*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.

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.

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.

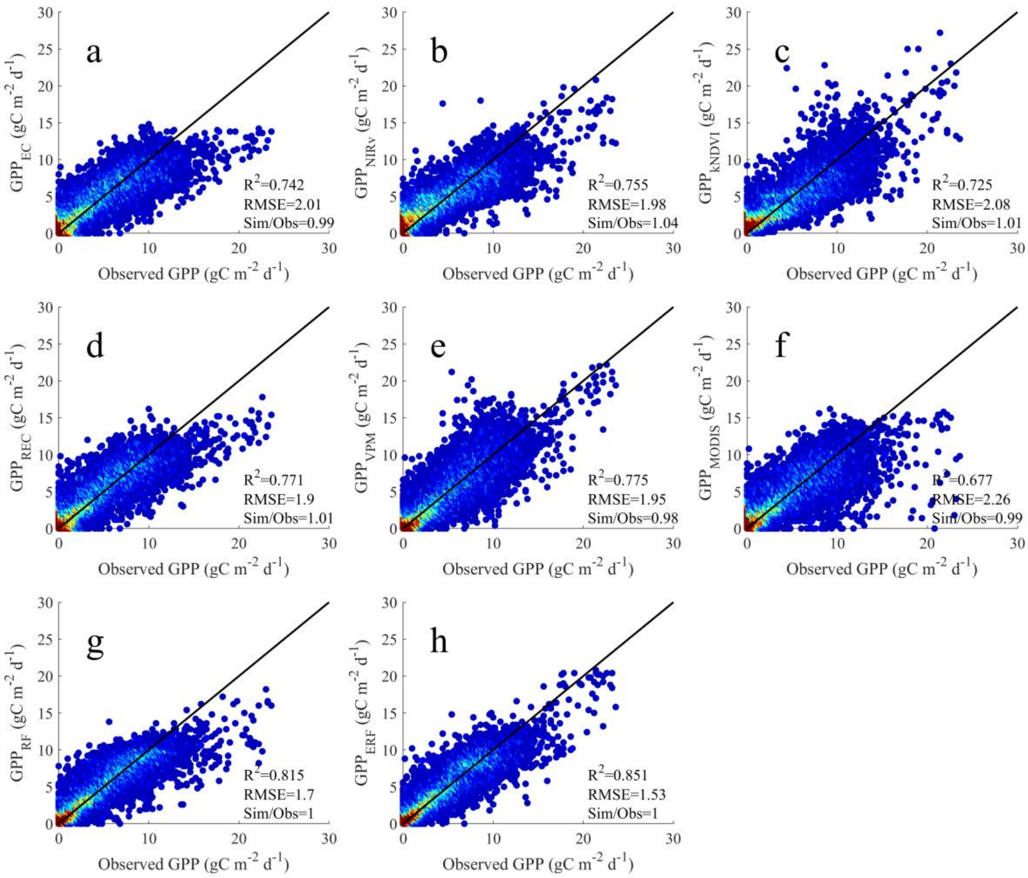


Figure R1. Comparison between the GPP simulations of the eight models and the GPP observations. a-h represents $GPP_{EC}$, $GPP_{NIRv}$, $GPP_{kNDVI}$, $GPP_{REC}$, $GPP_{VPM}$, $GPP_{MODIS}$, $GPP_{RF}$, $GPP_{ERF}$, respectively.

