https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

Biogeosciences Discussions

Evaluation of gradient boosting and random forest methods to model subdaily variability of the atmosphere–forest CO2 exchange

Matti Kämäräinen[1], Anna Lintunen[2,3], Markku Kulmala[3], Juha-Pekka Tuovinen[4], Ivan Mammarella[3], Juha Aalto[1], Henriikka Vekuri[4], Annalea Lohila[4,3]

[1]Weather and Climate Change Impact Research, Finnish Meteorological Institute, Helsinki, Finland

[2]Institute for Atmospheric and Earth System Research / Forest Sciences, Faculty of Agriculture and Forestry, University of Helsinki, Helsinki, Finland

[3]Institute for Atmospheric and Earth System Research / Physics, Faculty of Science, University of Helsinki, Helsinki, Finland

[4]Climate System Research, Finnish Meteorological Institute, Helsinki, Finland

Correspondence to: Matti Kämäräinen (matti.kamarainen@fmi.fi)

Abstract. Accurate estimates of the net ecosystem CO2 exchange (NEE) would improve the understanding of the natural carbon sources and sinks and their role in the regulation of the global atmospheric carbon. In this work, we use and compare the random forest (RF) and the gradient boosting (GB) machine learning (ML) methods for predicting the year-round 6 hourly NEE over 1996–2018 in a pine-dominated boreal forest in southern Finland and analyze the predictability of the NEE. Additionally, aggregation to weekly NEE values was applied to get information about longer term behavior of the method. The meteorological ERA5 reanalysis variables were used as predictors. Spatial and temporal neighborhood (predictor lagging) was used to provide the models more data to learn from, which was found to improve the accuracy compared to using only the nearest grid cell and time step. Both ML methods can explain the temporal variability of the NEE in the observational site of this study with the meteorological predictors, but the GB method was more accurate. It was more effective in separating the important predictors from non-important ones, showing no signs of overfitting despite many redundant variables. The accuracy of the GB (RF), here measured mainly using cross-validated Pearson correlation coefficient between the model result and the observed NEE, was high (good), reaching a best estimate value of 0.96 (0.94) and the root mean square value of 1.18 µmol m^{-2} s^{-1} (1.35 µmol m^{-2} s^{-1}). We recommend using GB instead of RF for modeling the CO2 fluxes of the ecosystems due to its better performance.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022


Biogeosciences Discussions

1 Introduction

Forests and other terrestrial carbon sinks remove about one third of the anthropogenic carbon dioxide (CO2) annually emitted to the atmosphere, and thus they constitute an important component of the global carbon balance (Friedlingstein et al., 2020). However, the existing observation network for estimating the total atmosphere–ecosystem CO2 exchange is sparse (Alton, 2020), and especially the historical coverage of observations over the past decades is poor. Among other biotypes and ecosystems, the boreal forests contribute significantly to the global atmospheric carbon stock, but how they do it in detail is still largely unknown, reflected in the wide range of estimates of the carbon storage of these forests (Bradshaw and Warkentin, 2015). Therefore, there is a need for accurate spatio-temporal modeling of carbon fluxes for improved monitoring and understanding the boreal, and ultimately, the global carbon cycles (Jung et al., 2020).

In boreal forests, the atmosphere–ecosystem CO2 flux shows strong seasonal and diurnal cycles, dominated by 1) the photosynthesis by plants (acting as a CO2 sink from the atmosphere), and 2) by the total ecosystem respiration, including plant respiration and organic decomposition processes by microorganisms (acting as a CO2 source into the atmosphere). In a homogeneous forest environment, the net flux generated by these processes can be accurately measured with the micrometeorological eddy covariance method, which has emerged as common standard for long-term ecosystem-scale flux measurements (Aubinet et al., 2012; Hicks and Baldocchi, 2020).

Both total respiration and photosynthesis are typically at their largest in the warm season in boreal forests (Ueyama et al., 2013; Wu et al., 2012; Kolari et al., 2007). On average, their net effect, i.e. the net ecosystem exchange of CO2 (NEE), is dominated by photosynthesis on the weekly scale in summer, but on the sub-daily scale, the total respiration turns NEE positive (i.e., into a source) during nights when photosynthesis of plants is switched off. In the cold season, the diurnal variability is mostly absent, and then NEE is again slightly positive as respiration still continues.

Various meteorological and local biotic factors and processes affect the NEE, and their importance is different in different seasons. Local conditions include soil type and properties, and plant species and their density distributions. Key meteorological variables, such as air temperature and short-wave radiation, typically have large seasonal and diurnal variations. These variables are observed globally using in-situ and remote sensing techniques, and the resulting large-scale data sets can be further post-processed and homogenized via data assimilation, employing numerical weather prediction (NWP) models, and presented in a spatio-temporal grid format. This product is called reanalysis, which can be considered a by-product of the NWP process (Parker, 2016).

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

In recent years, various machine learning (ML) approaches have been proposed and used to model the NEE (or related quantities) over various locations and globally (Besnard et al., 2019). In particular, the random forest (RF) has been popular, and it has been shown to be suitable for this task (Nadal-Sala et al., 2021; Reitz et al., 2021; Bodesheim et al., 2018; Tramontana et al., 2015).

We employ the RF algorithm to model the 6 hourly net CO2 exchange between the atmosphere and a boreal forest in Finland. In addition to the RF regression method, we use the gradient boosting (GB) regression (Friedman, 2001; Chapter 10 in Hastie et al., 2009; for examples of applications across a variety of fields, see http://kaggle.com) and compare their results. In addition to the comparisons of the methods, we investigate the meteorological controls on the CO2 exchange. Several meteorological predictors from the global ERA5 reanalysis (Hersbach et al., 2020) were used as input for the RF and GB regression models, including but not limited to (ground) temperatures, precipitation amounts, radiation quantities, and heat fluxes.

We investigate in detail whether the skill of the GB method could overcome the skill of the popular RF method in explaining the variability of the NEE when using the meteorological predictors. In addition to that, we rank the importance of the individual predictors in the study site and explore the effect of reducing both the number of samples and the number of predictors on the accuracy of the GB model. Finally, we discuss the significance of our results in a broader context.

2 Materials and methods

2.1 CO2 flux measurements as the target variable

The eddy covariance CO2 flux data, measured above a 50 year old Scots pine forest in Hyytiälä, Finland (61°51' N, 24°17' E) in 1996–2018 (Launiainen et al., 2022) and processed to represent the NEE, were acquired from https://smear.avaa.csc.fi/download (accessed 25 February 2021). Flux processing for the NEE was done using the EddyUH software (Mammarella et al., 2016; a summary of the data is shown in Fig. 1, presented as multi-year mean values). NEE is a sum of ecosystem carbon uptake in photosynthesis and carbon loss in respiration, and a negative NEE means that the forest takes up carbon, i.e., is a carbon sink. These data consist of 30 min averages which were aggregated for modeling to 6 h resolution using averaging with moving, non-overlapping windows. Only complete 6 hourly aggregates were accepted for the averaging process. The resulting data set contained 10500 non-missing data points and 22800 missing values. In addition to the preprocessed NEE data, the modeling was separately tested using the raw CO2 flux (i.e., measured by the eddy covariance system and without storage change flux correction and friction velocity filtering) as the target variable.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

Biogeosciences Discussions Open Access EGU

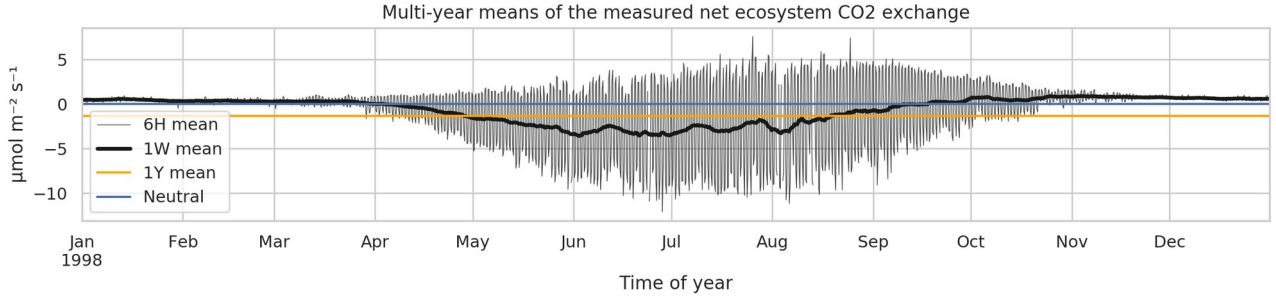


Figure 1: The 6 hourly (thin black), weekly (thick black), and annual (orange) multi-year means of observed net ecosystem CO2 exchange (NEE) at the Hyytiälä SMEARII site. Eddy covariance method with a 24-m tall tower was used for measurements. Years 1996–2018 were used in calculation of the mean values.

