Comments from the Reviewers:

Reviewer #1 (Formal Review for Authors):

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 observations and meteorological data, lacking direct measurements for LAI, FPAR, and surface reflectance, so only remote sensing data can be used. Considering the variety of remote sensing data sources, such as MODIS and AVHRR, it is evident that calibrating the same GPP model with different remote sensing data can yield varied 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 limited number of sites (e.g., 95 sites for Revised EC-LUE and 104 for NIRv) (Wang et al., 2021; Zheng et al., 2020). Therefore, to reduce the impact of the uncertainty of model parameters on simulation results, we did not use original parameters and conducted parameter calibration for GPP models across 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 pronounced. However, the ERF model maintained a consistent advantage, with R^2 significantly higher than other GPP models (0.856).

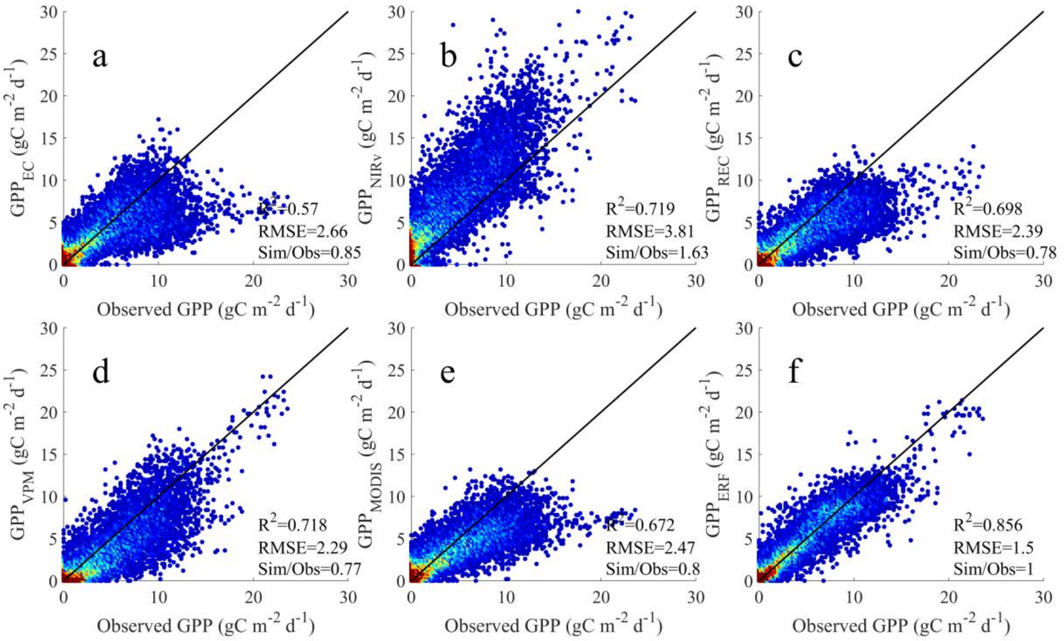


Figure R1. Comparison between the GPP simulations of the six models and the GPP observations. a-f represents GPP_{EC}, GPP_{NIRv}, GPP_{REC}, GPP_{VPM}, GPP_{MODIS}, GPP_{ERF}, respectively. The author of kNDVI did not provide model parameters, so this model was abandoned.

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).

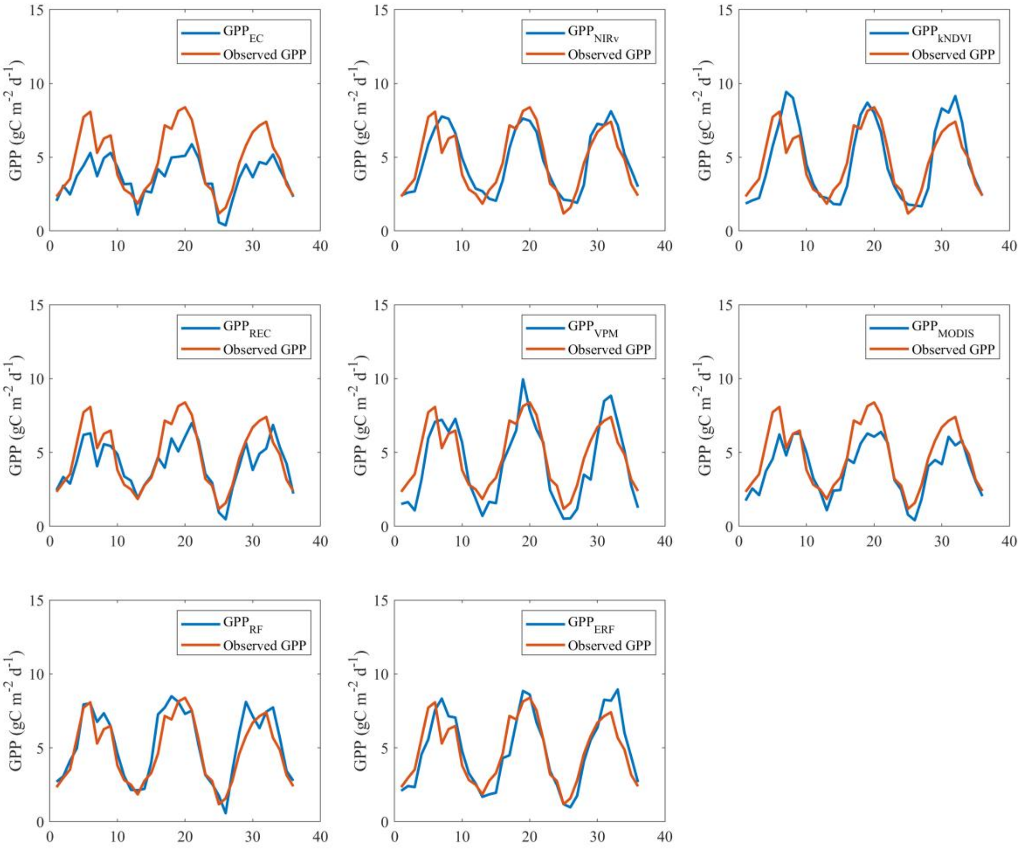


Figure R2. Performance of each GPP model on CN-Qia.

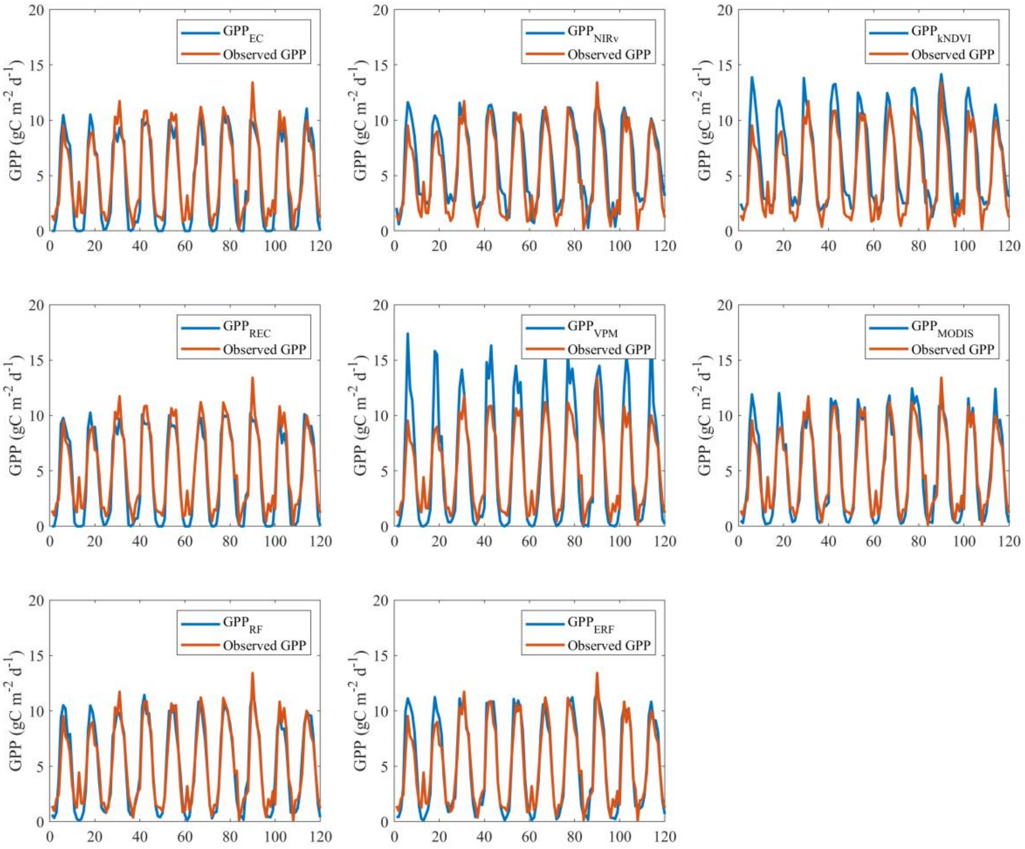


Figure R3. The performance of each GPP model on CH_Lae.

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. Among all the models, GPP_{VPM} demonstrated the best performance, with R^2 of 0.86 and RMSE of 1.34 gC m^{-2} d^{-1}. ERF_GPP also exhibited high generalization, with 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 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 were significantly underestimated, with the Sim/Obs being only 0.71-0.82.

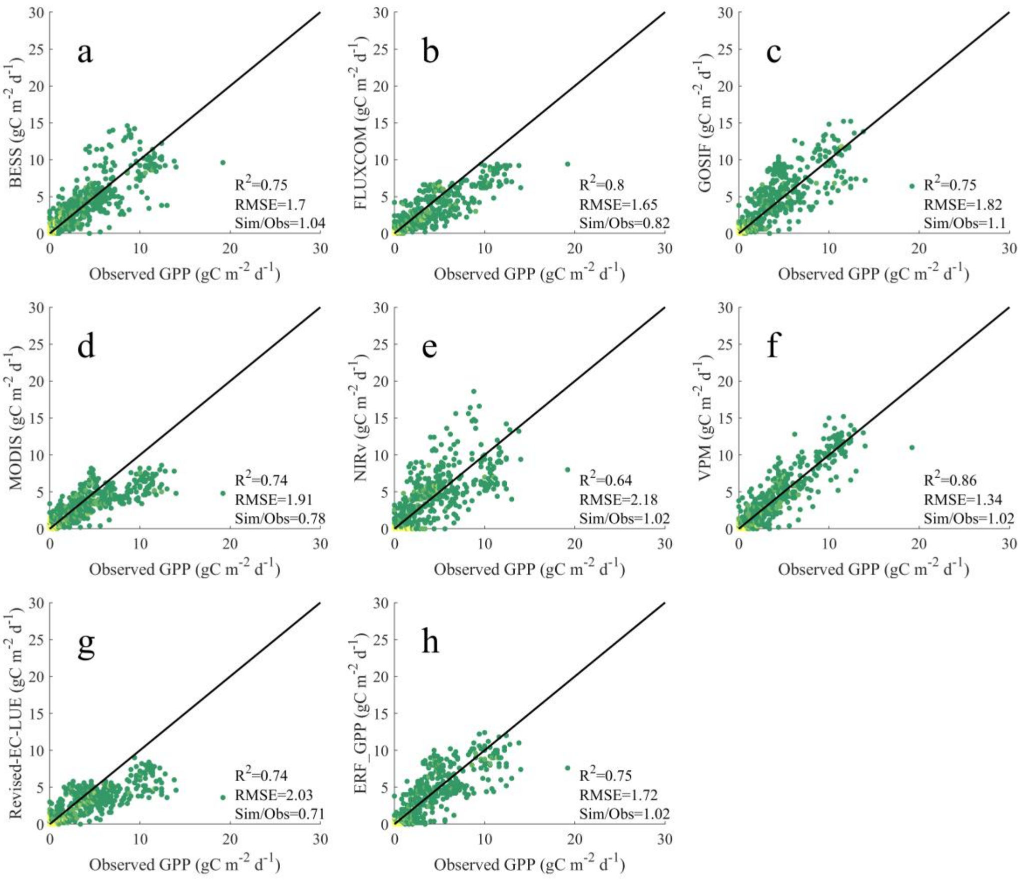


Figure R4. Comparison between the GPP datasets and the GPP observations from ChinaFlux. a-h represents BESS, FLUXCOM, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

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 exhibited different degrees of underestimation (Figure S8). Other GPP datasets demonstrated similar performance, with ERF_GPP being the best (R^2 = 0.74, RMSE = 2.26 gC m^{-2} d^{-1}).

