Answer to the Editor's comments

We are grateful to the Editor for raising the points that require some further discussion.

Please note, the rating of your manuscript is not overly high and this maybe due to weaknesses in the presentation that you now opt for improving.

It seems that the overall prevailing grade that the reviewers gave for our manuscript was 'good', but the second reviewer rated the 'significance of the topic' to be only 'fair' although this reviewer also considered the study 'interesting'. We hope the revised manuscript will succeed better in showing the topic's significance. In our study, we apply several popular machine learning models to model net ecosystem exchange in boreal forests. These machine learning models are often used as 'black box' models, meaning that their decision making is often not well understood. The novelty of our manuscript is that we visualize and examine the models' decisions on different input parameters based on the existing knowledge about ecosystem functioning. The study is truly multidisciplinary as the results produced by computer scientists are interpreted by atmospheric and forest scientists. We hope that the revised version of our manuscript accounting for the Reviewers' and Editor's comments will improve the low rating and make our study acceptable for publication.

But you may consider once more, whether you can discuss and summarise the most important achievements of the study a bit clearer. For example when discussing the importance values (Fig. 7 and Fig. C), add explanation, whether / why these findings make sense (and explainable ML a good choice). Did you learn anything unexpected form the explanations from ML?

We thoroughly went through the text and modified it following the Editor's suggestions. The discussion, addressing point by point the dependence of NEE on all input parameters for separate data sets, related to Figs. 7 and 8, was provided in subsection 3.1.2, and for the mixed data sets – in subsection 3.2.2 of the previous manuscript version. To improve the clarity of the discussions, we have split them into separate subsections focusing on feature importance (current subsections 3.1.2, 3.2.2) and ALE (current subsections 3.1.3, 3.2.3). We have also introduced similar logic in the subsections addressing ALE, starting the discussions with the most important variables as suggested by the feature importance diagrams.

In the case of separate sites and seasons, all the nonlinear models choose similar most powerful explanatory parameters (Lines 311-327), and compared to existing knowledge on ecosystem functioning, there were largely no surprises, which means that models in general work well. The choice of important variables becomes even more aligned when the mixed data sets are used (Lines 465-478). More surprising was to see how some models treat interdependent parameters (e.g., soil temperatures at different depths or VPD/RH), as they may produce strong but opposite dependencies of NEE on these parameters (Lines 373-387, 502-504). Another interesting finding was that for some models, soil water content that was low in importance among the set of features for separate sites suddenly becomes one of the most important variables for mixed data sets (Lines 471-473). This is accompanied by a drastic change of ALE for this variable (Lines 495-501) in all the models. As soil water content is the only variable that has a clear difference between Hyytiälä and Värriö data sets (Fig. A2), we interpret this new high position of soil water content in the feature importance diagrams so that the models treat it as a site parameter to distinguish between the Hyytiälä and Värriö data sets. This interpretation is supported by soil water content losing importance when site parameters are added as input variables.

We have also included text related to the added error bars due to various data splits (Lines 332-336): '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 Random Forest, which seems to be more confident than the other nonlinear models in its treatment of interdependent variables.'

For example you find that the intensity of diffuse radiation has a higher importance than the fraction of diffuse radiation. Why that?

We added the following discussion on this in the manuscript:

Lines 337-342: 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.

Lines 391-394: 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.

Then you explain that the VPD effect is rather explained by temperature than by relative humidity. Why that?

VPD is a function of both relative humidity and temperature as follows from eq. (2), and it is strongly correlated with both (Fig. A1). VPD influences photosynthesis via stomatal control. RH contribution to NEE is basically via this VPD effect on photosynthesis, whereas temperature, in addition, affects respiration.

We mention that VPD is dependent on both variables in various parts of the manuscript:

Lines 322-323: 'It is good to note that VPD is calculated based on air temperature (see Sec. 2.1), so these variables are not independent.'

Lines 332-333: 'For Värriö, Cubist and avNNet place interdependent VPD, RH and air temperature in the feature importance diagram within the error bar from each other'.

Lines 395-397: ‘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’.

Finally, you include the friction velocity as a variable, which doesn't yield a high importance score. This might be even trivial as the u* filtering is applied to the data sets, i.e. to exactly remove any relationship between u* and NEE.

