In their study the authors created two new datasets of gross primary productivity (GPP), one based on remote sensing and environmental predictors and one an ensemble of four existing GPP models. Both models connect predictors and observed GPP using Random Forests. To test the practicality of their approach, the authors compared their two products and the four existing models to FLUXNET site observations. Additionally, they created a global gridded GPP estimate using the ensemble-based approach and performed an independent evaluation using site observations from FluxChina. Improving estimates of global GPP is indeed an important scientific challenge. However, while the reported model metrics suggest a substantial improvement in particular for their ensemble-based model compared to existing models, I am not convinced of the novelty and whether there is indeed a real improvement. My main concerns are the following:

The methodology behind the model evaluation is unclear. It seems that all models "saw" the full FLUXNET data during parameter calibration and then final model evaluation (Fig. 1-4) was computed based on the full dataset? Model evaluation should be done on a separate test dataset. If no separate test dataset existed, the ensemble approach might just have learned the typical GPP values of this site and its fluctuations from the patterns in the four other models. There is an independent evaluation included in the paper which does not suffer from this issue (ChinaFlux), however, only 12 sites are included and other existing models show comparable prediction skills.

REPLY: Thanks for your comments. As you said, all models used FLUXNET data set in parameter calibration, but it should be noted that only 70% samples were selected in our parameter calibration each time, and the average value of 200 calibrated parameters was used as the final parameter, so as to avoid obtaining a parameter applicable to the complete FLUXNET. This is also a common practice for model parameter calibration (Badgley et al. 2019, Zheng et al. 2020). The practice of some previous studies was to use 70% of the sample for calibration and the remaining 30% for validation. But they only do this once, or choose the best many times (e.g. Wang et al. 2020), and it is entirely possible to get an accidental parameter that only applies to this one validation set.

Badgley, G., Anderegg, L. D., Berry, J. A., and Field, C. B.: Terrestrial gross primary production: Using NIRV to scale from site to globe, Global change biology, 25, 3731-3740, 2019.

Zheng, Y., Shen, R., Wang, Y., Li, X., Liu, S., Liang, S., Chen, J. M., Ju, W., Zhang, L., and Yuan, W.: Improved estimate of global gross primary production for reproducing its long-term variation, 1982–2017, Earth System Science Data, 12, 2725-2746, 2020.

Wang, S., Zhang, Y., Ju, W., Qiu, B., and Zhang, Z.: Tracking the seasonal and inter-annual variations of global gross primary production during last four decades using satellite near-infrared reflectance data, Science of the Total Environment, 755, 142569, 2021.

Secondly, the main purpose of our parameter calibration is to reduce the impact of the uncertainty of the model parameters on the validation results. The original parameters of these models were calibrated with only a small number of sites (e.g., 95 sites were

used for Revised EC-LUE and 104 for NIRv). Therefore, when we used the original parameters, the results validated by 170 sites (sorry, The 171 sites in the original text are typographical errors) in this study contain **not only the uncertainty of the model structure, but also the uncertainty of the model parameters.** In the revised version, we explain this in detail:

FLUXNET only provides GPP observation and meteorological data, while LAI, FPAR and surface reflectance are not provided, so only remote sensing data can be used. However, there are many sources of remote sensing data, such as MODIS, AVHRR, etc., so using different remote sensing data to calibrate the same GPP model may produce different model parameters. In addition, the number of sites used to calibrate model parameters is also an important influencing factor for model parameters. The original parameters of these models were calibrated with only a small number of sites (e.g., 95 sites were used for Revised EC-LUE and 104 for NIRv). Therefore, to reduce the impact of the uncertainty of the model parameters on simulation results, we did not use original parameters in the model, but carried out parameter calibration and for GPP models according to different vegetation types.

For the ensemble model, we used “5-fold cross-validation” method, which is the most common method for machine learning validation. That is to say, we divide all samples into 5 parts, select 4 of them for modeling each time, and validate the rest once, so that the cycle is repeated five times to obtain the complete validation result. These validation sets are independent, so the validation results are reliable.

To further dispel your doubts, we used the original parameters of these models for validation and the construction and validation of ensemble model. The author of kNDVI did not provide model parameters, so this model was abandoned. In addition, reviewer 2 suggested that MODIS and VPM be added. Therefore, the validation of 5 GPP models and the ensemble model built based on these 5 GPP models are shown in Figure R1, in the GPP simulation using the original parameters, the performance of these GPP models was significantly decreased, R^2 ranged from 0.570 to 0.719, RMSE ranged from 2.29 to 3.81 gC m^{-2} d^{-1}, in addition, the phenomenon of "high value underestimation and low value overestimation" was also serious. However, the ensemble model exhibited consistent advantages, with R^2 significantly higher than other GPP models (0.856). As just mentioned, these GPP models contain uncertainties in model parameters and model structure, which makes them perform poorly, and the excellence of the ensemble model also proves the reliability of the results of this study.

In the revised version, we have added the results of this section:

In order to further prove the robustness of the ERF model, we also used GPP models with original parameters for modeling and validation. As shown in Figure S3, the performance of these GPP models decreased significantly, with R^2 ranging from 0.570 to 0.719 and RMSE ranging from 2.29 to 3.81 gC m^{-2} d^{-1}. The phenomenon of "high underestimation and low overestimation" was also serious. However, the ERF model showed a consistent advantage, with R^2 significantly higher than other GPP models (0.856).

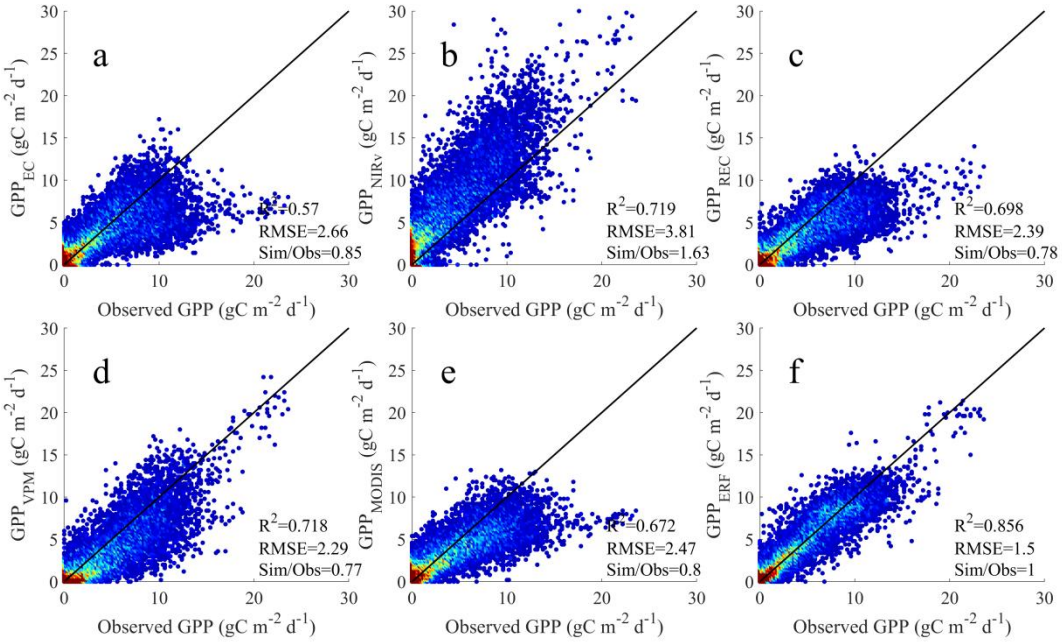

Figure R1. Comparison between the GPP simulations of the six models and the GPP observations. a-f represents GPP_{EC}, GPP_{NIRv}, GPP_{REC}, GPP_{VPM}, GPP_{MODIS}, GPP_{ERF}, respectively. The author of kNDVI did not provide model parameters, so this model was abandoned.

