Reply to the Reviewer 2 comments

We are grateful to the Reviewer for providing valuable insights, which helped to improve the clarity of the manuscript. Please find below our replies to specific comments:

The abstracts need to be improved by adding more details. The current version is a little vague. For example, the current version only mentioned four machine learning models. It is better to explicitly elaborate what kinds of four machine learning algorithms were used in the manuscript. Furthermore, model performance and statistics are also better to be presented in the abstract. In addition, the manuscript highlights the explainable machine learning to quantify NEE drivers. However, no details on which drivers are most important for NEE predictions in the abstract.

We thank the reviewer for the comment. We added the requested information in the abstract, which now reads:

'We apply four machine learning models (Cubist, Random Forest, artificial neural network and linear) 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 Värriö) and model NEE during the peak growing season and the whole year. All nonlinear models demonstrated similar results with R^2=0.88-0.90 for Hyytiälä, and R^2=0.70-0.76 for Värriö. Using Explainable Artificial Intelligence methods, we show that three most important input variables during the peak season are photosynthetically active radiation, diffuse radiation and vapour pressure deficit (or air temperature), whereas on the whole year scale, vapour pressure deficit is replaced by soil temperature'.

We are grateful to the reviewer for pointing out the source of uncertainties and potential overfitting, which showed that nonlinear ML models give comparable results, and we changed the text accordingly.

The manuscript was only conducted once for training and testing dataset splitting. As we know, there are always uncertainties in the data splitting. It is better to conduct data splitting multiple times to also present uncertainties of R2 and RMSE in Figures 1 and 2.

We thank the reviewer for the valuable suggestion. We agree that uncertainties in data splitting can affect the reliability of our results, and therefore, as was mentioned in Methods, we used k-fold cross-validation in our study. This approach improves the robustness of our findings and allows us to obtain more reliable estimates of R^2 and RMSE. However, we have now done model training with multiple different data splits, and observed that they are consistent with our results in Figures 1 and 2 (Figure A). We added uncertainties in Fig. 1-2.