We chose u* as one of the parameters following the setup in Moffat et al. (2010) to be able to compare with this study (and got the similar result that the variable is unimportant). While it is true that filtering is applied to exclude the lowest u* corresponding to non-turbulent conditions from the data sets, some relationship might still be there for higher values: actually, for Hyytiälä, there is some weak positive correlation, which we briefly mention in lines 405-409. The conclusion about NEE not depending on u* may serve as an additional checkpoint for the quality of the data set.

In your reply to Reviewer 2 you mention that fitting hyper-parameters reduced overall performance, which is to be expected. But you do not seem to explain, why and for which application a result less prone to over-fitting might be a better choice.

We are grateful to the Editor for stimulating the discussion on overfitting.

The K-fold cross-validation technique, which we have now used to find hyperparameters, can also be used to estimate the model performance on an unknown data set. 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. This procedure is repeated R times. Each time, we can calculate R2-scores and RMSE corresponding to the evaluation (holdout) data set. The average accuracy metrics obtained in such a procedure provide a reliable estimate of how the model is expected to perform on an unknown or test data set (e.g., Refailzadeh et al., 2009).

The R2-scores obtained from the cross-validation of our models' performance on the data sets from separate stations are now reported in Fig. 1 of this document. For all the models, the metrics agree with the scores reported in the manuscript for the test data set. The estimate obtained from cross-validation is even slightly lower than the results obtained on the test data set, likely due to the smaller size of the holdout subsets from the folding procedure. Therefore, we can conclude that all the models perform on the test data set with their expected scores.

However, when a final trained model is applied again to the training data set, the resulting scores sometimes can be high, as we see for both models based on regression trees, suggesting some overfitting. Nevertheless, it is not obvious if this degree of overfitting is necessarily bad, as mentioned by e.g., Zhang et al., 2023: ‘Note that overfitting is not always a bad thing. In deep learning especially, the best predictive models often perform far better on training data than on holdout data’. Based on the abovementioned arguments, we prefer keeping the hyperparameters in our models unchanged.

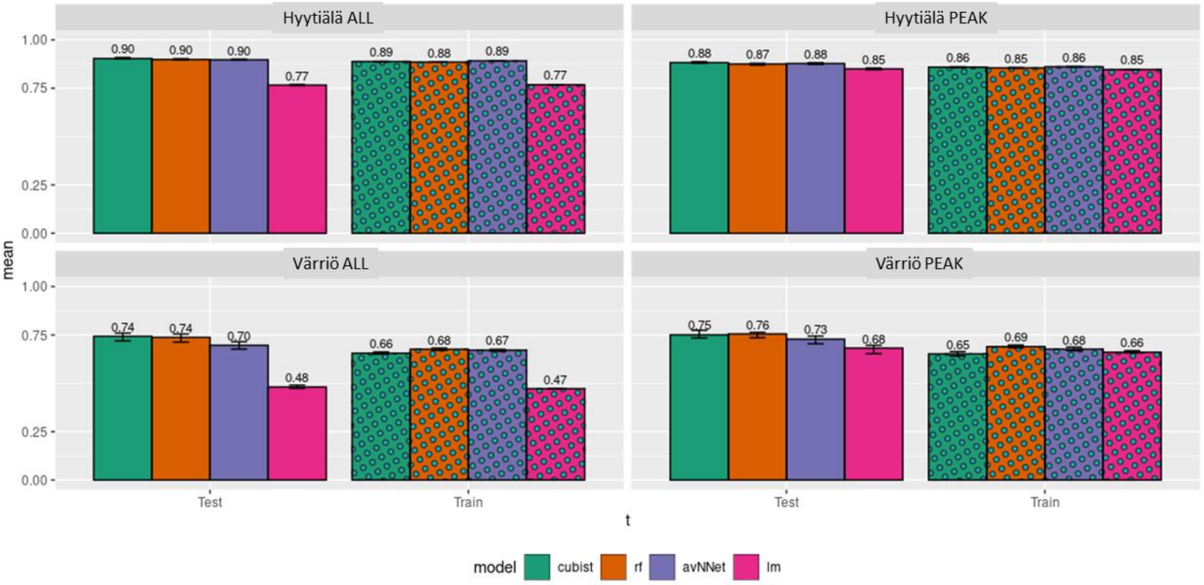

Fig. 1. R2-scores obtained by different ML models (see the legend) on test data (colored bars) and from the cross-validation procedure on the training data set (dotted bars). Panels correspond to the following setups: Hyytiälä ALL - the whole-year Hyytiälä data set, Hyytiälä PEAK – the peak season Hyytiälä data set, Värriö ALL – the whole-year Värriö data set, Värriö PEAK – the peak season Värriö data set.