In ChinaFlux's GPP validation, we did not validate these GPP models, but rather the published GPP data set and the results of the ensemble model in this study (at 0.05° grid). This is mainly due to the absence of meteorological data at some sites, which made it impossible for us to obtain the GPP simulation of all models at the site scale (500 m). In the revised version, we explain this in detail.

It should be noted that due to the absence of meteorological data from some sites in Chinaflux, we did not validate all GPP models at the site scale (500 m).

Even for the evaluation performed on a separate test dataset (i.e. ChinaFlux), I wonder whether the good prediction skill of GPPERF is mostly a result of spatial autocorrelation, i.e. by learning the patterns from the four GPP products RFERF basically finds the correct region and predicts the GPP values of the nearest FLUXNET site?

REPLY: Thanks for your comments. As mentioned above, ChinaFlux's site was not involved in the validation of simulation results of all GPP models, but is used to validation results of other GPP datasets and ensemble model at grid (0.05°). The good performance of GPP_{ERF} is not actually a spatial improvement, nor is it the result of spatial autocorrelation, because these GPP observations are a collection of different sites over the months, that is, high values actually indicate the GPP of the growing season, and low values indicate the non-growing season. Therefore, the improvement here is actually the simulation on the time series, which is to improve the phenomenon of "high value underestimation and low value overestimation" emphasized in this study. As shown in Figure R2 and R3, we show the simulation

results of each model at the two sites. It is obvious that GPP_{EC}, GPP_{REC} and GPP_{MODIS} on CN-Qia showed obvious underestimation during the growing season. On CH_Lae, GPP_{kNDVI} and GPP_{VPM} were significantly overestimated. In contrast, at both sites, GPP_{ERF} is more consistent with observations, meaning that the good performance of GPP_{ERF} is due to the correction on the time series (although it is not well corrected at all sites). The performance of each model is different at different sites, mainly because the process concerned by each model (meteorological constraints) is different. For example, NIRv and kNDVI do not use constraints in the modeling process, while other models add some constraints such as temperature. In the revised version, we have added the results of this section:

Further presentations were made at two typical sites, it was obvious that GPP_{EC}, GPP_{REC} and GPP_{MODIS} on CN-Qia showed obvious underestimation during the growing season (Figure S4). On CH_Lae, GPP_{kNDVI} and GPP_{VPM} were significantly overestimated (Figure S5). In contrast, at both sites, GPP_{ERF} was more consistent with observations, meaning that the good performance of GPP_{ERF} was due to the correction on the time series (although it was not well corrected at all sites).

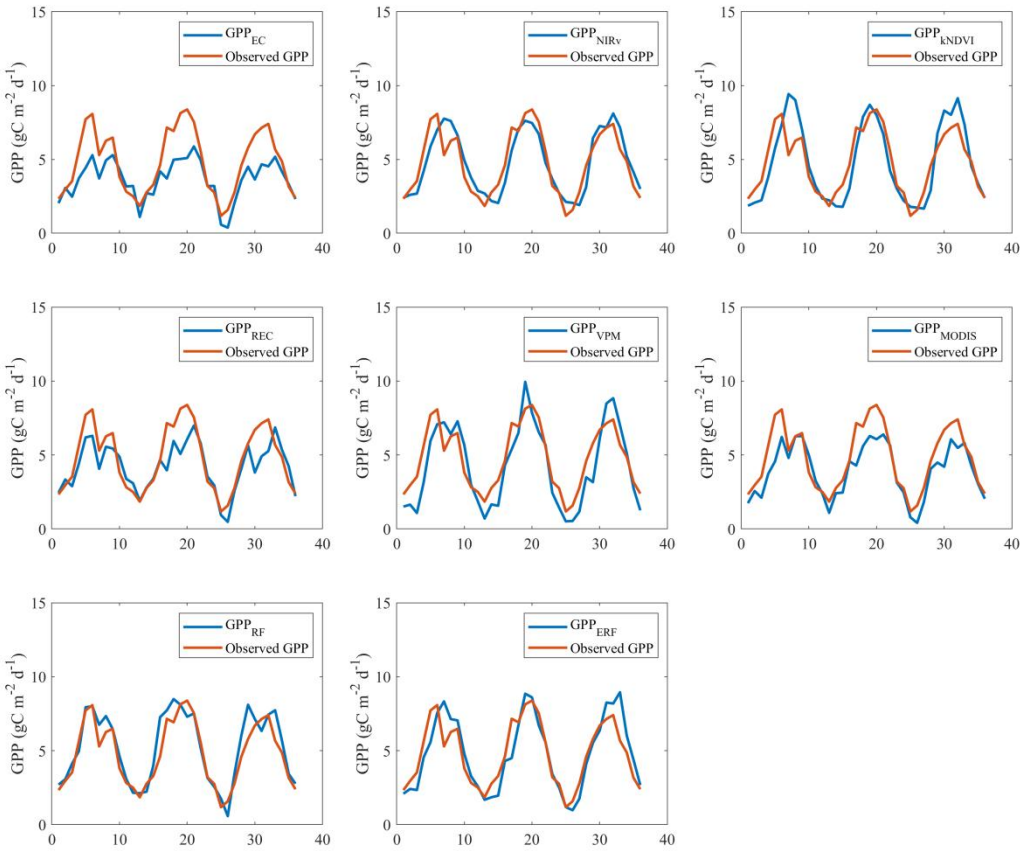

Figure R2. Performance of each GPP model on CN-Qia.

