Explainable machine learning for modelling of net ecosystem exchange in boreal forest

Ekaterina Ezhova[1,*], Topi Laanti[2,*], Anna Lintunen[3], Pasi Kolari[1], Tuomo Nieminen[1], Ivan Mammarella[1], Keijo Heljanko[2,4], and Markku Kulmala[1]

[1]INAR/Physics, University of Helsinki
[2]Department of Computer Science, University of Helsinki
[3]INAR/Agricultural and Forest Sciences, University of Helsinki
[4]Helsinki Institute for Information Techonology HIIT
[*]These authors contributed equally to this work.

Correspondence: Ekaterina Ezhova, ekaterina.ezhova@helsinki.fi

Abstract. There is a growing interest in applying machine learning methods to predict net ecosystem exchange (NEE) based on site information and climatic variables. We apply four machine learning models (Cubist, Random Forest, averaged Neural networks and Linear regression) to predict the NEE of boreal forest ecosystems based on climatic and site variables. We use data sets from two stations in the Finnish boreal forest (southern site Hyytiälä and northern site Värriö) and model NEE during the peak growing season and the whole year. For Hyytiälä, all nonlinear models demonstrated similar results with R^2=0.88 for the peak growing season and R^2=0.90 for the whole year. For Värriö, nonlinear models gave R^2=0.73-0.76 for the peak growing season; whereas Random Forest and Cubist with R^2=0.74 were somewhat better than averaged Neural networks with R^2=0.70 for the whole year. Using explainable artificial intelligence methods, we show that the most important input variables during the peak season are photosynthetically active radiation, diffuse radiation, and vapor pressure deficit (or air temperature), whereas, on the whole-year scale, vapor pressure deficit (or air temperature) is replaced by soil temperature. When the data sets from both stations were mixed, soil water content, the only variable clearly different between Hyytiälä and Värriö data sets, emerged as one of the most important variables, but its importance diminished when input variables labeling sites were added. In addition, we analyze the dependencies of NEE on input variables against the existing theoretical understanding of NEE drivers. We show that even though the statistical scores of some models can be very good, the results should be treated with caution, especially when applied to upscaling. In the model setup with several interdependent variables ubiquitous in atmospheric measurements, some models display strong opposite dependencies on these variables. This behavior might have adverse consequences if models are applied to the data sets in future climate conditions. Our results highlight the importance of explainable artificial intelligence methods for interpreting outcomes from machine learning models, particularly when a set containing interdependent variables is used as a model input.

1 Introduction

Forests play an important role in the global carbon cycle because they remove carbon from the atmosphere through photosynthesis and store it in the wood biomass and forest soil. Recent studies suggest that in the past several decades, the net carbon uptake of the boreal forest has been increasing and that of the tropical forest decreasing, making the boreal forest the largest terrestrial carbon sink on the planet (Tagesson et al., 2020). The dynamics of the forest carbon cycle and its interaction with various climatic drivers are generally well-understood; however, the complex responses of forests to climate change and their potential to mitigate its impacts keep boreal forests at the forefront of multidisciplinary research. This ongoing interest spans from observational studies to global modeling efforts (Artaxo et al., 2022; Petäjä et al., 2022; Kulmala et al., 2020, 2023; Tang et al., 2023). There is a growing need for more accurate models of carbon fluxes, providing reliable results in warming climate conditions (Kämäräinen et al., 2023). Hence, suitable models must correctly capture current carbon cycle dynamics using commonly measured ecosystem-level data and give reasonable predictions for, e.g., future higher temperatures. In other words, the models' performance should be adequate in the range of values currently underrepresented in the data sets.

In addition to traditional process-based models (Launiainen et al., 2022; Junttila et al., 2023), the use of machine learning (ML) models have become ubiquitous. ML models play an important role in providing an alternative for the hypothetic-deductive modeling approach, i.e., an inductive approach. This means no prior assumptions are made about the data, which is modeled with a purely empirical model with a general function class. Currently, there is plenty of carbon flux data available from the FLUXNET database, as well as extensive meteorological reanalysis data sets or measurements of many different variables directly from research stations. Data availability has boosted the application of data-intensive ML methods to carbon flux modeling (Dou and Yang, 2018; Zeng et al., 2020).

Using ML, the functional relationship between carbon flux (net ecosystem exchange, gross primary production or respiration) and the site and climatic variables, including radiation, meteorological and biospheric input parameters, can be obtained. There exists plenty of literature featuring the ML approach to quantify different components of the carbon cycle using site and climatic variables as input (Dou and Yang, 2018). In many studies (Cai et al., 2020; Wood, 2021; Zhu et al., 2023; Zeng et al., 2020), researchers identify 'the best model' which reproduces the carbon fluxes depending on available set of input parameters better than other models. Statistical accuracy metrics are typically used as a criterion for model assessment. Many different ML models have been tested, but Random Forest has appeared particularly popular (Liu et al., 2021; Reitz et al., 2021).

However, these empirical machine learning models are often "black box" in the sense that the parameters used by models to make the predictions can not be directly extracted from the model to provide a human understandable way to interpret them easily. The results, therefore, should be treated cautiously. Recently, Shirley et al. (2023) demonstrated with an example from Alaska that the boosted regression tree ML model gave inaccurate results in current and future carbon balance estimates at high latitudes. Increasing the data set by adding more stations from the same area improved the result for the current carbon sink. Still, future estimates were unreliable, ascribed to the fact that the data sets representing future conditions could not be used for model training.

In response to this need, various methods that attempt to make ML models more open and interpretable have emerged. They are called explainable artificial intelligence (XAI) methods (Dwivedi et al., 2023). With XAI techniques, researchers can explore and analyze the factors that influence the model outcomes, making it easier to interpret the results and enhance the utility of ML approaches, e.g., in the context of carbon cycle research.

In the present study, we model boreal forest NEE with subhourly time resolution, using an extensive set of input variables from two research stations at different latitudes: Hyytiälä at 61°51'N and Värriö at 67°46'N. Using the same time resolution, we use different data sets considering separately the peak growing season (defined as the period of maximum photosynthetic activity of an ecosystem) and the whole year. One of the two data sets is divided into pre- and post-thinning data periods because the thinning of a forest (i.e. cutting down the share of trees) significantly impacts not only the NEE, but also many site variables.

We expect an ML model to learn differently depending on the seasonality of the time series used for model training. For example, diffuse radiation is an essential input variable for photosynthesis on a subhourly scale during the peak growing season because ecosystem photosynthesis is enhanced under higher diffuse radiation conditions due to better light use efficiency (Gu et al., 2002; Ezhova et al., 2018). In winter, this effect is missing, which might make diffuse radiation not as crucial variable for the model trained on the whole year data set. Instead, other input variables, such as air or soil temperature, can be relevant when focusing on the seasonal cycle of carbon fluxes (Kolari et al., 2009). Moreover, besides time-related factors, a spatial factor represented by latitude is also expected to affect the model buildup. The first aim of this study is to analyze how ML models treat input variables related to temporal (peak season vs whole year) and spatial variability.

The second aim is to use different ML models to understand how the best model's outcome compares to what we know from process understanding of the carbon fluxes' dynamics. In addition to that, we compare different ML models and check if all of them reproduce CO_2 flux dynamics robustly, if they tend to choose the same important input variables, and if dependences on these variables are similar between the models.

Finally, we combine data sets from the two latitudes, include data from a post-thinning period in Hyytiälä, and use XAI to understand how the models perform on this mixed data set. We introduce additional variables (the site variables) distinguishing between the sites and model NEE with and without these variables.

In this study, we have several research goals: 1) compare the ML models' performance for two ecosystems from different latitudes but with the same main tree species using accuracy metrics and XAI (with a linear regression model as a baseline); assess the reliability of results based on the robustness of their reproduction by different models; 2) analyze the shift in the choice of model variables and their general performance depending on the seasonality (i.e., peak growing season or the whole year) and latitude; 3) study how combining the data sets from the two studied forest ecosystems at different latitudes and including post-thinning data affects model results.

Table 1. Summary of data sets: time periods and number of observations.

Site and case	Dates	N obs.
Hyytiälä, whole year	07/2008 - 09/2018	39096
Hyytiälä, peak season	Jul-Aug (2008 - 2018)	11730
Post-thinned Hyytiälä, whole year	02/2019 - 05/2021	11690
Post-thinned Hyytiälä, peak season	Jul-Aug (2019 - 2020)	1376
Värriö, whole year	05/2013 - 10/2019	26138
Värriö, peak season	Jul-Aug (2015 - 2019)	7172

2 Materials and methods

2.1 Stations and data sets

We used atmospheric observations from the SMEAR II station in Hyytiälä, Finland (Hari and Kulmala, 2005) and SMEAR I station in Värriö, Finland (Hari et al., 1994). The stations are located in boreal forest in central Finland (Hyytiälä: 61°51'N, 24°17'E, 80 m a.s.l.) and in Finnish subarctic region (Värriö: 67°46'N, 29°36'E, 180 m a.s.l.). The mean annual air temperature is 3.5^{o}C in Hyytiälä and -0.5^{o}C in Värriö (source: ICOS database). The mean annual precipitation in Hyytiälä is 710 mm, and in Värriö, it is 601 mm. Forest stands at both sites are dominated by 60-65-year-old Scots pines (*Pinus sylvestris* L.). However, the average tree height differs, being ca. 19.9 m at SMEAR II and 10 m at SMEAR I, as measured in 2023. The forest canopy at SMEAR II is closed, and at SMEAR I, it is open. Both sites are part of the Integrated Carbon Observation System (ICOS) and Integrated European Long-Term Ecosystem, critical zone, and socio-ecological Research (eLTER) networks, meaning continuous observations of carbon fluxes and other ecosystem parameters. Meteorological variables and radiation are also routinely measured at the stations. The data is publicly available to download from the SmartSMEAR database (https://smear.avaa.csc.fi/, accessed September 2022; latest updated data sets can be found at https://etsin.fairdata.fi/datasets/SmartSMEAR).

Data from Hyytiälä was divided into two separate data sets: pre-thinning, referred to just as Hyytiälä data (prior to 2019), and post-thinning (post 2019), referred to as post-thinning Hyytiälä data. The separation is due to the thinning of the forest at Hyytiälä station in 2019, which involved the removal of smaller trees from the forest understory, and additional thinning (from below) conducted from January to March 2020. In the thinning, 30% of tree basal area was removed (Aalto et al., 2023), which significantly changed NEE due to the decrease of biomass. The data set thus had too large differences to be treated as a direct continuation of the pre-thinning data set. The amount of data points and the time intervals for each data set can be seen in Table 1.

The data used in this study was at a 30-minute interval. The high frequency enables a more detailed study of the daily cycle of NEE. It allows for the analysis of the impact of such variables that affect the ecosystem processes on a short time scale, such as the impact of changes in radiation on photosynthesis. Raw data for the target variable (NEE) being modeled with the machine learning models is first captured using eddy covariance technique (Aubinet et al., 2012) and then processed to NEE

Table 2. List of input variables used for model training.

Abbrevation	Name	Units	Notes
PAR	Photosynthetically Active Radiation	μmol s^{-1} m^{-2}	Hyytiälä: Measured at 18 m height (radiation tower 12/2009-2/2017) or 35 m height (35 m tower 2/2017-). Värriö: - .
PAR_{dif}	Diffuse PAR	μmol s^{-1} m^{-2}	Hyytiälä: Measured at 18 m height (radiation tower 12/2009-2/2017) or 35 m height (35 m tower 2/2017-). Värriö: - .
F_{dif}	Diffuse Fraction	-	$F_{dif} = \frac{PAR_{dif}}{PAR}$
AirTemp	Air Temperature	^{o}C	Hyytiälä: Measured at 33.6 m height. Värriö: 9 m
SoilTempA	Soil Temperature	^{o}C	Hyytiälä: Measured 2-5 cm depth in the mineral soil). Värriö: 5cm.
SoilTempB	Soil Temperature	^{o}C	Hyytiälä: Measured 22-29 cm depth in the mineral soil (Only in Hyytiälä)
VPD	Vapor Pressure Deficit	Pa	Formula (2), section 2.1
SoilWatCont	Soil Water Content	%	Hyytiälä: 26-36 cm depth in the mineral soil. Värriö: - .
RH	Relative Humidity	%	Hyytiälä: Measured at 16 m height (4/1998-1/2017) or 35 m height (2/2017-). Värriö: 2m.
FricVel	Friction Velocity	m/s	Hyytiälä: Measured at 24 m height, 27 after 2019. Värriö: Measured at 16.6 m height

using the EddyUH software (Mammarella et al., 2016). Negative NEE corresponds to the ecosystem acting as a net carbon sink, while positive corresponds to the ecosystem acting as a net carbon source. We model NEE using meteorological variables such as air temperature, soil temperature, solar radiation, relative humidity, and soil moisture content. LAI is not used here as its seasonal variability in the chosen period is relatively small (Hyytiälä - about 30%, Värriö - 20%), which translates to below 10% change in canopy light interception and roughly the same percentage in GPP. For some input variables, minor differences exist in how the data is measured at the two stations (e.g., soil moisture is from slightly different depths). The data used was non-gapfilled to avoid the influence of models typically used for gapfilling. At Hyytiälä, photosynthetically active radiation (PAR) was not measured before 2009, and we used global radiation multiplied by the PAR quantum efficiency of 2 μmol s^{-1} W^{-1} (Ross and Sulev, 2000; Ezhova et al., 2018) to calculate missing values of PAR. All variables used are listed in Table 2.

In the pre-processing of the data, time points that contained missing values of any studied input variable were discarded. Also, all rows where the PAR value was less than 10 μmol s^{-1} m^{-2} were filtered out due to the interest being solely on modeling daytime NEE. We calculated the diffuse fraction:

Table 3. Overview of the training configurations for ML models across different datasets.

Set	Setup	Description
Set 1	Hyytiälä All	Models trained on the data from pre-thinned Hyytiälä, entire years
	Hyytiälä Peak	Models trained on the data from pre-thinned Hyytiälä, peak growing seasons
	Värriö All	Models trained on the data from Värriö, entire years
	Värriö Peak	Models trained on the data from Värriö, peak growing seasons
Set 2	All Without Site	Models trained on the mixed data set from both sites, including post-thinned Hyytiälä, entire years, no site labels
	All With Site	Models trained on the mixed dataset from both sites, including post-thinned Hyytiälä, entire years, sites labels included
	Peak Without Site	Models trained on the mixed dataset from both sites, including post-thinned Hyytiälä, peak growing seasons, no site labels
	Peak With Site	Models trained on the mixed dataset from both sites, including post-thinned Hyytiälä, peak growing seasons, site labels included

$$F_{dif} = \frac{\mathrm{PAR}_{dif}}{\mathrm{PAR}}, \tag{1}$$

and vapor pressure deficit (Monteith and Unsworth, 2013):

$$\mathrm{VPD} = e_s - e_a, \text{ where } e_s = 611 \exp(\frac{17.27 T_{air}}{237.7 + T_{air}}),\ e_a = e_s \frac{\mathrm{RH}}{100}. \tag{2}$$

In eq. (2), T_{air} is in units [^{o}C] and e_s, e_a are in units [Pa].

The machine learning models were trained in two sets of four setups (Table 3), and the results within a set were compared against each other. For both sets, four different machine learning models were trained for all of the four cases meaning total of thirty-two models trained. In the first set, models for data representing entire year and peak growth season (July and August) were trained using data from either pre-thinned Hyytiälä or Värriö. In the second set, models were trained by combining the data from two sites into a single mixed dataset and then training them with and without variables that denote from which site the data originates from ('Värriö', 'Hyytiälä' for Hyytiälä pre-thinned, 'HyytiäläT' for Hyytiälä post-thinned). Similarly to Set 1, setups included entire year and peak growing season. A summary of the configurations for all experiments can be seen in Table 3.

In all cases, the data was split into training and test data, where training data was used to train the models, while test data was used to evaluate the models' performance. For modeling NEE for pre-thinned Hyytiälä and Värriö, 75% of their respective data was used for training the model, while the rest was used as the test data to evaluate the model performance. In case of the mixed model, 80% of the each respective data set was used to train the model.

2.2 Machine learning models

To ensure robustness and reduce potential biases, we validate our findings across four distinct ML models, aiming to identify consistent patterns or insights and provide an overall picture of how well the models can use this data to predict NEE. Applying several models to the same data set provides a context on what input variables are consistently considered important across different models. The four models used were Cubist (Quinlan, 1992), Random Forest (Breiman, 2001), Averaged neural network (Kuhn, 2008), and basic Linear Regression (Kutner et al., 2004). All were implemented in R (v. 4.3.0: https://www.r-project.org/) using R's "caret" library (v. 6.0.94: https://github.com/topepo/caret/). Linear Regression served as the baseline model, while the other models were chosen due to their proven competence in solving various regression problems (Fernández-Delgado et al., 2019).

Random forest (RF) is a popular model that has been used in previous research (Cai et al., 2020; Liu et al., 2021; Abbasian et al., 2022; Zhu et al., 2023) due to its ease of use, high accuracy, and robustness. It is an ensemble model that uses the averaged output of random regression trees (Fernández-Delgado et al., 2019) by training different regression trees on different subsets of the data. The final prediction is the average result of the different tree predictions. The algorithm is quite robust as the different trees are trained with the different subsets of the training data. The randomForest library (Liaw and Wiener, 2002) implements the regression algorithm of RF used in this study.

Cubist is one of the best-performing regression models (Fernández-Delgado et al., 2019) across multiple types of data sets (i.e., type and size of data). Like RF, it is created from multiple individual regression trees, where each terminal leaf contains a smoothed linear regression model for prediction (Zhou et al., 2019). It creates a series of "if-then" rules that can be considered the branches of a tree, while the leaves are an associated multivariate linear model. The corresponding model is used to calculate the final predicted value as long as the set of covariates satisfies the conditions of the corresponding rule. Cubist also uses boosting with its training committees, which creates a series of trees with different weights and nearest-neighbors search to adjust the predictions better.