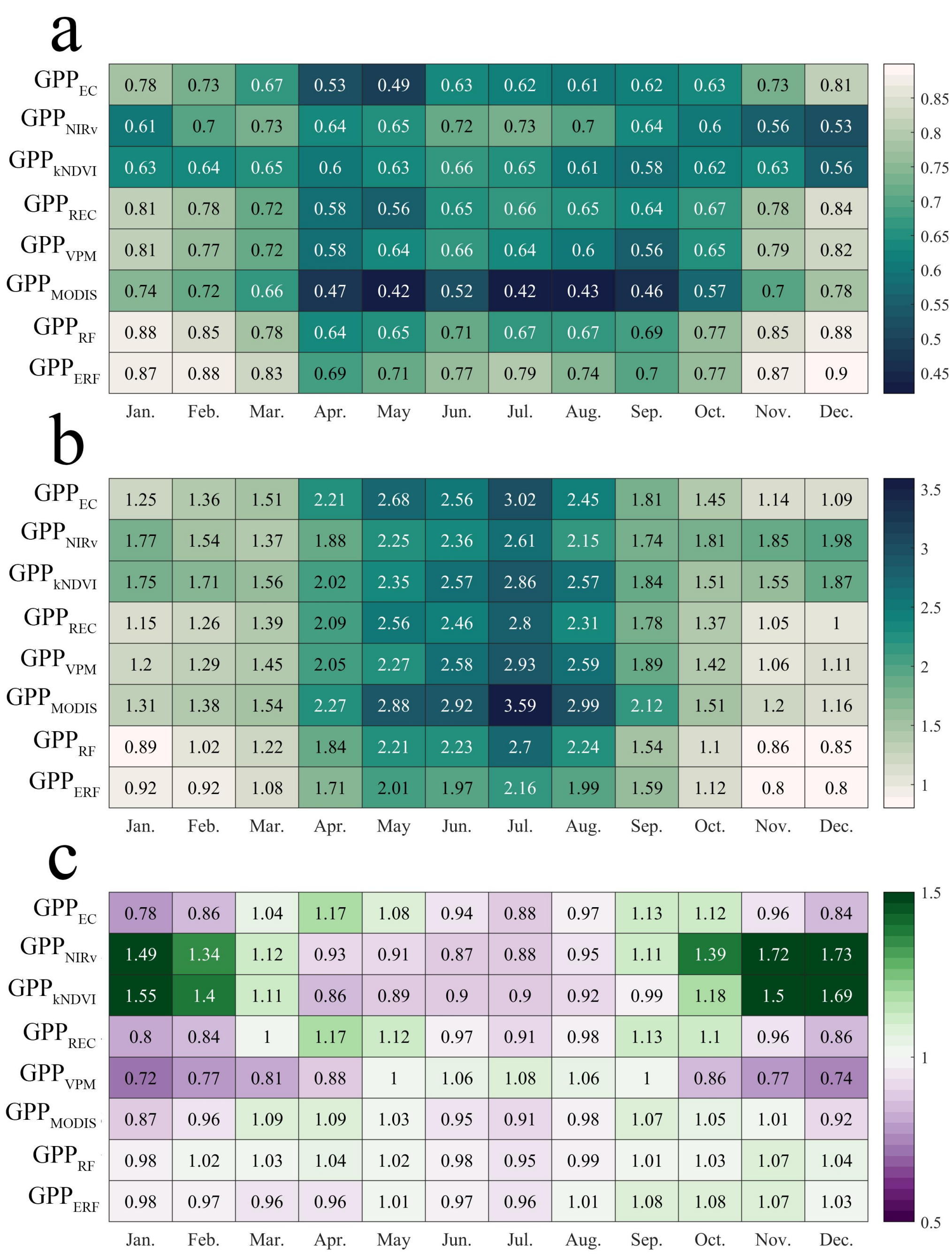


Figure R2. Performance of the eight models in each month. a, b and c represent $R^2$, RMSE, and Sim/Obs respectively.