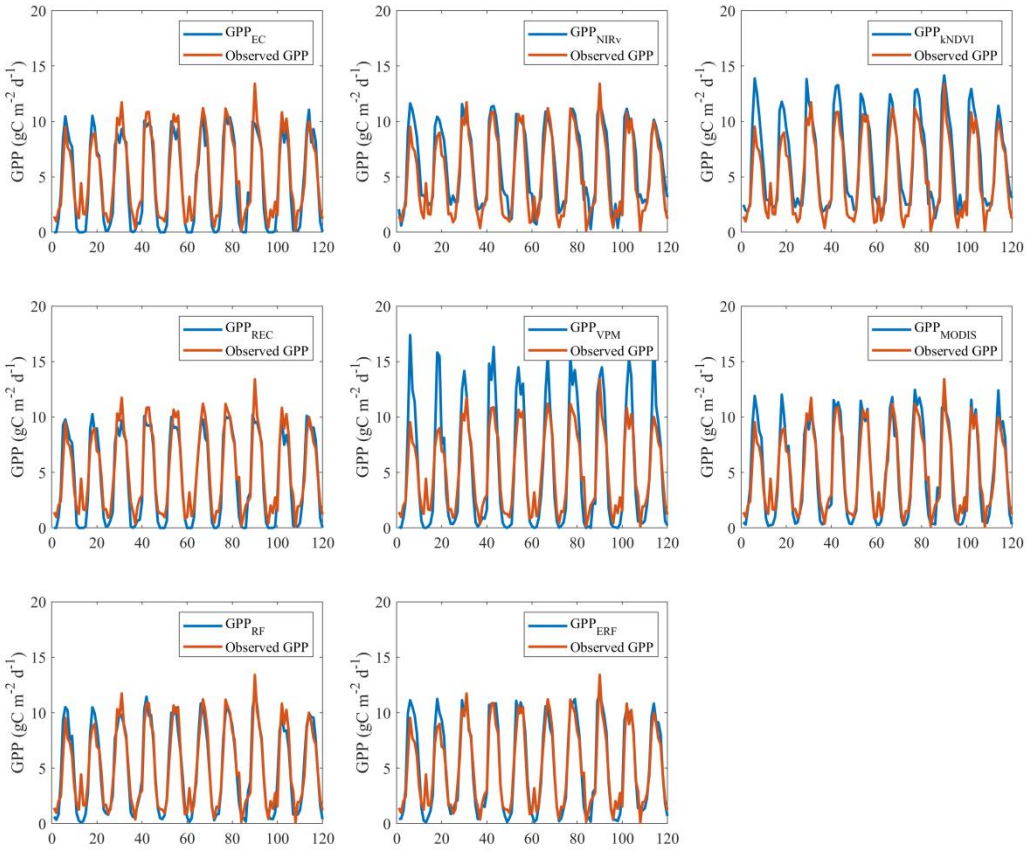

Figure R3. The performance of each GPP model on CH_Lae.

The authors' remote sensing and environmental predictors model seems to be similar to the FLUXCOM approach. I wonder what is the advantage and why FLUXCOM is not included in the comparison?

REPLY: Thanks for your comments. As you said, the RF model we used is actually one of the GPP estimation models in FLUXCOM. The purpose of using this model is to compare it with the ensemble model, nothing more, because both use the random forest method, then the difference in validation results is mainly due to the influence of the input data set (GPP_{RF}: remote sensing and environmental variables; GPP_{ERF}: GPP simulation). We did not use the FLUXCOM dataset when comparing the results of the ensemble model with other GPP datasets, mainly because it did not provide data with 0.05° resolution, and only one set of 0.5° data fit the Chinaflux validation set time range (2001-2018). Based on your comments, we have added a comparison of FLUXCOM in the revised version. We also added validation of the VPM and MODIS GPP datasets. As shown in Figure R4, in Chinaflux's validation, the accuracy of FLUXCOM is reliable, but it shows a certain underestimation. In contrast, VPM showed better performance, which may be related to their preprocessing of the input data. In the revised version, we have added a description of the relevant content in the results section:

As shown in Figure 6, ERF_GPP and other GPP datasets were validated using GPP observations from ChinaFlux. Of all the models, GPP_{VPM} has the best performance, with R^2 of 0.86 and RMSE of gC m^{-2} d^{-1}. ERF_GPP also had a high generalization, R^2 of 0.75, RMSE of 1.72 gC m^{-2} d^{-1}, there was no "high value underestimation and low value overestimation", which was comparable to the simulation accuracy of BESS and GOSIF. However, the simulation accuracy of the other GPP datasets in Chinaflux was relatively poor, with the R^2 of NIRv being only 0.64, while FLUXCOM, MODIS and Revised EC-LUE was significantly underestimated, with the Sim/Obs being only 0.71-0.82.

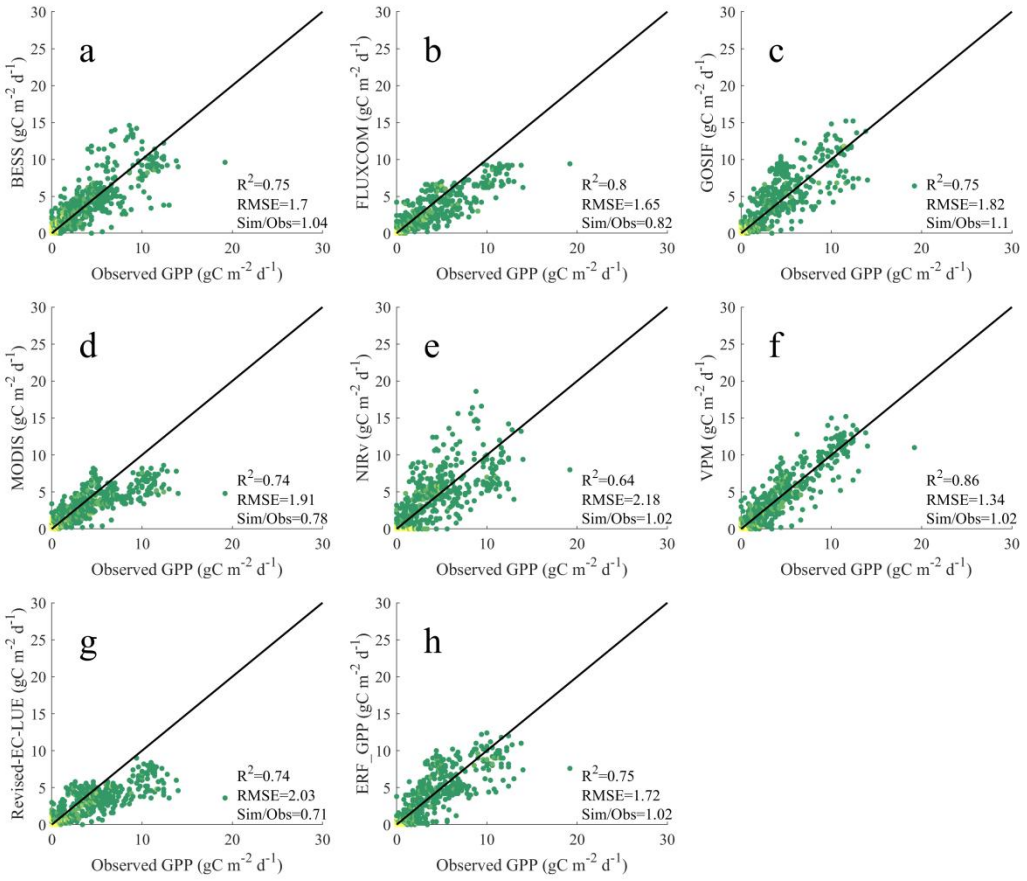

Figure R4. Comparison between the GPP datasets and the GPP observations from ChinaFlux. a-h represents BESS, FLUXCOM, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

The authors recalibrated the parameters underlying the four existing models but the justification for this action is unclear. I would like to see a comparison with the original models to see whether this indeed led to improvements in model performance.
REPLY: Thanks for your comments. As stated in the first point, the calibration parameters are used to compare the performance differences between models considering only the uncertainty of the model structure. For GPP simulation of original parameters, the ensemble model also showed superior performance (Figure R1).