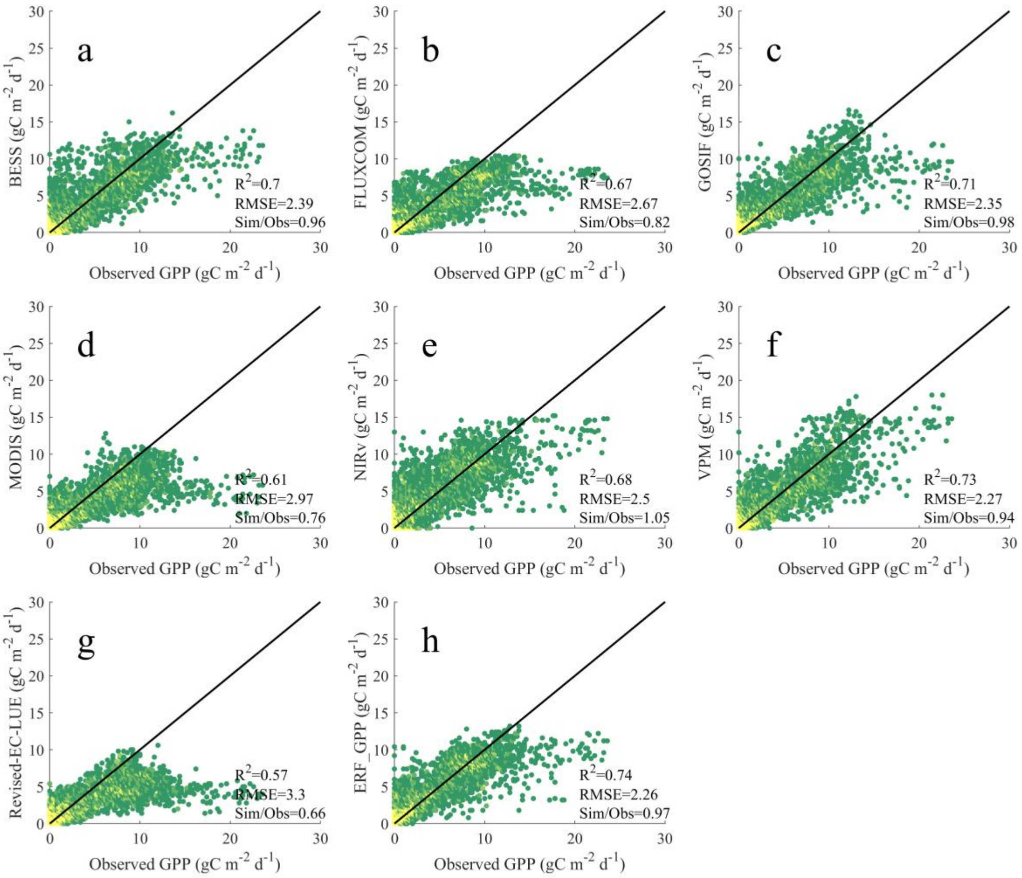


Figure R5. Comparison between the GPP datasets and the GPP observations from FLUXNET. a-h represents BESS, FLUXCOM, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

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.

We selected six independent models to estimate GPP in this study. These models are widely used with few model parameters and have demonstrated reliable accuracy in previous studies (Zheng et al., 2020; Zhang et al., 2017; Badgley et al., 2017).

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 serves as the primary 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 adopted methods 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, variations in these constraints varies significantly, leading to differences of over 10% in model explanatory power. (Yuan et al., 2014). Recent studies have proposed some novel vegetation indices that have been shown to be effective proxies for GPP through theoretical derivation and observed validation (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; Dechant et al., 2022). Both LUE and vegetation index models use a combination of linear mathematical formulas to estimate GPP. However, ecosystems are inherently complex, and the biases introduced by these numerical models increase the uncertainty in the estimates of the final product (GPP). Machine learning models has shown great potential for improving GPP estimates in previous studies (Jung et al., 2020; Guo et al., 2023). These model are 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 observed data and the representativeness of input data. Nevertheless, direct validation from flux towers of FLUXNET reveals that these models typically explain only about 70% of monthly GPP variations, with similar performance to other GPP models (Wang et al., 2021; Badgley et al., 2019; Zheng et al., 2020; Jung et al., 2020). Due to deviations in the model structure, a common limitation across these models is poor estimate of monthly extreme GPP, leading to the phenomenon of "high value overestimation and low value overestimation" (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 global 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 (Belgiu and Drăguţ, 2016). 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.

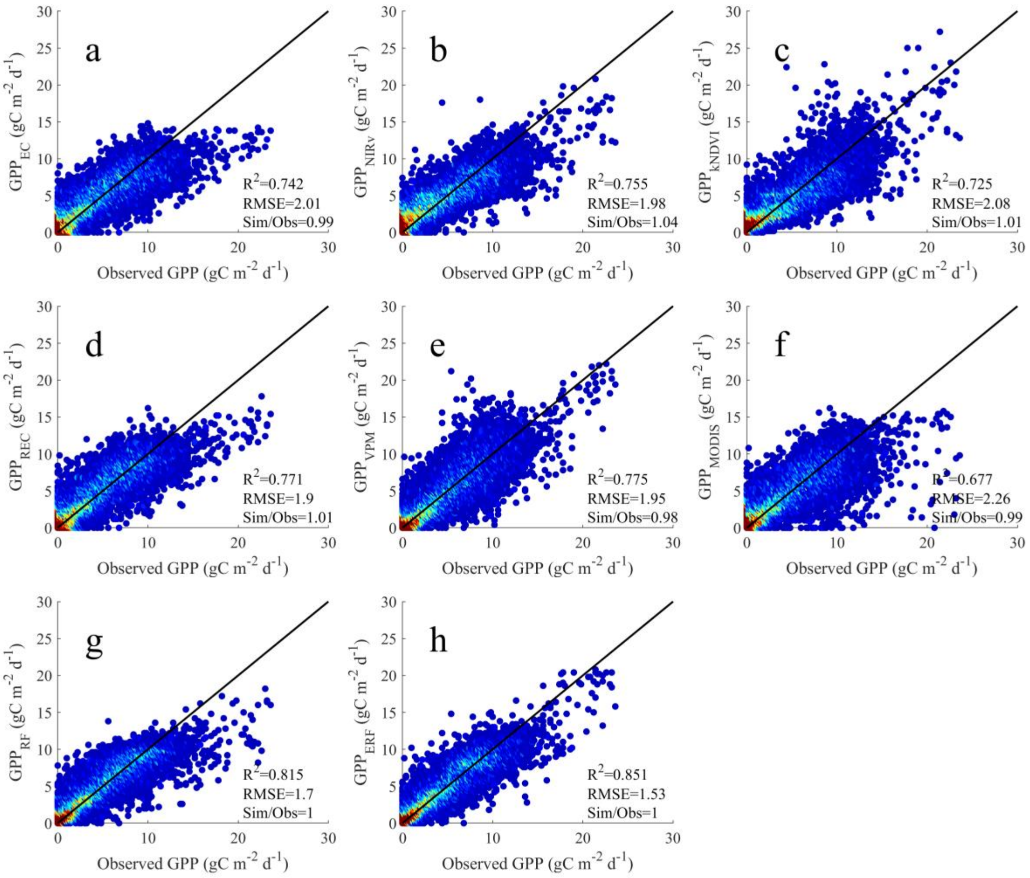


Figure R6. Comparison between the GPP simulations of the eight models and the GPP observations. a-h represents $\mathrm{GPP}_{\mathrm{EC}}$, $\mathrm{GPP}_{\mathrm{NIRv}}$, $\mathrm{GPP}_{\mathrm{kNDVI}}$, $\mathrm{GPP}_{\mathrm{REC}}$, $\mathrm{GPP}_{\mathrm{VPM}}$, $\mathrm{GPP}_{\mathrm{MODIS}}$,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

