➢ AC to Referee #1: General Comment

The objective of the article titled "CAMELE: Collocation-Analyzed Multi-source Ensembled Land Evapotranspiration Data" is to create a daily merged evaporation product using a collocation-based data ensemble method. This method takes into account non-zero error covariance conditions to merge multiple ET (Evapotranspiration) products, resulting in the Collocation Analyzed Multi-source Ensembled Land Evapotranspiration data. In general, the article is clear and well-written, and it falls within the scope of this journal. Below, I provide some points and comments that, in my opinion, can further enhance the manuscript:

AC:

We greatly appreciate the professional and constructive feedback provided by the reviewer. We will respond to each comment individually, and in the following responses, the line numbers corresponding to the added or revised content will be based on the updated version without highlights. You can open the PDF file's table of contents view to navigate to the relevant sections directly.

The responses will be in the following format:

- ➢ Reviewer's comments are shown in black.
- ➢ Our responses are shown in blue.
- ➢ The modifications to the manuscript are shown in orange.
- ➢ Previous contents in the old version (for comparison if needed) are shown in grey.

1 AC to Referee #1: Major Comment

1.1 Q1

It would be beneficial if the Scenario 1 product (at 0.10 degrees) could be extended until 2020. As far as I can see, PMLv2 is available until 2020 (please verify the link to the product). Additionally, it is important for the authors to outline their plans for updating the product and whether it will become operational. This is crucial, as many datasets become obsolete after publication.

AC:

We greatly appreciate the reviewer's suggestions. In fact, we have already utilized the latest PMLv2 data extended until 2020 in our research, and we have verified the link to the product, which is accurate. Table 2 in the original manuscript lists the combinations at 0.1° resolution, and the PMLv2 data extended to 2020 has been incorporated.

TABLE.2 Combination of inputs and accessible methods

Scenario 1 (0.1°)		
Period	**Selected Inputs**	**Method**
（2000.02.26-2000.12.31)	ERA5L/ PMLv2	IVD
（2001.01.01-2015.12.27)	ERA5L/ FluxCom/ PMLv2	EIVD
（2015.12.28-2020.12.26)	ERA5L/ PMLv2	IVD
Scenario 2 (0.25°)		
Period	**Selected Inputs**	**Method**
（1980.01.01-1999.12.31)	ERA5L/ GLDAS20/ GLEAMv3.7a	EIVD
（2000.01.01-2022.12.31)	ERA5L/ GLDAS21/ GLEAMv3.7a	

Furthermore, we have added Section 5.4 in the discussion, outlining our plans for future updates, which include:

- Updating the data used in this study to the most recent versions, ensuring more reliable results even with the use of newer data.
- Considering the inclusion of additional data and implementing extended collocation methods to further reduce estimation errors in ET.
- Improving the accuracy of CAMELE by integrating higher-resolution regional ET data.

These steps will address the issue of dataset obsolescence and enhance the long-term relevance and operational utility of our product.

New contents (Line 1034 to 1060):

"**5.4. Potential Applications and Future Enhancements**

In this section, we delve into the potential applications of our product and outline our commitment to future enhancements to maintain its accuracy and relevance.

Here, we identify three potential applications for our transpiration product: (1) Global ET Trends: Our product facilitates global-scale analysis of current ET patterns and long-term trends, essential for comprehending ecosystem responses to evolving environmental conditions in a warming climate; (2) Transpiration-to-Evapotranspiration Ratio: Our merging approach can fuse multi-source global gridded transpiration data, allowing for the examination of the transpiration-to-evapotranspiration ratio. This analysis can enhance water resource management and water availability predictions in diverse regions; (3) Attribution analysis: Our product is a valuable tool for attribution analysis, helping researchers identify the drivers of patterns. This knowledge is crucial for understanding the roles of climate variability, land-use changes, and other factors in shaping terrestrial water fluxes.

Furthermore, we are committed to enhancing our product proactively. Key strategies include: (1) Data Update and Validation: To ensure our product's continued accuracy and reliability, we will prioritize regularly updating the data used in this study to the latest versions. By adopting this approach, we aim to provide users with results that reflect the latest advancements in scientific knowledge; (2) Enhanced Integration and Error Reduction: We continually refine estimates by incorporating additional data sources and implementing extended collocation method to minimize errors; (3) Integration of High-Resolution Regional ET Data: Recognizing the significance of regional-scale insights, we will focus on improving the accuracy of CAMELE by integrating higher-resolution regional ET data. This integration will enable more precise regional estimation.

In summary, these endeavors collectively represent our commitment to maintaining our product's quality and relevance, ensuring its value for the scientific community."

1.2 Q2

I recommend expanding the introduction to clarify the implications of non-zero error covariance between different products. This will help readers better understand the importance of considering this aspect in merging strategies, especially when the assumption of error independence is violated.

AC:

We sincerely appreciate the reviewer's valuable feedback and have made appropriate revisions to the Introduction section in accordance with the suggestion. Specifically, we have focused on two key aspects:

- Emphasizing the impact of the violation of the zero-ECC assumption on collocation analysis.
- Highlighting the previous studies' neglect of adequately considering non-zero ECC, which we address in our research.

Revised contents (Line 104 to 130):

Although the above studies have demonstrated that collocation analysis can effectively assess the random error variance of ET products and integrate error information from multiple data sources, these studies have primarily overlooked a critical aspect: non-zero ECC between ET products. Li et al. (2022) global ET product evaluation research revealed clear non-zero ECC conditions between ERA5L, GLEAM, PMLv2, and FluxCom. In TC analysis, non-zero ECC can result in significant biases in TC-based results (Yilmaz and Crow, 2014). Furthermore, when using TC-based error information for fusion, it is crucial to consider the information related to ECC, as this can help improve the fusion accuracy (Dong et al., 2020b; Kim et al., 2021b).

It is worth noting that non-zero ECC conditions pose unique challenges. Unlike other violations of mathematical assumptions adopted by TC, they cannot be effectively mitigated through rescaling or compensated for by equal magnitude adjustments across inputs. Thus, the implications of non-zero ECC in the context of merging strategies are a critical consideration often overlooked in previous research. This oversight can lead to significant biases and inaccuracies. We aim to bridge this gap by systematically accounting for non-zero ECC in weight calculation, contributing to a more robust and accurate assessment.

1.3 Q3

Please consider using the modified Kling-Gupta efficiency proposed by Kling et al. (2012) instead of the KGE of Gupta et al. (2009).

- Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. Journal of hydrology, 424, 264-277.

AC:

We greatly appreciate the reviewer's constructive feedback and have revised the calculation method of KGE. We have made corresponding modifications to Section 3.6 as follows:

Revised contents (Line 500 to 504):

The modified KGE (Kling et al., 2012) offers insights into reproducing temporal dynamics and preserving the distribution of time series, which are increasingly used to calibrate and evaluate hydrological models (Knoben et al., 2019). For a better understanding of the KGE statistic and its advantages over the Nash-Sutcliffe Efficiency (NSE), please refer to Gupta et al. (2009). The equation is given by:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\sigma_{sim}/\mu_{sim}}{\sigma_{obs}/\mu_{obs}} - 1\right)^2} \qquad (27)$$

Furthermore, all calculations related to KGE have been updated accordingly, including:

Overall, the values of the modified KGE are slightly higher than the previous KGE values. The performance of the CAMELE fusion product remains superior to other products and combinations. Since multiple changes were involved and there were no adjustments to the conclusions, we will not list them individually here.

Relative contents in previous manuscript:

The KGE (Gupta et al., 2009) addressed several shortcomings in Nash-Sutcliffe Efficiency (NSE) and are increasingly used for calibration and evaluation (Knoben et al., 2019), given by:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \qquad (27)$$

1.4 Q4

If CAMELE performs similarly to other products, why should it be used? The goal in merging datasets is to outperform the products used in the merging procedure and thus better represent spatio-temporal evaporation patterns. The authors could focus on the fact that even though CAMELE may not outperform all products in all metrics, it performs better when considering all metrics. This suggests that it is a robust product and that the method can generate a product that leverages the complementary strengths of different datasets to some extent.

AC:

We greatly appreciate the reviewer's valuable feedback. As you rightly pointed out, CAMELE's performance in terms of accuracy metrics closely aligns with that of the input products. However, we have significantly improved error metrics, which is consistent with our strategy of calculating fusion weights based on collocation analysis to match random error variances. In response to your suggestion, we have revised the conclusion section to emphasize two key points:

- ➢ While CAMELE may not be the best performer in all metrics, it effectively reduces errors associated with the input products, making it highly robust when considering a comprehensive evaluation at the station scale.
- ➢ The weighting scheme that considers non-zero ECC (Error-Correction Coefficients) proves to be a more effective means of integrating the strengths and weaknesses of the input products, thus providing more reliable fusion results.

Revised contents (Line 1072 to 1086):

2. Compared to five input products, REA, and simple average, the CAMELE product performed well when evaluated against FluxNet flux tower data. While CAMELE may not excel in all individual metrics, it excels in effectively reducing errors associated with the input products. The result showed Pearson correlation coefficients (R) of 0.63 and 0.65, root-mean-square errors (RMSE) of 0.81 and 0.73 mm/d, unbiased root-mean-square errors (ubRMSE) of 1.20 and 1.04 mm/d, mean absolute errors (MAE) of 0.81 and 0.73 mm/d, and Kling-Gupta efficiency (KGE) of 0.60 and 0.65 on average over resolutions of 0.1° and 0.25°, respectively. This robust performance is especially evident when assessing its comprehensive station-scale evaluation.

3. For different plant functional types (PFTs), the CAMELE product outperformed the five input products, REA, and simple average in most PFTs. Although FluxCom and PMLv2 performed slightly better than CAMELE at some PFT sites, considering that both utilized FluxNet sites for product calibration, it indirectly demonstrates the promising and robust performance of CAMELE.

1.5 Q5

The multi-year comparison is interesting as it highlights variations in the datasets. The authors might consider excluding the trends comparison, as it may lack significance without a comparison with in-situ-based trends. This change would also help to reduce the manuscript.

AC:

We want to thank the reviewer for the valuable feedback. We have incorporated the analysis of trends by aligning the trend comparisons within the same period. Additionally, we have assessed the CAMELE and other products at the site scale, providing an evaluation of their estimations for multi-year linear trends and seasonality. This modification aims to address your concern and enhance the manuscript:

Revised contents (Line 798 to 873):

4.4. Assessment and comparison of linear trend and seasonality

In this section, we first validate and compare the performance of CAMELE with other products in estimating multi-year trends and seasonality at the site scale. Due to the inconsistent time lengths of FluxNet sites, trends at many sites are not significant. Therefore, we deliberately selected 13 sites with continuous evapotranspiration (ET) observations for the same 11-year period (2004 to 2014) and with significant trends. The annual ET values for each year were calculated as the mean of the 13 sites for that year, allowing the computation of linear trends and seasonality. We employed singular spectrum analysis (SSA), which assumes an additive decomposition A = LT + ST + R. In this decomposition, LT represents the long-term trend in the data, ST is the seasonal or oscillatory trend (or trends), and R is the remainder.

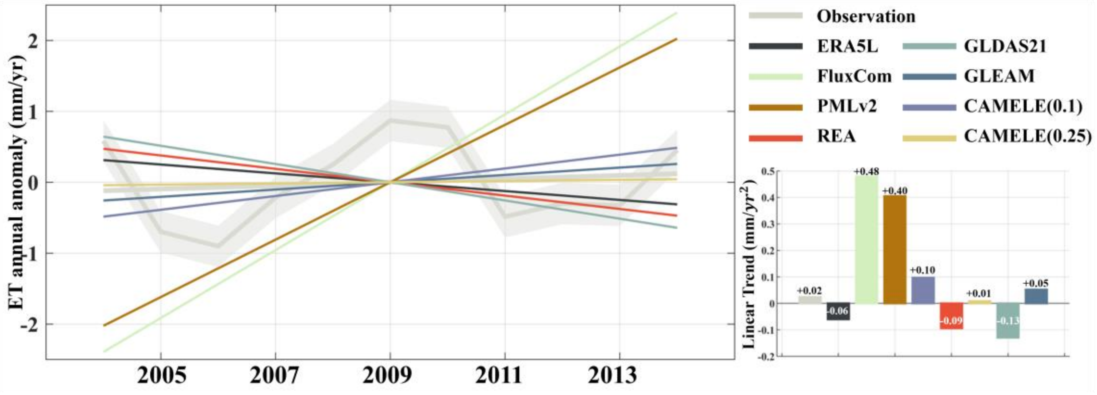


Figure 13 Comparison of linear trend from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The trends have been subjected to SSA decomposition, removing seasonality. The gray enveloping line represents the mean plus the standard deviation of the 13 sites.

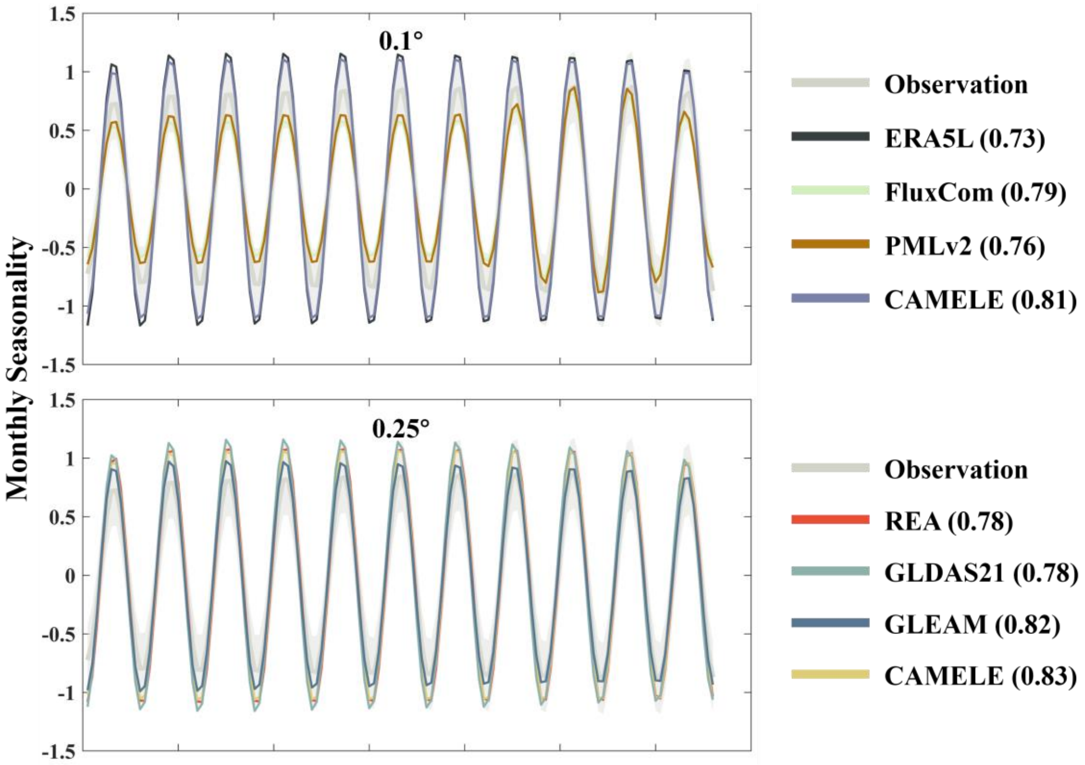


Figure 14 Comparison of seasonal variations from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The seasonality has been obtained through SSA decomposition, with the gray area representing the observed values. The parentheses in each product name indicate the KGE coefficient comparing with the observed values.

In Figure 13 and Figure 14, based on observations from FluxNet sites, we analyzed the performance of CAMELE and other products in estimating the linear trend and seasonality of ET over multiple years. It is important to note that we only present the analysis results for 13 sites with continuous 11-year observations, and the performance of different ET products in trend estimation at individual sites still varies, not fully reflecting the overall performance on all grids in terms of trend and seasonality. Nevertheless, such a comparison can still provide valuable insights.

Examining the results of the linear trend, both PMLv2 and FluxCom exhibit a significant upward trend, well above the observations. On the contrary, ERA5L, GLDAS, and REA show a noticeable downward trend, while CAMELE demonstrates a gradual upward trend closer to the observations. Additionally, GLEAM slightly outperforming CAMELE at a resolution of 0.25°. Overall, CAMELE shows good agreement with site observations in capturing the multi-year linear trend of ET.

Continuing with the analysis of seasonality, the KGE index comparing each product's results with observed values is provided in parentheses next to the product name. Generally, all products exhibit a good representation of ET's seasonal variations.

CAMELE's 0.1° seasonal results closely match FluxCom (with the two lines almost overlapping). However, the fluctuations it reflects are higher than the observed values.

This is likely due to keeping the 8-day average results of FluxCom consistent with PMLv2 every 8 days, and the variability in ET primarily originates from ERA5L results. This aspect may need improvement in subsequent research. At 0.25°, CAMELE's seasonal representation is closer to the observed results. The differences in CAMELE's performance at the two resolutions are mainly attributed to input variations, which we discuss in the following section as potential areas for improvement.

The results indicate that CAMELE effectively captures the multi-year changes in ET, but at 0.1°, it tends to overestimate seasonal fluctuations. We further generated global maps of multi-year linear trends in ET, estimating trends using Theil–Sen's slope method and testing significance with the Mann–Kendall method. The dotted areas indicate trends passing a significance test at a 5% level.

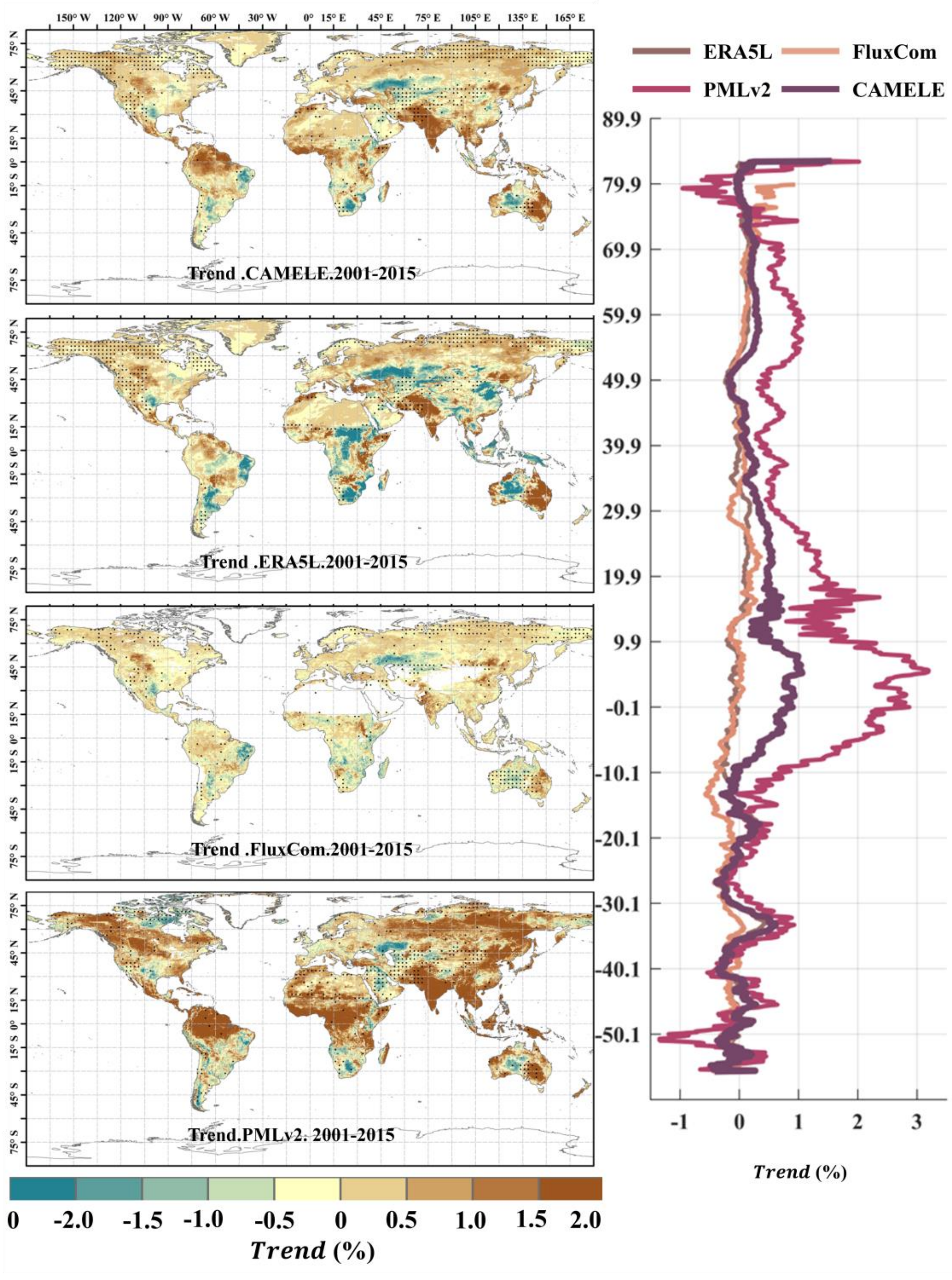


Figure 15 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area indicates that the trend has passed the significance test at 5 % level.

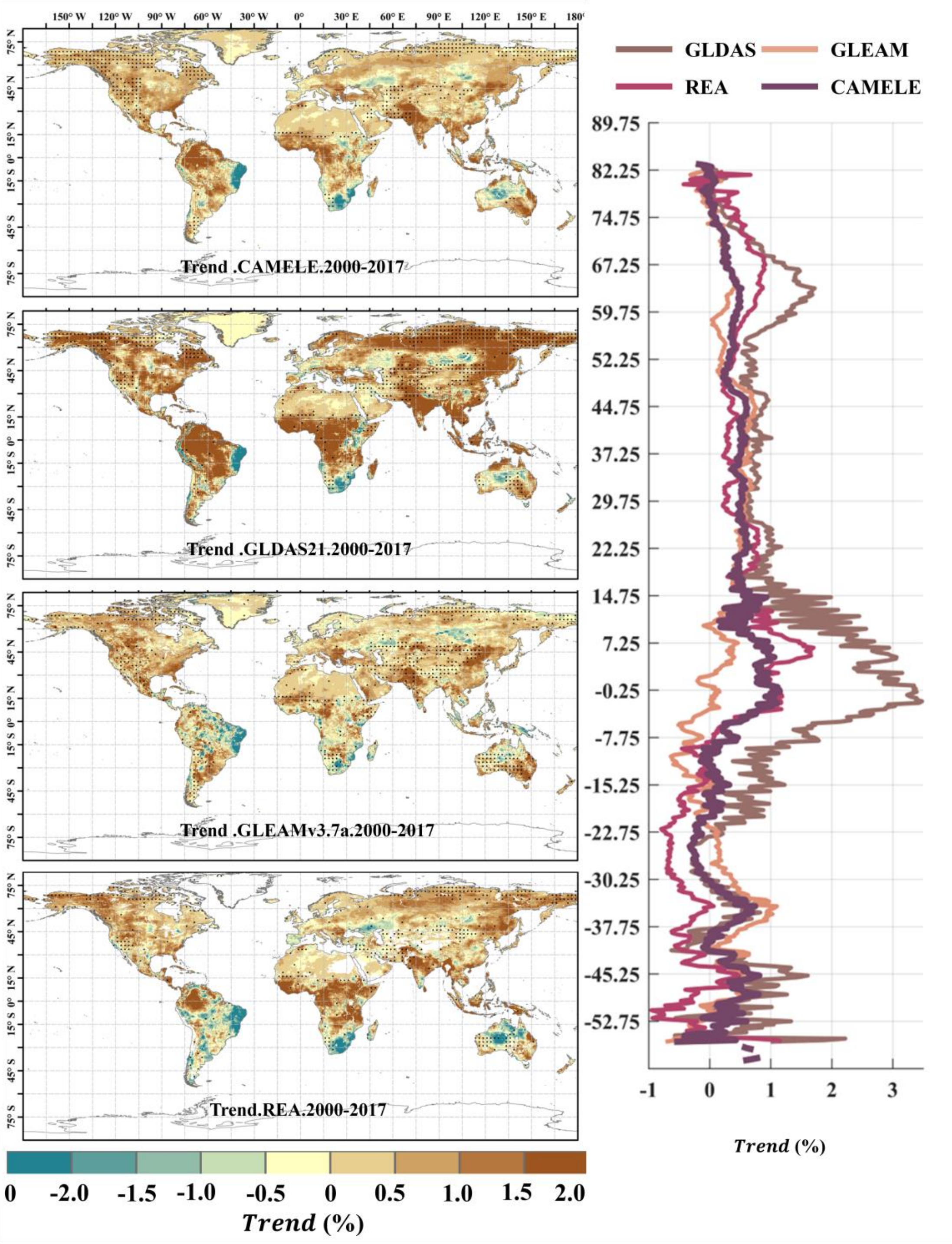


Figure 16 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen’s slope method, and the significance level is tested with the Mann–Kendall method. The dotted area indicates that the trend has passed the significance test at 5 % level.

Figure 15 and Figure 16 present the linear trends of multi-year daily scale evapotranspiration (ET) calculated for different products at resolutions of 0.1° and 0.25°, respectively. The corresponding latitude-dependent variations of the rate of change are shown on the right side. It can be observed that the differences in linear trends among the different products are more significant than the multi-year averages, and in some regions, they even exhibit opposite trends. For example, at 0.1° resolution, PMLv2 shows a global increase of 1.0% in ET in most regions, while the results from CAMELE, ERA5L, and PMLv2 indicate a milder increase in ET in the Amazon rainforest, southern Africa, and northwestern Australia. At 0.25° resolution, except for GLDAS2.1, which shows an apparent global increase in ET, the results from CAMELE, GLEAMv3.7a, and REA indicate milder variations in global ET.

1.6 Q6

The authors employ slightly different products for different periods in the development of the Scenario 1 and Scenario 2 CAMELE products. It's important to discuss the implications of this choice. Can the authors evaluate if there are changes in performance in the different periods selected to construct the datasets?