Related to the editor's comment 'for which application a result less prone to over-fitting might be a better choice', there is unfortunately no general solution.

We added the following discussion in the manuscript: '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. To tune hyperparameters and estimate expected model performance, we applied K-fold cross-validation, 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. Finally, we obtained similar accuracy metrics on the test data sets from different nonlinear ML models, suggesting that our results are robust.'

Please reconsider the Reviewer 2's comment on the advantages of density contour plots over scatter plots. I do not think that the main point was the file size, but the loss of information when too many points are covered and thus hidden by the points on top in scatter plots. Here the density contour plot gives more information than scatter plots. Your explanation on why the scatter plots are better is a bit vague. Please provide the two alternatives and then justify your choice.

We thank the reviewer and the editor for the comment. Below, we have provided a density contour plot in Figure A and a density scatter plot in Figure B to illustrate what the alternatives would look like compared to the plot we considered using, Figure C.

Overall, the problem of showing two scatter plots in one figure is hard to solve in this context as the distributions of points are very close to each other (Fig. C). The difference is in the noisy outlier points, which disappear entirely when we try contour plots and become almost invisible if we make one set of points transparent.

Our aim with scatter plot Fig. 3 was:

1. to see if the correlations follow 1:1 lines;
2. to illustrate obtained R2 and RMSE results for test and train data sets and all setups. The figure helps to understand that different R2 could simply be the result of different data range for Värriö and Hyytiälä, and it is also helpful when discussing possible overfitting issues.

In Fig. 6, we compare the performance of different models on the same setup. The figure clearly shows that the RF and linear models look a bit different from the other models. In the case of RF, there are fewer black points around orange points, indicating that the RMSE in the training data set is smaller compared to the test data, while there are visible clouds of black points around color points for other models.

Overall, our suggestion is to visualize density distributions of points on the sides of the figures and reduce the size of the figures as shown in Fig. C of the current document.