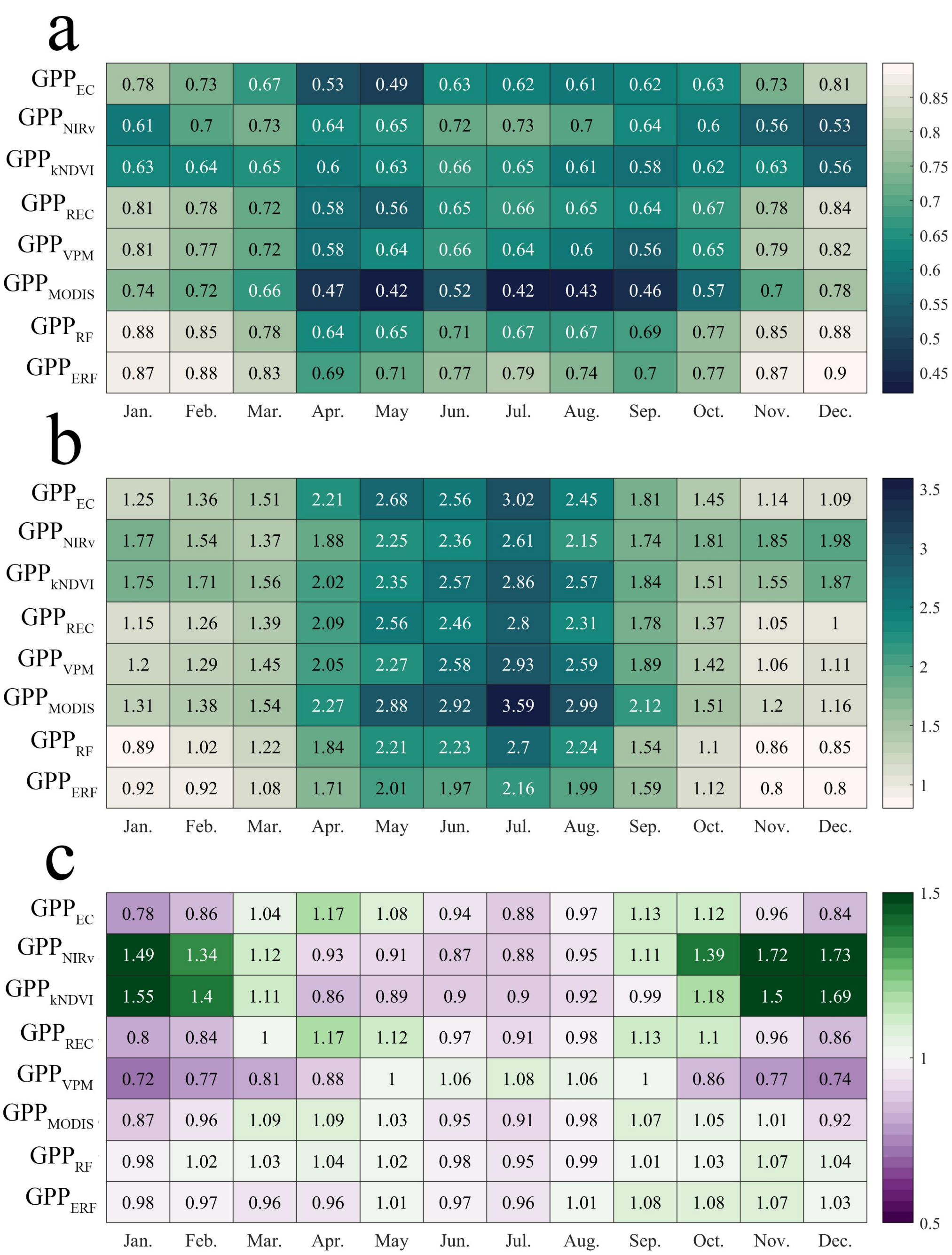
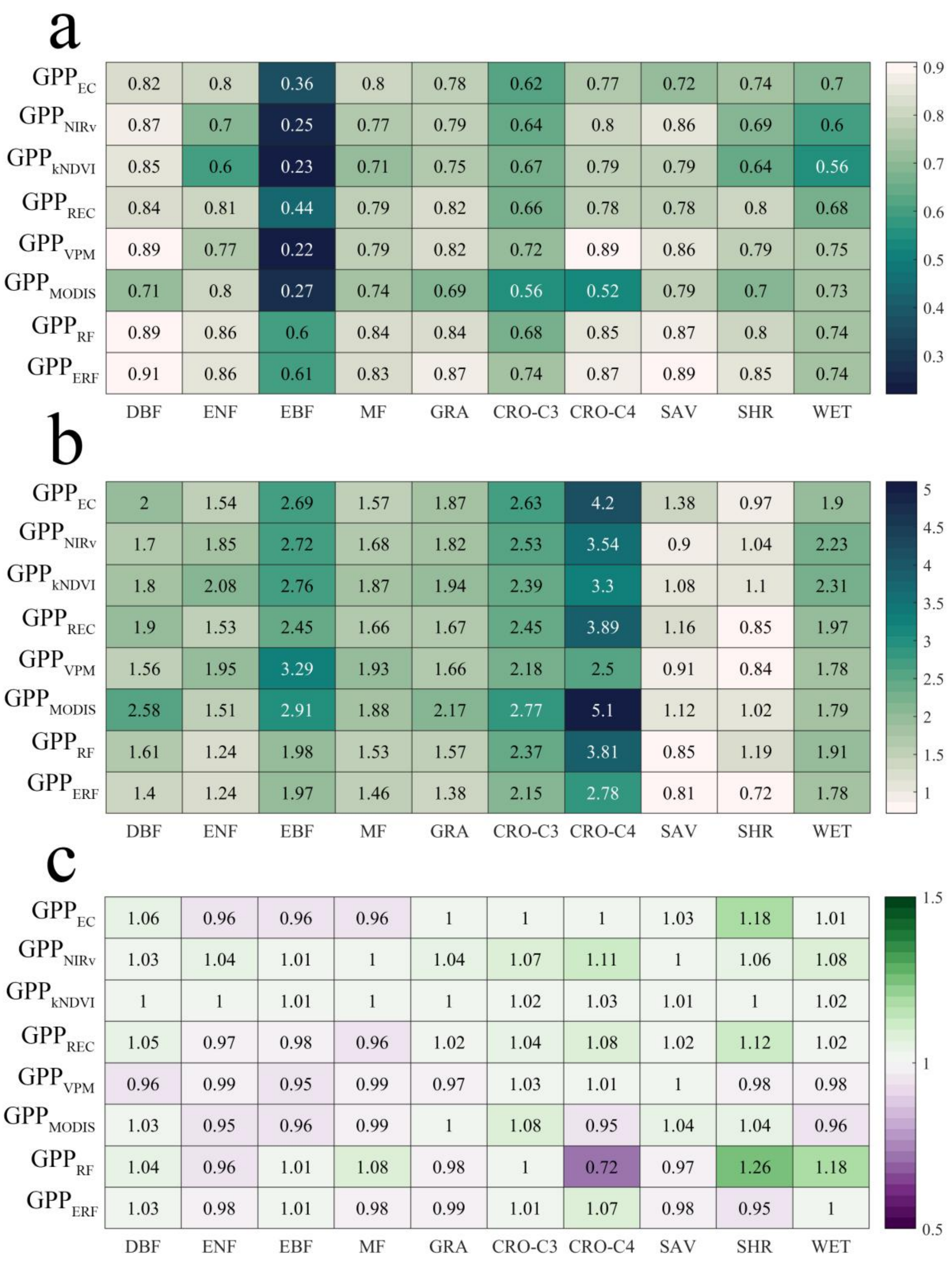


Figure R7. Performance of the eight models in each month. a, b and c represent R^2, RMSE, and Sim/Obs respectively.



	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
GPP_{EC}	0.82	0.8	0.36	0.8	0.78	0.62	0.77	0.72	0.74	0.7
GPP_{NIRv}	0.87	0.7	0.25	0.77	0.79	0.64	0.8	0.86	0.69	0.6
GPP_{kNDVI}	0.85	0.6	0.23	0.71	0.75	0.67	0.79	0.79	0.64	0.56
GPP_{REC}	0.84	0.81	0.44	0.79	0.82	0.66	0.78	0.78	0.8	0.68
GPP_{VPM}	0.89	0.77	0.22	0.79	0.82	0.72	0.89	0.86	0.79	0.75
GPP_{MODIS}	0.71	0.8	0.27	0.74	0.69	0.56	0.52	0.79	0.7	0.73
GPP_{RF}	0.89	0.86	0.6	0.84	0.84	0.68	0.85	0.87	0.8	0.74
GPP_{ERF}	0.91	0.86	0.61	0.83	0.87	0.74	0.87	0.89	0.85	0.74

	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
GPP_{EC}	2	1.54	2.69	1.57	1.87	2.63	4.2	1.38	0.97	1.9
GPP_{NIRv}	1.7	1.85	2.72	1.68	1.82	2.53	3.54	0.9	1.04	2.23
GPP_{kNDVI}	1.8	2.08	2.76	1.87	1.94	2.39	3.3	1.08	1.1	2.31
GPP_{REC}	1.9	1.53	2.45	1.66	1.67	2.45	3.89	1.16	0.85	1.97
GPP_{VPM}	1.56	1.95	3.29	1.93	1.66	2.18	2.5	0.91	0.84	1.78
GPP_{MODIS}	2.58	1.51	2.91	1.88	2.17	2.77	5.1	1.12	1.02	1.79
GPP_{RF}	1.61	1.24	1.98	1.53	1.57	2.37	3.81	0.85	1.19	1.91
GPP_{ERF}	1.4	1.24	1.97	1.46	1.38	2.15	2.78	0.81	0.72	1.78

	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
GPP_{EC}	1.06	0.96	0.96	0.96	1	1	1	1.03	1.18	1.01
GPP_{NIRv}	1.03	1.04	1.01	1	1.04	1.07	1.11	1	1.06	1.08
GPP_{kNDVI}	1	1	1.01	1	1	1.02	1.03	1.01	1	1.02
GPP_{REC}	1.05	0.97	0.98	0.96	1.02	1.04	1.08	1.02	1.12	1.02
GPP_{VPM}	0.96	0.99	0.95	0.99	0.97	1.03	1.01	1	0.98	0.98
GPP_{MODIS}	1.03	0.95	0.96	0.99	1	1.08	0.95	1.04	1.04	0.96
GPP_{RF}	1.04	0.96	1.01	1.08	0.98	1	0.72	0.97	1.26	1.18
GPP_{ERF}	1.03	0.98	1.01	0.98	0.99	1.01	1.07	0.98	0.95	1

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

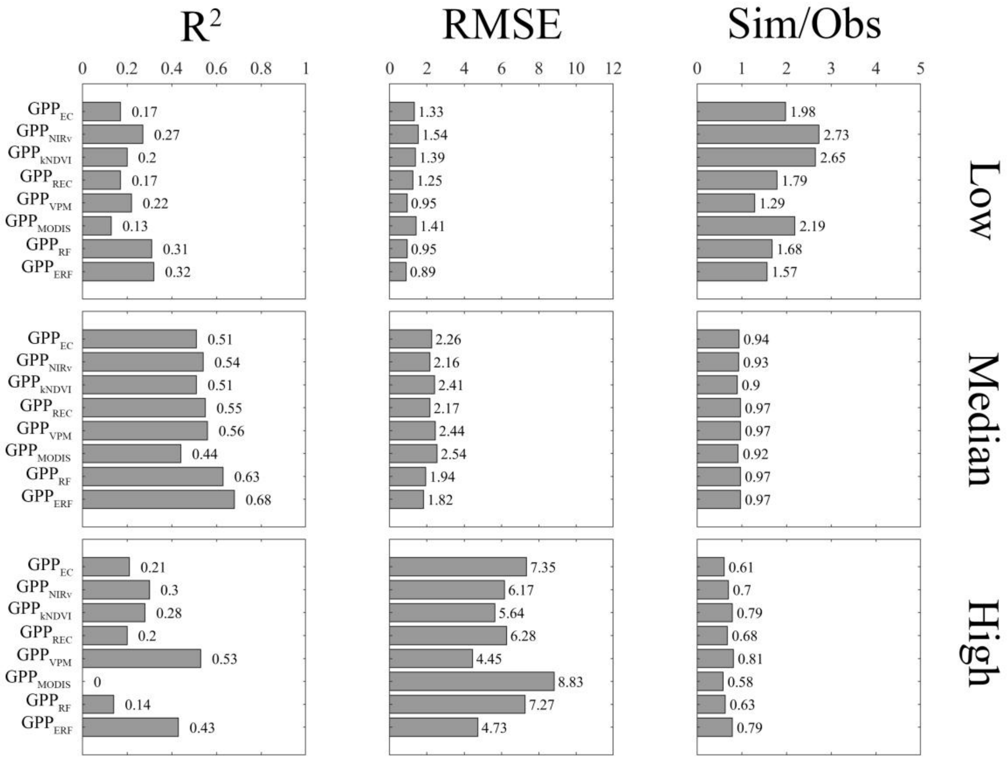


Figure R9. Performance of eight models in different subvalues.

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.