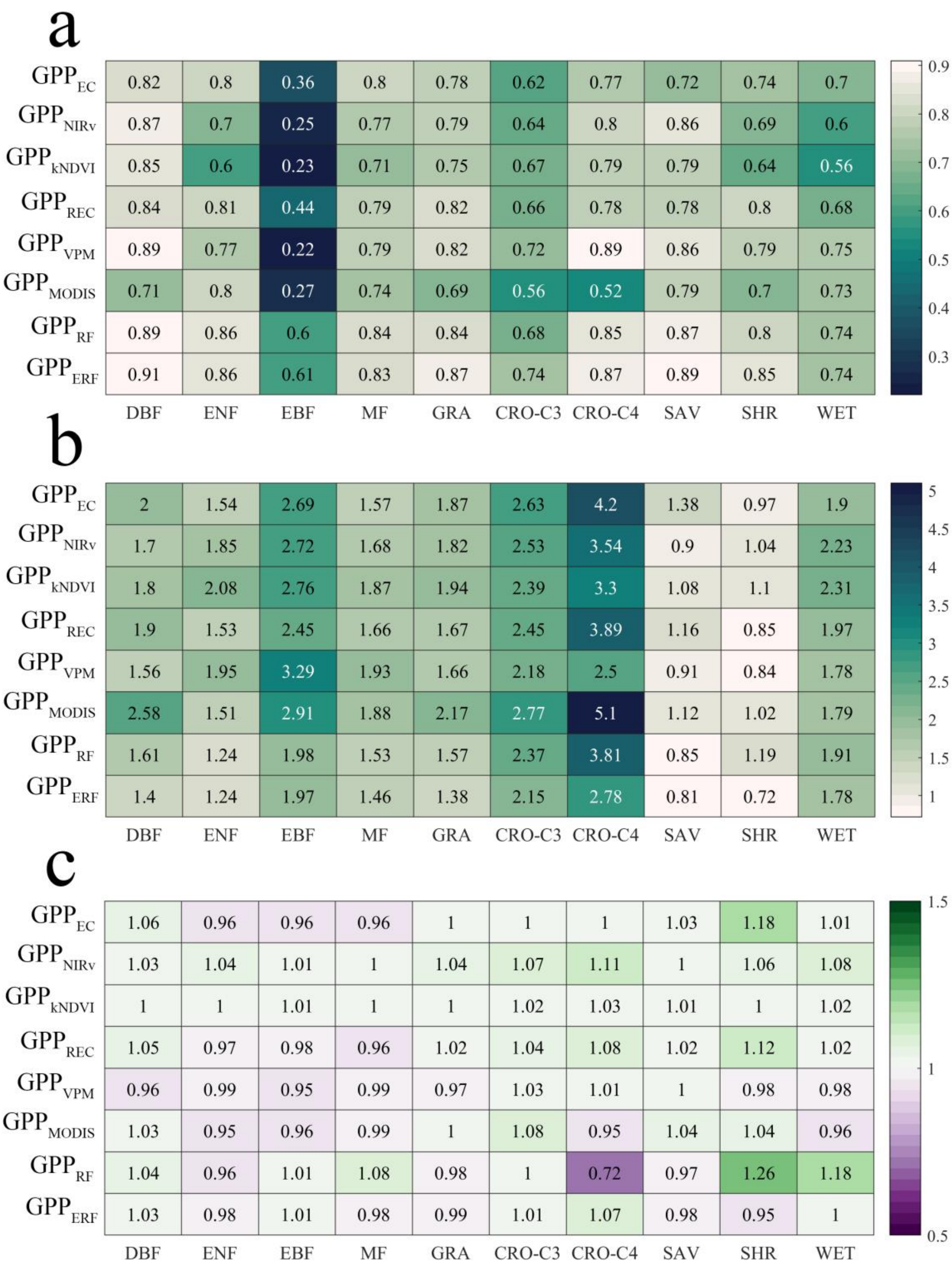


Figure R3. The performance of the eight models on different vegetation types. a, b and c represent $R^2$, RMSE, and Sim/Obs respectively.