Additionally, weekly means were calculated from the 6 h data for validation purposes. For this, a moving, overlapping and centered windowing was used to preserve the same number of samples as in the 6 h data. Missing data inside the window were accepted not to discard almost all of the samples. When validating the model, the missing 22800 time steps were also rejected from the model results for consistency.

2.2 Variables from the ERA5 reanalysis as predictors

Typically, air and soil temperatures, short-wave (photosynthetically active) radiation, and relative humidity are the key meteorological variables used in modeling the CO2 flux (eg., Nadal-Sala et al., 2021). In addition to these, we included a large set of other variables 1) to search for new, unexpected relationships between the flux and these less common variables, and 2) to study how much these variables can either improve or deteriorate the accuracy of the model. Altogether 19 meteorological variables from the global ERA5 reanalysis product (Hersbach et al., 2020) were selected (Table 1).

The ERA5 reanalysis data for 1996–2018 were downloaded from https://cds.climate.copernicus.eu/ (accessed 15 March 2021) in the 1°×1° spatial and 1 h temporal resolution. The data were downsampled to 6 hourly using moving averaging with non-overlapping windows.

Table 1. Gridded parameters from the ERA5 reanalysis product.

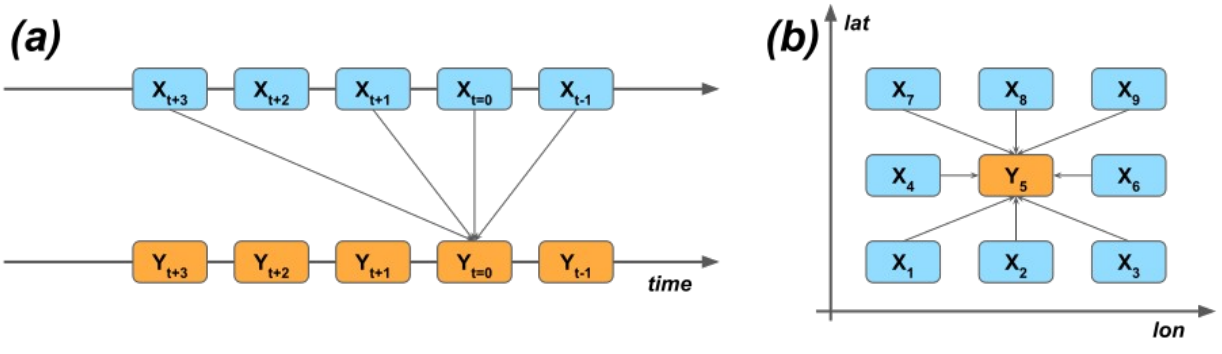
Variable	Abbreviation
Evaporation	e
Mean surface direct short-wave radiation flux	msdrswrf
Mean sea level pressure	msl

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

Mean surface latent heat flux	mslhf
Mean surface sensible heat flux	msshf
Relative humidity at 1000 hPa	r
Snow depth	sd
Soil temperature, level 1 (7 cm)	stl1
Soil temperature, level 2 (28 cm)	stl2
Soil temperature, level 3 (100 cm)	stl3
Volumetric soil water, layer 1 (0–7 cm)	swvl1
Volumetric soil water, layer 2 (7–28 cm)	swvl2
Volumetric soil water, layer 3 (28–100 cm)	swvl3
2-meter temperature	t2m
Total cloud cover	tcc
Total precipitation	tp
10-meter u-component of the wind	u10
10-meter v-component of the wind	v10
Geopotential at 150 hPa	z

2.3 Temporal lagging and spatial neighbourhoods of the predictor data

As the first approximation, the modeling could be carried out by using the grid point closest to the Hyytiälä site. Similarly, temporal synchronization of the predictor data and the target variable could be used. On the other hand, many processes happen sequentially in time and their effect on the target variable could be seen as delayed. For example, meteorological conditions in the night-time can affect the plant photosynthesis the following day (Kolari et al., 2007), and advection of humid or dry air from nearby regions can increase or decrease photosynthesis. We wanted to give the ML models the opportunity to take advantage of these relationships happening in space and time. For this, we selected the 25 closest grid cells around the site and five closest time steps around each of the time steps (t=0) of the target variable. Note that lagging was applied both to forward and delay the predictors in time (Fig. 2a). Then, in total, we had 19 variables × 25 grid cells × 5 temporal lags = 2375 individual predictors for modeling.



https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022


Biogeosciences Discussions Open Access EGU

Figure 2: Examples of using (a) temporal lagging and (b) spatial neighbourhoods of predictor variables X to model the target variable Y. In the study, the temporal lags [-2, -1, 0, +1, +2] were used. In b), only nine grid cells are shown for clarity, but 25 nearest grid cells were used in the experiments.

Technically, the calculation of the correlation matrix was too laborious a task with 23752 ≈ 5.6 × 106 operations. However, the predictor set necessarily contains highly correlated variables: for example, the temperature time series of neighbouring grid cells are correlated. Such collinearity can hamper the robustness and reliability of statistical models (Lavery et al., 2019). To deal with the collinearity, the principal component analysis method (Jolliffe and Cadima, 2016) using 1) all components and 2) reduced number of components was tested as a preprocessing step to make the predictors orthogonal, i.e., non-correlated, but it was found that this dimension reduction method was unnecessary, as the results were slightly better without it (not shown), and thus it was not used here.

2.4 Gradient boosting and random forest regressions

For the machine learning of this study, the xgboost package (version 1.4.2: https://xgboost.readthedocs.io/en/latest/python/index.html) of the Python language (v. 3.7.6: https://www.python.org/) was used to fit both the GB and the RF regression models.

Compared to, for example, deep learning methods, GB and RF models can fit properly with relatively small data sets, do not necessarily require graphical processing units to fit fast, have only a small set of tunable hyperparameters, do not require heavy preprocessing of the predictor or the target data, such as removal of the seasonality, and are generally easier to use. That said, one preprocessing step was found to improve the model accuracy: quantile transformation with 10^5 quantiles was used to make the target variable, i.e., the CO2 flux, strictly Gaussian distributed. Validation of the model was performed using the inverse transformed (non-Gaussian) flux data.

Both the GB and RF are ensemble based tree methods, which means that the final prediction of the model is formed by calculating the mean of weak learners, trees, constituting the ensemble. Variation between the ensemble members is created by fitting the members to random subsamples of the predictor matrix X: these subsamples are formed by sampling randomly both the predictor and the time step dimensions. While the members of the RF are just the trees fitted independently to different subsamples, the GB takes an additional step by fitting the models hierarchically one by one, such that each member tree reduces the prediction error of the previous one. In other words, each new member is forced to concentrate on those observations that are the most difficult to predict correctly (Chapter 10 in Hastie et al., 2009), and in this sense, GB learns

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

more than the RF. We further improved the robustness and accuracy of GB by using a hybrid approach, where small RF models with 10 members were used in boosting instead of single trees.

2.5 Cross-validation framework and parameter tuning

K-fold cross-validation with shuffling and five splits was used to fit five separate ensemble models (Hawkins et al., 2003). In this method, the entire data set is split K (here five) times such that each of the K models has its own validation set, and the remaining data is used to fit the model. Finally, the predictions from all models were combined to form a continuous time series covering all time steps in 1996–2018.

As an important variation to the standard K-fold cross-validation method, we randomly sampled years instead of individual time steps. Sampling randomly time steps would lead to sampling from the same weather events, i.e., from serially correlated data, which would lead to overestimation of the model accuracy in the validation. This can be avoided by sampling sufficiently large, continuous blocks in time, such as years.

For measuring the goodness of fit, the root mean squared error (RMSE) and the Pearson correlation coefficient (CORR) have been used as metrics of model skill, and they were calculated from the 6 hourly and weekly data separately. Bootstrapping with 10^3 samples was used to estimate the sampling errors.