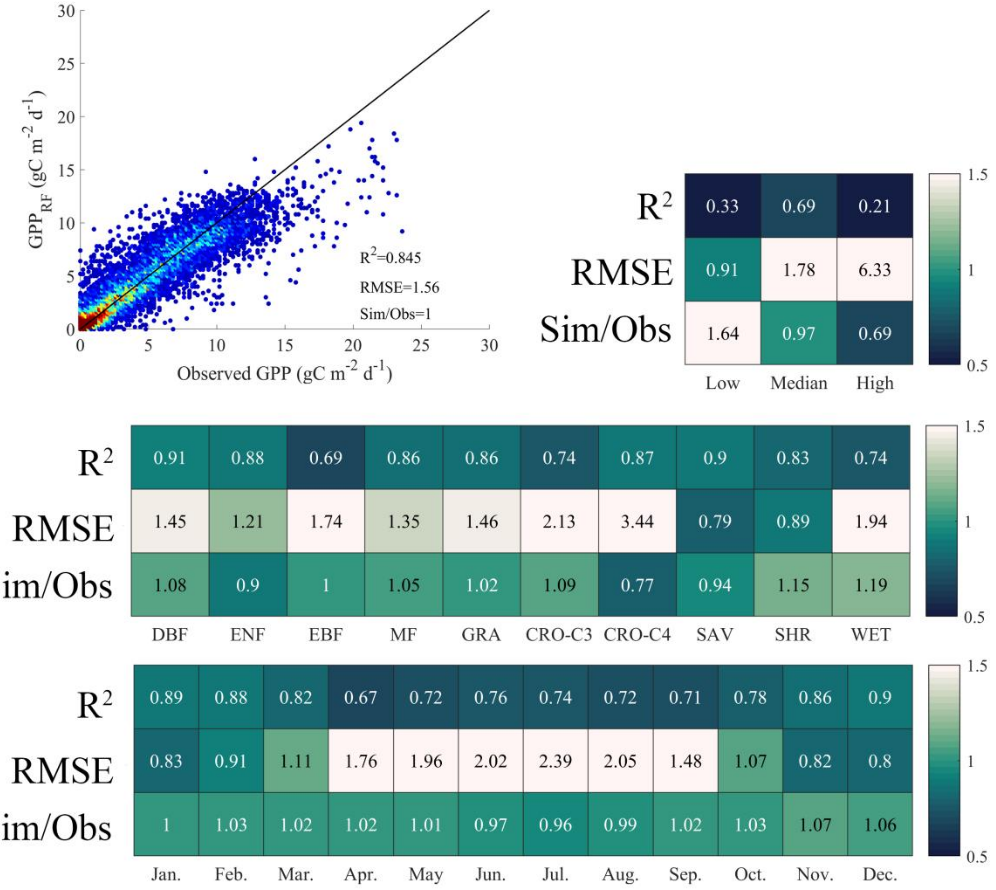


Figure R10. Performance of the random forest model using 9 variables.

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.

In addition, we selected monthly average air temperature, total solar radiation and VPD. The site observations do not provide direct solar radiation, so we extracted data from the ERA5 covering the flux tower. Monthly minimum air temperature was derived from hourly air temperature. Since some required model data are not directly available at flux sites, LAI and FPAR were extracted from MOD15A2H (500 m), and surface reflectance data (red band, near infrared band, blue band and shortwave infrared band) were derived from MCD43A4 (500 m) and MOD09A1 (500 m).

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 observations and meteorological data, lacking direct measurements for LAI, FPAR, and surface reflectance, so only remote sensing data can be used. Considering the variety of remote sensing data sources, such as MODIS and AVHRR, it is evident that calibrating the same GPP model with different remote sensing data can yield varied 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.

Reviewer #2 (Formal Review for Authors):

This study offers a contribution to global gross primary production (GPP) mapping, developing an ensemble model based on random forest algorithm. This model inputs GPP estimations from various remote sensing-based models, showing superior accuracy by explaining 83.7% of GPP variations across 171 sites, outperforming traditional models. It estimates the global GPP to be 131.2 PgC yr-1 from 2001-2022, with an increasing trend. While the authors have done a lot of work and the work is significant, the paper could benefit from a more comprehensive consideration of certain details and improvements in writing clarity.

In Section 2.3, the authors selected specific models as input variables for the ERF model. However, other widely applied models such as the P model, VPM model, MODIS GPP algorithm, and NIRvP for vegetation indices have not been considered. What was the rationale behind selecting these four models? Furthermore, in comparing global results, why were certain products chosen, such as VPM, MODIS, and FLUXCOM data, especially considering FLUXCOM also employs machine learning methods and has released a new version of its data (FLUXCOMX)? Additionally, it appears the ECGC has only recently been launched and may not be as "widely used" as mentioned in the manuscript.

REPLY: Thanks for your comments. 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.

We selected six independent models to estimate GPP in this study. These models are widely used with few model parameters and have demonstrated reliable accuracy in previous studies (Zheng et al., 2020; Zhang et al., 2017; Badgley et al., 2017).

At the same time, according to your suggestion, we have also added VPM and MODIS in the revised version. In other words, there are 6 GPP models in the ensemble model in the latest version. As shown in Figure R1-R4, the result is similar to the original paper. In all respects, the performance of the ensemble model is best.

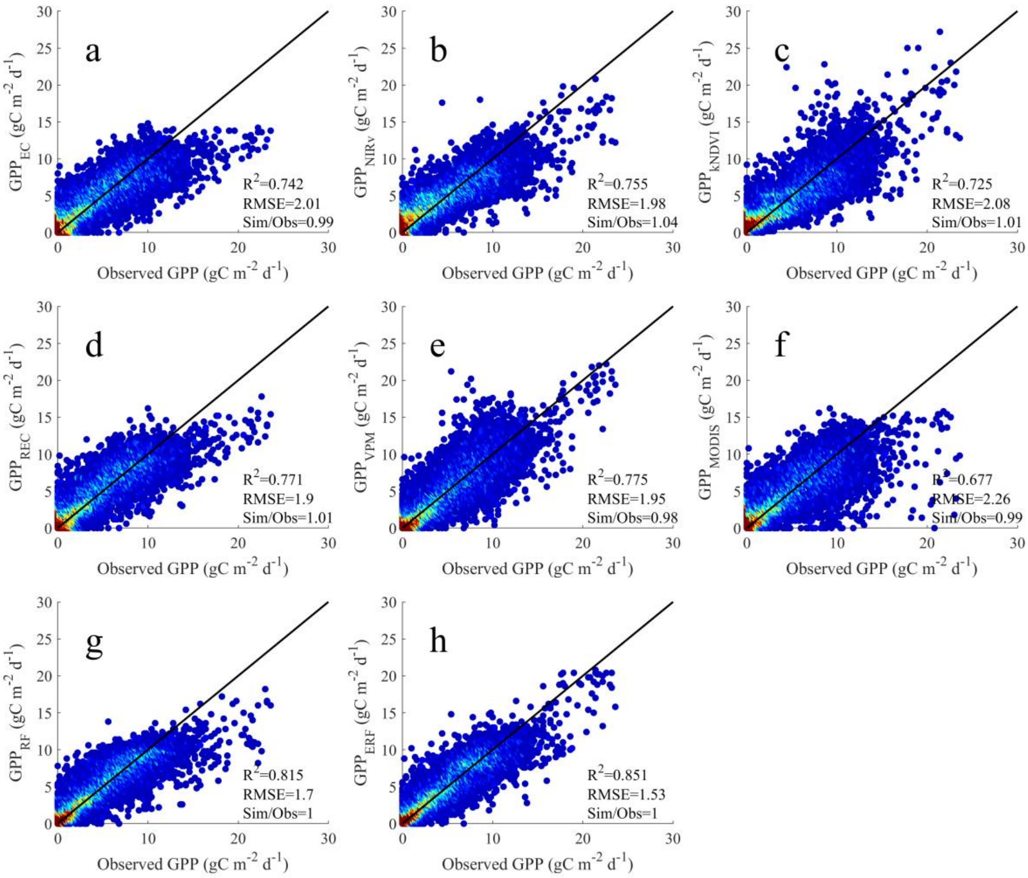
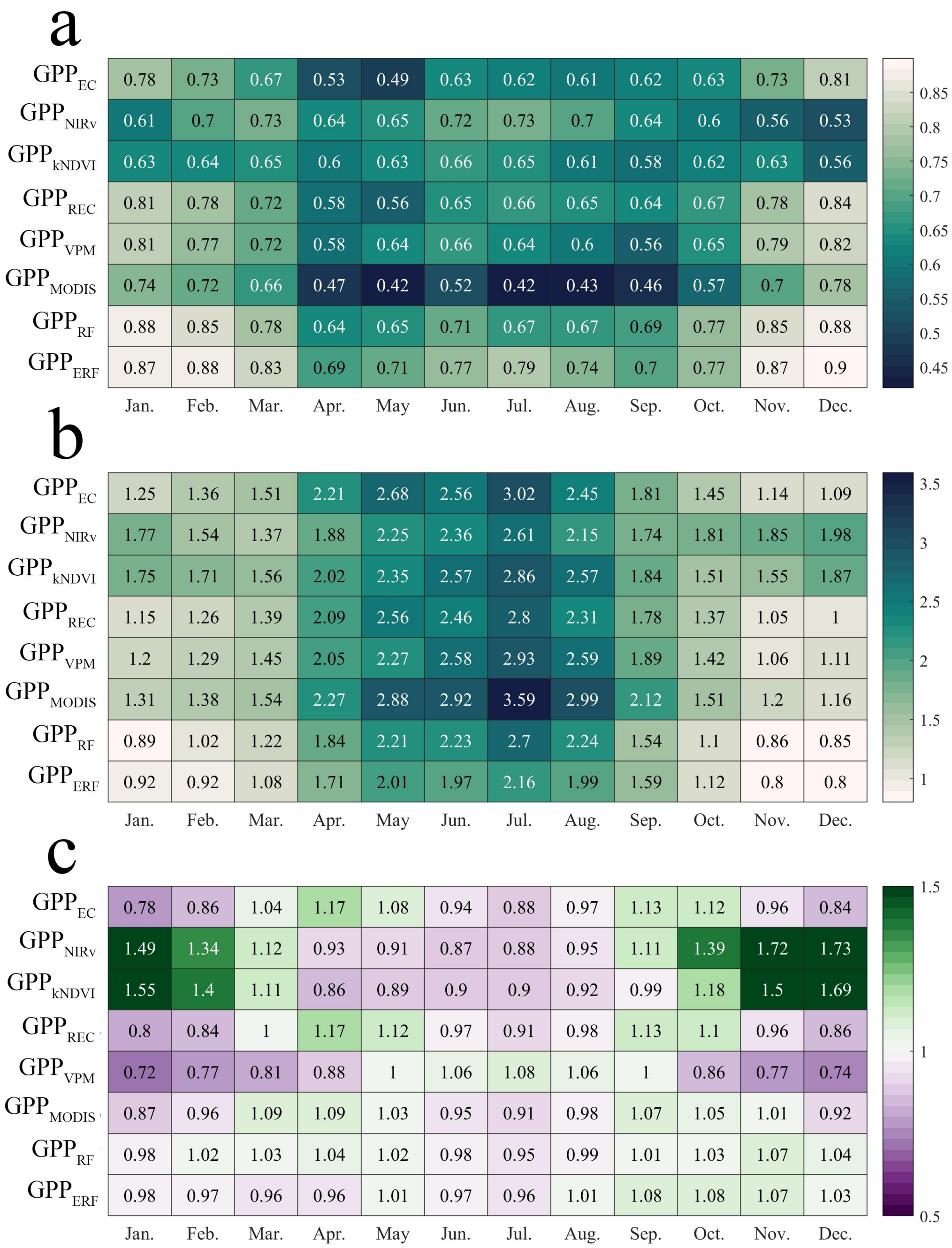


Figure R1. Comparison between the GPP simulations of the eight models and the GPP observations. a-h represents GPP_{EC}, GPP_{NIRv}, GPP_{kNDVI}, GPP_{REC}, GPP_{VPM}, GPP_{MODIS}, GPP_{RF}, GPP_{ERF}, respectively.



a

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
GPP_{EC}	0.78	0.73	0.67	0.53	0.49	0.63	0.62	0.61	0.62	0.63	0.73	0.81
GPP_{NIRv}	0.61	0.7	0.73	0.64	0.65	0.72	0.73	0.7	0.64	0.6	0.56	0.53
GPP_{kNDVI}	0.63	0.64	0.65	0.6	0.63	0.66	0.65	0.61	0.58	0.62	0.63	0.56
GPP_{REC}	0.81	0.78	0.72	0.58	0.56	0.65	0.66	0.65	0.64	0.67	0.78	0.84
GPP_{VPM}	0.81	0.77	0.72	0.58	0.64	0.66	0.64	0.6	0.56	0.65	0.79	0.82
GPP_{MODIS}	0.74	0.72	0.66	0.47	0.42	0.52	0.42	0.43	0.46	0.57	0.7	0.78
GPP_{RF}	0.88	0.85	0.78	0.64	0.65	0.71	0.67	0.67	0.69	0.77	0.85	0.88
GPP_{ERF}	0.87	0.88	0.83	0.69	0.71	0.77	0.79	0.74	0.7	0.77	0.87	0.9

