Comments from the Reviewers:

Reviewer #1 (Formal Review for Authors):

I appreciate the author's additional analysis. The authors now tested a very simple RF model which only uses latitude, longitude, month and year as predictors. They found that this model has a prediction skill comparable to the ERF model for FLUXNET sites but that the predicted global maps look unrealistic. The authors are right that this simple model cannot be used to make global GPP predictions. However, this was exactly the point I was trying to make: an excellent model performance for the FLUXNET sites does not guarantee this this model skill can be transferred to the global scale. Instead, the excellent model performance at site level is inherent to way RF works, which gives it an advantage to other approaches. This does not mean your global GPP maps from the ERF model are useless but the issue has to be made very clear in the paper to avoid giving an unrealistic level of confidence in the predicted maps.

REPLY: Thanks for your comments. We totally agree with you. You're thoughtful. Machine learning models have an inherent advantage, which is that even if we use variables unrelated to GPP (such as months) to estimate GPP, as long as they have the same variation characteristics, the model accuracy will be high. However, as we mentioned before, the ERF model contains remote sensing information, and some vegetation indices such as LAI have a good correlation with GPP, so the global GPP estimated by this model will not be so outrageous. In the revised version, we emphasized this point in discussions.

Due to the inherent advantages of the RF method, the accuracy of the model was comparable to that of the ERF model, even if a very simple model that used longitude, latitude, month, and year as explanatory variables (Figure S11 a). However, the global GPP estimated by this model was not reliable (Figure S11 b). This means that it is unknown whether site-scale model can be fully applied to global GPP estimates. ERF model can overcome this limitation well. On the one hand, the explanatory variables used in the model are derived from GPP simulation in which contain a lot of remote sensing information, which can ensure that the global GPP estimated by the model is reliable. On the other hand, the second validation method also further shows that the ERF model has good generalization and has greater potential than other models in estimating global GPP.

The findings from the simple model also reinforce my second suggestion which would help to report model performance metrics that would likely be more realistic for the global maps. My suggestion was to use ALL the data from a single site exclusively either for model training or for model testing, e.g. if there are 100 sites use the data from 70 (or 80) sites for model training and test the model on the remaining 30 sites. However, the authors text reads as if they again split the data randomly. Separating the data non-randomly is important to make sure the model cannot use any data from this site to make predictions for a specific site. The prediction skill in this task would then be a better indicator of the global prediction skill of the model.

REPLY: Thanks for your comments. In the previous version, we used the random split. Based on your suggestion, we conducted additional analysis. We randomly selected the data of 70% sites for model training, and validated the data of 30% sites. As shown in Figure R1, GPP_{ERF} still maintains optimal model accuracy, indicating that the ERF model has greater potential for estimating global GPP than other models.
In the revised version, we have added this method:
In addition, we used a second validation method in which all data from 70% of the sites were selected for modeling and only all data from the remaining 30% of the sites were validated, a process that was repeated 200 times. This validation will further illustrate the generalization of the model, i.e. its potential for estimating GPP without observations.
Combining the results of all flux sites, GPP_{ERF} explained 85.1% of the monthly GPP variations, while the seven GPP estimate models only explained 67.7%-81.5% of the monthly GPP variations (Figure 2). Another validation method also showed similar results, the average R^2 and RMSE of 200 validation results of ERF model were 0.822 and 1.68gC m^{-2} d^{-1}, which were obviously better than other models (Figure S3).

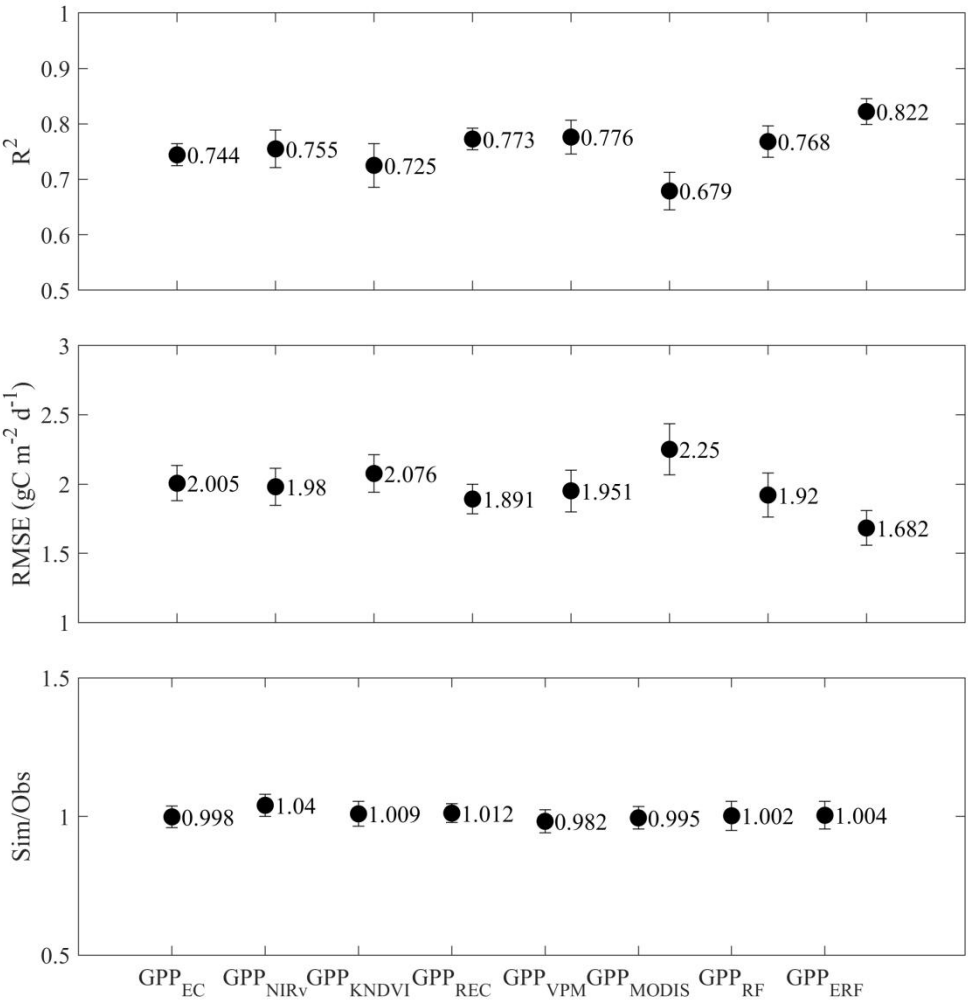

Figure R1. Validation results for each model on all data at 30% of the sites. The black dots represent the mean of the 200 validation results, and the upper and lower boundaries represent the standard deviation.

Finally, our paper took a lot of your time and effort, and we thank you again.