For tuning, multiple rounds of cross-validation were performed with different settings of the parameters, and the best combinations, in the RMSE sense, were selected. The hyperparameters of the models, and their tuned values, are listed in Table 2. A slight improvement in the accuracy of the models could still be achieved by a more thorough exploration of the hyperparameter space using, for example, exhaustive grid search (Pedregosa et al., 2011) or Bayesian optimization (Snoek et al., 2012), but major improvements in the accuracy, which is already close to the optimum with the presented values, might not be likely.

Table 2. Optimized hyperparameters of the GB and RF regression models. The squared error was used as the cost function in the optimization. Optimization was based on multiple rounds of 5-fold cross-validation. Constant default values of other model parameters were used, and they are not presented here. Using num_parallel_tree = 10 > 1 for GB increases the robustness of the model by fitting 10 trees – which equals a small RF – instead of one tree.

Model parameter	Explanation	Optimized value for GB	Optimized value for RF
learning_rate	Step size of the optimization process	0.075	–
max_depth	Maximum depth of a single tree	7	14

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

alpha	The L1 regularization parameter	0.01	0.01
subsample	Random sample size of a tree (proportion of time steps)	0.75	0.50
colsample_bytree	Random sample size of a tree (proportion of predictors)	0.75	–
colsample_bynode	Random sample size of each layer inside a tree (proportion of predictors)	–	0.50
n_estimators	Number of boosting rounds	500	–
num_parallel_tree	Number of random forest samples	10	500

185

3 Results

3.1 Goodness of fit of the machine learning approaches

For the 6 h GB data, the 95% confidence intervals (CIs), based on bootstrapping with 103 samples, were 1.14–1.22 µmol
190 m^{-2} s^{-1} for the RMSE, and 0.955–0.960 for CORR (Fig. 3). For the weekly data, the 95% CIs were 0.455–0.481 µmol m^{-2} s^{-1} and 0.981–0.984 for RMSE and CORR, respectively. The RF performance was also good, but did not reach the GB skill, as the RMSE CIs for the 6 hourly (weekly) data were 1.31–1.40 µmol m^{-2} s^{-1} (0.511–0.536 µmol m^{-2} s^{-1}) for the RMSE, and 0.941–0.947 (0.979–0.981) for CORR, respectively.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

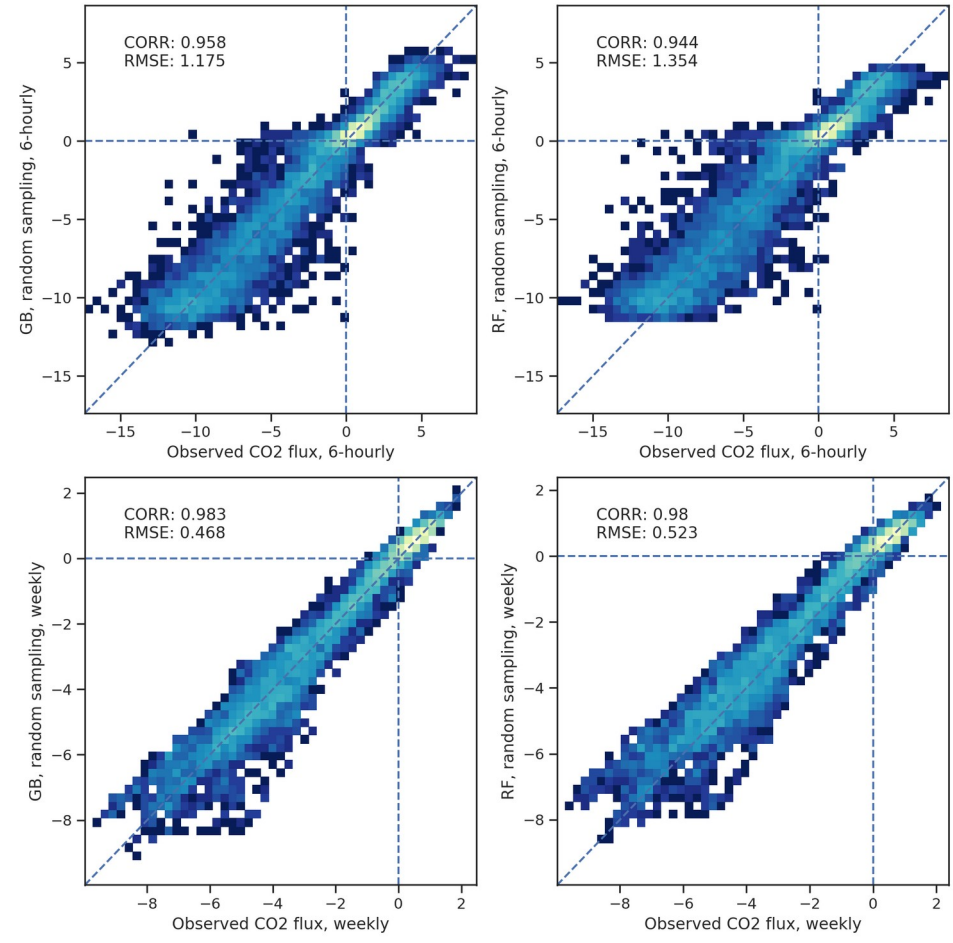


Figure 3: Two-dimensional probability density histograms of the 6 h (upper row) and weekly mean (bottom) observed and modeled CO2 fluxes (GB shown on the left column; RF on right). Color shading indicates qualitatively the density of the observed–modelled value pairs inside each pixel. Estimated Pearson correlation coefficients (CORR) and the root mean square errors (RMSE) of the fit are also shown. See text for the confidence limits of these values.

To study the effect of diurnal and seasonal cycles, monthly and 6 hourly grouping were used simultaneously, and RMSE and CORR were calculated for these groups for the GB algorithm (Fig. 4a). The lowest correlation was found in August at 00 UTC (CORR = 0.49; 95% CIs 0.31–0.64) and the highest in April at 06 UTC (0.89; 0.86–0.91). In general, both the absence of diurnal variation and the small absolute values of the flux in general increase the correlation uncertainty in winter. The largest variation of the target variable in the summer daytime (06–12 UTC; 09–15 local time) yields the largest RMSE.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

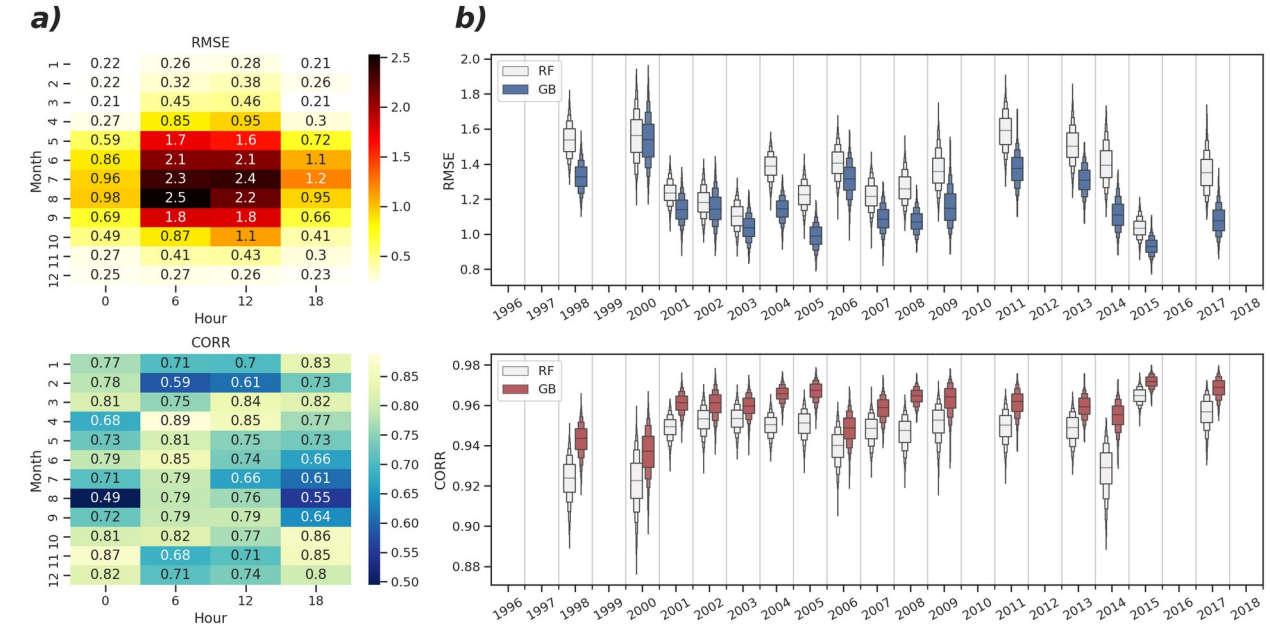


Figure 4: Estimates of root mean square error (upper panels) and Pearson correlation coefficient (lower panels) derived from 103 bootstrap samples. a: Monthly and 6 hourly decompositions of RMSE and CORR for GB. Median values of bootstrap distributions are shown. b: Annual confidence intervals of RMSE and CORR for the RF and GB. Median is shown with a horizontal line in the center of each figure unit. Deciles of distributions are shown with box edges. Years with more than 40% of missing data were excluded.

