➢ 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:

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

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:

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

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.

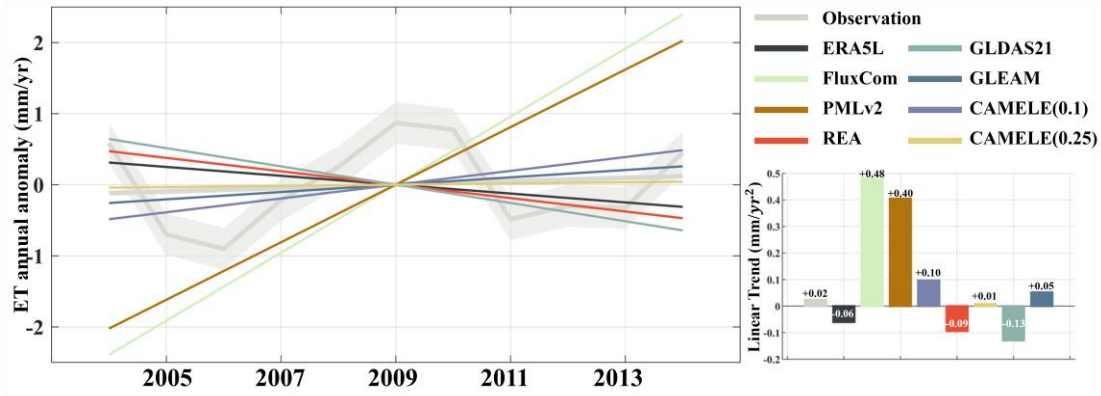


Figure 13 Comparison of linear trend from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The trends have been subjected to SSA decomposition, removing seasonality. The gray enveloping line represents the mean plus the standard deviation of the 13 sites.