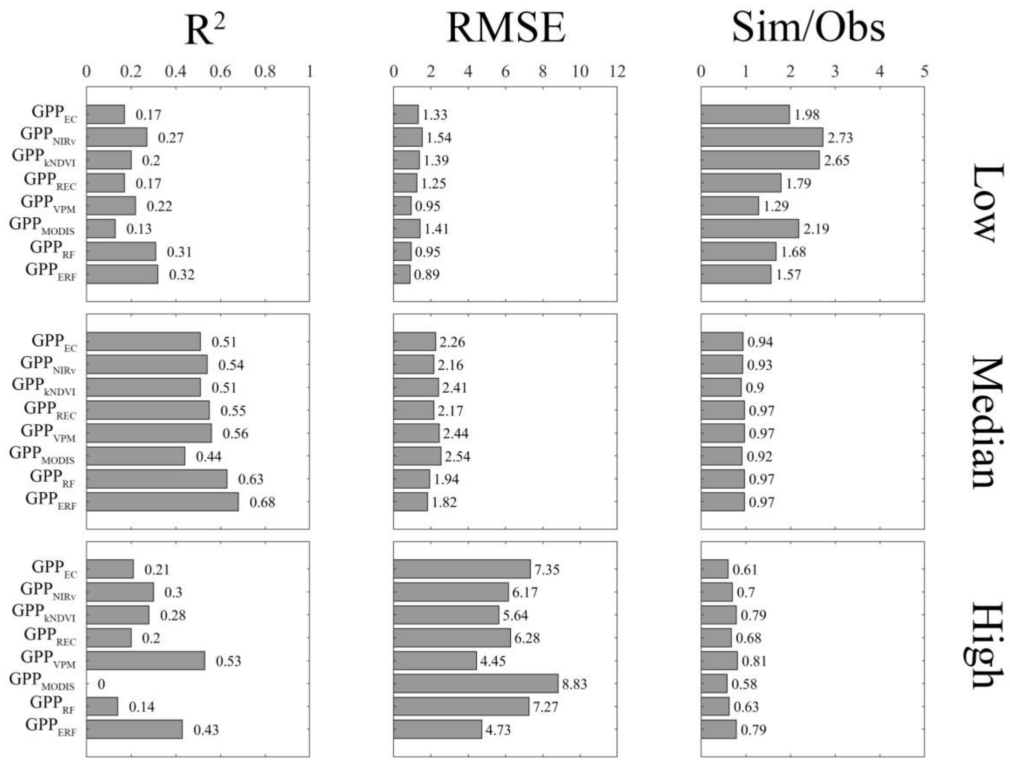


Figure R4. Performance of eight models in different subvalues.

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.

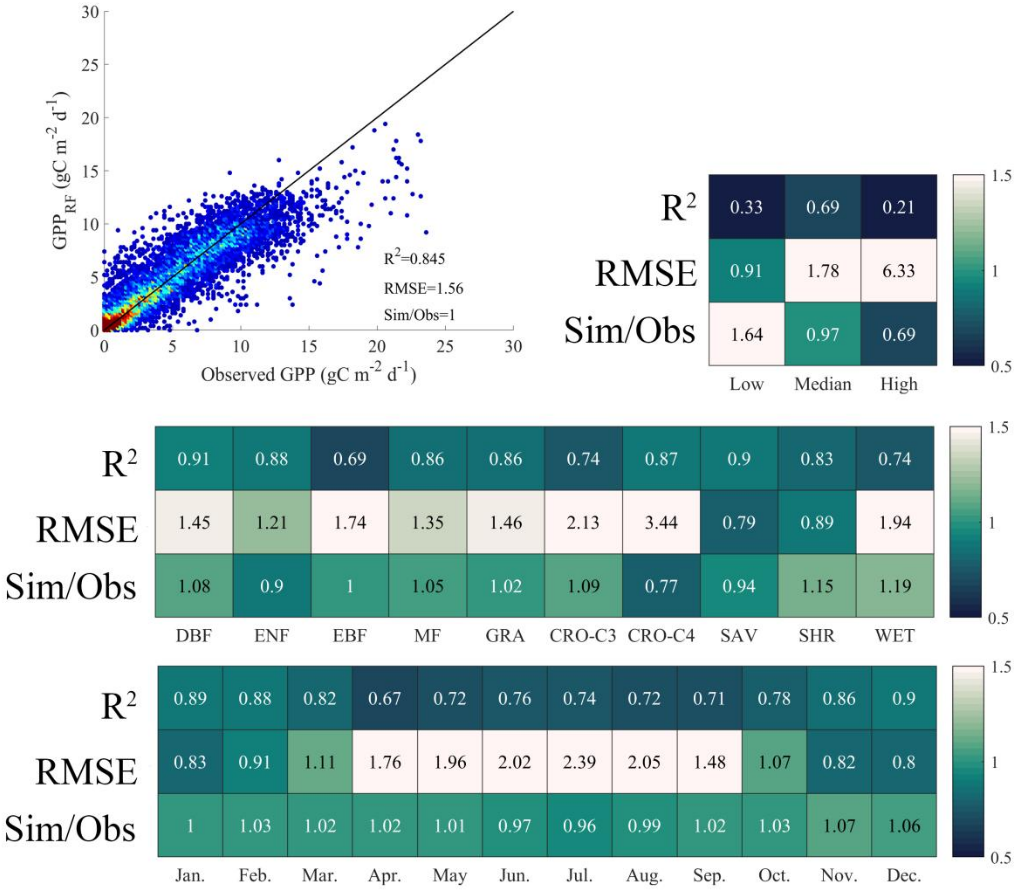


Figure R5. Performance of the random forest model using 9 variables.