When excluding the winter months, the day-time NEE was better predicted (CORR = 0.66–0.89) than the night-time exchange (CORR = 0.49–0.77). Interestingly, an opposite result was achieved when the raw CO2 flux was modeled instead of the preprocessed NEE: in that case the night-time (18–00 UTC) fluxes were better predicted (CORR = 0.74–0.85) than the morning (06 UTC; CORR = 0.62–0.83) and afternoon (12 UTC; CORR = 0.67–0.86) fluxes. Analysis of the results of the different target data imply that the sampling error emerging from a rather large share of missing samples in the NEE data could explain the differences (not shown).

Additionally, annual grouping of the data was used to obtain annual estimates of CORR, RMSE and their confidence intervals, again using the bootstrapping technique for both ML algorithm results (Fig. 4b). These estimates show an increasing temporal trend for CORR, implying either 1) a quality improvement in observed fluxes or 2) in the ERA5 predictor data over the years, or 3) changes in the environment as the forest grows. The highest CORR was achieved in 2015 (median CORR for GB = 0.969), and the lowest in 2000 (median CORR for GB = 0.930). Because of smaller samples, the annual CI estimates were wider than when bootstrapping the whole dataset of 23 years.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

3.2 Temporal distribution of the fitting data affects the goodness of fit

The time series of the CO2 flux observations in Hyytiälä is exceptionally long and complete in time. Therefore, it is important to study the sensitivity of modeling to the amount and distribution of fitting data to assess whether the methods could be used for sites with less data. For this, additional, in-sample sampling was used to reduce the amount of data prior to fitting of the models in the cross-validation framework. Sampling with sample sizes of 10%, 20%, … 100% were used to resample the data within each fold of the cross-validation.

The in-sample sampling was implemented with two different strategies. First, the ordinary *random sampling* was used. This strategy mimics the cases in which the time series of a site is incomplete, i.e., contains missing observations randomly distributed over the study period. Second, *non-random sampling* was used to study the cases in which the study period is short but more complete, implemented by using the same percentage shares as with the random sampling, but selecting continuous blocks of data from the beginning of the cross-validation samples. This strategy was used to simulate the cases in which the observational data is more complete but its total length is shorter.

The results indicate that the GB can cope with less data compared to RF (Fig. 5). For example, when considering the non-random sampling, the GB achieves the same skill with 20% data as the RF with 50%. Additionally, the CORR results reveal that the selection of the ML algorithm is more important in determining the goodness of fit of the result than the selection of the sampling strategy at each percentage level. The differences between the random and non-random sampling results also indicate that lengthening the time series by adding more years to it might be a better strategy to further improve the model than gap filling the missing values in the existing observational time series. This can be seen in the larger changes in the non-random sampling results as the amount of data increases: with the random sampling approach, the changes, and hence the algorithm improvements, become quite small with larger than 60% amounts.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

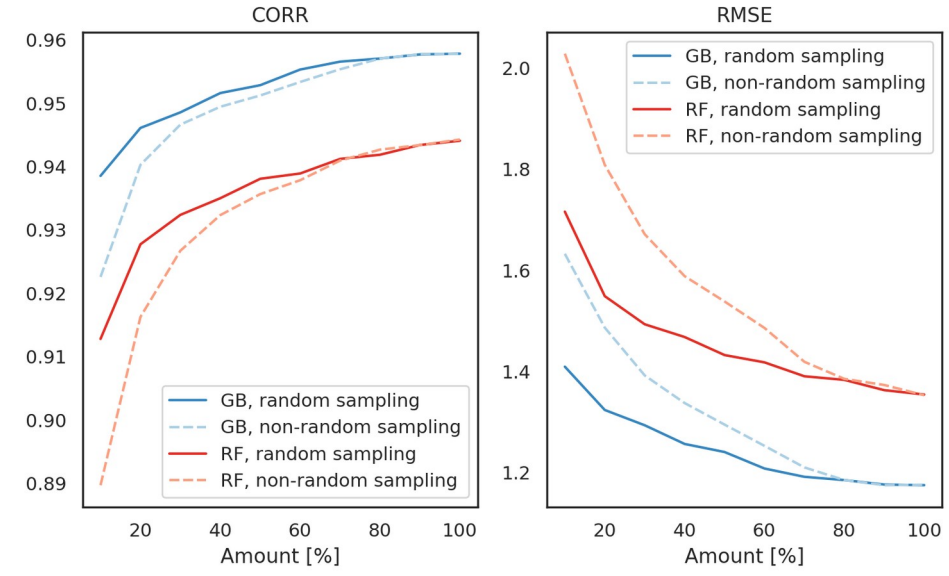


Figure 5: Cross-validated Pearson correlation coefficient (left) and root mean square error (right) as a function of the amount of fitting data for the gradient boosting approach (dark tones) and the random forest approach (light tones). Different in-sample sampling approaches (random versus non-random) are also shown with different dashes. 6 h averages were used in this experiment.

3.3 Analysis of predictor importance

For measuring the predictor importance, the *gain* metric of the xgboost package was used. The gain implies the relative contribution of each predictor to the model, and it is calculated by measuring each predictor's contribution to each tree of the model. When comparing the predictors, a higher gain value implies that the predictor is more important for generating a prediction.

Figure 6 presents the 40 most important individual predictors in each of the five fitted GB models (which differ from each other by the cross-validation years used in fitting), and Figures A1–A4 summarize the mean results for different parameters and grid cells–lag combinations. Sensible heat flux turned out to be the most important of the input parameters. Also, the soil temperature of the uppermost layer, and the short-wave radiation were among the most important predictors. They were followed by the 2-meter temperature, soil temperatures of deeper layers, and evaporation rate. The non-lagged variables were more important than the lagged ones, and so were the positively lagged predictors compared to the negatively lagged ones. However, the nearest grid cell did not contain the most important predictor data on average.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

