Biogeosciences Discussions

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 "Spatiotemporal lagging of predictors improves machine learning estimates of atmosphere–forest CO_2 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.

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.

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.

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).

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?

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.

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.

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.

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.

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

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

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.

L173: Use 1000 instead of 103 for easy reading.

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?

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.

Results
L189: Do you mean 1,000 samples?

Figure 4: 1000 bootstrap samples?

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.

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

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

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.

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?

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.

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.