Model Averaged Neural Networks (avNNet) is a single hidden layer feed-forward neural network characterized by its architecture and training approach. The network consists of interconnected neurons arranged in layers, with the final layer outputting the prediction (Ripley, 2007). During the training phase, initial weights, which influence predictions, are randomly assigned. These weights are then iteratively updated, enabling the network to capture nonlinear relationships. Given the randomness in predictions due to these initial weight assignments, avNNet constructs multiple neural network models and averages their results. This averaging process promotes a more robust and stable prediction by minimizing the impact of any single model's randomness.

The basic multivariate Linear Regression (LinRegr) is used as a baseline to understand how much impact and improved results more advanced models can provide. LinRegr finds a linear relationship between the independent and dependent variables determined by minimizing the sum of the squared differences between the predicted and the actual values (Hastie et al., 2009).

Table 4. List of the final model hyperparameters with their respective values for each modelling setup. Values of parameters are listed in the following order corresponding to different setups: Hyytiälä All, Hyytiälä Peak, Värriö All, Värriö Peak, Mixed data sets with site label and Mixed data sets without site label.

Method	Hyperparameter	Description	Values
Cubist	committees	Number of committees (models) to be fitted.	100, 90, 100, 100, 100, 100
	neighbors	Number of nearest neighbors used in prediction.	9, 9, 6, 3, 9, 6
Random Forest	mtry	Number of variables sampled at each split.	3, 3, 6, 2, 11, 8
	min node size	Minimum size of terminal nodes (leaves).	5, 5, 5, 5, 5, 5
avNNet	size	Number of units in the hidden layer(s).	13, 13, 13, 13, 13, 13
	decay	Weight decay parameter for regularization.	0.1, 0.1, 0.1, 0.1, 0.1, 0.1
	bag	Boolean flag for using bootstrap aggregating (bagging).	False, False, False, False, False, False

2.3 Cross-validation framework, hyperparameter tuning and validation metrics

K-fold cross-validation is a resampling method for validating model efficiency, which generally results in less biased models (Jung, 2018). K-fold cross-validation method shuffles the data set randomly and splits it into K groups or folds. First, each fold is taken as a holdout, while the model is fit on the rest of the folds, and then the model is evaluated on the holdout set. The score is retained, and the model is discarded. In repeated K-fold cross-validation, this process is done R times on different splits. K-Fold cross-validation also effectively prevents model overfitting, where a machine learning model has learned to model the inherent noise of a dataset, to a point where it fails to model for points not included in the training dataset (Berrar, 2019).

During the model training, repeated K-fold cross-validation was used with Caret librarys (Kuhn, 2023) grid hyperparameter search. This method trains and evaluates a model using all possible combinations of specified hyperparameter values to identify the combination that yields the best model performance. It was used to tune the models' hyperparameters and configuration settings that are external to the model and can be adjusted to optimize performance. Values $R = 5$ repeats and $K = 10$ folds were used to fit each model. The tuned hyperparameters can be seen in Table 4. The train and test data as well as the folds of the K-fold cross-validation were split using a predetermined random split to ensure repeatability. However, due to technical limitations, in-depth hyperparameter tuning was not used on the models that contained data from all sites. Instead, hyperparameters based on the results from the single-site models were used.

In evaluating the performance of our machine learning models, we primarily relied on two key metrics to assess the models' goodness of fit: the coefficient of determination (R^2) and the root mean squared error (RMSE). RMSE measures the differences between the values predicted by a model and the actual values and provides an understanding of the magnitude of error the

model might make in its predictions. A lower RMSE indicates a better fit to the data, implying that the model's predictions are more precise. The models' hyperparameters were tuned specifically based on the RMSE score.

In addition, each model was trained on five different data splits to account for variability and reduce the influence of any single fortunate or unfortunate split on the results. The performance metrics, R^2 and RMSE, were averaged across these splits to ensure a robust and reliable assessment of model performance.

2.4 Explainable AI Methods

As machine learning models have been used more in research and industry, the demand for more transparent and interpretable models has grown (Dwivedi et al., 2023). As model accuracy has risen, so has model complexity. The highly accurate and complex models have many hyperparameters that can not be made human-understandable. To be trustworthy, the ML model must produce interpretable or transparent results. Relying on unexplained or inaccurate predictions can lead to critical errors. Accuracy metrics do not always portray the true prediction capability of a model, so it is vital to critically evaluate the results against existing knowledge or theories. XAI methods aim to provide machine learning models and methods that enable users to better understand, analyze, and evaluate the models' decision-making.

In this study, we used two XAI methods: permutation feature importance and accumulated local effect (ALE) plots (Molnar, 2020). They provide insight into how the input variables affect a model's output. Both are model-agnostic global methods, meaning they can be used regardless of the selected model and provide interpretations on the data set as a whole rather than individual points (Molnar, 2020). Both of these methods were implemented using R's "iml" library (v.0.11.1: https://github.com/christophM/iml/, Molnar et al. (2018)).

2.4.1 Permutation Feature Importance

Permutation feature importance is a method that aims to measure the increase in the prediction error of a model after the input variables (features) are permuted. In permutation feature importance, the relationship between a specific input variable and the variable the model tries to predict is deliberately disrupted to understand how the models' prediction accuracy is affected (Molnar, 2020). If an input variable is important, randomly rearranging its values increases the model error, as the model then relies on that specific input variable for an accurate prediction. Trained model is denoted as f, input variable matrix as $\mathbf{X}$, target vector as $\mathbf{y}$, and error measure $L(\mathbf{y}, f(\mathbf{X}))$. The algorithm works as follows:

1. Estimates the original model error $e = L(\mathbf{y}, f(\mathbf{X}))$
2. For each input variable with index $i \in \{1, ..., p\}$, where p is the total number of input variables, the following is done:
    2.1 Generates a permutated input variable matrix $\hat{\mathbf{X}}$ by permuting input variable i in the data $\mathbf{X}$, which breaks the association between input variable i and the true outcome $\mathbf{y}$.
    2.2 Estimates the error caused by the permutation by predicting with it $\hat{e} = L(\mathbf{y}, f(\hat{\mathbf{X}}))$.
    2.3 Calculates permutation input variable importance as quotient $Imp_i = \hat{e}/e$.

3. Sorts input variables by descending Imp.

Only test data is used to calculate the permutation feature importance. Assessing feature importance using the training data might result in too inflated scores due to a model overfitting on training data. That said, the features with very high scores might not be as important for making accurate predictions on new, unseen data. As with the metrics R^2 and RMSE, the Permutation Feature Importance was calculated on multiple different datasplits to ensure robustness of the results.

2.4.2 ALE Plots

Accumulated local effect (ALE) plots describe how input variables influence the prediction of a machine learning model on average (Molnar, 2020). ALE reduces a complex machine learning function to a function that depends on only one, as in our case, or two input variables and visualizes the effects between a selected variable and the prediction of the target variable of a machine learning model. The idea is to remove the unwanted effects of other input variables, take partial derivatives (local effects) of the prediction function with respect to the feature of interest, and integrate (accumulate) them with respect to the same feature.

The value of ALE at a certain point can be thought of as the effect of the selected variable at a specific value compared to the average prediction made on the data. To calculate ALE value for input variable s at point $x \in [min(\mathbf{x_s}),\ max(\mathbf{x_s})]$, with $\mathbf{x}_s$ being the vector of this variables values, the input variable values $\mathbf{x}_s$ are divided into K intervals, where the start of the first interval is the lowest value $z_0 = min(\mathbf{x}_s)$, and the differences of predictions between the sequential intervals is calculated. While the exact ALE formula requires a model with a derivative, an approximate version is used here that is more widely adopted and works for models without a derivative. Initially, an uncentered effect is computed:

$$\bar{f}_{s,ALE}(x) = \sum_{k=1}^{k_s(x)} \frac{1}{n_s(k)} \sum_{i:\ x_s^{(i)} \in]z_{k-1,s}, z_{k,s}]} \left[f(z_{k,s}, \mathbf{x}_{-s}^{(i)}) - f(z_{k-1,s}, \mathbf{x}_{-s}^{(i)}) \right].$$

The input variable of interest is replaced with grid values $\mathbf{z}$, where the grid values represent the edges of the intervals. The interval index an input variable value $x \in \mathbf{x}_s$ falls in is denoted as $k_s(x)$, while $n_s(k)$ denotes the number of observations inside the k-th interval of $\mathbf{x}_s$. A single data point is denoted as $\mathbf{x}^{(i)} = (x_s^{(i)}, \mathbf{x}_{-s}^{(i)}))$, where $x_s^{(i)}$ denotes the i-th value for the selected input variable, and $\mathbf{x}_{-s}^{(i)}$ is the vector of all the other features of a single data point that are kept constant. The ML predicting function is denoted as f.

The differences between the predictions $f(z_{k,s}, \mathbf{x}_{-s}^{(i)}) - f(z_{k-1,s}, \mathbf{x}_{-s}^{(i)})$ are the effect that the input variable s has for an individual data point to predicting the dependent variable (NEE in our case) when using the upper and lower values of an certain interval. The sum $\sum_{i:\ x_s^{(i)} \in]z_{k-1,s}, z_{k,s}]}$ adds up the effects of all instance within an interval $x_s^{(i)} \in]z_{k-1,s}, z_{k,s}]$. This is then divided by the number of observations in this interval $n_s(k)$ to obtain the average difference of the predictions of this interval. The sum $\sum_{k=1}^{k_s(x)}$ accumulates the average effects across all intervals, meaning that the uncentered ALE of an input variable of interest is accumulated by all its previous intervals. After that, the effect is centered, making the mean effect zero:

$$f_{s,ALE}(x) = \bar{f}_{s,ALE}(x) - \frac{1}{n}\sum_{i=1}^{n} \bar{f}_{s,ALE}(x_s^{(i)}).$$

The value of ALE can be thought of as the main effect of the input variable at a certain value compared to the average prediction of the data. ALE plot has the advantage that it generates valid interpretations even if the variables are correlated, an issue that persists in other methods that reduce a prediction function f to a function that depends on a single input variable such as PDP or M-plots (Molnar, 2020). As with permutation feature importance, only the test data set was used to reduce the chance of inflating scores due to a model overfitting on the training data set.

3 Results and discussion

3.1 NEE modelling for Hyytiälä and Värriö data sets

In this section, we report the results obtained with different models from the Set 1 in Table 3 (pre-thinned Hyytiälä and Värriö, whole year and peak growing season). First, we assess models' performance with routinely used accuracy metrics (R^2 and RMSE), visualize diurnal and annual NEE cycles, and then use XAI methods. In each subsection, we start the discussion with the peak growing season results and continue with the whole season results.

3.1.1 Assessing model performance using accuracy metrics

Figs. 1 and 2 show coefficients of determination and RMSE, respectively, for all the models, two stations, the peak growing season, and the whole year (Set 1 in Table 3). In general, the models perform better if trained on the Hyytiälä data set compared to the Värriö data set, as seen from higher R^2-coefficients. If the model is used on the training data set, the R^2-coefficients and RMSE are somewhat better than when used on the test data set, as expected. This effect is especially pronounced for RF and Cubist models, which achieve high scores (R^2>0.85), largely because they are regression tree-based models that tend to produce overly optimistic results on the data they were trained on. These high training scores do not reflect Out-of-Bag (OOB) performance, which typically provides a more accurate estimate of the model's true predictive ability on the data it was trained on (Kuhn and Johnson, 2013), due to it not being available to use on all models. The difference between the train and test scores is larger for Värriö than for Hyytiälä, as can be expected because Värriö data set is smaller (Zhang et al., 2023). LinRegr and avNNet have almost identical scores on training and test data sets. The difference in scores between the training and test data sets is called generalization error. In some cases, large generalization error points to overfitting, i.e., the model learns the training data set too well and then performs poorly on the test data set. We applied K-fold cross-validation to avoid overfitting when choosing hyperparameters; see subsection 2.3. Additionally, we tried different splits of the data into training and test data sets, which showed that the variation of the resulting R^2-coefficients and RMSE was small (Fig. 1, 2). In addition, we obtained similar accuracy metrics on the test data sets using different nonlinear ML models, which also suggests that our results are robust. In what follows, the results are reported for the test data sets if not stated otherwise.

For the peak growing season, all four models perform well, including LinRegr, which is only slightly worse than the more complex models. For Hyytiälä, all nonlinear ML models give similarly high R^2-coefficients, close to 0.9, and RMSE values almost do not differ between these models. For Värriö, RF is slightly better than other ML models demonstrated by both higher R^2-coefficient and lower RMSE. Compared to Hyytiälä's $R^2 = 0.87$, Värriö's R^2-coefficient is lower, $R^2 \simeq 0.70-0.74$, which could be related to the higher share of outliers in the data or a smaller range of the predicted variable. The predictors vary within similar ranges in Hyytiälä and Värriö, whereas the predicted variable NEE has a larger value range in Hyytiälä compared to Värriö (corresponding to a weaker carbon sink in Värriö) because Värriö ecosystem is less productive. It is also possible that the difference in R^2-coefficients could be because the available predictors have a more significant effect on the forest carbon balance in Hyytiälä than in Värriö. A decrease in R^2-coefficients for the cases when the predicted variable had a smaller value range was reported by other ML studies, e.g., Liu et al. (2021) and Abbasian et al. (2022). Also, for process-based models, reproducing carbon fluxes at less productive forests with low leaf area index is challenging (Mäkelä et al., 2019).

Scatter plots of measured vs. modeled data for training and test data sets are shown in Fig. 3 using one of the best performing models, RF. The lowest modelled NEE values tend to be overestimated, and the highest underestimated. This is seen best in the training data sets (because they are much larger) deviating from 1:1 lines at the extremes of the data. In Fig. 2, it is visible that RMSE values for Värriö are lower than those for Hyytiälä, which means that Värriö values in Fig. 3 are closer to the best-fit lines. Still, it does not mean that the model is better because the best-fit line of the measured vs. modeled data points is not 1:1. By high accuracy scores, the mean diurnal cycle of NEE within the peak growing season is almost perfectly reproduced by the RF model (Fig. 4) with slightly smaller standard deviations in the modeled than measured data.

In the case of the whole-year data sets, the performance of LinRegr drastically decreases when compared to the peak growing season data sets (Figs. 1, 2). This could be expected because, on the whole-year scale, NEE dependence on many variables becomes nonlinear. Especially for the Värriö data set, LinRegr R^2-coefficient falls below 0.5, and RMSE increases by 40% compared to the nonlinear ML models, meaning that more complex models are needed and justified. Figs. 3 - 5 show scatter plots and annual daytime NEE cycle for Hyytiälä and Värriö. The same conclusions as for the peak growing season data sets apply here as the mean values were almost perfectly reproduced and extreme values missing. The models captured the essential features of the annual NEE cycle, including ecosystem spring and autumn phenological transitions (Fig. 5).

It is interesting to consider different models' performance for the same setup. Here we show an example for Hyytiälä All setup (Fig. 6). The test cases for all nonlinear ML models look similar. Note orange points (test RF) covering black points (training RF) illustrating the smaller RMSE for the training data set in Fig. 6. LinRegr plot is more scattered, and the points are not organized along one line (in agreement with reported low R^2-coefficients and high RMSE).

Compared to other studies, Dou and Yang (2018) demonstrated that in modeling whole-year NEE of forest ecosystems, the R^2-coefficients as high as 0.64-0.80 can be reached on the test data sets for separate evergreen needleleaf forest ecosystems. Our scores are within this interval for Värriö and significantly higher (0.90) for Hyytiälä. However, we used a different, more diverse set of input variables and modeled half-an-hour fluxes compared to daily fluxes in the study mentioned above. On a similar data set (deciduous forest in Germany, summer time, half-an-hour resolution), Moffat et al. (2010) got $R^2 = 0.93$ and RMSE of 2.3 μmol s^{-1} m^{-2} using artificial neural network, which is close to Hyytiälä results.

3.1.2 Which variables explain NEE: feature importance

We now consider feature importance, allowing us to analyze how the models rank input variables by their explanatory power. For the peak growing season, all nonlinear ML models agree for both stations (Fig. 7, Table A2) that the variables with the most explanatory power are PAR and diffuse PAR. Moreover, PAR typically comes first, except for Cubist in the case of Värriö. Overall, during the peak growing season in boreal forests, a daytime CO_2 flux due to photosynthesis prevails over that due to respiration, at least in Hyytiälä (Kolari et al., 2009). Therefore, one can expect that parameters controlling photosynthesis also dominate the NEE response. PAR is theoretically the most important variable during the peak growing season to explain photosynthesis (Palmroth and Hari, 2001; Moffat et al., 2010), and the stimulating effect of diffuse radiation on the peak season photosynthesis (diffuse radiation fertilization) is also well-known (Gu et al., 2002). Accordingly, the models consider light-related variables to be the most important. Interestingly, LinRegr chooses diffuse PAR as the most important variable to explain NEE, likely because the dependence of photosynthesis on diffuse PAR can be considered closer to linear.

The third variable in importance after PAR and diffuse PAR, as seen by nonlinear models, is VPD (3 cases), air temperature (2 cases), or soil temperature A (1 case). It is good to note that VPD is calculated based on air temperature (see Sec. 2.1), so these variables are not independent. PAR, diffuse PAR, and VPD are confirmed as essential drivers of carbon assimilation in numerous studies on photosynthesis in different ecosystems (Gu et al., 2002; Larcher, 2003; Grossiord et al., 2020). Particularly for Hyytiälä during the growing season, a statistical model showed that daily photosynthesis is most sensitive to light and VPD (Peltoniemi et al., 2015). However, as NEE is the net result of photosynthesis and respiration, and respiration is highly sensitive to temperature, it makes sense that the models pick either VPD or temperature as the third important variable. Ecosystem respiration is the sum of aboveground and belowground respiration, but soil temperature is sometimes considered a better parameter for modelling ecosystem respiration than air temperature (Kolari et al., 2009; Lasslop et al., 2012).

