Reply to Reviewers

Dear Reviewers,

Thank you for taking the time to review our submission titled "Evaluation of annual trends in carbon cycle variables simulated by CMIP6 Earth system models in China" to *Geoscientific Model Development*. We are grateful for your feedback and suggestions, which have significantly strengthened the manuscript. All the comments have been carefully considered. We believe that the revisions being made have addressed the concerns raised by the reviewers. We hope it meets the standards for publication in *Geoscientific Model Development*.

Below, we provide point-to-point replies. Each of reviewers' comments is first presented, followed by our reply and changes in the paper. In the revised manuscript, all the changes are marked in blue. Thank you once again for the time and expertise in reviewing our manuscript.

Sincerely,

Ziyang Li

Reviewer 1:

General Comment:

This manuscript presents a valuable and timely assessment of the performance of CMIP6 Earth System Models (ESMs) in simulating key carbon cycle variables (LAI, GPP, NPP, NEP, LST) over China during the historical period. The systematic comparison against satellite-derived observational datasets provides crucial insights into model biases and uncertainties. The study is well-structured, addresses a significant gap in evaluating regional ESM performance, and offers meaningful contributions towards understanding model limitations for future projections. The identification of spatial patterns of misestimation and the discussion linking deficiencies to model processes are particularly insightful. Overall, this is a solid piece of research with important implications for the carbon cycle modeling community.

While the manuscript is strong in its current form, I have several suggestions aimed at enhancing its comprehensiveness, clarity, and robustness:

Reply:

Thanks for the comprehensive and professional comments. They are extremely helpful and beneficial for improving our paper. Below, we have carefully prepared point-to-point replies for the comments, particularly about the influences that affect the accuracy of the models. We hope that the updated manuscript can address all your concerns.

Specific Comments (1):

The analysis focuses effectively on interannual trends. However, the seasonal cycle is a fundamental aspect of ecosystem carbon dynamics and ESM performance. Could the authors provide some analysis or discussion regarding how well the evaluated CMIP6 models capture the seasonal patterns of carbon cycle indicators for terrestrial ecosystems across different regions of China?

Reply:

We thank the reviewer for raising this important point regarding the seasonal cycle in ecosystem carbon dynamics and model performance. We acknowledge that the seasonal cycle is indeed a fundamental aspect of terrestrial carbon cycling. However, the primary focus of this particular study is on the interannual trends and variability of the carbon cycle indicators across China. Our analysis and model evaluation framework were specifically designed to address questions related to long-term, year-to-year changes and, therefore, do not encompass a detailed assessment of seasonal patterns. We agree that evaluating how well CMIP6 models capture regional seasonal dynamics represents a valuable and complementary research direction. To address the reviewer's suggestion, we will explicitly include a discussion in the manuscript highlighting the importance of assessing seasonal cycles in future work to further improve our understanding of model capabilities in simulating phenological responses and intra-annual carbon fluxes across different regions of China.

Changes in the paper:

Line 4331-434

"However, it is important to note that while this study focused on evaluating model performance in capturing interannual trends and variability, the assessment of seasonal cycle dynamics (e.g., phenological timing, amplitude of seasonal fluctuations) across China's diverse regions represents a critical avenue for future research to further refine our understanding of model capabilities in simulating terrestrial carbon cycle processes."

Specific Comments(2):

Figure 3 effectively presents the evaluation of simulated interannual variability for LAI and GPP against observations. To ensure consistency and provide a comparable level of detail for all key variables, I recommend extending this type of analysis to NPP, NEP, and LST. Creating a new figure (or adapting Figure 4) to include Taylor diagrams (or similar metrics) assessing interannual variability for NPP, NEP, and LST, analogous to Figure 3, would greatly enhance the paper's completeness and allow for a direct comparison of variability performance across all five variables.

Reply:

We thank the reviewer for this excellent suggestion. We agree that applying a consistent evaluation framework across all key variables is crucial for a comprehensive and comparable assessment of model performance regarding interannual variability.

In direct response to this comment, we have now extended our analysis of interannual variability to include NPP, NEP, and LST, following the same methodology used for LAI and GPP in the original Figure 3. To present these results clearly and maintain a comparable level of detail: The Taylor diagrams evaluating the interannual variability of NPP and NEP have been incorporated into a revised Figure 3 (which now also includes LAI and GPP). The Taylor diagram evaluating the interannual variability of LST has been incorporated into a revised Figure 4 (alongside its temporal trend).

Changes in the paper:

Line 267-270

"Figure 3: Overall annual average (a) NPP and (c) NEP trends in China during 2003-2014. The asterisk (*) indicates the significant trend ($p<0.05$). The Taylor diagrams compare the remotely-sensed and model-simulated annual mean (b) NPP and (d) NEP for the historical period (2003-

2014). The standard deviation shows the interannual variability of the observed and simulated LAI. The dashed green lines show centered root mean square difference (RMSD) between model simulations and satellite observations."

Line 271-277

"The Taylor diagram (Fig. 3b, 3d) evaluates interannual variations of NPP and NEP across individual models, the MME, and observations. The SD of ESM-simulated NPP ranges from 7.70 (ACCESS-ESM1-5) to 24.95 (BCC-CSM2-MR), while RMSD values span from 8.82 (INM-CM4-8) to 25.66 (INM-CM5-0). For NEP, SD ranges from 9.19 (CanESM5) to 20.24 (BCC-CSM2-MR) with RMSD values between 13.55 (IPSL-CM6A-LR) and 21.41 (BCC-CSM2-MR). The MME demonstrates notably lower variability, with SD values of 5.65 for NPP and 6.19 for NEP, and RMSD values of 9.71 for NPP and 11.28 for NEP. Generally, the MME-simulated SD and RMSD for both variables are lower than those of all individual models."