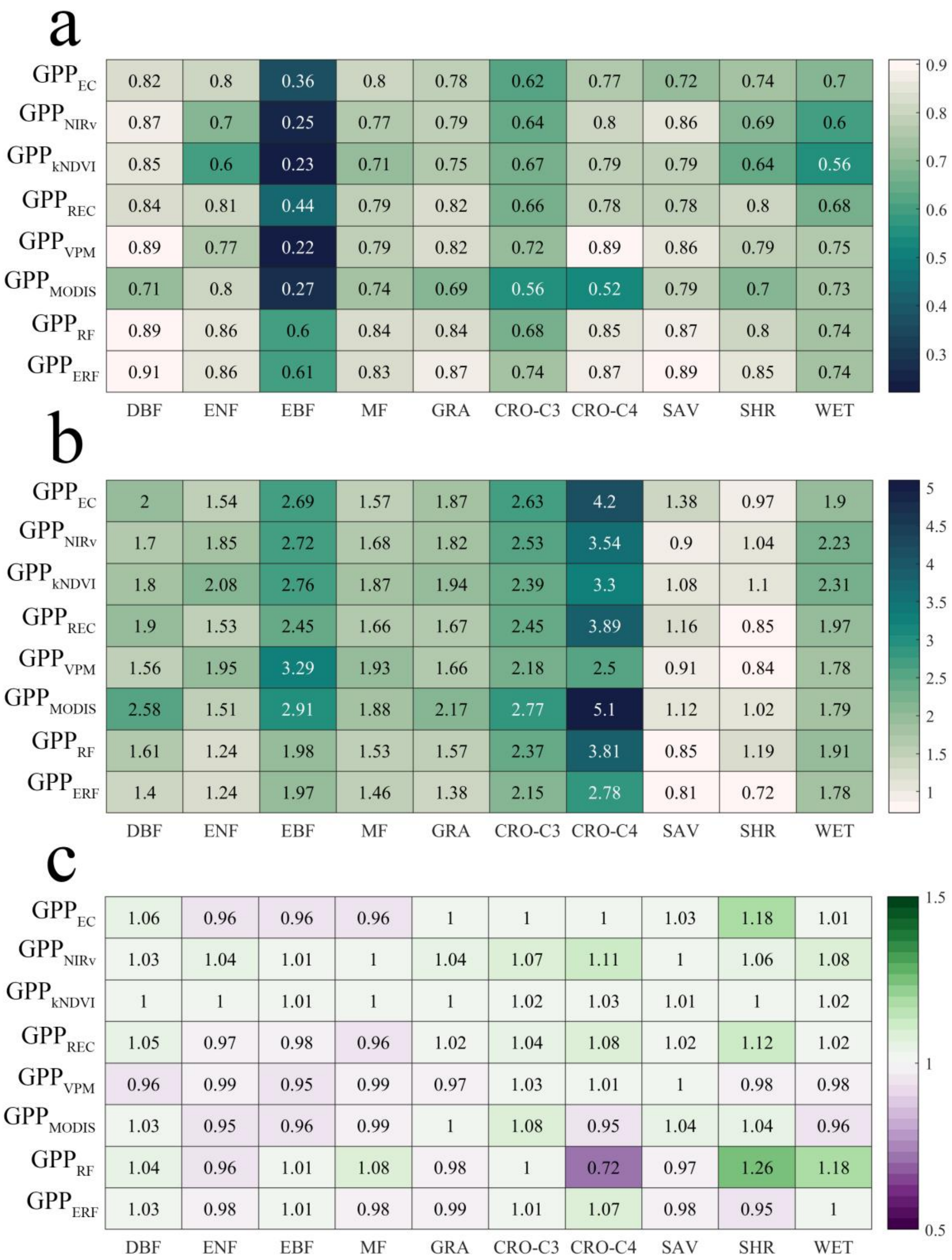
b

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
GPP_{EC}	1.25	1.36	1.51	2.21	2.68	2.56	3.02	2.45	1.81	1.45	1.14	1.09
GPP_{NIRv}	1.77	1.54	1.37	1.88	2.25	2.36	2.61	2.15	1.74	1.81	1.85	1.98
GPP_{kNDVI}	1.75	1.71	1.56	2.02	2.35	2.57	2.86	2.57	1.84	1.51	1.55	1.87
GPP_{REC}	1.15	1.26	1.39	2.09	2.56	2.46	2.8	2.31	1.78	1.37	1.05	1
GPP_{VPM}	1.2	1.29	1.45	2.05	2.27	2.58	2.93	2.59	1.89	1.42	1.06	1.11
GPP_{MODIS}	1.31	1.38	1.54	2.27	2.88	2.92	3.59	2.99	2.12	1.51	1.2	1.16
GPP_{RF}	0.89	1.02	1.22	1.84	2.21	2.23	2.7	2.24	1.54	1.1	0.86	0.85
GPP_{ERF}	0.92	0.92	1.08	1.71	2.01	1.97	2.16	1.99	1.59	1.12	0.8	0.8

c

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
GPP_{EC}	0.78	0.86	1.04	1.17	1.08	0.94	0.88	0.97	1.13	1.12	0.96	0.84
GPP_{NIRv}	1.49	1.34	1.12	0.93	0.91	0.87	0.88	0.95	1.11	1.39	1.72	1.73
GPP_{kNDVI}	1.55	1.4	1.11	0.86	0.89	0.9	0.9	0.92	0.99	1.18	1.5	1.69
GPP_{REC}	0.8	0.84	1	1.17	1.12	0.97	0.91	0.98	1.13	1.1	0.96	0.86
GPP_{VPM}	0.72	0.77	0.81	0.88	1	1.06	1.08	1.06	1	0.86	0.77	0.74
GPP_{MODIS}	0.87	0.96	1.09	1.09	1.03	0.95	0.91	0.98	1.07	1.05	1.01	0.92
GPP_{RF}	0.98	1.02	1.03	1.04	1.02	0.98	0.95	0.99	1.01	1.03	1.07	1.04
GPP_{ERF}	0.98	0.97	0.96	0.96	1.01	0.97	0.96	1.01	1.08	1.08	1.07	1.03

Figure R2. Performance of the eight models in each month. a, b and c represent R^2, RMSE, and Sim/Obs respectively.



	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
GPP_{EC}	0.82	0.8	0.36	0.8	0.78	0.62	0.77	0.72	0.74	0.7
GPP_{NIRv}	0.87	0.7	0.25	0.77	0.79	0.64	0.8	0.86	0.69	0.6
GPP_{kNDVI}	0.85	0.6	0.23	0.71	0.75	0.67	0.79	0.79	0.64	0.56
GPP_{REC}	0.84	0.81	0.44	0.79	0.82	0.66	0.78	0.78	0.8	0.68
GPP_{VPM}	0.89	0.77	0.22	0.79	0.82	0.72	0.89	0.86	0.79	0.75
GPP_{MODIS}	0.71	0.8	0.27	0.74	0.69	0.56	0.52	0.79	0.7	0.73
GPP_{RF}	0.89	0.86	0.6	0.84	0.84	0.68	0.85	0.87	0.8	0.74
GPP_{ERF}	0.91	0.86	0.61	0.83	0.87	0.74	0.87	0.89	0.85	0.74

	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
GPP_{EC}	2	1.54	2.69	1.57	1.87	2.63	4.2	1.38	0.97	1.9
GPP_{NIRv}	1.7	1.85	2.72	1.68	1.82	2.53	3.54	0.9	1.04	2.23
GPP_{kNDVI}	1.8	2.08	2.76	1.87	1.94	2.39	3.3	1.08	1.1	2.31
GPP_{REC}	1.9	1.53	2.45	1.66	1.67	2.45	3.89	1.16	0.85	1.97
GPP_{VPM}	1.56	1.95	3.29	1.93	1.66	2.18	2.5	0.91	0.84	1.78
GPP_{MODIS}	2.58	1.51	2.91	1.88	2.17	2.77	5.1	1.12	1.02	1.79
GPP_{RF}	1.61	1.24	1.98	1.53	1.57	2.37	3.81	0.85	1.19	1.91
GPP_{ERF}	1.4	1.24	1.97	1.46	1.38	2.15	2.78	0.81	0.72	1.78

	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
GPP_{EC}	1.06	0.96	0.96	0.96	1	1	1	1.03	1.18	1.01
GPP_{NIRv}	1.03	1.04	1.01	1	1.04	1.07	1.11	1	1.06	1.08
GPP_{kNDVI}	1	1	1.01	1	1	1.02	1.03	1.01	1	1.02
GPP_{REC}	1.05	0.97	0.98	0.96	1.02	1.04	1.08	1.02	1.12	1.02
GPP_{VPM}	0.96	0.99	0.95	0.99	0.97	1.03	1.01	1	0.98	0.98
GPP_{MODIS}	1.03	0.95	0.96	0.99	1	1.08	0.95	1.04	1.04	0.96
GPP_{RF}	1.04	0.96	1.01	1.08	0.98	1	0.72	0.97	1.26	1.18
GPP_{ERF}	1.03	0.98	1.01	0.98	0.99	1.01	1.07	0.98	0.95	1

Figure R3. The performance of the eight models on different vegetation types. a, b and c represent R^2, RMSE, and Sim/Obs respectively.

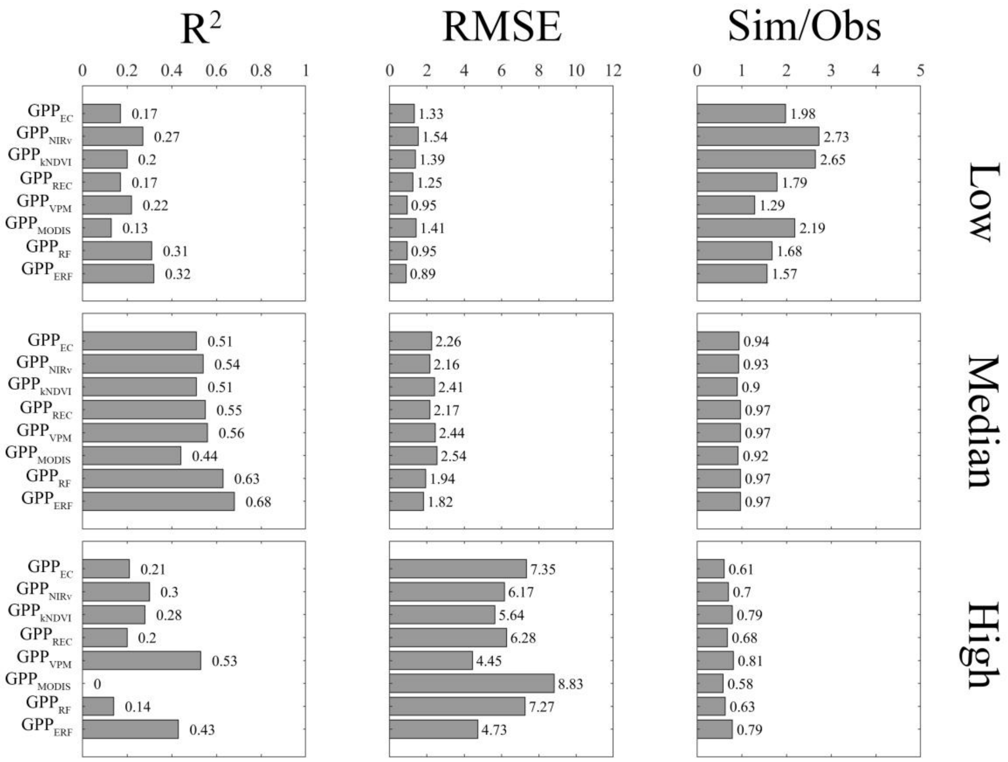


Figure R4. Performance of eight models in different subvalues.