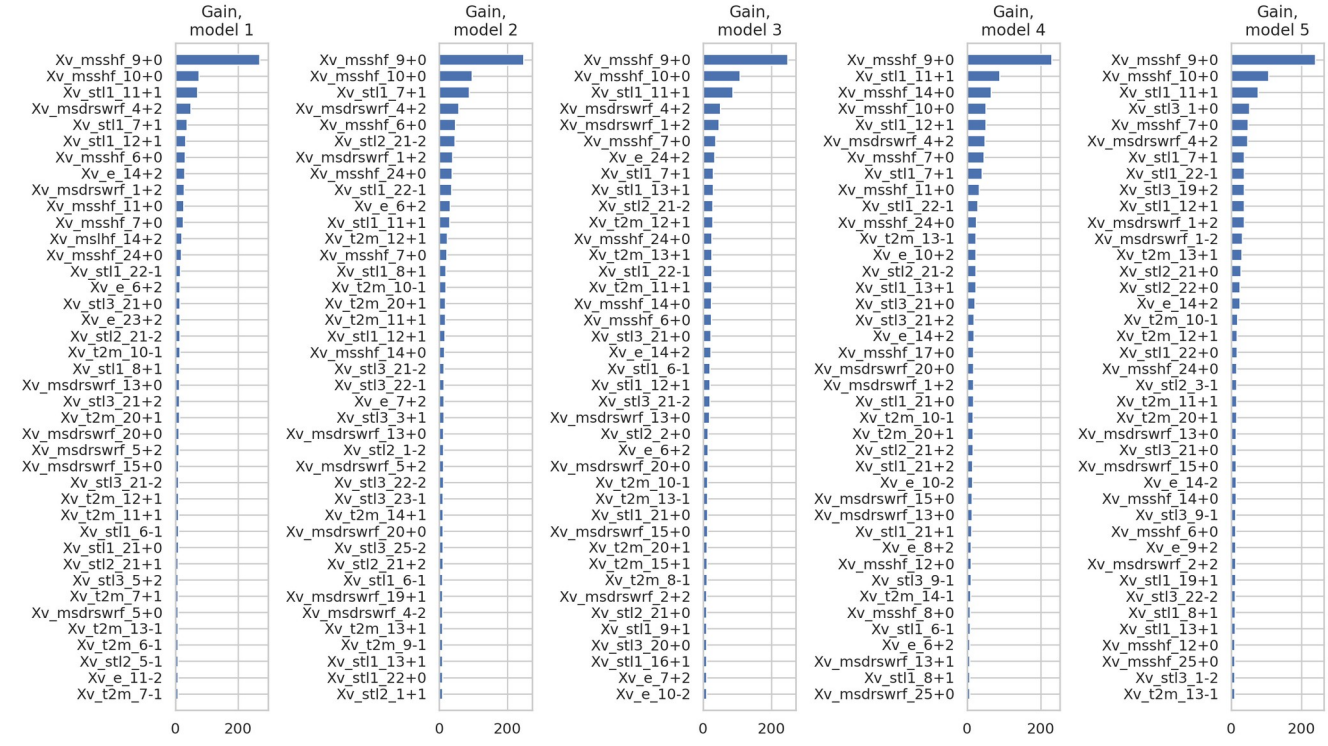


Figure 6: The 40 most important individual predictors in each of the fitted gradient boosting models based on the gain metric value. Predictor name coding: Xv_VARIABLE_GRIDCELL-ID_TIMELAG. For example, in each of the models, the mean sensible heat flux (msshfl) from the grid cell number 9 without temporal lagging (+0) was the most important predictor of the CO2 flux variability. See Table 1 for explanations of abbreviations.

To study the overall relevance of the input variables, we conducted an experiment in which we excluded them one by one, beginning from the worst (total cloud cover, tcc; Fig. A1), and measured the accuracy of GB after each drop until it started to decrease significantly. It turned out that half of the variables originally included were redundant, i.e., they did not improve the model accuracy at all. Importantly, however, they did not worsen the model either. Relative humidity was the first variable to add significant value to the model, and those with a smaller average gain could be discarded without virtually any effect to the results. Interestingly, using only the two most important variables – sensible heat flux and the soil temperature at the uppermost layer – yielded a model with a relatively good accuracy (CORR = 0.947).

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

4 Discussion and conclusions

Many local factors affecting the CO2 exchange between the atmosphere and a boreal forest either vary only slowly over time, as is the case for the plant distribution and growth and soil microorganism populations, or are effectively constant (e.g., soil properties and shape of the terrain). In contrast, the variability of meteorological factors is prominent and happens in short time scales and, partly for these reasons, dominates the variability of the flux response (Sierra et al., 2009). Indeed, the vast majority of the CO2 flux variation in the studied forest can be explained by using only meteorological factors, of which the most important ones were, in order, sensible heat flux, soil temperature, short-wave radiation, air temperature, evaporation rate, latent heat flux, snow depth, air pressure, and relative humidity. Out of all 19 variables included in the analysis, these are the ones which significantly contributed to the GB model skill. It is worth noting that some of the variables included in the analysis are not completely independent of the physical and biophysical processes: to some extent, many of them are regulated by the plants themselves, and the environment in general. The most important of such variables are the latent and sensible heat fluxes, evaporation, relative humidity, and the near-surface temperature.

At least to some extent, if not completely, the ML methods employed here might be able to account for slow changes in the response happening over the years if 1) they are caused by the meteorological variables, and 2) the current period of the study contains clear enough signals of these changes. For example, the increasing trend in temperature is one of the most important variables explaining the CO2 variability both in the short and long term (Huntingford et al., 2017; Pulliainen et al., 2017). However, it is unclear how well the present methods can handle cases in which the values of predictor variables fall outside of the range used in fitting the models. It is likely that the temperature extremes exceed the observed variability in the near future along with the warming local and global climate. Eventually, the ecosystem changes become so large that the accuracy of the method will necessarily deteriorate.

When interpreting the results, it is important to distinguish the conceptual difference between the negative and positive temporal lags. A strong correlation between the response variable and positively lagged predictor is an indicator of the predictor driving the CO2 flux, either directly or indirectly. Intuitively, a correlation between the flux and a negatively lagged predictor variable is more difficult to understand. In these cases the relationship must be indirect and more of a proxy-like: for example, horizontal advection can carry properties to or away from the study site, and these properties can be identified from both upwind (corresponding to the positive time lags) and downwind (negative time lags) grid cells. It is also possible that because of spatial biases and other inaccuracies in the gridded form of the variables, some of the neighboring grid cells might better represent the local conditions than the nearest cell.

In general, machine learning methods seek for relationships between the response variable and the predictor data, and they cannot distinguish whether these relationships are truly causal. Even though the identified relationships and interaction

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

mechanisms may not be intuitive and even causally coherent, they can still be used to improve the model accuracy. To be beneficial for the modeling, such a relationship just needs to be sufficiently robust and strong, and constant in time. Even though the predictor dataset contained many redundant variables, the GB method effectively excluded them, and the cross-validated correlation remained high. In addition to this, the method proved to be skilful even in cases in which the amount of fitting samples was heavily reduced. With less powerful statistical methods, overfitting would be much more likely, leading to poorer cross-validation results when using redundant and/or collinear predictor variables and/or small fitting samples (Chapter 7 in Wilks, 2011; Chapters 3 and 7 in Hastie et al., 2009; Lavery et al., 2019).

Both the efficiency of the GB method in rejecting the non-optimal predictors and the ability to cope with small fitting samples are especially encouraging considering its application to other locations: all variables can be used, letting the model decide about the redundancy. It is likely that the same variables that were found important at our study site might not constitute an optimal choice in other ecosystems and locations; vice versa, the predictors found redundant in Hyytiälä, such as soil moisture, can be important in other environments (Nadal-Sala et al., 2021; Zhou et al., 2019).

Because 1) the meteorological predictors can explain almost all of the variability of the observed atmosphere-ecosystem CO2 flux, 2) gradient boosting regression is efficient in modeling that variability, and 3) CO2 flux is measured globally at a large number of sites representing different climates and ecosystems (Hicks and Baldocchi, 2020), this work could act as a first step in creation of a multi-purpose, national, regional, or global flux model (Jung et al., 2020). These meteorological variables, derived here from the ERA5 reanalysis product, are easily and freely available globally in a spatially and temporally dense, complete, and homogeneous format, and they extend back to the 1950s. However, because local biotic conditions may dominate the variation among different locations, they should be included in the model as well.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