Line 284-287

"Figure 4: (a) Overall annual average LST trends in China during 2003-2014. The asterisk (*) indicates the significant trend ($p<0.05$). (b) The Taylor diagrams compare the remotely-sensed and model-simulated global annual mean LST for the historical period (2003-2014). The standard deviation shows the interannual variability of the observed and simulated LAI. The dashed green lines show centered root mean square difference (RMSD) between model simulations and satellite observations."

Line 288-291

"The Taylor diagram evaluates interannual variations of LST across individual models, the MME, and observations (Fig. 4). The SD of ESM-simulated LST ranges from 0.20 (INM-CM4-8) to 0.48 (IPSL-CM6A-LR), while RMSD values span from 0.30 (MPI-ESM1-2-HR) to 0.59 (IPSL-CM6A-LR). The MME demonstrates notably lower variability and error, with an SD of 0.10 and RMSD of 0.36."

Specific Comments(3):

Presenting the observed interannual trends for the satellite datasets (LAI, NPP, NEP, LST, CSIF-derived GPP) is essential context. However, to streamline the main manuscript narrative and focus it more directly on the model evaluation results and analysis, I suggest moving the detailed presentation of the observed trend maps (e.g., Figure 1 or its equivalent) to the Supplementary Material. The main text can then succinctly summarize the key features of the observed trends before delving into the model comparison. This would improve the flow and conciseness of the core results section.

Reply:

We thank the reviewer for this constructive suggestion to improve the narrative flow and conciseness of the manuscript. We agree that focusing the main text more directly on the model evaluation results enhances the clarity of our core analysis. Accordingly, we have moved the detailed observational trend maps for all key variables (specifically, the original Figures 2, 9, and 10, and Table 2) to the Supplementary Material. The main text now provides a succinct summary of the key observed trends before delving into the model comparison, which allows readers to grasp the essential context without interrupting the primary narrative focused on the model assessment.

Changes in the paper:

Add Table S1, Figures S1-S8 to Supplement Materials.

Line 210-227

"From 2003 to 2014, most regions across China exhibited increasing trends in MODIS LAI (82.60% of the area; Fig. S1) and CSIF (86.50%; Fig. S2), with significant increases covering 44.10% for LAI and 48.34% for CSIF concentrated in southern forests, northeastern China, and the Loess Plateau. Notably, CSIF demonstrated stronger rising trends than LAI in the North China Plain. MODIS NPP trends (Fig. S3) showed minimal significant changes, as 84.57% of the area exhibited non-significant variability, while significant increases and decreases covered only 15.23% and 0.20% respectively. These patterns closely mirrored MODIS NEP trends (Fig. S4), where 85.07% of the area showed non-significant changes with merely 0.28% significant decreases. For MODIS LST during 2003–2019 (Fig. S5), 47.34% and 52.66% of the study area exhibited increasing and decreasing LST trends, respectively, yet 97.89% of the area showed non-significant changes, with significant warming limited to 0.78%.

Spatially coherent trends emerged across northwestern China (Xinjiang grasslands/croplands, central Inner Mongolia, Tibetan Plateau), where declining LAI and CSIF aligned with significant LST increases, while vegetation productivity declines were also observed in parts of southern Tibet. Conversely, northeastern China consistently exhibited significant increases in LAI, CSIF, NPP, and NEP, coinciding with declining LST trends. The Loess Plateau (northern Shaanxi/Ningxia) similarly demonstrated concurrent increases in LAI, CSIF, NPP, and NEP. Southern forested regions showed LAI and CSIF gains, though central Yunnan experienced notable CSIF reductions alongside NPP and NEP declines. Contrasting patterns characterized eastern China: the North China Plain had significant CSIF increases but prominent CSIF reductions later, alongside NEP declines and LST warming, while southeastern provinces (Guangdong, Fujian) and the Yangtze River Delta featured NPP and NEP reductions alongside localized CSIF decreases."

Specific Comments(4):

Section 4.3, "Uncertainty of the observed LAI and CSIF in China," rightly addresses uncertainties in these specific observational products. However, the study relies on multiple satellite-derived datasets (LAI, NPP, NEP, LST, CSIF). To provide a more comprehensive assessment of uncertainty sources affecting the benchmark itself, the discussion in this section should be expanded to explicitly consider the uncertainties associated with all the primary observational datasets used (NPP, NEP, and LST, in addition to LAI and CSIF).

Reply:

We sincerely thank the reviewer for this insightful suggestion. We agree that a comprehensive discussion of uncertainties across all primary observational datasets is crucial for a robust interpretation of the model evaluation results. In direct response to this comment, we have significantly expanded Section 4.3. The section has been retitled to reflect its broader scope (e.g., "Uncertainties in the Observational Benchmark Datasets"), and the discussion now explicitly incorporates a detailed analysis of the uncertainties associated with the MODIS NPP, NEP, and LST products, in addition to the existing analysis of LAI and CSIF-derived GPP uncertainties.