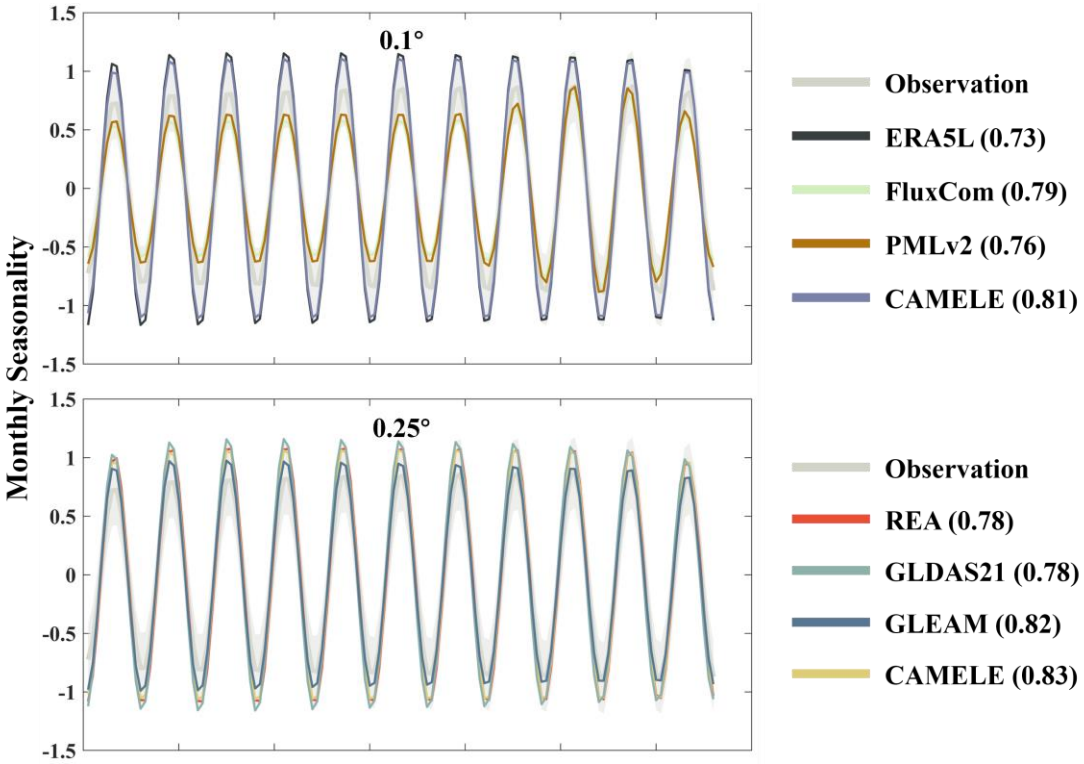


Figure 14 Comparison of seasonal variations from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The seasonality has been obtained through SSA decomposition, with the gray area representing the observed values. The parentheses in each product name indicate the KGE coefficient comparing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