What are the implications of adding FluxCom for 2001-2015 in Scenario 1? It might be more suitable to use FluxCom as a benchmark and produce the Scenario 1 product solely with ERA5-Land and PMLv2, using only the IVD method. If the authors choose not to follow this suggestion, they should explain, evaluate, and discuss the implications of using two different methods with an additional product for different periods in the Scenario 1 product.

Similarly, it would be beneficial to address the transition from GLDASv20 to GLDASv21 in 1999 for Scenario 2. Can the authors discuss the implications of changing the product versions?

AC:

We sincerely appreciate the valuable comments provided by the reviewer. Since the three questions raised in the comments are closely related, we will address them collectively. These responses pertain to the comparisons between various fusion schemes, as elaborated in Section 5.3 Comparison of different fusion scheme. Following your suggestions, we have incorporated additional comparative results. To begin with, we will address your questions:

RC1: What is the impact of transitioning from GLDASv2.0 to GLDASv2.1, and why was this transition made in 1999?

AC1: The GLDASv2 product series comprises versions 2.0, 2.1, and 2.2. GLDAS-2.2 product suites employ data assimilation (DA), whereas GLDAS-2.0 and GLDAS-2.1 products are considered "open-loop" with no data assimilation **(Rui et al., n.d.)**. The GLDAS-2.1 simulation utilizes conditions from the GLDAS-2.0 simulation, with upgraded models driven by a combination of datasets. Previous research has shown that GLDAS-2.1 offers improvements in the simulation of hydrological variables at the regional scale compared to GLDAS-2.0 **(Qi et al., 2018, 2020)**. Therefore, we opted to use GLDAS-2.1 data for as much time series as possible, resulting in the transition from GLDAS-2.0 to GLDAS-2.1 after 1999. Updated comparisons in Section 5.3 (CAMELE vs Comb2) also indicate that the fusion results using GLDAS-2.1 have more minor errors. Furthermore, we analyzed alternative scenarios, and the comparative results suggest that the approach employed in our study is optimal.

References:

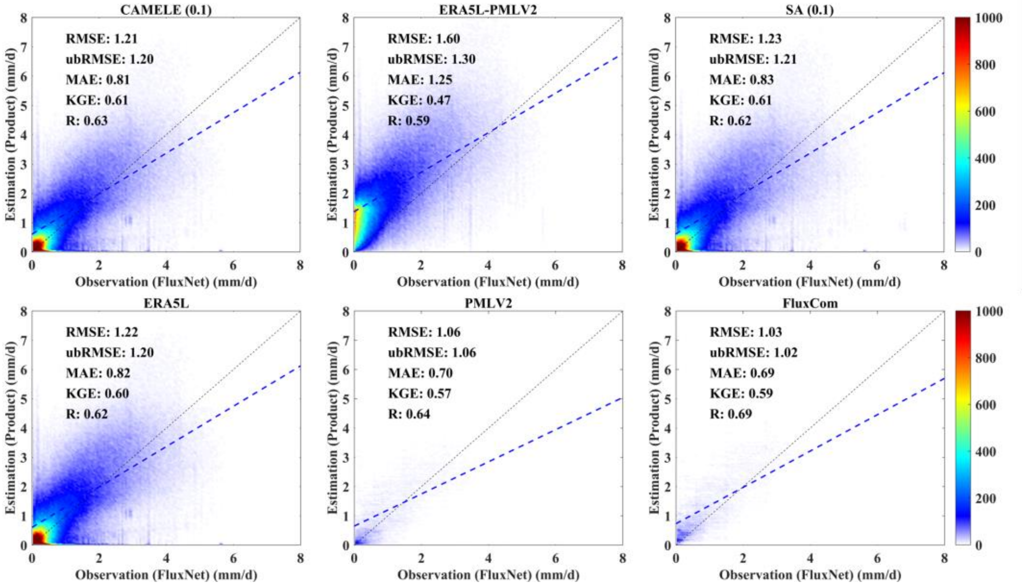
Qi, W., Liu, J., and Chen, D.: Evaluations and Improvements of GLDAS2.0 and GLDAS2.1 Forcing Data's Applicability for Basin Scale Hydrological Simulations in the Tibetan Plateau, JGR Atmospheres, 123, https://doi.org/10.1029/2018JD029116, 2018.

Qi, W., Liu, J., Yang, H., Zhu, X., Tian, Y., Jiang, X., Huang, X., and Feng, L.: Large Uncertainties in Runoff Estimations of GLDAS Versions 2.0 and 2.1 in China, Earth and Space Science, 7, e2019EA000829, https://doi.org/10.1029/2019EA000829, 2020.

Rui, H., Beaudoing, H., and Loeser, C.: README for NASA GLDAS Version 2 Data Products, n.d.

QC2: Why did we not consider using the IVD method exclusively to merge ERA5-Land and PMLv2 at 0.1° resolution?

AC2: We greatly appreciate your observation. We initially compared the IVD fusion results using only ERA5-Land and PMLv2, as illustrated by the scatterplot below.



The fusion results exhibited a positive bias and did not perform as well as individual products or the simple average. Several factors contributed to this phenomenon, including the possibility of significant errors in either ERA5-Land or PMLv2. More importantly, the limitation of using only two datasets prevented us from effectively obtaining error information through collocation analysis **(Dong et al., 2019, 2020)**. Hence, we decided to ensure that we had three datasets as inputs, enabling the application of the EIVD method and ensuring consistency in the methods used for 0.1° and 0.25° resolutions.

References:

Dong, J., Crow, W. T., Duan, Z., Wei, L., and Lu, Y.: A double instrumental variable method for geophysical product error estimation, Remote Sensing of Environment, 225, 217–228, https://doi.org/10.1016/j.rse.2019.03.003, 2019.

Dong, J., Wei, L., Chen, X., Duan, Z., and Lu, Y.: An instrument variable based algorithm for estimating cross-correlated hydrological remote sensing errors, Journal of Hydrology, 581, 124413, https://doi.org/10.1016/j.jhydrol.2019.124413, 2020.

The above responses provide a summary of our answers to your questions. Based on your suggestions, we have updated the content in Section 5.3, which includes the information provided in the responses to clarify further the optimality of the fusion approach selected in this study.

Revised Section 5.3 Comparison of different fusion scheme (Line 965 to 1017):

In this section, we conducted comparisons in three aspects: (1) comparing the performance of CAMELE at different resolutions; (2) comparing the performance of different change fusion schemes, explicitly changing the input products' versions (GLDAS21 to GLDAS20 or GLDAS22, GLEAMv3.7a to v3.7b); and (3) comparing the performance of the results obtained without considering the ECC impact.

We conducted a comprehensive comparison of our fusion approach with several alternative schemes. Specifically, these schemes encompassed utilizing only ERA5L and PMLV2 at 0.1° based on the IVD method (Comb1), changing the versions of GLDAS2 and GLEAM at 0.25° based on the EIVD method (Comb2-5), and two TC fusion approaches at 0.1° and 0.25°, which did not incorporate ECC.

It should be noted that the Comb2 scheme, which includes GLDAS20, covers the period from 1980 to 2014, while the other 0.25° comparison schemes (Comb3-5) span from 2003 to 2022. The combinations based on TC (assuming zero ECC) had the same inputs as CAMELE at both resolutions.

Table 7 Average metrics for CAMELE and other fusion schemes at all sites. The bolded sections indicate the schemes with the best performance in their respective metrics.

Product	RMSE	ubRMSE	MAE	KGE	R

	(mm/d)	(mm/d)	(mm/d)		
CAMELE (0.1)	**0.83**	**0.71**	**0.64**	**0.57**	**0.71**
CAMELE (0.25)	1.03	0.87	0.75	0.51	0.67
ER5L+PMLV2 *(Comb1-0.1 | IVD)*	1.13	1.00	0.89	0.46	0.61
ER5L+GLDAS20+GLEAMv3.7a *(Comb2-0.25 | EIVD)*	1.09	0.89	0.87	0.44	0.66
ER5L+GLDAS22+GLEAMv3.7a *(Comb3-0.25 | EIVD)*	1.20	0.95	0.94	0.44	0.68
ER5L+GLDAS22+GLEAMv3.7b *(Comb4-0.25 | EIVD)*	1.19	0.94	0.93	0.44	0.69
ER5L+GLDAS21+GLEAMv3.7b *(Comb5-0.25 | EIVD)*	1.05	0.90	0.80	0.49	0.69
ER5L+FluxCom+PMLv2 *(Zero-ECC-0.1 | TC)*	1.06	0.91	0.80	0.46	0.60
ER5L+GLDAS21+GLEAMv3.7a *(Zero-ECC-0.25 | TC)*	1.26	1.03	0.99	0.39	0.61

According to the information in the table, CAMELE (0.1°) results were superior in all indicators. Firstly, when comparing the performance of CAMELE at resolutions of 0.1° and 0.25°, it was observed that the fused product performed slightly worse at the 0.25° resolution. This could be attributed to the variations in the input products. Additionally, the representative of FluxNet sites at the 0.25° resolution decreased, leading to degraded statistical indicators.

At the 0.1° resolution, we conducted a comparison of results obtained by exclusively fusing ERA5-Land and PMLv2. Multiple indicators indicated that this approach did not enhance the accuracy of ET estimates and fell significantly short of the scheme employed in CAMELE. This implies that using only two product sets as input did not allow for effective error analysis through collocation analysis, resulting in suboptimal fusion results. More importantly, the limitation of employing only two datasets prevented us from effectively acquiring error information through collocation analysis (Dong et al., 2020a, 2019). Consequently, we made the strategic decision to ensure the inclusion of three datasets as inputs, facilitating the utilization of the EIVD method and maintaining methodological consistency between the 0.1° and 0.25° resolutions.

Furthermore, when comparing the results of different fusion schemes between CAMELE and Comb2-5 at the 0.25° resolution, CAMELE performed better regarding error metrics (RMSE, ubRMSE, MAE). The differences in fitting metrics (KGE, R) were insignificant, indicating that the choice of fusion scheme primarily affected the errors of the fusion results. The relatively poor performance of other fusion schemes could be due to the lack of consideration for non-zero ECC. For example, GLEAMv3.7b and GLDAS2.2 employed the satellite data from MODIS, introducing random error homogeneity between the two datasets.

For the comparative analysis of the GLDAS2.0 and GLDAS2.1 schemes, the usage of GLDAS2.1 yielded better performance. The GLDAS-2.1 simulation leverages conditions from the GLDAS-2.0 simulation, with improved models driven by a combination of datasets. Previous research has demonstrated that GLDAS-2.1 offers improvements in the regional-scale simulation of hydrological variables compared to GLDAS-2.0 (Qi et al., 2018, 2020). Consequently, we chose to incorporate GLDAS-2.1 data for as much of the time series as possible.

Moreover, when comparing the fusion effects with and without considering non-zero ECC conditions, it was evident that considering ECC information could effectively improve the performance of the fused product, which further demonstrated the reliability and advantages of the fusion method employed in this study.

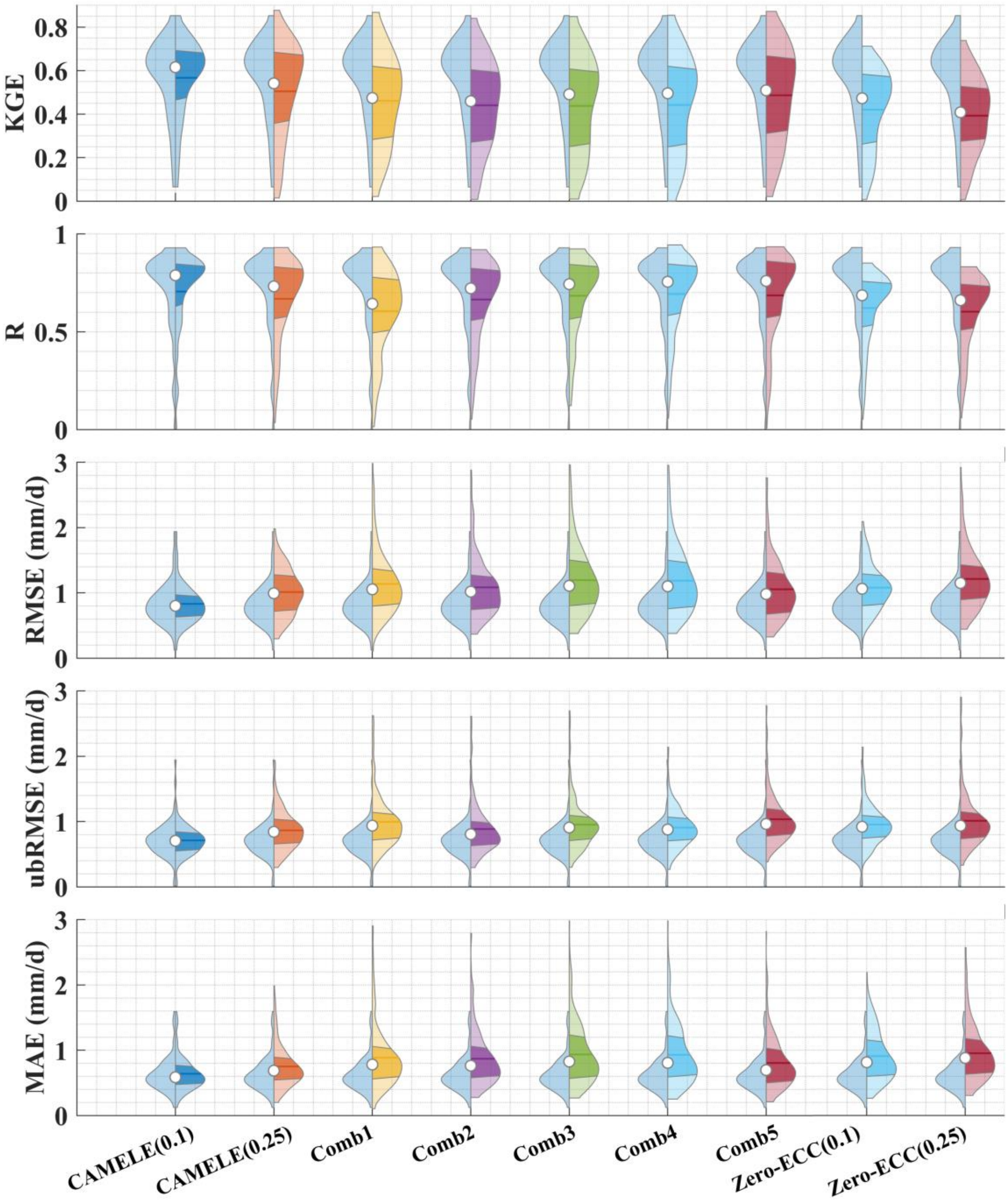


Figure 13 Violin plot comparing KGE, R, RMSE, ubRMSE and MAE of CAMELE with other fusion schemes. The right half of each violin plot represents the distribution, with shaded areas indicating the box plot, where the horizontal line corresponds to the median and the dot represents the mean. The left half represents the results of CAMELE (0.1°) for comparison.

We further provided violin plots for different metrics, comparing the results of each fusion scheme to CAMELE (0.1°). The results indicated that the fusion schemes adopted were significantly superior to other schemes based on the distribution of results for all metrics across all sites. Regarding KGE and R, CAMELE's results were concentrated near 1 for most sites. Regarding RMSE, ubRMSE, and MAE, their results were concentrated below one mm/d. The results in the plots also suggested that CAMELE performed slightly worse at 0.25° compared to 0.1° but still outperformed other combination results. Additionally, comparing CAMELE and the zero-ECC scheme in the plots further highlighted the importance of considering non-zero ECC conditions.

References:

Dong, J., Crow, W. T., Duan, Z., Wei, L., and Lu, Y.: A double instrumental variable method for geophysical product error estimation, Remote Sensing of Environment, 225, 217–228, https://doi.org/10.1016/j.rse.2019.03.003, 2019.

Dong, J., Wei, L., Chen, X., Duan, Z., and Lu, Y.: An instrument variable based algorithm for estimating cross-correlated hydrological remote sensing errors, Journal of Hydrology, 581, 124413, https://doi.org/10.1016/j.jhydrol.2019.124413, 2020.

Qi, W., Liu, J., and Chen, D.: Evaluations and Improvements of GLDAS2.0 and GLDAS2.1 Forcing Data's Applicability for Basin Scale Hydrological Simulations in the Tibetan Plateau, JGR Atmospheres, 123, https://doi.org/10.1029/2018JD029116, 2018.

Qi, W., Liu, J., Yang, H., Zhu, X., Tian, Y., Jiang, X., Huang, X., and Feng, L.: Large Uncertainties in Runoff Estimations of GLDAS Versions 2.0 and 2.1 in China, Earth and Space Science, 7, e2019EA000829, https://doi.org/10.1029/2019EA000829, 2020.

1.7 Q7

The discussion regarding the impact of underlying assumptions in collocation analysis could be more closely related to the development of CAMELE. As it currently stands, it seems to be a comparison of evaporation datasets. Readers would benefit from a more direct connection between the performance of CAMELE and the assumptions of the methods used in its development.

It's worth exploring why the merging scheme did not significantly improve the performance of CAMELE. Could this be attributed to non-linear relationships between evaporation magnitudes and their respective errors? The authors should consider expanding on this in the discussion section.

AC:

We sincerely appreciate the reviewer's inquiries, and we acknowledge that both of your questions are closely related to the mathematical assumptions underlying collocation analysis. Therefore, we will address both issues in our response.

Firstly, it's important to clarify that Section 5.1, titled "Impact of underlying assumptions in collocation analysis," is intended to provide a detailed analysis of mathematical assumptions' impact on collocation analysis. It is not meant to be a direct comparison of evaporation datasets. In Section 5.1, we individually analyze their effects on the results. We emphasize the significance of non-zero ECC. This analysis naturally leads to Section 5.2, where we delve into a more comprehensive examination of ECC. Overall, we believe that the analysis in Sections 5.1 and 5.2 is quite thorough and adequately addresses the underlying assumptions of collocation analysis.

Secondly, concerning the issue of linear relationships, we would like to provide two points for clarification:

1. The relatively limited improvement observed in CAMELE with the merging scheme might be attributed to the fact that the initial set of inputs chosen for CAMELE already demonstrated good performance (as indicated in Figure 4). In this context, the merging framework effectively reduced errors. However, for further improvements, we acknowledge that incorporating regional ET products or site-specific data could enhance precision, as discussed in Section 5.4.
2. In collocation analysis, the consideration of non-linear relationships is primarily implemented through the multiplicative error model, involving a logarithmic transformation of inputs. However, such relationships have been more commonly identified in rainfall products **(Li et al., 2018)**, whereas collocation analysis in the context of ET products often indicates that linear relationships are reasonable **(Li et al., 2022; Park et al., 2023)**. ET products may contain systematic errors, and if

collocated anomalies are merged with reliable average values, it may yield more desirable data. However, the precondition for this improvement is the availability of reliable average values, as mentioned in the analysis presented in Section 5.1.

References:

Li, C., Tang, G., and Hong, Y.: Cross-evaluation of ground-based, multi-satellite and reanalysis precipitation products: Applicability of the Triple Collocation method across Mainland China, Journal of Hydrology, 562, 71–83, https://doi.org/10.1016/j.jhydrol.2018.04.039, 2018.

Li, C., Yang, H., Yang, W., Liu, Z., Jia, Y., Li, S., and Yang, D.: Error Characterization of Global Land Evapotranspiration Products: Collocation-based approach, Journal of Hydrology, 128102, 2022.

Park, J., Baik, J., and Choi, M.: Triple collocation-based multi-source evaporation and transpiration merging, Agricultural and Forest Meteorology, 331, 109353, 2023.

As per your feedback, we have integrated the above responses into the updated Section 5.1 to provide a more comprehensive and direct connection between the performance of CAMELE and the assumptions of collocation analysis. We hope this addresses your concerns adequately.

Revised contents (Line 884 to 897):

The linearity assumption shapes the error model by including additive and multiplicative biases and zero-mean random error. Although some studies have explored the application of a non-linear rescaling technique (Yilmaz and Crow, 2013; Zwieback et al., 2016), those efforts are primarily limited to soil moisture signals and often fail to accurately represent the true signal unless all datasets share a similar signal-to-noise ratio (SNR). However, it is worth noting that after rescaling processes, such as cumulative distribution function (CDF) matching or climatology removal, the resulting time series (anomalies) are often considered linearly related to the truth since higher-order error terms are removed. In addition, multiplicative relationships have been more commonly identified in rainfall products (Li et al., 2018). In contrast, collocation analysis within the context of ET products frequently suggests that linear relationships are reasonable (Li et al., 2022; Park et al., 2023). Therefore, the linear error model remains a robust implementation, though it has the potential for improvement through rescaling techniques.

2 AC to Referee #1: Minor Comments

2.1 Line 28

I would recommend caution in using qualitative terminology like "excellent performance." Additionally, I find this statement a bit misleading because the merged products performed closely to the products used in their merging. Please revise carefully the manuscript to avoid these overstatements.

AC:

We appreciate the reviewer's feedback and agree that the previous description lacked objectivity. We have now revised to adopt a neutral tone and have replaced "excellent" with "promising" in the original statement:

Revised contents (Line 28):

"CAMELE exhibits promising performance across various vegetation coverage types, as validated against in-situ observations."

2.2 Lines 58-59

The authors mention the following: "...previous research has predominantly focused on regional-scale ET estimation, necessitating a more straightforward and reliable global simulation method." It would be helpful for the authors to clarify what they mean by a "straightforward and reliable simulation method."

AC:

We appreciate the reviewer's suggestion. There was an inaccuracy in the description in question as previous research has not only focused on regional-scale ET but has also included gridded ET estimations. Since this section is not closely related to the surrounding content, we have removed it in the revised version.

2.3 Line 224

A space before the reference is missing. It should be added for proper formatting.

AC:

Thanks for the notification. We have revised it accordingly.

2.4 Line 254

Was the IGBP classification obtained from a dataset? If so, how were the functional types calculated? Do they change during the period of analysis? Please provide details regarding the source and methodology for classifying the functional types.

AC:

We greatly appreciate the reviewer's suggestion. The previous description was inaccurate. The IGBP information for each site was obtained from metadata provided on the FLUXNET official website sourced from observations made by the data providers at each site.

However, the official website did not provide a specific description of the methodology for classifying functional types at each site. Furthermore, information regarding changes in IGBP classifications for the sites was not publicly available. Our study utilized the latest FLUXNET 4.0 data available for download from the official website until February 2020. The data change log indicated, "No new sites, and for current sites, no new data, only new metadata." (Data Change Log - FLUXNET) As a result, it is hard to determine whether there were any changes in functional types at the sites during the study period.

We acknowledge the possibility of such changes and have revised the description accordingly to indicate that IGBP classifications were determined based on the metadata from the FluxNet official website, and changes during the study period, if any, are not publicly accessible. Interested parties can obtain relevant information by directly contacting the site coordinators.

Revised contents (Line 293-302):

"The International-Geosphere–Biosphere Program (IGBP) land cover classification system (Loveland et al., 1999) was employed to distinguish the 13 Plant Functional Types (PFTs) across sites. The IGBP classification was determined based on metadata from the FluxNet official website, including evergreen needle leaf forests (ENF, 49 sites), evergreen broadleaf forests (EBF, 15 sites), deciduous broadleaf forests (DBF, 26 sites), croplands (CRO, 20 sites), grasslands (GRA, 39 sites), savannas (SAV, nine sites), mixed forests (MF, nine sites), closed shrublands (CSH, three sites), deciduous needle leaf forests (DNF, one site), open shrublands (OSH, 13 sites), snow and ice (SNO, one site), and permanent wetland (WET, 21 sites). Changes in the IGBP classification during the study period are possible, but such information is not publicly available. Interested parties can obtain relevant information by directly contacting the site coordinators."

