➢ 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.

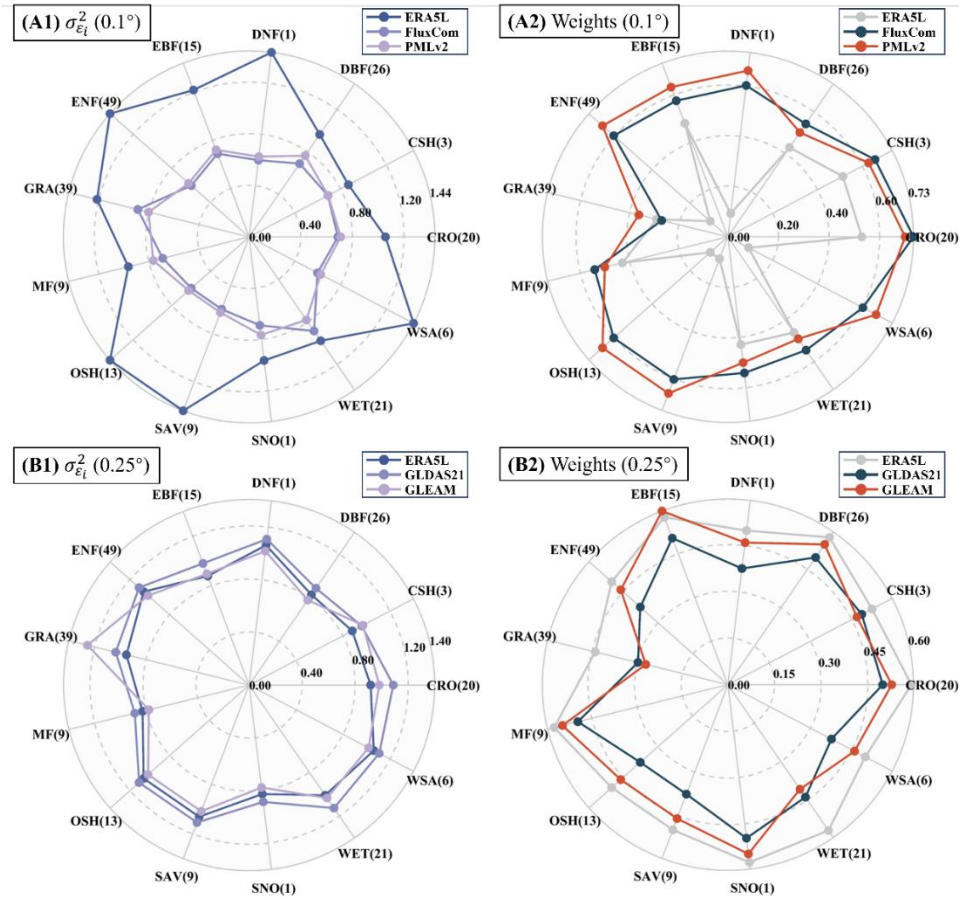


Figure 9 Mean collocation-based errors and weights of different products at various PFTs sites at (A) 0.1° and (B) 0.25° resolutions. The parentheses next to each PFTs name denote the corresponding number of sites.

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.