We didn't consider the P model and NIRvP. For the P model, although it is the structure of the LUE model, the calculation of the Photo respiratory compensation point parameter of this model is actually very complicated, which is similar to the process model. This point violates the basic criteria for selecting GPP models in this study. For NIRvP, in a recent study, we found that the model underestimated the impact of drought on GPP by not taking into account environmental constraints (Chen et al, 2024). That is, in dry years, the negative anomaly of GPP is very small, which is obviously inconsistent with the observation. Due to this shortcoming, we do not consider using this model to estimate the global GPP, although its performance may be similar to other models. 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

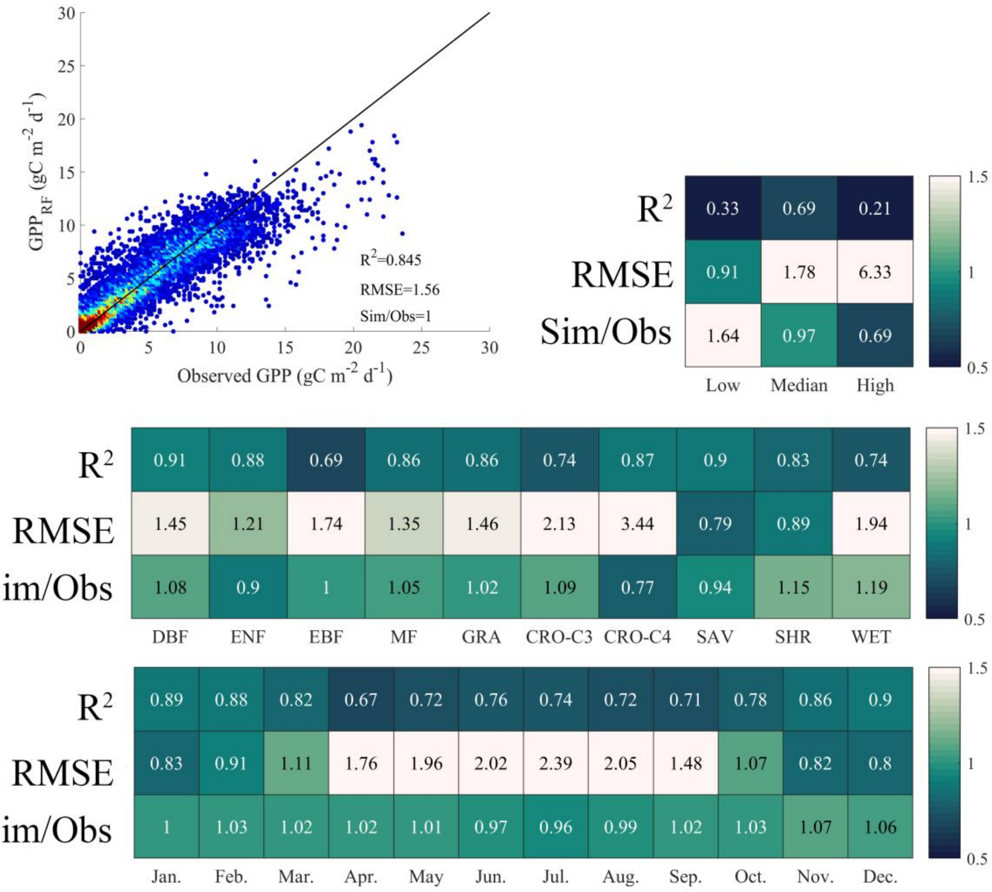
Chen, X., Chen, T., Liu, S., Chai, Y., Guo, R., Dai, J., ... & Wei, X. (2024). Vegetation Index-Based Models Without Meteorological Constraints Underestimate the Impact of Drought on Gross Primary Productivity. Journal of Geophysical Research: Biogeosciences, 129(1), e2023JG007499.

The authors compare the ERF model with a traditional random forest (RF) model. Table 2 indicates that the traditional RF model used only 4 variables, while the ERF model incorporates several GPP estimation models. However, it actually includes even more variables, such as kNDVI, NIRv, FPAR, CO2, dif/dir SR, etc. The ERF model contains more variables than the RF model, but for a fair comparison, the same data should be used. Would the accuracy of the ERF model still surpass that of the RF model if an RF model were constructed using all data inputs from the ERF model?

REPLY: Thanks for your comments. 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 R1-R4, 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.

As shown in Figure R5, 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.



	Low	Median	High
R^2	0.33	0.69	0.21
RMSE	0.91	1.78	6.33
Sim/Obs	1.64	0.97	0.69

	DBF	ENF	EBF	MF	GRA	CRO-C3	CRO-C4	SAV	SHR	WET
R^2	0.91	0.88	0.69	0.86	0.86	0.74	0.87	0.9	0.83	0.74
RMSE	1.45	1.21	1.74	1.35	1.46	2.13	3.44	0.79	0.89	1.94
Sim/Obs	1.08	0.9	1	1.05	1.02	1.09	0.77	0.94	1.15	1.19

	Jan.	Feb.	Mar.	Apr.	May	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
R^2	0.89	0.88	0.82	0.67	0.72	0.76	0.74	0.72	0.71	0.78	0.86	0.9
RMSE	0.83	0.91	1.11	1.76	1.96	2.02	2.39	2.05	1.48	1.07	0.82	0.8
Sim/Obs	1	1.03	1.02	1.02	1.01	0.97	0.96	0.99	1.02	1.03	1.07	1.06

Figure R5. Performance of the random forest model using 9 variables.

Why did the authors opt to estimate monthly GPP instead of daily? Are the estimation results from different models in the ERF model aggregated from daily to monthly, or are they directly estimating monthly GPP? If monthly, how are parameters like Solar Zenith Angle adjusted when optimizing the rECLUE model?

REPLY: Thanks for your comments. All the results of the model simulation were carried out on the monthly scale. If it is a daily scale GPP simulation, even at 0.05 resolution, it will take a lot of time, so we did not do daily scale GPP simulation. For the solar zenith angle parameter in Revised-EC-LUE, we use the solar zenith angle in the middle of each month as the solar zenith angle of the current month, which is a simplification. Compared with several important parameters that affect GPP simulation, the effect of this parameter is negligible.

In Table 2, the EC-LUE model considers VPD and CO2, which the original model does not. The supplementary documents indicate that the authors modified the EC-LUE model, thus it is no longer the original EC-LUE model. The only difference between it and the rECLUE model seems to be the consideration of sunlit and shaded leaves. Given that Figure 1 shows minimal differences between them, does including it as an input for the ERF model result in redundancy with rECLUE?

REPLY: Thanks for your comments. First of all, we did not modify the EC-LUE, we used the version published by Yuan et al (2019). As you said, based on our results, the difference between the two models is really not obvious. However, we wish to retain this result because a secondary purpose of our study was to compare the performance differences of these models after parameter calibration.

To address your concerns, our study adds an additional analysis of using different numbers of GPP models in the ensemble model to further compare the performance differences in the final results. As shown in Table R1, As the number of GPP in the ensemble model increases, the model performance gains gradually decrease.

Yuan, W., Zheng, Y., Piao, S., Ciais, P., Lombardozzi, D., Wang, Y., ... & Yang, S. (2019). Increased atmospheric vapor pressure deficit reduces global vegetation growth. Science advances, 5(8), eaax1396.

The introduction requires careful revision as many uncertainties or current issues listed by the authors seem not to be addressed in this manuscript.

REPLY: Thanks for your comments. The revised introduction highlights the uncertainties of several GPP models and introduces ensemble model.

The light use efficiency (LUE) model is one of the most widely adopted methods 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, variations in these constraints varies significantly, leading to differences of over 10% in model explanatory power. (Yuan et al., 2014). Recent studies have proposed some novel vegetation indices that have been shown to be effective proxies for GPP through theoretical derivation and observed validation (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; Dechant et al., 2022). Both LUE and vegetation index models use a combination of linear mathematical formulas to estimate GPP. However, ecosystems are inherently complex, and the biases introduced by these numerical models increase the uncertainty in the estimates of the final product (GPP). Machine learning models has shown great potential for improving GPP estimates in previous studies (Jung et al., 2020; Guo et al., 2023). These model are 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 observed data and the representativeness of input data. Nevertheless, direct validation from flux towers of FLUXNET reveals that these models typically explain only about 70% of monthly GPP variations, with similar performance to other GPP models (Wang et al., 2021; Badgley et al., 2019; Zheng et al., 2020; Jung et al., 2020). Due to deviations in the model structure, a common limitation across these models is poor estimate of monthly extreme GPP, leading to the phenomenon of "high value overestimation and low value overestimation" (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 global carbon cycle.

Some detailed comments:

L41: The authors suggest poor estimation accuracy partly because remote sensing models cannot fully represent photosynthesis. Does the ERF model overcome this limitation?

REPLY: Thanks for your comments. The ERF model also does not fully address this problem, but only improves the estimation of the GPP. In the revised version, this sentence has been deleted.

L46-47: What does "this process may be missing" refer to? Is it the CO2 fertilization effect or a negative trend influenced by CO2? If it's the fertilization effect, many models already consider its impact. If it refers to a negative trend, what improvements have been made in the ERF model? I think this negative trend might not be incorporated into the model.

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.

L52: The authors note significant differences in the same vegetation types across different regions, but it seems the ERF model did not address this variability when optimizing parameters and developing the model.

REPLY: Thanks for your comments. We agree with you that this sentence has been deleted.

L54-55: It's unclear what this typical example refers to. Parameters for C3 and C4 vegetation inherently need to be considered separately, representing two different vegetation types.

REPLY: Thanks for your comments. Although C3 and C4 are two types of planting, C3 and C4 crops were not divided in many previous studies. Here we want to emphasize the difference between C3 and C4 in the growing season, in the revised version, this sentence has been deleted.

L56-60: Environmental factors add to GPP estimation uncertainty. How have the authors improved or reduced this uncertainty, given that most models already account for environmental factors?

REPLY: Thanks for your comments. In the ERF model, the uncertainty of these environmental constraints has actually been propagated into the simulated GPP, that is, during the modeling process, the model only needs to consider the uncertainty of the simulated GPP. Accordingly, for other GPP models, there is still the influence of the uncertainty of environmental constraints. In the discussion section of the revised version, we have added an explanation of the relevant content.

In other words, the ERF model does not need to take into account the uncertainties of the model structure (such as meteorological constraints) and model parameters (such as maximum light use efficiency), but rather focuses on the uncertainties inherent in the simulated GPP.

L69-70: Tian et al. (2023) also applied ML models to multi-model ensembles. What are the innovative aspects of this study compared to their research?

REPLY: Thanks for your comments. Compared with Tian et al. (2023), our study is a further extension of applying an ensemble model to GPP estimation. There is a big difference compared to their study. Firstly, parameter calibration was carried out in our study so that the final validation results were comparable, that is, the difference in model performance was mainly due to the uncertainty of the model structure. Secondly, our research focuses on the phenomenon of "low value overestimation and high value underestimation" of the GPP model, and the research results show that the ensemble model has a good performance in different vegetation types, different months, and different subvalues. Finally, the ERF model was used to estimate the global GPP and validated on different observational data sets, which further proves the robustness of the ERF model in GPP estimation. In the discussion section of the revised version, We explained the differences between the results of this study and theirs.