Changes in the paper:

Line 464-475

"4.3 Uncertain of the observed datasets in China

Accurate observational data are essential for determining and improving the precision of

models (Luo et al., 2016). In this study, two remote sensing datasets that better reflect actual vegetation growth, namely the reprocessed MODIS dataset and the CSIF dataset, were selected for evaluation. However, satellite remote sensing data in China are subject to considerable uncertainty due to various factors. High cloud cover during the rainy season and snow cover in high-latitude areas during winter can introduce inaccuracies. Additionally, satellite sensors are prone to degradation over time, leading to reduced sensitivity. Although the reprocessed MODIS and CSIF datasets utilize spatiotemporal filtering and machine learning techniques to enhance data quality (Zhang et al., 2018; Yuan et al., 2011), significant uncertainties remain. Similarly, limitations in satellite remote sensing-based carbon accounting (Araza et al., 2023), contribute to substantial uncertainties in MODIS NPP and NEP products (Sun et al., 2021; Huang et al., 2018; Ma et al., 2016). Concurrently, studies attribute uncertainties in MODIS LST primarily to spatial inconsistencies and surface emissivity uncertainties, the latter resulting from inadequate global representativeness in land cover classification (Ma et al., 2021; Wan et al., 2002; Duan et al., 2019).”

Specific comments(5):

The analysis focuses on the period 2003-2014. Could the authors please provide a more explicit justification for selecting this specific timeframe? Clarifying the rationale is important for interpreting the results and their broader applicability.

Reply:

We appreciate the reviewer's suggestion for clarifying the rationale behind our selected timeframe (2003-2014). The start year of 2003 was chosen primarily to ensure the stability and consistency of the MODIS sensor data used in our analysis. Following the launch of the MODIS Aqua satellite in October 2002, which formed the dual-satellite constellation (Terra and Aqua) essential for improved temporal coverage, we avoided the immediate post-launch period (2002) to mitigate potential transient effects or calibration uncertainties that could introduce abrupt changes in the vegetation monitoring record. Therefore, 2003 marks the beginning of a stable, dual-sensor era for MODIS data. The end year of 2014 aligns with the termination point of the historical experiment simulations provided by the CMIP6 models, which constitute a core component of our comparative analysis. This timeframe selection ensures both the reliability of the benchmark remote sensing data and direct comparability with the available CMIP6 historical climate model outputs.

Changes in the paper:

Line 137-139

“The analysis period spans 2003 to 2014. This timeframe was selected to utilize the stable post-launch era of the MODIS Terra-Aqua dual-satellite constellation for vegetation monitoring, commencing after potential initial sensor calibration transients, and to align with the end year of the CMIP6 historical experiment simulations used for comparison.”

Technical corrections(1):

On page 2, line 33, “...regulating stomatal conductance, to effectively respond to rising atmospheric CO_2 concentrations…” should be “...regulating stomatal conductance to effectively respond to rising atmospheric CO_2 concentrations…". Please check and change accordingly.

Reply:

Thanks for pointing this mistake out. We revised the sentence to correct the mistake.

Changes in the paper:

Line 32-34

"Vegetations maximize water-use efficiency at the leaf scale by dynamically regulating stomatal conductance to effectively respond to rising atmospheric CO_2 concentrations and global warming (Fu et al., 2022)."

Technical corrections(2):

On page 5, line 112, "where the inter-model errors is mitigated through model averaging, thereby amplifying underlying true signals." should be "where the inter-model errors are…". Please check and change accordingly.

Reply:

Thanks for pointing this mistake out. We revised the sentence to correct the mistake.

Changes in the paper:

Line 129-130

"This approach combines the true and noise signals, where the inter-model errors are mitigated through model averaging, thereby amplifying the underlying true signals."

Technical corrections(3):

On page 20, line 350, "(Anav et al., 2013; Song et al., 2021)," should be "(Anav et al., 2013; Song et al., 2021).". Please check and change accordingly.

Reply:

Thanks for pointing this mistake out. We revised the sentence to correct the mistake.

Changes in the paper:

Line 366-367

"Previous studies have found that, whether for CMIP5 or CMIP6, models have not effectively captured the long-term trends of vegetation (Anav et al., 2013; Song et al., 2021)."

Technical corrections(4):

On page 23, line 427, "tree cover prcentage" should be "tree cover percentage". Please check and change accordingly.

Reply:

Thanks for pointing this mistake out. We revised the sentence to correct the mistake.

Changes in the paper:

Figure S8

"Figure S8: Evaluation of the tree cover percentage trend performance of the CMIP6 ESM from 2003 to 2014 in China."

Reviewer 2:

Dear Prof. Li,

Thank you sincerely for dedicating your time to provide the Community Comment on our submission titled "Evaluation of annual trends in carbon cycle variables simulated by CMIP6 Earth system models in China" to Geoscientific Model Development. Your insights and feedback are highly valued, as they have offered critical guidance to further refine and strengthen the scientific rigor of our manuscript.

We have carefully and thoroughly reviewed every point raised in your comment, and have made targeted revisions to address the concerns you noted—we are confident these adjustments have enhanced the quality and clarity of the work. It is our sincere hope that the revised manuscript now aligns well with the publication standards of Geoscientific Model Development.

Below, we present a detailed point-to-point response to your comment: first, we restate your original comment for clarity, followed by our specific reply and a description of the corresponding revisions made to the paper. For ease of reference, all modifications in the revised manuscript are highlighted in blue.