We note that nonlinear ML models typically place several variables close to the third position in the feature importance diagram. For Hyytiälä, RF places diffuse fraction close to VPD, followed by air temperature and RH; Cubist and avNNet place intercorrelated soil temperature A and B (R = 0.98, Fig. A1) high. For Värriö, Cubist and avNNet place interdependent VPD, RH, and air temperature in the feature importance diagram within the error bar from each other. Relatively large error bars for these variables suggest that the models seem to have difficulties ranking them, as their order may likely change depending on the data split. At the same time, the error bars are smallest for RF, which seems more confident than other nonlinear models in its treatment of interdependent variables.

Suppose the model chooses one variable before another correlated one. In that case, the second one can be placed low in the feature importance diagram, as the model, in principle, does not need it anymore. This does not mean, however, that one of the correlated variables explains NEE clearly better than the other: for example, Moffat et al. (2010) showed, using an artificial neural network, that intercorrelated diffuse fraction and diffuse radiation (as well as intercorrelated VPD and RH) have the same explanatory power for the summertime forest NEE, and can be used interchangeably. However, all our models place diffuse PAR higher than diffuse fraction, and they typically place VPD higher than RH.

Feature importance for the whole-year setups (Fig. 7) shows another set of most relevant variables lifting soil temperature at the expense of air temperature or VPD (nonlinear models, see Table A2). In many cases, soil temperature becomes the second important variable, sometimes even the first (avNNet, Hyytiälä). The increasing importance of the temperature-related variable is expected because, in the whole-year case, the model needs to capture the seasonality of carbon flux (Mäkelä et al., 2004, 2006), and soil temperature grows in summer and decreases in winter. However, the models' choice of soil temperature over air temperature requires additional explanation. Presumably, the soil temperature is positive during the warm season and nearly constant during winter in the presence of snow. This behavior is in line with NEE, which is also inhibited in winter. Air temperature, in contrast, may display significant variability also in winter. In addition, soil temperature limits plant water use and photosynthesis in spring and autumn (Wu et al., 2012; Lintunen et al., 2020). In the case of the LinRegr, PAR is no longer among the three most important variables, replaced by another soil temperature or diffuse fraction.

3.1.3 How the models use input variables: ALE

Proceeding with ALE, we discuss dependencies of NEE on input variables as seen by the models, focusing on the peak growing season so far. ALE demonstrates that NEE decreases with increasing PAR and diffuse PAR for all the models (Fig. 8). Nonlinear models show the nonlinear dependence of NEE on PAR, which is most pronounced for the RF model. This model shows that NEE saturates at higher PAR values, resembling the light response curve, and for the Värriö data set, NEE levels off at the largest diffuse PAR. This could be because high diffuse PAR is observed under a cloudy sky, and in Värriö, the corresponding PAR level can already be close to the light saturation point (Ezhova et al., 2018) inhibiting photosynthesis.

RF and Cubist also capture a nonlinear dependence of NEE on VPD, which has an optimum value between the low and high values of VPD. At very high VPD, stomatal closure prevents plants from losing water (Running, 1976), also affecting photosynthesis. Besides, high VPD is often associated with high temperature, which increases NEE due to increased respiration. At low VPD, when water vapor pressure at the leaf level and in the atmosphere is about the same, there is no driving force to sustain transpiration. This inhibits water uptake by the roots and generally slows down plant metabolism, affecting photosynthesis. Moreover, low VPD is associated with lower PAR and higher diffuse fraction (Fig. A1), pointing at overcast cloudy conditions when photosynthesis is light-limited.

Note that dependencies of NEE on PAR, diffuse PAR, and VPD are qualitatively similar in all used nonlinear models, though quantitatively, sensitivity to the corresponding variables somewhat differs. However, the dependence of NEE on air temperature is not the same in all models. In Hyytiälä, RF and Cubist feature an increase of NEE with air temperature, whereas LinRegr and avNNet demonstrate a decrease. In Värriö, all models except avNNet suggest a positive dependence of NEE on air temperature. The positive dependence is in line with the stomatal control at high temperatures (stomatal closure dampening photosynthesis) and higher soil respiration during the peak growing season.

It is interesting to analyze ALE from different models trained on the input data sets with several temperature variables. Both soil and air temperature are typically included in modeling studies of NEE based on machine learning (Dou and Yang, 2018; Liu et al., 2021; Abbasian et al., 2022). Cai et al. (2020) and Wood (2021) include average, minimum, and maximum air and soil temperature in their studies, adding more interdependent variables in the data sets. Hyytiälä's data set includes

air temperature and temperature from A and B soil horizons. In the peak season, all these temperature-related variables show quite similar dynamics. With soil depth, the mean temperature and amplitude of the diurnal temperature cycle decrease, and the time lag between the temperature signals increases. However, horizon B is not too deep and the lag remains generally smaller than half a day. All the models, besides RF, treat soil temperatures A and B as important variables and demonstrate strong but opposite dependencies on these variables (Fig.8). As soil temperatures A and B are correlated (Appendix A, Fig. A1), opposite NEE dependencies must outweigh each other. Strong opposite dependencies on correlated variables should be treated cautiously as the models might use them to tune towards higher scores on given data sets. In the case of correlated soil temperatures, there is no guarantee that this compensation or tuning will work for even higher temperature, which is currently not represented in the data set. The same conclusion applies to using the model developed for a particular site on the data sets from other sites (Peltoniemi et al., 2015). In contrast, RF shows a strong association of NEE only with air temperature and a weak association with two soil temperature variables.

Now we briefly discuss other variables that have a more minor effect on NEE. Diffuse fraction demonstrates a consistent impact across all models, leading to some increase in NEE with its rise. This effect likely stems from the reduction of photosynthesis under an overcast sky with low radiation and high diffuse fraction. Note that diffuse fraction and diffuse PAR contain the same information provided PAR is included in the data set. Gross primary production in Hyytiälä has its minimum at the low diffuse PAR and a maximum at the high diffuse PAR compared to the weak parabolic dependence on diffuse fraction (Ezhova et al., 2018; Neimane-Šroma et al., 2024). That may be why the models choose diffuse PAR over diffuse fraction. Most models could then deem the diffuse fraction relatively unimportant as they already use diffuse PAR.

RH directly influences VPD through a linear relationship (eq. (2), Fig. A1). The higher the RH, the closer ambient air is to saturation, and VPD, in this case, is small. Low RH, vice versa, favors higher VPD values. Having VPD as one of the powerful explaining variables should, in principle, diminish the role of RH, as is the case for RF and Cubist. However, RH is placed relatively high in the feature importance for avNNet and LinRegr, which is also reflected in the significant range of NEE variability due to RH.

In Hyytiälä, all nonlinear models feature an increase in NEE with decreasing soil water content. In Värriö, all models feature an increase of NEE with increasing soil water content, and in Hyytiälä, Cubist and avNNet demonstrate similar behavior. Note, however, that sensitivity to this variable is quite low for all models, indicating that soil moisture does not limit ecosystem functioning in current conditions. However, this could change in the future, which would perhaps not be captured by the models.

For friction velocity, all models indicate a consistent trend in Hyytiälä, where an increase in friction velocity leads to a decrease in NEE, suggesting that NEE flux is somewhat sensitive to changes in turbulence levels. On the one hand, this could indicate an eddy covariance problem (Moffat et al., 2010). On the other hand, this dependence might reflect physical processes: friction velocity has a weak increasing trend in Hyytiälä due to trees getting taller, which coincides with the weak, increasing trend in carbon uptake but not in respiration (Launiainen et al., 2022). In Värriö, there is no clear dependence between friction velocity and NEE. Generally, this variable holds limited importance in the overall model predictions, which is to be expected as filtering by friction velocity is applied to the data sets routinely during quality checks.

We proceed with the whole-year ALE plots: the dependencies of NEE on light variables (PAR, diffuse PAR) remain largely similar to those for the peak growing season setup (Fig. A3). Most nonlinear models (except avNNet on Hyytiälä data set) predict that the NEE dependence on air temperature has a minimum in the presence of negative temperatures in the data set, suggesting larger NEE during the cold season and the warmest summer periods. This might reflect the absence of photosynthesis in the cold season and the increased respiration accompanied by inhibited photosynthesis for the highest temperatures. In Hyytiälä, NEE dependence on soil temperature A also has a minimum. In Värriö, NEE decreases with increasing soil temperature until it plateaus at around 15°C in the case of RF and avNNet. Note that for Hyytiälä, NEE dependencies on soil temperatures B and A are again of opposite sign for all models except RF. The LinRegr fails on the Hyytiälä data set, showing a weak association of NEE with air temperature but featuring lower NEE and stronger carbon sink at low, even negative soil temperatures. The failure of LinRegr on the whole-year data could be due to its inability to capture the nonlinear dependence of NEE on temperature, which becomes significant on a whole-season scale.

Considering less critical variables, the dependencies remain mainly the same. In some cases, however, avNNet demonstrates dependencies inconsistent with expected behavior, e.g., featuring a stronger carbon sink under low RH conditions. It is worth mentioning that the dependence on soil water content is quite complicated in Hyytiälä, with a minimum and a maximum. This could be related to data containing subsets with high water content at low temperatures when photosynthesis is inhibited, e.g., during snowmelt or late autumn. In any case, as for the peak growing season setup, the sensitivity of NEE to this variable is low.

Finally, if the most important input variables for the studied sites are the same and the dependencies of NEE on these variables are similar in the case of RF and, to a lesser extent, Cubist, one could expect that it is possible to build a more generic model, which would be able to give reasonable results for many different boreal forest sites. We, therefore, built one model based on all the data in the following section.

3.2 NEE modelling: mixed data set

In this section, we report the results of NEE modeling using Set 2 (Table 3), which consists of mixed data from pre-thinning Hyytiälä (referred to just as Hyytiälä), Värriö and post-thinning Hyytiälä. We aimed to understand how the models perform in the following cases: 1) mixed data set, containing data from both sites without any separation or benchmarking the data (setups Peak Without Site and All Without Site, Table 3); 2) mixed data set, but we introduce three binary dummy variables that identify the site (setups Peak With Site and All With Site, Table 3). Three binary variables were used instead of a single categorical one due to some models requiring real numbers as input (Hancock and Khoshgoftaar, 2020).

3.2.1 Assessing model performance on a mixed data set using accuracy metrics

The determination coefficients for mixed data sets are shown in Fig.9, separately for the model runs with and without the variables for the site identity. Adding site variables to the data set slightly improves the correlation coefficient R^2 (within 3.5%), which remains high for the best models, RF and Cubist (0.84-0.87 for the peak season, 0.86-0.89 for the whole season). Comparing this result to the results for the separate stations (Fig. 1), we note that the scores are closer to those for Hyytiälä.

This could be because Hyytiälä data prevails in the compiled data set. However, a trial run with equal inputs from different sets (Hyytiälä pre-thinned + Hyytiälä post-thinned + Värriö) shows that R^2 was only marginally lower, by 0.02 for the nonlinear ML models (Figs. A5, A6). This finding suggests that factors other than the prevalence of the Hyytiälä data set may be important: for example, the value range of the data. Hyytiälä data set has a larger NEE value range compared to Värriö, and that could be the reason for better Hyytiälä R^2-coefficient, as mentioned in sec. 3.1.1. Therefore, one could expect larger R^2-coefficients for any mixed sets containing a sufficient amount of Hyytiälä data when compared to the Värriö data set. Interestingly, LinRegr performs worse than other models on a compiled data set, even for the peak growing season. The LinRegr R^2-coefficient on the mixed data set is clearly lower than on the Hyytiälä data set (drop from 0.85 to 0.80 for the peak growing season and from 0.76 to 0.68 for the whole year).

As said, site variables do not have a significant effect on R^2-coefficients, but the advantage is more evident for RMSE (Fig. 10). RMSE for the peak growing season data is generally larger than for the whole season, likely because the fluctuations and errors of NEE measurements outside the growing season are relatively small. In addition, the flux random error increases with the flux magnitude within the growing season. If we compare the models with and without site variables, we see that adding site variables reduces RMSE by 10-13%: from about 2.4 to 2.15 μmol s^{-1} m^{-2} for the peak growing season and from about 1.8 to 1.6 μmol s^{-1} m^{-2} for the whole year. Considering models trained on data of separate sites, RMSE scores in the models with site variables are somewhat smaller compared to the models trained on the Hyytiälä data set, probably due to the presence of Värriö data with smaller RMSE than Hyytiälä data (Fig. 2) or due to the larger size of the mixed data set. Overall, introducing the site variables in the mixed data set barely improves the correlation between measured and modeled points but reduces the scatter in the plot presenting measured vs modeled points.

3.2.2 Feature importance for the mixed data set

We assess the feature importance diagrams provided by the models on the mixed data sets, paying special attention to the ranking of the site variables (Fig. 11, Table A2). It follows quite clearly from Table A2 that the models' choice of the most important input parameters becomes more aligned when they are trained on the mixed data set. For example, all the models, without exception, choose PAR as the most important variable in both peak-season and whole-year setups. During the peak season, the second variable in the feature importance diagrams is diffuse PAR (6 setups out of 8) or VPD (2 out of 8). Continuing with the peak season, site parameter 'Värriö' appears only as the third variable in the corresponding setups (replaced in some models with VPD or diffuse PAR). In the setup without site parameters, the third important variable is VPD or diffuse PAR or soil water content in case of RF. The latter observation is interesting as Värriö has different soil characteristics: soil moisture is lower there (Fig. A2), and RF might have used it as a replacement of the site variable.

For the whole-year setups, the second variable after PAR in the feature importance diagrams is almost always soil temperature A (7 out of 8 cases) or diffuse PAR (1 out of 8 cases). (Recall that the Hyytiälä feature importance set contained Soil temperature B, not A. Replacement of this variable by the temperature at A horizon is because Soil temperature B is not in the data set anymore as it was not measured in Värriö). The third variable is diffuse PAR (5 out of 8 cases) or VPD/soil temperature/air temperature (one case each). Note that site variables are not among the three most important in the whole-season

setups. Instead, the models retain NEE dependence on soil temperature, which allows them to reproduce a seasonal cycle and choose over the site variables diffuse PAR, VPD or air temperature. Nevertheless, site variables appear among the six highest input variables in the feature importance diagrams, and as follows from Fig.10, they help to reduce the RMSE.

Another observation is that among site variables, the models put 'Värriö' highest in the peak-season setups but 'Hyytiälä' in the whole-year setups. However, as mentioned before, it should be possible for the models to use them interchangeably.

3.2.3 ALE for the mixed data set

Judging by ALE (Figs. 12, A4), dependencies of NEE on light variables (PAR and diffuse PAR) for all setups in Set 2 are similar to those for separate stations (Figs. A3, A3). In the peak season, the nonlinear models suggest that the third important variable is 'Värriö' for the setups with site parameters. From Fig. 12, it can be seen that the modeled NEE increases if 'Värriö' changes from zero to one. The models then use this site variable to make all NEE values at Värriö somewhat higher than the general mean value for all three sites, which is the case due to lower tree biomass. Similarly, models use the variable 'Hyytiälä' when it is equal to one to decrease NEE, and this decrease is less pronounced for Cubist and RF than for the other models. Finally, when 'HyytiäläT' variable is equal to one, RF and Cubist slightly increase NEE, whereas the other models decrease NEE. Because the prevailing data set is still Hyytiälä pre-thinned, this data set likely dictates the base values chosen by the models. Therefore, a moderate increase of NEE for the Hyytiälä thinned data set and a stronger NEE increase for the Värriö station is reasonable. Interestingly, LinRegr does not use the site variable 'Värriö' at all.

In the peak-season setup without site variables, soil water content is one of the relevant variables chosen by the models, especially RF. Judging by ALE (Fig. 12), the models prescribe higher values of NEE to the drier cases, which is in line with how the ecosystem functions under drier conditions (reduction of photosynthesis). Similarly, for the whole-season setup without site variables, we note that NEE decreases strongly with increasing soil water content (Fig. A4), in contrast to what was observed when we modeled separate sites. ALE plots for both the peak-season and whole-season setups (Figs. 12, A4) demonstrate clearly how soil water content loses its strong position when site variables are introduced and how the NEE dependence on this variable again becomes complex, in line with what is observed for separate stations.

NEE dependencies on VPD are qualitatively similar for mixed data sets in those for separate sites. LinRegr and avNNet still have strong and opposite NEE dependencies on VPD and RH, similar to their performance on the Hyytiälä data set. These models might use variable RH to compensate for a too-strong modelled effect of VPD on NEE.

Interestingly, all models display a positive dependence of NEE on air temperature for the peak growing season and setup with site variables, unlike avNNet and LinRegr on Hyytiälä data. Positive dependence is in line with theoretical expectations due to increasing respiration and reduced photosynthesis with increasing temperature within the peak season. At the same time, NEE somewhat decreases with increasing soil temperature A for all the models except RF; however, this effect of soil temperature on NEE, as captured by the models, is much weaker than that of air temperature.

On the whole year scale, all nonlinear models demonstrate rather similar NEE dependencies on different variables (except strong NEE dependencies on VPD and RH partially outweighing each other as modelled by avNNet), which was also the case

for the separate Värriö setup. This could be due the data from Värriö that has a long dormant season: one of the main tasks of the models is to reproduce seasonal cycle, for which the nonlinear models use soil temperature in a similar manner.

Generally, RF performed more in line with theoretical expectations from ecophysiological research than other models when trained on the data set containing interdependent variables. LinRegr and the avNNet demonstrate strong dependencies of NEE on VPD, which they likely compensate for by relatively strong dependencies of NEE on air temperature and RH. Due to that, some ALE may appear counterintuitive (e.g., strengthening of carbon sink with increasing air temperature during the peak season), contradicting the expectations based on general knowledge of ecosystem functioning. In addition, all models except RF demonstrate strong opposite associations with soil temperature A and B when both variables are available (Fig. 8).

4 Conclusions

We modeled NEE at two sites in boreal forest: one in central Finland and one in the Finnish subarctic. We focused on the peak growing season and whole-year data sets. Peak growing season NEE for separate sites can be modeled reasonably well even with a simple linear regression model. However, Linear Regression performs significantly worse than nonlinear ML models in the case of the mixed data sets from both sites or whole-year data sets.

