Comments from the Reviewers:

Reviewer #1 (Formal Review for Authors):

In my initial assessment, I raised concerns about the evaluation as the models "saw" the full FLUXNET data. Unfortunately, this concern has not been adequately addressed, possibly due to a lack of clarity in my communication. My main concern is that the RF algorithm may have an inherent advantage over the other models insofar that the authors selected the training data randomly, i.e. most of the data of a site goes into the training data while some into the validation. In simple words, when the RF model aims to predict e.g. the November 2010 value of site 1, it might just predict the average November value of this site (temporal autocorrelation). To address this issue and increase confidence in the performance of the ensemble model, I propose the following analyses:

REPLY: Thanks for your comments. We appreciate your time and effort. We are sorry that We did not fully understand your meaning in the first revision. First, the good performance of the RF model is related to the time change, which is mainly due to the seasonal change of vegetation. However, we do not consider our simulation results to be related to temporal autocorrelation. In RF models, estimates of GPP depend heavily on the characteristics of the inputs. In the initial modeling learning, GPP has established a good relationship with these features (because GPP, LAI and meteorological data have similar seasonal changes), that is, GPP=f (LAI,T,P). In the validation set, when there is a low (high) LAI input, the GPP estimate will also be low (high). That is, when we predict the value of site 1 in November 2010, the predicted value is actually determined by LAI, T, P, not just a multi-year mean. To further address your concerns, we also performed the analysis you presented.

- How does the author's ensemble model compare to the prediction skill of a very simple RF model which only uses longitude, latitude, month, and possibly year as predictors? This comparison would clarify whether the complexity of the ensemble model significantly improves prediction skill compared to a simpler approach.

REPLY: Thanks for your comments. As mentioned above, RF models that only consider longitude, latitude, year, and month are also able to achieve good simulation accuracy due to seasonal changes in vegetation (Figure R1a). In the importance analysis, we find that the importance of the month is as high as 64% (Figure R1b). However, this approach is not advisable. Because our goal in modeling at the site scale is to obtain the spatial distribution of GPP at the global or regional scale. Without the addition of remote sensing and meteorological data, the spatial distribution of GPP obtained by this simple model is completely unreliable (Figure R1c). Therefore, in the current GPP simulation, it is only meaningful to compare the simulation performance of different GPP models when remote sensing (and meteorological data) are added. In addition, we further demonstrate the simulation accuracy of this simple model at two sites (GPP_{test} in Figures R2 and R3). Due to the absence of remote sensing and meteorological data, the model simulates a very small difference in the inter-annual variation and the seasonal variation (because the model is mainly driven by months), so it is actually similar to the case where the simulated results are an average as you mentioned.