Once again, we would like to express our gratitude for your professional expertise and the time you have invested in reviewing our work.

Sincerely,

Ziyang Li

Overview

Leaf Area Index (LAI), Gross Primary Productivity (GPP), Net Primary Productivity (NPP), Net Ecosystem Productivity (NEP) and Land Surface Temperature (LST) as key indicators of carbon cycle in terrestrial ecosystems. In this paper, this research provides a quite interesting question about the interannual trends performance of LAI, GPP, NPP, NEP and LST simulated by 12 CMIP6 Earth System Models.

This study is significant as it provides information on how Earth System Models fail to capture the trends of CMIP6 ESM-simulated variables in China. The manuscript falls within the scope of the journal and should be considered for publication subject to minor revisions.

General comment (1)

In the introduction, the author does not clearly present the research significance, particularly the practical relevance of evaluating five indices from CMIP6 simulations.

Reply:

We thank the reviewer for this valuable comment. In response, we have expanded the Introduction to provide a clearer and more detailed explanation of the scientific definitions and practical significance of the five key indicators (LAI, GPP, NPP, NEP, and LST) evaluated from the CMIP6 simulations. These additions highlight the practical relevance of our study in understanding ecosystem carbon dynamics and model performance, thereby strengthening the motivation and context of our research.

Changes in the paper:

Lines 47-61

"The LAI, defined as the total one-sided leaf area per unit ground surface area, is a key

parameter of vegetation canopy structure that directly influences light interception, transpiration, and the spatial heterogeneity of GPP. GPP refers to the total amount of carbon dioxide fixed into organic compounds by vegetation through photosynthesis, serving as a core indicator of an ecosystem's carbon sequestration capacity. NPP represents the net carbon accumulation after subtracting autotrophic respiration from total photosynthetic fixation, reflecting the primary production potential and health of ecosystems as influenced by GPP and plant physiological regulation. NEP denotes the net carbon exchange between the ecosystem and the atmosphere by further subtracting heterotrophic respiration from NPP, making it a crucial measure for assessing regional carbon source/sink status under the influence of environmental factors such as atmospheric CO_2 and climate (Fang et al., 2001). LST is the thermodynamic temperature at the land-atmosphere interface, playing a key role in surface energy and water exchange while jointly affecting ecological processes through interactions with solar radiation, soil properties, vegetation, and atmospheric conditions (Li et al., 2023). In-depth research on LST facilitates a deeper understanding of surface-atmosphere exchange processes at global and regional scales and provides high-quality quantitative indicators of surface conditions for scientific applications. Consequently, LST has been designated as an indispensable observation indicator for the International Geosphere and Biosphere Program (IGBP) and the Global Climate Observing System (GCOS) (Townshend et al., 1994; Hollmann et al., 2013).”

General comment (2)

Among the 12 CMIP6 models evaluated, which one performs best?

Reply:

Our comprehensive evaluation of 12 CMIP6 models reveals that no single model performs consistently best across all variables (LAI, GPP, NPP, NEP, LST) and regions in China. It is noteworthy that while certain models may appear to capture the overall interannual trend of a specific variable reasonably well at the aggregate scale, this apparent skill often masks substantial and widespread spatial biases, where large geographical areas exhibit significant misestimation of both the magnitude and direction of the trend. This discrepancy between a seemingly accurate broad-scale trend and poor spatial realism is a common limitation among the models. Consequently, we caution against endorsing any single model as universally reliable. Instead, despite the noted performance limitations of the Multi-Model Ensemble mean (MME) across China's complex landscapes, we still recommend its use as a pragmatic approach to balance out individual model biases and provide a more conservative projection for assessing regional vegetation and climate dynamics.

Changes in the paper:

Lines 125-129

“To provide a more robust and integrated assessment than any single model can offer, we primarily utilize the Multi-Model Ensemble (MME) mean, calculated as the arithmetic mean of all available models, as a central benchmark for evaluating the simulated spatial patterns and interannual variability of vegetation and climatic variables over China. The MME mean-based approach, an established statistical integration technique, synthesizes outputs from diverse models through averaging (Zeng et al., 2016).”

General comment (3)

It is recommended to evaluate an ensemble model.

Reply:

We thank the reviewer for raising this important point regarding ensemble model evaluation. We fully agree with the recommendation, and we would like to clarify that the core of our analysis is precisely built upon the systematic evaluation of a Multi-Model Ensemble (MME) mean, which is constructed from the simple average of all 14 available model outputs. The performance of this MME for all key ecosystem carbon cycle variables (LAI, GPP, NPP, NEP and LST) across China is not only a central focus of our results but also forms the primary basis for our main conclusions. We contend that the MME serves a dual purpose: it functions both as a robust integrated benchmark that collectively represents the current modeling capability and as a diagnostic tool that reveals common biases and uncertainties, thereby providing a critical reference point for assessing the overall model performance and guiding future model development.

Changes in the paper:

Line 125-129

"To provide a more robust and integrated assessment than any single model can offer, we primarily utilize the Multi-Model Ensemble (MME) mean, calculated as the arithmetic mean of all available models, as a central benchmark for evaluating the simulated spatial patterns and interannual variability of vegetation and climatic variables over China. The MME mean-based approach, an established statistical integration technique, synthesizes outputs from diverse models through averaging (Zeng et al., 2016)."

General comment (4)

It is recommended that the author focus on specific regions (e.g., the North China Plain, the Pearl River Delta) and provide targeted suggestions for improving the simulations.