Appendices

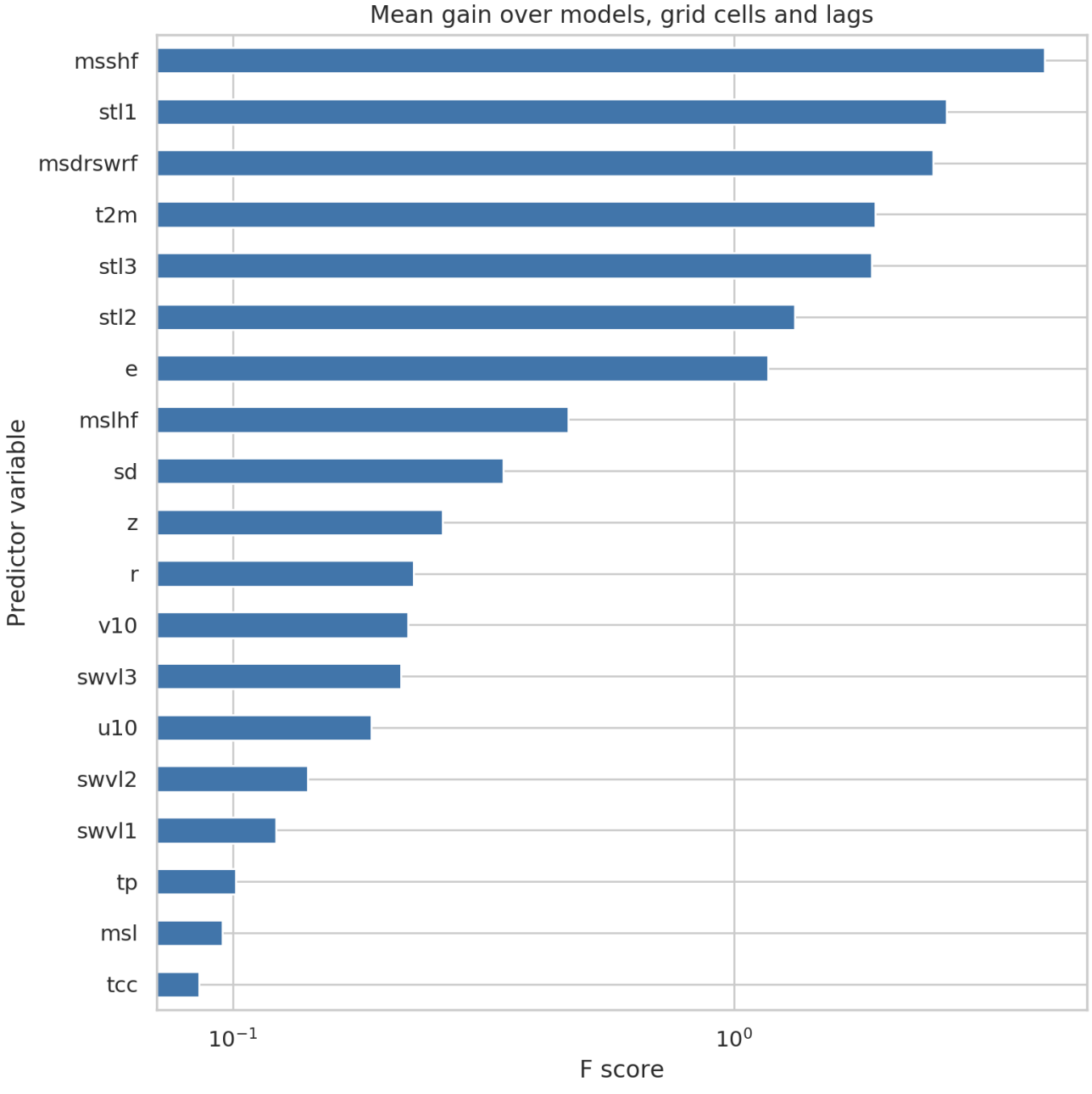


Figure A1. Mean gain of ERA5 gridded parameters averaged over the gradient boosting models, grid cells, and lags. See Table 1 for explanations of abbreviations. Note the logarithmic x-axis.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

350

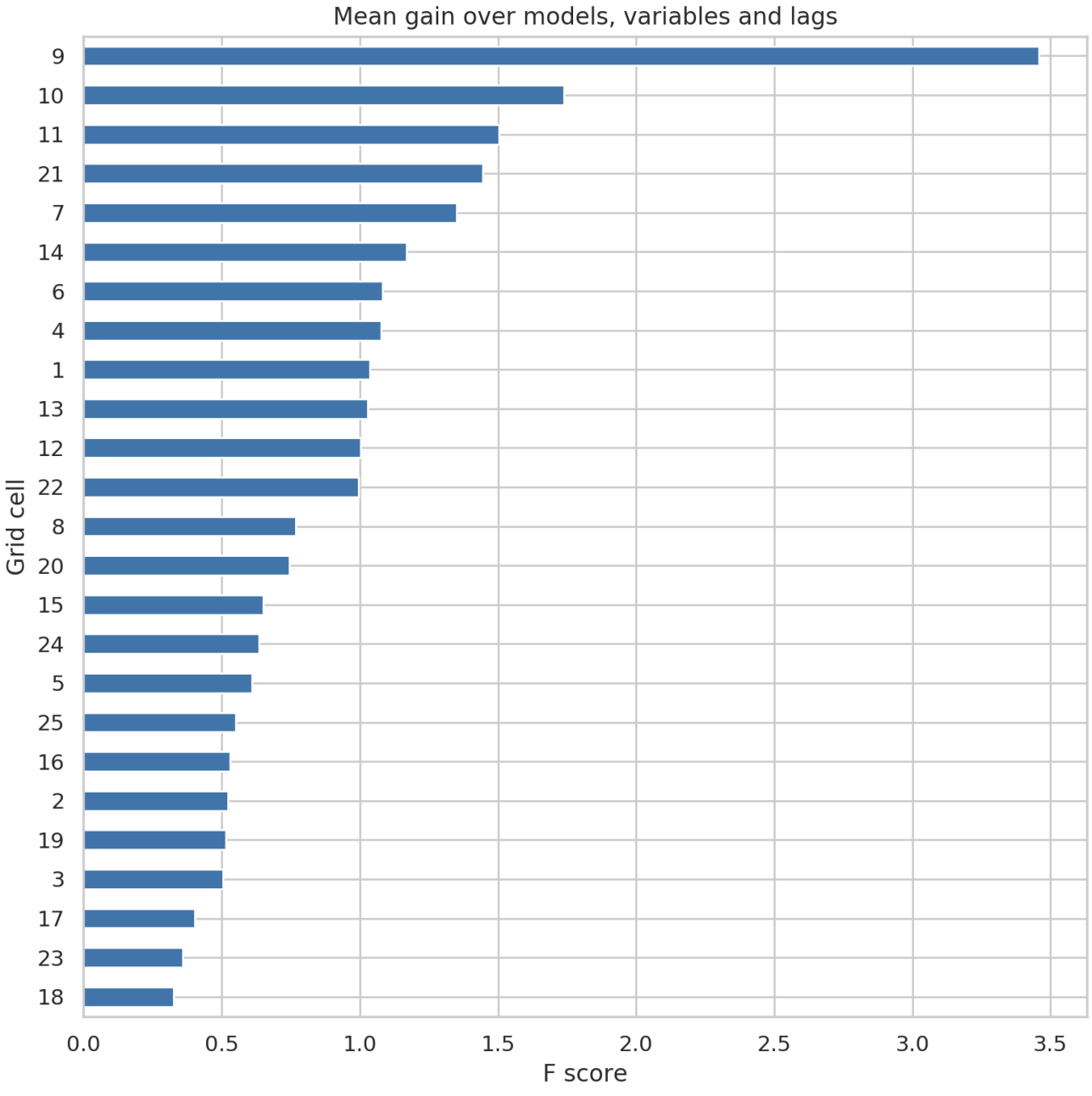


Figure A2. Mean gain of ERA5 grid cells averaged over predictor variables, gradient boosting models, and lags. The numbering logic is so that the cell in the bottom left corner is number 1, the next one to the right is number 2, and so on. The first cell of the

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

next row upward is the number 6. The center cell of the domain, which is closest to the Hyytiälä site, is number 13. See Fig. 2 for visualization of the numbering principle.