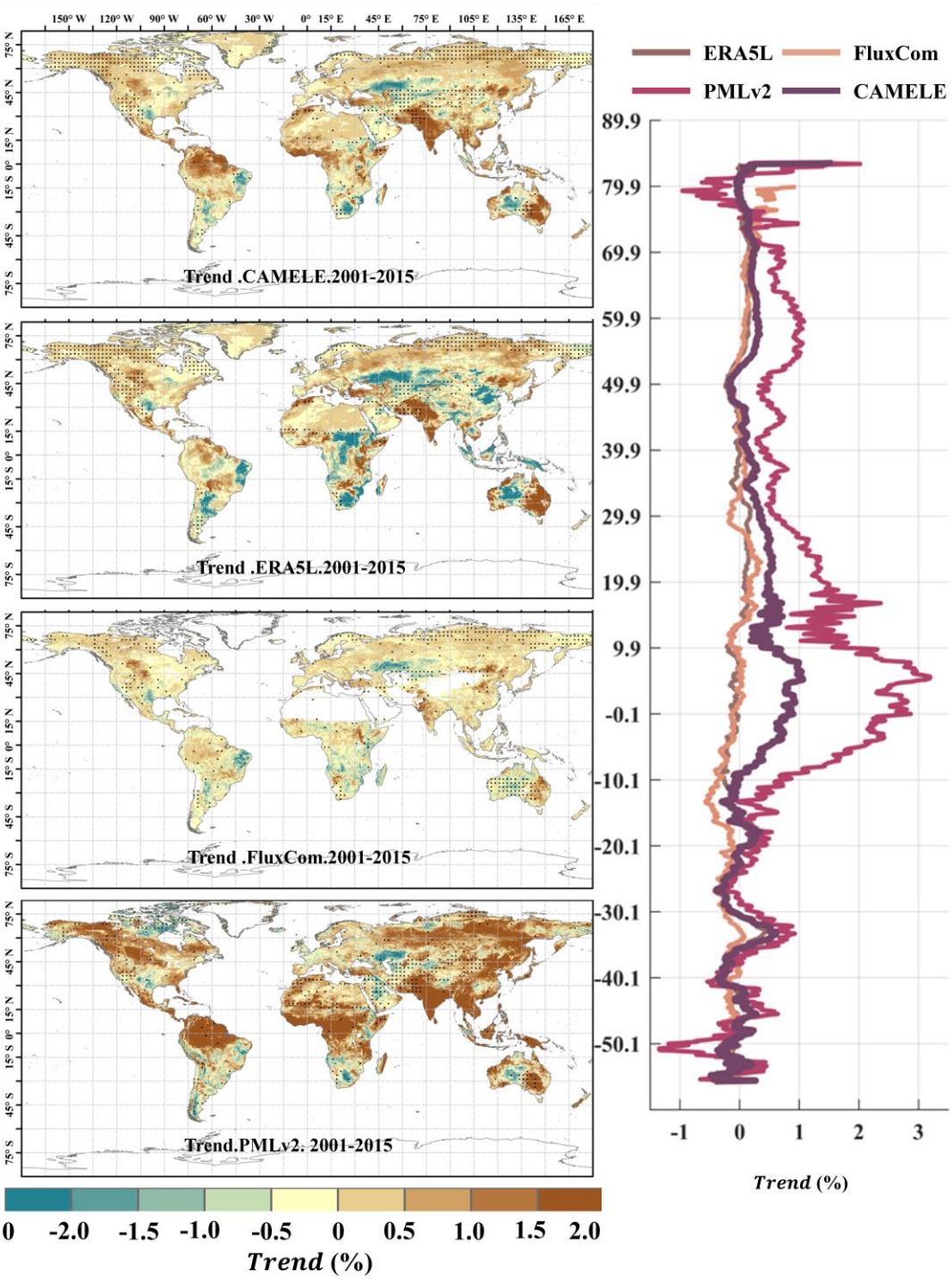


Figure 15 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area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