Reply:

We sincerely thank the reviewer for this constructive suggestion to enhance the practical value of our study. We agree that providing region-specific insights is crucial for guiding future model development. In our initial analysis, we indeed assessed model performance across several key regions, including the Tibetan Plateau, the Loess Plateau, northeastern China, the North China Plain, and southern China, as reflected in the spatial evaluation. However, we acknowledge that the conclusion section lacked targeted discussions for these specific areas. In direct response to your comment, we have now supplemented the conclusion with a focused analysis and specific recommendations for three representative regions exhibiting distinct model biases: the Tibetan Plateau (where models show a systematic overestimation), the Loess Plateau (characterized by a general underestimation), and the Pearl River Basin (where simulations show significant and complex errors).

Changes in the paper:

Lines 491-495

"The model simulations exhibit pronounced regional biases, including a systematic overestimation of ecological variables on the cold-arid Tibetan Plateau, a general underestimation over the extensively vegetated Loess Plateau likely linked to misrepresented ecological restoration processes, and widespread simulation errors in the rapidly urbanizing Pearl River Basin, potentially due to unaccounted-for anthropogenic pressures."

Reviewer 3:

Dear Reviewers,

Thank you sincerely for dedicating your time to provide the Community Comment on our submission titled “Evaluation of annual trends in carbon cycle variables simulated by CMIP6 Earth system models in China” to Geoscientific Model Development. Your insights and feedback are highly valued, as they have offered critical guidance to further refine and strengthen the scientific rigor of our manuscript.

We have carefully and thoroughly reviewed every point raised in your comment, and have made targeted revisions to address the concerns you noted—we are confident these adjustments have enhanced the quality and clarity of the work. It is our sincere hope that the revised manuscript now aligns well with the publication standards of Geoscientific Model Development.

Below, we present a detailed point-to-point response to your comment: first, we restate your original comment for clarity, followed by our specific reply and a description of the corresponding revisions made to the paper. For ease of reference, all modifications in the revised manuscript are highlighted in blue.

Once again, we would like to express our gratitude for your professional expertise and the time you have invested in reviewing our work.

Sincerely,

Ziyang Li

Overview

This manuscript evaluates annual trends in LAI, GPP, NPP, NEP, and LST over China using CMIP6 Earth system models. While the topic is relevant, the analysis does not provide new insight into the causes of model–observation discrepancies or pathways to reduce them. The contribution would be stronger with a clear attribution of biases and concrete recommendations for model improvement. In addition, the manuscript contains numerous grammatical errors and would benefit from thorough language editing.

General comment (1)

CMIP6 models simulate their own meteorology, which can diverge from observed climate. In contrast, TRENDY simulations are driven by observed forcings (e.g., CRUJRA). Please justify evaluating interannual trends with fully coupled CMIP6 output and clarify/discuss how much of the trend mismatch stems from (i) differences in simulated climate versus observations, vs. (ii) process parameterizations or structural choices. Current discussion on the model-observation mismatch is too vague (e.g. section 4.1).

The authors should also discuss whether the large inter-model spread arises primarily from differing simulated climate trajectories or from carbon-cycle parameterizations/structures.

Reply:

We thank the reviewers for these insightful comments regarding the respective roles of simulated climate versus model structural differences in driving the inter-model spread and trend mismatches. We agree that this is a critical aspect of interpreting our results.

In response, we have significantly expanded our discussion in Section 4.1 to address this point more concretely. We acknowledge that fully and precisely attributing the discrepancies to climatic

versus structural factors with the standard CMIP6 output is challenging, as it would require idealized model experiments. However, we have incorporated new analyses to provide clearer insights.

First, to evaluate the role of simulated climate, we now include an assessment of the core climate drivers (temperature, precipitation, and solar radiation) from the CMIP6 models against observational datasets (CRUNCEP, ERA5) using Taylor diagrams (Figure S9) and a summary table of biases (Table S2). This analysis quantitatively shows that the models exhibit substantial biases in their simulated climate over China, for instance, generally overestimating temperature while underestimating precipitation and radiation, with high RMSD values.

Second, to address the role of model structure and parameterization, we have synthesized information from key literature to create a new summary table (Table S3) detailing the key land surface components and related parameters for the selected ESMs. This table visually underscores the considerable diversity in model structures and parameterizations, which is a fundamental source of the inter-model spread in carbon cycle simulations.

While we cannot provide a precise quantitative partition, the new figures and tables allow for a more informed discussion, suggesting that both erroneous climatic trajectories and divergent carbon-cycle representations are significant and intertwined contributors to the overall model-observation mismatch and the large inter-model spread.

Changes in the paper:

Lines 167-172

"To verify the accuracy of CMIP6 ESMs in simulating climate data (temperature, precipitation, and solar radiation), we employed historical temperature and precipitation data sourced from the reanalyzed CRUNCEP dataset. Specifically, we utilized monthly data from the atmospheric stress component of CRUNCEP (https://rda.ucar.edu/datasets/ds) and standardized its spatial resolution to 0.5° × 0.5°. Given that CRUNCEP does not include solar radiation data, this study also uses monthly-scale net solar radiation data from ERA5 (https://cds.climate.copernicus.eu/cdsapp#!/search?type=dataset), with its spatial resolution resampled to 0.5° × 0.5°."