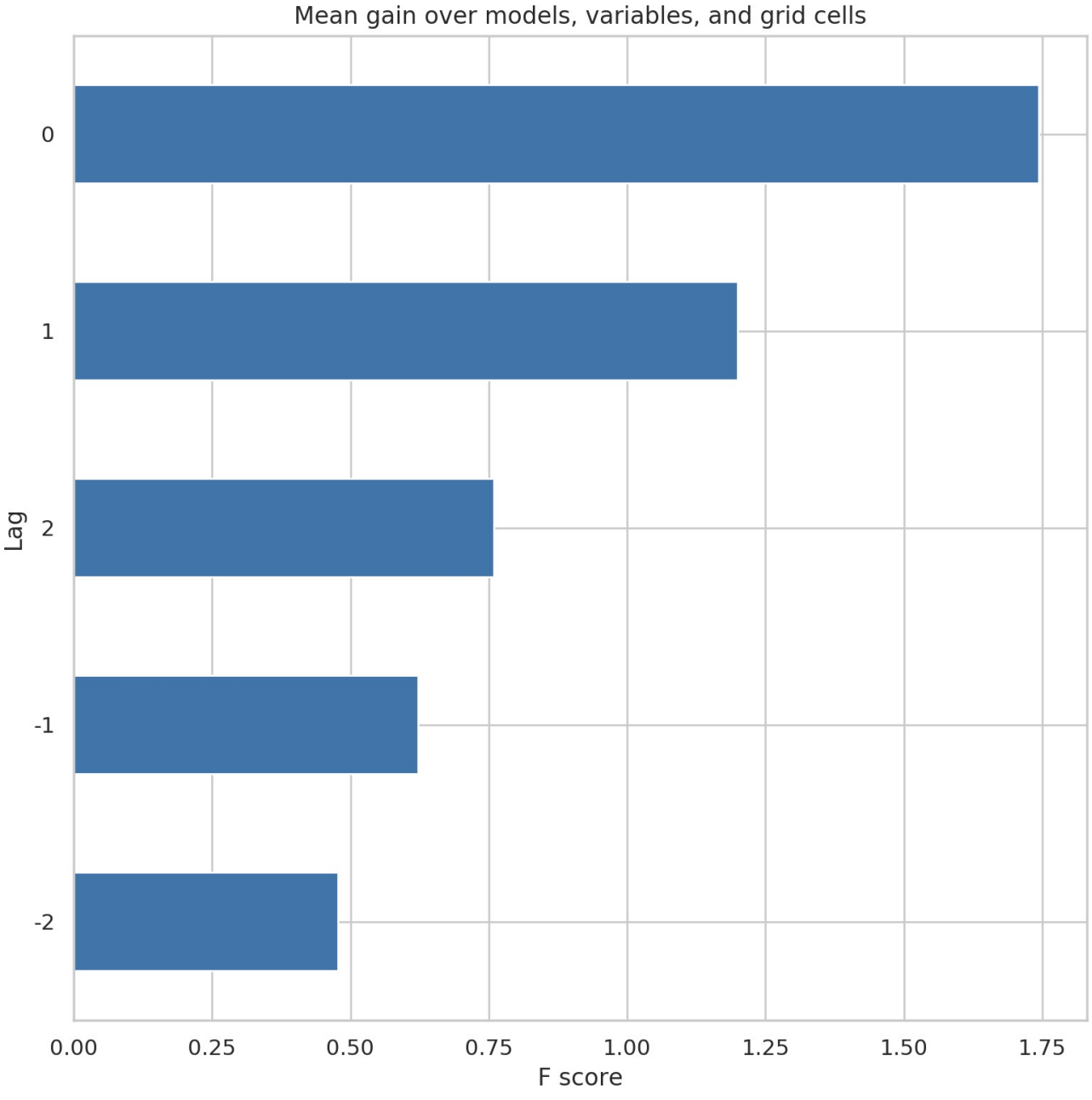


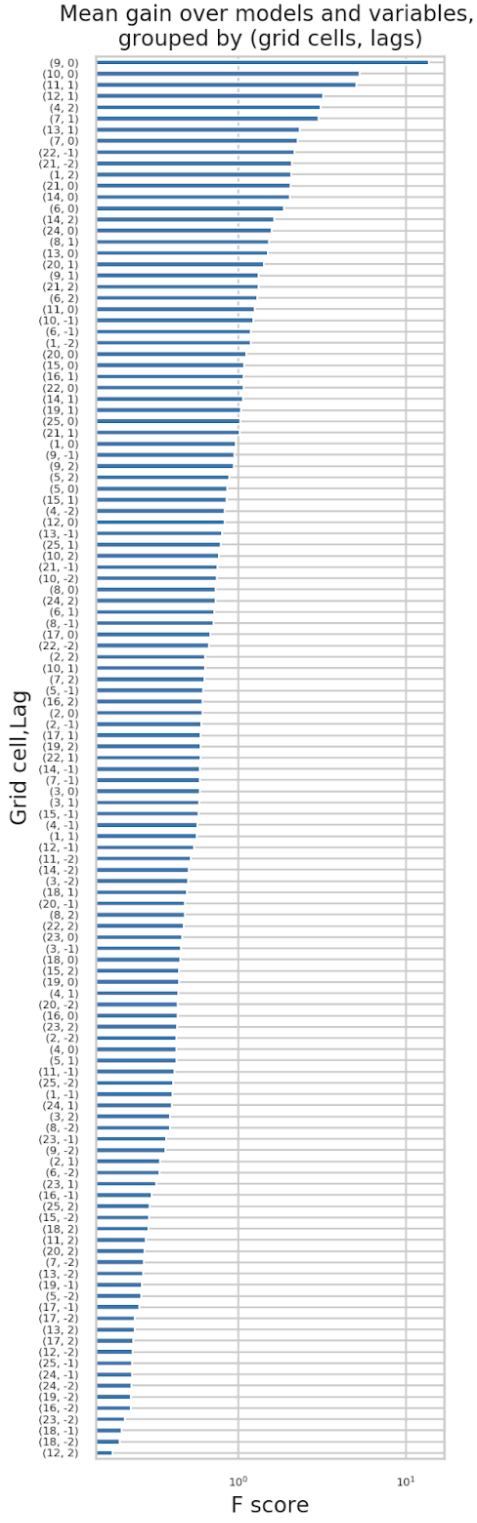
Figure A3. Mean gain of different lags of the ERA5 predictors, averaged over gradient boosting models, grid cells and lags.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022


Biogeosciences
Discussions
Open Access
EGU

Positive lags indicate forwarding the predictors relative to the target variable, and negative lags indicate postponing (delaying) the predictors. Zero indicates non-lagged data.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022



https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

Figure A4. Mean gain of different lags and grid cells of the ERA5 predictors, averaged over gradient boosting models and variables, and using groupings for (grid cells, lags). Negative lags indicate postponing (delaying) the predictors relative to the target variable, and negative lags indicate forwarding the predictors. Zero indicates non-lagged data. See Fig. 2 and Fig. A2 for the logic of the numbering of the grid cells. Note the logarithmic x-axis.

Code and data availability

The code for reproducing the results from experiments and analyses is available at Kämäräinen et al. (2022; https://zenodo.org/badge/latestdoi/368864113). The code can be used to download and preprocess also the ERA5 predictor data: other data, including the NEE data, are included in the repository.

Author contribution

MKä designed the experiments and the structure and content of the manuscript, wrote and executed the code, and composed the text. ALi participated in the planning of the manuscript content and made major suggestions during the writing process, and helped significantly with the references. JT contributed significantly to the content of the reference list and commented the text. IM was responsible for the EC measurements at the study site. HV tested the code and made suggestions how to improve it. MKu, JA, and ALo commented the manuscript.

Competing interests

Authors declare that there are no competing or conflicting interests affecting the work.

Acknowledgements and financial support

We thank the creators and maintainers of the ERA5 reanalysis for providing this invaluable data freely available for the research community. We also thank Hyytiälä SMEAR II staff, ICOS research infrastructure and the responsible researchers for maintaining the eddy covariance data and providing it openly available online. We acknowledge the following projects for the funding of the work: ACCC Flagship funded by the Academy of Finland (337549), Academy professorship funded by the Academy of Finland (302958), research projects funded by the Academy of Finland (342890, 325656, 316114,

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

325647, 347782), Jane and Aatos Erkko Foundation (project Quantifying carbon sink, CarbonSink+ and their interaction with air quality) and the European Research Council project ATM-GTP (742206).

References

Alton, P. B.: Representativeness of global climate and vegetation by carbon-monitoring networks; implications for estimates of gross and net primary productivity at biome and global levels, Agric. For. Meteorol., 290, https://doi.org/10.1016/j.agrformet.2020.108017, 2020.

Aubinet, M., Vesala, T., and Papale, D. (Eds.): Eddy Covariance: A Practical Guide to Measurement and Data Analysis, Springer Science+Business Media B.V, 438 pp., https://doi.org/10.1007/978-94-007-2351-1, 2012.

Besnard, S., Carvalhais, N., Arain, M. A., Black, A., Brede, B., Buchmann, N., Chen, J., Clevers, J. G. P. W., Dutrieux, L. P., Gans, F., Herold, M., Jung, M., Kosugi, Y., Knohl, A., Bewerly, L. E., Paul-Limoges, E., Lohila, A., Merbold, L., Roupsard, O., Valentini, R., Wolf, S., Zhang, X., and Reichstein, M.: Memory effects of climate and vegetation affecting net ecosystem CO2 fluxes in global forests, PLoS One, 14, https://doi.org/https://doi.org/10.1371/journal. pone.0211510, 2019.

Bodesheim, P., Jung, M., Gans, F., Mahecha, M., and Reichstein, M.: Upscaled diurnal cycles of land-atmosphere fluxes: a new global half-hourly data product, Earth Syst. Sci. Data, 10, 1327–1365, https://doi.org/10.5194/essd-2017-130, 2018.

Bradshaw, C. J. A. and Warkentin, I. G.: Global estimates of boreal forest carbon stocks and flux, Glob. Planet. Change, 128, 24–30, https://doi.org/10.1016/j.gloplacha.2015.02.004, 2015.