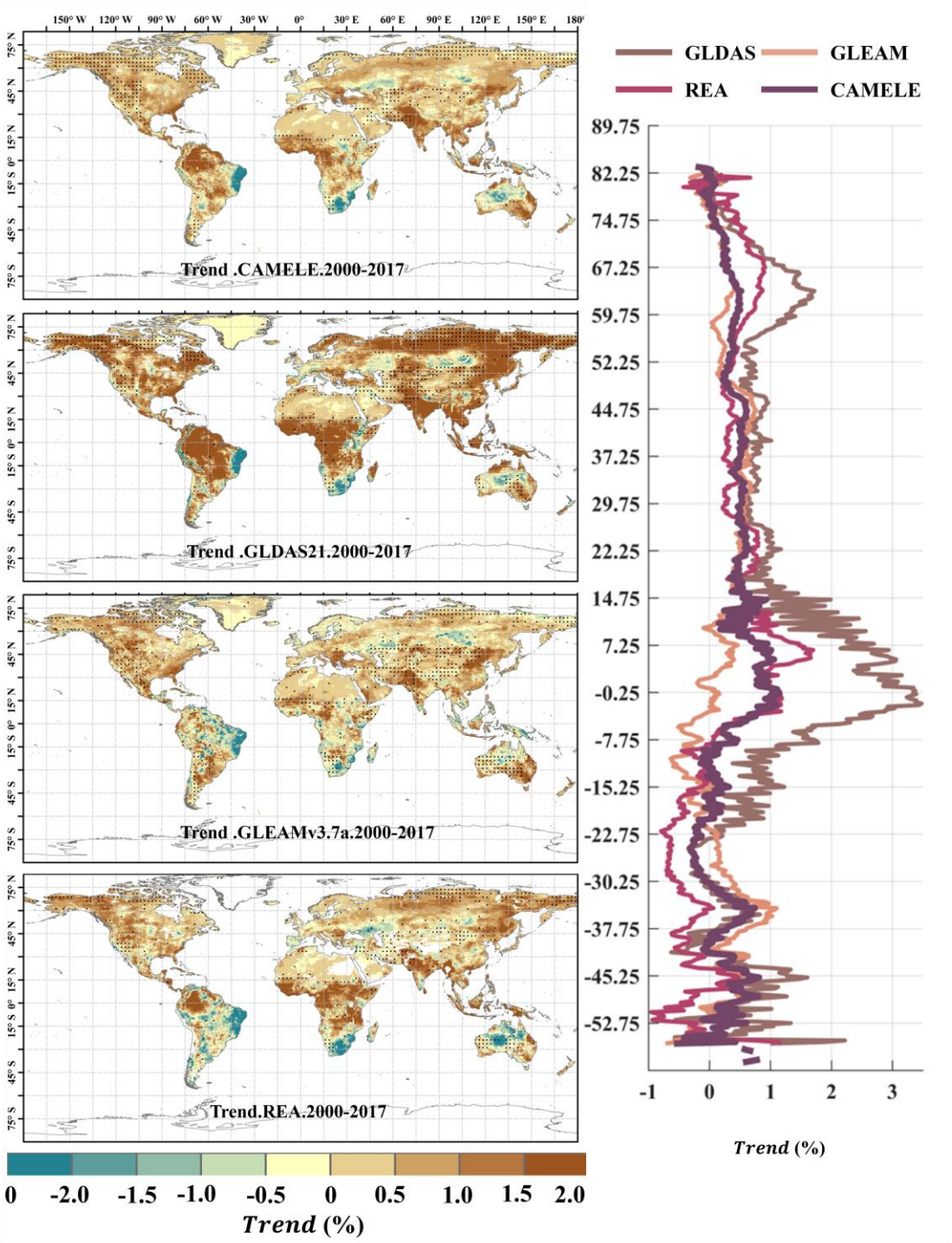


Figure 16 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted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

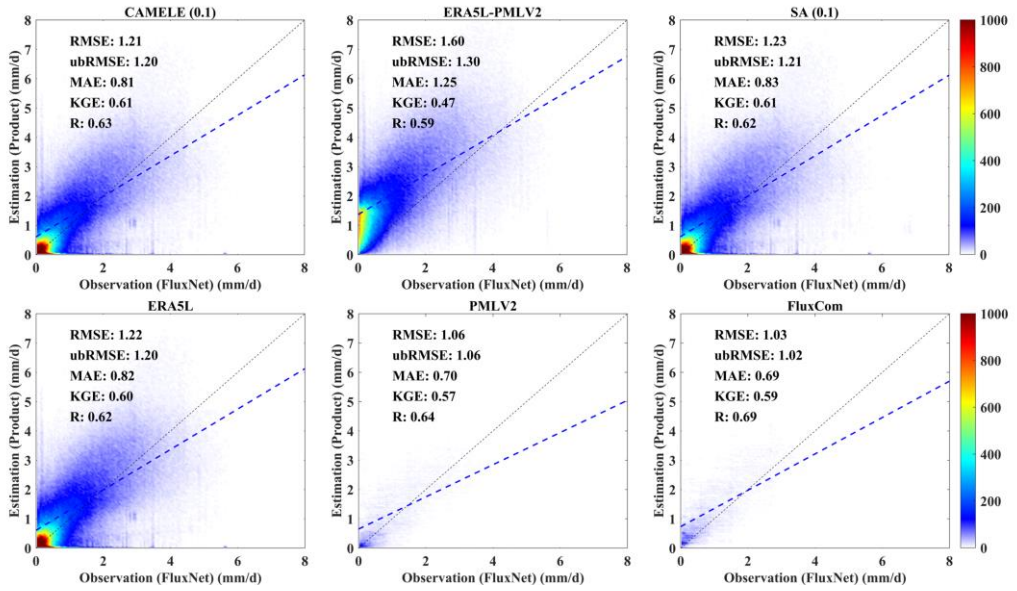
Qi, W., Liu, J., and Chen, D.: Evaluations and Improvements of GLDAS2.0 and GLDAS2.1 Forcing Data's Applicability for Basin Scale Hydrological Simulations in the Tibetan Plateau, JGR Atmospheres, 123, https://doi.org/10.1029/2018JD029116, 2018.

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.