Lines 421-431

"The large inter-model spread and the model-observation mismatch in carbon cycle trends can be attributed to two primary sources: biases in the simulated climate and differences in model structure/parameterization. Our evaluation of the models' climatic outputs (Figure S9, Table S2) reveals that models did not accurately reproduce the observed climate over China for 2003-2014, generally overestimating mean temperature while underestimating mean precipitation and solar radiation. The high RMSD values, particularly for precipitation and radiation, indicate substantial errors in the simulated climatic drivers that propagate into the carbon cycle simulations. Furthermore, parameterization and model structure are fundamental for ecosystem models to generate realistic projections, playing a critical role in their accuracy (Luo et al., 2016). As synthesized from previous studies (Table S3), the selected ESMs exhibit considerable diversity in their key land surface components and related parameters (Spafford and Macdougall, 2021; Arora et al., 2020; Pan et al., 2025). This structural and parametric heterogeneity is a major factor contributing to the divergent performances in simulating the trends of carbon-cycle variables among the models."

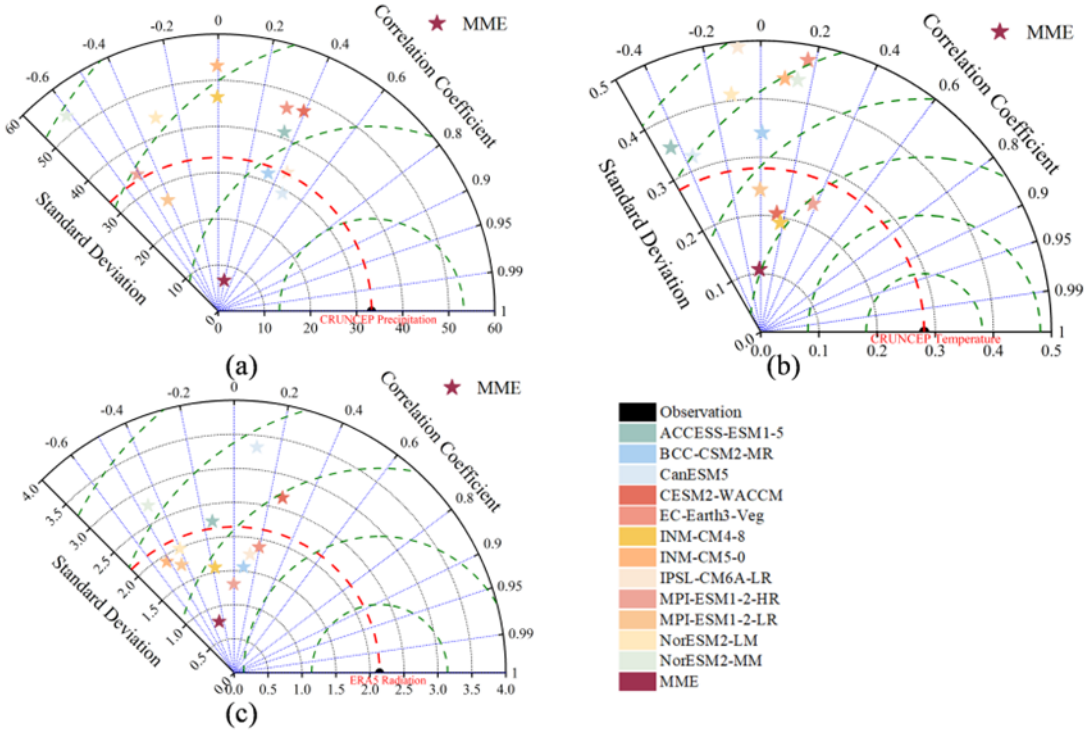

Figure S9: The Taylor diagram compares the observed and model-simulated annual means of climate factors (a) precipitation, (b) temperature, and (c) solar radiation for the historical period (2003-2014). The standard deviation shows the interannual variability of the observed and simulated LAI. The dashed green lines show centered root mean square difference (RMSD) between model simulations and satellite observations.

Table S2: Summary of Statistical Metrics (Mean, SD, RMSD) for Observed and Model-Simulated Annual Mean Precipitation, Temperature, and Radiation

	Precipitation(mm)			Temperature(°C)			Radiation(W/m²)		
	Mean	SD	RMSD	Mean	SD	RMSD	Mean	SD	RMSD
Remote sensing observation	613.64806	33.26141	/	7.63093	0.28095	/	178.18524	2.14209	/
MME	867.4114	6.73035	31.29347	6.52192	0.10744	0.29024	194.01037	0.78118	2.3726
ACCESS-ESM1-5	1040.18351	41.31607	41.28819	7.16172	0.35231	0.51566	184.97961	2.2438	3.17482
BCC-CSM2-MR	727.37508	31.7226	35.76094	4.82402	0.34208	0.42253	183.79863	1.55593	2.42488
CanESM5	747.40222	29.02126	30.65384	3.78862	0.32382	0.47857	188.30924	3.32677	3.61066