*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 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.

*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 only the uncertainties of 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. On this basis, our research is further extended. 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 results show that the ERF model had 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.
*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.

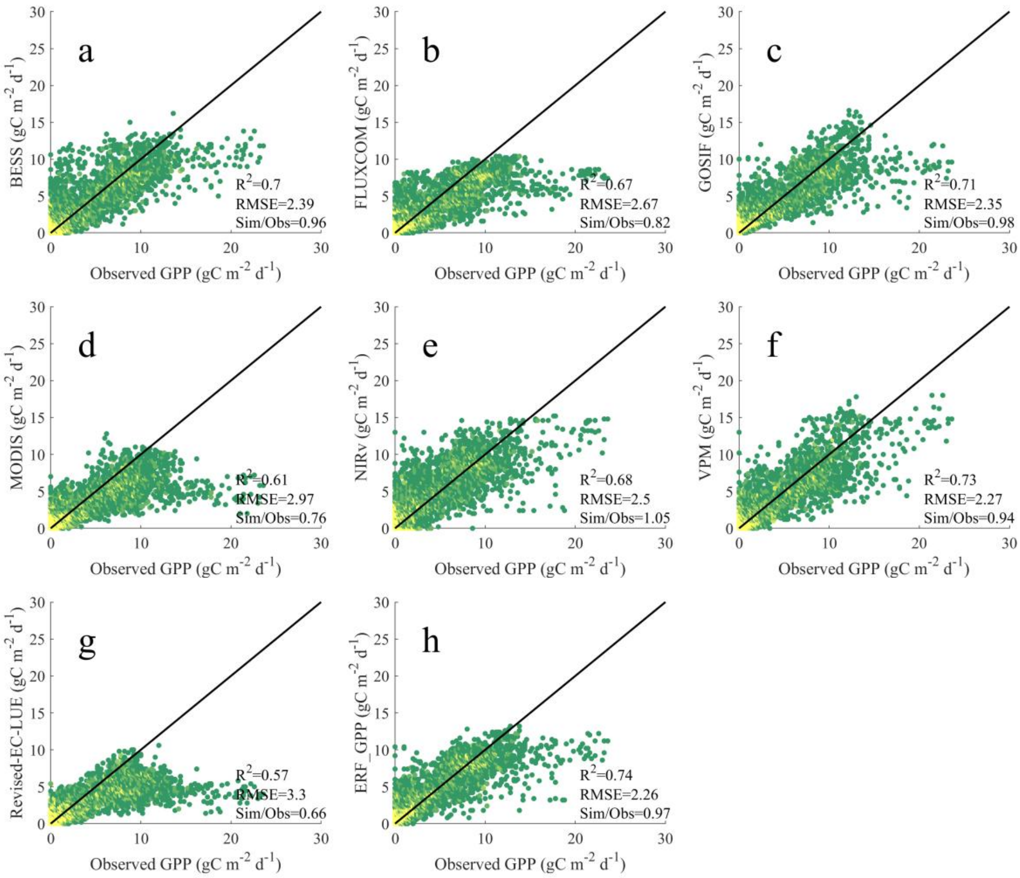

Figure R6. Comparison between the GPP datasets and the GPP observations from FLUXNET. a-h represents BESS, FLUXCOM, GOSIF, MODIS, NIRv, VPM, Revise-EC-LUE, ERF_GPP, respectively.

*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).