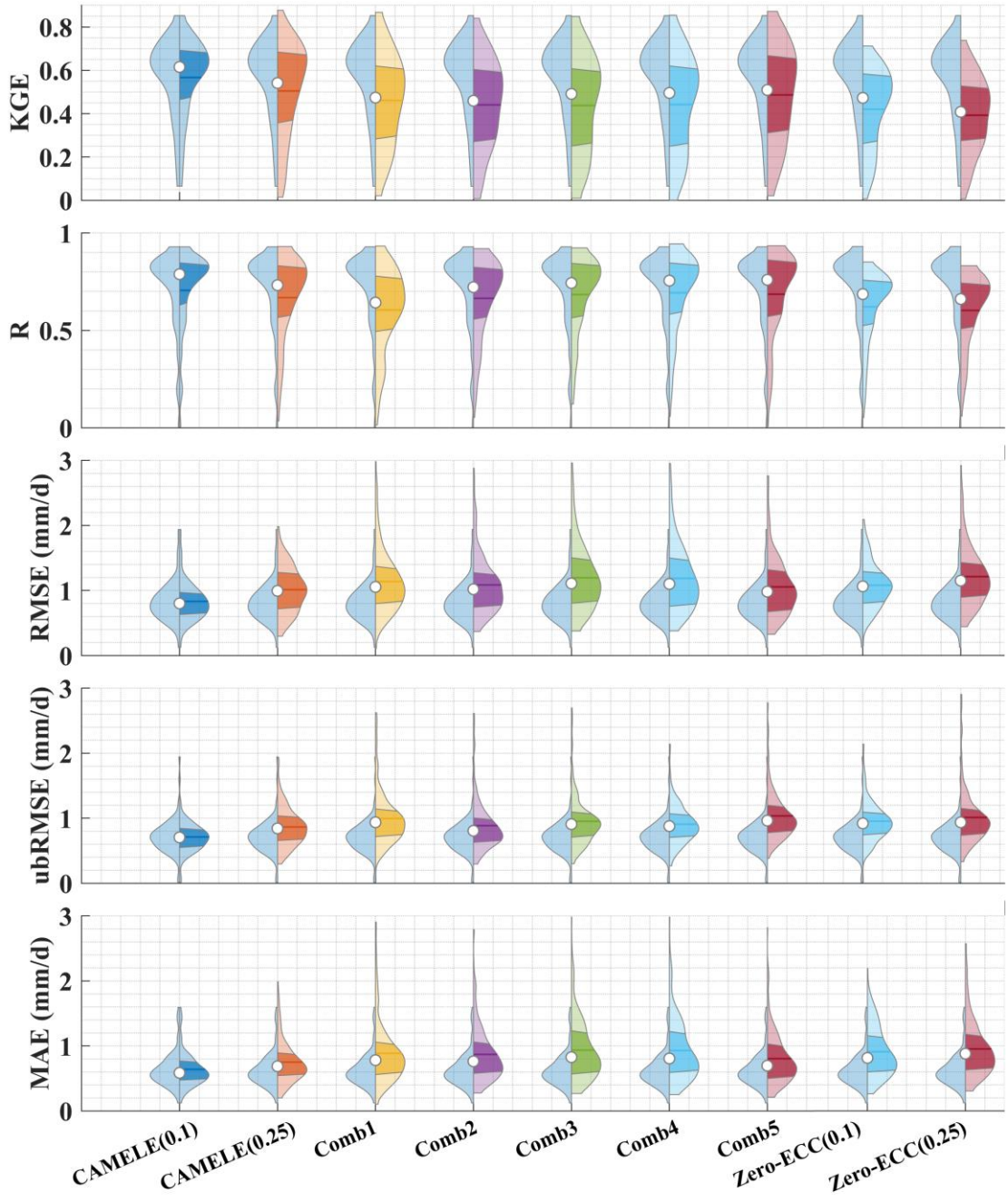


Figure 13 Violin plot comparing KGE, R, RMSE, ubRMSE and MAE of CAMELE with other fusion schemes. The right half of each violin plot represents the distribution, with shaded areas indicating the box plot, where the horizontal line corresponds to the median and the dot represents the mean. The left half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