It is worth noting that in the study of Tian et al. (2023), the ERF model was also used to improve the GPP estimation. Our research extends this work in several ways. Firstly, parameter calibration was carried out in our study so that the final validation results are comparable, that is, differences in model performance are mainly due to the uncertainty of the model structure. Secondly, our study focuses on the phenomenon of "high value underestimation and low value overestimation" of GPP model, with results indicating that the ERF model performed well across various vegetation types, months, and subvalues. Finally, we generated the ERF_GPP dataset and validate it on different observational datasets, further confirming the robustness of the ERF model in GPP estimation.

L85: How is ERA5-LAND data procesed in coastal regions? What is the reason for choosing temperature and radiation data from ERA5-Land and ERA5 respectively (this distinction should be made clear in Table 1)?

REPLY: Thanks for your comments. For coarser data conversions to 0.05°, we used the nearest neighbor resampling method. We do the same in the coastal areas. There is no direct radiation in ERA-land, so we used ERA5 monthly data on single levels. In the revised version, we illustrate this in Table1.

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.

L104: What does "reference year" mean? How are different datasets aggregated to 0.05 degrees?

REPLY: Thanks for your comments. In the process of calculating the global GPP, land use data is needed. For 2001-2022, we all use data from the same year (i.e., reference year). The simulation was conducted at a resolution of 0.05°, so the effect of land use change on GPP can be negligible. 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.

Section 2.5: Why not utilize all available Fluxnet sites for validation instead of limiting to only Chinese sites? Would this not lead to a smaller dataset and reduce the representativeness for validating a global product?

REPLY: Thanks for 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 R6, 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.

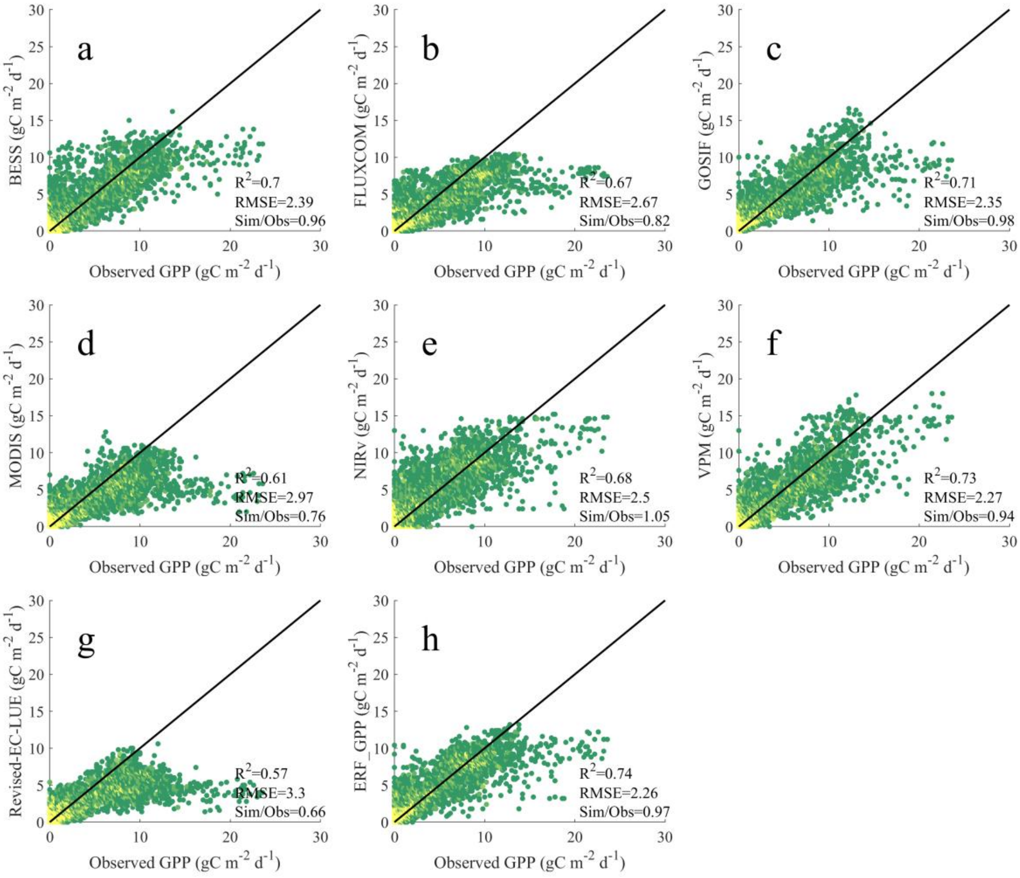


Figure R6. Comparison between the GPP datasets and the GPP observations from FLUXNET. a-h represents BESS, FLUXCOM, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

Figures 1 and 2: It's recommended to include units for GPP, and RMSE should also specify units.

REPLY: Thanks for your comments. In the revised version, we have added units.

Figure 3: Adding seasonal variation for representative sites of different vegetation types could better highlight the model's advantages.

REPLY: Thanks for your comments. In the revised version, we have added two typical sites to illustrate that the ensemble model's improvements to GPP are improvements to time series. We did not select a typical site analysis for all vegetation types because the ensemble model showed similar improvements for most sites.

As shown in Figure R7 and R8, 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 show obvious underestimation during the growing season. On CH_Lae, GPP_{kNDVI} and GPP_{VPM} are 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 calibrated at all sites). The performance of each model is different at different sites, mainly because the process concerned by each model (environmental constraints) is different. For example, NIRv and kNDVI do not use environmental 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).

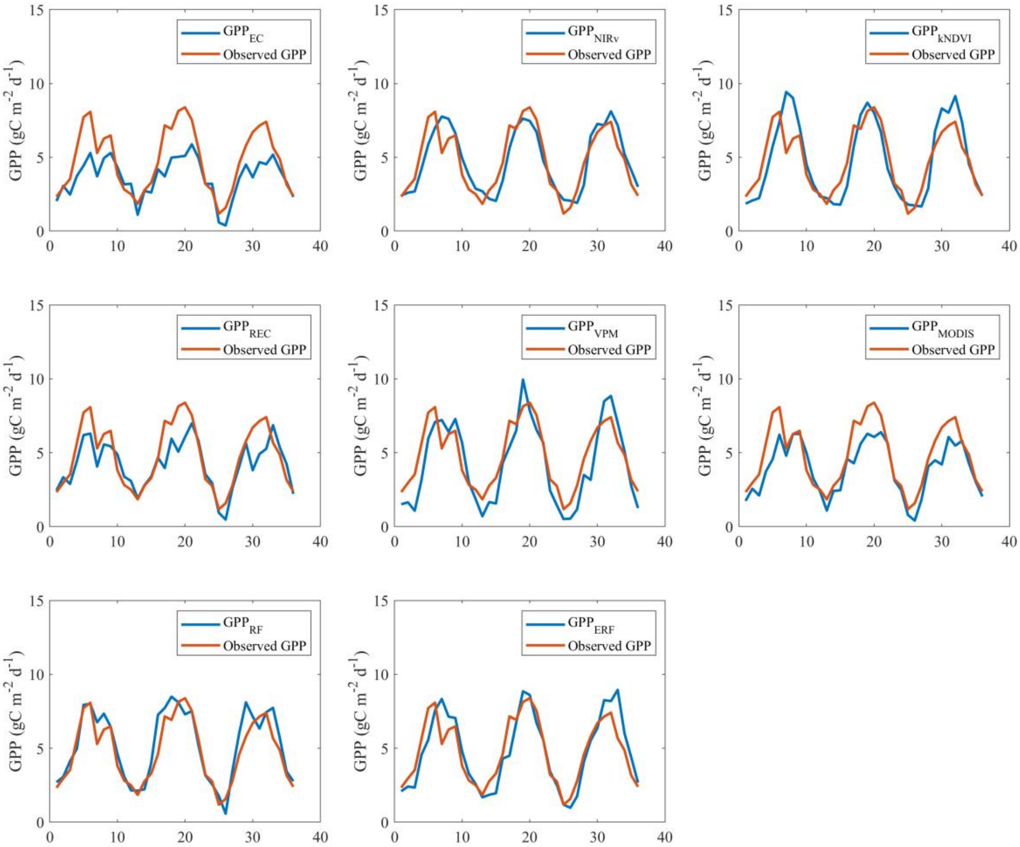


Figure R7. Performance of each GPP model on CN-Qia.

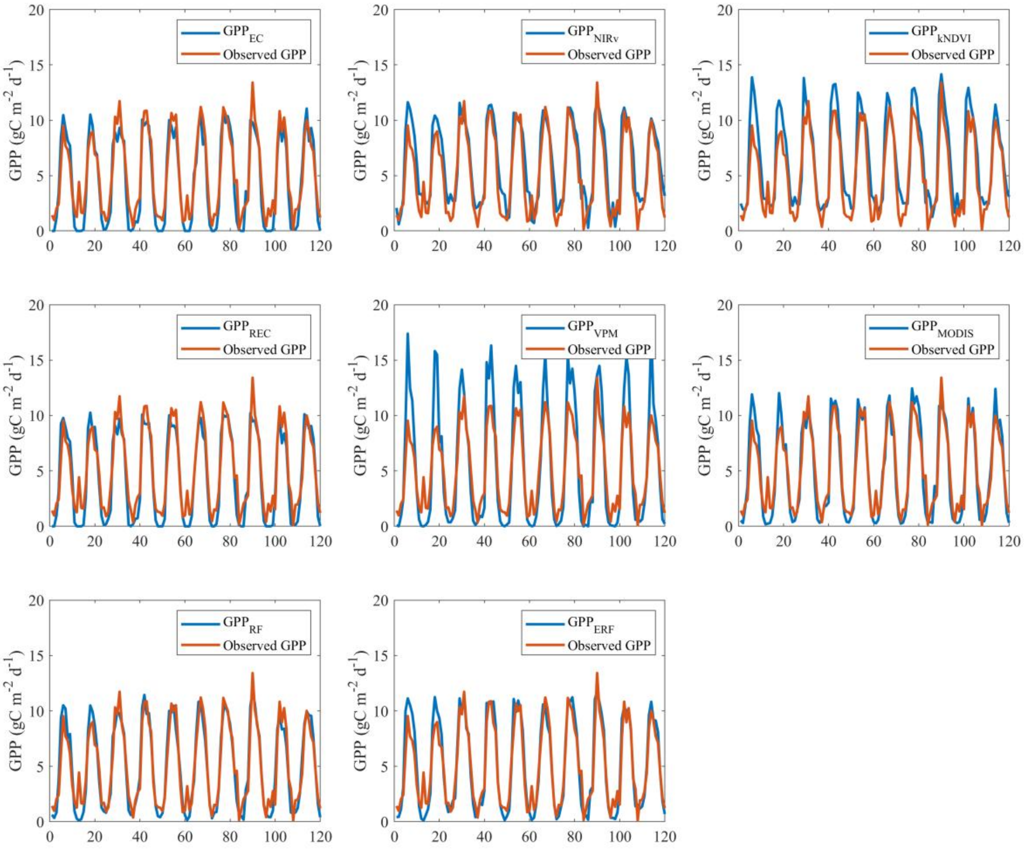


Figure R8. The performance of each GPP model on CH_Lae.

L228: Does ERF_GPP refer to the global product, while GPPERF denotes site estimation values?

REPLY: Thanks for your comments. As you said, GPP_{ERF} represents the site simulation and ERF_GPP represents the global GPP. In the revised version, we defined these.

L257, NIRV should be corrected to NIRv.

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

In Figure S4, discrepancies with Figure 6 are noted. Is it reasonable to directly average accuracy across various sites, given differences in data quantity and the range of GPP values at different sites?

REPLY: Thanks for your comments. This average is indeed not very reasonable, in the revised version, we deleted this part of the content.