Relative contents in previous manuscript:

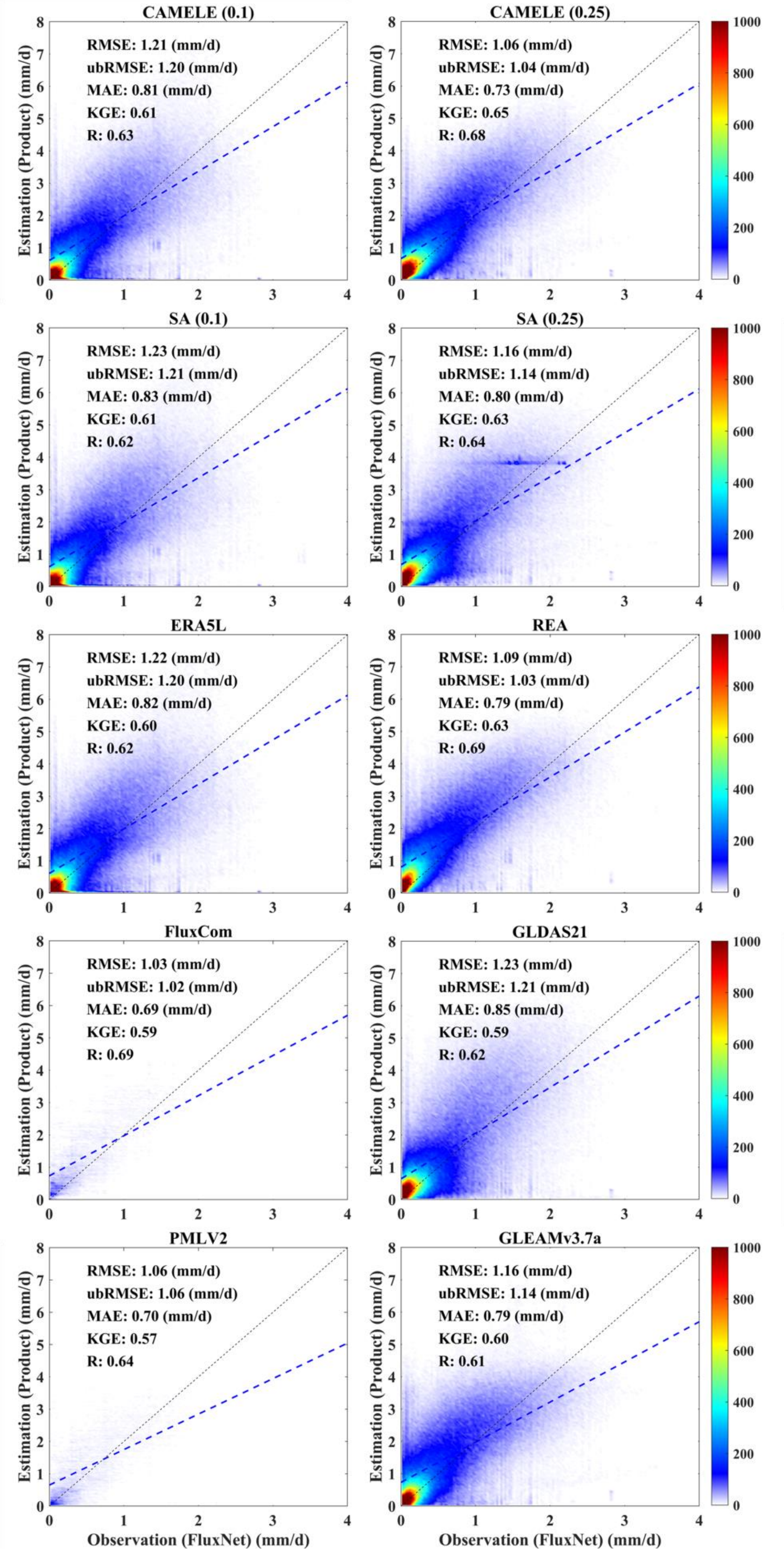
"The International-Geosphere–Biosphere Program (IGBP) land cover classification system (Loveland et al., 1999) was employed to distinguish the 13 Plant Functional Types (PFTs) across sites, including evergreen needle leaf forests (ENF, 49 sites), evergreen broadleaf forests (EBF, 15 sites), deciduous broadleaf forests (DBF, 26 sites), croplands (CRO, 20 sites), grasslands (GRA, 39 sites), savannas (SAV, nine sites), mixed forests (MF, nine sites), closed shrublands (CSH, three sites), deciduous needle leaf forests (DNF, one site), open shrublands (OSH, 13 sites), snow and ice (SNO, one site), and permanent wetland (WET, 21 sites)."

2.5 Figure 4

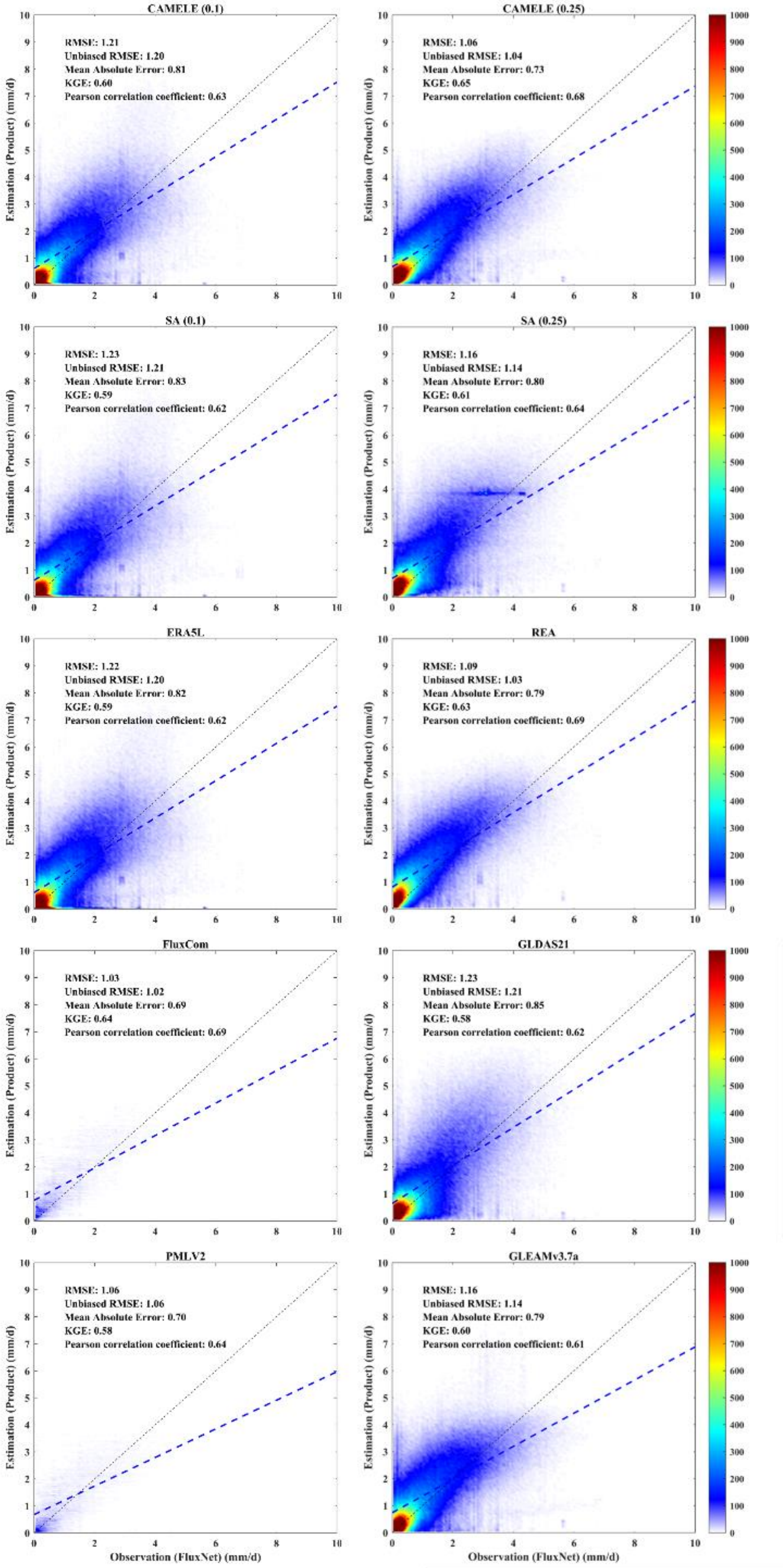
The quality of Figure 4 could be improved. Consider enhancing the clarity and readability of the figure. You might want to simplify the information presented or consider moving some details to a supplementary figure.

AC:

We appreciate the reviewer's feedback. We acknowledge that the original Figure 4 had issues with information overload, small font size, and suboptimal axis scaling. In response to these concerns, we have made the improvements to enhance the clarity and readability of Figure 4 (now is Figure 6) and better convey the intended information:


CAMELE (0.1)
RMSE: 1.21 (mm/d)
ubRMSE: 1.20 (mm/d)
MAE: 0.81 (mm/d)
KGE: 0.61
R: 0.63
CAMELE (0.25)
RMSE: 1.06 (mm/d)
ubRMSE: 1.04 (mm/d)
MAE: 0.73 (mm/d)
KGE: 0.65
R: 0.68
SA (0.1)
RMSE: 1.23 (mm/d)
ubRMSE: 1.21 (mm/d)
MAE: 0.83 (mm/d)
KGE: 0.61
R: 0.62
SA (0.25)
RMSE: 1.16 (mm/d)
ubRMSE: 1.14 (mm/d)
MAE: 0.80 (mm/d)
KGE: 0.63
R: 0.64
ERA5L
RMSE: 1.22 (mm/d)
ubRMSE: 1.20 (mm/d)
MAE: 0.82 (mm/d)
KGE: 0.60
R: 0.62
REA
RMSE: 1.09 (mm/d)
ubRMSE: 1.03 (mm/d)
MAE: 0.79 (mm/d)
KGE: 0.63
R: 0.69
FluxCom
RMSE: 1.03 (mm/d)
ubRMSE: 1.02 (mm/d)
MAE: 0.69 (mm/d)
KGE: 0.59
R: 0.69
GLDAS21
RMSE: 1.23 (mm/d)
ubRMSE: 1.21 (mm/d)
MAE: 0.85 (mm/d)
KGE: 0.59
R: 0.62
PMLV2
RMSE: 1.06 (mm/d)
ubRMSE: 1.06 (mm/d)
MAE: 0.70 (mm/d)
KGE: 0.57
R: 0.64
GLEAMv3.7a
RMSE: 1.16 (mm/d)
ubRMSE: 1.14 (mm/d)
MAE: 0.79 (mm/d)
KGE: 0.60
R: 0.61
Estimation (Product) (mm/d)
Observation (FluxNet) (mm/d)
0 1 2 3 4
0 1 2 3 4 5 6 7 8
0 200 400 600 800 1000

Previous Figure 4 for comparison:



2.6 Line 557

The authors mention that based on the results of Figure 4, CAMELE performs well at 0.10 and 0.25 degrees, and all products have similar performance. The phrase "performs well" may sound like it performs better compared to other products, which could be misleading. Consider rephrasing this to clarify that CAMELE performs similarly to other products.

AC:

We appreciate the valuable feedback from the reviewer. We have revised the text to avoid any potential confusion. The term "performs well" has been replaced with "exhibited consistent performance" to clarify that CAMELE's performance is like that of other products. We believe these changes enhance the accuracy and clarity of our findings:

Revised contents (Line 623-635):

The scatter plots in Figure 6 demonstrate that CAMELE consistently performs at 0.1° and 0.25° resolutions. At 0.1° resolution, FluxCom and PMLv2 showed superior performance with fewer data points due to their original 8-day average resolution. CAMELE exhibited a performance like ERA5L. At 0.25° resolution, CAMELE performed comparably to the other datasets, demonstrating reasonable accuracy. Notably, there was an improvement in the KGE and R indices. The fitted line closely approximated the 1:1 line, indicating a solid agreement with the observed values. Moreover, the results obtained from the simple average were also acceptable, but SA (0.25°) had a concentration of data points between (2-4 mm/d), possibly due to the inputs having a high concentration within that range. The assumption that a simple average implies equal performance of each product on every grid cell is inaccurate; variations in performance exist among different products across distinct grid cells (regions).

2.7 Line 585:

While it's understandable that the authors want to promote their product, it might seem a bit odd to say that CAMELE performs exceptionally well and closely resembles two of the products used in the merging scheme. The desired outcome in merging datasets is to outperform the products used in the merging procedure. Consider rephrasing this to maintain objectivity.

AC:

We appreciate the reviewer's feedback. The previous description lacked objectivity, and we have made the suggested changes to convey the results better. The revised statement emphasizes the improvement in performance without directly comparing CAMELE to the merging scheme products, ensuring a more balanced and objective representation of our findings.

Revised contents (Line 655 to 658):

"CAMELE demonstrates a notable enhancement in performance at the 0.1° level. This suggests that the fusion method effectively reduces errors, aligning with the original intention of weight calculation, and it compares favorably with the products used in the merging scheme."

2.8 Figure 6:

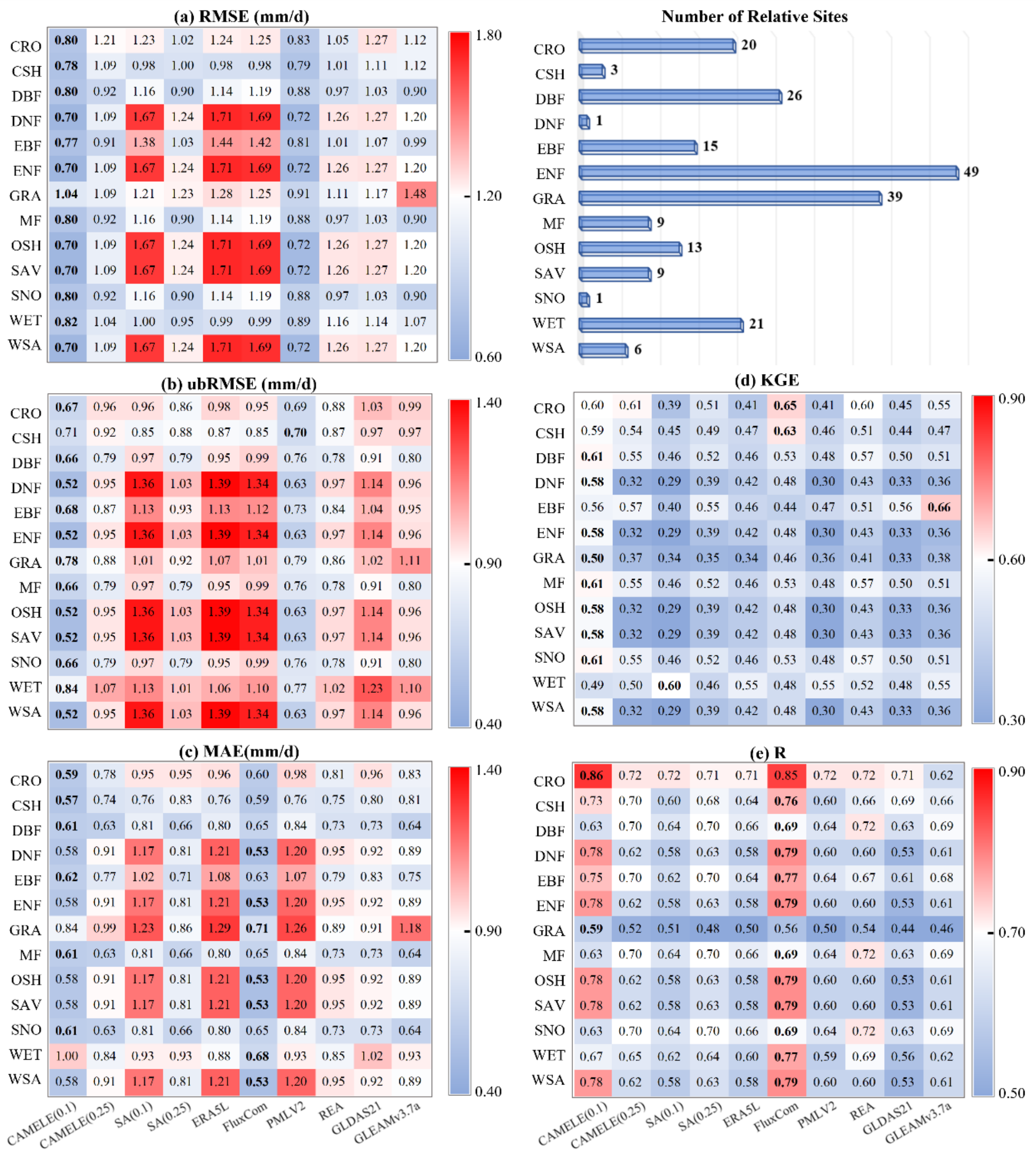
The quality of Figure 6 could be improved for better clarity and readability. Consider reducing the information presented in the figure or moving some details to a supplementary figure. Another option is to highlight the top three performing products for each IGBP class with color coding.

AC:

We greatly appreciate the reviewer's suggestions. In response to the comments, we have made the following improvements to Figure 6 (now Figure 8):

- We have changed the color bar in the original figure to a more distinct red-blue color scheme.
- We have added labels to each subfigure and provided corresponding explanations in the figure captions.
- We have highlighted the top-performing product in each row (corresponding to a specific PFT) with bold formatting. We chose this based on our experimentation, as highlighting the top three products made the information too cluttered and less reader-friendly.

The modified figure and its captions are presented below. We hope that this revised version conveys the information more clearly:



(a) RMSE (mm/d)

	CAMELE(0.1)	CAMELE(0.25)	SA(0.1)	SA(0.25)	ERA5L	FluxCom	PMLV2	REA	GLDAS21	GLEAMv3.7a
CRO	**0.80**	1.21	1.23	1.02	1.24	1.25	0.83	1.05	1.27	1.12
CSH	**0.78**	1.09	0.98	1.00	0.98	0.98	0.79	1.01	1.11	1.12
DBF	**0.80**	0.92	1.16	0.90	1.14	1.19	0.88	0.97	1.03	0.90
DNF	**0.70**	1.09	1.67	1.24	1.71	1.69	0.72	1.26	1.27	1.20
EBF	**0.77**	0.91	1.38	1.03	1.44	1.42	0.81	1.01	1.07	0.99
ENF	**0.70**	1.09	1.67	1.24	1.71	1.69	0.72	1.26	1.27	1.20
GRA	**1.04**	1.09	1.21	1.23	1.28	1.25	0.91	1.11	1.17	1.48
MF	**0.80**	0.92	1.16	0.90	1.14	1.19	0.88	0.97	1.03	0.90
OSH	**0.70**	1.09	1.67	1.24	1.71	1.69	0.72	1.26	1.27	1.20
SAV	**0.70**	1.09	1.67	1.24	1.71	1.69	0.72	1.26	1.27	1.20
SNO	**0.80**	0.92	1.16	0.90	1.14	1.19	0.88	0.97	1.03	0.90
WET	**0.82**	1.04	1.00	0.95	0.99	0.99	0.89	1.16	1.14	1.07
WSA	**0.70**	1.09	1.67	1.24	1.71	1.69	0.72	1.26	1.27	1.20

(b) ubRMSE (mm/d)

	CAMELE(0.1)	CAMELE(0.25)	SA(0.1)	SA(0.25)	ERA5L	FluxCom	PMLV2	REA	GLDAS21	GLEAMv3.7a
CRO	**0.67**	0.96	0.96	0.86	0.98	0.95	0.69	0.88	1.03	0.99
CSH	0.71	0.92	0.85	0.88	0.87	0.85	**0.70**	0.87	0.97	0.97
DBF	**0.66**	0.79	0.97	0.79	0.95	0.99	0.76	0.78	0.91	0.80
DNF	**0.52**	0.95	1.36	1.03	1.39	1.34	0.63	0.97	1.14	0.96
EBF	**0.68**	0.87	1.13	0.93	1.13	1.12	0.73	0.84	1.04	0.95
ENF	**0.52**	0.95	1.36	1.03	1.39	1.34	0.63	0.97	1.14	0.96
GRA	**0.78**	0.88	1.01	0.92	1.07	1.01	0.79	0.86	1.02	1.11
MF	**0.66**	0.79	0.97	0.79	0.95	0.99	0.76	0.78	0.91	0.80
OSH	**0.52**	0.95	1.36	1.03	1.39	1.34	0.63	0.97	1.14	0.96
SAV	**0.52**	0.95	1.36	1.03	1.39	1.34	0.63	0.97	1.14	0.96
SNO	**0.66**	0.79	0.97	0.79	0.95	0.99	0.76	0.78	0.91	0.80
WET	**0.84**	1.07	1.13	1.01	1.06	1.10	0.77	1.02	1.23	1.10
WSA	**0.52**	0.95	1.36	1.03	1.39	1.34	0.63	0.97	1.14	0.96

(c) MAE(mm/d)

	CAMELE(0.1)	CAMELE(0.25)	SA(0.1)	SA(0.25)	ERA5L	FluxCom	PMLV2	REA	GLDAS21	GLEAMv3.7a
CRO	**0.59**	0.78	0.95	0.95	0.96	0.60	0.98	0.81	0.96	0.83
CSH	**0.57**	0.74	0.76	0.83	0.76	0.59	0.76	0.75	0.80	0.81
DBF	**0.61**	0.63	0.81	0.66	0.80	0.65	0.84	0.73	0.73	0.64
DNF	0.58	0.91	1.17	0.81	1.21	**0.53**	1.20	0.95	0.92	0.89
EBF	**0.62**	0.77	1.02	0.71	1.08	0.63	1.07	0.79	0.83	0.75
ENF	0.58	0.91	1.17	0.81	1.21	**0.53**	1.20	0.95	0.92	0.89
GRA	0.84	0.99	1.23	0.86	1.29	**0.71**	1.26	0.89	0.91	1.18
MF	**0.61**	0.63	0.81	0.66	0.80	0.65	0.84	0.73	0.73	0.64
OSH	0.58	0.91	1.17	0.81	1.21	**0.53**	1.20	0.95	0.92	0.89
SAV	0.58	0.91	1.17	0.81	1.21	**0.53**	1.20	0.95	0.92	0.89
SNO	**0.61**	0.63	0.81	0.66	0.80	0.65	0.84	0.73	0.73	0.64
WET	1.00	0.84	0.93	0.93	0.88	**0.68**	0.93	0.85	1.02	0.93
WSA	0.58	0.91	1.17	0.81	1.21	**0.53**	1.20	0.95	0.92	0.89

(d) KGE

	CAMELE(0.1)	CAMELE(0.25)	SA(0.1)	SA(0.25)	ERA5L	FluxCom	PMLV2	REA	GLDAS21	GLEAMv3.7a
CRO	0.60	0.61	0.39	0.51	0.41	**0.65**	0.41	0.60	0.45	0.55
CSH	0.59	0.54	0.45	0.49	0.47	**0.63**	0.46	0.51	0.44	0.47
DBF	**0.61**	0.55	0.46	0.52	0.46	0.53	0.48	0.57	0.50	0.51
DNF	**0.58**	0.32	0.29	0.39	0.42	0.48	0.30	0.43	0.33	0.36
EBF	0.56	0.57	0.40	0.55	0.46	0.44	0.47	0.51	0.56	**0.66**
ENF	**0.58**	0.32	0.29	0.39	0.42	0.48	0.30	0.43	0.33	0.36
GRA	**0.50**	0.37	0.34	0.35	0.34	0.46	0.36	0.41	0.33	0.38
MF	**0.61**	0.55	0.46	0.52	0.46	0.53	0.48	0.57	0.50	0.51
OSH	**0.58**	0.32	0.29	0.39	0.42	0.48	0.30	0.43	0.33	0.36
SAV	**0.58**	0.32	0.29	0.39	0.42	0.48	0.30	0.43	0.33	0.36
SNO	**0.61**	0.55	0.46	0.52	0.46	0.53	0.48	0.57	0.50	0.51
WET	0.49	0.50	**0.60**	0.46	0.55	0.48	0.55	0.52	0.48	0.55
WSA	**0.58**	0.32	0.29	0.39	0.42	0.48	0.30	0.43	0.33	0.36

(e) R

	CAMELE(0.1)	CAMELE(0.25)	SA(0.1)	SA(0.25)	ERA5L	FluxCom	PMLV2	REA	GLDAS21	GLEAMv3.7a
CRO	**0.86**	0.72	0.72	0.71	0.71	0.85	0.72	0.72	0.71	0.62
CSH	0.73	0.70	0.60	0.68	0.64	**0.76**	0.60	0.66	0.69	0.66
DBF	0.63	0.70	0.64	0.70	0.66	**0.69**	0.64	0.72	0.63	0.69
DNF	0.78	0.62	0.58	0.63	0.58	**0.79**	0.60	0.60	0.53	0.61
EBF	0.75	0.70	0.62	0.70	0.64	**0.77**	0.64	0.67	0.61	0.68
ENF	0.78	0.62	0.58	0.63	0.58	**0.79**	0.60	0.60	0.53	0.61
GRA	**0.59**	0.52	0.51	0.48	0.50	0.56	0.50	0.54	0.44	0.46
MF	0.63	0.70	0.64	0.70	0.66	**0.69**	0.64	0.72	0.63	0.69
OSH	0.78	0.62	0.58	0.63	0.58	**0.79**	0.60	0.60	0.53	0.61
SAV	0.78	0.62	0.58	0.63	0.58	**0.79**	0.60	0.60	0.53	0.61
SNO	0.63	0.70	0.64	0.70	0.66	**0.69**	0.64	0.72	0.63	0.69
WET	0.67	0.65	0.62	0.64	0.60	**0.77**	0.59	0.69	0.56	0.62
WSA	0.78	0.62	0.58	0.63	0.58	**0.79**	0.60	0.60	0.53	0.61

Figure.8 Heatmaps of five statistical indicators, where each row corresponds to the mean value for all sites of the specific PFT, and each column corresponds to a product. The product with the best performance for that PFT is highlighted in bold within each row. (a)-(c) represent three error indicators: RMSE, ubRMSE, and MAE; (d)-(e) represent two goodness-of-fit indicators: KGE and R.