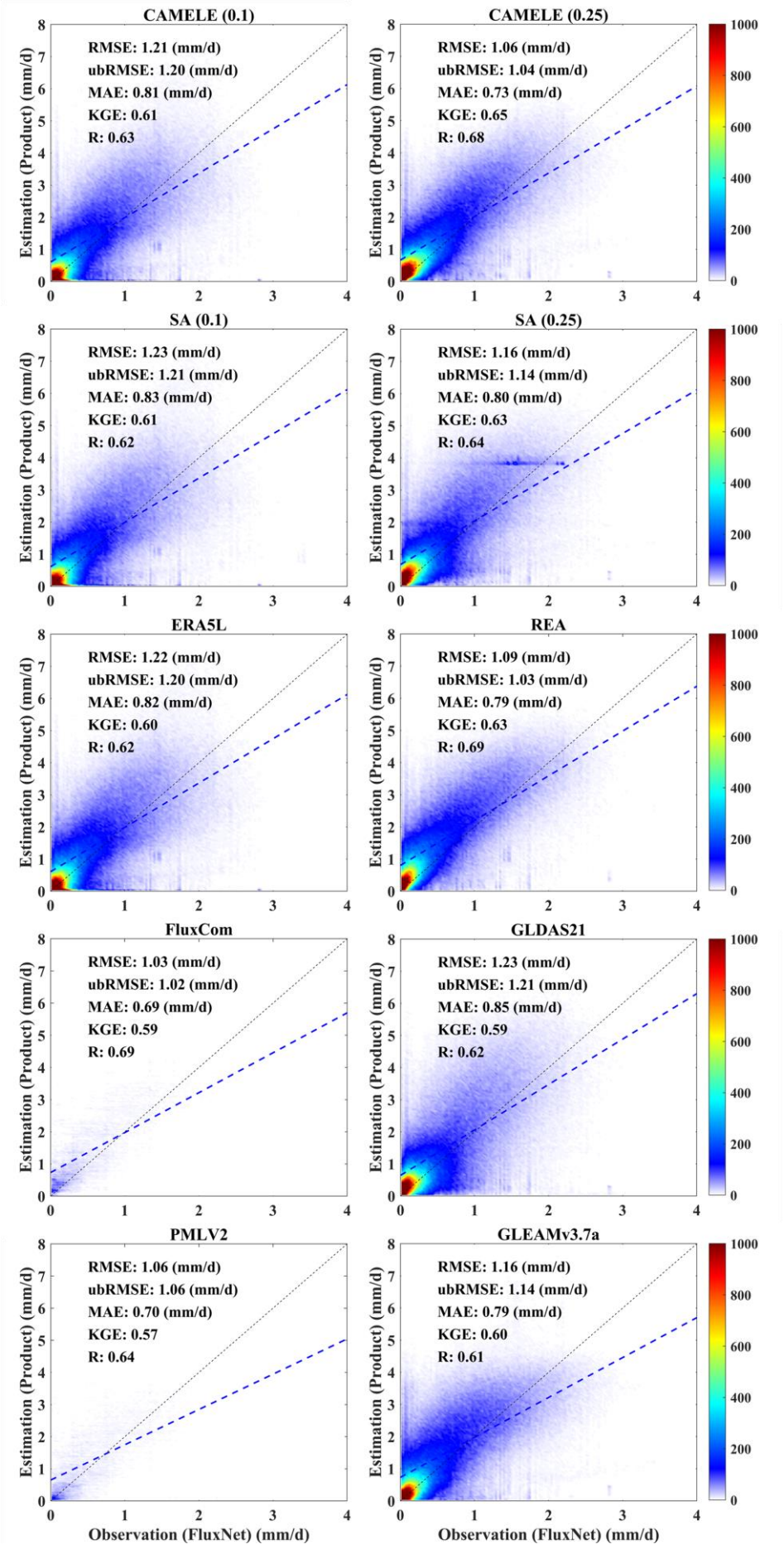
"The International-Geosphere–Biosphere Program (IGBP) land cover classification system (Loveland et al., 1999) was employed to distinguish the 13 Plant Functional Types (PFTs) across sites, including evergreen needle leaf forests (ENF, 49 sites), evergreen broadleaf forests (EBF, 15 sites), deciduous 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