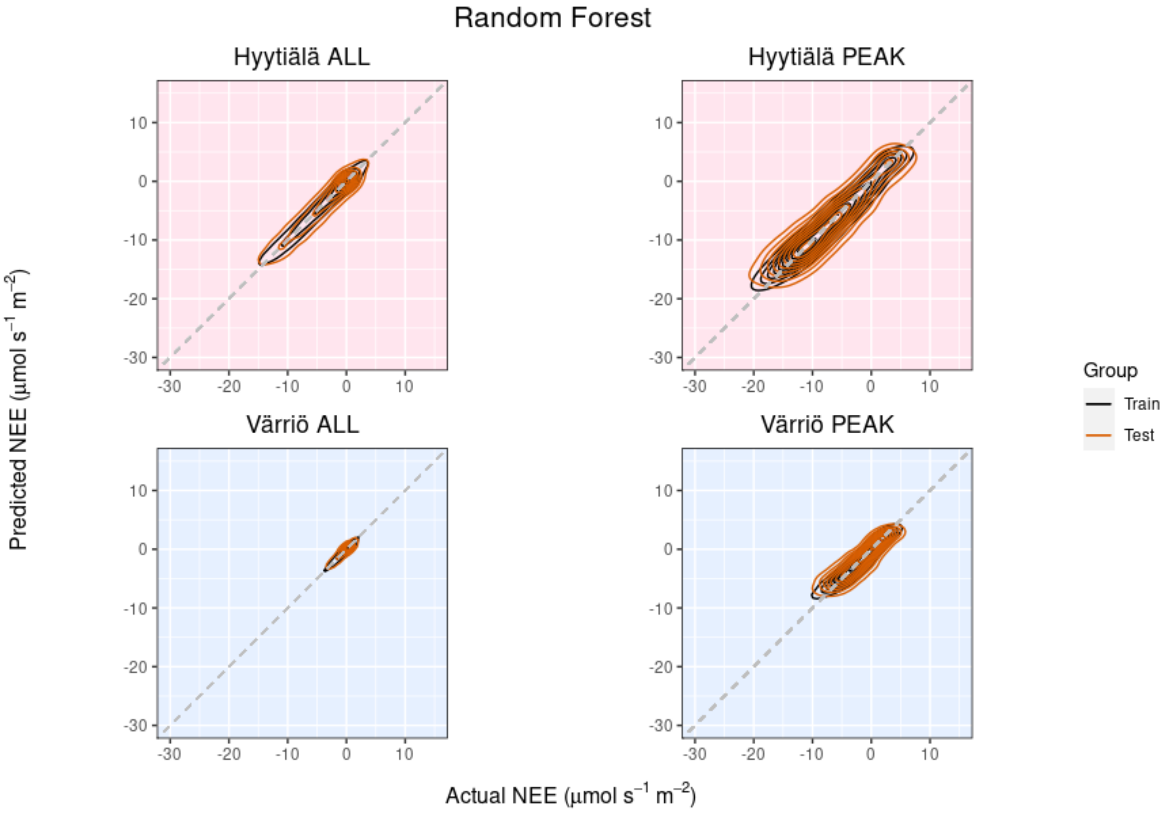

Figure A: A density contour plot alternative to Figure 3 of the manuscript.

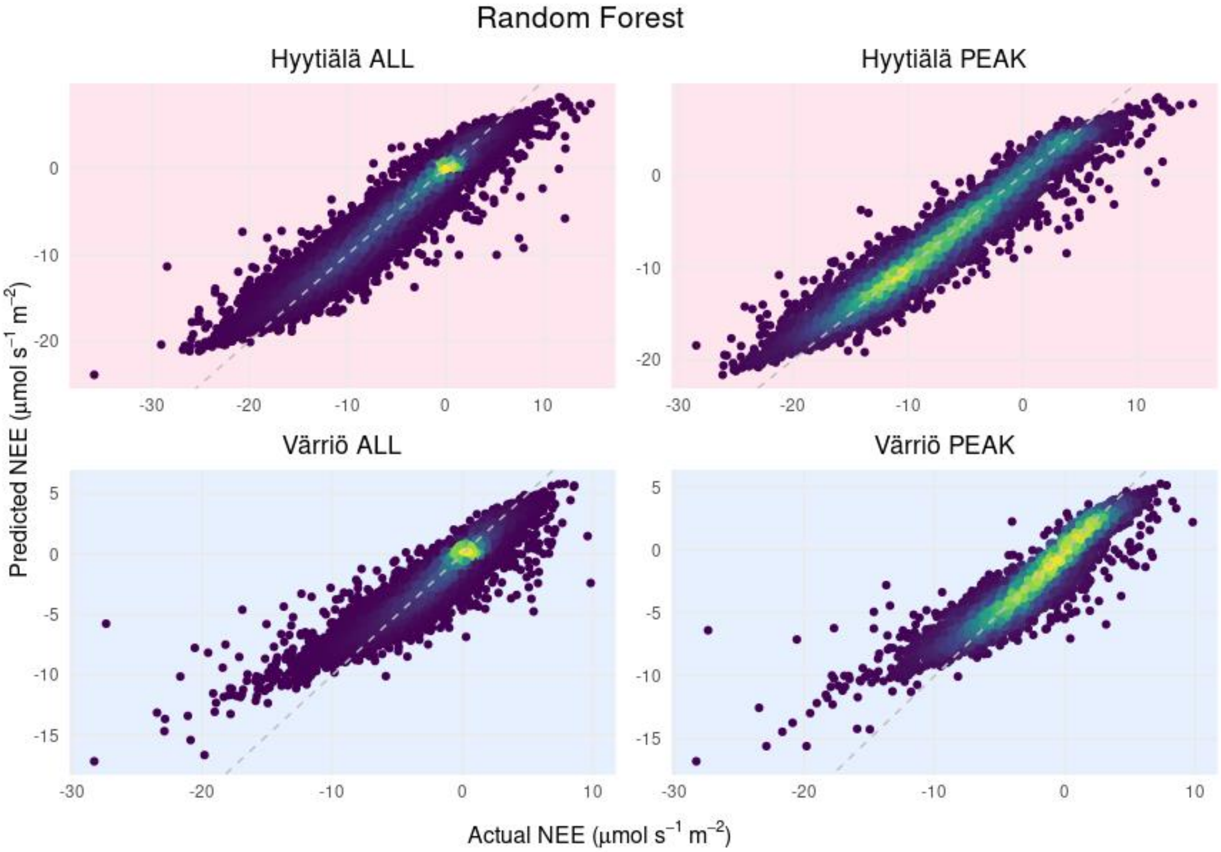

Figure B: A density scatter plot alternative to Figure 3 of manuscript.