Several existing GPP datasets are only shown in the comparison to ChinaFlux but were not included in the ensemble-based product. Vice versa, two of the models used in the FLUXNET comparison were omitted from the ChinaFlux comparison. I wonder why the authors selected these four models (EC-LUE, Revised-EC-LUE, GPP-kNDVI, GPP-NIRv) in the ensemble approach even though the comparison in Fig. 6 suggests other products perform much better? If the reason is the spatial resolution this should be better explained.

REPLY: Thanks for your comments. In the first point, we explain why GPP models and ensemble model are not validated using ChinaFLUX. In response to your comments, we have added the results of validation of GPP datasets and ensemble models using FLUXNET data in the revised version. Similarly, we extracted 0.05° MODIS land use covering the flux tower and used the site for analysis when the vegetation types of the flux tower were consistent with MODIS land use. In the end, 52 sites from FLUXNET were used. As shown in Figure R5, the validation results of the ensemble model are significantly better than those of other GPP datasets. However, underestimation is shown in the high value, which may be due to the inconsistency between the 0.05° coarse resolution and the flux tower footprint. In the revised version, we have added a description of the relevant content in the results section:

In the validation of FLUXNET, the R^2 of FLUXCOM, MODIS, and Revised EC-LUE ranged from 0.57 to 0.67, and the RMSE ranged from 2.67 to 3.3 gC m^{-2} d^{-1}, and showed different degrees of underestimation (Figure S8). Other GPP datasets showed similar performance, with ERF_GPP being the best (R^2=0.74, RMSE = 2.26 gC m^{-2} d^{-1}). Notably, in the high values, all models exhibited significant underestimation, which may be caused by the 0.05° resolution being inconsistent with the flux tower footprint.

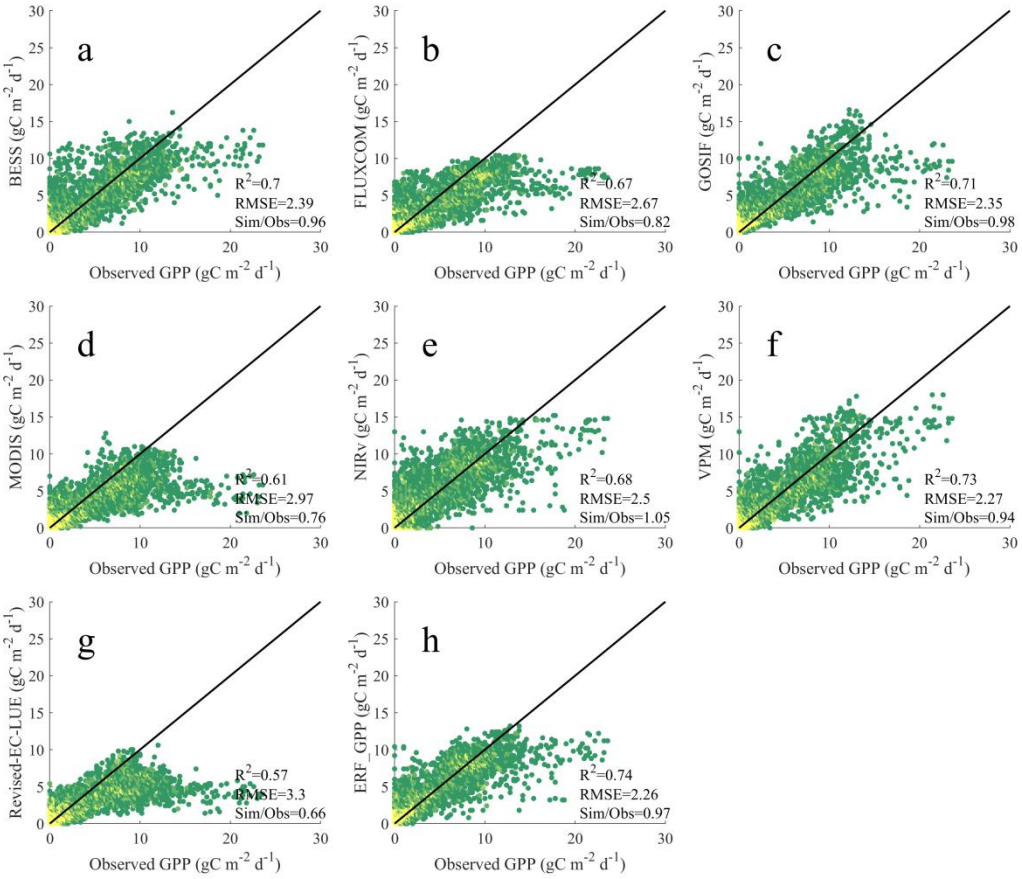

Figure R5. Comparison between the GPP datasets and the GPP observations from FLUXNET. a-h represents BESS, FLUXCOM, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

For the four (now is six) GPP models selected in the ensemble model, this is justified and sorry not to be mentioned in the original article. The GPP models mainly include process model, light use efficiency model, vegetation index model and machine learning model. The process model is very complex, many parameters are considered, and the accuracy of the models is not very outstanding, although they are more suitable for the process of photosynthesis. We expect the ensemble model to improve the performance of the model without being too complex, so we mainly chose a few representative models that are widely used. In the revised version, we explain this in detail. At the same time, at the suggestion of reviewer 2, we also added VPM and MODIS in the revised version. In other words, there are 6 GPP models in the ensemble model in the revised version.

In this study, six independent models were selected to estimate GPP. These models are widely used with few model parameters and have shown reliable model accuracy in previous studies.

In addition, we added a section on the effect of the amount of GPP on the accuracy of the ensemble model. As shown in Table R1, as the number of GPP in the ensemble model increases, the model performance gains gradually decrease.

Table R1. Effect of the GPP number in the ERF model on model performance

GPP number	2	3	4	5
R^2	0.793±0.024	0.824±0.011	0.836±0.004	0.845±0.001
RMSE	1.798±0.104	1.658±0.052	1.600±0.022	1.556±0.009
Sim/Obs	1±0.001	0.999±0.000	1±0.000	1±0.000

Minor comments:

L18: Remove "a".

REPLY: Thanks for your comments. We have corrected this error in the revised version.

L33: I think you mean "to the terrestrial carbon cycle".

REPLY: Thanks for your comments. This sentence has since been modified to Gross primary productivity (GPP) is the largest carbon flux in the global carbon cycle, and it is also the input of carbon into the terrestrial carbon cycle.

L38: Unclear, is this about remote sensing-based estimates or GPP estimates in general? Also it is unclear how the approach applied in this study helps with the problems mentioned in the following sentences. Overall the introduction lacks connectivity.

REPLY: Thanks for your comments. This refers to the models that involve remote sensing data in the estimation of GPP.

In this paragraph, we want to emphasize the problems existing in these GPP models. Of course, many of the problems mentioned have not been solved in our research. Therefore, we have sorted out this paragraph again, focused on the uncertainty of several GPP models, and introduced the ensemble model.