Friedlingstein, P., O'Sullivan, M., Jones, M. W., Andrew, R. M., Hauck, J., Olsen, A., Peters, G. P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J. G., Ciais, P., Jackson, R. B., Alin, S., Aragão, L. E. O. C., Arneth, A., Arora, V., Bates, N. R., Becker, M., Benoit-Cattin, A., Bittig, H. C., Bopp, L., Bultan, S., Chandra, N., Chevallier, F., Chini, L. P., Evans, W., Florentie, L., Forster, P. M., Gasser, T., Gehlen, M., Gilfillan, D., Gkritzalis, T., Gregor, L., Gruber, N., Harris, I., Hartung, K., Haverd, V., Houghton, R. A., Ilyina, T., Jain, A. K., Joetzjer, E., Kadono, K., Kato, E., Kitidis, V., Korsbakken, J. I., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Liu, Z., Lombardozzi, D., Marland, G., Metzl, N., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S. I., Niwa, Y., O'Brien, K., Ono, T., Palmer, P. I., Pierrot, D., Poulter, B., Resplandy, L., Robertson, E., Rödenbeck, C., Schwinger, J., Séférian, R., Skjelvan, I., Smith, A. J. P., Sutton, A. J., Tanhua, T., Tans, P. P., Tian, H., Tilbrook, B., Van Der Werf, G., Vuichard, N., Walker, A. P., Wanninkhof, R., Watson, A. J., Willis, D., Wiltshire, A. J., Yuan, W., Yue, X., and Zaehle, S.: Global Carbon Budget 2020, Earth Syst. Sci. Data, 12, 3269–3340, https://doi.org/10.5194/essd-12-3269-2020, 2020.

Friedman, J.: Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat., 29, 1189–1232, 2001.

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edi., Springer Series in Statistics, 745 pp., 2009.

https://doi.org/10.5194/bg-2022-108
Preprint. Discussion started: 11 May 2022

Hawkins, D. M., Basak, S. C., and Mills, D.: Assessing model fit by cross-validation, J. Chem. Inf. Comput. Sci., 43, 579–586, https://doi.org/10.1021/ci025626i, 2003.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J. N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hicks, B. B. and Baldocchi, D. D.: Measurement of Fluxes Over Land: Capabilities, Origins, and Remaining Challenges, Boundary-Layer Meteorol., 177, 365–394, https://doi.org/10.1007/s10546-020-00531-y, 2020.

Huntingford, C., Atkin, O. K., Martinez-De La Torre, A., Mercado, L. M., Heskel, M. A., Harper, A. B., Bloomfield, K. J., O'Sullivan, O. S., Reich, P. B., Wythers, K. R., Butler, E. E., Chen, M., Griffin, K. L., Meir, P., Tjoelker, M. G., Turnbull, M. H., Sitch, S., Wiltshire, A., and Malhi, Y.: Implications of improved representations of plant respiration in a changing climate, Nat. Commun., 8, 1–11, https://doi.org/10.1038/s41467-017-01774-z, 2017.

Jolliffe, I. T. and Cadima, J.: Principal component analysis: A review and recent developments, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci., 374, https://doi.org/10.1098/rsta.2015.0202, 2016.

Jung, M., Schwalm, C., Migliavacca, M., Walther, S., Camps-valls, G., Koirala, S., Anthoni, P., Besnard, S., Bodesheim, P., Carvalhais, N., Chevallier, F., and Gans, F.: Scaling carbon fluxes from eddy covariance sites to globe: synthesis and evaluation of the FLUXCOM approach, 17, 1343–1365, https://doi.org/https://doi.org/10.5194/bg-17-1343-2020, 2020.

Kolari, P., Lappalainen, H. K., Hänninen, H., and Hari, P.: Relationship between temperature and the seasonal course of photosynthesis in Scots pine at northern timberline and in southern boreal zone, 59, 542–552, https://doi.org/10.1111/j.1600-0889.2007.00262.x, 2007.

Kämäräinen, M., Lintunen, A., Kulmala, M., Tuovinen, J., Mammarella, I., Aalto, J., Vekuri, H., and Lohila, A.: Gradient boosting and random forest tools for modeling the NEE 2022, Zenodo/Github [code] https://zenodo.org/badge/latestdoi/368864113.

Launiainen, S., Katul, G. G., Leppä, K., Kolari, P., Aslan, T., Grönholm, T., Korhonen, L., Mammarella, I., and Vesala, T.: Does growing atmospheric CO 2 explain increasing carbon sink in a boreal coniferous forest? , Glob. Chang. Biol., 1–20, https://doi.org/10.1111/gcb.16117, 2022.

Lavery, M. R., Acharya, P., Sivo, S. A., and Xu, L.: Number of predictors and multicollinearity: What are their effects on error and bias in regression?, Commun. Stat. Simul. Comput., 48, 27–38, https://doi.org/10.1080/03610918.2017.1371750, 2019.

Mammarella, I., Peltola, O., Nordbo, A., and Järvi, L.: Quantifying the uncertainty of eddy covariance fluxes due to the use of different software packages and combinations of processing steps in two contrasting ecosystems, Atmos. Meas. Tech., 9, 4915–4933, https://doi.org/10.5194/amt-9-4915-2016, 2016.

Nadal-Sala, D., Grote, R., Birami, B., Lintunen, A., Mammarella, I., Preisler, Y., Rotenberg, E., Salmon, Y., Tatarinov, F., Yakir, D., and Ruehr, N. K.: Assessing model performance via the most limiting environmental driver in two differently stressed pine stands, Ecol. Appl., 31, 1–16, https://doi.org/10.1002/eap.2312, 2021.

Parker, W. S.: Reanalyses and observations: What's the Difference?, Bull. Am. Meteorol. Soc., 97, 1565–1572, https://doi.org/10.1175/BAMS-D-14-00226.1, 2016.

Pedregosa, F., Thirion, G., Gramfort, A., Michel, V., and Thirion, B.: Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res., 12, 2825–2830, 2011.

Pulliainen, J., Aurela, M., Laurila, T., Aalto, T., Takala, M., Salminen, M., Kulmala, M., Barr, A., Heimann, M., Lindroth, A., Laaksonen, A., Derksen, C., Mäkelä, A., Markkanen, T., Lemmetyinen, J., Susiluoto, J., Dengel, S., Mammarella, I., Tuovinen, J. P., and Vesala, T.: Early snowmelt significantly enhances boreal springtime carbon uptake, Proc. Natl. Acad. Sci. U. S. A., 114, 11081–11086, https://doi.org/10.1073/pnas.1707889114, 2017.

Reitz, O., Graf, A., Schmidt, M., Ketzler, G., and Leuchner, M.: Upscaling Net Ecosystem Exchange Over Heterogeneous Landscapes With Machine Learning, J. Geophys. Res. Biogeosciences, 126, 1–16, https://doi.org/10.1029/2020JG005814, 2021.

Sierra, C. A., Loescher, H. W., Harmon, M. E., Richardson, A. D., Hollinger, D. Y., and Perakis, S. S.: Interannual variation of carbon fluxes from three contrasting evergreen forests: The role of forest dynamics and climate, Ecology, 90, 2711–2723, https://doi.org/10.1890/08-0073.1, 2009.

Snoek, J., Larochelle, H., and Adams, R. P.: Practical Bayesian Optimization of Machine Learning Algorithms, Adv. Neural Inf. Process. Syst., 25, 2960–2968, https://doi.org/10.1163/15685292-12341254, 2012.

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

Ueyama, M., Iwata, H., Harazono, Y., Euskirchen, E. S., Oechel, W. C., and Zona, D.: Growing season and spatial variations of carbon fluxes of Arctic and boreal ecosystems in Alaska (USA), Ecol. Appl., 23, 1798–1816, https://doi.org/10.1890/11-0875.1, 2013.

Wilks, D.: Statistical methods in the atmospheric sciences, Third Edit., edited by: DMOWSKA, R., HARTMANN, D., and ROSSBY, H. T., Elsevier Inc., Oxford, 676 pp., 2011.

Wu, S. H., Jansson, P. E., and Kolari, P.: The role of air and soil temperature in the seasonality of photosynthesis and transpiration in a boreal Scots pine ecosystem, Agric. For. Meteorol., 156, 85–103, https://doi.org/10.1016/j.agrformet.2012.01.006, 2012.

Zhou, Q., Fellows, A., Flerchinger, G. N., and Flores, A. N.: Examining Interactions Between and Among Predictors of Net Ecosystem Exchange: A Machine Learning Approach in a Semi-arid Landscape, Nat. Sci. Reports, 9, 1–11, https://doi.org/10.1038/s41598-019-38639-y, 2019.