Previous Figure 6 for comparison:

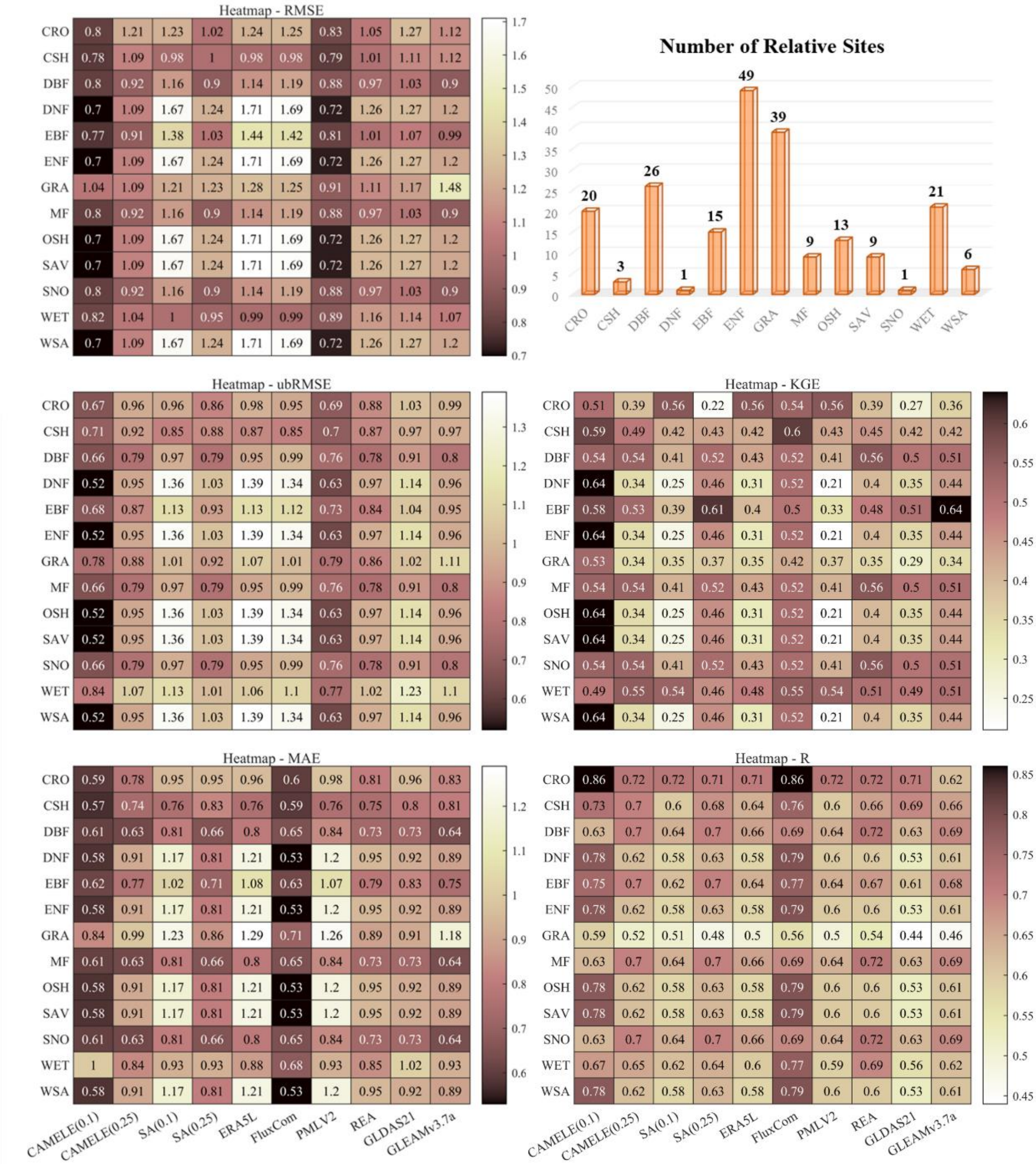


Figure.6 Heatmap of five indicators calculated separately for each site, classified by PFTs. The top right corner indicates the number of sites corresponding to each type.

2.9 Line 863:

Remove the word "excellent" from this line to maintain a more neutral tone.

AC:

Thank you for your suggestion. Updated:

Revised Content (Line 1083 to 1085):

"Although FluxCom and PMLv2 performed slightly better than CAMELE at some PFT sites, considering that both utilized FluxNet sites for product calibration, it indirectly demonstrates the promising performance of CAMELE."

➢ AC to Referee #2: General Comment

The manuscript aims to develop a new evapotranspiration (ET) product using collocation techniques. It has the potential to be a useful contribution to the literature and to the broad userbase of ET products. Nonetheless, a few major issues need to be addressed before it can be considered for publication.

AC:

We greatly appreciate the professional and constructive feedback provided by the reviewer. We will respond to each comment individually, and in the following responses, the line numbers corresponding to the added or revised content will be based on the updated version without highlights. You can open the PDF file's table of contents view to navigate to the relevant sections directly.

The responses will be in the following format:

- Reviewer's comments are shown in black.
- Our responses are shown in blue.
- The modifications to the manuscript are shown in orange.
- Previous contents in the old version (for comparison if needed) are shown in grey.

1 AC to Referee #2: Major Comments

1.1 Q1

First, the construction of the products should be justified by a clearly outlined rationale. The new ET product is built from multiple ET solutions with different temporal coverage. How are those individual products selected for each analyzed period, and when three instead of two products are selected, what is the corresponding gain in terms of performance?

AC:

Thank you for your inquiry. We have provided justification and description for the product selection in the beginning of the datasets section, based on three considerations: (1) Maintaining consistent original spatiotemporal resolution among the products to minimize potential downscaling operations and avoid introducing additional errors; (2) Ensuring three or more products within the same resolution or period, aligning with the collocation method where lag-1 sequences from two products, are typically selected as the third input, aiming to incorporate more information for effective fusion; (3) Choosing products with relatively high visibility, widespread usage, and global coverage. In addition, we also address the existence of high-resolution regional ET product, which could be used for further update of CAMELE.

The relevant explanations have been added to the beginning of the datasets section:

Revised Contents (Line 145 to 159) :

"…We selected five widely used ET products that spanned the period from 1980 to 2022. When selecting these products, our aims are to ensure: (1) consistency in original spatiotemporal resolution among the products: minimize potential downscaling operations and avoid introducing additional errors; (2) having three or more products within the same resolution or period: incorporate more information for effective fusion; (3) products with extensive global observational sequences: gain basic recognition from the community. While we acknowledge the existence of other higher-precision products, their integration would require either downscaling or upscaling other products, potentially introducing uncertainties. Therefore, we chose the combination outlined in the manuscript. Despite its relatively lower resolution compared to some products, it still contributes to our understanding of ET variations, facilitating advantageous exploration. Furthermore, we incorporated in-situ observations and Lu et al. (2021) 's global 0.25° daily-scale ET product derived using Reliability Ensemble Averaging (denoted as REA) to compare our merged product comprehensively…"

1.2 Q2

Second, the manuscript demonstrates the consistency of the proposed product with its peers, but I believe it is more important to highlight the unique strength and weakness of the new product. When and where does the new product outperform or underperform its peers? Does it improve upon its individual constituents in terms of characterizing the long-term trend, seasonality, inter-annual variability, or the extremes of ET?

AC:

Thank you for your valuable suggestions. We have conducted further analysis of the performance of the CAMELE product and have emphasized several strengths:

1. It effectively captures the multi-year linear trend.
2. Enhances the accuracy of estimating multi-year mean values.
3. Better characterizes extreme values of ET (5th and 95th percentiles at monthly scale).

We have also acknowledged the limitations of CAMELE:

1. lower resolution compared to regional high-resolution ET products, limiting its potential for regional analysis.
2. potential overestimation of seasonality.

We have included additional analyses in the results section 4.3 and 4.4 (new subsections) focusing on trend, seasonality, multi-year average, and extreme values to address these aspects. A new section discussing future improvements has been added (5.4, new subsection). Modifications have been made to the abstract and conclusion to reflect these changes.

1.2.1 Revision in Abstract (Line 34 to 38)

"…In addition, comparisons indicate that CAMELE can effectively characterize the multi-year linear trend, mean average, and extreme values of ET. However, it exhibits a tendency to overestimate seasonality. In summary, we propose a reliable set of ET data that can aid in understanding the variations in the water cycle…"

1.2.2 New Contents regarding regional ET data (2 Datasets, Line 151 to 156):

"…While we acknowledge the existence of other higher-precision products, their integration would require either downscaling or upscaling other products, potentially introducing uncertainties. Therefore, we chose the combination outlined in the manuscript. Despite its relatively lower resolution compared to some products, it still contributes to our understanding of ET variations, facilitating advantageous exploration…"

1.2.3 New Contents about multi-year mean and extreme ET value (4.3 Results, Line 732 to 761):

"…For site comparisons, we have selected monthly mean ET values and three quantiles (5th, 50th, and 95th) to represent the products' performance in estimating ET's average and extreme values.

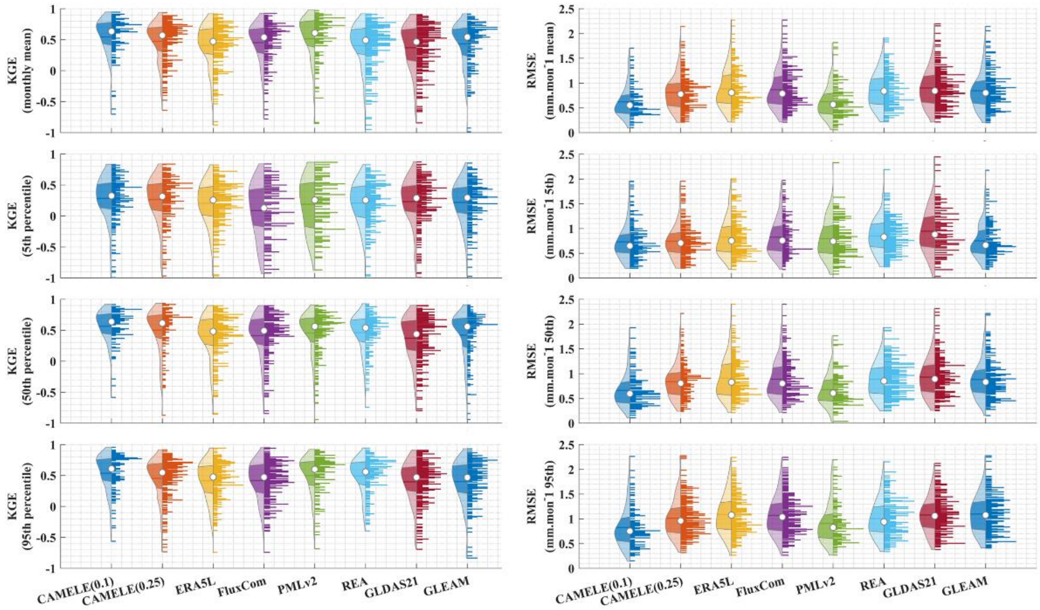


Figure 10 Violin plots depicting the KGE and RMSE metrics calculated for CAMELE and other products based on the monthly mean, 5th, 50th, and 95th percentiles at each FluxNet site. The left four columns represent KGE plots, while the right four columns represent RMSE plots. The dots in the violin plots represent the median, and the horizontal lines represent the mean.

Table 6 Average values of KGE and RMSE corresponding to different products, calculated based on the results obtained for each site. The bolded sections indicate the schemes with the best performance in their respective metrics.

Product		**KGE**			
		Mean	**5th**	**50th**	**95th**
0.1°-daily	**CAMELE**	**0.54**	**0.28**	**0.57**	**0.54**
	ERA5L	0.41	0.21	0.40	0.42
	FluxCom	0.45	0.09	0.42	0.42
	PMLv2	0.52	0.19	0.46	0.50
0.25°-daily	**CAMELE**	**0.47**	**0.26**	**0.50**	0.45
	REA	0.40	0.21	0.46	**0.50**
	GLDAS21	0.37	0.23	0.37	0.40
	GLEAMv3.7a	0.43	0.22	0.42	0.40
Product		**RMSE (mm/mon)**			
		Mean	**5th**	**50th**	**95th**

0.1°-daily	**CAMELE**	0.63	**0.73**	**0.66**	**0.83**
	ERA5L	0.89	0.83	0.91	1.09
	FluxCom	0.87	0.83	0.89	1.07
	PMLv2	**0.63**	0.80	0.68	0.91
0.25°-daily	**CAMELE**	**0.81**	**0.74**	**0.84**	1.01
	REA	0.86	0.85	0.88	**1.01**
	GLDAS21	0.90	0.95	0.93	1.08
	GLEAMv3.7a	0.85	0.75	0.88	1.10

The information in Figure 10 corresponds to the data presented in Table 6, which involves the calculation of KGE and RMSE at each site, followed by statistical analysis. From the distribution of the violin plots, it can be observed that a violin plot with a closer belly to 1 indicates better results in terms of the KGE.

The results show that CAMELE outperforms other products in the estimation of monthly averages and the 5th, 50th, and 95th percentiles at both 0.1° and 0.25° resolutions. Its performance in capturing monthly averages is noteworthy, with a noticeable improvement in the KGE and RMSE metrics relative to the inputs. Examining the results for percentiles, CAMELE shows a relatively poorer estimation for shallow values (5th percentile) but still demonstrates some improvement compared to the input data, albeit influenced by input errors.

At 0.1°, PMLv2 and FluxCom perform just below the fusion result, aligning with the previous error and weight analysis. At 0.25°, GLEAM and REA closely follow CAMELE, with REA exhibiting slightly better estimation results for extremely high values (95th percentile) than CAMELE. Despite this, the analysis results still indicate that the products obtained reflect well the multi-year averages and extremes of ET, holding promise as reliable products for analyzing ET variations…”

1.2.4 New Contents about multi-year trend and seasonality (4.4 Results, Line 798 to 846):

“…In this section, we first validate and compare the performance of CAMELE with other products in estimating multi-year trends and seasonality at the site scale. Due to the inconsistent time lengths of FluxNet sites, trends at many sites are not significant. Therefore, we deliberately selected 13 sites with continuous evapotranspiration (ET) observations for the same 11-year period (2004 to 2014) and with significant trends. The annual ET values for each year were calculated as the mean of the 13 sites for that year, allowing the computation of linear trends and seasonality. We employed singular spectrum analysis (SSA), which assumes an additive decomposition A = LT + ST + R. In this decomposition, LT represents the long-term trend in the data, ST is the seasonal or oscillatory trend (or trends), and R is the remainder.

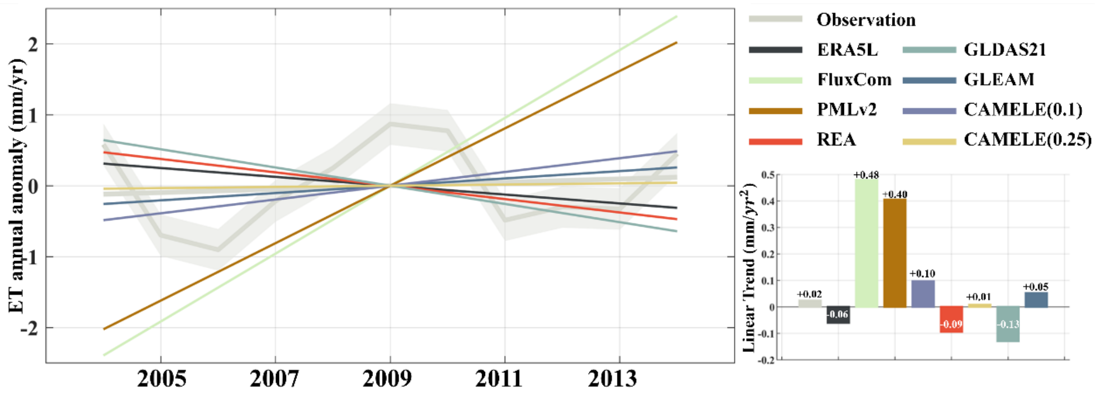


Figure 13 Comparison of linear trend from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The trends have been subjected to SSA decomposition, removing seasonality. The gray enveloping line represents the mean plus the standard deviation of the 13 sites.

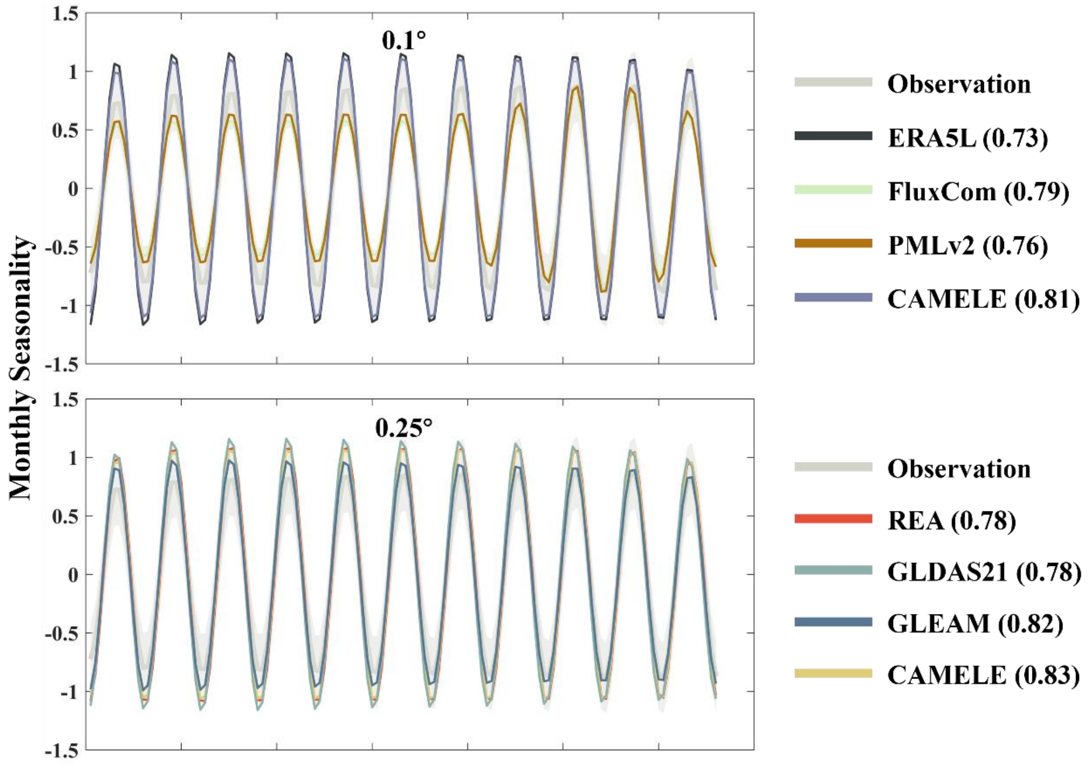


Figure 14 Comparison of seasonal variations from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The seasonality has been obtained through SSA decomposition, with the gray area representing the observed values. The parentheses in each product name indicate the KGE coefficient comparing with the observed values.

In Figure 13 and Figure 14, based on observations from FluxNet sites, we analyzed the performance of CAMELE and other products in estimating the linear trend and seasonality of ET over multiple years. It is important to note that we only present the analysis results for 13 sites with continuous 11-year observations, and the performance of different ET products in trend estimation at individual sites still varies, not fully reflecting the overall performance on all grids in terms of trend and seasonality. Nevertheless, such a comparison can still provide valuable insights.

Examining the results of the linear trend, both PMLv2 and FluxCom exhibit a significant upward trend, well above the observations. On the contrary, ERA5L, GLDAS, and REA show a noticeable downward trend, while CAMELE demonstrates a gradual upward trend closer to the observations. Additionally, GLEAM slightly outperforming CAMELE at a resolution of 0.25°. Overall, CAMELE shows good agreement with site observations in capturing the multi-year linear trend of ET.

Continuing with the analysis of seasonality, the KGE index comparing each product's results with observed values is provided in parentheses next to the product name. Generally, all products exhibit a good representation of ET's seasonal variations. CAMELE's 0.1° seasonal results closely match FluxCom (with the two lines almost overlapping). However, the fluctuations it reflects are higher than the observed values. This is likely due to keeping the 8-day average results of FluxCom consistent with PMLv2 every 8 days, and the variability in ET primarily originates from ERA5L results. This aspect may need improvement in subsequent research. At 0.25°, CAMELE's seasonal representation is closer to the observed results. The differences in CAMELE's performance at the two resolutions are mainly attributed to input variations, which we discuss in the following section as potential areas for improvement.

The results indicate that CAMELE effectively captures the multi-year changes in ET, but at 0.1°, it tends to overestimate seasonal fluctuations…”

1.2.5 New Contents for future update (5.4 Discussion, Line 1034 to 1060):

“…In this section, we delve into the potential applications of our product and outline our commitment to future enhancements to maintain its accuracy and relevance.

Here, we identify three potential applications for our transpiration product: (1) Global ET Trends: Our product facilitates global-scale analysis of current ET patterns and long-term trends, essential for comprehending ecosystem responses to evolving environmental conditions in a warming climate; (2) Transpiration-to-Evapotranspiration Ratio: Our merging approach can fuse multi-source global gridded transpiration data, allowing for the examination of the transpiration-to-evapotranspiration ratio. This analysis can enhance water resource management and water availability predictions in diverse regions; (3) Attribution analysis: Our product is a valuable tool for attribution analysis, helping researchers identify the drivers of patterns. This knowledge is crucial for understanding the roles of climate variability, land-use changes, and other factors in shaping terrestrial water fluxes.

Furthermore, we are committed to enhancing our product proactively. Key strategies include: (1) Data Update and Validation: To ensure our product's continued accuracy and reliability, we will prioritize regularly updating the data used in this study to the latest versions. By adopting this approach, we aim to provide users with results that reflect the latest advancements in scientific knowledge; (2) Enhanced Integration and Error Reduction: We continually refine estimates by incorporating additional data sources and implementing extended collocation method to minimize errors; (3) Integration of High-Resolution Regional ET Data: Recognizing the significance of regional-scale insights, we will focus on improving the accuracy of CAMELE by integrating higher-resolution regional ET data. This integration will enable more precise regional estimation.

In summary, these endeavors collectively represent our commitment to maintaining our product's quality and relevance, ensuring its value for the scientific community…”

1.3 Q3

Third, the organization of the manuscript can be improved, too. I list some of my suggestions in the detailed comments below. For example, the discussion of the non-zero ECC spreads across two subsections in the Discussion, and I think they should be merged and moved to the result section. I think addressing these issues will strengthen the scientific robustness of the manuscript and facilitate the adoption of the new product.

AC:

Thank you very much for your detailed suggestions. We will address each of your points in the "detailed comments" section.

2 AC to Referee #2: Detailed Comments

2.1 Line 173-175

L173-175. The rationale of choosing GLDAS-2.0/2.1 is questionable. Given the difference in the underlying modeling/reanalysis schemes, the error structures of GLDAS and ECMWF-based ET estimates are inherently different regardless of what sets of meteorological forcing are used. When selecting the GLDAS products, one could arguably choose the ET products that are driven by the more reliable forcing.

AC:

Thank you very much for your valuable suggestion. We acknowledge the significant differences between different versions of GLDAS-2, and our choice here was aimed at covering the period from 1980 to 2022. We have added explanations regarding the differences in GLDAS-2 versions and cited recent literature highlighting non-zero ECC between GLDAS2.2 and ERA5L.

Revised Contents (Line 199 to 213):

"…This study aimed to cover the research period from 1980 to 2022. Non-zero ECC between the transpiration estimates of GLDAS-2.2 and ERA5L has been reported in a recent study (Li et al., 2023a). Considering the similarities in the calculation of ET and transpiration of GLDAS and ERA5L, this report partially indicates a correlation. Therefore, GLDAS-2.0 and GLDAS-2.1 were selected as inputs instead. The "Evap_tavg" parameter representing evapotranspiration is derived from the original products and aggregated to a daily scale. For more detailed information on the GLDAS-2 models, please refer to NASA's Hydrology Data and Information Services Center at http://disc.sci.gsfc.nasa.gov/hydrology.

Despite the same forcing between GLDAS-2.1 and GLDAS-2.2, significant differences exist between the model results of different GLDAS versions (Qi et al., 2020, 2018; Jiménez et al., 2011). The non-zero ECC will generally still be met between different versions. Thus, we still need to analyze the non-zero ECC situations between ERA5L and GLDAS-2.0 and 2.1, which will be assessed in the discussion sections…"

References:

Li, C., Liu, Z., Tu, Z., Shen, J., He, Y., and Yang, H.: Assessment of global gridded transpiration products using the extended instrumental variable technique (EIVD), Journal of Hydrology, 623, 129880, https://doi.org/10.1016/j.jhydrol.2023.129880, 2023a

Jiménez, C., Prigent, C., Mueller, B., Seneviratne, S. I., McCabe, M. F., Wood, E. F., Rossow, W. B., Balsamo, G., Betts, A. K., Dirmeyer, P. A., Fisher, J. B., Jung, M., Kanamitsu, M., Reichle, R. H., Reichstein, M., Rodell, M., Sheffield, J., Tu, K., and Wang, K.: Global intercomparison of 12 land surface heat flux estimates, J. Geophys. Res., 116, D02102, https://doi.org/10.1029/2010JD014545, 2011.

Qi, W., Liu, J., and Chen, D.: Evaluations and Improvements of GLDAS2.0 and GLDAS2.1 Forcing Data's Applicability for Basin Scale Hydrological Simulations in the Tibetan Plateau, JGR Atmospheres, 123, https://doi.org/10.1029/2018JD029116, 2018.

Qi, W., Liu, J., Yang, H., Zhu, X., Tian, Y., Jiang, X., Huang, X., and Feng, L.: Large Uncertainties in Runoff Estimations of GLDAS Versions 2.0 and 2.1 in China, Earth and Space Science, 7, e2019EA000829, https://doi.org/10.1029/2019EA000829, 2020.

2.2 Line 252

L252. It would be ideal to show the distribution of the selected sites on a map.

AC:

Thank you very much for your suggestion. Updated.

New Figure (Line 303 to 304):

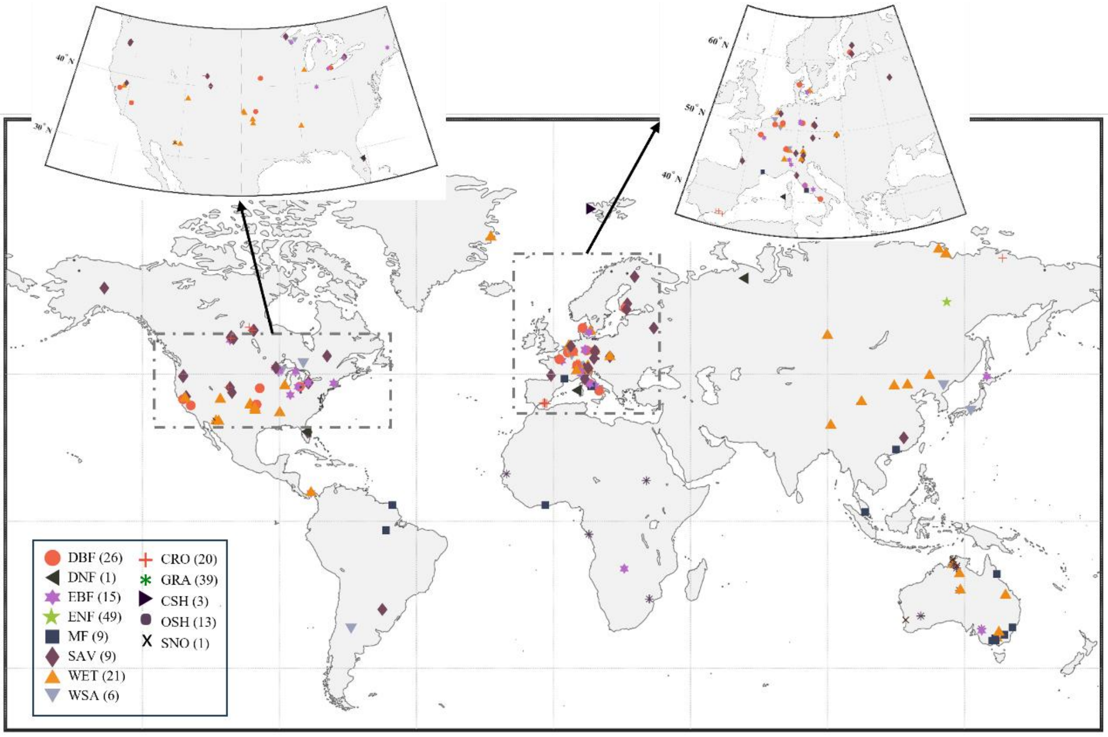


Figure 1 Global distribution of selected FluxNet Sites.