The most powerful explaining variables in the peak growing season setups are PAR, diffuse PAR, and vapor pressure deficit (or air temperature); in the case of the whole-year setups, such variables are PAR, soil temperature and diffuse PAR. This is a robust result reproduced by most of the models used in this study. High vapor pressure deficit dampens photosynthesis and, hence, makes NEE increase. This effect is essential during the peak growing season. The models presumably used soil temperature to account for the change in NEE within a seasonal cycle.

To build a joint model for several sites, we added site variables. The model is more sensitive to these variables within the peak growing season, whereas soil temperature retains its importance for the whole-year data sets. In the absence of site-specific variables, Random Forest ranks soil water content, the variable that differs most between the sites, as the third most important in the feature importance diagram. NEE dependence on soil water content and the importance of this variable for NEE predictions change drastically for the models built on the data sets, including and excluding site variables.

Our ALE results suggest that Cubist and especially Random Forest display more robust behavior modeling complex non-linear dependence of net ecosystem exchange on the set of interconnected variables. They could qualitatively reproduce the theoretically expected dependencies of NEE on the major climatic drivers of ecosystem processes under different conditions and for several sites. This result aligns with many studies that used Random Forest based on its best performance compared with other models. Additionally, Linear Regression and Model Averaged Neural Networks tend to overemphasize certain variables while compensating with other interdependent variables. In our modeling study, Linear Regression and Model Averaged Neural Networks compensated using variables like air temperature and relative humidity, which are highly sensitive to changing climate conditions.

All in all, it should be noted that the models' performance changes depending on a given setup, so no single recommendation suggesting or prohibiting a specific model can be given. This is, instead, a case-by-case issue. Therefore, we call for broader

usage of Explainable Artificial Intelligence methods when applying ML methods, especially when choosing the most suitable model. Feature importance and ALE plots together allow for a direct comparison between ML model functioning and process-based models.

Finally, we showed that even a simple way to account for the difference between the sites decreases RMSE and improves the model. The next step is to introduce a more suitable variable, allowing us to distinguish the ecosystems from each other. As Hyytiälä data are split into pre-thinned and post-thinned, we need a variable that could account for this change in the vegetation. The best candidates for this could be satellite-based NDVI and LAI (Launiainen et al., 2022; Zhu et al., 2023), which we plan to add to our data set instead of site variables.

Code and data availability. The code and data are available at Zenodo repository [DOI: 10.5281/zenodo.1401314], https://doi.org/10.5281/zenodo.140131

Author contributions. EE, AL, KH and MK designed and conceptualized the study. TL performed modeling and prepared figures, wrote the manuscript (Introduction and Section 2). EE interpreted results and wrote the manuscript (Introduction, Section 3 and Conclusion). AL, PK, IM, KH and MK contributed with results interpretation, review and editing. All the authors commented on the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge the following projects: ACCC Flagship funded by the Academy of Finland grant number 337549 (UH) and 337552 (FMI), Academy professorship funded by the Academy of Finland (grant no. 302958), Academy of Finland projects no. 1325656, 311932, 334792, 316114, 325647, 325681, 347782, "Quantifying carbon sink, CarbonSink+ and their interaction with air quality" INAR project funded by Jane and Aatos Erkko Foundation, and HORIZON EUROPE (Project 101056921 — GreenFeedBack). University of Helsinki support via ACTRIS-HY is acknowledged. University of Helsinki Doctoral Programme in Atmospheric Sciences is acknowledged. Support of the technical and scientific staff in Hyytiälä is gratefully acknowledged.

References

Aalto, J., Anttila, V., Kolari, P., Korpela, I., Isotalo, A., Levula, J., Schiestl-Aalto, P., and Bäck, J.: Hyytiälä SMEAR II forest year 2020 thinning tree and carbon inventory data, https://doi.org/10.5281/zenodo.7639833, 2023.

Abbasian, H., Solgi, E., Hosseini, S. M., and Kia, S. H.: Modeling terrestrial net ecosystem exchange using machine learning techniques based on flux tower measurements, Ecological Modelling, 466, 109 901, 2022.

Artaxo, P., Hansson, H.-C., Andreae, M. O., Bäck, J., Alves, E. G., Barbosa, H. M. J., Bender, F., Bourtsoukidis, E., Carbone, S., Chi, J., Decesari, S., Després, V. R., Ditas, F., Ezhova, E., Fuzzi, S., Hasselquist, N. J., Heintzenberg, J., Holanda, B. A., Guenther, A., Hakola, H., Heikkinen, L., Kerminen, V.-M., Kontkanen, J., Krejci, R., Kulmala, M., Lavric, J. V., De Leeuw, G., Lehtipalo, K., Machado, L. A. T., McFiggans, G., Franco, M. A. M., Meller, B. B., Morais, F. G., Mohr, C., Morgan, W., Nilsson, M. B., Peichl, M., Petäjä, T., Praß, M., Pöhlker, C., Pöhlker, M. L., Pöschl, U., Von Randow, C., Riipinen, I., Rinne, J., Rizzo, L. V., Rosenfeld, D., Silva Dias, M. A. F., Sogacheva, L., Stier, P., Swietlicki, E., Sörgel, M., Tunved, P., Virkkula, A., Wang, J., Weber, B., Yáñez-Serrano, A. M., Zieger, P., Mikhailov, E., Smith, J. N., and Kesselmeier, J.: Tropical and boreal forest – atmosphere interactions: A review, Tellus B Chem. Phys. Meteorol., 74, 24, 2022.

Aubinet, M., Vesala, T., and Papale, D.: Eddy covariance: a practical guide to measurement and data analysis, Springer Science & Business Media, 2012.

Berrar, D.: Cross-Validation, in: Encyclopedia of Bioinformatics and Computational Biology, edited by Ranganathan, S., Gribskov, M., Nakai, K., and Schönbach, C., pp. 542–545, Academic Press, Oxford, https://doi.org/https://doi.org/10.1016/B978-0-12-809633-8.20349-X, 2019.

Breiman, L.: Random Forests, Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Cai, J., Xu, K., Zhu, Y., Hu, F., and Li, L.: Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest, Applied energy, 262, 114 566, 2020.

Dou, X. and Yang, Y.: Estimating forest carbon fluxes using four different data-driven techniques based on long-term eddy covariance measurements: Model comparison and evaluation, Science of the Total Environment, 627, 78–94, 2018.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al.: Explainable AI (XAI): Core ideas, techniques, and solutions, ACM Computing Surveys, 55, 1–33, 2023.

Ezhova, E., Ylivinkka, I., Kuusk, J., Komsaare, K., Vana, M., Krasnova, A., Noe, S., Arshinov, M., Belan, B., Park, S.-B., Lavric, J. V., Heimann, M., Petaja, T., Vesala, T., Mammarella, I., Kolari, P., Bäck, J., Rannik, U., Kerminen, V.-M., and Kulmala, M.: Direct effect of aerosols on solar radiation and gross primary production in boreal and hemiboreal forests, Atmospheric Chemistry and Physics, 18, 17 863–17 881, 2018.

Fernández-Delgado, M., Sirsat, M. S., Cernadas, E., Alawadi, S., Barro, S., and Febrero-Bande, M.: An extensive experimental survey of regression methods, Neural Networks, 111, 11–34, 2019.

Grossiord, C., Buckley, T. N., Cernusak, L. A., Novick, K. A., Poulter, B., Siegwolf, R. T. W., Sperry, J. S., and McDowell, N. G.: Plant responses to rising vapor pressure deficit, New Phytologist, 226, 1550–1566, https://doi.org/https://doi.org/10.1111/nph.16485, 2020.

Gu, L., Baldocchi, D., Verma, S. B., Black, T., Vesala, T., Falge, E. M., and Dowty, P. R.: Advantages of diffuse radiation for terrestrial ecosystem productivity, Journal of Geophysical Research: Atmospheres, 107, ACL–2, 2002.

Hancock, J. T. and Khoshgoftaar, T. M.: Survey on categorical data for neural networks, J. Big Data, 7, 28, https://doi.org/10.1186/S40537-020-00305-W, 2020.

Hari, P. and Kulmala, M.: Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II), Boreal Environment Research, 10, 315–322, 2005.

Hari, P., Kulmala, M., Pohja, T., Lahti, T., Siivola, E., Palva, L., Aalto, P., Hämeri, K., Vesala, T., Luoma, S., and Pulliainen, E.: Air pollution in eastern Lapland : challenge for an environmental measurement station, Silva Fennica 1994. 28(1): 29–39., 28, https://doi.org/10.14214/sf.a9160, 1994.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction, vol. 2, Springer, 2009.

Jung, Y.: Multiple predicting K-fold cross-validation for model selection, Journal of Nonparametric Statistics, 30, 197–215, 2018.

Junttila, V., Minunno, F., Peltoniemi, M., Forsius, M., Akujärvi, A., Ojanen, P., and Mäkelä, A.: Quantification of forest carbon flux and stock uncertainties under climate change and their use in regionally explicit decision making: Case study in Finland, Ambio, https://doi.org/10.1007/s13280-023-01906-4, 2023.

Kämäräinen, M., Lintunen, A., Kulmala, M., Tuovinen, J.-P., Mammarella, I., Aalto[1], J., Vekuri, H., and Lohila, A.: Evaluation of gradient boosting and random forest methods to model subdaily variability of the atmosphere–forest CO2 exchange, Biogeosciences Discussions, 2022, 1–24, 2023.

Kolari, P., Kulmala, L., Pumpanen, J., Launiainen, S., Ilvesniemi, H., Hari, P., and Nikinmaa, E.: CO_2 exchange and component CO_2 fluxes of a boreal Scots pine forest, Boreal Environment Research, 14, 761–783, 2009.

Kuhn, M.: Building predictive models in R using the caret package, Journal of Statistical Software, 28, 1–26, 2008.

Kuhn, M.: caret: Classification and Regression Training, https://cran.r-project.org/web/packages/caret/index.html, r package version 6.0-90, 2023.

Kuhn, M. and Johnson, K.: Applied Predictive Modeling, Springer, 2013.

Kulmala, M., Ezhova, E., Kalliokoski, T., Noe, S., Vesala, T., Lohila, A., Liski, J., Makkonen, R., Bäck, J., Petäjä, T., and Kerminen, V.-M.: CarbonSink+: Accounting for multiple climate feedbacks from forests, Boreal Environment Research, 25, 145–159, 2020.

Kulmala, M., Cai, R., Ezhova, E., Deng, C., Stolzenburg, D., Dada, L., Guo, Y., Yan, C., Peräkylä, O., Lintunen, A., Nieminen, T., Kokkonen, T., Sarnela, N., Petäjä, T., and Kerminen, V.-M.: Direct link between the characteristics of atmospheric new particle formation and Continental Biosphere-Atmosphere-Cloud-Climate (COBACC) feedback loop, Boreal Environment Research, 28, 1, 2023.

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W.: Applied Linear Statistical Models, McGraw-Hill, New York, 5th edn., 2004.

Larcher, W.: Physiological plant ecology: ecophysiology and stress physiology of functional groups, Springer Science & Business Media, 2003.

Lasslop, G., Migliavacca, M., Bohrer, G., Reichstein, M., Bahn, M., Ibrom, A., Jacobs, C., Kolari, P., Papale, D., Vesala, T., et al.: On the choice of the driving temperature for eddy-covariance carbon dioxide flux partitioning, Biogeosciences, 9, 5243–5259, 2012.

Launiainen, S., Katul, G. G., Leppä, K., Kolari, P., Aslan, T., Grönholm, T., Korhonen, L., Mammarella, I., and Vesala, T.: Does growing atmospheric CO2 explain increasing carbon sink in a boreal coniferous forest?, Global Change Biology, 28, 2910–2929, https://doi.org/https://doi.org/10.1111/gcb.16117, 2022.

Liaw, A. and Wiener, M.: Classification and Regression by randomForest, R News, 2, 18–22, https://CRAN.R-project.org/doc/Rnews/, 2002.

Lintunen, A., Paljakka, T., Salmon, Y., Dewar, R., Riikonen, A., and Hölttä, T.: The influence of soil temperature and water content on belowground hydraulic conductance and leaf gas exchange in mature trees of three boreal species, Plant, Cell & Environment, 43, 532–547, 2020.

Liu, J., Zuo, Y., Wang, N., Yuan, F., Zhu, X., Zhang, L., Zhang, J., Sun, Y., Guo, Z., Guo, Y., et al.: Comparative analysis of two machine learning algorithms in predicting site-level net ecosystem exchange in major biomes, Remote Sensing, 13, 2242, 2021.

Mäkelä, J., Knauer, J., Aurela, M., Black, A., Heimann, M., Kobayashi, H., Lohila, A., Mammarella, I., Margolis, H., Markkanen, T., Susiluoto, J., Thum, T., Viskari, T., Zaehle, Z., and Aalto, T.: Parameter calibration and stomatal conductance formulation comparison for boreal forests with adaptive population importance sampler in the land surface model JSBACH, Geoscientific Model Development, 12, 4075–4098, 2019.

Mammarella, I., Peltola, O., Nordbo, A., Järvi, L., and Rannik, Ü.: Quantifying the uncertainty of eddy covariance fluxes due to the use of different software packages and combinations of processing steps in two contrasting ecosystems, Atmospheric Measurement Techniques, 9, 4915–4933, 2016.

Moffat, A. M., Beckstein, C., Churkina, G., Mund, M., and Heimann, M.: Characterization of ecosystem responses to climatic controls using artificial neural networks, Global Change Biology, 16, 2737–2749, 2010.

Molnar, C.: Interpretable machine learning, Lulu. com, 2020.

Molnar, C., Casalicchio, G., and Bischl, B.: iml: An R package for interpretable machine learning, Journal of Open Source Software, 3, 786, 2018.

Monteith, J. and Unsworth, M.: Principles of environmental physics: plants, animals, and the atmosphere, Academic Press, 2013.

Mäkelä, A., Hari, P., Berninger, F., Hänninen, H., and Nikinmaa, E.: Acclimation of photosynthetic capacity in Scots pine to the annual cycle of temperature, Tree Physiology, 24, 369–376, 2004.

Mäkelä, A., Kolari, P., Karimäki, J., Nikinmaa, E., Perämäki, M., and Hari, P.: Modelling five years of weather-driven variation of GPP in a boreal forest, Agricultural and Forest Meteorology, 139, 382–398, 2006.

Neimane-Šroma, S., Durand, M., Lintunen, A., Aalto, J., and Robson, T. M.: Shedding light on the increased carbon uptake by a boreal forest under diffuse solar radiation across multiple scales, Global Change Biology, 30, e17 275, https://doi.org/https://doi.org/10.1111/gcb.17275, e17275 GCB-23-3034.R2, 2024.

Palmroth, S. and Hari, P.: Evaluation of the importance of acclimation of needle structure, photosynthesis, and respiration to available photosynthetically active radiation in a Scots pine canopy, Canadian Journal of Forest Research, 31, 1235–1243, 2001.

Peltoniemi, M., Pulkkinen, M., Aurela, M., Pumppanen, J., Kolari, P., and Mäkelä, A.: A semi-empirical model of boreal-forest gross primary production, evapotranspiration, and soil water — calibration and sensitivity analysis, Boreal Environment Research, 20, 151–171, 2015.

Petäjä, T., Tabakova, K., Manninen, A., Ezhova, E., O'Connor, E., Moisseev, D., Sinclair, V. A., Backman, J., Levula, J., Luoma, K., Virkkula, A., Paramonov, M., Räty, M., Äijälä, M., Heikkinen, L., Ehn, M., Sipilä, M., Yli-Juuti, T., Virtanen, A., Ritsche, M., Hickmon, N., Pulik, G., Rosenfeld, D., Worsnop, D., Bäck, J., Kulmala, M., and Kerminen, V.-M.: Influence of biogenic emissions from boreal forests on aerosol–cloud interactions, Nature Geoscience, 15, 42–47, 2022.

Quinlan, J. R.: Cubist: Rule- and Instance-Based Regression Modeling, https://CRAN.R-project.org/package=cubist, R package version 0.4.2.1, 1992.

Reitz, O., Graf, A., Schmidt, M., Ketzler, G., and Leuchner, M.: Upscaling net ecosystem exchange over heterogeneous landscapes with machine learning, Journal of Geophysical Research: Biogeosciences, 126, e2020JG005 814, 2021.

Ripley, B. D.: Pattern recognition and neural networks, Cambridge university press, 2007.

Ross, J. and Sulev, M.: Sources of errors in measurements of PAR, Agricultural and Forest Meteorology, 100, 103–125, 2000.

Running, S. W.: Environmental control of leaf water conductance in conifers, Canadian Journal of Forest Research, 6, 104–112, 1976.

Shirley, I. A., Mekonnen, Z. A., Grant, R. F., Dafflon, B., and Riley, W. J.: Machine learning models inaccurately predict current and future high-latitude C balances, Environmental Research Letters, 18, 014 026, 2023.

Tagesson, T., Schurgers, G., Horion, S., Ciais, P., Tian, F., Brandt, M., Ahlström, A., Wigneron, J.-P., Ardö, J., Olin, S., Fan, L., Wu, Z., and Fensholt, R.: Recent divergence in the contributions of tropical and boreal forests to the terrestrial carbon sink, Nature Ecology & Evolution, 4, 202–209, 2020.

Tang, J., Zhou, P., Miller, P. A., Schurgers, G., Gustafson, A., Makkonen, R., Fu, Y. H., and Rinnan, R.: High-latitude vegetation changes will determine future plant volatile impacts on atmospheric organic aerosols, npj Climate and Atmospheric Science, 6, 147, 2023.

Wood, D. A.: Net ecosystem carbon exchange prediction and insightful data mining with an optimized data-matching algorithm, Ecological Indicators, 124, 107 426, 2021.

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, Agricultural and Forest Meteorology, 156, 85–103, 2012.

Zeng, J., Matsunaga, T., Tan, Z.-H., Saigusa, N., Shirai, T., Tang, Y., Peng, S., and Fukuda, Y.: Global terrestrial carbon fluxes of 1999–2019 estimated by upscaling eddy covariance data with a random forest, Scientific data, 7, 313, 2020.

Zhang, A., Lipton, Z. C., Li, M., and Smola, A. J.: Dive into deep learning, Cambridge University Press, 2023.