The light use efficiency (LUE) model is one of the most widely used models for estimating GPP. It assumes that GPP is proportional to the photosynthetically active radiation absorbed by vegetation, and optimizes the spatio-temporal pattern of GPP through meteorological constraints such as temperature and water (Pei et al., 2022). However, the form of these meteorological constraints varies greatly, and this difference alone can lead to a difference of more than 10% in the explanatory power of the models (Yuan et al., 2014). Recent studies have proposed some new vegetation indices that have been shown to be effective proxies for GPP through theoretical derivation and validation by observations (Badgley et al., 2017; Camps-Valls et al., 2021). However, these vegetation indices often use only remote sensing data as an input for estimating long-term GPP without considering meteorological factors, which has led to some controversy (Chen et al., 2024; Dechant et al., 2020). Both the LUE model and the vegetation index model use a combination of linear mathematical formulas to estimate GPP. However, ecosystems are highly complex and the biases introduced into a process by this numerical model increase the uncertainty in the estimates of the final product (GPP). The machine learning model has shown in previous studies that it has great potential to improve GPP estimates (Jung et al., 2020). This model is trained by non-physical means directly using GPP observations and selected environmental and vegetation variables, and the performance of the model depends on the number and quality of the observed data and the

representativeness of the input data. Machine learning has also been widely used in recent years due to its advantages such as the fact that no parameter calibration is required and the reliable model accuracy. Nevertheless, direct validation from flux tower of FLUXNET shows that the model typically explains only about 70% of the monthly variations in GPP, with similar performance to other models (Wang et al., 2021; Badgley et al., 2019; Zheng et al., 2020; Jung et al., 2020). Due to the deviation of the model structure, there is a common problem in these models, that is, the estimation of monthly extreme GPP is poor, and the phenomenon of "high value overestimation and low value overestimation" occurs (Zheng et al., 2020). Especially for extremely high values, which usually occur during the growing season and largely determine the annual value and interannual fluctuations of GPP, this underestimation may hinder our understanding of the entire carbon cycle.

L48: Unclear, do you mean the models assume a positive relationship between CO2 and GPP while it is actually negative? Or that CO2 fertilization started to saturate?

REPLY: Thanks for your comments. As you said, it means that the effect of CO_2 fertilization tends to be saturated, that is, the positive impact of CO_2 fertilization on GPP is weakening. Considering that the ensemble model in this study also did not include this saturation CO_2 fertilization effect, we deleted this sentence to avoid misunderstanding.

L55: Is this for the same region?

REPLY: Thanks for your comments. This is true in some areas where C3 and C4 are grown alternately. This sentence has been deleted due to changes in the introduction.

L73: "low"?

REPLY: Thanks for your comments. We have corrected this error in the revised version.

L85: "ERA". Also references are missing.

REPLY: Thanks for your comments. We have corrected this error and added references in the revised version.

L108: How were they resampled?

REPLY: Thanks for your comments. For higher resolution data, we gridded the dataset to 0.05° by averaging all pixels whose center fell within each 0.05° grid cell for upscaling. For lower resolution data, we used the nearest neighbor resampling to 0.05°. In the revised version, we explained the resampling method in detail.

Finally, for higher resolution data, we gridded the dataset to 0.05° by averaging all pixels whose center fell within each 0.05° grid cell for upscaling. For lower resolution data, we used the nearest neighbor resampling to 0.05°. In addition, MODIS data were aggregated to a monthly scale to ensure spatio-temporal consistency.

L115: Why only 171 sites? Did the other sites not contain any high-quality years?

REPLY: Thanks for your comments. As you said, there are some sites that only have one or two years of data, so it's not unusual to have years without quality data. For example, AU-Lox only has data for 2008-2009, and US-Wi1 only has data for 2003. Therefore, based on the quality screening criteria, only 170 sites were used in the end (sorry, The 171 sites in the original text are typographical errors).

L120: The paper often mentions "remote sensing models" but the atmospheric data is actually from a reanalysis (ERA5) or FLUXNET.

REPLY: Thanks for your comments. There is also a class of models that estimate GPP driven only by climate and land use data, known as dynamic global vegetation models. These models do not involve remote sensing data in the estimation of GPP, so they are fundamentally different from the models mentioned in this study. In the revised edition, we call these models "GPP models", "LUE models," and "Vegetation Index models".

L121: What is "traditional random forest model"? The authors often mix the nature of the data (e.g. remote sensing) and modelling approach (e.g. random forests).

REPLY: Thanks for your comments. The traditional random forest model refers to the model using remote sensing data and environmental factors in previous studies. In the revised version, we redefine this concept. In addition, we define these models uniformly as GPP models under different methods.

L125: Table 1 says EC-LUE also considers CO2.

REPLY: Thanks for your comments. This is a mistake, as the Revised EC-LUE model simply divides the leaves into sunlit and shaded leaves. In the revised version, we have corrected this error.

L127: SIF was not mentioned previously.

REPLY: Thanks for your comments. The SIF here is sun-induced chlorophyll fluorescence, and in the revised version, we have corrected this error.

L129 A brief summary of random forests is needed. Also why did you choose these four predictors? I assume adding more variables would increase model performance.

REPLY: Thanks for your comments. We briefly introduce random forest methods in the revised version.

Random forest is an ensemble learning algorithm that combines the outputs of multiple decision trees to produce a single result, and is commonly used for classification and regression problems. In the regression problem, the output result of each decision tree is a continuous value, and the average of the output results of all decision trees is taken as the final result.

Following your suggestions, we adjusted the input data in the random forest model, including LAI, FPAR, T, TMIN, VPD, DifSR and DirSR, a total of 7 variables. The addition of CO_2 does not make sense because it does not characterize the effect of CO_2 fertilization. In addition, NIRv and kNDVI are not included in the model because these two inputs are proxies for GPP and are converted to GPP using only a linear equation. If these two variables are included, the model is essentially the same as the ensemble model. To further dispel your doubts, we present the results of models incorporating NIRv and kNDVI, but to avoid repetitive results, this part is not presented in the paper.

As shown in Figure R6-R9, the R^2 of the random forest model using 7 variables is 0.815. Although it is slightly better than other GPP models, it still lags behind the ensemble model. In addition, the performance of the model in different months, different vegetation types and different subvalues is also worse than that of the ensemble model. In other words, the result is similar to the original paper.