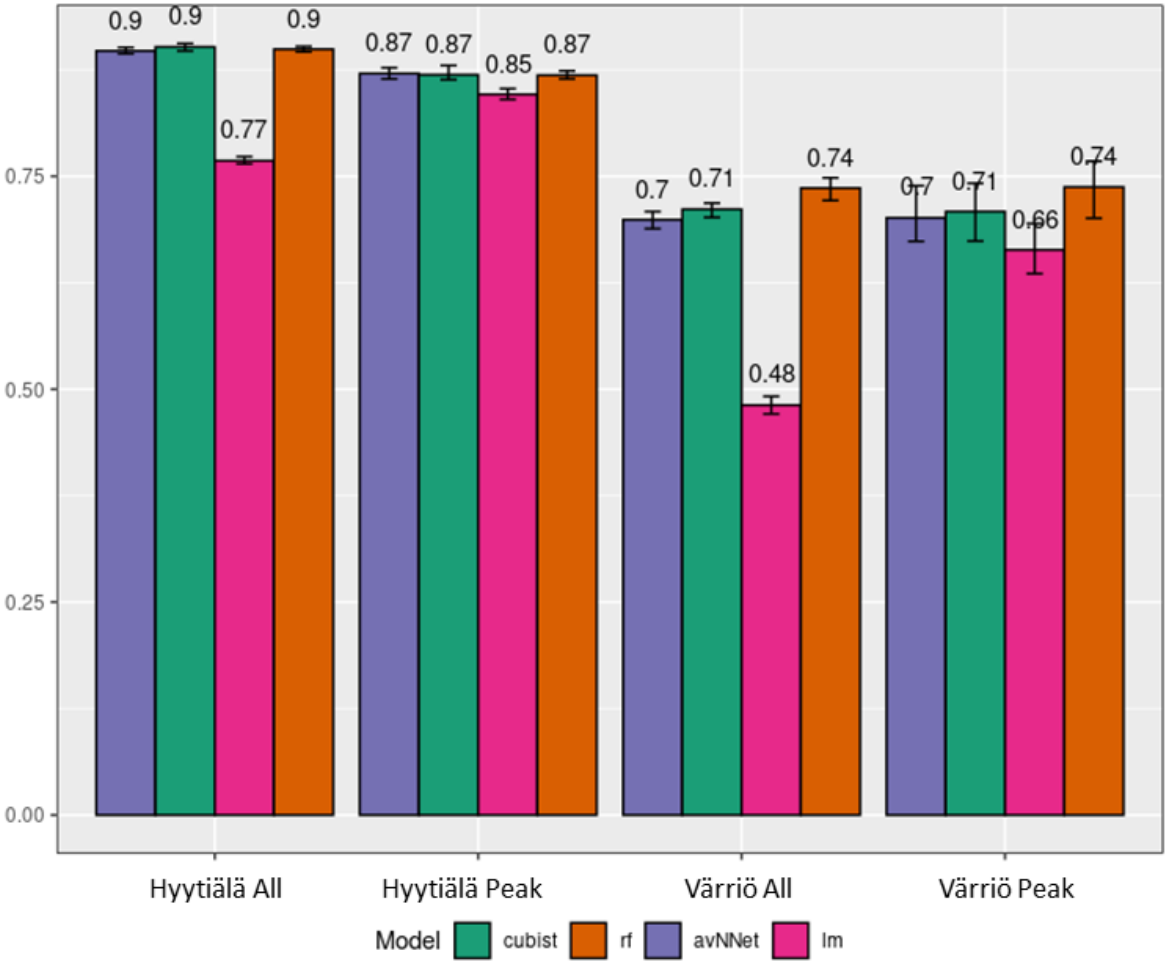

Figure A: ML models trained on different splits of the data. The most variance seen in the scores is in Värriö Peak, which is most likely due to it being the smallest data set. Overall, the scores are consistent with the results shown in the manuscript.

From Figures 1 and 2, it seems that some machine learning models are overfitting, which means that training performance is much better than testing performance. It is better to tune hyperparameters of machine learning models to avoid machine learning model overfitting.

We thank the reviewer for this comment. We have used k-fold cross-validation, as well as internal model hyperparameter tuning, which both reduce the likelihood of overfitting [1,2]. We observed the mean 0.16 difference in R^2 score between the testing and training performance of Värriö setup for Random Forest and Cubist. Model hyperparameter tuning resulted in poorer training performance would always result in poorer test performance.

Figure B below illustrates the Random Forest performance on Värriö All setup when different hyperparameters that can have effect on overfitting are tuned. The most crucial hyperparameter in regards of overfitting, min.node.size [2], shows that increasing it decreases both testing and training model performance, as well as the difference between them. This points at decreasing overfitting. At the largest value of min.node.size used here, we got the model test R^2 score of 0.70, i.e., closer to the neural network result, which was the lowest among the nonlinear models.

