrics are sometimes unstable and difficult to interpret for

tree ensemble methods, especially when features are highly correlated. I suggest evaluating feature importance using SHAP as an additional metric to get a more rigorous quantification of importance. See some discussions about feature importance here (Yasodhara et al., 2021, https://link.springer.com/chapter/10.1007/978-3-030-84060-0_19#Sec), here (https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27), and an example using SHAP here (Green et al., 2022, https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.16139).

We thank Referee #2 for the references and for the idea of using SHAP values as a complementary or alternative measure of predictor importance. However, we are not sure whether we have enough resources to complete the study with SHAP values. We agree that sometimes interpretation of the gain values can be difficult, but we believe that the subsampling procedure that we used makes the gain analysis more stable: see our response to Referee #1 (the comment about gain values, referring to line 258 in the text).

We will consider, but do not promise, either testing only shortly or also reporting SHAP values in the article. If we can not use them, we will then warn the readers about the potential stability problems of the gain values.

5. Data-driven models of carbon fluxes often use satellite observed structural vegetation information as a major input. Therefore, it is interesting to see in this paper, that climate variables (from ERA5) could explain 95% of temporal dynamics of NEE in a site. Moreover, the level of accuracy from this paper is considerably higher than those from similar studies, both from a single site and from spatial upscaling over multiple sites. Could you please provide more discussion on the model performance and feature selection of this study in the context of previous results from the literature?

We will provide more discussion about the accuracy of the results.

The most important reason explaining the good result is the direct availability of the observational data in the study site, which allows the model to learn the site-specific temporal distribution and other details accurately.

Building a full 3-dimensional model (with dimensions (time, latitude, longitude)) would have required more measurements of NEE from sites representing different bioclimatic conditions. That kind of modeling would enable eg. LOO cross-validation over different study sites, yielding estimates of spatial uncertainty, which are not possible to get using only one measurement site.

Comparisons of our results with single site studies are occasionally difficult because of different time resolutions, different skill metrics, and different bioclimates, but we will try it as well as possible.

See also our response to the second general comment of Referee #1 about details required to make a global or regional model of spatiotemporal variability of NEE.

Specific comments:

Abstract

L18-19: This is an informative finding. But the manuscript doesn’t have an experiment that directly compares a model with spatial and temporal information to a model without such features.

We can perform such an experiment and report the results in some appropriate way.

L20-22: Both GB and RF rely on the same theoretical approach to identify features that are important to minimize the loss function since they are both tree-based algorithms. The fact that GB is more accurate than RF demonstrates the effectiveness of the “boosting” technique, but there is no direct evidence that GB identifies “more important features” than RF or is more resistant to overfitting.

We can repeat our experiment about inclusion of input variables one by one also for the RF to see whether direct signs of overfitting from that approach could be found.

Introduction

L50-56: Background on the reanalysis is informative, but is this necessary for this paper, given that most readers may already have a general knowledge.

Among the authors of the manuscript it was considered important. The educational background of readers of Biogeosciences might not be very homogeneous, and therefore, we think these lines might be good to have there. However, we can also remove the explanation, if requested by the Referee.

Methods

L134-135: Does this result apply to both RF and GB? This is an interesting finding to me and could be highlighted in the result/discussion/conclusion.

We will highlight this result in other parts of the text as well.

L145: Could you please elaborate on the benefits of transforming the target variable to Gaussian?

Even though it is not completely clear to us why the results are better with transformed (and back-transformed after modeling) data, it might be related to better simulation of the non-extreme values. See this example and the related discussion: https://stats.stackexchange.com/questions/447863/log-transforming-target-var-for-training-a-random-forest-regressor

As the great majority of the data is non-extreme, even a slight enhancement of simulation of the “major bulk” of the data can lead to overall skill improvements – despite the slightly less accurate simulation of the tails.

We will add this explanation in the text.

Figure2: Showing 25 grid cells would be helpful (maybe remove the notations “X” since the plot will be more compact.)

We will change the figure as suggested by both Referees.

L170: I suggest adding bias and the Nash-Sutcliffe model efficiency (NSE) (https://en.wikipedia.org/wiki/Nash–Sutcliffe_model_efficiency_coefficient) (or R2 score, coefficient of determination, common in machine learning applications) to the evaluation metrics, so it is easy to compare the results in this study with other papers.

See our response earlier: the second comment in Major comments of this document.

L173: Use 1000 instead of 103 for easy reading.

We will change 10^3 to 1000.

L175: Hyperparameter tuning through a grid search or other techniques is a common procedure to obtain the optimal accuracy of a machine learning model. It is an essential step to create a fair game when benchmarking different models. Often hyperparameters are determined for each cross-validation fold (see Tramontana et al., 2016 for an example). Although it might be true that significant improvement in the model performance is not likely, it is important to include sufficient justification about your tuning process. For example, what was the search space of parameters? How many sets of parameters were evaluated?