2.3 Line 271

L271. The methodological detail for Sections 3.1 can be trimmed as they are widely available. Highlighting aspects that are either implemented or discussed in this study would be sufficient, e.g. the assumptions, especially regarding the cross-correlated errors.

AC:

Thank you very much for your suggestion. We have trimmed approximately 25% of the content in Section 3.1, retaining essential formulas and crucial mathematical assumptions discussed in this study.

2.4 Line 393-400

L393-L400. This should go to the intro. Overall the method session needs to focus on clarifying the rationale of the chosen methodology and how they are directly implemented for this study.

AC:

Thank you very much for your suggestion. We have revised and relocated the content to the introduction section as recommended.

Revised Contents (Line 103 to 112):

"…Moreover, error information derived from collocation analysis is valuable for merging multi-source data. This was initially applied by Yilmaz et al. (2012) in the fusion of multi-source soil moisture products and later improved by Gruber et al. (2017) and further applied in the production of the European Space Agency Climate Change Initiative (ESA CCI) global soil moisture product (Gruber et al., 2019). Dong et al. (2020b) also adopted this approach to fusing multi-source precipitation products. In the study of evapotranspiration, Li et al. (2023c) and Park et al.(2023) utilized a weight calculation method that does not consider non-zero ECC and fused multiple ET products in the Nordic and East Asia, respectively, achieving satisfactory fusion results…"

2.5 Line 430

L430. One of the PMLv2 should be GLDAS.

AC:

Thank you for pointing out the mistake. Updated.

Revised Contents (Line 461):

"…analyze the performance of five sets of ET products (ERA5L/ PMLv2/FluxCom/GLDAS2/GLEAMv3) at the global scale…"

2.6 Line 430-439

L430-439. I understand this is a prior work that is directly related to this study, but the summary of this prior finding should go to the intro. Only the key assumptions made based on this prior work need to be highlighted in the method section (in this case, the non-zero ECC pairs).

AC:

Thank you for your suggestion. Firstly, the conclusions of this prior work are mentioned in the Introduction:

Unchanged Contents (Line 116):

"…Li et al. (2022) global ET product evaluation research revealed clear non-zero ECC conditions between ERA5L, GLEAM, PMLv2, and FluxCom…"

Considering that this part of the Introduction primarily discusses the application of collocation methods in ET, we chose to leave out non-zero ECC pairs in this section. Secondly, the current placement in the Method section, specifically in the Combinations subsection, is deemed appropriate.

2.7 Line 454

L454. I don't fully understand the rationale of grouping different products within each scenario. For Scenario 1, e.g., is the goal here to include as many products as possible within a given period? If that's the case, it will be helpful to clarify the gain in terms of performance of doing so.

AC:

Thank you for your valuable feedback. In response to your **Major Comments Q1**, we have updated the Datasets section to clarify product selection and matching. Additionally, we have revised the corresponding explanation following the table to specify the goal of including three or more products whenever possible. This aims to optimize the performance within a given period.

Revised Contents (Line 486 to 493):

"…It should be noted that the same product can have different versions. In this study, appropriate versions are selected based on the following principles: (1) Selection based on the corresponding data coverage duration and ensuring more products to gain more information; (2) Choosing the latest version while considering the assumption of non-zero ECC conditions; (3) Making efforts to select the exact product versions for different periods, to avoid uncertainties caused by version changes. We selected a subset of sites to compare the fusion results using different versions, and the corresponding details will be presented in the discussion section…"

2.8 Line 477-484

L477-484. This should go to the Method section.

AC:

Many thanks to your suggestions. This part has been moved to the Datasets section.

Revised Contents (Line 156 to 159):

"…Furthermore, we incorporated in-situ observations and Lu et al. (2021) 's global 0.25° daily-scale ET product derived using Reliability Ensemble Averaging (denoted as REA) to compare our merged product comprehensively…"

2.9 Line 485

L485. What about the correlated errors?

AC:

The results of correlated errors are discussed in the Discussion section, where we believe it is more appropriate to address them.

2.10 Line 621-629

L621-629. This is more informative than the global statistics. I think it will be useful for the readers to adopt the new product if the authors can highlight when/where and over what spatiotemporal scales that CAMELE outperforms other products substantially. From a practical standpoint, establishing the *unique* strength of the proposed product is more important than showing its consistency with its peers.

AC:

Thank you once again for your valuable suggestion. We have incorporated additional analysis addressing your concern, and the relevant content can be found in the response to your **Major Comment 1.2**, eliminating the need for duplication here. Furthermore, we have provided an in-depth analysis highlighting the consistent superior performance of CAMELE across various PFTs, emphasizing the reliability of the fusion approach.

Revised Contents (Line 701 to 721):

"…From the results, it is evident that CAMELE performs well across various vegetation types. To delve deeper into the reasons behind this performance, we conduct site-scale analyses at two resolutions, evaluating errors and computed weights for different PFTs sites. These are visualized in radar chart format in Figure 8.

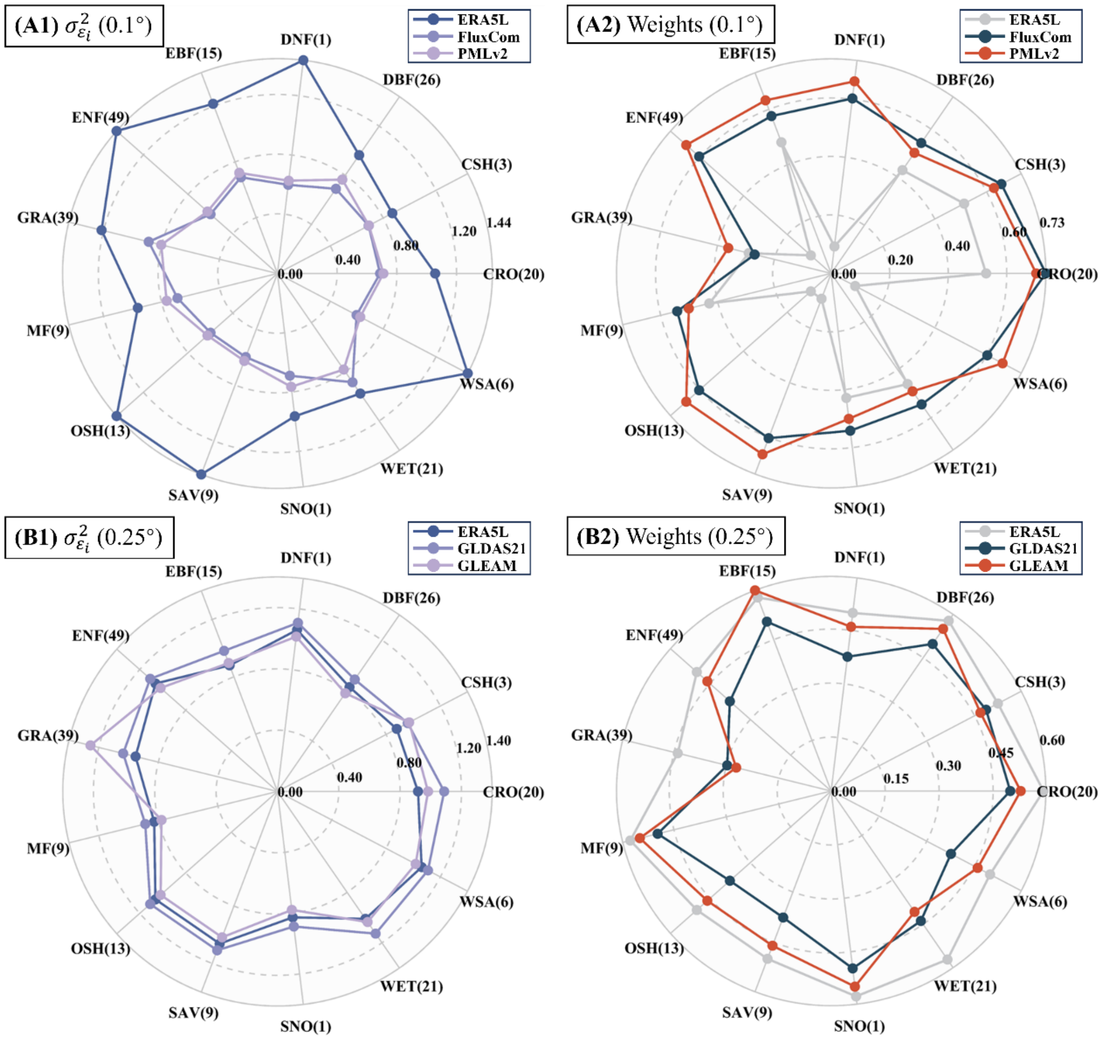


Figure 9 Mean collocation-based errors and weights of different products at various PFTs sites at (A) 0.1° and (B) 0.25° resolutions. The parentheses next to each PFTs name denote the corresponding number of sites.

The results from Figure 9 demonstrate that the error-weighting calculation method based on collocation effectively considers the error situation of inputs, thereby providing reasonable weight assignments. At 0.1° resolution, ERA5L's error is significantly higher across all PFTs than FluxCom and PMLv2, resulting in relatively lower corresponding weights. FluxCom and PMLv2 exhibit closer performance, with higher weights at most PFT sites. At 0.25° resolution, ERA5L, GLDAS21, and GLEAM perform more evenly, with minimal differences, resulting in closer weights. The weights for different inputs vary noticeably with changes in PFTs, depending on the performance of other products within the same combination. Products with more significant errors correspondingly have lower weights, affirming the rationale behind the fusion method. However, it is essential to note that the presented results depict the mean values of errors and weights across all sites; there might be variations among sites with the same PFTs…”

2.11 Line 635

L635. How does the proposed product and its peers compare with the FluxNet in terms of long-term average and trend?

AC:

Thank you again for your suggestion. We have incorporated additional analysis comparing the proposed product and its peers with FluxNet in terms of long-term average and trend. The relevant details can be found in our responses to your Major Comments

1.2.3 New Contents about multi-year mean and extreme ET value (4.3 Results, Line 732 to 761):

1.2.4 New Contents about multi-year trend and seasonality (4.4 Results, Line 798 to 846):

2.12 Line 677

L677. I think only the statistically significant trends should be shown.

AC:

We sincerely appreciate your insightful comments, which are crucial for the accurate calculation of trends. We have re-plotted the trends for various products, including 0.1° (2001-2015) and 0.25° (2000-2017) datasets, along with CAMELE, highlighting regions with significant changes. The trends are estimated using Theil–Sen's slope method, and their significance is tested with the Mann–Kendall method. The dotted areas indicate trends passing the significance test at a 5% level.

Additionally, we have rectified the coding error in the original 0.1° trend plot, where latitude variation was incorrectly portrayed as the dependent variable. Please find the corrected trend for CAMELE, demonstrating consistency among input ensemble members. Furthermore, modifications have been made to the figure captions for clarity.

Revised Figures (Line 850 to 862):

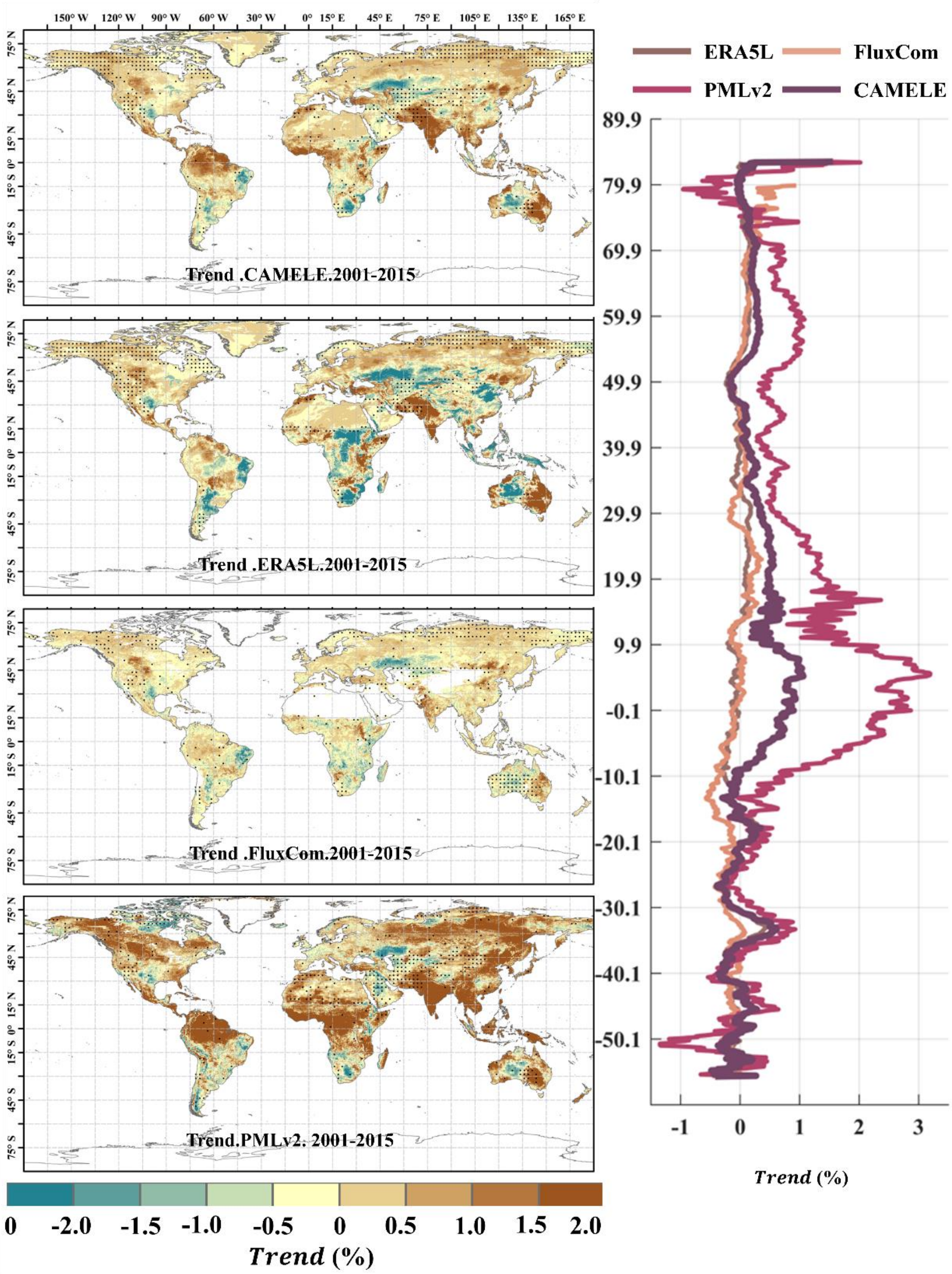


Figure 15 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area indicates that the trend has passed the significance test at 5 % level.

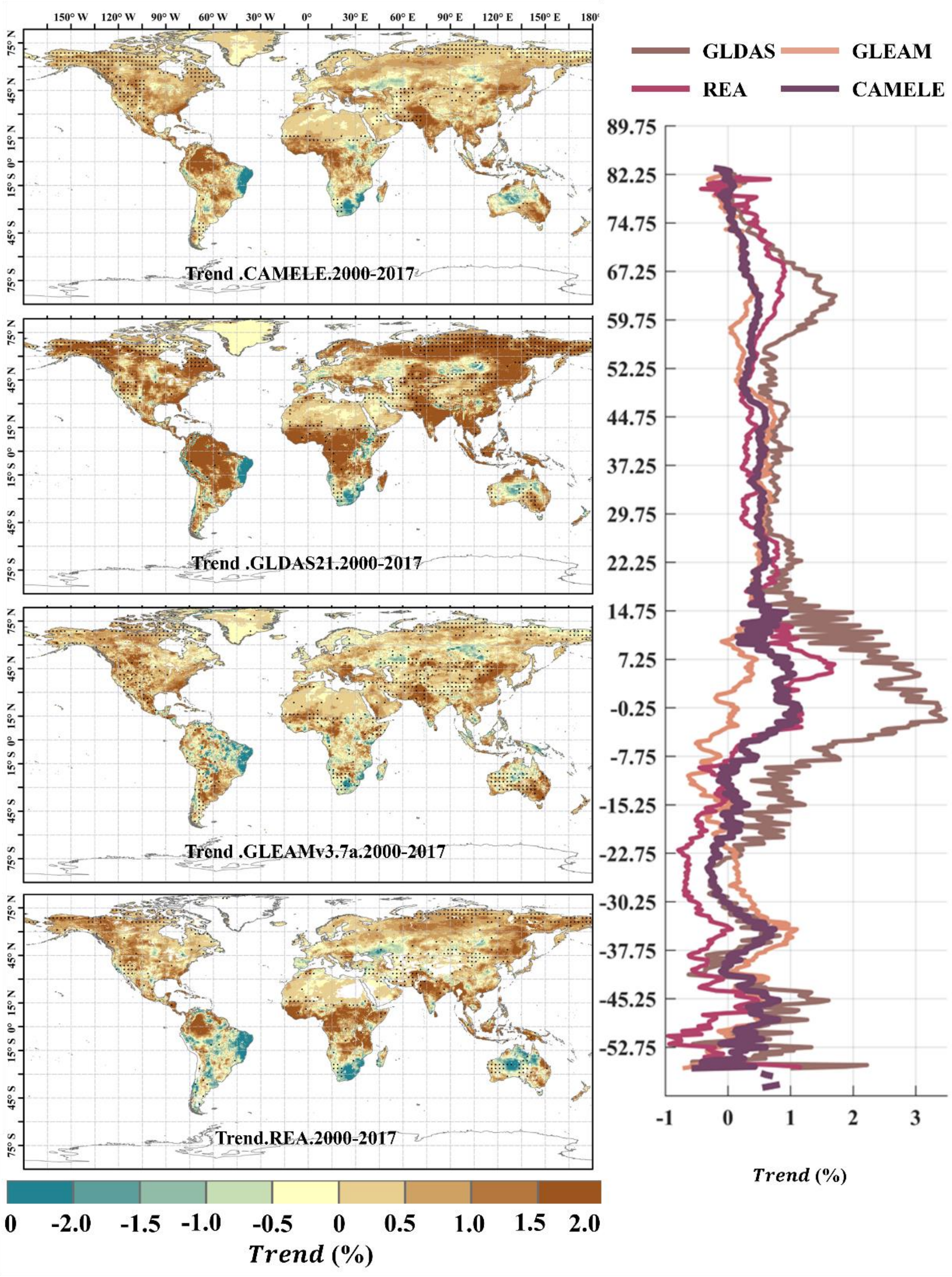


Figure 16 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area indicates that the trend has passed the significance test at 5 % level.

Previous Figures

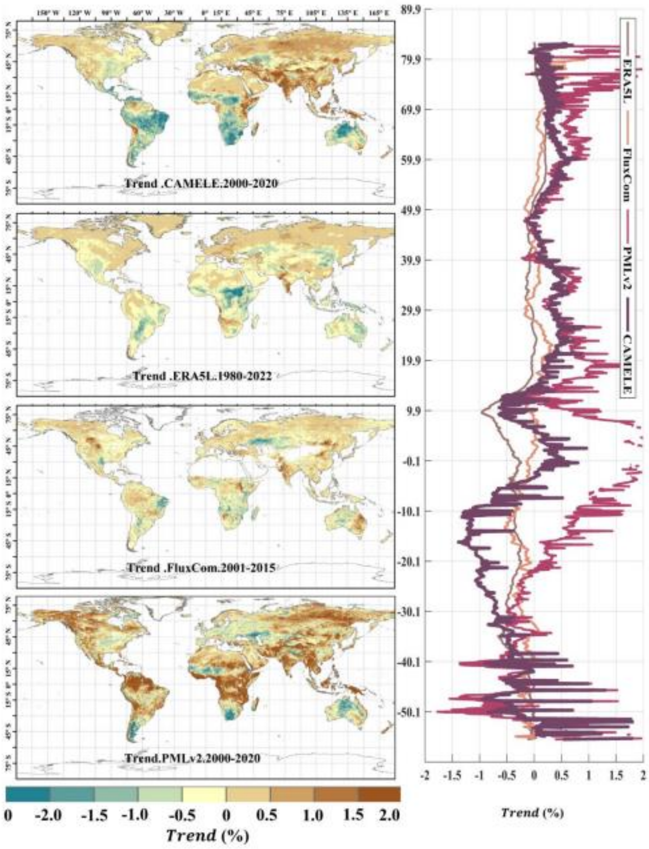


Figure 9 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

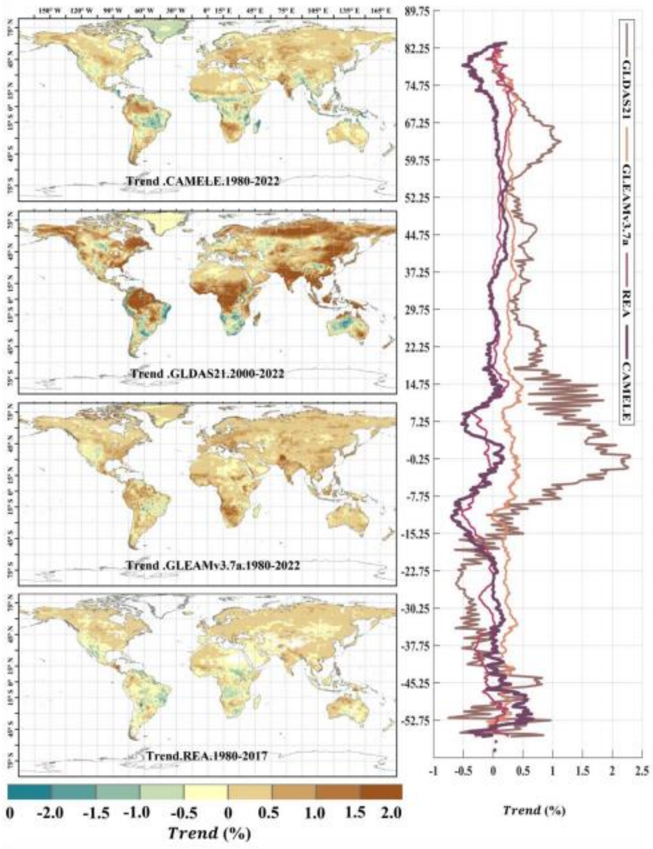


Figure 10 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

2.13 Line 745

L745. This should go to the result section. It is not clear to me how different treatment of the cross-correlation pairs will impact the final product.

AC:

Thank you very much for your suggestion. We still believe that placing it in the discussion section is more appropriate for two reasons:

(1). The results section primarily focuses on discussing the evaluation results of CAMELE and comparing its performance with other products. As mentioned earlier, we have already included a detailed analysis of trend, seasonality, mean, and extreme, making the section quite extensive. Adding the ECC part might seem abrupt to readers solely interested in ET.

(2) We have specified which two groups have non-zero ECC, serving as a test of the validity of our hypothesis.

Setting non-zero ECC in collocation calculations requires careful consideration. In our case, we are validating the correctness of our non-zero ECC pairs, not attempting to compare the effects of all possible pairs, which would be impractical. Furthermore, we have discussed the scenario without setting non-zero ECC in Section 5.3 (i.e., the results of the traditional TC method), providing a comprehensive discussion on non-zero ECC in this context.

➢ AC to Referee #3: General Comment

The manuscript "CAMELE: Collocation-Analyzed Multi-source Ensembled Land Evapotranspiration Data" presents an ensemble product compiled from collocation analysis/weighting of five global evapotranspiration (ET) products (ERA5-Land/GLEAM/GLDAS/FluxCom/PML). The authors illustrate that by using non-zero ECC collocation weighting, multiple independently-sourced ET products can be merged resulting in enhanced accuracy. Generally, the paper is properly written and structured. It is well suited for this journal.

AC:

We greatly appreciate the professional and constructive feedback provided by the reviewer. We will respond to each comment individually, and in the following responses, the line numbers corresponding to the added or revised content will be based on the updated version without highlights. You can open the PDF file's table of contents view to navigate to the relevant sections directly.

The responses will be in the following format:

- Reviewer's comments are shown in black.
- Our responses are shown in blue.
- The modifications to the manuscript are shown in orange.
- Previous contents in the old version (for comparison if needed) are shown in grey.

1 AC to Referee #3: Some Remarks

1.1 Q1

The authors should consider [consistently] defining all abbreviations before use (more below). While many of the abbreviations may be obvious to the authors (and for many in the sub-field), they may be misinterpreted by other readers. Some terms, such as EC may be misinterpreted by most interested in the (experimental) observation and modeling of the surface energy budget.

AC:

Thank you for your valuable suggestion. We have revised the manuscript to consistently provide full abbreviations upon their first use.

1.2 Q2

5 ET products (ERA5L/GLEAMv3/GLDAS/FluxCom/PMLv2) are applied in this study. What was the criteria used to select these 5? Have the authors considered including other ET products, such as the MERRA, MOD16, WaPOR, SSEBop, …, in their analyses. If not, why?

All the ET products described here (and consequently the ensemble CAMELE product – ~1°, 0.25°) are rather coarse. Most of the local characteristics that influence the local flux interactions are therefore averaged out. For purposes that involve local/field-scale applications, and in terms of accuracy (i.e., evaluation scale mismatch with FluxNet local footprints), a discussion of the scale limitations is necessary.