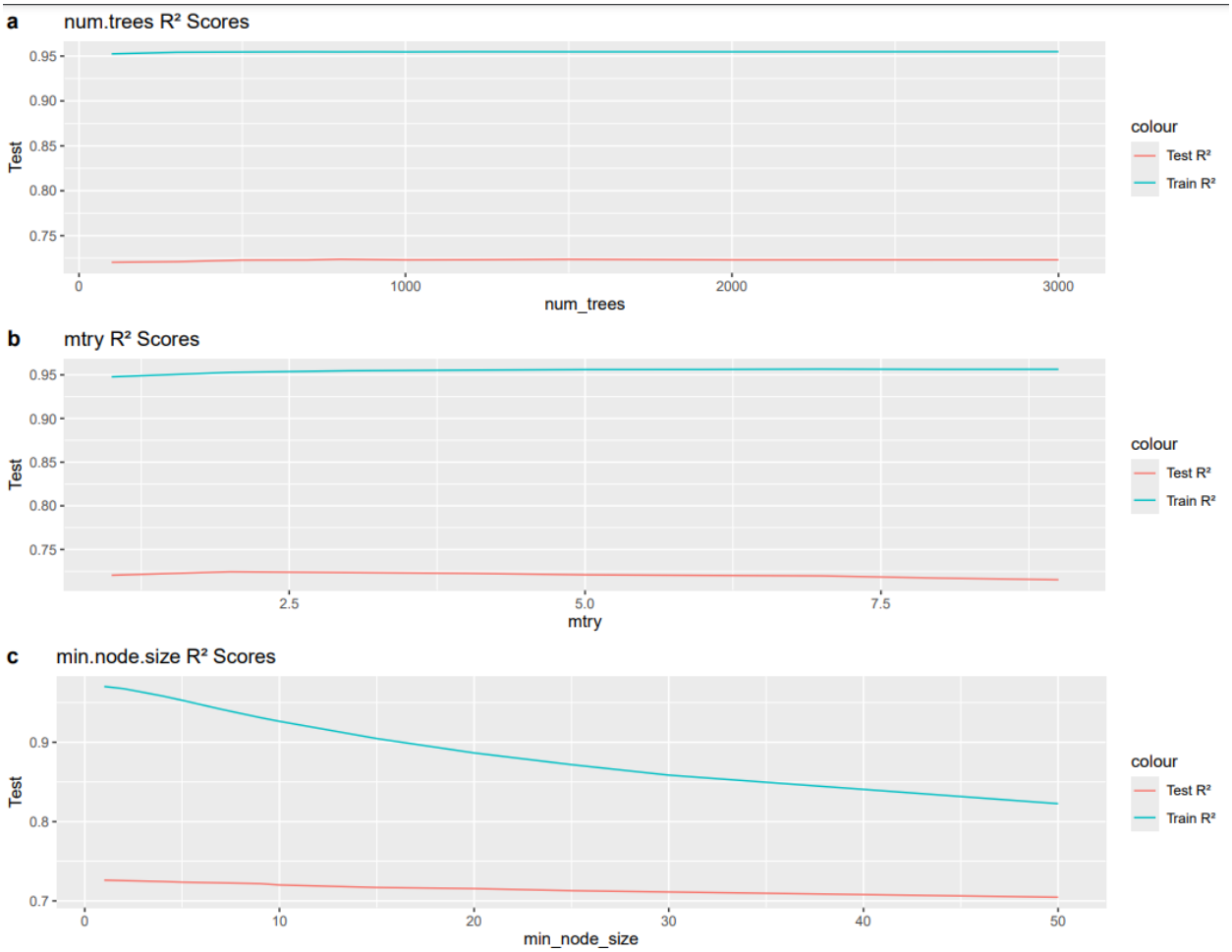

Figure B: Experimental tuning of Random Forest hyperparameters. Tuning these hyperparameters can reduce overfitting, but in this case reducing the overfitting resulted in poorer training and test scores, indicating overall poorer fit while the difference between training and test scores did become smaller.

Figures 3 and 6. There are too many points in the scatter plots. It is better to use density scatter plots to illustrate the results.

We agree with the notion that there are too many points in the scatter plots to the degree that loading the images takes too long. However, after comparison of scatter and density plots, we felt that scatter plots better tell the story of how the different datasets deviate from each other than density plots. We have reduced the file size instead to make inspecting the images smoother.

***Figures 4 and 5. It is better to add units to the y-axis*.**

We added the units to the y-axis in both images.

***Figures 7 and 11. Better to add uncertainty bars, once you have done different data splitting*.**

We thank the reviewer for the suggestion. While we recognize the importance of representing variability with uncertainty bars, our study uses a consistent data split across all XAI methods to ensure comparability. This approach helps us to maintain control and attribute differences directly to the XAI techniques rather than data variability.
However, as the R2 and RMSE scores have some deviation for Värriö PEAK data set, we decided to add make feature importance for various different splits to ensure that the results are consistent with different data splits. The results can be seen in Figure C:

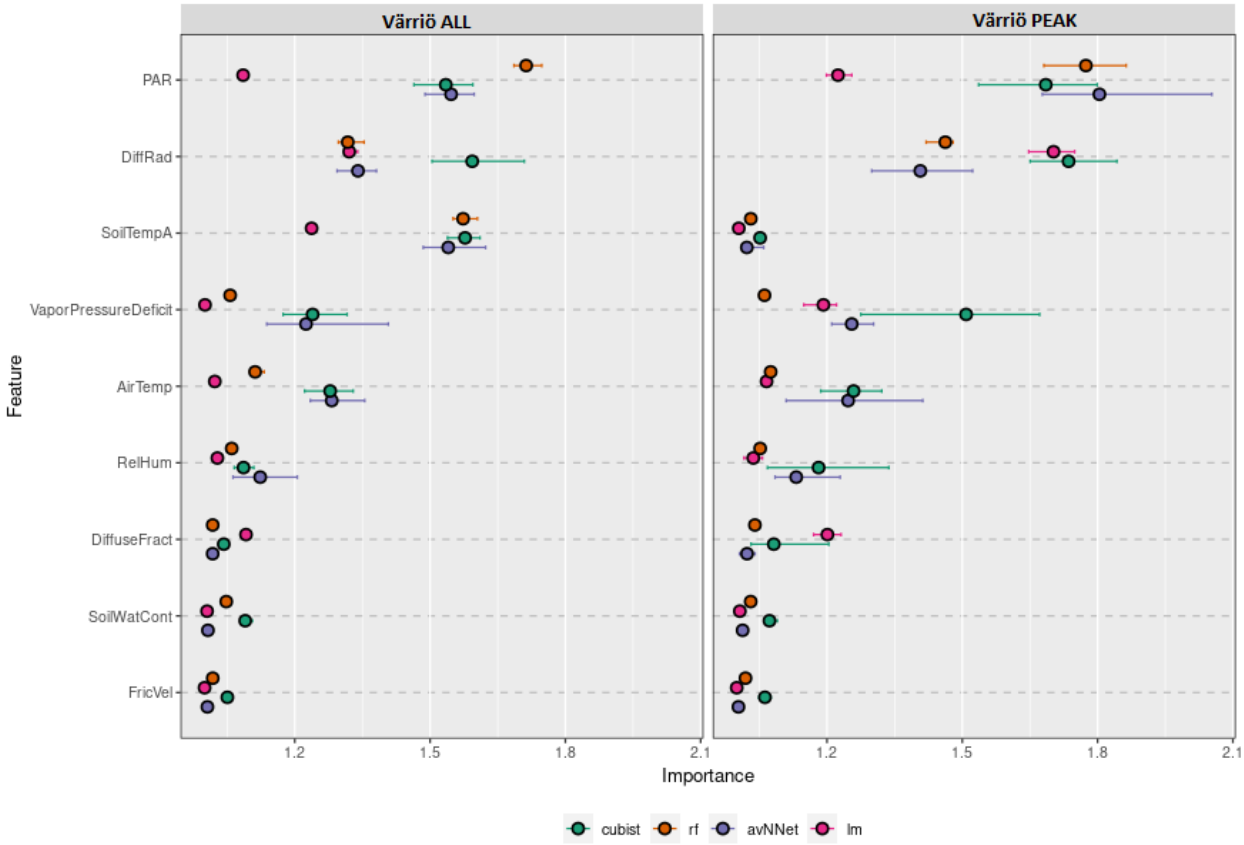

Figure C: Feature importance for Värriö ALL and Värriö PEAK with uncertainty bars, where Feature Imporance was done to multiple different data splits of these data sets. All were done on test data, data that the model was not used to train on. The points display the mean value, while the error bars display the maximum and minimum values. Here the ordering is based on the mean importance acrossa all models, instead of just Random Forests, as is in the manuscript. The results are consistent with what we present in the manuscript.

Figure A1. It is better to add the significance levels for the correlation analysis.

We thank the reviewer for this suggestion. Given that almost all the correlation coefficients in our analysis show p-values significantly lower than the conventional threshold (e.g., $p < 0.05$), indicating statistical significance, we added asterisks in Figure A1 where needed to mark nonsignificant values.

We thank the reviewer again for the useful suggestions. We hope that our answers address all the comments raised.

References

[1] Bengio Y, Grandvalet Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. Journal of Machine Learning Research 5 (2004) 1089–1105.
[2] Barreñada, L., Dhiman, P., Timmerman, D., Boulesteix, A. L., & Van Calster, B. (2024). Understanding random forests and overfitting: a visualization and simulation study. *arXiv preprint arXiv:2402.18612*.