CESM2-WACCM	931.57702	47.06833	43.754	8.61381	0.20517	0.31106	191.1325	2.66486	2.81495
EC-Earth3-Veg	697.68418	46.3696	45.6074	5.62987	0.4737	0.48608	190.16296	1.87844	2.44838
INM-CM4-8	1116.61609	46.36924	54.74785	7.04681	0.19034	0.29721	210.0205	1.56934	2.75106
INM-CM5-0	1082.96623	53.14569	60.16718	6.52697	0.4369	0.47534	210.44639	1.90865	3.38243
IPSL-CM6A-LR	738.0079	33.52197	33.67346	3.04972	0.49094	0.56015	201.9655	1.75335	2.47112
MPI-ESM1-2-HR	733.44302	34.4346	56.33748	7.25265	0.23723	0.27916	194.34585	1.30037	2.40424
MPI-ESM1-2-LR	794.64687	26.41971	48.19112	6.4484	0.24422	0.35745	186.96089	1.76221	3.17683
NorESM2-LM	851.89759	44.01298	60.15218	8.31717	0.41054	0.50303	193.13471	1.98927	3.31409
NorESM2-MM	834.14687	53.5992	75.2135	7.8185	0.43701	0.46347	197.56405	2.76213	4.02235

Table S3: Summary of key land surface components and related parameters for selected ESMs

ESM name	Modelling group	Land surface model	Fire/N cycle/Dynamic vegetation	Prognostic LAI/leaf phenology	Number of live&dead carbon pools	Number of PFTs
ACCESS-ESM1-5	CSIRO	CABLE2.4	No/Yes/No	Yes/No	3 & 6	13
BCC-CSM2-MR	BCC	BCC-AVIM2	No/No/No	Yes/Yes(for deciduous)	3 & 8	16
CanESM5	CCCma	CLASS-CTEM	No/No/No	Yes/Yes	3 & 2	9
CESM2-WACCM	CESM	CLM5	-/Yes/No	Yes/Yes	22 & 7	22
EC-Earth3-Veg	EC-Earth	H-TESSEL & LPJ-GUESS	-	-	-	-
INM-CM4-8	INM	-	-	-	-	-

INM-CM5-0	INM	-	-	-	-	-
IPSL-CM6A-LR	IPSL	ORCHIDEE v2.0	No/No/No	Yes/Yes	8 & 3	15
MPI-ESM1-2-HR	MPI	JSBACH3.2	Yes/Yes/Yes	Yes/Yes	3 & 18	13
MPI-ESM1-2-LR	MPI	JSBACH3.2	Yes/Yes/Yes	Yes/Yes	3 & 18	13
NorESM2-LM	NorESM	CLM5	Yes/Yes/No	Yes/Yes	22 & 7	22
NorESM2-MM	NorESM	CLM5	Yes/Yes/No	Yes/Yes	22 & 7	22

General comment (2)

The authors should provide a rationale for the chosen variables. If the aim is the carbon cycle, biomass (above-ground/below-ground) is a key integrator and should be evaluated or explicitly justified as out of scope. Also note that LST primarily reflects the surface energy balance; explain how LST evaluation informs carbon-cycle processes, or consider adding energy-balance variables (e.g., sensible/latent heat, ET) for context.

Reply:

We thank the reviewer for these insightful comments regarding variable selection. In response to the first point, we have enhanced the Introduction to provide a clearer rationale for our chosen variables (LAI, GPP, NPP, NEP), which collectively represent key processes in the terrestrial carbon cycle from canopy structure (LAI) to photosynthetic input (GPP), net biomass production (NPP), and ecosystem-level carbon flux (NEP). While we acknowledge biomass as an important integrated carbon pool, our study specifically focuses on flux-related and vegetation-influenced variables that are more directly comparable across CMIP6 models. The exclusion of biomass was primarily due to the considerable challenges in obtaining reliable, spatially explicit observational data for benchmarking, as well as significant structural uncertainties in model representations of biomass partitioning.

Regarding LST, we have expanded our explanation to clarify its role as a critical environmental regulator—rather than a direct carbon flux—that strongly influences carbon-cycle processes by modulating plant physiological activity, soil respiration rates, and ecosystem-scale carbon exchanges. We use LST as a widely available and physically consistent proxy for assessing model capability in reproducing surface climatology, which indirectly supports the interpretation of carbon flux simulations.

Changes in the paper:

Lines 47-61

"The LAI, defined as the total one-sided leaf area per unit ground surface area, is a key parameter of vegetation canopy structure that directly influences light interception, transpiration, and the spatial heterogeneity of GPP. GPP refers to the total amount of carbon dioxide fixed into organic compounds by vegetation through photosynthesis, serving as a core indicator of an ecosystem's carbon sequestration capacity. NPP represents the net carbon accumulation after

subtracting autotrophic respiration from total photosynthetic fixation, reflecting the primary production potential and health of ecosystems as influenced by GPP and plant physiological regulation. NEP denotes the net carbon exchange between the ecosystem and the atmosphere by further subtracting heterotrophic respiration from NPP, making it a crucial measure for assessing regional carbon source/sink status under the influence of environmental factors such as atmospheric CO_2 and climate (Fang et al., 2001). LST is the thermodynamic temperature at the land-atmosphere interface, playing a key role in surface energy and water exchange while jointly affecting ecological processes through interactions with solar radiation, soil properties, vegetation, and atmospheric conditions (Li et al., 2023). In-depth research on LST facilitates a deeper understanding of surface-atmosphere exchange processes at global and regional scales and provides high-quality quantitative indicators of surface conditions for scientific applications. Consequently, LST has been designated as an indispensable observation indicator for the International Geosphere and Biosphere Program (IGBP) and the Global Climate Observing System (GCOS) (Townshend et al., 1994; Hollmann et al., 2013)."