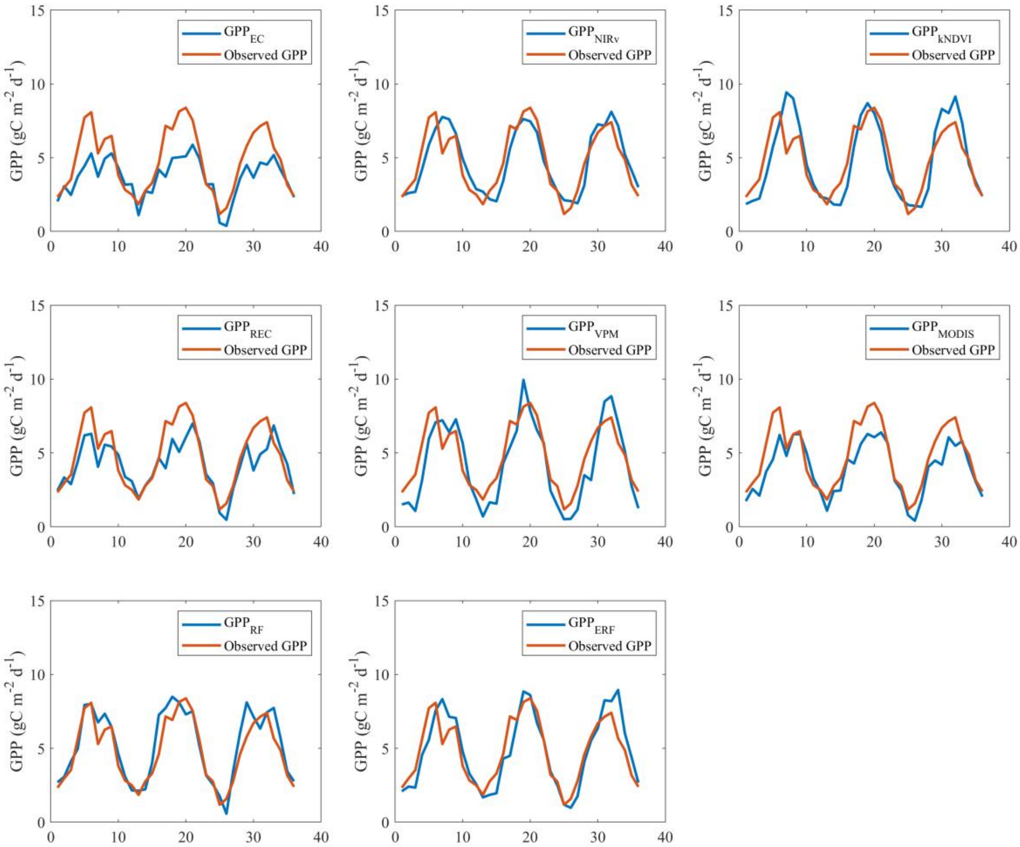


Figure R7. Performance of each GPP model on CN-Qia.