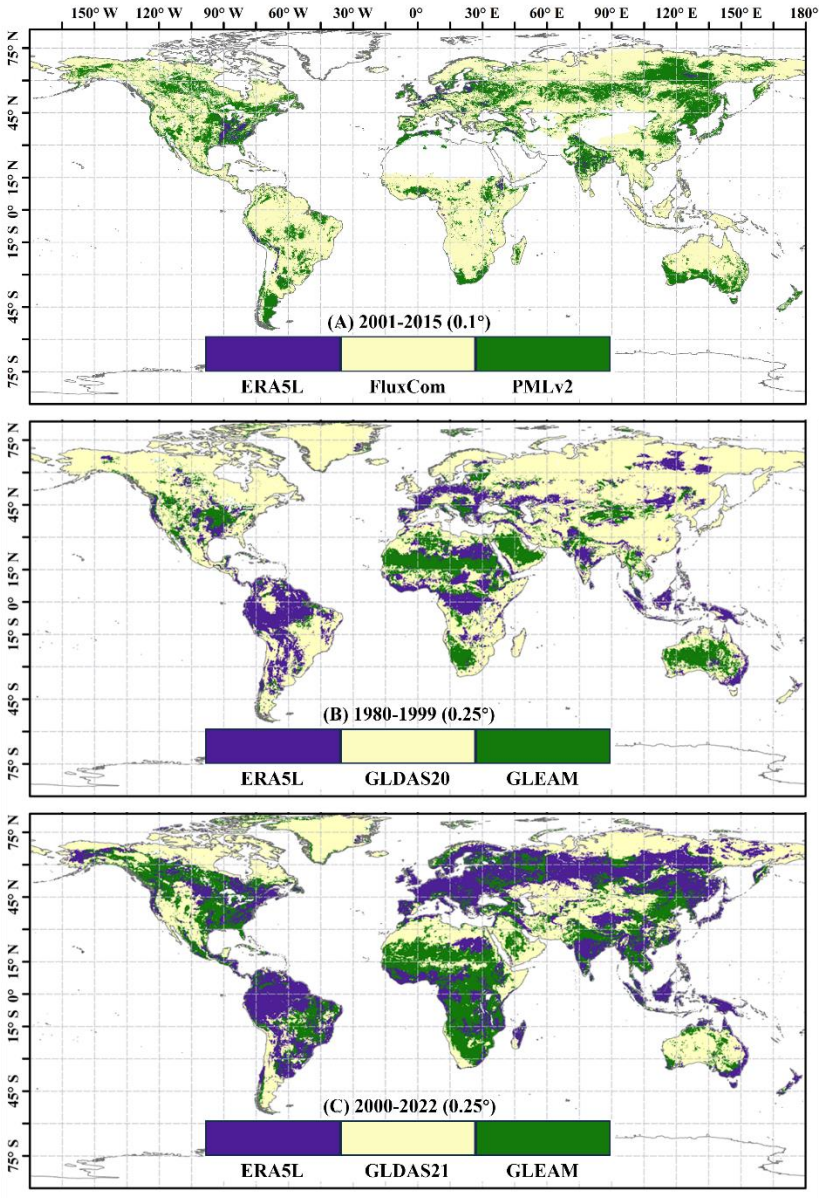


Figure 5 Map of the prevailing product at individual pixels based on scenario-specific weights.

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.

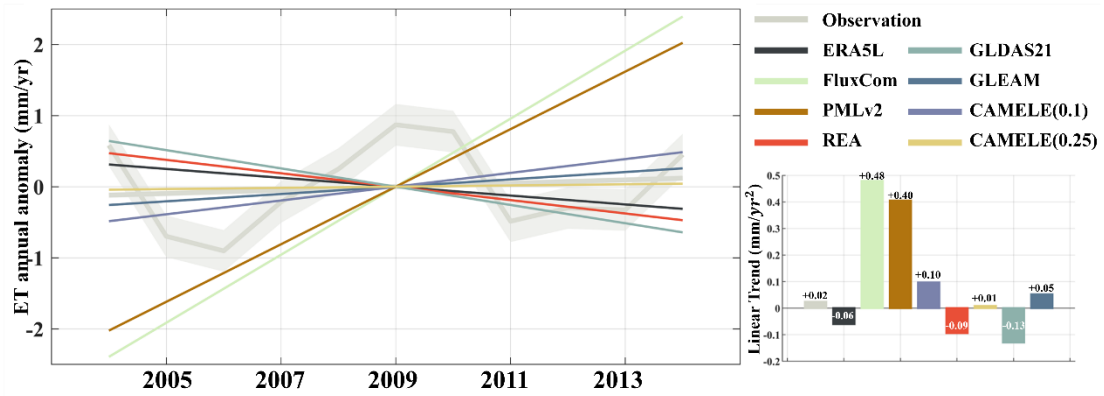


Figure 12 Comparison of linear trend from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The trends have been subjected to SSA decomposition, removing seasonality. The gray enveloping line represents the mean plus the standard deviation of the 13 sites.