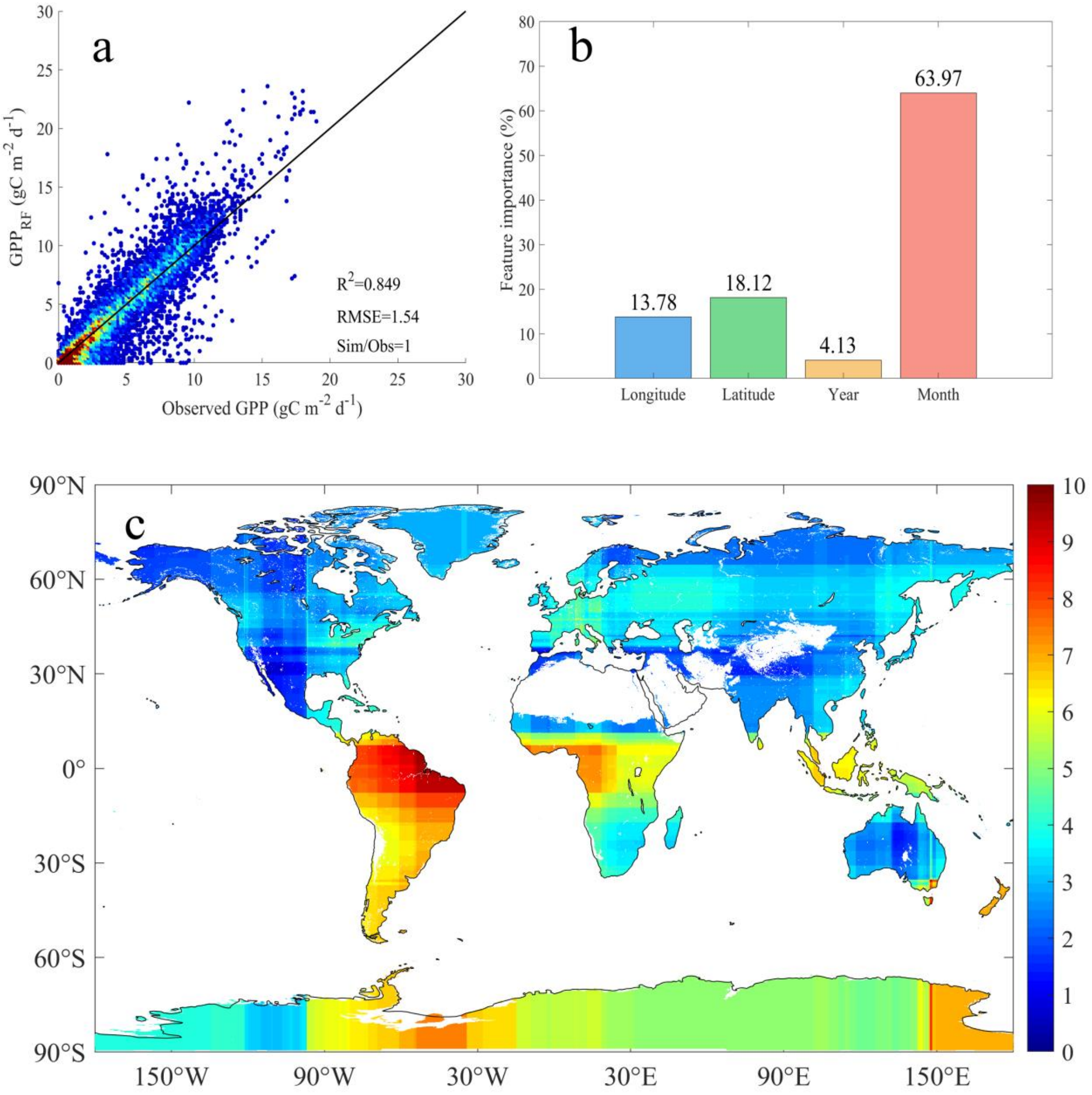

Figure R1. Simulation performance of a random forest model with only longitude, latitude, year, and month. a represents the result of the 5-fold-cross-validation, b represents the relative importance of the random forest model, and c is the multi-year mean estimated by the model for 2001-2022.