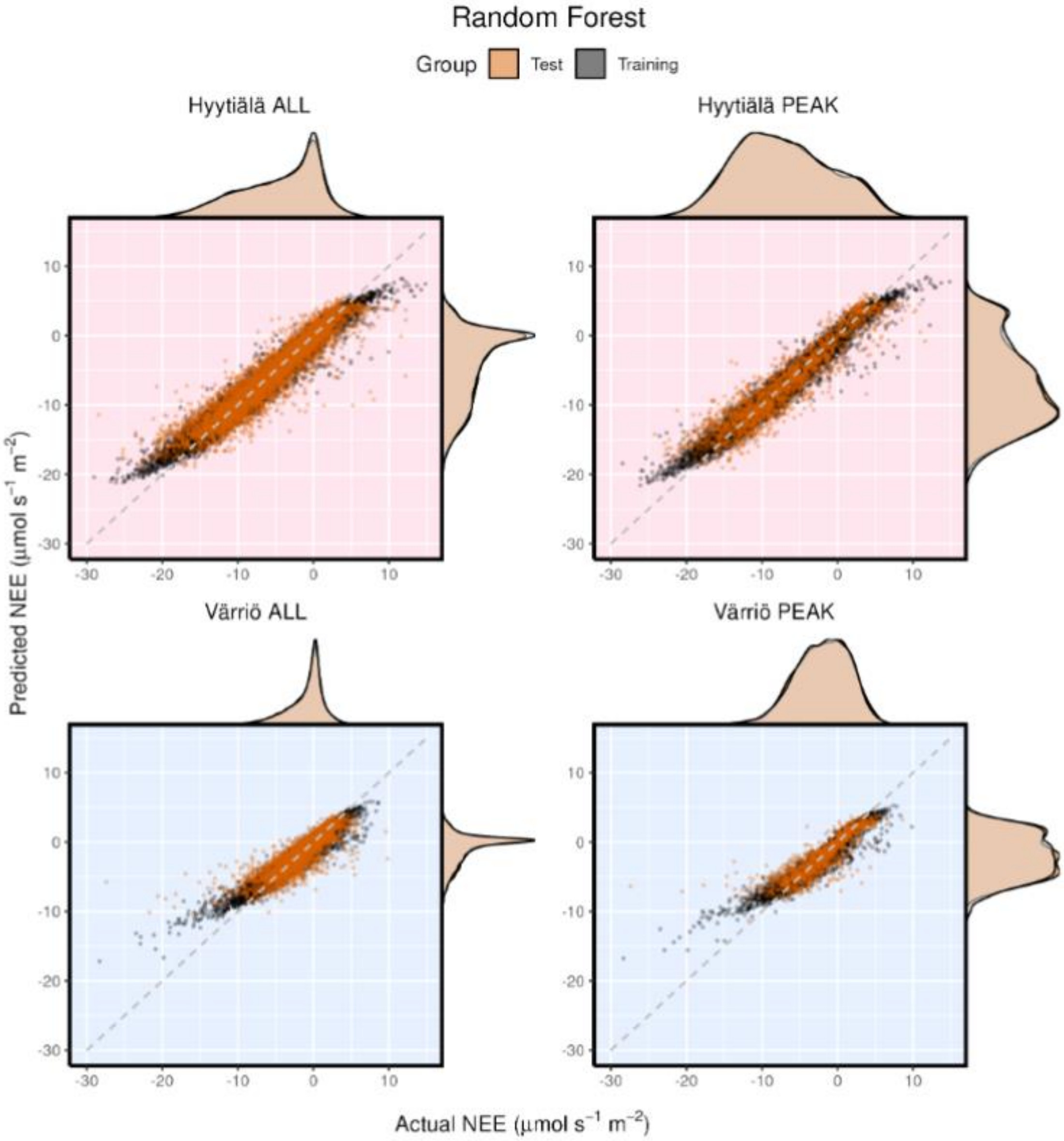

Figure C: The scatter plot used in the revised manuscript as Figure 3.

Finally, we have modified the abstract compared to the version sent to Reviewer 2. We have improved the clarity and included a sentence about the mixed data set results, which were previously omitted. In the abstract, we have also updated the R-scores compared to how they were presented in the response to Reviewer 2. In this response document, we had a small error in Figure A illustrating the results for Värriö Peak, with the scores lower (0.71-0.74 while the correct numbers were 0.73-0.76) compared to the original manuscript and to its current revised version.

We thank again the Editor for the useful suggestions. We hope that you will find that the present manuscript addresses all the comments raised.

References

Refaeilzadeh, P., Tang, L., Liu, H. (2009). Cross-Validation. In: LIU, L., ÖZSU, M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023). *Dive into deep learning*. Cambridge University Press.