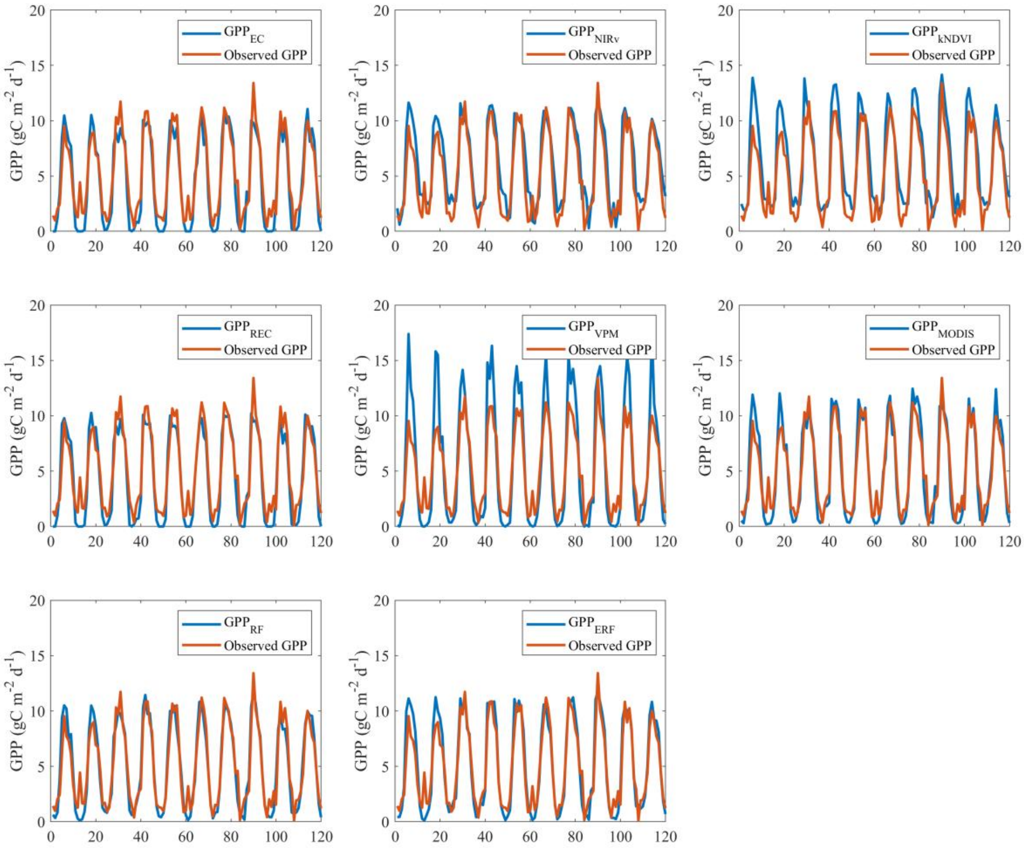


Figure R8. The performance of each GPP model on CH_Lae.

*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 $GPP_{MODIS}$, which had the highest relative importance, was no more than 25%, so it looks more like a weighted average of multiple GPP simulations.

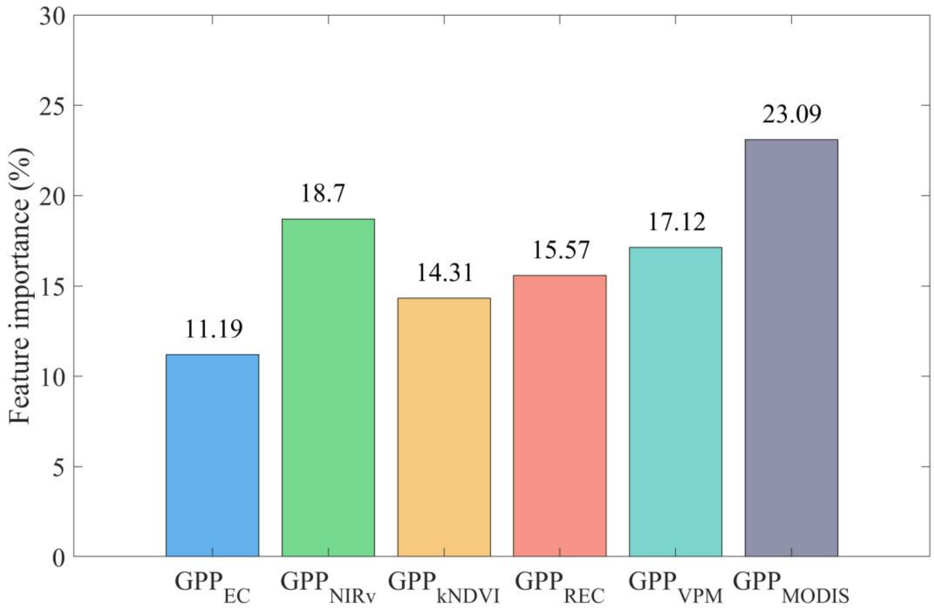


Figure R9. Average of 200 feature importance in the ERF model.

*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 site-scale models, we estimated global GPP for 2001-2022 using ERF model (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.

The results of the two uncertainty analyses consistently show that ERF_GPP presents a high uncertainty in the tropical region (Figure S6 and S7), and the uncertainty of ERF caused by the number of GPP observations is relatively small, the standard deviation of 100 simulations is 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 caused by the number of features is much more uncertain, especially if the number of features is small. It is worth noting that when the number of features is 5, the uncertainty is already substantially less, and the standard deviation is generally lower than 0.5 gC $m^{-2}$ $d^{-1}$.

ERF_GPP showed high uncertainty in the tropical regions, similar reports have been made in previously published GPP datasets (Badgley et al., 2019; Guo et al., 2023). There are very few observations of flux in these regions, so both in terms of annual totals and long-term trends, 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 (Badgley et al., 2019), which further exacerbates the uncertainty in GPP estimates for the regions.