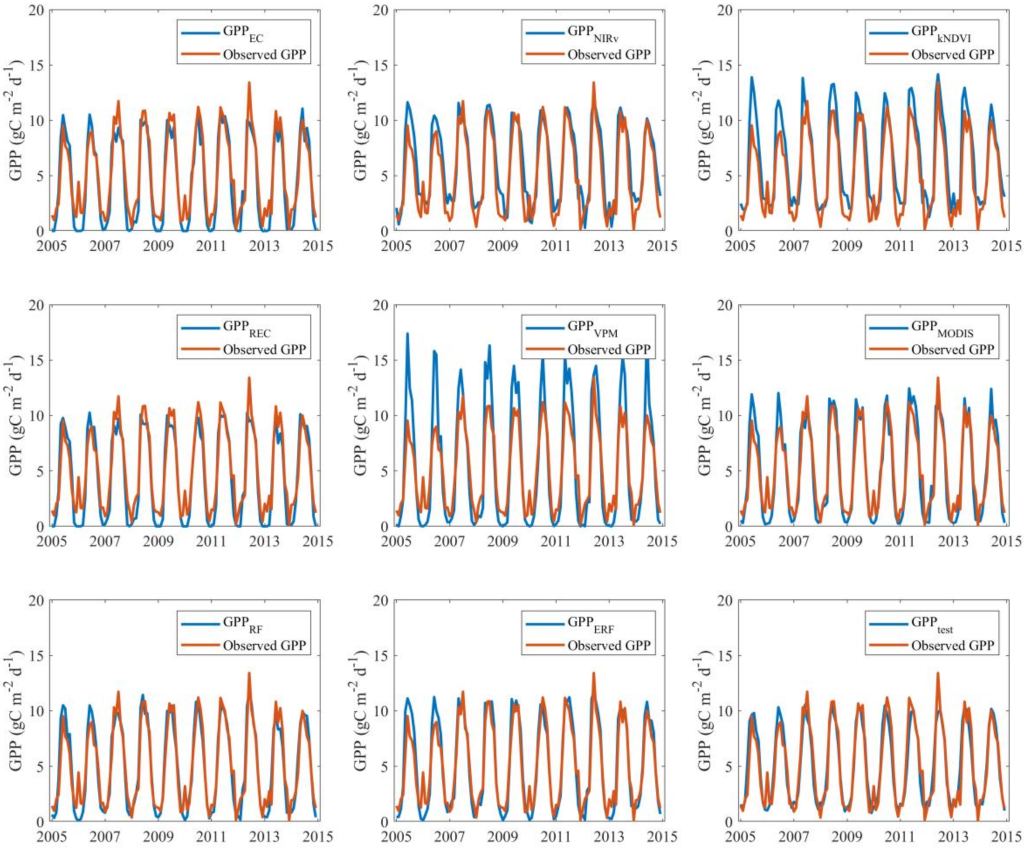

Figure R2. Performance of each GPP model on CN-Qia. The last one is a simple random forest model with only longitude, latitude, year and month.

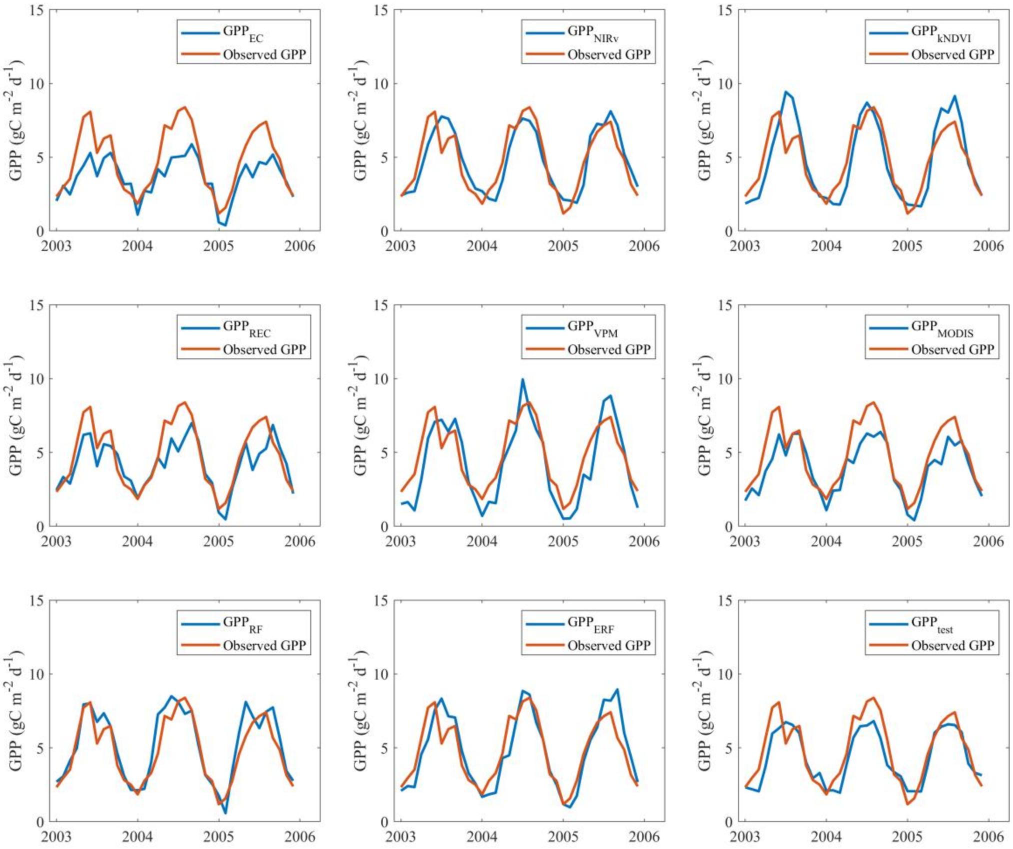

Figure R3. The performance of each GPP model on CH_Lae. The last one is a simple random forest model with only longitude, latitude, year and month.

- Use all the data from each site exclusively for either training (~70%) or validation (30%), instead of employing random splits. Then repeat the model creation and check model performances only on the 30% of validation subset.