Zhou, J., Li, E., Wei, H., Li, C., Qiao, Q., and Armaghani, D. J.: Random forests and cubist algorithms for predicting shear strengths of rockfill materials, Applied sciences, 9, 1621, 2019.

Zhu, X.-J., Yu, G.-R., Chen, Z., Zhang, W.-K., Han, L., Wang, K.-F., Chen, S.-P., Liu, S.-M., Wang, H.-M., Yan, J.-H., Tan, J.-L., Zhang, F.-W., Zhao, F.-H., Li, Y.-N., Zhang, Y.-P., Shi, P.-L., Zhu, J.-J., Wu, J.-B., Zhao, Z.-H., Hao, Y.-B., Sha, L.-Q., Zhang, Y.-C., Jiang, S.-C., Gu, F.-X., Wu, Z.-X., Zhang, Y.-J., Zhou, L., Tang, Y.-K., Jia, B.-R., Li, Y.-K., Song, Q.-H., Dong, G., Gao, Y.-H., Jiang, Z.-D., Sun, D., Wang, J.-L., He, Q.-H., Li, X.-H., Wang, F., Wei, W.-X., Deng, Z.-M., Hao, X.-X., Li, Y., Liu, X.-L., Zhang, X.-F., and Zhu, Z.-L.: Mapping Chinese annual gross primary productivity with eddy covariance measurements and machine learning, Science of The Total Environment, 857, 159 390, https://doi.org/https://doi.org/10.1016/j.scitotenv.2022.159390, 2023.

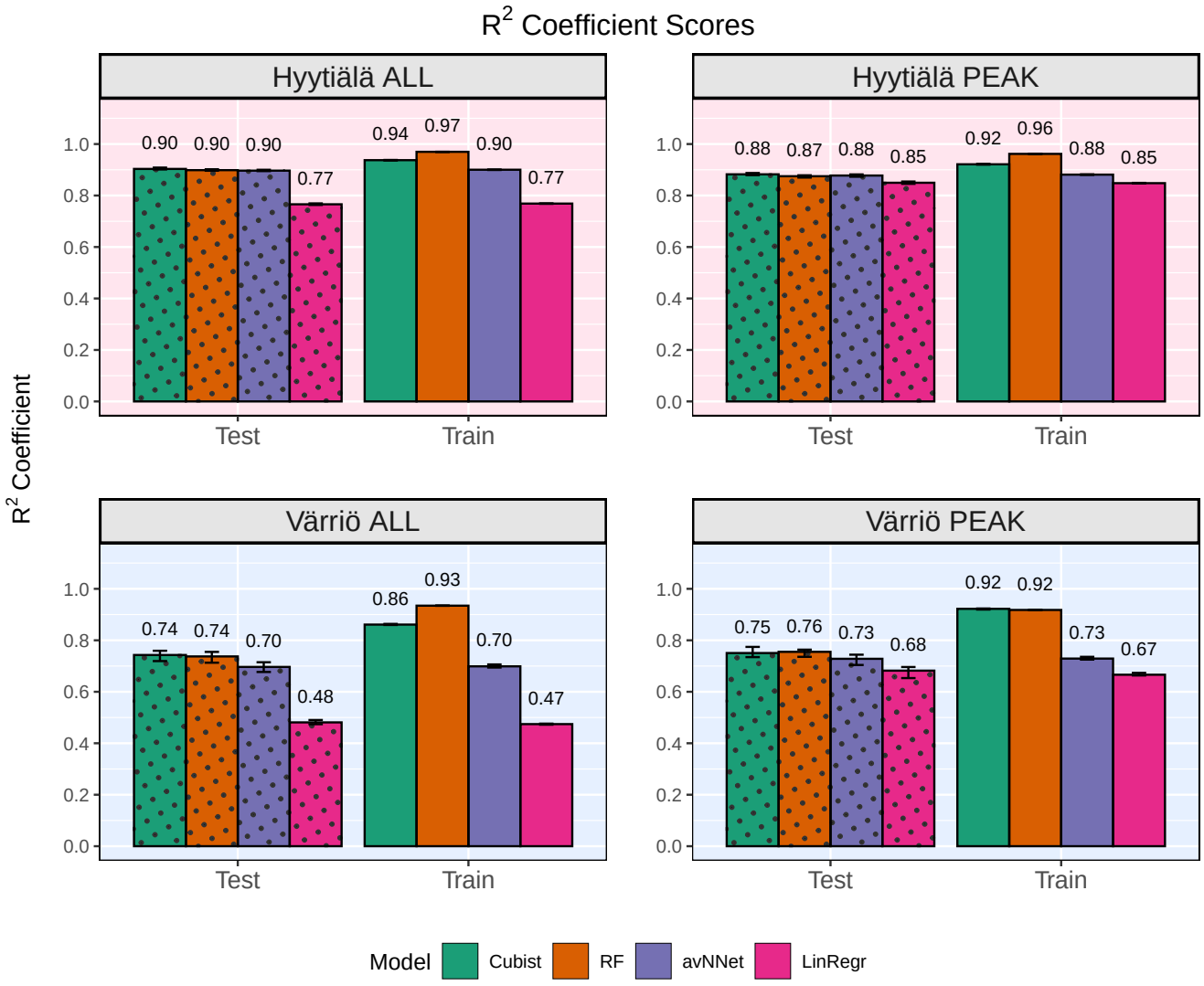

Figure 1. R^2-coefficients for all the models and different setups from Set 1 (Table 3). In each of the four panels, the results for the training data set are shown on the right (marked 'Train'), and the results for the test data set are shown on the left (dotted bars, marked 'Test'). Different colors are used to distinguish between the ML models, see legend.'ALL' denotes the scores for the models trained on the whole year data sets; 'PEAK' - for the models trained on the peak growing season data sets. The black error bars show the min and max, and the bars show the mean of the scores trained on different splits of the data.

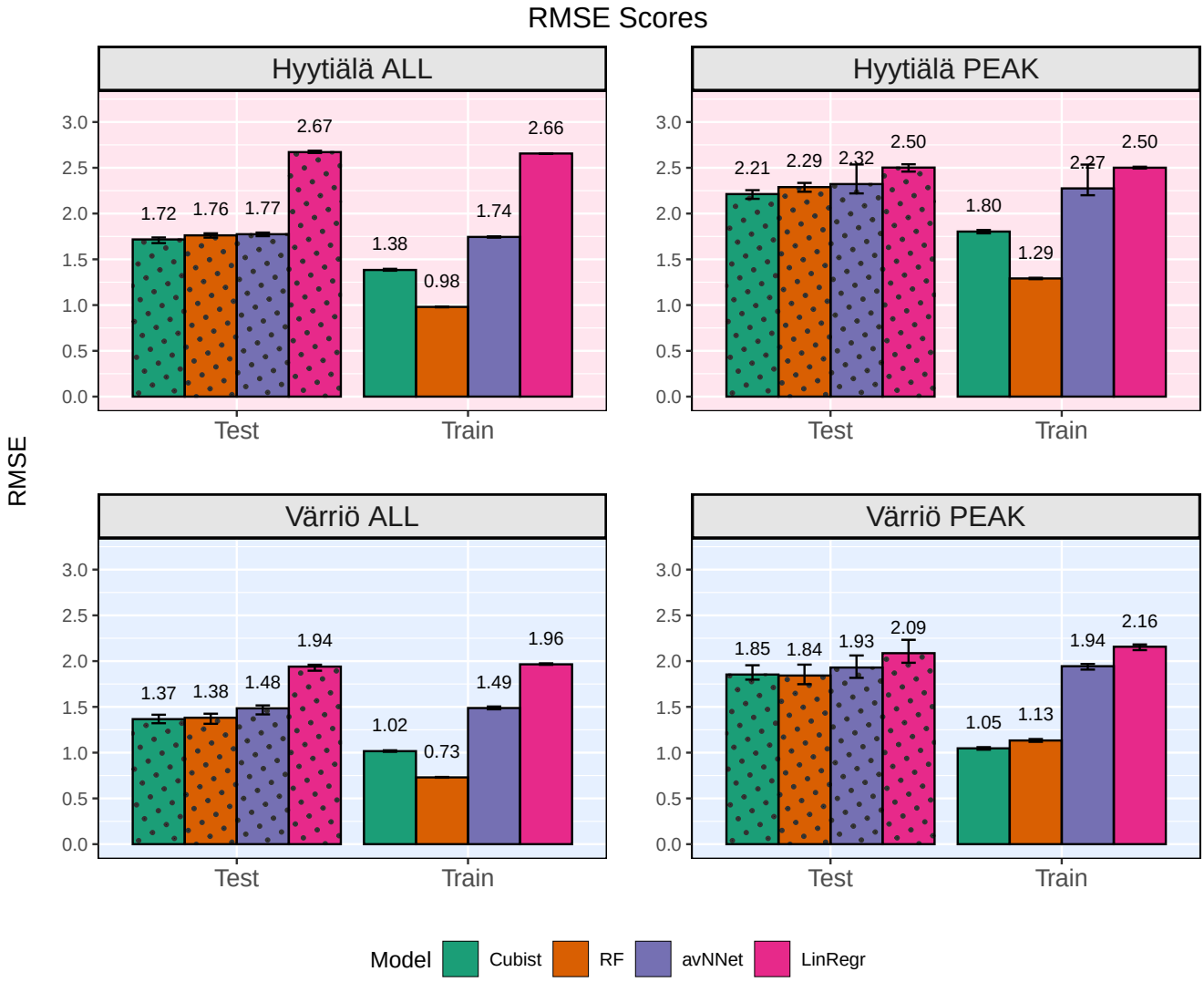

Figure 2. RMSE for all models and different setups from Set 1 (Table 3). In each of the four panels, the results for the training data set are shown on the right (marked 'Train'), and the results for the test data set are shown on the left (dotted bars, marked 'Test'). Different colors are used to distinguish between the ML models, see legend. 'ALL' denotes the scores for the models trained on the whole year data sets; 'PEAK' - for the models trained on the peak growing season data sets. The black error bars show the min and max, and the bars show the mean of the scores trained on different splits of the data.

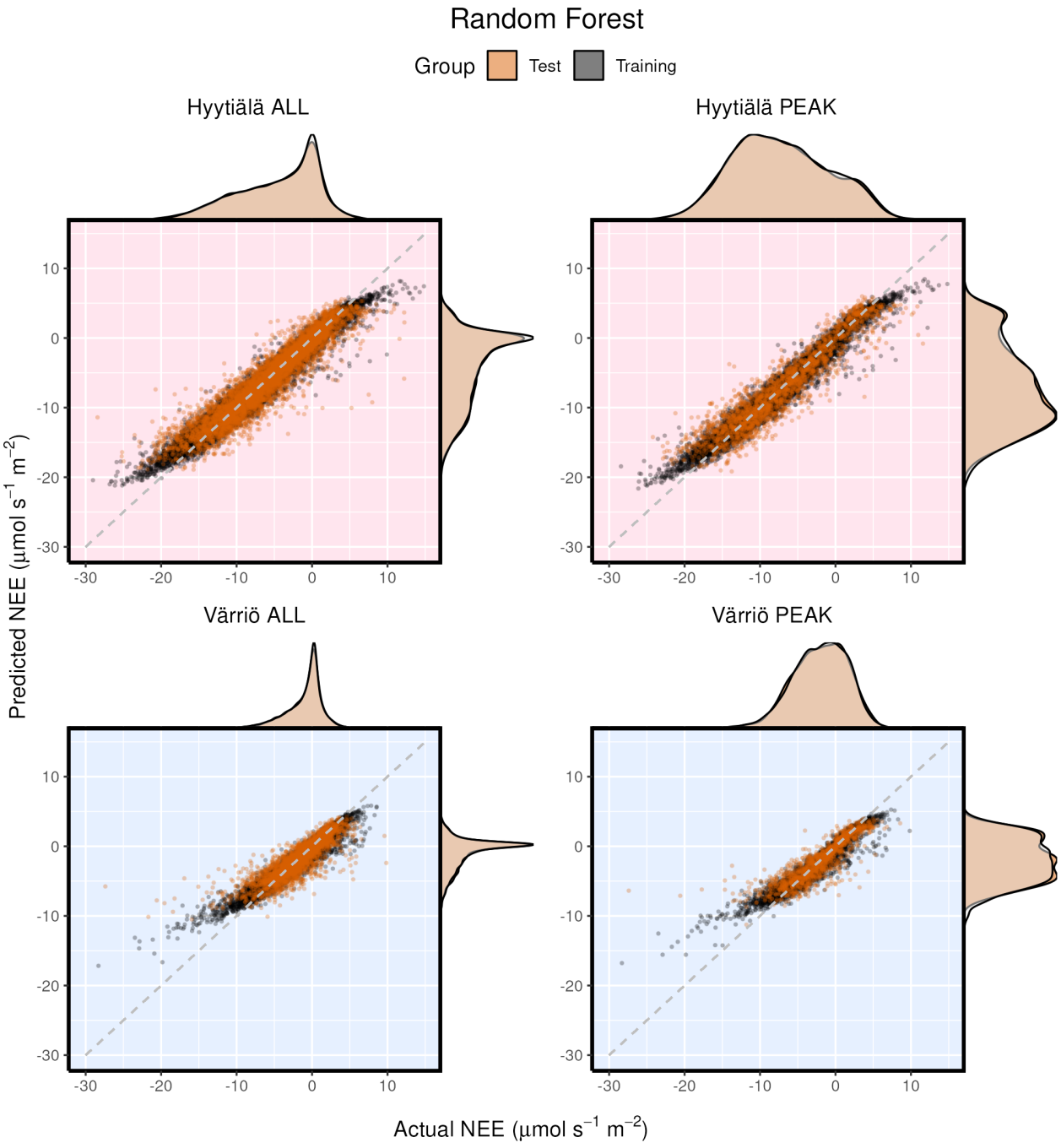

Figure 3. Modeled vs measured NEE for Hyytiälä and Värriö on the example of Random Forest model. Black points indicate training data sets, orange - test data sets. 'ALL' denotes the plots based on the whole year data sets; 'PEAK' - on the peak growing season data sets. The density distributions of the actual NEE and predicted NEE are shown on top and right side of the plots, respectively, with colored being the test, and translucent being the training data.

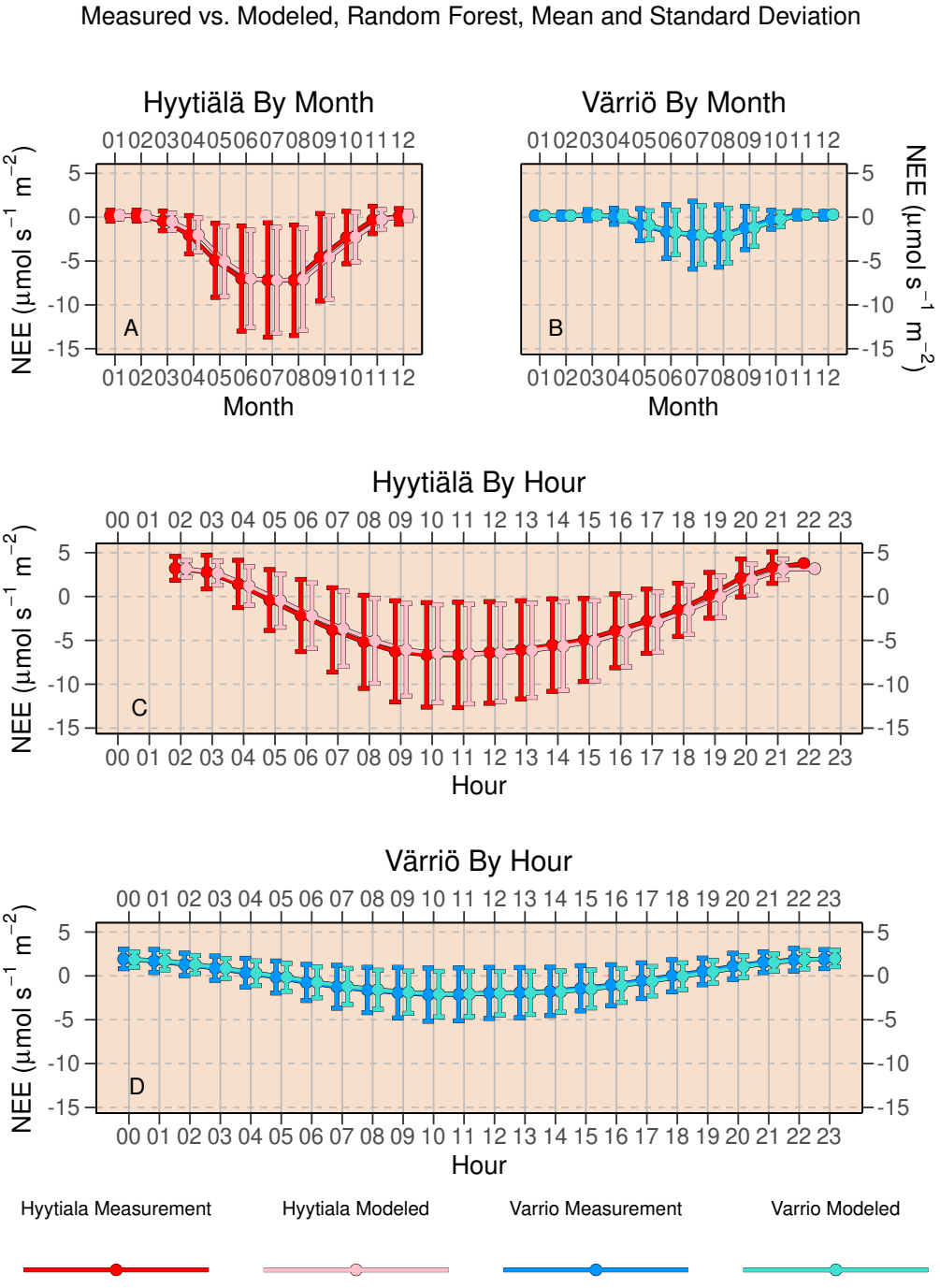

Figure 4. Mean diurnal and monthly cycles of daytime NEE as reproduced by Random Forest compared to actual NEE. Error bars denote standard deviation.

