Reply to the Referee's comments

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

Introduction: The introduction provides sufficient background information to understand the aim of the paper, however, it often comes across as overly explanatory and some sections could be synthesized and written more concisely to cut down on the word count. Consider revising this section to reduce unnecessary information and improve the flow of the background towards the objectives of the study.

We agree with the comment and have now made the information flow better. Redundant information was removed, and we have focused the text on our research goals.

Line 19-23: Awkward wording. The sentence could flow better, and currently does not come to a satisfying conclusion.

We agree with the comment. We have rearranged the sentence as follows: *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.*

Line 26: The word "variable" from variable values can be removed, it flows better without and feels like it is already implied from how the sentence is written.

We agree with the comment and have removed this word from the sentence.

Line 49-50: This sentence is jarring and does not give any obvious rationale for why you are now mentioning measurement of different temporal scales. I would consider revising to improve the flow of the paragraph and maybe add a rationale as to why this matters for the paper. Maybe start by discussing the difference between gapfilling and upscaling studies/questions and then discuss how they are typically measured at different temporal scales, so if a researcher would want to look at both they would need multiple scales of data.

We have removed the discussion about the temporal scales from the introduction, because, as the reviewer pointed out, this is not the focus of our manuscript.

Line 74: Referring to subhourly time resolution - Does this relate to the findings of the paper, since you state earlier that upscaling studies typically use longer timescales and upscaling is the part you had the hardest time modeling? Why did you not try using data from multiple temporal resolutions to model both the gapfilling and upscaling if you knew that different resolutions were better for modeling different kinds of data?

We have removed the discussion about temporal scales from the Introduction.

Line 103-105: It is recommended to be consistent with the ordering of the sites throughout the methods. If you start with SMEAR I (Värriö) then SMEAR II (Hyytiälä), it would be best to always refer to them in that order to avoid confusion.

We agree with the comment and rearranged the text for the ordering to be more consistent, i.e. Hyytiälä is discussed first and Värriö after it in all notations.

Line 142: Referring to training vs test data - You should mention what the test sets were as well and how they were selected. It seems like it was 4% of observations used for testing, except for post-thinning Hyytiälä which used 3%, why?

We understand the confusion, and added clarifications on how much of the data was used for the test/training data: 75% for training and 25% for test for Hyytiälä and Värriö. In the case of the mixed model, 80% of the data was used for training and 20% for testing. These are standard portions used in ML modeling.

Line 143: Referencing the phrase "individual sites" - Does this separate pre- and post-thinning Hyytiälä, so there are three sets of all season data, and three sets of peak season data, then one set of all data combined, correct? It may be good to be more explicit about what constitutes as individual sites since pre and post-thinning Hyytiälä are from the same site.

We apologize for not being clear. We added a table that summarizes information about all numerical experiments. We hope that the table will help to distinguish between different research cases. The table is as follows:

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

Set	Site/Data Period	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 Site All	Models trained on the mixed data set from both sites, including post-thinned Hyytiälä, entire years, no site labels
	All Site All (Label)	Models trained on the mixed dataset from both sites, including post-thinned Hyytiälä, entire years, sites labels included
	All Site Peak	Models trained on the mixed dataset from both sites, including post-thinned Hyytiälä, peak growing seasons, no site labels
	All Site Peak (Label)	Models trained on the mixed dataset from both sites, including post-thinned Hyytiälä, peak growing seasons, site labels included

Line 210: Another small comment, start with the ALE plot paragraph since the method is mentioned first, or mention Permutation Feature Importance first in the preceeding paragraph. It is helpful to be consistent with the order things are discussed.

We agree with this comment and have fixed the ordering to be more consistent regarding the order in which the concepts are discussed, with Feature Importance first, and ALE plots second.

Line 217: Estimates

We included this missing word in the text.

Line 273: Refer to either R^2, R-squared, or R-scores throughout the document, do not switch between them.

We thank the referee for the comment. This is fixed to be more consistent across the text, now being referred to exclusively as R^2-coefficient.

Line 283: remove ":", instead use ";". Also "accounted here", should be "accounted for here".

We thank the referee for the observation and fixed these typos.

Line 365: I would change the formatting used for this section.

We agree with this note. Formatting was changed here to be consistent with the other parts of the paper.

Line 418-421: I think this distinction is unnecessary, you could just state that you distinguished between the sites by coding them to three dummy variables.

We agree with this comment and have now changed the text as follows: *we introduce three binary variables that identify the site. Three binary variables were used instead of a single categorical one due to some models requiring real numbers as input.*

Line 427: Have you tried running the models after standardizing the number of observations included from each site? Ideally twice to compare the same time periods for Värriö and pre-thinning Hyytiälä, and Värriö and post-thinning Hyytiälä. This could help prevent the scores from following a single site just because it was sampled more.

We thank the referee for this important question. We ran experiments using as a training dataset a balanced dataset with an equal number of data points from Hyytiälä, Värriö and post-thinning Hyytiälä. The results were consistent with the results obtained from a non-balanced dataset, still having R^2-coefficient close to that of the site with the better score, Hyytiälä, even though the number decreased marginally (by about 0.02 for non-linear models). RMSE calculated using the unbalanced data set (1.60-1.68, all season) are lower than RMSE calculated using only Hyytiälä data (1.74-1.79), possibly because RMSE for Värriö and Hyytiälä post-thinned data sets are smaller. For the balanced data set, RMSE increases (1.67-1.73) but still below

RMSE for Hyytiälä alone. We therefore conclude that adding more data points from other sites does not necessarily make the predictions worse, especially if there is a site identifier, but makes the predictions of the sites that have additional data points somewhat more accurate. We have added a brief discussion about the balanced data set in Results and score figures in the Supplementary material.

Line 466-468: Generally, this should either be in brackets inside the other sentence, or have the brackets removed.

We thank the referee for this observation, the brackets have been removed.

Figure captions: I believe figures should stand on their own without requiring the reader to have read either the main text or other figure captions. Most of your figure captions are a single line and not very descriptive. Even, as in figure 2, where you ask the reader to refer to an earlier caption is missing from the other figure captions. It would be best if you wrote full captions for all figures.

We agree with this comment; we have made the figure captions more descriptive.

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