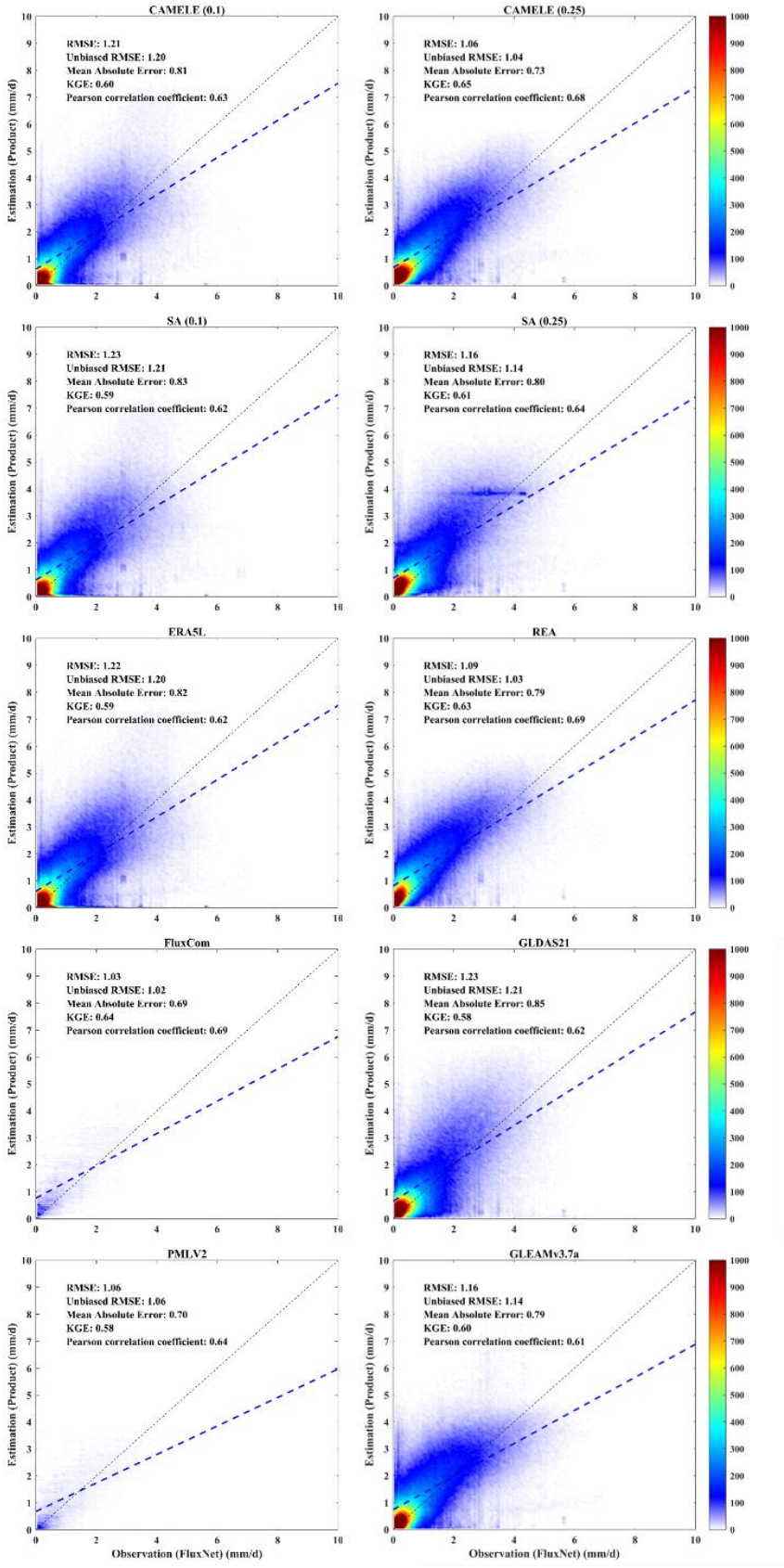
The quality of Figure 4 could be improved. Consider enhancing the clarity and readability of the figure. You might want to simplify the information presented or consider moving some details to a supplementary figure.

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 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:

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."