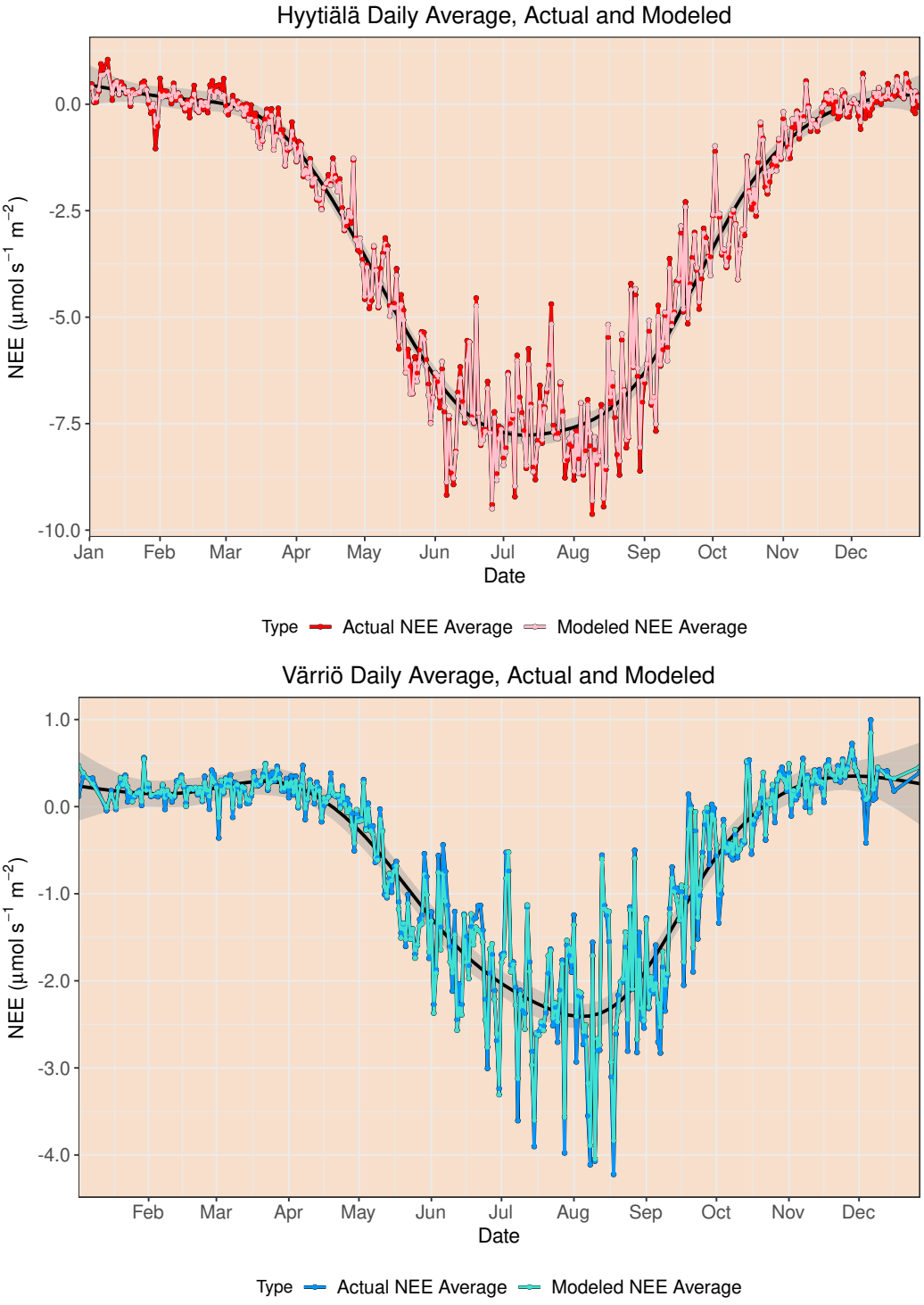

Figure 5. Mean annual cycle of daily NEE as reproduced by Random Forest compared to actual NEE.

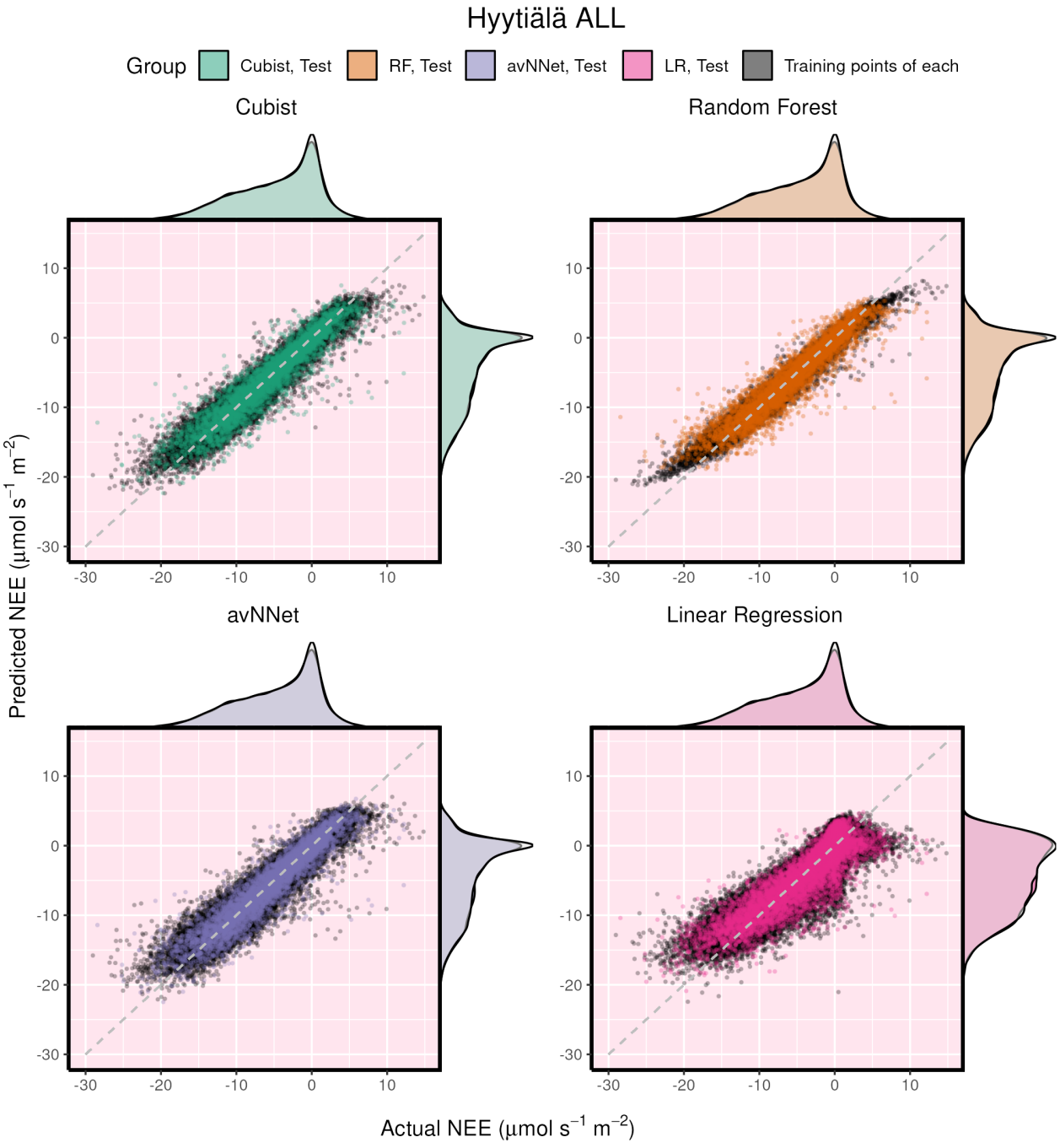

Figure 6. Modeled vs measured NEE illustrating performance of all the models on the whole year Hyytiälä data set. Black points indicate training data sets, orange - test data sets. The density distributions of the actual NEE and predicted NEE are shown on top and right side of the plots, respectively, with colored being the test, and translucent being the training data.

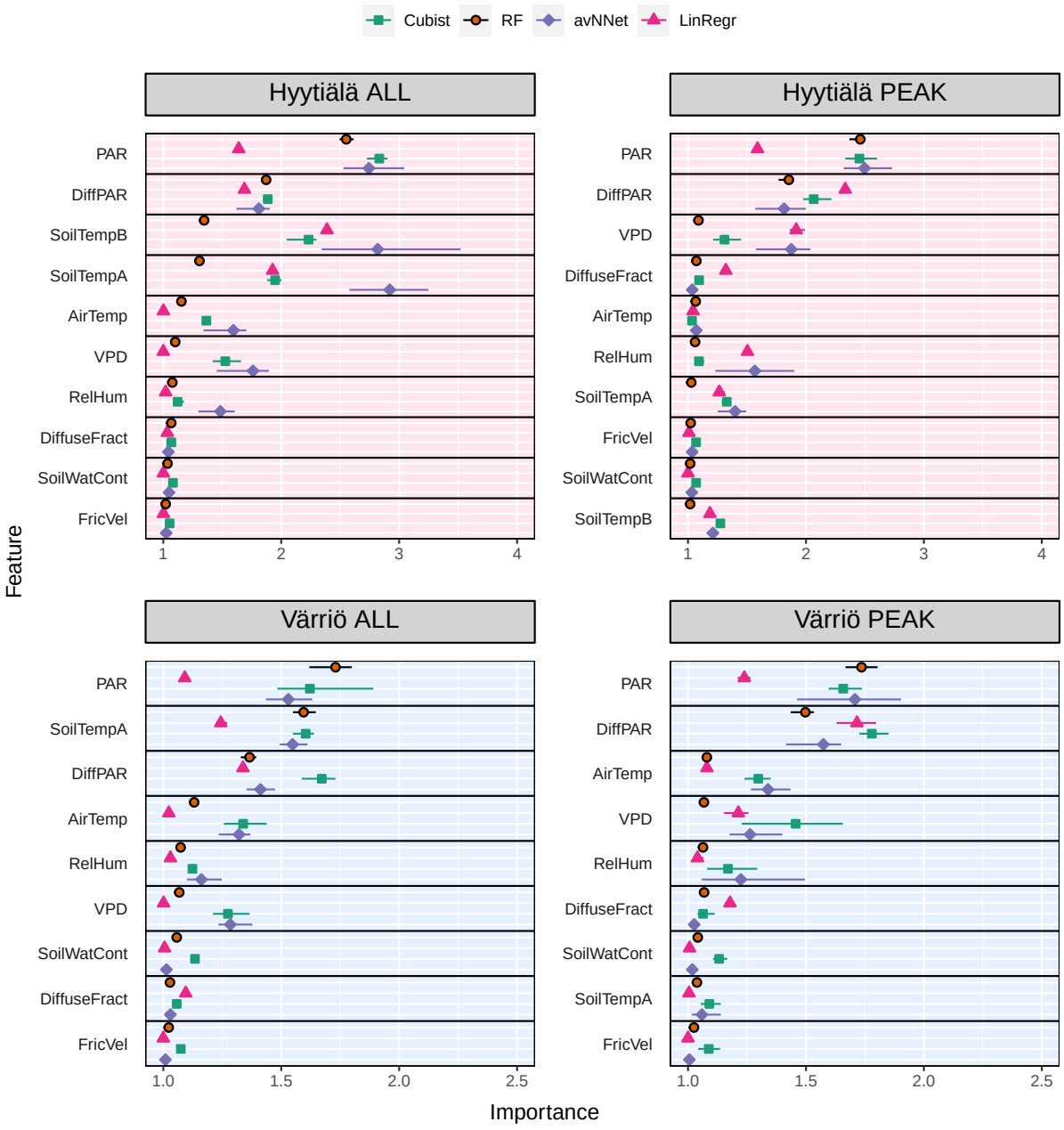

Figure 7. Feature importance for all the models and different setups from Set 1 (Table 3). The order of features is in accordance with the outcome of the Random Forest model. 'ALL' denotes the plots based on the whole year data sets; 'PEAK' - on the peak growing season data sets. The points indicate the mean of the FI score on across multiple datasets, while the bars show the min and max, respectively.

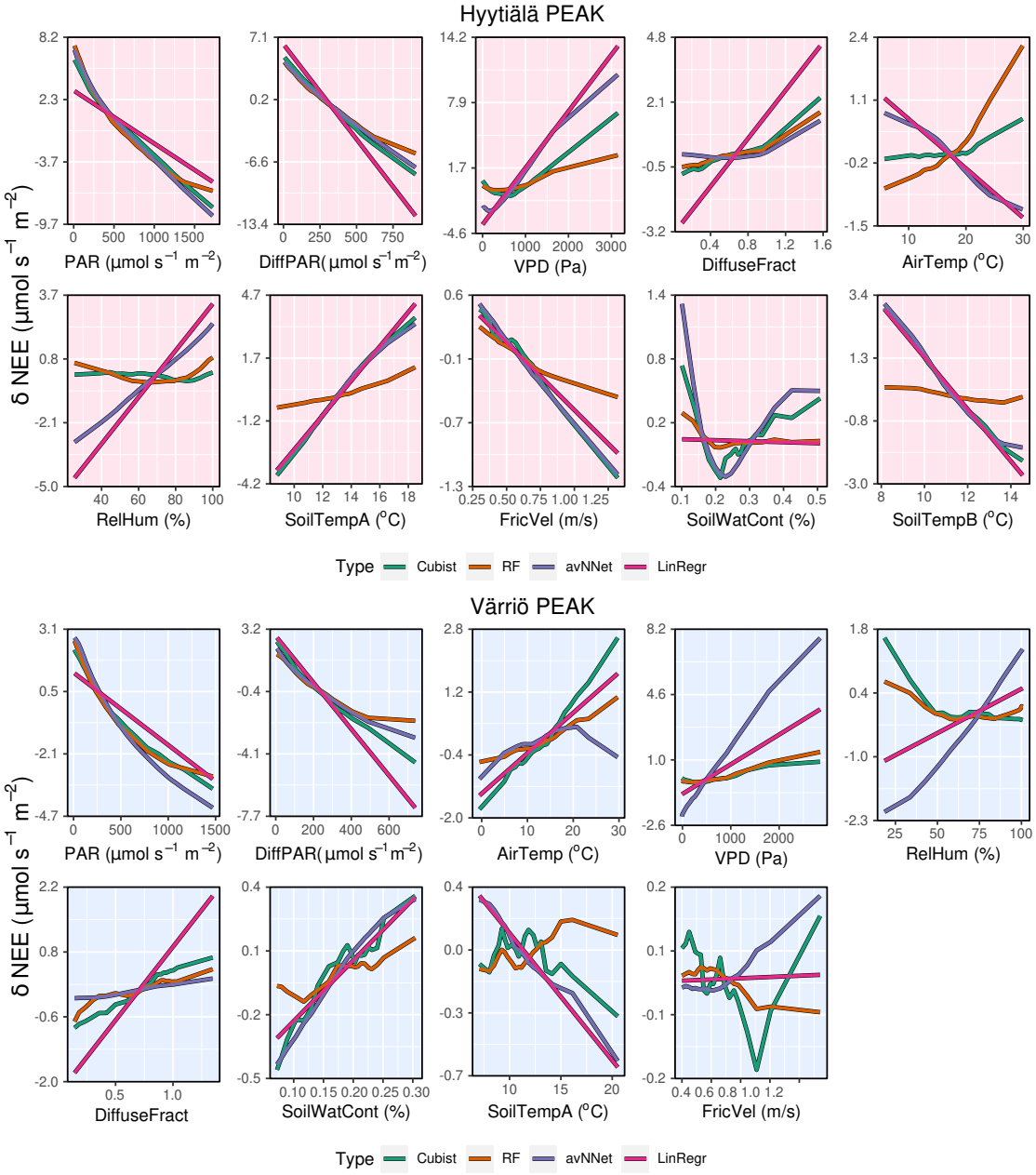

Figure 8. ALE plots for all the models (see legend), data sets correspond to the peak growing season in Hyytiälä (upper panels) and Värriö (lower panels).