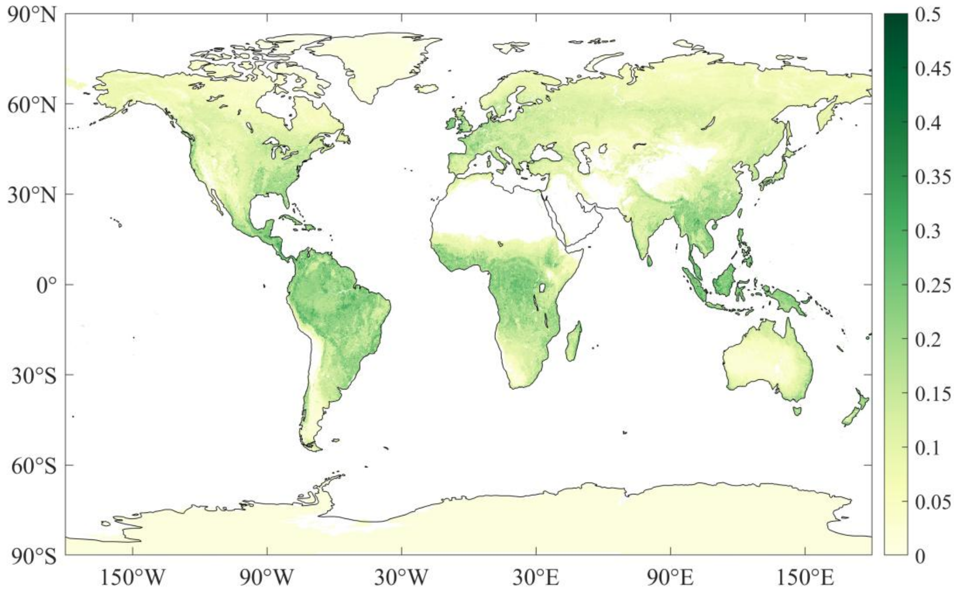


Figure R10. Uncertainty of ERF_GPP caused by the number of GPP observations.

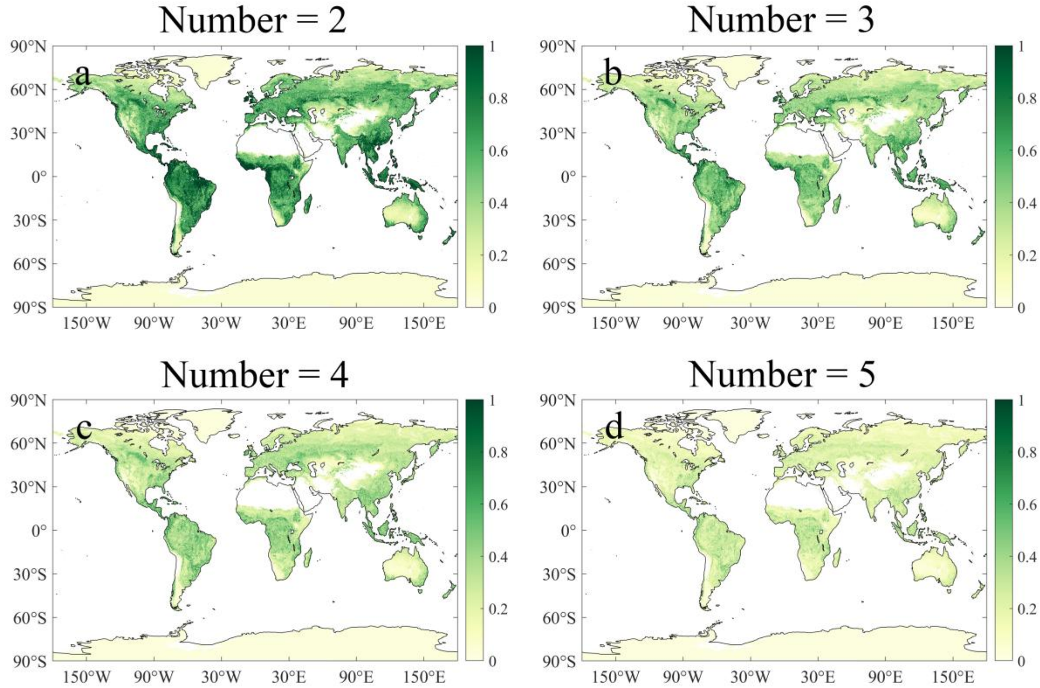


Figure R11. Uncertainty of ERF_GPP due to the number of features (simulated GPP).