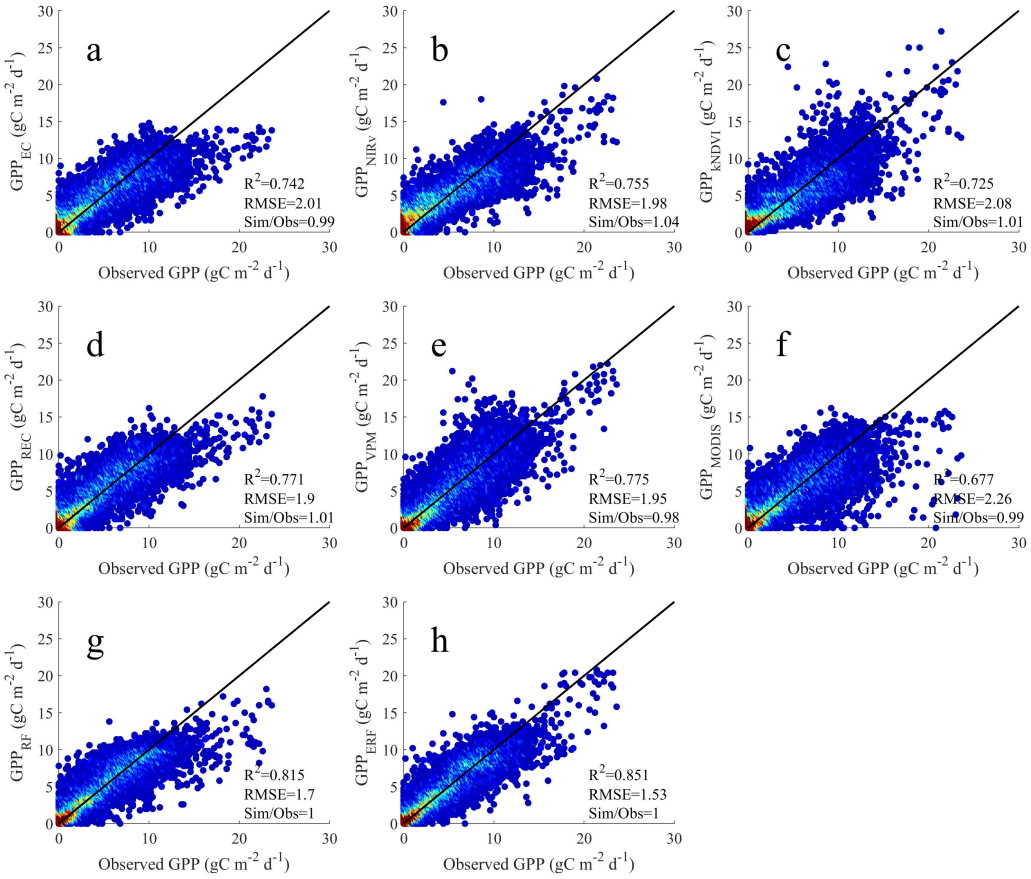

Figure R6. Comparison between the GPP simulations of the eight models and the GPP observations. a-h represents GPP_{EC}, GPP_{NIRv}, GPP_{kNDVI}, GPP_{REC}, GPP_{VPM}, GPP_{MODIS}, GPP_{RF}, GPP_{ERF}, respectively.

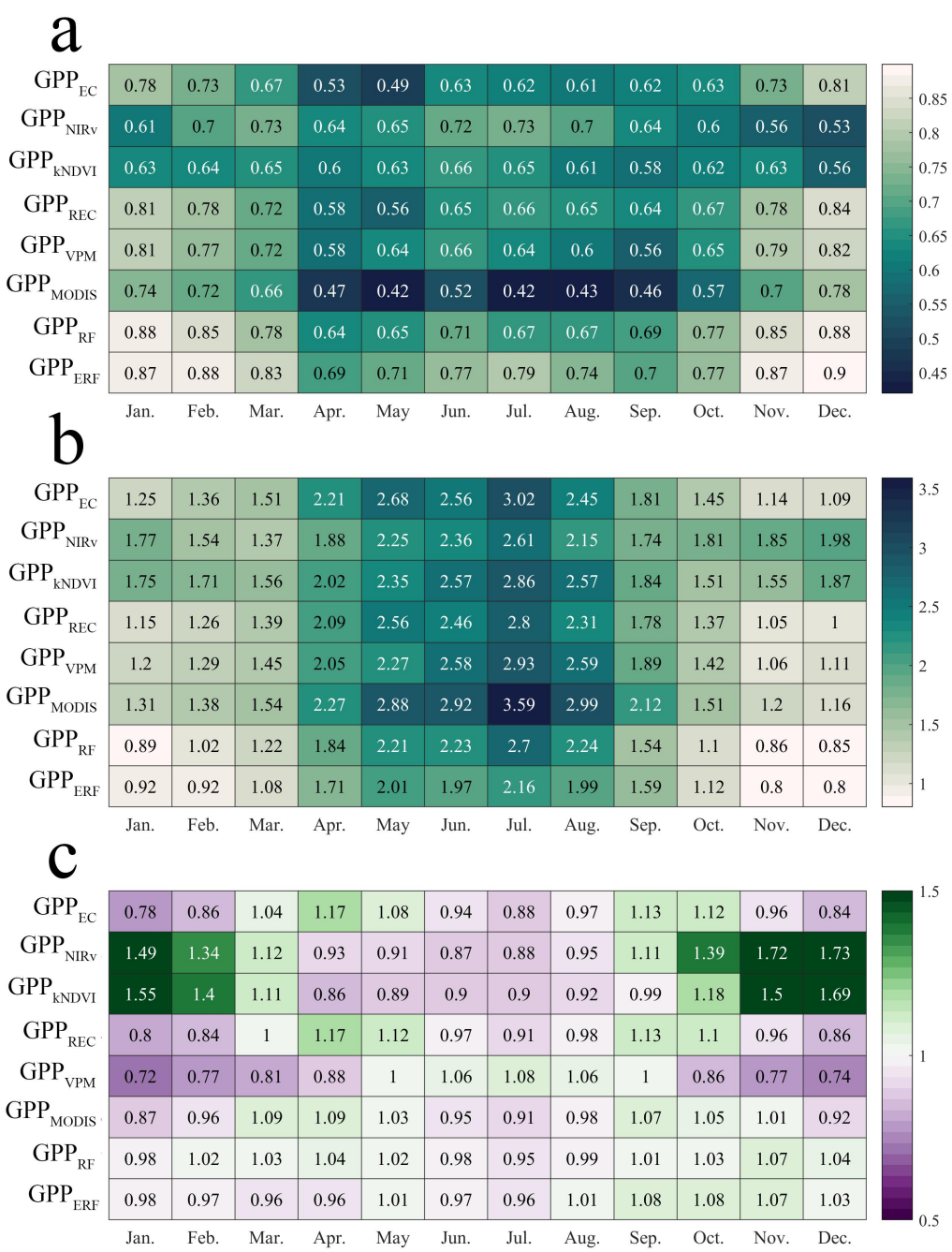

Figure R7. Performance of the eight models in each month. a, b and c represent R^2, RMSE, and Sim/Obs respectively.