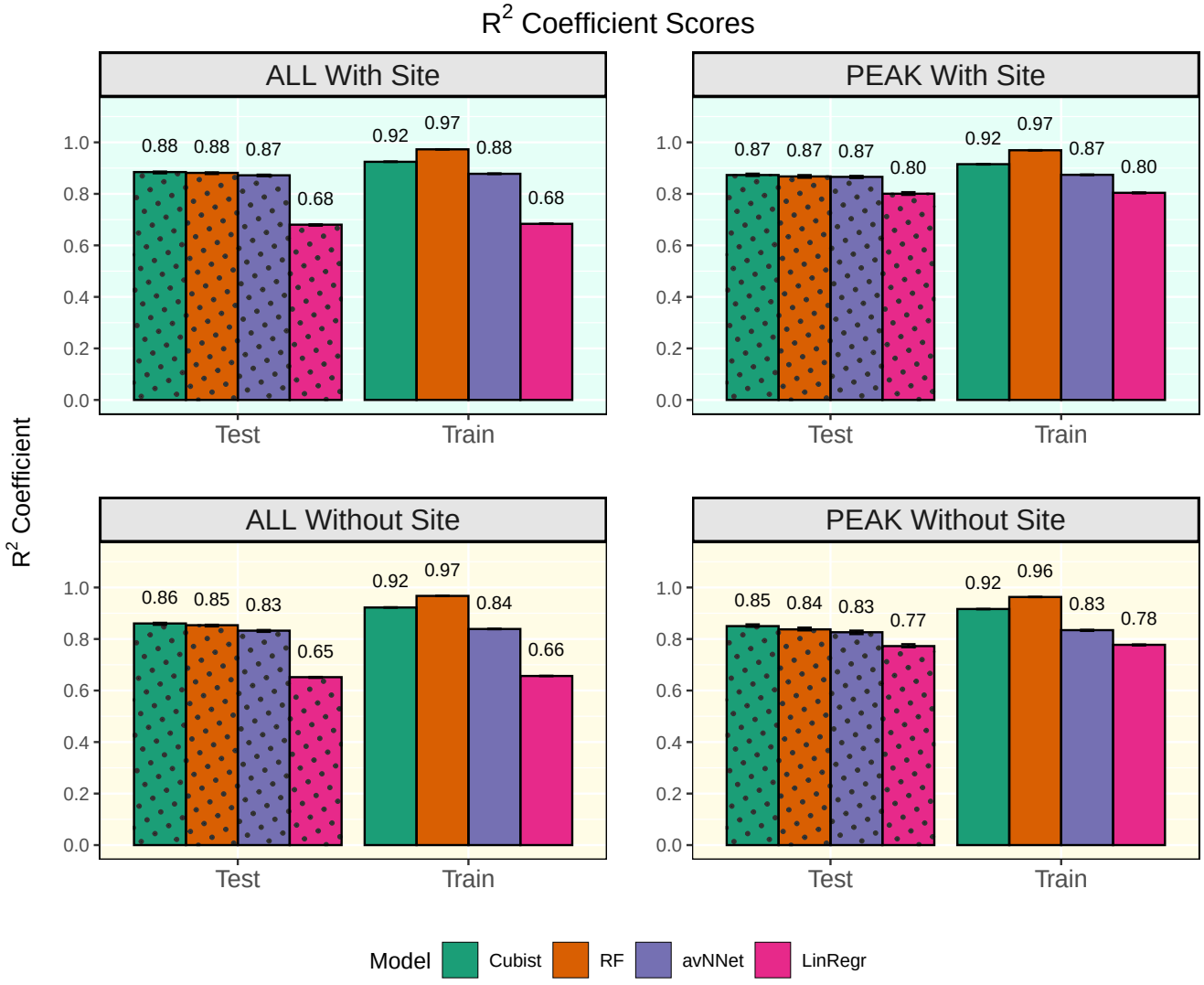

Figure 9. R^2-coefficients for all the models and different setups from Set 2 (Table 3). In each of the four panels, the results for the training data set are shown on the right (marked 'Train'), and the results for the test data set are shown on the left (dotted bars, marked 'Test'). 'ALL' denotes the scores for the models using the whole year data sets; 'PEAK' - for the models using the peak growing season data sets. 'With Site' - the input variables contain the information about site, 'Without Site' - no information about site. The black error bars show the min and max, and the bars show the mean of the scores trained on different splits of the data.

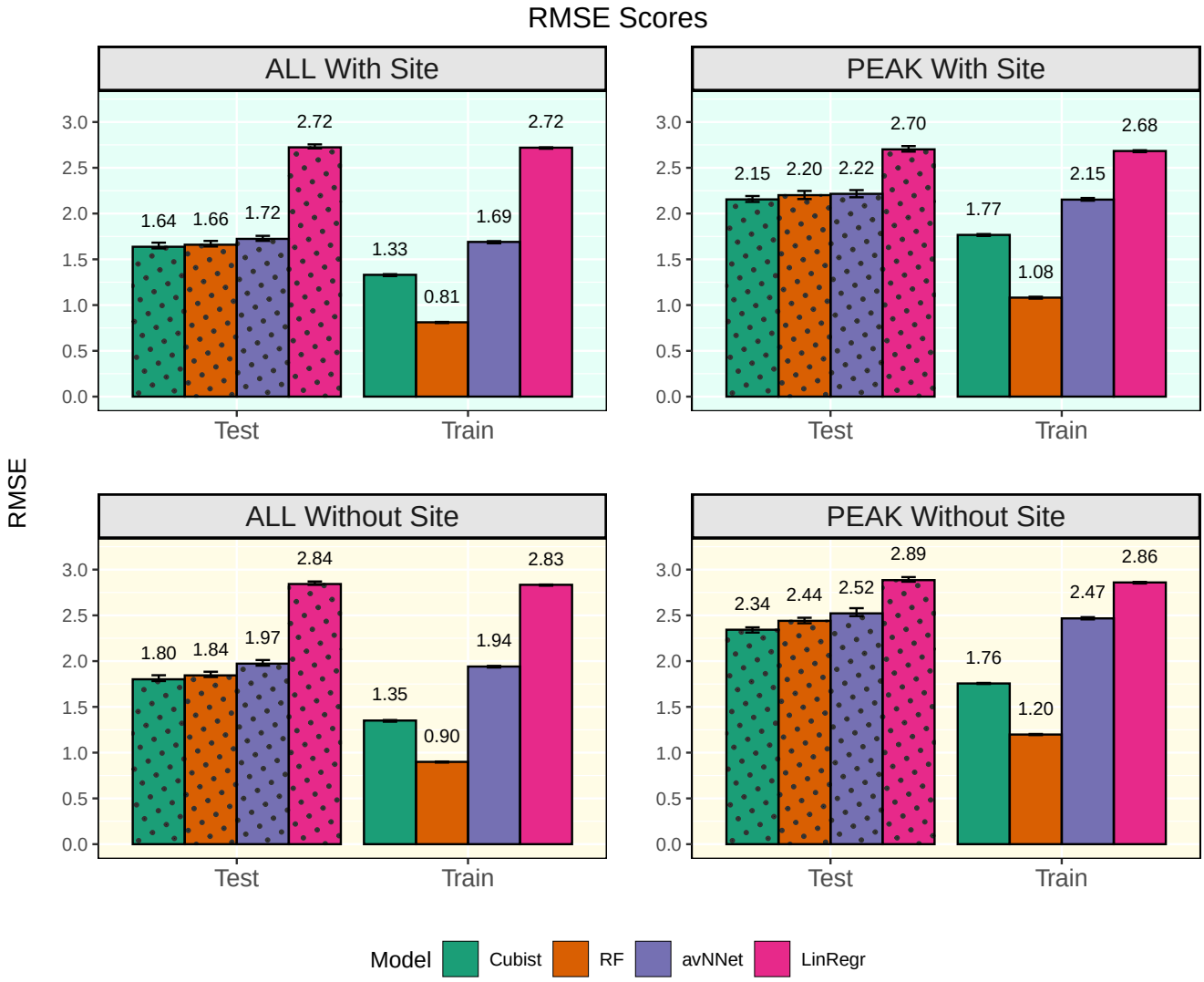

Figure 10. RMSE for all the models and different setups from Set 2 (Table 3). In each of the four panels, the results for the training data set are shown on the right (marked 'Train'), and the results for the test data set are shown on the left (dotted bars, marked 'Test'). 'ALL' denotes the scores for the models using the whole year data sets; 'PEAK' - for the models using the peak growing season data sets. 'With Site' - the input variables contain the information about site, 'Without Site' - no information about site. The black error bars show the min and max, and the bars show the mean of the scores trained on different splits of the data.

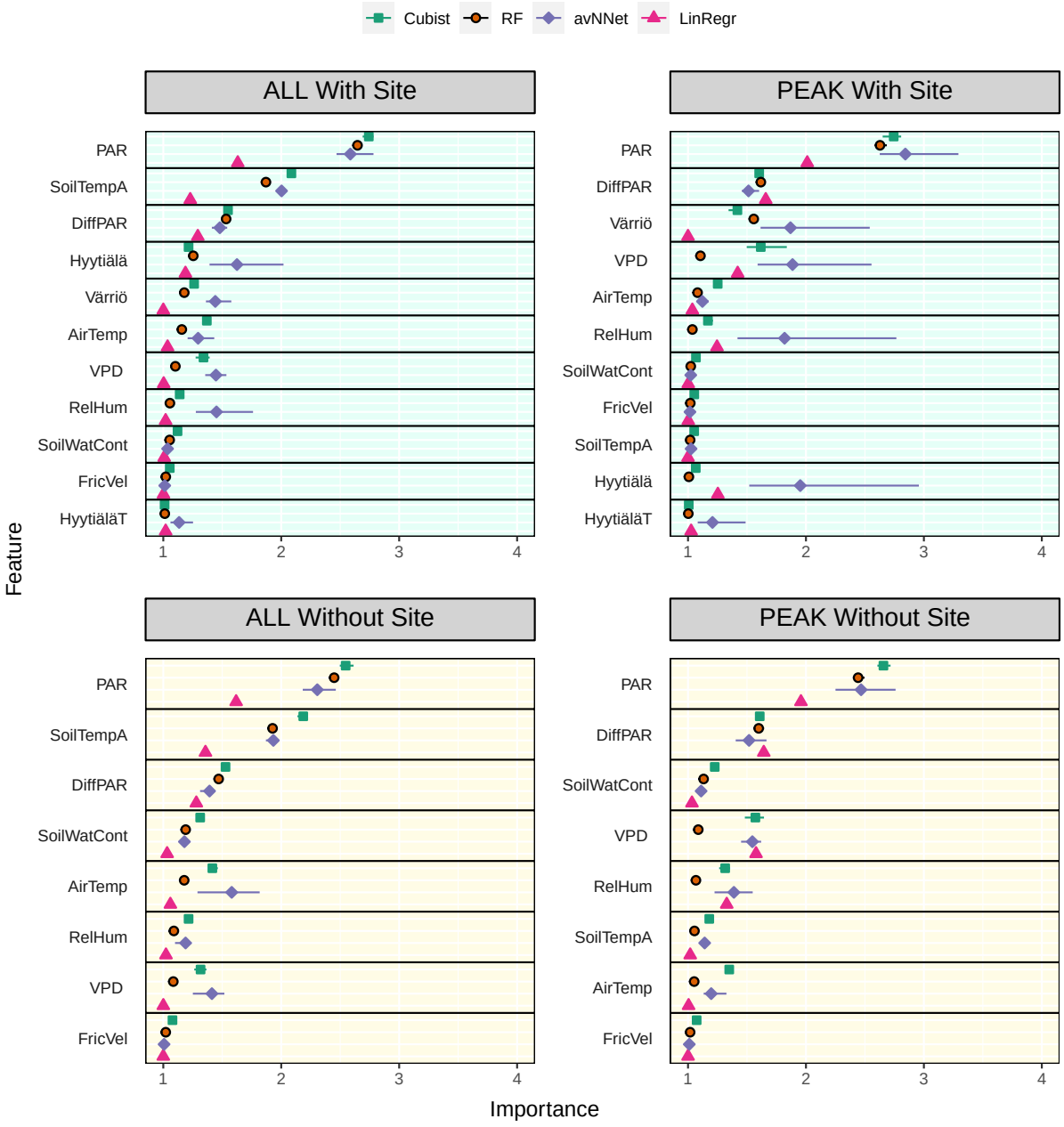

Figure 11. Feature importance for all the models trained on the mixed data sets containing ('With Site') and not containing ('Without Site') site variables. The order of features is in accordance with the outcome of the Random Forest model. 'ALL' denotes the plots based on the whole year data sets; 'PEAK' - on the peak growing season data sets. The points indicate the mean of the FI score on across multiple datasets, while the bars show the min and max, respectively.

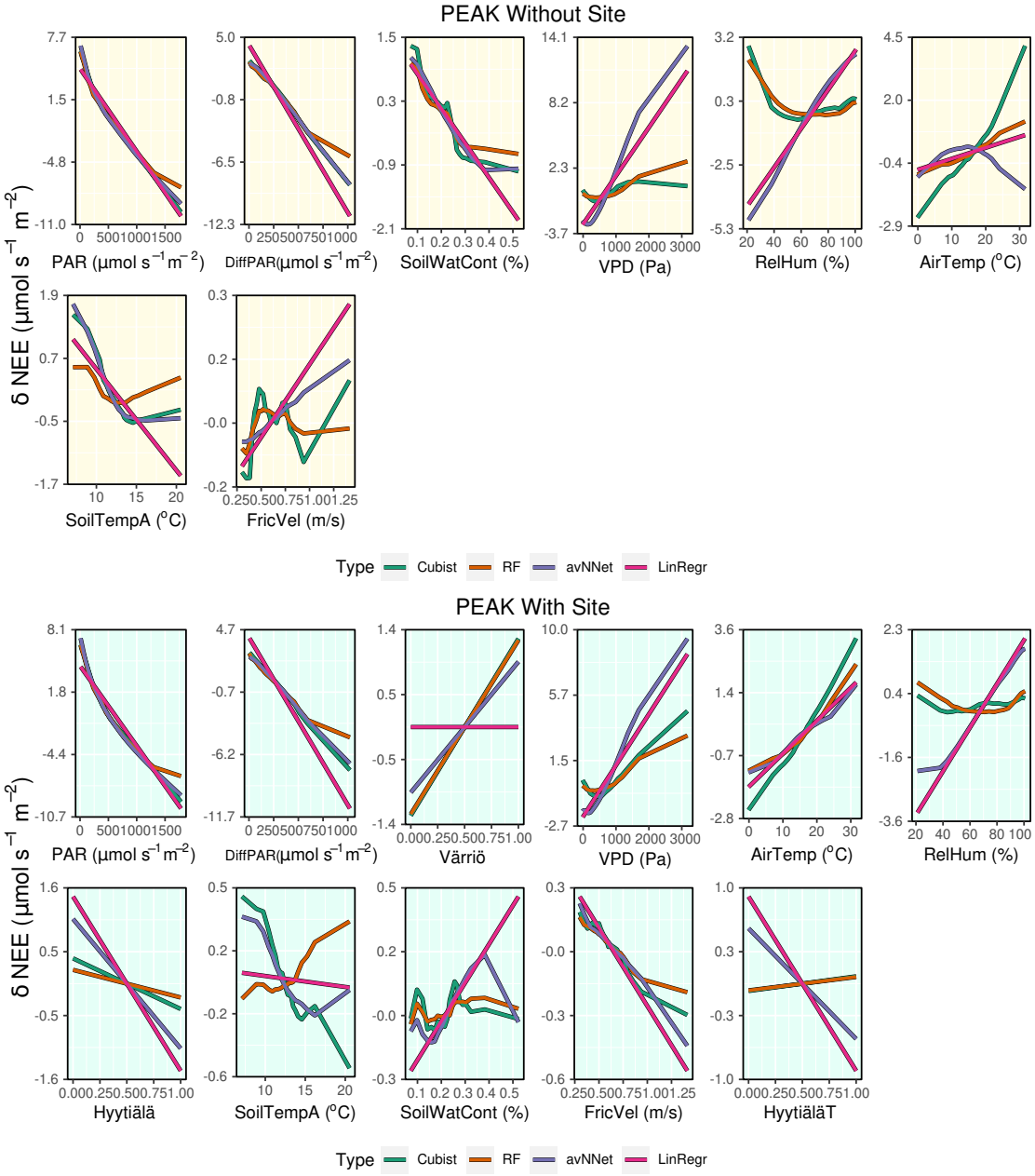

Figure 12. ALE plots for all the models trained on the mixed data sets containing ('With Site', lower panels) and not containing ('Without Site', upper panels) site variables. The data sets are from the peak growing season.

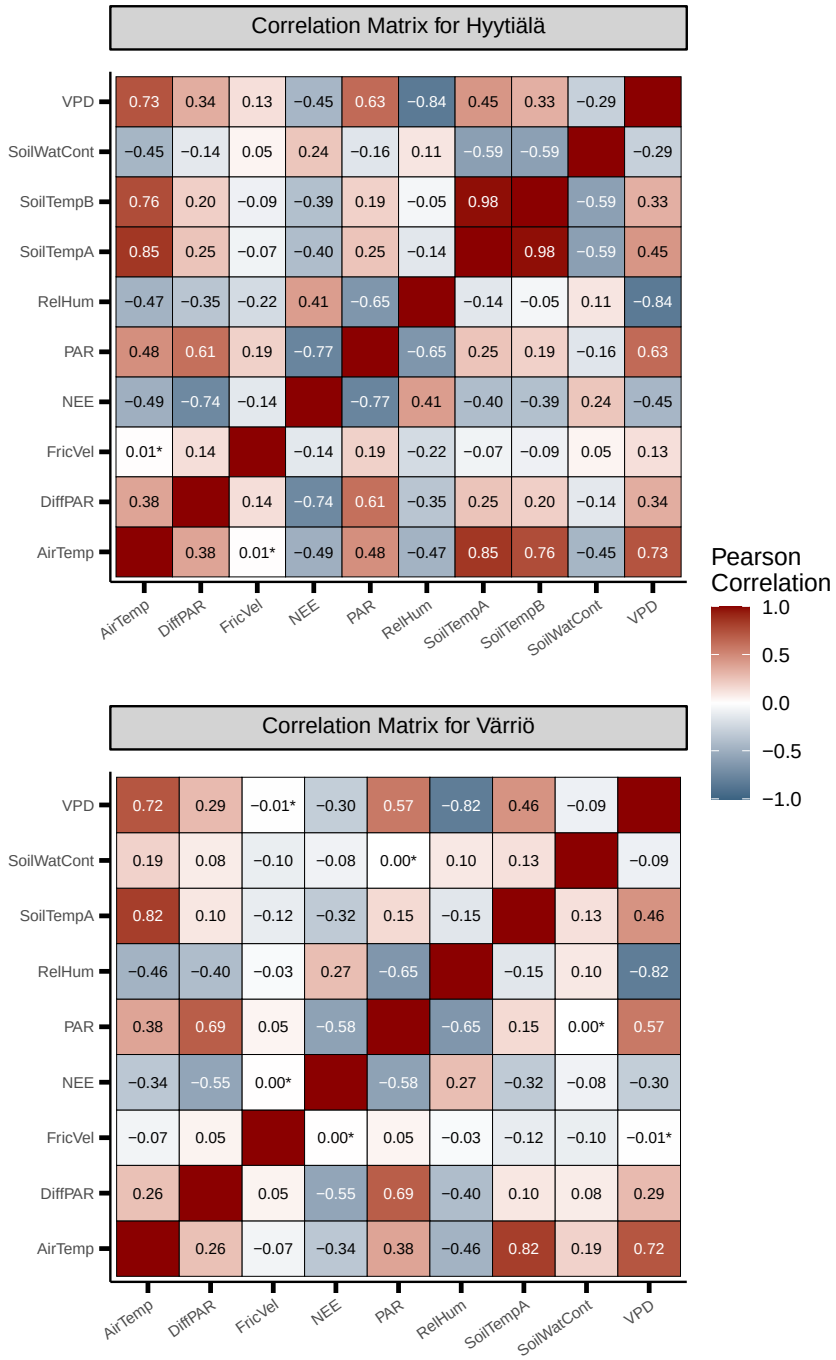

Figure A1. Heat maps illustrating linear correlation between input variables in Hyytiälä and Värriö. Statistically insignificant correlations are marked with *.