REPLY: Thanks for your comments. According to your suggestion, we only selected 70% of the data for training and reserve the remaining 30% for validation. This was repeated 200 times, as shown in Figure R4, and the result was very similar to the result of the 5-fold-cross-validation, with GPPERF still maintaining the highest accuracy. In the revision, we have added this validation method:

In addition, we used a second validation method where 70% of the data was selected for modeling and only the remaining 30% was validated, a process that was repeated 200 times.

Combining the results of all flux sites, GPP$_{ERF}$ explained 85.1% of the monthly GPP variations, while the seven GPP models only explained 67.7%-81.5% of the monthly GPP variations (Figure 2). Another validation method also showed similar results (Figure S3).

As for the method for evaluating the ERF model, we used the 5-fold-cross-validation, which was also used in the validation of the FLUXCOM dataset (Tramontana et al., 2016). Here we reinterpret the 5-fold-cross-validation, we divide all the data into five pieces, select four of them (80%) for modeling, then validate the remaining one (20%), and repeat this five times to get the complete validation set. In fact, the method is similar to the one you mentioned, except that we do a loop to get the full validation result, and the above method only retains 30%. Therefore, it is inevitable that the validation results of the two methods are similar.

Tramontana G, Jung M, Schwalm C R, et al. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms[J]. Biogeosciences, 2016, 13(14): 4291-4313.

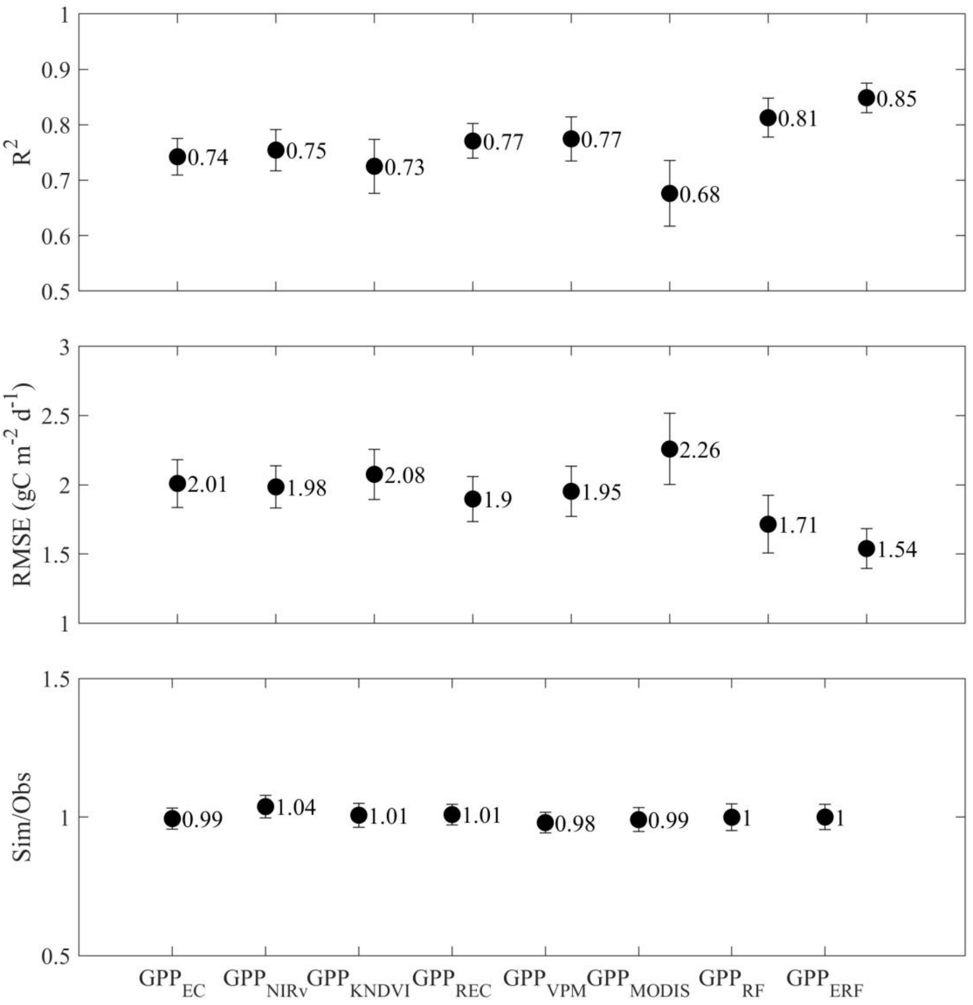

Figure R4. Validation results for each model on 30% validation sets. The black dots represent the mean of the 200 validation results, and the upper and lower boundaries represent five times the standard deviation.