We will add information about the details of the hyperparameter tuning, such as the searched space of parameters.

Determining hyperparameters separately for each cross-validation fold sounds potentially computationally too heavy, but we can consider it. Also, we can consider applying some automated approach, such as Bayesian optimization (https://scikit-optimize.github.io/stable/auto_examples/bayesian-optimization.html).

L180: Another suggestion is to perform repetition experiments (e.g. 30 or 50 repeated experiments for each algorithm, each with a different random split, and random state in the models) to estimate uncertainties from randomness in the cross-validation split and model initializations. See Besnard et al. (2019) for an example. In this way, the model comparison is robust to algorithm and splitting randomness. Confidence intervals of RMSE/R2 can also be derived this way, instead of bootstrapping within the samples.

We will also consider this approach, which sounds promising. Most likely reducing the num_parallel_tree from 10 to 1 is necessary to accomplish the heavier computation, which will slightly deteriorate the skill of the GB model.

Results

L189: Do you mean 1,000 samples?

Yes – formatting of the number (10^3) was lost at some stage of the text processing.

Figure 4: 1000 bootstrap samples?

Yes – formatting of the number (10^3) was lost at some stage of the text processing.

L222-224 (Figure 4.): The variation of accuracy between years can also be related to the random split of years during cross-validation. For a test year, if years with similar climate conditions are in the training set, the testing accuracy is likely higher than otherwise. To this end, the repeated model runs would help eliminate this effect.

We can consider this approach.

L232: do you mean “sub-sampling” here?

Yes. We will include the word “sub-sampling” in this sentence.

L230-240: The description of methods should be in Section 2, and here you may present the results.

We will move the description to Section 2.

L265: I suggest placing Figure A1-3 to the main text, and Figure 6 can be presented in Appendix. Figure A1-3 summarizes the importance of individual ERA5 variables, different spatial grid cells, and information from different temporal windows respectively. They are easier to interpret and provide a clearer comparison than Figure 6.

We agree that Figures A1–A3 could be included in the main text, and Figure 6 could be moved to Appendix instead.

L269: It is interesting but somewhat surprising that the nearest grid cell is not the most important in the model. Further investigation and explanation would be needed here. What is the size of the tower footprint? How heterogeneous is this area? Is the tower close to cell 9, which may have a similar plant composition as the tower footprint? Is this related to lateral flows? What is the dominant wind direction?

See Figure 2 and its explanation in the comments of Referee #1, showing 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. The dominant wind direction is from the South-West.

Considering the 12-hourly 4DVar approach of the assimilation of ERA5 (https://www.ecmwf.int/en/about/media-centre/news/2017/20-years-4d-var-better-forecasts-through-better-use-observations), the relatively large share of derived or simulated (not assimilated) variables, and the sparseness of the observations, it is not perhaps that surprising that there are uncertainties both in space and in time of the reanalysis. We believe that tree-based methods can take into account these biases and weight the different cells so that the NEE variability can be optimally modeled based on the combination of the data from different cells.

Note also the spatial coarseness of the ERA5 data that we used. We are planning to redownload the data in the original, denser resolution. How much this will change the results is yet to be seen.

L282-283: It is interesting that sensible heat and soil temperature alone could explain 90% of the variance in NEE. Is this for the 6-hourly or weekly model? This could be because diurnal and seasonal cycles dominate the temporal dynamics of NEE. Could you please provide more information on this analysis? For example, provide a figure like the heatmaps in Figure 4 to show if the accuracy of interannual variabilities drops when using only two variables.

The result is for the 6-hourly model. It is indeed likely that the temporal cycles can explain the good result: those two variables might be sufficient to describe the cycles. Additionally, in this time resolution the unexplained (small-scale) variability is likely to be small, as it has been smoothed out by the temporal averaging. Both these reasons probably enhance the skill metrics. This might be an important explanation for the skill of the model, which is seemingly high compared to other single-site studies, typically using either denser (1 min, 30 min, 1 hour) or lower (eg. 1 month) time resolution.

We can plot the heatmaps as suggested, and either include them in the article text or in the Appendix if interesting results can be seen from them. We will also make clear the effect of the selected temporal time resolution on the results.

Discussion and conclusions

L324: By "exclude", do you mean that the redundant variables have low feature importance? It might be misleading to say the model excludes a variable.

We will reformulate the wording of the sentence. It is true that even though some variables might get near zero feature importances, and thus do not effectively participate in the prediction, they still can be found at least in some of the trees.