Comments to the author:

The Referee recommends minor revisions and I agree with their assessment. Please address the remaining comments in a brief letter addressed to me and improve the manuscript accordingly and I would be happy to recommend its acceptance for publication in Biogeosciences.

Response to the editor: In the last round of minor revisions, we have addressed three minor comments from Referee #3. In light of your decision to address the remaining comments, in this round of minor revisions, we further refined our response to the third comment of referee #3 and improved the manuscript accordingly.

In addition, we have checked the manuscript for possible typos, etc.

Response to remaining comments from Referee #3

Comment 3. Please add an explanation for the reason why MLR, RF, SVM, and ANN are separately compared instead of all models and why PLSR with high R2 is removed in the Methodology section. This response is similar to the response for L198.

Last version Response: elaborated as 'Subsequently, the model accuracies corresponding to different levels of various features are compared in a cross-study fashion. In the evaluation of algorithms and time scales, we also implement comparisons within individual studies. For example, in the evaluation of the effects of the algorithms, we compare the accuracy of models using the same training data and keeping other features as constants in individual studies. In this intra-study comparison step, only algorithms with relatively large sample sizes in the cross-study comparisons were selected.'

Updated version response and revision: elaborated as 'Subsequently, the model accuracies corresponding to different levels of various features are compared in a cross-study fashion. In the evaluation of algorithms and time scales, we also implement comparisons within individual studies. For example, in the evaluation of the effects of the algorithms, we compare the accuracy of models using the same training data and keeping other features as constants in individual studies. In this intra-study comparison step, only algorithms with relatively large sample sizes in the cross-study comparisons were selected. In this study, algorithms with less than 10 available model records are not considered to have a sufficient sample size and we do not give further conclusive opinions on the accuracy of these algorithms due to their small samples (e.g., PLSR and BART with high R-squared but very few records as evidence). MLR, RF, SVM, and ANN were found to have large sample sizes (Fig. 5a), and thus their accuracies can be comparable. Based on this, in the intra-study comparison step, we only compare the accuracy differences between MLR, RF, SVM, and ANN in the context of using the same data and the same other model features (Fig. 5b).' **(Line 188)**