AC:

Thank you for the insightful feedback. Our selection criteria aimed to ensure: (1) consistency in original spatiotemporal resolution among the products; (2) having three or more products within the same resolution or period; (3) products with extensive global observational sequences. Among the products mentioned, MERRA has a resolution of 0.625x0.5, requiring downscaling for pairing; MOD16, with its 500m resolution, offers higher accuracy but would entail down sampling other products, leading to potential errors; WaPOR and SSEBop provide global monthly data, with SSEBop's daily data limited to the continental United States, mismatching in temporal resolution with other products. Hence, considering these aspects, we opted for the ensemble mentioned in the paper. While it lacks the precision of other products, it still aids in understanding ET variations and serves as a beneficial dataset.

In Section 2, "Datasets," we have included explanations regarding the selection of the products.

Modified Contents (Line 146 to 156):

"…When selecting these products, our aims are to ensure: (1) consistency in original spatiotemporal resolution among the products; (2) having three or more products within the same resolution or period; (3) products with extensive global observational sequences. While we acknowledge the existence of other higher-precision products, their integration would require either downscaling or upscaling other products, potentially introducing uncertainties. Therefore, we chose the combination outlined in the manuscript. Despite its relatively lower resolution compared to some products, it still contributes to our understanding of ET variations, facilitating advantageous exploration…"

Certainly, we acknowledge the coarseness of the obtained data compared to regionally high-resolution products, presenting apparent limitations. In the newly added Section "5.4. Potential Applications and Future Enhancements", we address this drawback and introduce prospects, aiming to leverage the strengths of regional high-precision products to further enhance CAMELE.

New Contents (Line 1052 to 1058):

"… (2) Enhanced Integration and Error Reduction: We continually refine estimates by incorporating additional data sources and implementing extended collocation method to minimize errors; (3) Integration of High-Resolution Regional ET Data: Recognizing the significance of regional-scale insights, we will focus on improving the accuracy of CAMELE by integrating higher-resolution regional ET data. This integration will enable more precise regional estimation…"

1.3 Q3

One interesting outcome from this study is that the CAMELE product appears to perform comparatively well over most of IGBP-based plant functional types (PFTs). While commendable, the authors only touch on this without really discussing why it performs better. What are the implications of selecting one product over the other over different PFTs, especially with respect to real applications.

AC:

Thank you for the insightful suggestion. We have expanded upon the analysis of why CAMELE performs better across various PFTs in the respective section of the manuscript. In essence, our findings highlight that error analysis in collocation and the methodology for weight computation effectively capture product inaccuracies in inputs, thus yielding reasonable weights.

New Contents (Line 701 to 721):

"… From the results, it is evident that CAMELE performs well across various vegetation types. To delve deeper into the reasons behind this performance, we conduct site-scale analyses at two resolutions, evaluating errors and computed weights for different PFTs sites. These are visualized in radar chart format in Figure 9.

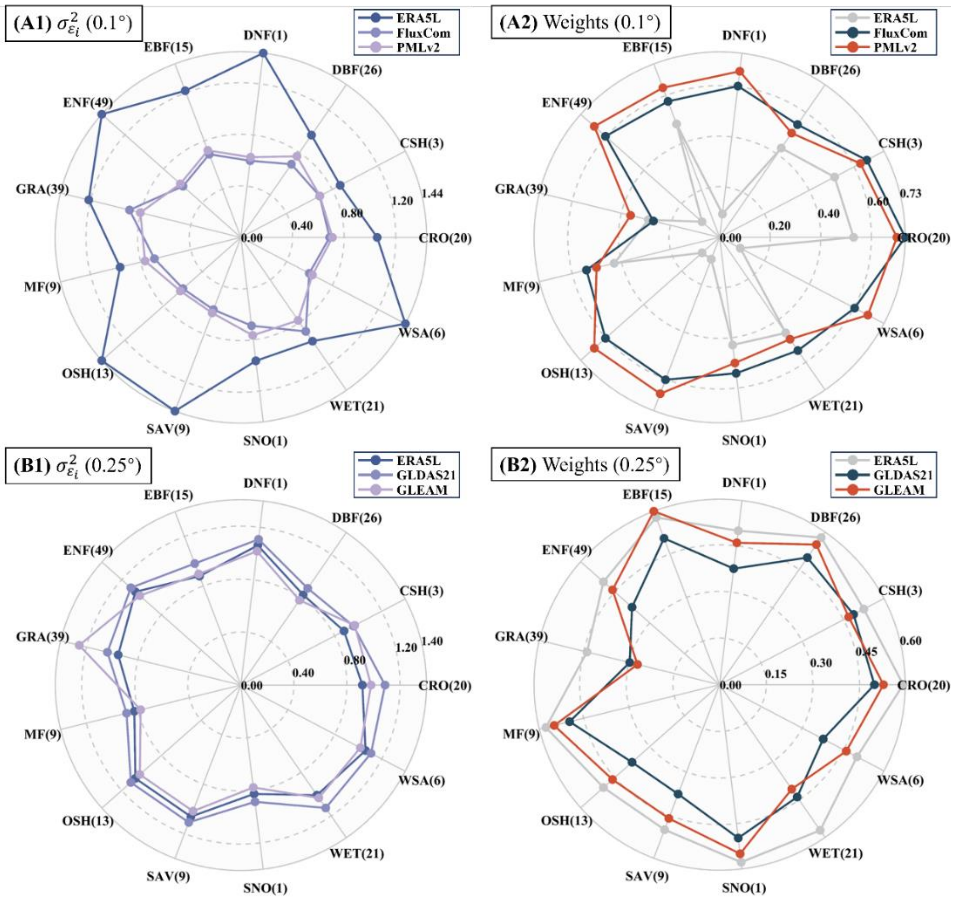


Figure 9 Mean collocation-based errors and weights of different products at various PFTs sites at (A) 0.1° and (B) 0.25° resolutions. The parentheses next to each PFTs name denote the corresponding number of sites.

The results from **Figure 9** demonstrate that the error-weighting calculation method based on collocation effectively considers the error situation of inputs, thereby providing reasonable weight assignments. At 0.1° resolution, ERA5L's error is significantly higher across all PFTs than FluxCom and PMLv2, resulting in relatively lower corresponding weights. FluxCom and PMLv2 exhibit closer performance, with higher weights at most PFT sites. At 0.25° resolution, ERA5L, GLDAS21, and GLEAM perform more evenly, with minimal differences, resulting in closer weights. The weights for different inputs vary noticeably with changes in PFTs, depending on the performance of other products within the same combination. Products with more significant errors correspondingly have lower weights, affirming the rationale behind the fusion method. However, it is essential to note that the presented results depict the mean values of errors and weights across all sites; there might be variations among sites with the same PFTs…”

1.4 Q4

A weighted average of the ensemble members (ET products) is described and discussed. Nowhere, though, do the authors detail the weights quantitatively, even for a few example scenarios. In Park et al. (2023), for instance, the weighting factors calculated in a triple-collocation study (with no consideration of non-zero ECC) were analyzed. It would be interesting to provide the readers with a sense of such weights between the 5 products used within CAMELE, especially given that non-zero ECC is considered here. Are they close to equal weighting? Are different weights assigned depending on the season, e.g. as is/was done for the DOLCE product? How much do the weights vary with the plant functional type?

AC:

We appreciate the valuable feedback provided by the reviewer. In response to the suggestion, we have included additional discussions in Section 4.1, referring to the presentation style of Park et al. (2023). This section now addresses the distribution of dominant products in each grid under three fusion scenarios, where the dominant product refers to the product with a higher weight in each grid. Additionally, due to the numerous figures illustrating the weight distribution for each product, we have placed them in the appendix for clarity.

Upon examining the weight calculation results and the distribution of dominant products, it is evident that equal weighting is not employed. This is further emphasized by the comparisons with simple averaged results presented in Sections 4.2 and 4.3. The weights for each grid in every scenario are determined through collocation analysis of inputs over all periods. Hence, these weights remain constant along seasons, representing the optimal weight scheme based on the minimum MSE for the respective inputs. Moreover, it is worth noting that the weights vary with PFTs, as discussed in Section 4.2, addressing your third question (Q3).

By your recommendation, we have added the following content in Section 4.1, aiming to address your inquiries:

New Contents (Line 582 to 602):

"… Next, in Figure 5, we present the dominant product for each grid cell in the three scenarios, where dominance refers to the product with the highest assigned weight. The results in Figure 5 indicate that at 0.1° resolution, the weights for FluxCom and PMLv2 are significantly higher than ERA5L, aligning with the error calculations presented in Figure 2. This underscores the effectiveness of error and weight analysis based on collocation in reflecting product performance, thereby allowing for a rational adaptation of weights. At 0.25° resolution, the dominant regions for ERA5L, GLDAS-2, and GLEAM products are relatively balanced. In the fusion scenario from 1980 to 1999, GLDAS20 predominantly covers the Northern Hemisphere, while GLEAM dominates the Southern Hemisphere, with ERA5L prevalent in the Amazon region. However, in the fusion scenario from 2000 to 2022, GLEAM's dominant region significantly expanded, primarily covering the central United States and southeastern China. The Amazon region continues to be dominated by ERA5L. The variation in dominant products highlights that the calculation of product weights evolves with changes in the fusion scenario. The error and weight computation methods based on collocation can only provide the minimum MSE solution for a given combination of inputs. It is important to note that changes in inputs will impact the results.

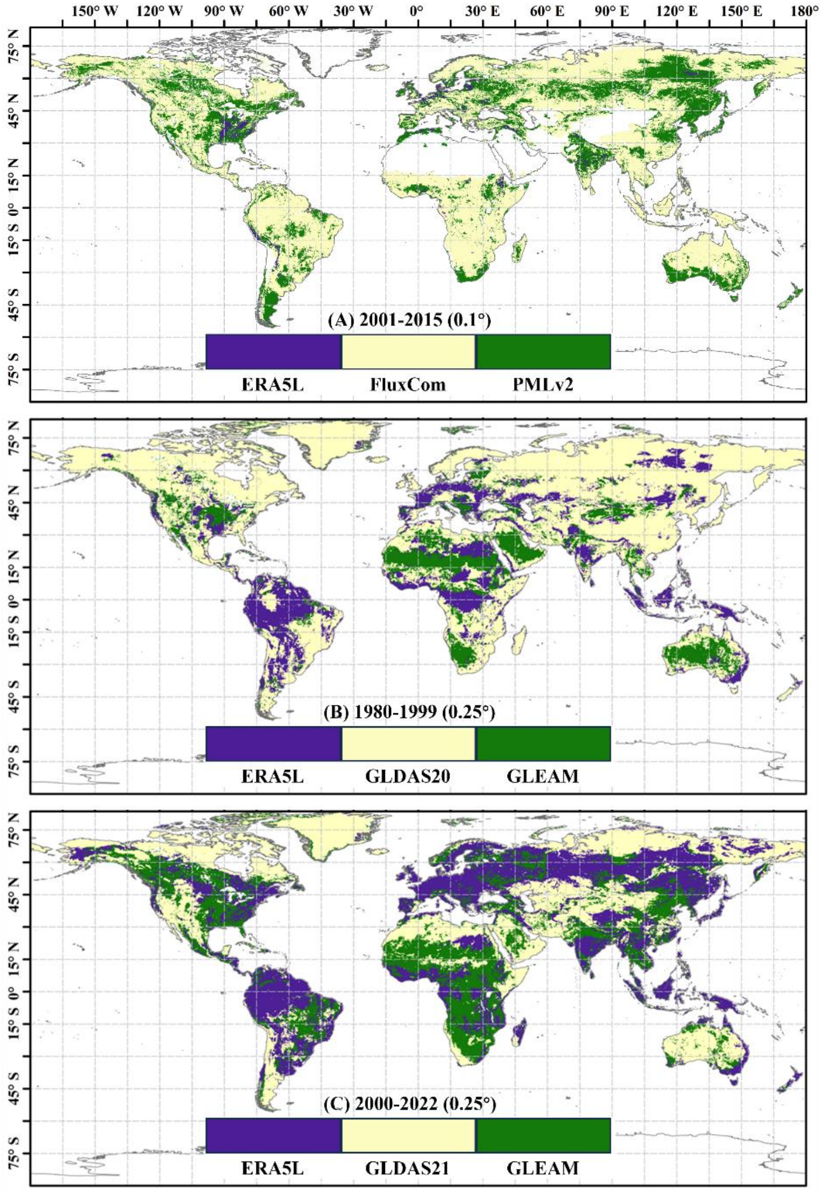


Figure 5 Map of the prevailing product at individual pixels based on scenario-specific weights.

2 AC to Referee #3: Specific Comments

2.1 Line 34

L34: Do these metrics imply the CAMELE ensemble ET is fit to be applied as a benchmark reference of choice? maybe state that it may be a suitable 'reference' candidate for ET product evaluations.

AC:

Thank you for your suggestion. We have incorporated your feedback by adding a statement in the Abstract:

New Contents (Line 37 to 39):

"… In summary, we propose a reliable set of ET data that can aid in understanding the variations in the water cycle and has the potential to serve as a benchmark for various applications.…"

2.2 Line 42

L42: add to read "soil moisture and air temperature/humidity.

AC:

Updated.

2.3 Line 45

L45: "… evapotranspiration, resulting in many datasets" - Maybe some citation is necessary here?

AC:

Thank you for your suggestion. We have incorporated a recent comprehensive review article on ET published in Nature, which addresses the citation gap you pointed out.

Revised Contents (Line 50 to 51):

"… In recent decades, numerous studies have focused on estimating global land evapotranspiration, resulting in many datasets (Yang et al., 2023) …"

Reference:

Yang, Y., Roderick, M. L., Guo, H., Miralles, D. G., Zhang, L., Fatichi, S., Luo, X., Zhang, Y., McVicar, T. R., Tu, Z., Keenan, T. F., Fisher, J. B., Gan, R., Zhang, X., Piao, S., Zhang, B., and Yang, D.: Evapotranspiration on a greening Earth, Nat Rev Earth Environ, https://doi.org/10.1038/s43017-023-00464-3, 2023.

2.4 Line 66

L66: TC and EIVD used before being defined - Full names given further below (lines [71] and [79]). Consider describing the abbreviations here instead.

AC:

Thank you for your suggestion. Updated.

2.5 Line 96

L96: “(i.e., IVS, IVD, TC, EIVD, and EC)” – note that some of the abbreviations here have not been described earlier (e.g. EC)

AC:

Thank you for your suggestion. All related abbreviations have been corrected.

2.6 Line 109

L109: “… error covariance (ECC)” - defined in L78 as “error cross-correlation”. Consistency.

AC:

Thank you for your suggestion. Updated.

2.7 Line 137, 139

L137, 139: “referred to as ERA5L” – you call it ERA5L instead of the common ERA5-Land. Ok.

“… ERA5-Land …” – Consistency. Continue using ‘ERA5L’ since that is how it is abbreviated in this study

AC:

Thank you for your suggestion. Updated.

2.8 Line 174-175

L174-175: “… potential error homogeneity issues between GLDAS-2.2 and ERA5L” - Have these potential ‘homogeneity errors‘ due to use of equivalent meteorological forcings been documented anywhere? There should still be differences between the two ET estimates/products since: 1) GRACE data is assimilated (L171-172), and 2) different LSMs are used (i.e. lines [159-160] and [141-143] for GLDAS and ERA5-Land, respectively)

Looking at Figure1 of Jiménez et al. (2011) where 3 GLDAS models (NOA, Mosaic, CLM) are inter-compared -among others) shows that relatively large variations can be observed between the NOA, MOS, CLM flux estimates; these can generally be attributed to the differences in the models (parameterization, structure, physics, …). As such, the non-homogeneous error condition (as required in TC) will generally still be met between different LSMs - even with equivalent forcings.

AC:

Thank you for pointing out the issue at this section. The correlation between GLDAS-2.2 and ERA5L has been documented in Li et al., 2023. However, it is important to note that their focus was on the estimation of transpiration. Considering the similarities in the calculation of ET and T of GLDAS and ERA5L, this report partially indicates a correlation. Additionally, regarding the correlation among different models within GLDAS-2, we have added relevant explanations in this section.

Revised Contents (Line 200 to 214):

"… This study aimed to cover the research period from 1980 to 2022. Non-zero ECC between the transpiration estimates of GLDAS-2.2 and ERA5L has been reported in a recent study (Li et al., 2023a). Considering the similarities in the calculation of ET and transpiration of GLDAS and ERA5L, this report partially indicates a correlation. Therefore, GLDAS-2.0 and GLDAS-2.1 were selected as inputs instead. The "Evap_tavg" parameter representing evapotranspiration is derived from the original products and aggregated to a daily scale. For more detailed information on the GLDAS-2 models, please refer to NASA's Hydrology Data and Information Services Center at http://disc.sci.gsfc.nasa.gov/hydrology.

Despite the same forcing between GLDAS-2.1 and GLDAS-2.2, significant differences exist between the model results of different GLDAS versions (Qi et al., 2020, 2018; Jiménez et al., 2011). The non-zero ECC will generally still be met between different versions. Thus, we still need to analyze the non-zero ECC situations between ERA5L and GLDAS-2.0 and 2.1, which will be assessed in the discussion sections…"

Reference:

Li, C., Liu, Z., Tu, Z., Shen, J., He, Y., and Yang, H.: Assessment of global gridded transpiration products using the extended instrumental variable technique (EIVD), Journal of Hydrology, 623, 129880, https://doi.org/10.1016/j.jhydrol.2023.129880, 2023a

Jiménez, C., Prigent, C., Mueller, B., Seneviratne, S. I., McCabe, M. F., Wood, E. F., Rossow, W. B., Balsamo, G., Betts, A. K., Dirmeyer, P. A., Fisher, J. B., Jung, M., Kanamitsu, M., Reichle, R. H., Reichstein, M., Rodell, M., Sheffield, J., Tu, K., and Wang, K.: Global intercomparison of 12 land surface heat flux estimates, J. Geophys. Res., 116, D02102, https://doi.org/10.1029/2010JD014545, 2011.

Qi, W., Liu, J., and Chen, D.: Evaluations and Improvements of GLDAS2.0 and GLDAS2.1 Forcing Data's Applicability for Basin Scale Hydrological Simulations in the Tibetan Plateau, JGR Atmospheres, 123, https://doi.org/10.1029/2018JD029116, 2018.

Qi, W., Liu, J., Yang, H., Zhu, X., Tian, Y., Jiang, X., Huang, X., and Feng, L.: Large Uncertainties in Runoff Estimations of GLDAS Versions 2.0 and 2.1 in China, Earth and Space Science, 7, e2019EA000829, https://doi.org/10.1029/2019EA000829, 2020.

2.9 Line 181

L181: Note that, while not yet documented, they now have v3.8a available

AC:

Thank you for your suggestion. At the time of submission, version 3.8 was not publicly available. It is now accessible, and we have removed the term "latest" accordingly.

2.10 Line 185

L185: "…from 1980 to 2022" - Note that v3.7b (based on satellite data) only runs from 2003

AC:

Thank you for your suggestion. We have removed the phrase "from 1980 to 2022" as it is clarified later that the scope applies to both 3.7a and 3.7b.

Unchanged content (Line 221 to 224):

"…Two datasets that differ only in forcing and temporal coverage are provided: GLEAMv3.7a-43-year period (1980 to 2022) based on satellite and reanalysis (ECMWF) data; GLEAMv3.7b-20-year period (2003 to 2022) based on only satellite data…"

2.11 Line 194, 196

L194, 196: "into actual transpiration or bare soil evaporation" – maybe replace 'or' with 'and'? for total actual ET. "by (Martens et al., 2017)" >> "by Martens et al. (2017)"

AC:

Thank you for your suggestion. Updated.

2.12 Line 198

L198: Add this abbreviation in L194 above or define Actual Evapotranspiration (AET) here

AC:

Thank you for your suggestion. Updated.

2.13 Line 210

L210: "…, white sky albedo, …" - Do they really only use the white sky albedo in their computations of available energy ? Normally the broadband albedo is applied, which is a combination of white- (diffuse) and black-sky (direct) albedos (see MODIS albedo data for reference - https://lpdaac.usgs.gov/documents/97/MCD43_ATBD.pdf, i.e. pg.11, EQ 32).

In Figure1 of Zhang et al. (2019), they indeed indicate "White Sky Shortwave Albedo", but the same is not mentioned anywhere else in their article. Since it might have been a misplaced error in that figure, you should drop 'white sky' here unless you can confirm from them that only WS albedos are used in PMLv2 calculations - which would then mean an additional source of uncertainty in PMLv2 ET products.

AC:

Thank you very much for pointing out the error. We have verified with Prof. Yongqiang Zhang, the author of PMLv2, and confirmed that they indeed use broadband albedo in their calculations. We have accordingly revised the manuscript to reflect this clarification.

Revised Contents (Line 244):

"… The daily inputs for this model include leaf area index (LAI), broadband albedo…"

2.14 Line 213

L213: "($Psurf$, Pa, U, q), and" - These [meteo] variables have not been defined elsewhere.

AC:

Thank you very much for pointing out this issue. The relevant descriptions have been added:

Revised Contents (Line 244 to 250):

"… The daily inputs for this model include leaf area index (LAI), broadband albedo, and emissivity obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS), as well as temperature variables (daily maximum temperature-T_{max} , daily minimum temperature-T_{min}, daily mean temperature-T_{avg}), instantaneous variables (surface pressure-P_{surf}, atmosphere pressure-P_a, wind speed at 10-meter height-U, specific humidity-q), and accumulated variables (precipitation-P_{rcp}, inward longwave solar radiation-R_{ln}, inward shortwave solar radiation-R_s) from GLDAS-2.0…"

2.15 Line 241

replace "evaporation" in L241 with evapotranspiration.

AC:

Thank you for your suggestion. Updated.

2.16 Line 251

L251: "ET data were corrected." - Maybe clarify how? residual method? bowen? …?

AC:

Thank you very much for pointing out our issue. We utilized the energy balance-based correction method proposed by Twine et al. (2000), specifically employing the residual method. To provide clarity, we have added a brief explanation:

Revised Contents (Line 287 to 289)

"… Therefore, following the method proposed by Twine et al. (2000), the measured ET data were corrected using the residual method based on energy balance…"

2.17 Line 252-261

L252-261: That makes 12 PFTs and 206 sites. What of the other 212-206 sites (13-12=1 PFT)?

AC:

Thank you for pointing out the mistake. We missed the six WSA sites. The relevant descriptions have been revised:

Revised Contents (Line 291 to 303)

"… The International-Geosphere–Biosphere Program (IGBP) land cover classification system (Loveland et al., 1999) was employed to distinguish the 13 Plant Functional Types (PFTs) across sites. The IGBP classification was determined based on metadata from the FluxNet official website, including evergreen needle leaf forests (ENF, 49 sites), evergreen broadleaf forests (EBF, 15 sites), deciduous broadleaf forests (DBF, 26 sites), croplands (CRO, 20 sites), grasslands (GRA, 39 sites), savannas (SAV, 9 sites), mixed forests (MF, 9 sites), closed shrublands (CSH, 3 sites), deciduous needle leaf forests (DNF, 1 site), open shrublands (OSH, 13 sites), snow and ice (SNO, 1 site), woody savannas (WSA, 6 sites) and permanent wetland (WET, 21 sites). Changes in the IGBP classification during the study period are possible, but such information is not publicly available. Interested parties can obtain relevant information by directly contacting the site coordinators…"

2.18 Line 270

L270: Again, EC here has yet to be defined. It is defined further below [L342]. Note that EC in ET circles may be interpreted to mean Eddy Covariance, so consider defining EC further up to avoid confusion.

AC:

Thank you for your suggestion. Updated.

2.19 Line 315

L315, Equation 7: NSR is Noise to signal ratio? why write it here if it will not be used elsewhere?

AC:

Thank you for your suggestion. Indeed, NSR is not used later, so we have removed the subsequent derivation step.

Revised Contents (Line 354)

Following similar ideas, Mccoll et al. (2014) extended the framework to estimate the data-truth correlation, known as the ETC:

$$R_i^2 = \frac{\beta_i^2 \sigma_\Theta^2}{\beta_i^2 \sigma_\Theta^2 + \sigma_{\varepsilon_i}^2} = \frac{SNR_i}{1 + SNR_i} \tag{7}$$

$$R_i^2 = 1 - fMSE_i$$

2.20 Line 316

L316: “In comparison to the conventional coefficient of determination Rij” - Is it common to write the standard/conventional coefficient of determination as R instead of R^2?. R is generally reserved for correlation.

AC:

This is a generally used expression in triple collocation analysis. R_i^2 is the data-truth correlation, which incorporates the dependency on the chosen reference.

2.21 Line 395

L395: “… CCI” is not defined

AC:

Thank you for the notice. Updated.

Revised Contents (Line 105 to 108)

“…This was initially applied by Yilmaz et al.(2012) in the fusion of multi-source soil moisture products and later improved by Gruber et al. (2017) and further applied in the production of the European Space Agency Climate Change Initiative (ESA CCI) global soil moisture product (Gruber et al., 2019)…”

2.22 Line 410

L410: “… superior …” – your ensemble ET product performs somewhat similarly to the others, so the authors should be a bit modest here. Use another word; otherwise detail the aspects that make it superior.

AC:

Thank you for your suggestion. We have changed “superior” to “promising”, indicating our anticipation for better fusion results.

Revised Contents (Line 439 to 441)

“…The merging technique employed in this study provides a more explicit characterization of product errors and facilitates the derivation of more reliable weight coefficients, thereby achieving promising fusion outcomes…”

2.23 Line 418

L418: “… PMLv2 and FluxCom have an original resolution of 0.083° and an 8-day average - note that for FluxCom, energy balance fluxes are also available at the daily scale, i.e. denoted ‘RS_METEO’.

AC:

Thank you for your suggestion. We have specified here that FluxCom-RS is used for the 8-day average data. FluxCom-RS_METEO provides different inputs, including ensemble daily scale data, all at 0.5° (720_360), which does not match the spatial resolution of other inputs. We believe that interpolating directly from 0.5° to 0.1° is a large span and may introduce errors. Therefore, we used FluxCom-RS here.