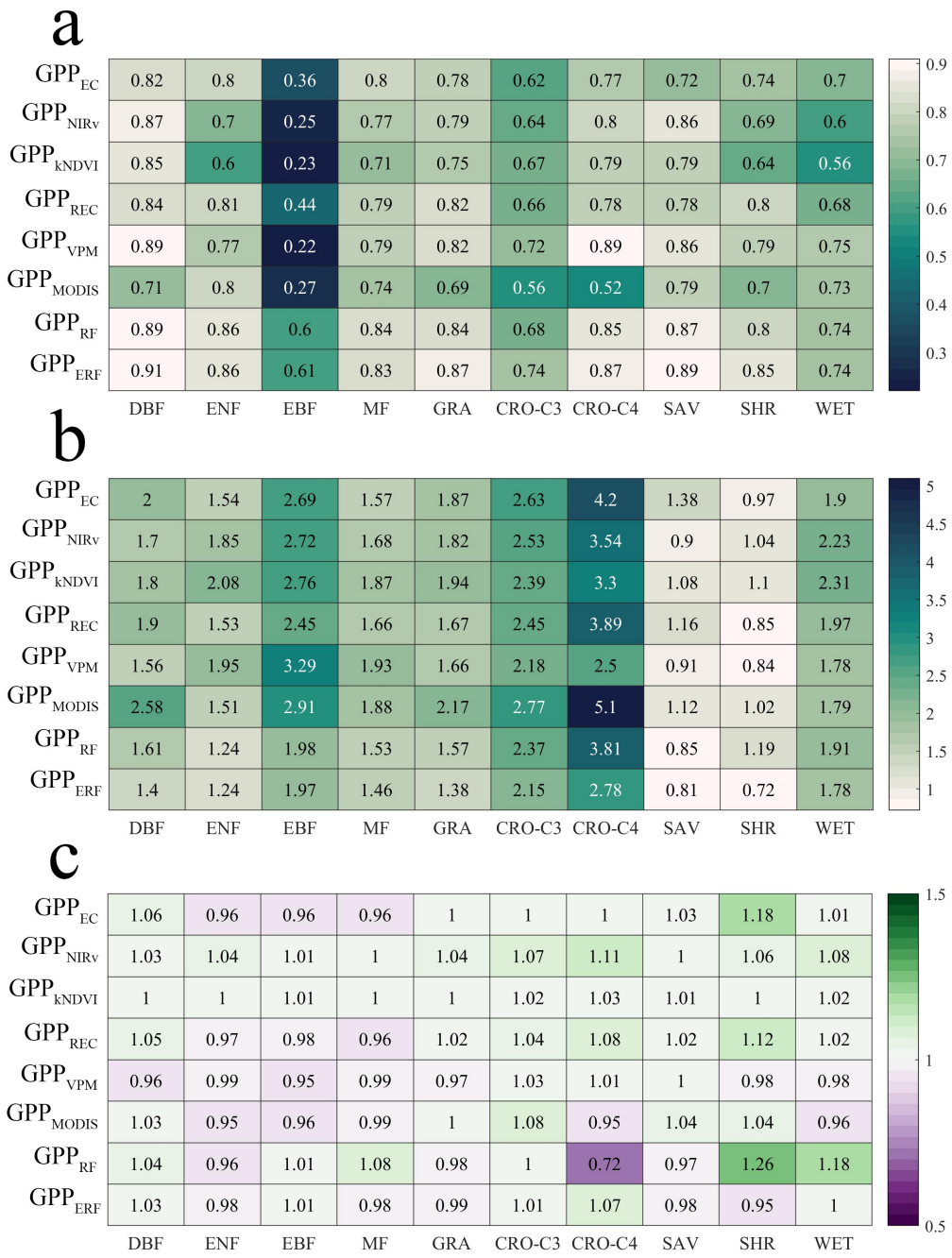

Figure R8. The performance of the eight models on different vegetation types. a, b and c represent R^2, RMSE, and Sim/Obs respectively.

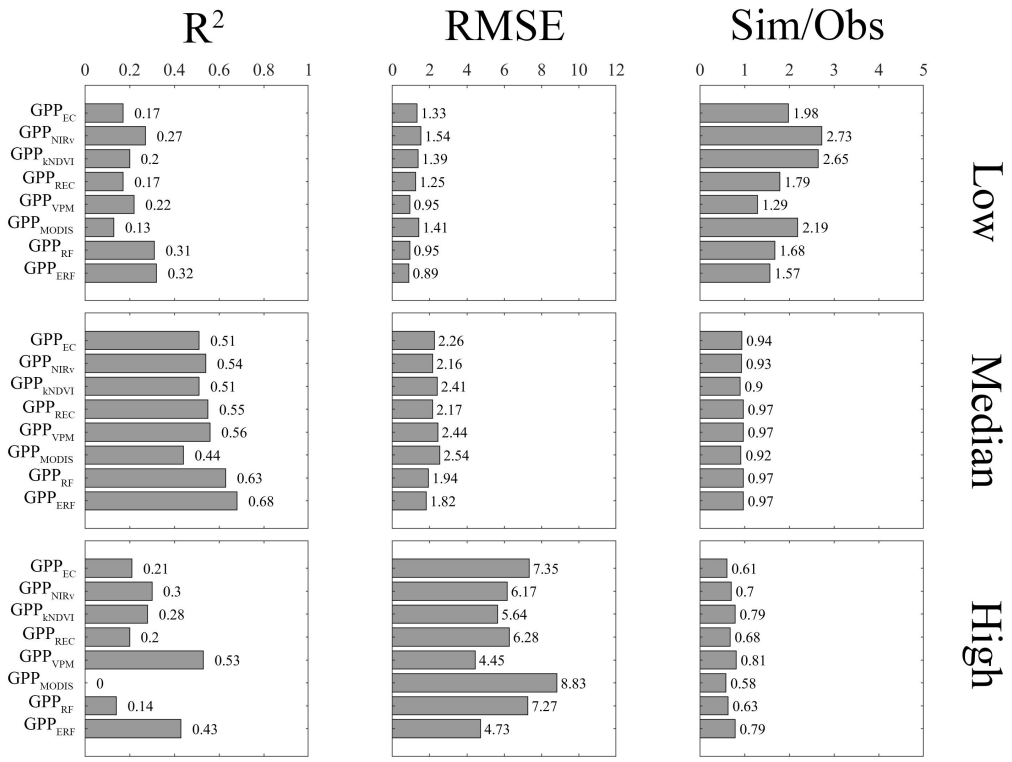

Figure R9. Performance of eight models in different subvalues.

As shown in Figure R10, R^2 of the random forest model using 9 variables is 0.845, which is similar to the performance of the ensemble model, as mentioned earlier, the two models are essentially the same. However, in terms of vegetation type (underestimation of C4 crops, overestimation of SHR and WET), and subvalues (underestimation of high value), the performance of the model also remained gap with that of the ensemble model.