General comment (3)

The study focuses on interannual trends but does not evaluate absolute levels. Please assess model skill for the magnitude (bias, RMSE, correlation) alongside trends, or cite and synthesize prior evaluations that establish these baselines. Without this, it is difficult to judge whether trend errors arise from mean-state biases or not.

Reply:

We thank the reviewer for raising this critical point regarding the assessment of model skill in simulating absolute magnitudes. In response, we have expanded our evaluation beyond interannual trends to comprehensively assess model performance in representing the mean state of all variables. Specifically, we have incorporated Taylor diagrams into the revised Figures 3 (for LAI, GPP, NPP, and NEP) and 4 (for LST), which simultaneously visualize three key statistical metrics—the correlation coefficient, centered root-mean-square error (RMSE), and the standard deviation (representing the amplitude of variability). This multi-metric approach allows for a direct and quantitative assessment of how well the models replicate both the spatial patterns and the absolute magnitudes of the observed benchmarks. By integrating this analysis, we can now more robustly discern whether potential trend errors are linked to underlying mean-state biases or to inaccuracies in representing temporal dynamics, thereby strengthening the foundation for interpreting our trend-based results.

Changes in the paper:

Line 267-270

"Figure 3: Overall annual average (a) NPP and (c) NEP trends in China during 2003-2014. The asterisk (*) indicates the significant trend ($p<0.05$). The Taylor diagrams compare the remotely-sensed and model-simulated annual mean (b) NPP and (d) NEP for the historical period (2003-2014). The standard deviation shows the interannual variability of the observed and simulated LAI. The dashed green lines show centered root mean square difference (RMSD) between model simulations and satellite observations."

Line 271-277

"The Taylor diagram (Fig. 3b, 3d) evaluates interannual variations of NPP and NEP across individual models, the MME, and observations. The SD of ESM-simulated NPP ranges from 7.70

(ACCESS-ESM1-5) to 24.95 (BCC-CSM2-MR), while RMSD values span from 8.82 (INM-CM4-8) to 25.66 (INM-CM5-0). For NEP, SD ranges from 9.19 (CanESM5) to 20.24 (BCC-CSM2-MR) with RMSD values between 13.55 (IPSL-CM6A-LR) and 21.41 (BCC-CSM2-MR). The MME demonstrates notably lower variability, with SD values of 5.65 for NPP and 6.19 for NEP, and RMSD values of 9.71 for NPP and 11.28 for NEP. Generally, the MME-simulated SD and RMSD for both variables are lower than those of all individual models."

Line 284-287

"Figure 4: (a) Overall annual average LST trends in China during 2003-2014. The asterisk (*) indicates the significant trend ($p<0.05$). (b) The Taylor diagrams compare the remotely-sensed and model-simulated global annual mean LST for the historical period (2003-2014). The standard deviation shows the interannual variability of the observed and simulated LAI. The dashed green lines show centered root mean square difference (RMSD) between model simulations and satellite observations."

Line 288-291

"The Taylor diagram evaluates interannual variations of LST across individual models, the MME, and observations (Fig. 4). The SD of ESM-simulated LST ranges from 0.20 (INM-CM4-8) to 0.48 (IPSL-CM6A-LR), while RMSD values span from 0.30 (MPI-ESM1-2-HR) to 0.59 (IPSL-CM6A-LR). The MME demonstrates notably lower variability and error, with an SD of 0.10 and RMSD of 0.36."

General comment (4)

The authors should justify using CSIF alone for GPP evaluation. CSIF is not a purely observational GPP product and carries its own assumptions. Consider evaluating with additional GPP datasets (e.g., GOSIF-GPP, FLUXCOM-GPP). The authors should show whether conclusions are robust across benchmarks.

Reply:

We thank the reviewer for this valuable suggestion regarding GPP benchmark selection. We fully acknowledge that CSIF is not a purely observational product and carries inherent assumptions. In response, we have now added justification in the Methods section explaining our selection of CSIF as the primary GPP benchmark. This decision is further supported by our prior empirical comparisons with alternative GPP products (including GOSIF and FLUXCOM-GPP), which indicated that the temporal trends in CSIF align more closely with observed vegetation dynamics across China. Compared with GOSIF-GPP and FLUXCOM-GPP, the CSIF demonstrates superior temporal continuity by effectively mitigating data gaps from clouds and sensor revisits, possesses a more direct physiological link to photosynthesis through chlorophyll fluorescence, and offers more consistent spatial scalability. Although minor quantitative differences exist, the core conclusions regarding model performance remain robust. A statement to this effect has been added to the Discussion to acknowledge the potential influence of benchmark choice.

Changes in the paper:

Lines 140-145

"Solar-Induced Chlorophyll Fluorescence (SIF), which integrates the complex physiological functions of plants and can directly reflect the dynamic changes in plants' actual photosynthetic process, exhibits a strong linear relationship with GPP, thereby serving as a direct observational indicator for GPP (Mohammed et al., 2019; Frankenberg et al., 2014; Walther et al., 2016).

Compared with GOSIF-GPP and FLUXCOM-GPP, CSIF exhibits higher temporal resolution and more robust spatial data gap-filling capability. Therefore, the CSIF dataset was used to validate the GPP outputs of the models under the historical scenario."

Lines 470-471

"Although the reprocessed MODIS and CSIF datasets utilize spatiotemporal filtering and machine learning techniques to enhance data quality (Zhang et al., 2018; Yuan et al., 2011), significant uncertainties remain."