Revised Contents (Line 446 to 447)

"…In this study, we employ five commonly used global land surface ET products as described in the datasets section. PMLv2 and FluxCom-RS have an original resolution of 0.083° and an 8-day average…"

2.24 Line 420

L420: "… and the values for each data period of 8 days are kept consistent. For example, the values for March 5 to March 12, 2000, are the same"

It is not clear what '8 days' means here. Going by L420 it is appears like one value is replicated for the 8 days. [Actual] ET is influenced by radiation, atmospheric vapour demands as well as surface water availability. These do not usually remain constant throughout an 8-day period. So using an 8-day average to represent the temporal dynamics should ideally introduce further uncertainties.

Also, why do the authors only use the FluxCom 8-day dataset (which employs only remote sensing data)? there is also the 'RS_METEO', which is available at daily timesteps.

AC:

Thank you very much for pointing out the issue. Firstly, the daily scale data of FluxCom's RS_METEO is at 0.5° (720_360), which significantly differs in spatial resolution from other products. Interpolating from 0.5° to 0.1° would introduce considerable errors, so we opted for the higher resolution RS data. Additionally, we acknowledge your concern about the variation in ET over an 8-day period. Assigning the same value for each 8-day period in FluxCom and PMLv2 indeed introduces errors. We have added clarification regarding the errors:

Revised Contents (Line 448 to 455)

"…In this research, they are interpolated to 0.1° resolution, and the values for each data period of 8 days are kept consistent. For example, the values for March 5 to March 12, 2000, are the same. ET values often exhibit variability over an 8-day period, making the use of an 8-day average to represent temporal dynamics potentially introducing further uncertainties. This operation is performed to ensure adequate data for the collocation analysis (Kim et al., 2021a). We openly acknowledge the possible sources of error and express our commitment to addressing and improving them in future work…"

The goal is to achieve results with higher temporal resolution. From the site assessment results, CAMELE's performance remains promising. We have included an analysis of linear trends and seasonality, identifying a potential overestimation of seasonality at 0.1°. We honestly acknowledge the possible sources of error and express our commitment to improving them in future work.

New Contents (Line 799 to 847)

"…4.4. Assessment and comparison of linear trend and seasonality

In this section, we first validate and compare the performance of CAMELE with other products in estimating multi-year trends and seasonality at the site scale. Due to the inconsistent time lengths of FluxNet sites, trends at many sites are not significant. Therefore, we deliberately selected 13 sites with continuous evapotranspiration (ET) observations for the same 11-year period (2004 to 2014) and with significant trends. The annual ET values for each year were calculated as the mean of the 13 sites for that year, allowing the computation of linear trends and seasonality. We employed singular spectrum analysis (SSA), which assumes an additive decomposition A = LT + ST + R. In this decomposition, LT represents the long-term trend in the data, ST is the seasonal or oscillatory trend (or trends), and R is the remainder.

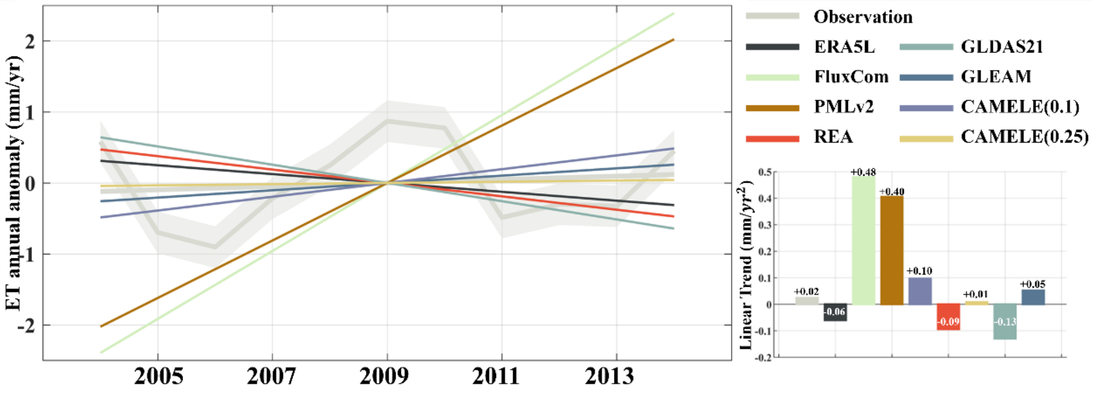


Figure 12 Comparison of linear trend from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The trends have been subjected to SSA decomposition, removing seasonality. The gray enveloping line represents the mean plus the standard deviation of the 13 sites.

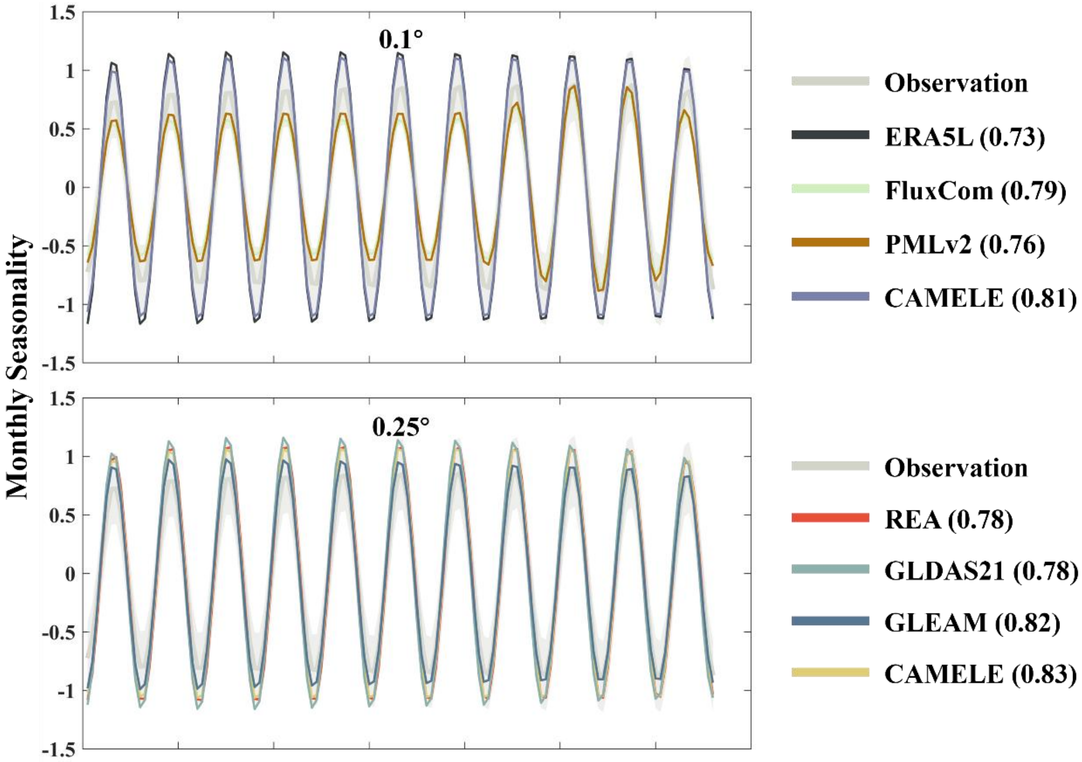


Figure 13 Comparison of seasonal variations from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The seasonality has been obtained through SSA decomposition, with the gray area representing the observed values. The parentheses in each product name indicate the KGE coefficient comparing with the observed values.

In Figure 12 and Figure 13, based on observations from FluxNet sites, we analyzed the performance of CAMELE and other products in estimating the linear trend and seasonality of ET over multiple years. It is important to note that we only present the analysis results for 13 sites with continuous 11-year observations, and the performance of different ET products in trend estimation at individual sites still varies, not fully reflecting the overall performance on all grids in terms of trend and seasonality. Nevertheless, such a comparison can still provide valuable insights.

Examining the results of the linear trend, both PMLv2 and FluxCom exhibit a significant upward trend, well above the observations. On the contrary, ERA5L, GLDAS, and REA show a noticeable downward trend, while CAMELE demonstrates a gradual upward trend closer to the observations. Additionally, GLEAM slightly outperforming CAMELE at a resolution of 0.25°. Overall, CAMELE shows good agreement with site observations in capturing the multi-year linear trend of ET.

Continuing with the analysis of seasonality, the KGE index comparing each product's results with observed values is provided in parentheses next to the product name. Generally, all products exhibit a good representation of ET's seasonal variations.

CAMELE's 0.1° seasonal results closely match FluxCom (with the two lines almost overlapping). However, the fluctuations it reflects are higher than the observed values. This is likely due to keeping the 8-day average results of FluxCom consistent with PMLv2 every 8 days, and the variability in ET primarily originates from ERA5L results. This aspect may need improvement in subsequent research. At 0.25°, CAMELE's seasonal representation is closer to the observed results. The differences in CAMELE's performance at the two resolutions are mainly attributed to input variations, which we discuss in the following section as potential areas for improvement.

The results indicate that CAMELE effectively captures the multi-year changes in ET, but at 0.1°, it tends to overestimate seasonal fluctuations…"

2.25 Line 430

L430: "ERA5L/GLEAMv3/PMLv2/FluxCom/PMLv2" - PMLv2 appears twice

AC:

Thank you for pointing out the mistake. Updated.

Revised Contents (Line 461)

"…analyze the performance of five sets of ET products (ERA5L/ PMLv2/FluxCom/GLDAS2/GLEAMv3) at the global scale…"

2.26 Line 468, 469

L468, 469: "shortcomings in Nash-Sutcliffe" such as? Where>>where

AC:

Thank you for your reminder. Firstly, we have switched to the modified KGE (Kling et al., 2012) index based on the suggestions of other reviewers, and briefly explained the advantages of the modified KGE in a sentence. The comparison between the KGE and NSE indices can be found in the literature by Kling et al., and we have added explanations in the relevant sections.

Revised Contents (Line 501 to 505)

"…The modified KGE (Kling et al., 2012) offers insights into reproducing temporal dynamics and preserving the distribution of time series, which are increasingly used to calibrate and evaluate hydrological models (Knoben et al., 2019). For a better understanding of the KGE statistic and its advantages over the Nash-Sutcliffe Efficiency (NSE), please refer to Gupta et al. (2009). The equation is given by:

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\sigma_{sim}/\mu_{sim}}{\sigma_{obs}/\mu_{obs}} - 1\right)^2} \qquad (27)$$

2.27 Line 491

L491: "variation curves of average with latitude" - Variation curves of which state variable ? add *ET after 'average'. Are these metrics curves really fair since there is quite a bit of missing data, especially over North Africa (and other sub-Saharan regions below the Equator & Australia - e.g. for Fluxcom)

AC:

Thank you very much for your suggestion. The variation curves presented here depict the mean errors at each latitude (0.1° interval) and not for ET. We have clarified this in the caption accordingly.

Revised Contents (Line 522)

"…**Figure 2** Global distribution of absolute error variances ($\sigma_{\varepsilon_i}^2$) of ERA5L, FluxCom, and PMLv2 using EIVD at 0.1° from 2001 to 2015, depicted alongside corresponding variation curves of average $\sigma_{\varepsilon_i}^2$ with latitude…"

Regarding the issue of missing data, particularly over North Africa and other sub-Saharan regions below the Equator & Australia, as raised for FluxCom, we have addressed this concern by incorporating a description of the uncertainty associated with these gaps in our error analysis.

New Contents (Line 538 to 544)

"…It is important to note that due to missing data in specific regions at 0.1°, such as Northern Africa, the Sahara Desert region, Northwestern China, and Australia, the error results obtained may not accurately reflect the performance of FluxCom and PMLv2 in these areas. Considering the current results, we can cautiously conclude that FluxCom and PMLv2 demonstrate better performance. Future data supplementation in these regions would further enhance our ability to analyze the products' accuracy…"

2.28 Line 509, 510

L509,510: "(0.59±0.58 mm/d), GLDAS2.0 (0.37±0.44 mm/d), and GLEAMv3.7a (0.38±0.36 mm/d)…" - Are these reported global (mean±standard deviation) values more or less equivalent to the latitude-averaged values ? looking at the variation curves in Figure 2, I am not really sure, as the GLEAM and GLDAS products qualitatively appear to have higher variation than ERA5 especially from latitude -45_to_-35 and -15_to_10

AC:

Regarding your inquiry, we have examined the relevant results and confirmed that there are no calculation errors. The higher mean error of ERA5L is primarily observed in the East Asia and Australia regions, where there is a higher density of grid points.

2.29 Line 517-519

L517-519: "average distribution with latitude" – average distribution of variation with latitude

“ERA5L demonstrates a more even distribution, whereas GLDAS and GLEAM exhibit relatively higher uncertainties in tropical regions” – the authors could consider discussing in terms of the model theoretical basing/assumptions/inputs – i.e., surface characteristics/physics.

AC:

Thank you for your valuable suggestions. We appreciate the insight you provided. The reason we did not delve into the analysis of why GLDAS and GLEAM exhibit higher errors in tropical regions compared to ERA5L is that a thorough understanding would require further model experiments or sensitivity analyses. Therefore, we focused on describing the observed phenomena. In response to your suggestion, we have added a brief analysis of the possible reasons for the errors in the two model products, presenting them as potential factors:

New Contents (Line 560 to 566)

“…The ET calculations in both GLDAS and GLEAM involve complex surface parameterization processes. In tropical regions, the high non-heterogeneity in land covers poses a challenge, and the 0.25° resolution grid may not capture the intricacies of the underlying surface conditions. This mismatch could impact the parameterization process, leading to errors. Future work could involve in-depth model analyses or sensitivity experiments to identify sources of error in complex ET models, facilitating improvements…”

2.30 Line 529

L529: “during this timeframe” - Figure 2 reports on period 1980-1999 while Figure 3 reports for 2000-2022 [both ERA5, GLEAM, GLDAS2.0/2.1]. Is selection of ~20 year periods deliberate? What about 1980-2022?

AC:

Thank you very much for your suggestion. Unfortunately, it is not feasible to display the error analysis for the period 1980-2022 for ERA5L here. The collocation analysis presents the errors for each combination (in this case, the triplet) within the available period for that specific combination of products. With the switch from GLDAS2.0 to GLDAS2.1, the triplets have changed, and it is not valid to simply combine the errors of ERA5L for the two timeframes mathematically. As you mentioned earlier, there is a significant difference in results between different LSMs of GLDAS2, so changing the version of GLDAS2 naturally affects the error results for ERA5L. Nevertheless, we have obtained crucial error information that can be utilized in weight calculations.

2.31 Line 543

L543: “In this subsection, …” - you move directly into CAMELE without mentioning how the weighting between the ensemble members is done. At least, refer the reader to Section 3.4. Also note that Section 3.4 does not mention CAMELE even once.

AC:

Thank you very much for your valuable suggestion. As addressed in response to Q4, we have incorporated an analysis of the dominant product based on weighted drawing during different ensemble stages in this subsection (New Figure 4). Please refer to the answer to Q4 for more details. Additionally, Section 3.4 focuses on the mathematical methods of fusion and the combinations involved, without specifically addressing the performance analysis of CAMELE.

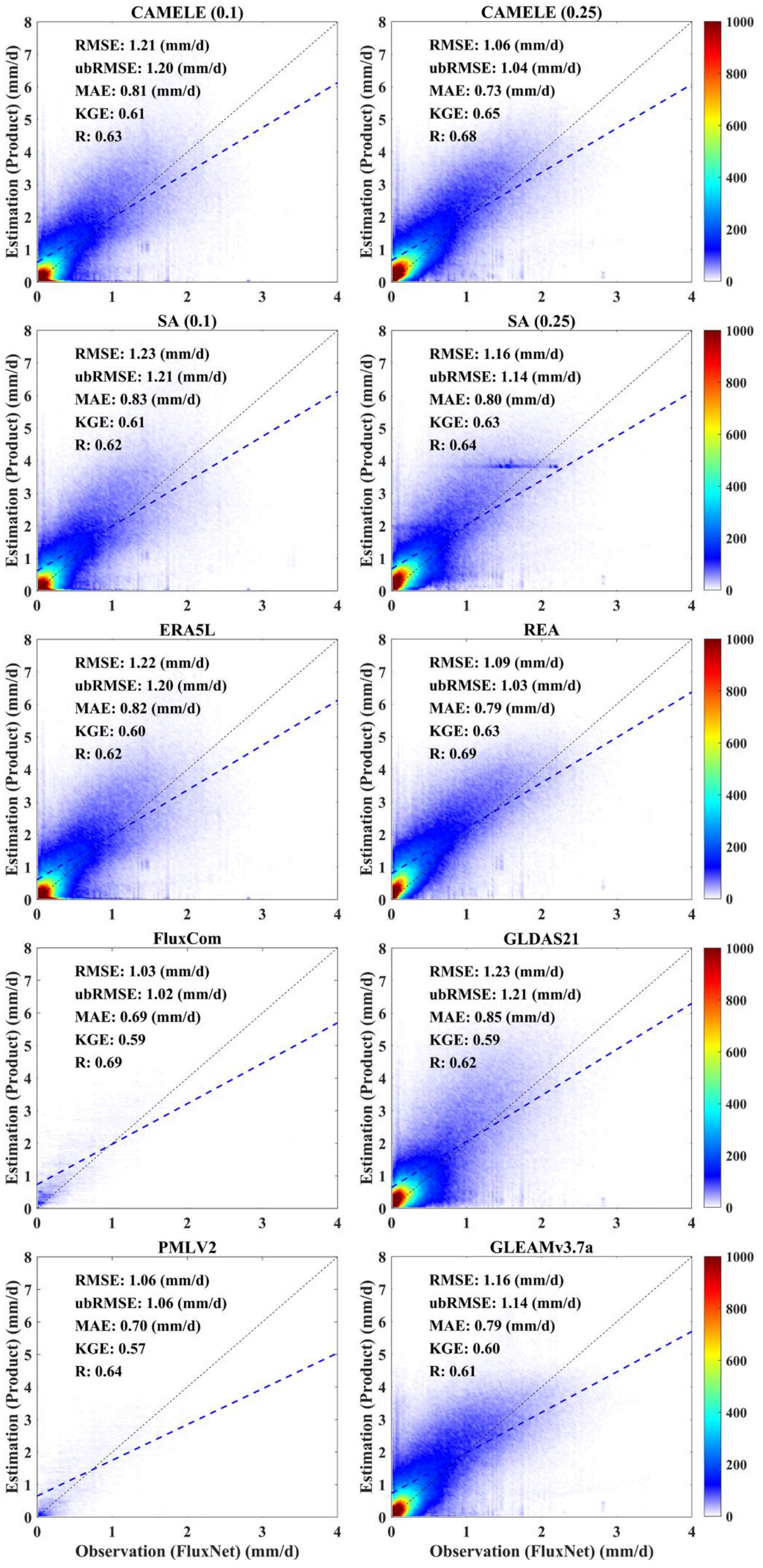
2.32 Line 552

L552 - not all performance metrics in the Figure 4 are unit-less

AC:

Many thanks for pointing out our mistakes. Figure 4 (Now Figure 6) has been updated with a larger font size and correct unit.

Revised Figure (Line 618)


CAMELE (0.1)
RMSE: 1.21 (mm/d)
ubRMSE: 1.20 (mm/d)
MAE: 0.81 (mm/d)
KGE: 0.61
R: 0.63
CAMELE (0.25)
RMSE: 1.06 (mm/d)
ubRMSE: 1.04 (mm/d)
MAE: 0.73 (mm/d)
KGE: 0.65
R: 0.68
SA (0.1)
RMSE: 1.23 (mm/d)
ubRMSE: 1.21 (mm/d)
MAE: 0.83 (mm/d)
KGE: 0.61
R: 0.62
SA (0.25)
RMSE: 1.16 (mm/d)
ubRMSE: 1.14 (mm/d)
MAE: 0.80 (mm/d)
KGE: 0.63
R: 0.64
ERA5L
RMSE: 1.22 (mm/d)
ubRMSE: 1.20 (mm/d)
MAE: 0.82 (mm/d)
KGE: 0.60
R: 0.62
REA
RMSE: 1.09 (mm/d)
ubRMSE: 1.03 (mm/d)
MAE: 0.79 (mm/d)
KGE: 0.63
R: 0.69
FluxCom
RMSE: 1.03 (mm/d)
ubRMSE: 1.02 (mm/d)
MAE: 0.69 (mm/d)
KGE: 0.59
R: 0.69
GLDAS21
RMSE: 1.23 (mm/d)
ubRMSE: 1.21 (mm/d)
MAE: 0.85 (mm/d)
KGE: 0.59
R: 0.62
PMLV2
RMSE: 1.06 (mm/d)
ubRMSE: 1.06 (mm/d)
MAE: 0.70 (mm/d)
KGE: 0.57
R: 0.64
GLEAMv3.7a
RMSE: 1.16 (mm/d)
ubRMSE: 1.14 (mm/d)
MAE: 0.79 (mm/d)
KGE: 0.60
R: 0.61
Estimation (Product) (mm/d)
Observation (FluxNet) (mm/d)
0
1
2
3
4
5
6
7
8
0
200
400
600
800
1000

2.33 Line 568

L568 “not align with the actual situation” - Needs to be discussed a bit more than this. What do the authors mean by ‘actual situation’?

AC:

Thank you for bringing up this concern. We intended to convey that "simple average assumes that each product performs equally on each grid cell" is inaccurate. We have revised the corresponding description to clarify that different products exhibit variations in performance across different grid cells (regions).

Revised Contents (Line 633 to 636)

“…The assumption that a simple average implies equal performance of each product on every grid cell is inaccurate; variations in performance exist among different products across distinct grid cells (regions)…”

2.34 Line 585, 588

L585, 588: “…exceptionally” – it performs well, not sure if “exceptionally”. “suggests more minor errors” – what do you mean “more minor”?

AC:

Thank you for bringing up this concern. We have accordingly revised the description.

Revised Contents (Line 653 to 656)

“…CAMELE performs well overall, closely resembling PMLv2 and FluxCom. On the other hand, the results obtained from the Simple Average are relatively poorer. Regarding the RMSE, ubRMSE, and MAE indicators, a violin plot with a closer belly to 0 suggests less errors…”

2.35 Line 643

L643: “multi-year…” - as you have shown in Figures 2 and 3, different estimates can exhibit varied performance when different periods are considered. Why compare estimates from mixed periods here?

AC:

We sincerely appreciate your observation regarding the inconsistency in time periods. We have addressed this concern by incorporating new multi-year average distribution figures. Specifically, the 0.1° plot spans from 2001 to 2015, while the 0.25° plot covers the period from 2000 to 2017. The updated results exhibit minimal variations from the previous figures, with discrepancies primarily observed in specific regions. We encourage you to compare the previous figures for a detailed assessment.