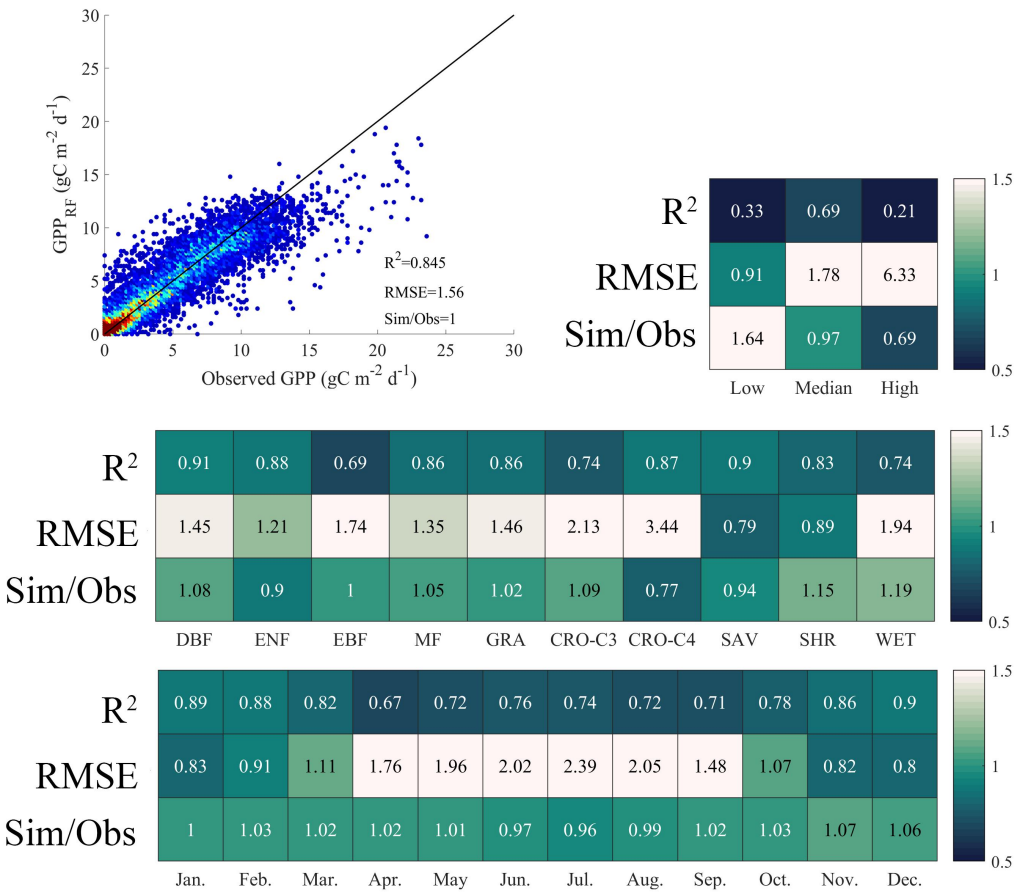

Figure R10. Performance of the random forest model using 9 variables.

L132: "multi-model".

REPLY: Thanks for your comments. We have corrected this error in the revised version.

L137: Provide information about data source. If I understood correctly, e.g. FPAR is from MODIS (500m) while AT from FLUXNET? And ERA5 AT is only used for the global prediction? This is confusing. Also where is the NIR data from?

REPLY: Thanks for your comments. In the revised version, we further clarified the source of the data. FLUXNET not only provided GPP observations, it also provided meteorological data, and ERA5-land was used for global GPP estimates. In addition, the red-band and near-infrared data were also from MODIS.

GPP observations from the FLUXNET 2015 dataset, which includes carbon fluxes and meteorological variables from more than 200 flux sites around the world (Pastorello et al., 2020). GPP cannot be obtained directly from the flux site and usually needs to be obtained by dismantling the Net Ecosystem Exchange. We chose a monthly level GPP based on the nighttime partitioning method and retained only high quality data (NEE_VUT_REF_QC > 0.8) for every year, and finally selected 170 sites with 10932 monthly values for this study. In addition, average air temperature, total solar radiation and VPD on the monthly scale were selected.

Since part of the data required for the model is not directly available at the flux site, LAI and FPAR were extracted from MOD15A2H (500 m), surface reflectance (red band, near infrared band, blue band and shortwave infrared band) are derived from MCD43A4 (500m) and MOD09A1 (500m).

L140: What differences do you mean?

REPLY: Thanks for your comments. FLUXNET only provides GPP observation and meteorological data, while LAI, FPAR and other data are not provided, so only remote sensing data can be used. However, there are many sources of remote sensing data, such as MODIS, AVHRR, etc., so using different remote sensing data to calibrate the same GPP model may produce different model parameters. In addition, the number of sites used to calibrate model parameters is also an important influencing factor for model parameters. Therefore, in the revised version, this sentence has been modified to

FLUXNET only provides GPP observation and meteorological data, while LAI, FPAR and surface reflectance are not provided, so only remote sensing data can be used. However, there are many sources of remote sensing data, such as MODIS, AVHRR, etc., so using different remote sensing data to calibrate the same GPP model may produce different model parameters.

L155: The model overestimates or underestimates.

REPLY: Thanks for your comments. We have corrected this error in the revised version.

F160: How many? Again, references are missing.

REPLY: Thanks for your comments. The flux observations provided by Chinaflux are not consolidated in a single article, so it is still being updated, and it is difficult to specify how many sites are available. For references, we show them in Table S1 because every site has one reference.

L166: Lack of consistency, GPPERF, ERF_GPP or "random forest-based ensemble model"? Or does GPPERF refer to the site predictions while ERF_GPP to the global ones? Again, why are some models thrown out in this step while others are included for the first time?

REPLY: Thanks for your comments. GPP_{ERF} represents the site simulation and ERF_GPP represents the global GPP. Random forest-based ensemble model represents GPP simulation method. In the revised version, we define these.

In this step, we aim to compare ERF_GPP with some of the GPP datasets that are widely used, including GPP datasets generated by other models because these datasets are generated by other methods, such as BESS, which is based on process models, and GOSIF, which is GPP generated by Sun-induced chlorophyll fluorescence. The models used in this study are not all compared in this step, because not all models have relevant data sets, such as kNDVI.

L185: What do you mean by changes in cropland? Do you mean seasonal changes in cropland GPP?

REPLY: Thanks for your comments. As you said, this refers to the seasonal variation of GPP in cropland. In the revised version, this sentence has been modified to

It is worth noting that compared to other vegetation types, the RMSE was highest for cropland, with 6 out of 8 models for C4 crop exceeding 3 gC m^{-2} d^{-1}, suggesting that these existing GPP models may not properly capture the seasonal changes in cropland GPP.

Fig. 2+Fig. S3 Why are the metrics different? Is Fig. S3 the mean of the individual sites while Fig. 2 the mean of all data?

REPLY: Thanks for your comments. As you said, Fig. S3 is the mean of the individual sites, and Fig.2 is all the data. We found that the mean of the individual sites was not very reasonable, which was deleted in the revised version.

L207: "models". This error occurs several times in the manuscript.

REPLY: Thanks for your comments. We have corrected this error in the revised version.

L215: What do you mean by extreme? The highest values (>10 gC/m2/d)? Does this represent 33% of all data?

REPLY: Thanks for your comments. This extreme is actually more of an empirical distinction, such as the high value (>15 gC m^{-2} d^{-1}, redefined in the revised version), which means that the GPP for that month >450 gC m^{-2} $month^{-1}$, no doubt only some sites can achieve the extreme high value. In addition, it can also be found in Figure 2 that some sites have a significant underestimation in the high value, which is also one of the criteria for empirical discrimination. The use of percentages also requires an empirical discrimination, as there is no precedent for validating GPP in extreme.

Fig. S2: Why is there an extra panel for site 1? Why don't you also show the FLUXNET sites?

REPLY: Thanks for your comments. In the revised version, we show all GPP observation sites in Fig.S2.

In general, having a native English speaker review the text would enhance its quality.

REPLY: Thanks for your comments. In the revised version, we have made corrections to the language section.