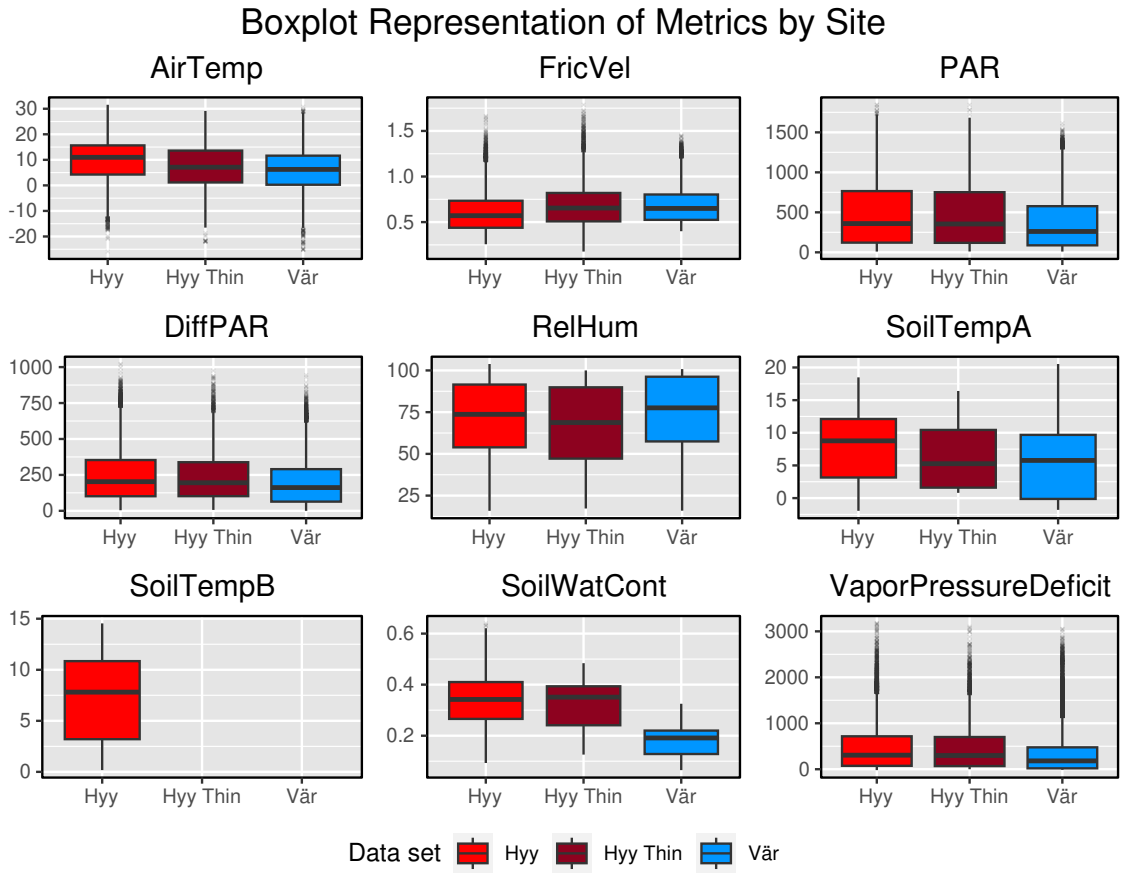

Figure A2. Box plots of input variables comparing Hyytiälä and Värriö data sets on the whole year time scale. 'Hyy' refers to Hyytiälä pre-thinned data set, 'Hyy Thin' - to Hyytiälä post-thinned, 'Vär' - to Värriö data set.

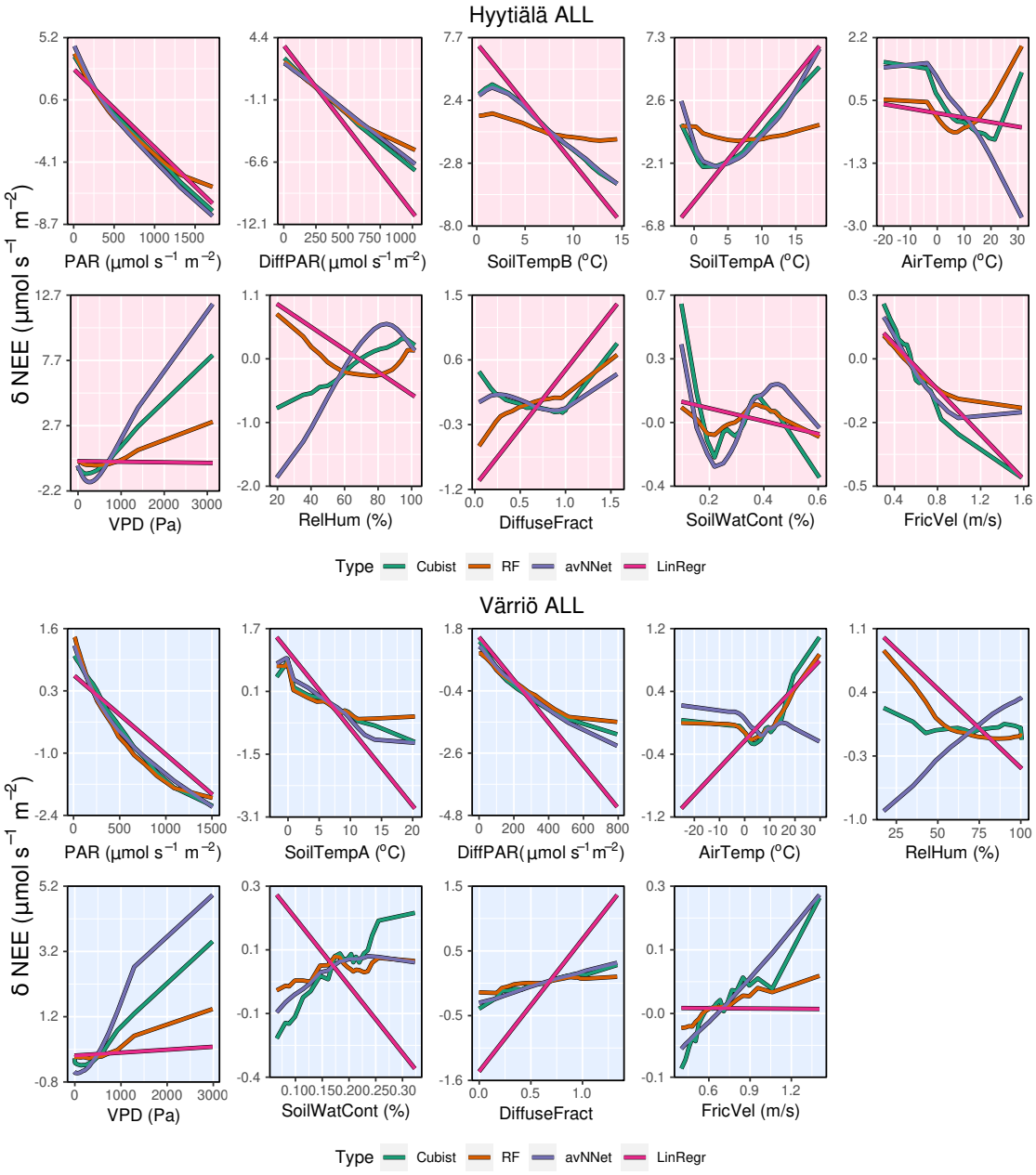

Figure A3. ALE plots for all the models trained on the whole-year data sets from Hyytiälä (upper panels) and Värriö (lower panels).

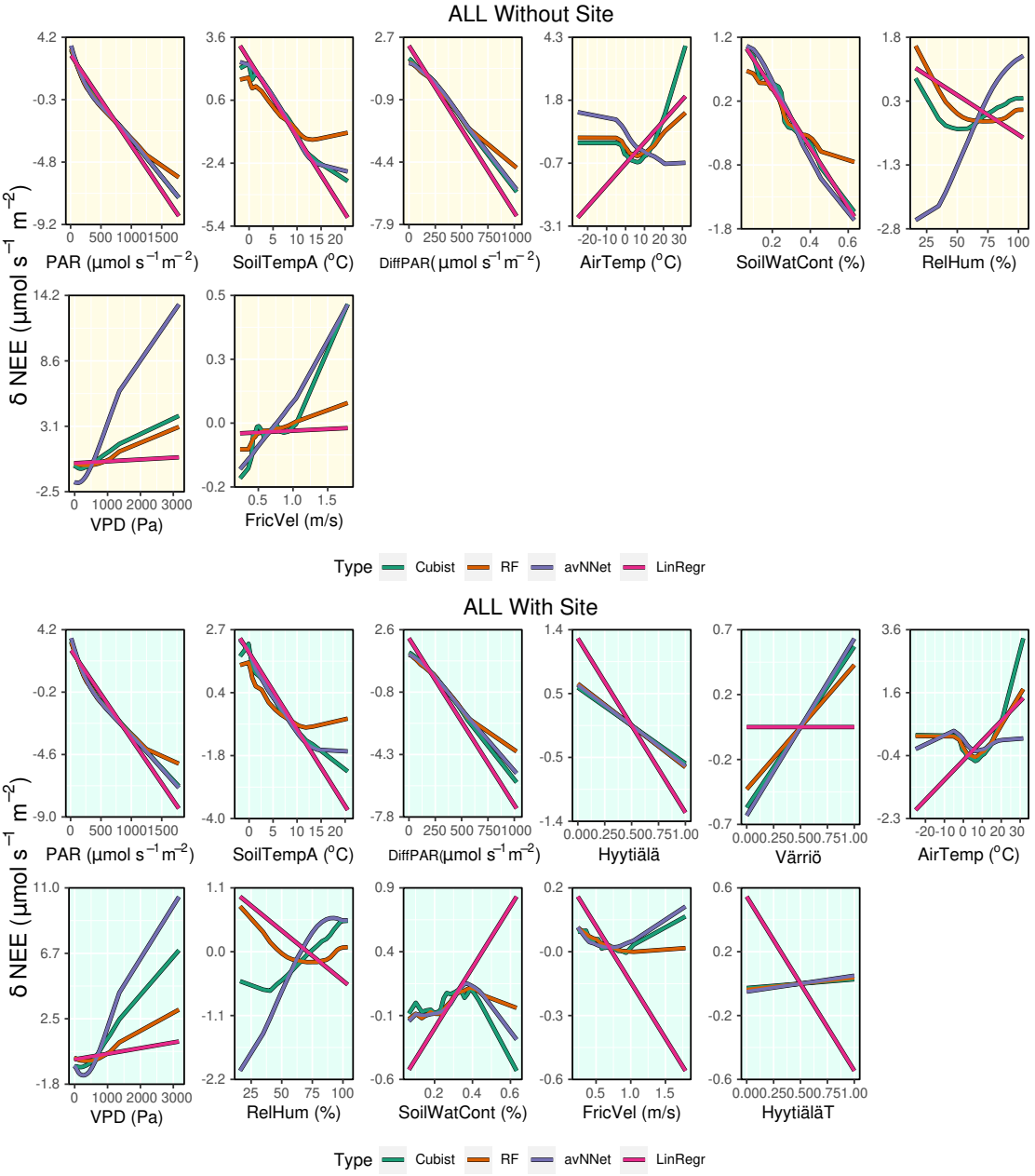

Figure A4. ALE plots for all the models trained on the mixed data sets containing ('With Site') and not containing ('Without Site') site variables. The data sets are from the whole year.

Figure A5. R^2-coefficients for all the models and different setups from Set 1 (Table 3). Here, the models were trained using equal amount of data points from Hyytiälä, Värriö and post-thinning Hyytiälä (balanced data sets). In each of the four panels, the results for the training data set are shown on the right (marked 'Train'), and the results for the test data set are shown on the left (dotted bars, marked 'Test'). 'ALL' denotes the scores for the models using the whole year data sets; 'PEAK' - for the models using the peak growing season data sets. 'With Site' - the input variables contain the information about site, 'Without Site' - no information about site.

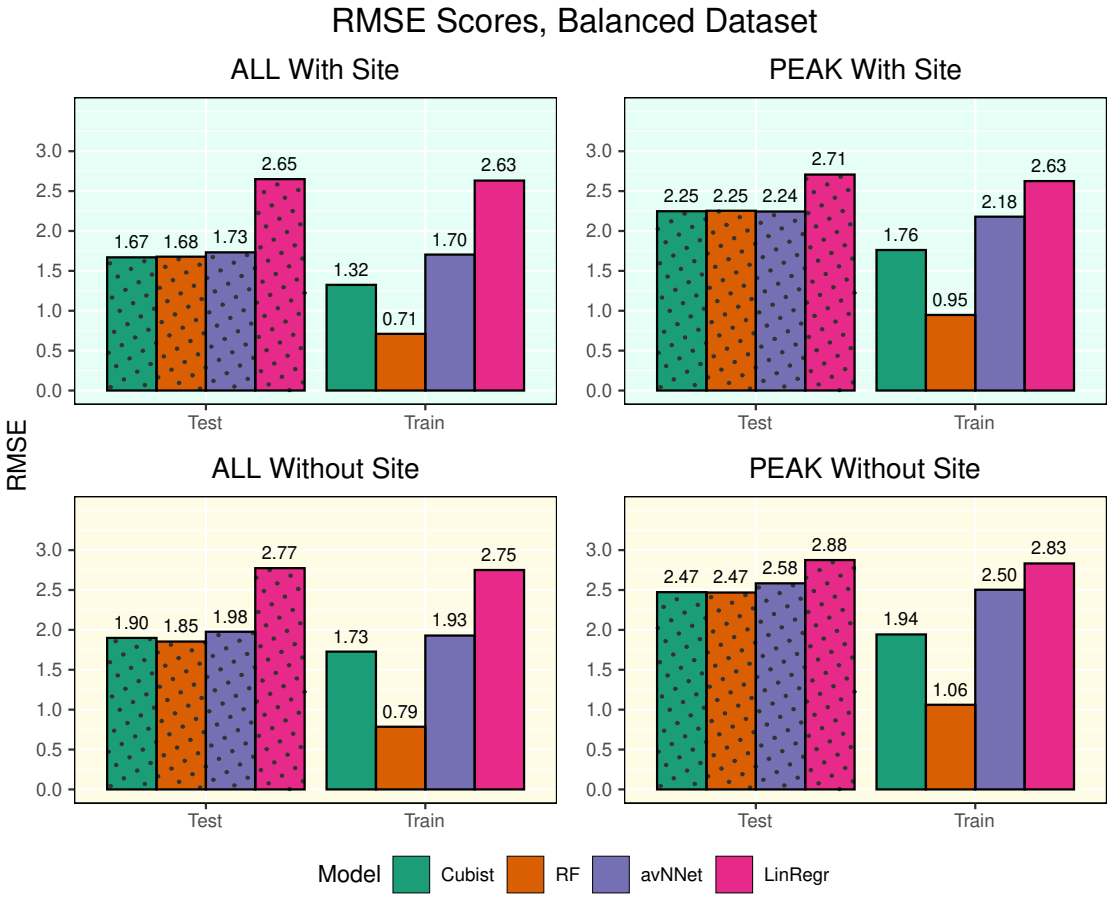

Figure A6. RMSE for all the models and different setups from Set 1 (Table 3). Here, the models were trained using equal amount of data points from Hyytiälä, Värriö and post-thinning Hyytiälä (balanced data sets). In each of the four panels, the results for the training data set are shown on the right (marked 'Train'), and the results for the test data set are shown on the left (dotted bars, marked 'Test'). 'ALL' denotes the scores for the models using the whole year data sets; 'PEAK' - for the models using the peak growing season data sets. 'With Site' - the input variables contain the information about site, 'Without Site' - no information about site.

Table A1. Three most important features for different models and Set 1 (Table 3)

Model	RF	Cubist	AvNNet	LinRegr	RF	Cubist	AvNNet	LinRegr
Peak	Hyytiälä				Värriö			
P1	PAR	PAR	PAR	PAR_{dif}	PAR	PAR_{dif}	PAR	PAR_{dif}
P2	PAR_{dif}	PAR_{dif}	PAR_{dif}	VPD	PAR_{dif}	PAR	PAR_{dif}	PAR
P3	VPD	$SoilTemp_A$	VPD	PAR	AirTemp	AirTemp	VPD	VPD
All	Hyytiälä				Värriö			
P1	PAR	PAR	$SoilTemp_B$	$SoilTemp_B$	PAR	PAR_{dif}	PAR	PAR_{dif}
P2	PAR_{dif}	$SoilTemp_B$	PAR	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$
P3	$SoilTemp_B$	PAR_{dif}	PAR_{dif}	PAR_{dif}	PAR_{dif}	PAR	PAR_{dif}	F_{dif}

Table A2. Three most important features for different models and Set 2 (Table 3)

Model	RF	Cubist	AvNNet	LinRegr	RF	Cubist	AvNNet	LinRegr
Peak	Without Site				With Site			
P1	PAR	PAR	PAR	PAR	PAR	PAR	PAR	PAR
P2	PAR_{dif}	VPD	PAR_{dif}	PAR_{dif}	PAR_{dif}	PAR_{dif}	VPD	PAR_{dif}
P3	SoilWatCont	PAR_{dif}	VPD	VPD	Värriö	Värriö	PAR_{dif}	VPD
All	Without Site				With Site			
P1	PAR	PAR	PAR	PAR	PAR	PAR	PAR	PAR
P2	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	$SoilTemp_A$	PAR_{dif}
P3	PAR_{dif}	PAR_{dif}	AirTemp	PAR_{dif}	PAR_{dif}	PAR_{dif}	VPD	$SoilTemp_A$