L275: What does "representative" refer to in this context?

REPLY: Thanks for your comments. In the ERF model, we performed a feature importance analysis (Figure R9). From the average of 200 times, the results of the ensemble model do not depend on a single GPP simulation. Even the GPP_{MODIS} with the highest relative importance does not exceed 25%, and it looks more like a weighted average of multiple GPP simulations. There is no mechanism for machine learning, so we do not know the specific reason for this result. Therefore, the term "representative" here refers to the multiple GPP simulations, not a single GPP simulation. In the revised version, we have added a description of the relevant content in the discussion.

To further clarify the impact of explanatory variables on the ERF model, we conducted a feature importance analysis (Figure S9). From an average of 200 times, the results of the ERF model did not depend on a single GPP simulation. Even GPPMODIS, with the highest relative importance, accounted for no more than 25%, suggesting that the ERF model behaves more like a weighted average of multiple GPP simulations.

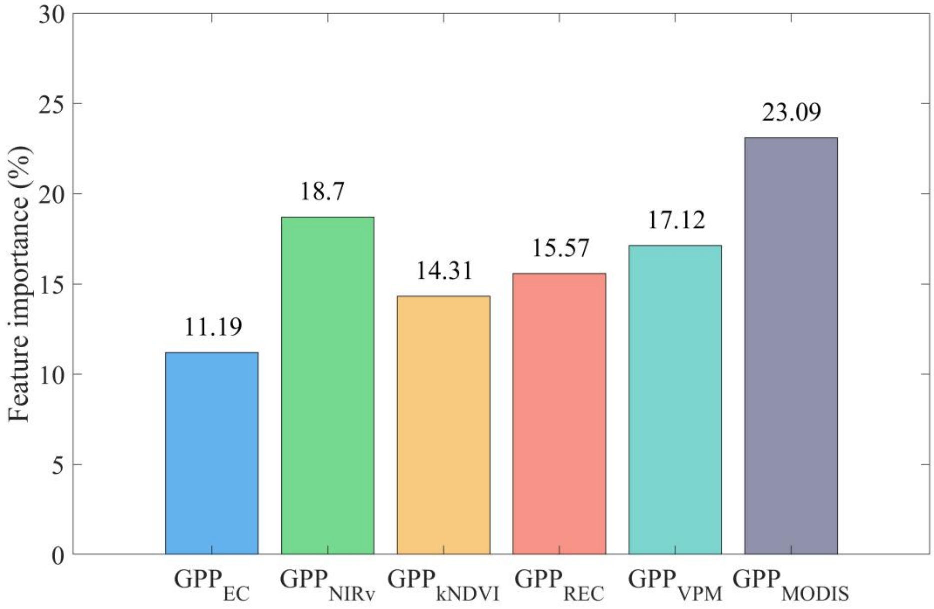


Figure R9. Average of 200 feature importance in the ERF model.

L280-282: Some models and products already utilize dynamic temperature parameters, which the authors have not mentioned or compared.

REPLY: Thanks for your comments. After searching, we found relevant study. We cite this study in the revised version and show that this refinement has the potential to improve global GPP estimates.

Previous study has shown that the estimation of GPP can be effectively improved by using dynamic temperature parameters (Chang et al., 2021).

L283-293: Could the overestimation of low values be due to scale issues, even at the site scale, considering the used LAI is 500 m?

REPLY: Thanks for your comments. The LAI of 500m is actually quite consistent with the range of the flux tower. It is possible to attribute the problem of overestimation of low values to scale problems, that is, modeling with 30m or 100m data may not have this problem. However, 30m and 100m are not in line with the observation range of the flux tower, and we believe that the modeling results under real conditions (although LAI of 500m itself is uncertain) are more reliable, that is, the high underestimation is attributed to the problem of the model structure.

In the ERF model, is it possible to output the importance of different models during the estimation process?

REPLY: Thanks for your comments. As mentioned above, the results of the ensemble model do not depend on a single GPP simulation.

Section 4.2: Supplementing the spatial distribution of product uncertainty is recommended.

REPLY: Thanks for your comments. According to your comments, we have added the spatial distribution of the uncertainty of ERF_GPP. The uncertainty of ERF_GPP mainly comes from two aspects, one is the influence of the number of GPP observations, and the other is the influence of the number of features (that is, the simulated GPP) used in the modeling process. For the first uncertainty, we randomly selected 80% of the data to build a model and simulate the multi-year average of global GPP. The process was repeated 100 times, and 100 groups of multi-year averages of ERF_GPP were obtained. Their standard deviations were considered to be the uncertainty of ERF_GPP caused by the number of GPP observations. For the second uncertainty, we choose different number of features to build models and simulate the multi-year average of global GPP. A total of 56 groups of multi-year averages of ERF_GPP are obtained. The standard deviation of different combinations is considered to be the uncertainty of ERF_GPP caused by the number of features. R10 and R11 show two types of uncertainty of ERF_GPP, similar to the spatial distribution, and ERF_GPP shows high uncertainty in the tropical regions, which has been reported in previous studies. There are very few observations of flux in these regions, both in terms of annual totals and long-term trends, and tropical regions are currently the most controversial areas in global GPP estimates. In addition, the problem of cloud pollution in remote sensing data in the tropics is well known, which further exacerbates the uncertainty in GPP estimates for the regions. In the revision, we have added a description of the relevant content and discussed it.

2.5 Global GPP estimation based on ERF model and its uncertainty.

Based on the ERF model, we estimated global GPP for 2001-2022 (ERF_GPP). The uncertainties of ERF_GPP can be attributed to two primary factors, one is the influence of the number of GPP observations, and the other is the influence of the number of features (that is, the simulated GPP). For the first type of uncertainty, we randomly selected 80% of the data to build a model and simulate the multi-year average of global GPP. The process was repeated 100 times, yielding 100 sets of multi-year averages of ERF_GPP. Their standard deviations were considered as the uncertainty of ERF_GPP caused by the number of GPP observations. For the second type of uncertainty, we selected different number of features to build a model and simulate the multi-year average of global GPP. A total of 56 sets of multi-year averages for ERF_GPP were obtained. The standard deviation of different combinations was considered to be the uncertainty of ERF_GPP caused by the number of features.

The results of the two uncertainty analyses consistently indicated that ERF_GPP exhibited a high uncertainty in tropical regions (Figures S6 and S7), and the uncertainty of ERF_GPP caused by the number of GPP observations was relatively small, the standard deviation of 100 simulations was about 0.3 gC m^{-2} d^{-1} in the tropics and lower in other regions, below 0.1 gC m^{-2} d^{-1}. In contrast, the ERF_GPP caused by the number of features was much more uncertain, especially when the number of features was small. It is worth noting that when the number of features was 5, the uncertainty was already substantially less, and the standard deviation was generally lower than 0.5 gC m^{-2} d^{-1}.

ERF_GPP exhibited high uncertainty in tropical regions, similar reports have been made in previously published GPP datasets (Badgley et al., 2019; Guo et al., 2023). The scarcity of flux observations in these regions (Pastorello et al., 2020), coupled with the well-known issue of cloud pollution and saturation in remote sensing data in the tropics (Badgley et al., 2019), exacerbates the uncertainty in GPP estimates for these regions. Therefore, in future studies, on the one hand, more flux observations in tropical regions are needed, and on the other hand, attempts can be made to combine optical and microwave data to improve the estimation of GPP.

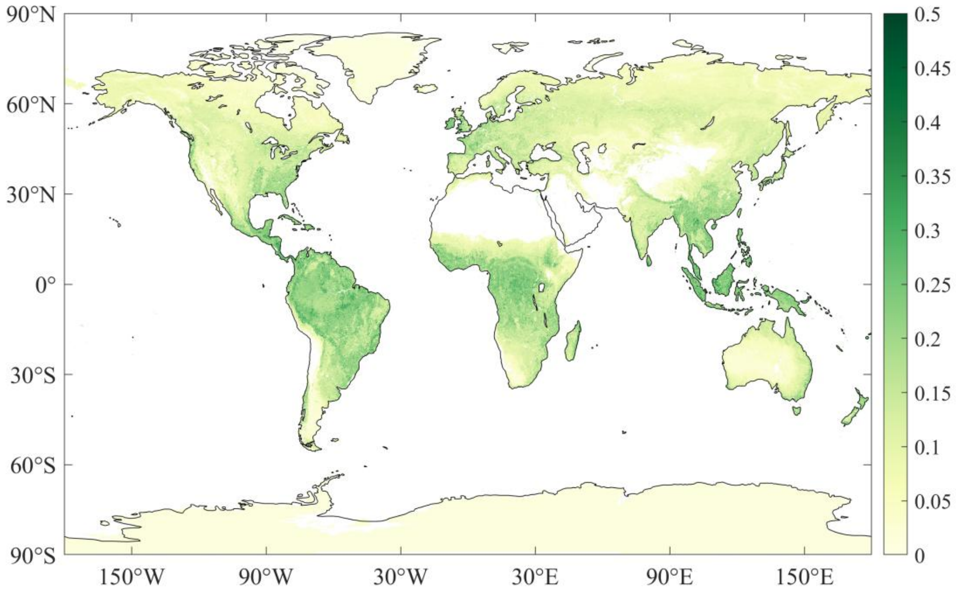


Figure R10. Uncertainty of ERF_GPP caused by the number of GPP observations.

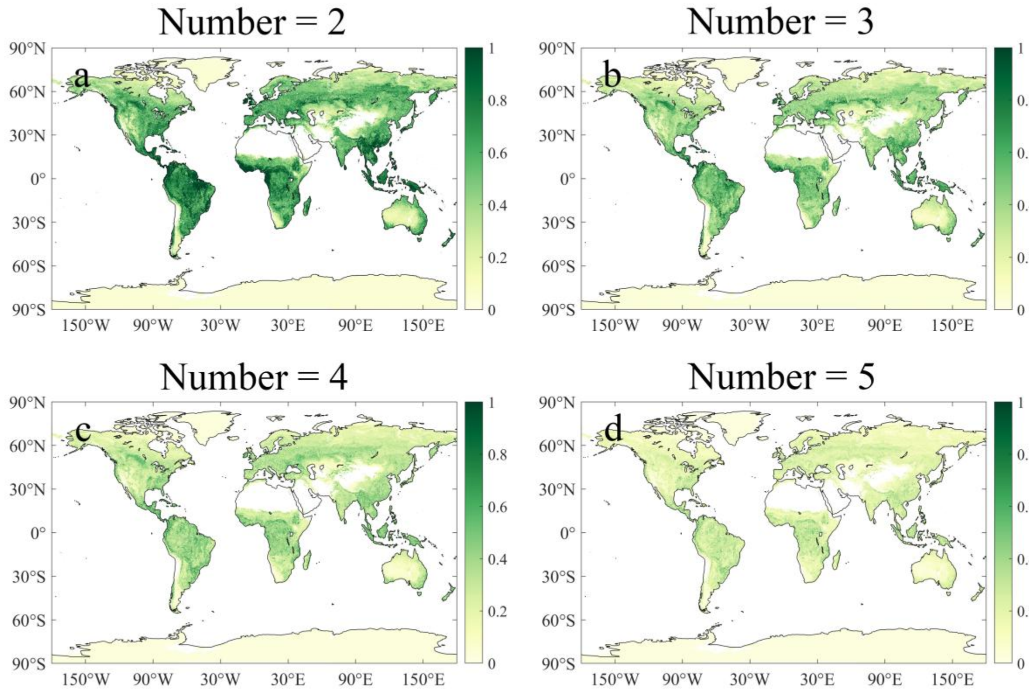


Figure R11. Uncertainty of ERF_GPP due to the number of features (simulated GPP).