New Figures (Line 763 to 766)

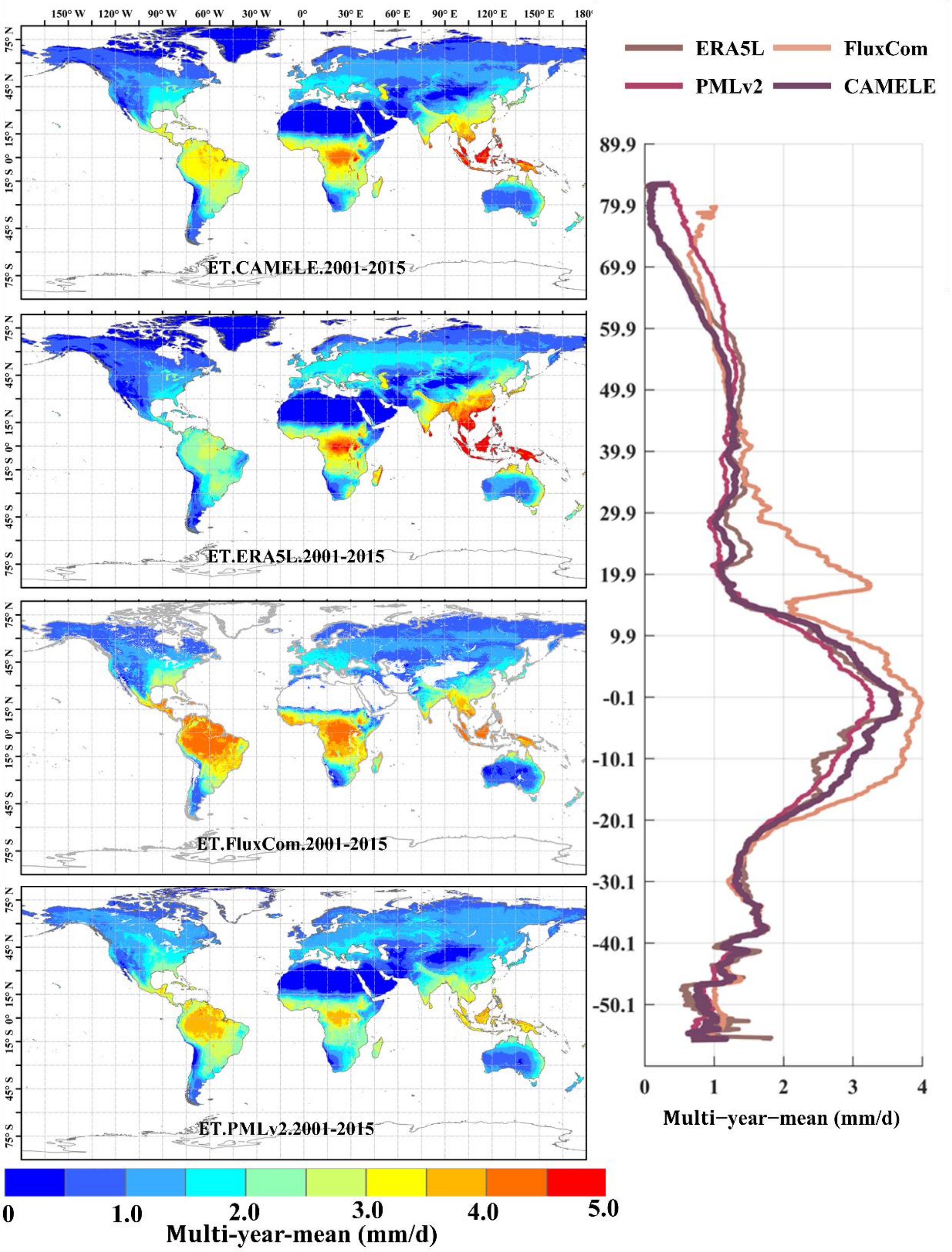


Figure 11 Global distribution of multi-year daily average ET at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

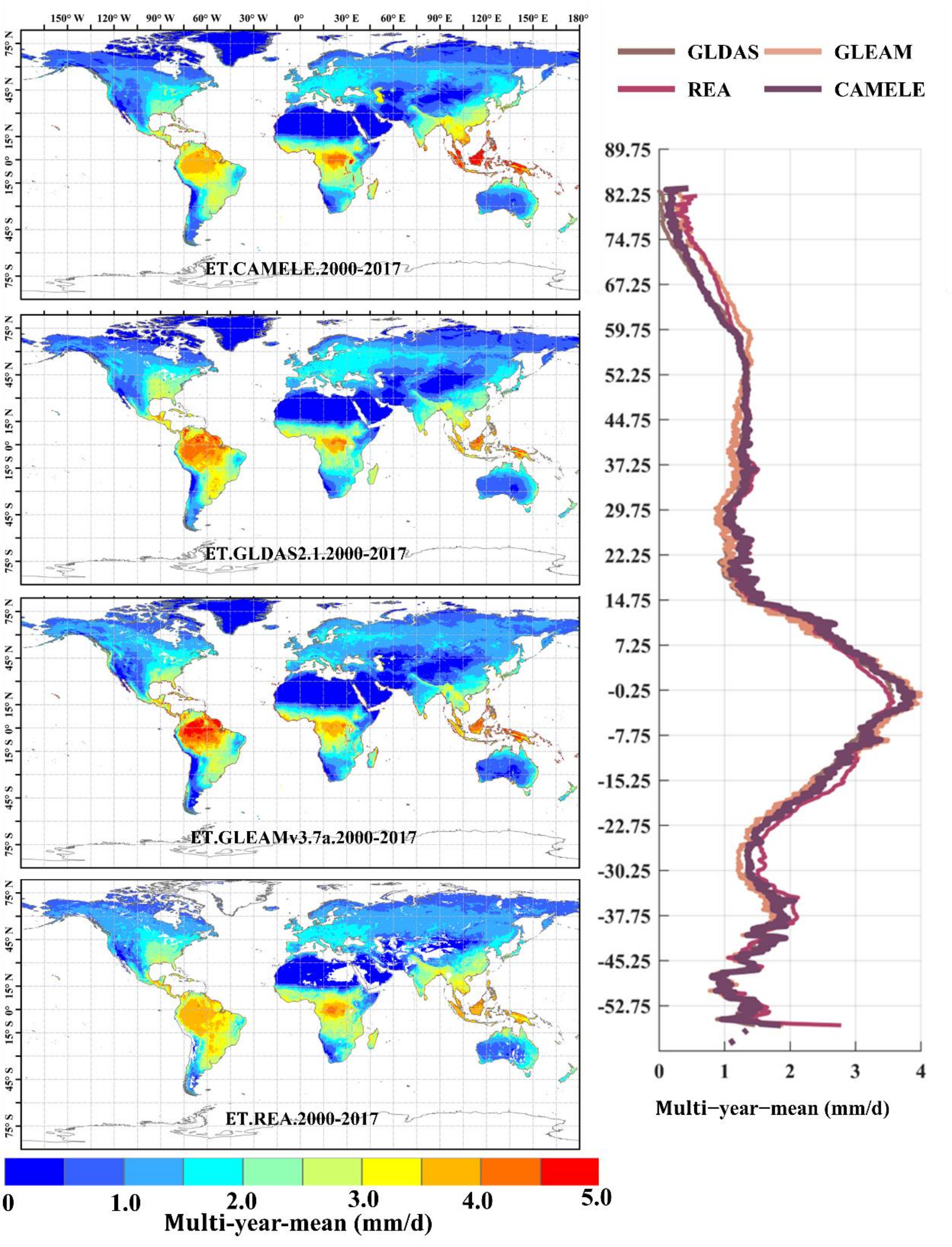


Figure 1 Global distribution of multi-year daily average ET at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

Previous Figures

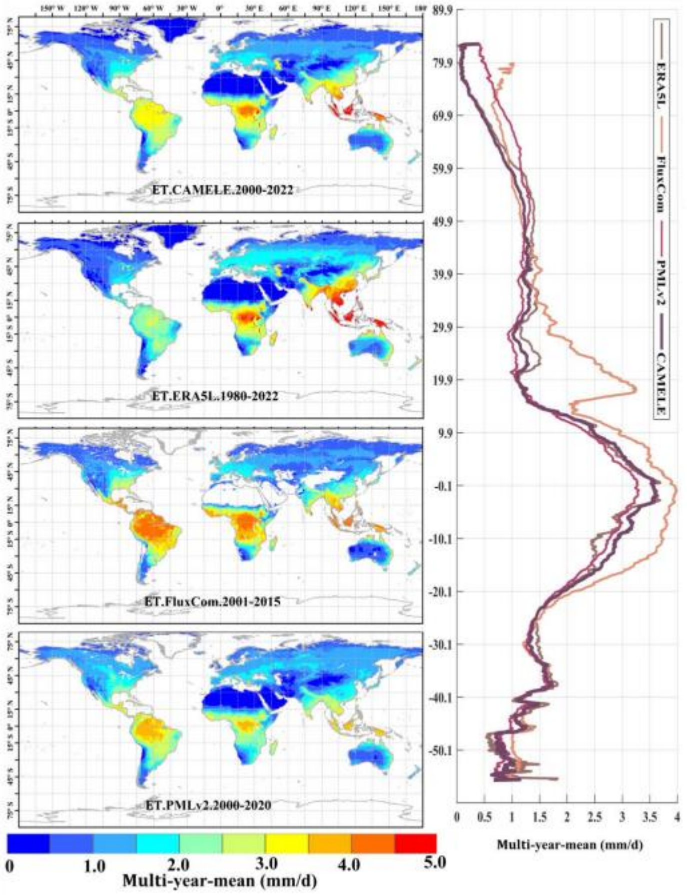


Figure 9 Global distribution of multi-year daily average ET at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

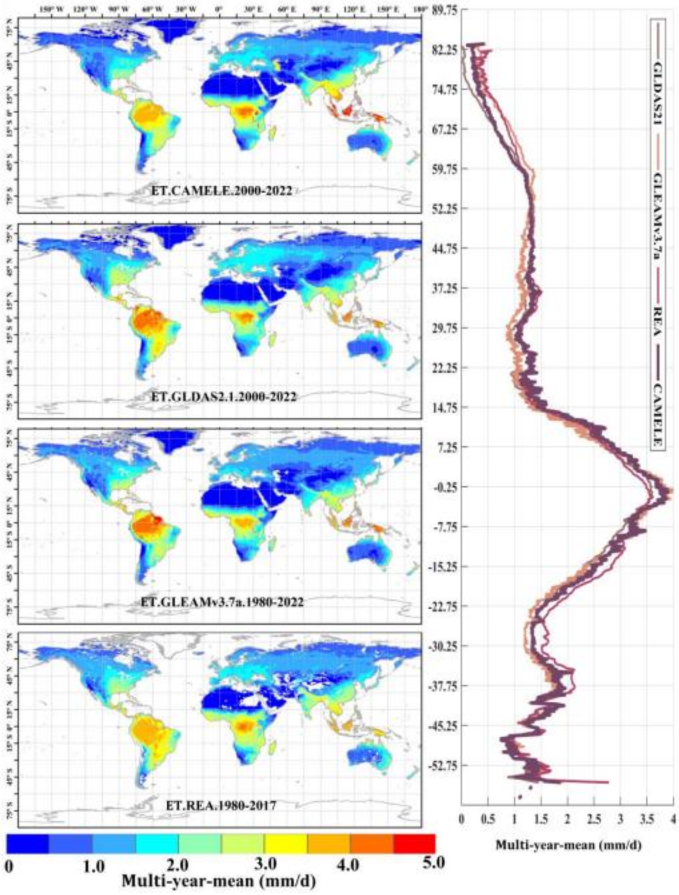


Figure 10 Global distribution of multi-year daily average ET at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

2.36 Line 661

L661 – “while GLDAS and GLEAM have weights of approximately close to 1/3 each” - 1/3 each ? not clear. Also could the authors consider providing the readers with a quantitative illustration of the weights between the 5 products used within CAMELE?

AC:

Thank you very much for pointing out the unclear expression. In this context, when we mentioned "weights of approximately close to 1/3 each" for MERRA2, GLDAS, and GLEAM in the calculation of REA for the Congo Basin and Amazon Rainforests, we were referring to their approximate equal contributions, resulting in REA values distributed among them in a roughly equal manner over multiple years. (found in Lu et al., (2021)) It is important to note that this does not pertain to the fusion weights within CAMELE (as we did not use MERRA2 in CAMELE). We have accordingly clarified our wording.

Additionally, we have enhanced the analysis of weights by introducing a discussion on dominant products based on weights in our response to Query 4 (learning the ways expressed in Park et al., (2023)). Furthermore, detailed information on the weight distribution in different combinations is provided in the appendix.

References:

Lu, J., Wang, G., Chen, T., Li, S., Hagan, D. F. T., Kattel, G., Peng, J., Jiang, T., and Su, B.: A harmonized global land evaporation dataset from model-based products covering 1980–2017, Earth System Science Data, 13, 5879–5898, https://doi.org/10.5194/essd-13-5879-2021, 2021.

Park, J., Baik, J., and Choi, M.: Triple collocation-based multi-source evaporation and transpiration merging, Agricultural and Forest Meteorology, 331, 109353, 2023.

Revised Contents (Line 782 to 786)

"…The assigned weights for REA's inputs (MERRA2, GLDAS, and GLEAM.) are approximately equal in these two regions, each contributing about one-third to the overall calculation (Lu et al., 2021). This balanced allocation results in the REA being distributed among them roughly equally over multiple years in these two regions…"

2.37 Line 672

L672: "…of average with latitude" – do the authors mean the "average trend with latitude"? how is this trend over the different periods calculated also why is there no consistency in the periods considered? the trend in ERA5L which is from 1980-2022 appears to have a negative trend at the tropics (especially in Africa close to the Equator). Additionally, unlike all the other products, CAMELE appears to have pronounced negative trends in the southern hemisphere, why? How can the weighted output (CAMELE) have higher negative trends than the input/ensemble members? can the authors provide the trends of the other 2 ensemble products to aid with interpretation?

AC:

We sincerely appreciate your insightful comments, which are crucial for the accurate calculation of trends. We have re-plotted the trends for various products, including 0.1° (2001-2015) and 0.25° (2000-2017) datasets, along with CAMELE, highlighting regions with significant changes. The trends are estimated using Theil–Sen's slope method, and their significance is tested with the Mann–Kendall method. The dotted areas indicate trends passing the significance test at a 5% level.

Additionally, we have rectified the coding error in the original 0.1° trend plot, where latitude variation was incorrectly portrayed as the dependent variable. Please find the corrected trend for CAMELE, demonstrating consistency among input ensemble members. Furthermore, modifications have been made to the figure captions for clarity.

Revised Figures (Line 851 to 863)

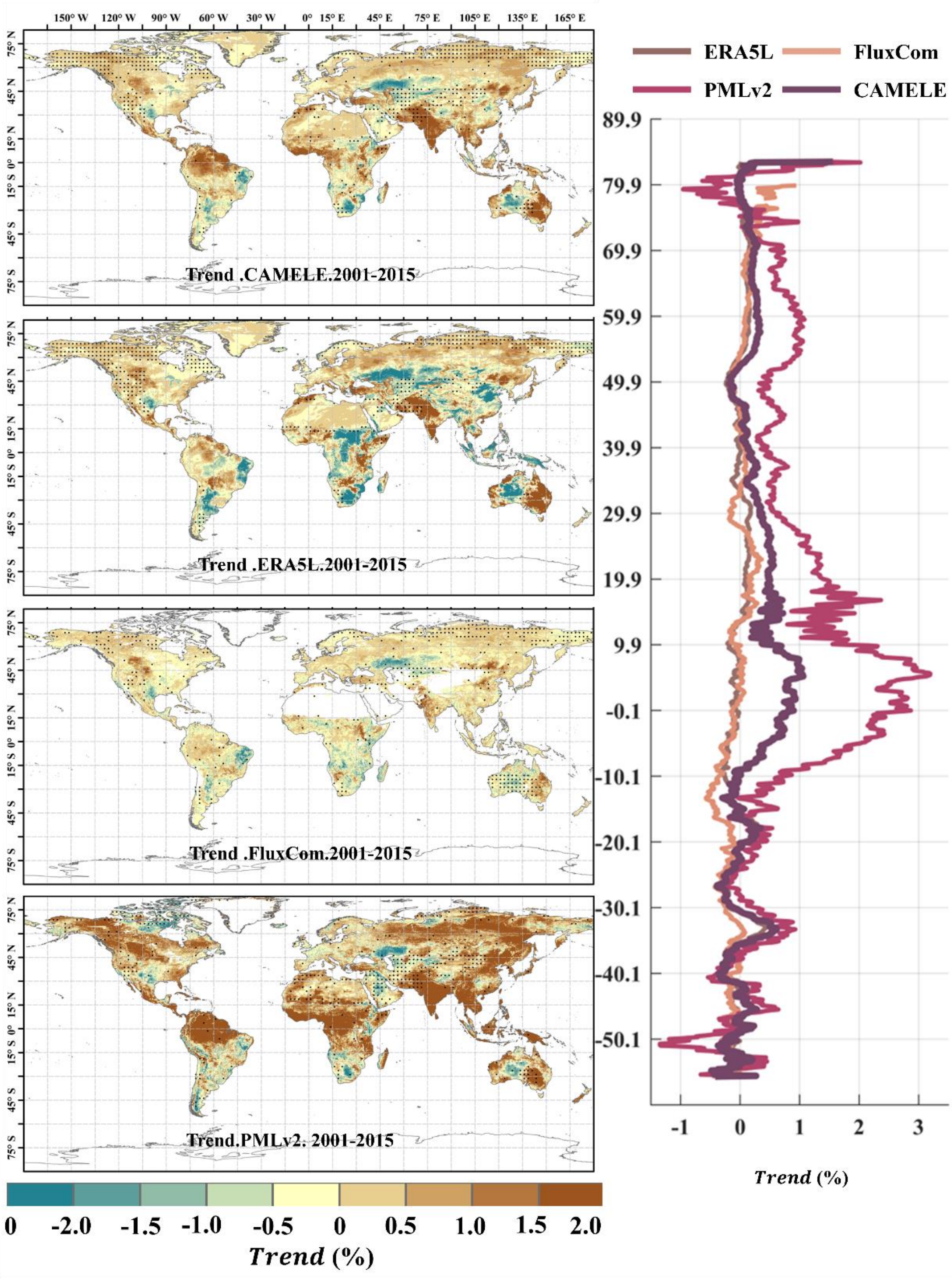


Figure 15 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area indicates that the trend has passed the significance test at 5 % level.

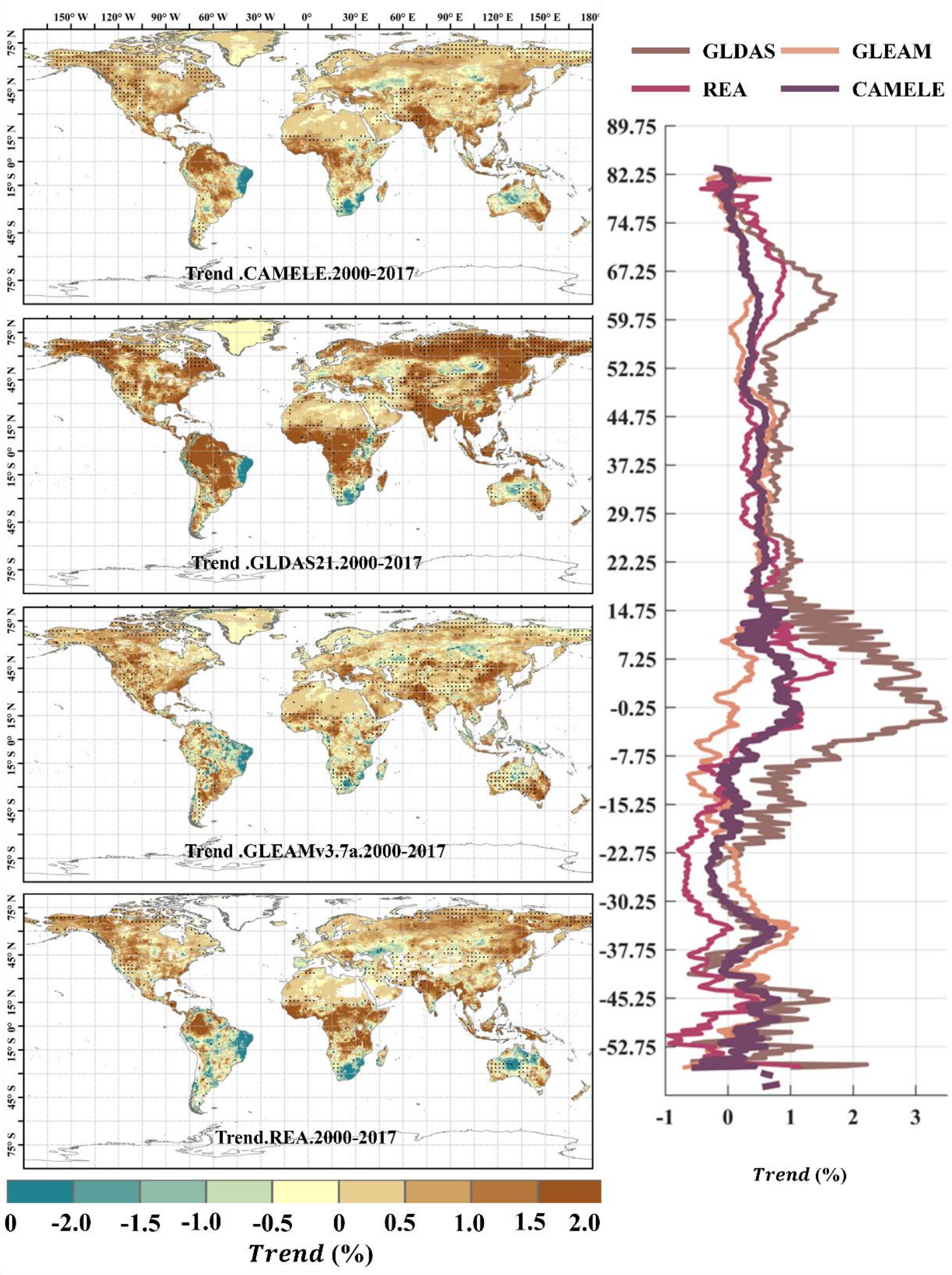


Figure 16 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area indicates that the trend has passed the significance test at 5 % level.

Previous Figures

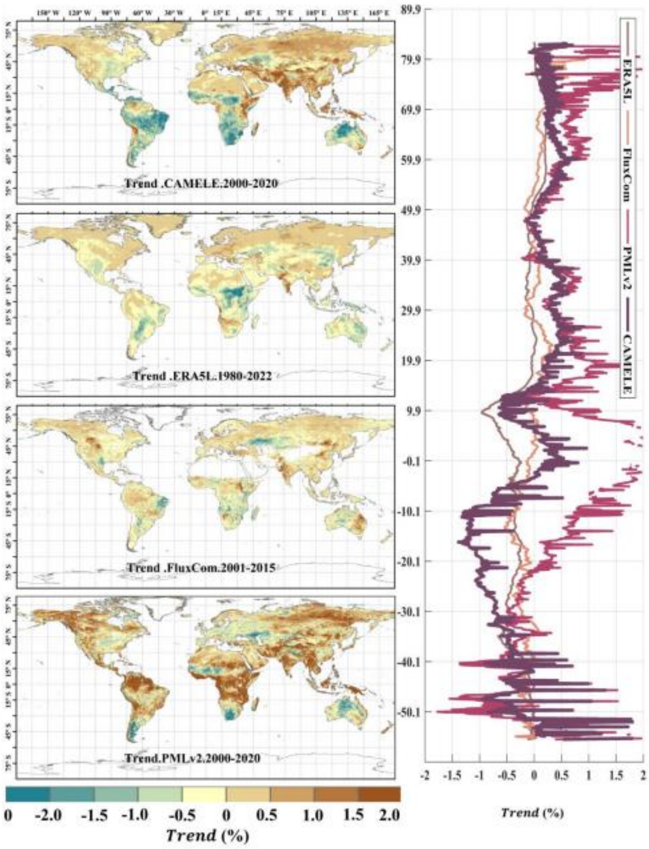


Figure 9 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

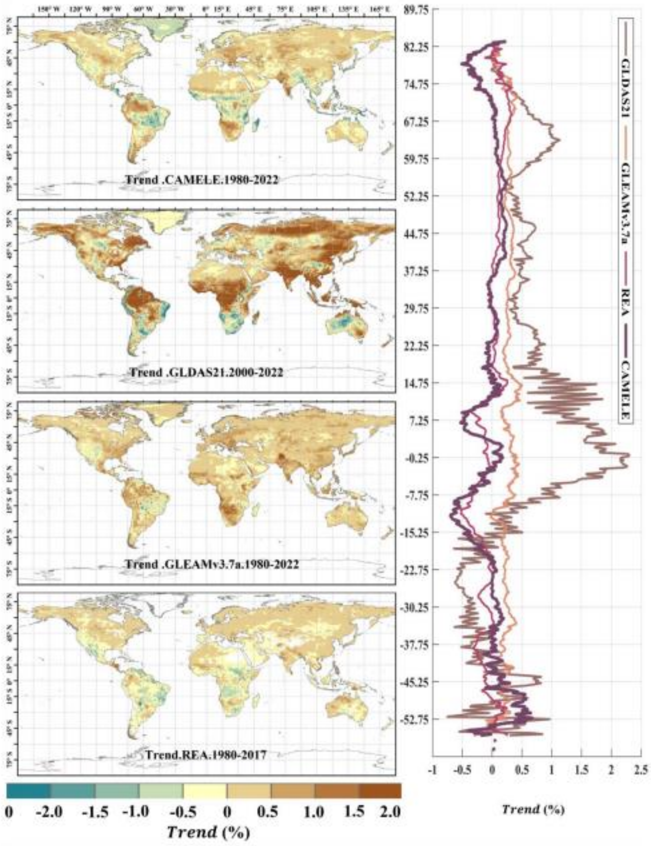


Figure 10 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

2.38 Line 691

L691: “introduce specific impacts” - what specific impacts? using consistent comparison periods would help readers and the broader scientific community better interpret the results.

AC:

The periods under different resolutions are now consistent. Therefore, the mentioned sentence has been removed.

2.39 Line 693

L693: “characteristics of the data itself influence this” - maybe explain how the grid and data characteristics influence the temporal trend or point to the section where it is discussed.

AC:

Regarding your comment, the phrase "characteristics of the data itself influence this" has been removed in the revised version as we encountered difficulty recalling the original intention behind it.

2.40 Line 695-744

L695-744: This section (5.1) is too general …and the authors are really not discussing the results as presented in the previous chapter. Shorten the section OR consider moving to introduction or methodology section as justification of the algorithms selected in this study.

AC:

Thank you for your valuable suggestion. We have reduced one quarter of the length in Section 5.1. Considering that placing this part in the methods section would make it excessively long and it is indeed more appropriate for the discussion, we have opted for shortening only.

2.41 Line 808

L808: "This could be attributed to the variations in the input products" – what variations?

AC:

The ambiguity in our expression has been addressed, and the statement has been removed for clarity.

2.42 Line 817-818

L817-818: "GLEAMv3.7b and GLDAS2.2 employed the satellite data from MODIS, introducing rand" - Citation needed. Also, in L174 you talk of error homogeneity arising from ERA5L and GLDAS (due to meteorological inputs) but the same is not discussed here.

Reference from the reviewer

Jiménez, C., Prigent, C., Mueller, B., Seneviratne, S. I., McCabe, M. F., Wood, E. F., … Wang, K. (2011). Global intercomparison of 12 land surface heat flux estimates. *Journal of Geophysical Research Atmospheres*, *116*(2), 1–27. https://doi.org/10.1029/2010JD014545

Park, J., Baik, J., & Choi, M. (2023). Triple collocation-based multi-source evaporation and transpiration merging. *Agricultural and Forest Meteorology*, *331*(February), 109353. https://doi.org/10.1016/j.agrformet.2023.109353

Zhang, Y., Kong, D., Gan, R., Chiew, F. H. S., McVicar, T. R., Zhang, Q., & Yang, Y. (2019). Coupled estimation of 500 m and 8-day resolution global evapotranspiration and gross primary production in 2002–2017. *Remote Sensing of Environment*, *222*(December 2018), 165–182. https://doi.org/10.1016/j.rse.2018.12.031

AC:

Thank you for pointing out the oversight. The correct statement should address the correlation between GLDAS2.1 and ERA5L, and we have accordingly made the necessary revision.

Revised Contents (Line 1004 to 1007)

"…The relatively poorer performance of other fusion schemes could be due to the lack of consideration for non-zero ECC. For example, non-zero ECC between GLDAS-2.2 and ERA5L has been reported in a recent study (Li et al., 2023a) …"

Reference:

Li, C., Liu, Z., Tu, Z., Shen, J., He, Y., and Yang, H.: Assessment of global gridded transpiration products using the extended instrumental variable technique (EIVD), Journal of Hydrology, 623, 129880, https://doi.org/10.1016/j.jhydrol.2023.129880, 2023a