Additionally, I appreciate the authors' inclusion of FLUXCOM for comparison, but it remains unclear which FLUXCOM product was used. Was it the RF model (which I think would be most comparable) or an ensemble of several machine learning approaches? Maybe show both? I also wonder why FLUXCOM performs so good for CHINAFLUX (Fig. R4) but much worse for FLUXNET (Fig. R5)?

REPLY: Thanks for your comments. In the last revision, we used an ensemble version. in the new version, we supplemented the Random forest-based dataset, and surprisingly, in the validation, we found that the random forest-based dataset (FLUXCOM-RF) performed better than the ensemble version (FLUXCOM-ENS) (Figures R5 and R6). This may be because the ensemble version only uses simple multi-model averaging and does not get good results. Of course, even with the more accurate FLUXCOM-RF, ERF_GPP is comparable. In the revision, we compared both FLUXCOM datasets.

On the second question, you can actually see that with the exception of NIRv, the accuracy of the other products in CHINAFLUX is higher than that of FLUXNET, so the difference is obviously independent of the model structure. We think this may be because in CHINAFLUX, the input data sets (some remote sensing indicators such as LAI) of these models are strongly correlated with GPP. In contrast, in FLUXNET, the relationship between LAI and GPP is much weaker (Hu et al, 2022). Of course, this is just our guess, and it's also an interesting question that could be studied further in the future.

Hu, Z., Piao, S., Knapp, A. K., Wang, X., Peng, S., Yuan, W., ... & Yu, G. (2022). Decoupling of greenness and gross primary productivity as aridity decreases. Remote Sensing of Environment, 279, 113120.

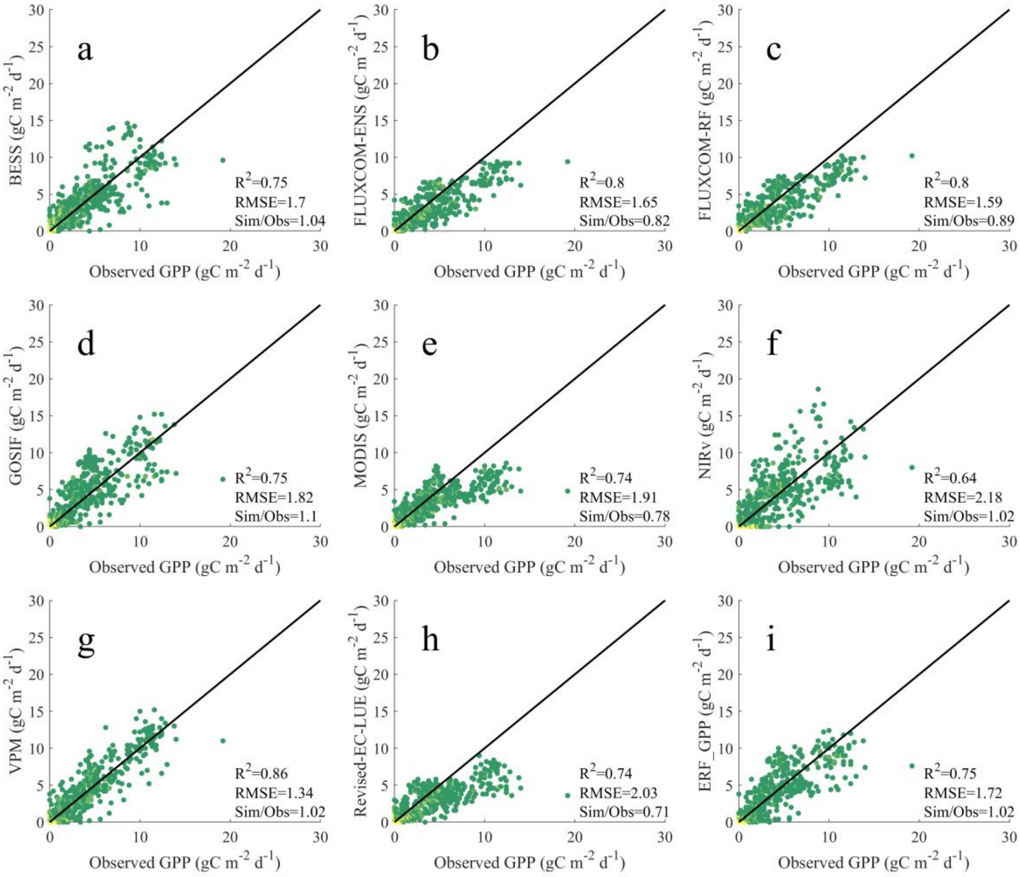

Figure R5. Comparison between the GPP datasets and the GPP observations from ChinaFlux. a-i represents BESS, FLUXCOM-ENS, FLUXCOM-RF, GOSIF, MODIS,

NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

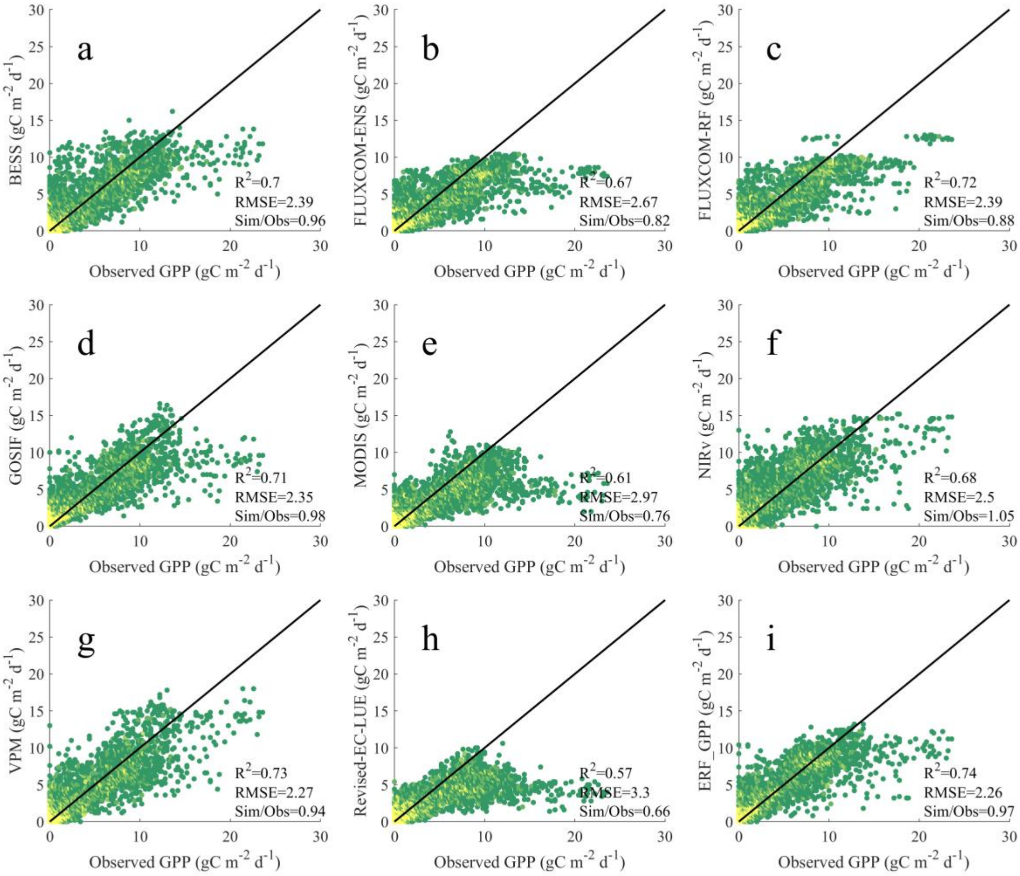

Figure R6. Comparison between the GPP datasets and the GPP observations from FLUXNET. a-i represents BESS, FLUXCOM-ENS, FLUXCOM-RF, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

Overall, there are still areas lacking in clarity. For example, in Table R1 "GPP number" should be revised to e.g. "number of GPP models" or "number of GPP products". Which models were added in which step? Fig. R2/R3 have no x-label.

REPLY: Thanks for your comments. Sorry that there are still some errors, we have further checked the full text. Table R1 is added to further illustrate the robustness of the ERF model, which is explained in the main text.

In addition, we tested the effect of the number of GPP models on the accuracy of the ERF model. As shown in Table S8, as the number of GPP in the ERF model increased, the performance gain of the model gradually decreased.