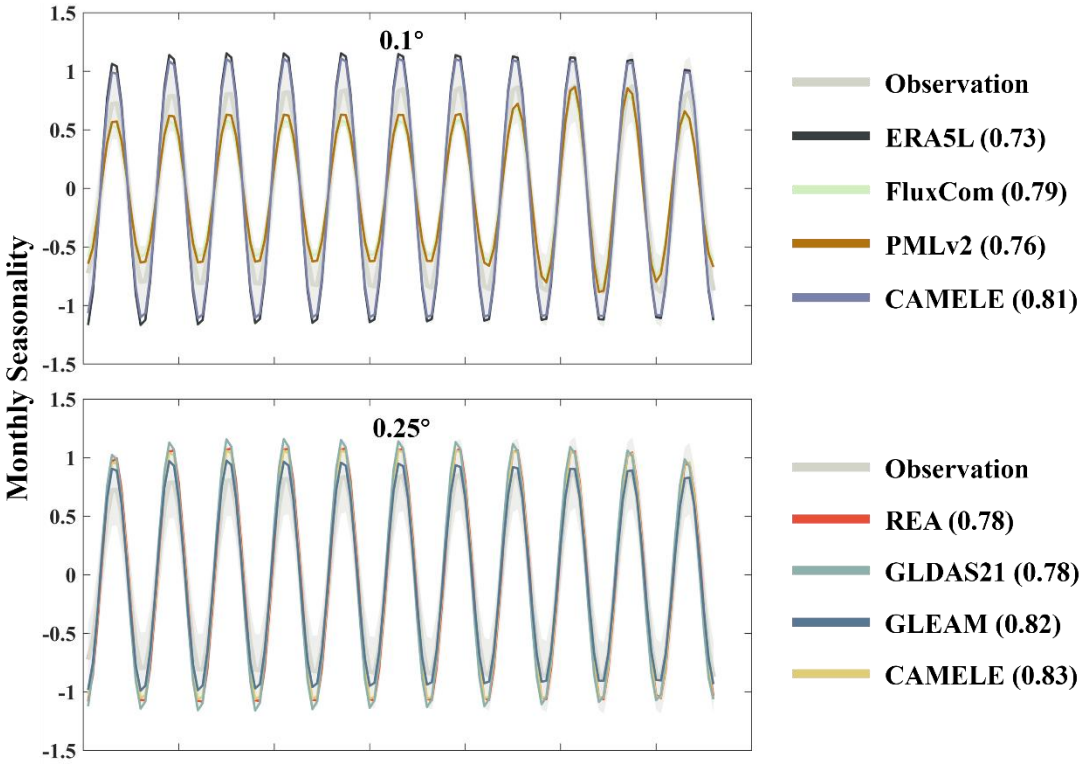


Figure 13 Comparison of seasonal variations from 2004 to 2014 among 13 FluxNet sites using CAMELE and other products. The seasonality has been obtained through SSA decomposition, with the gray area representing the observed values. The parentheses in each product name indicate the KGE coefficient comparing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

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

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.

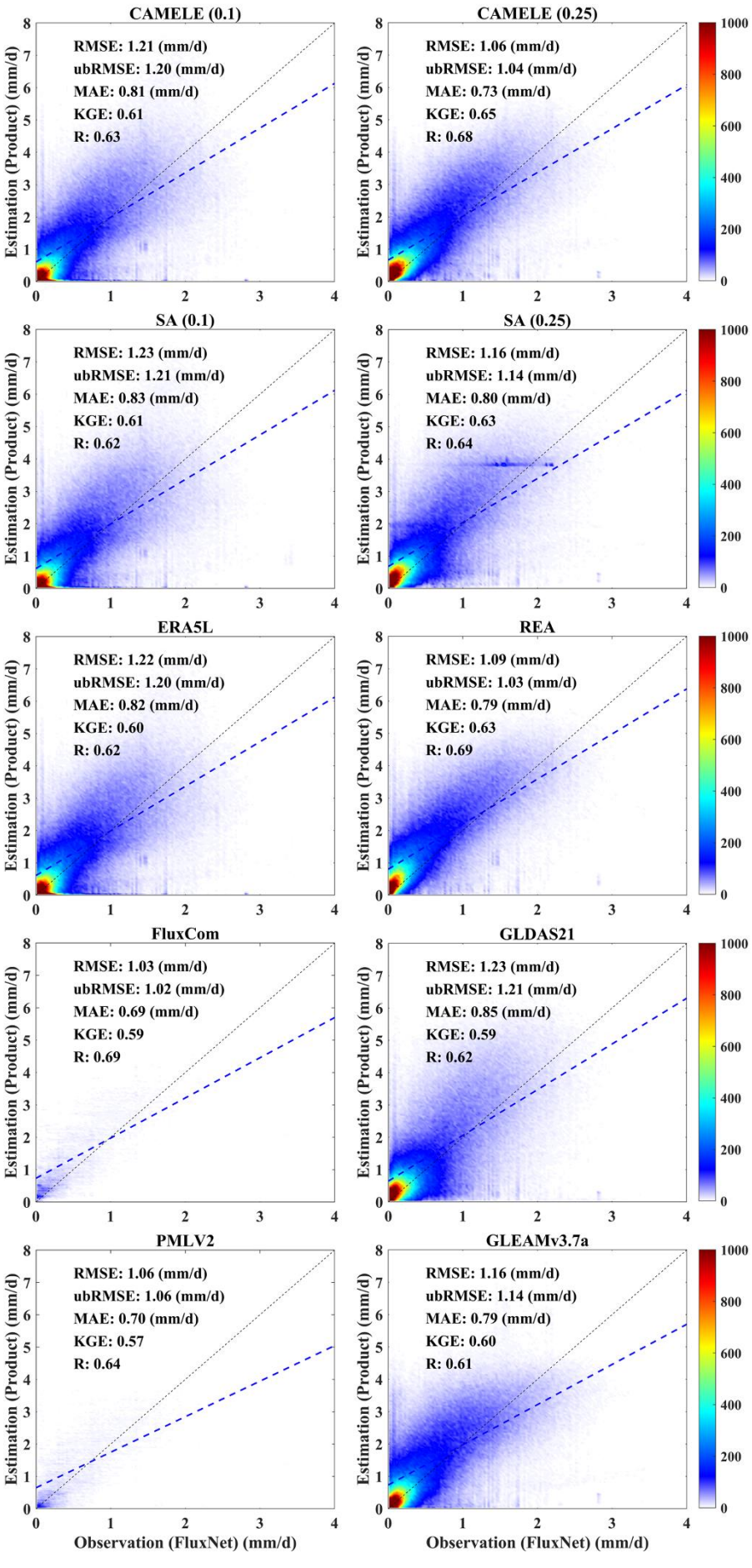
2.32 Line 552

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)

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)

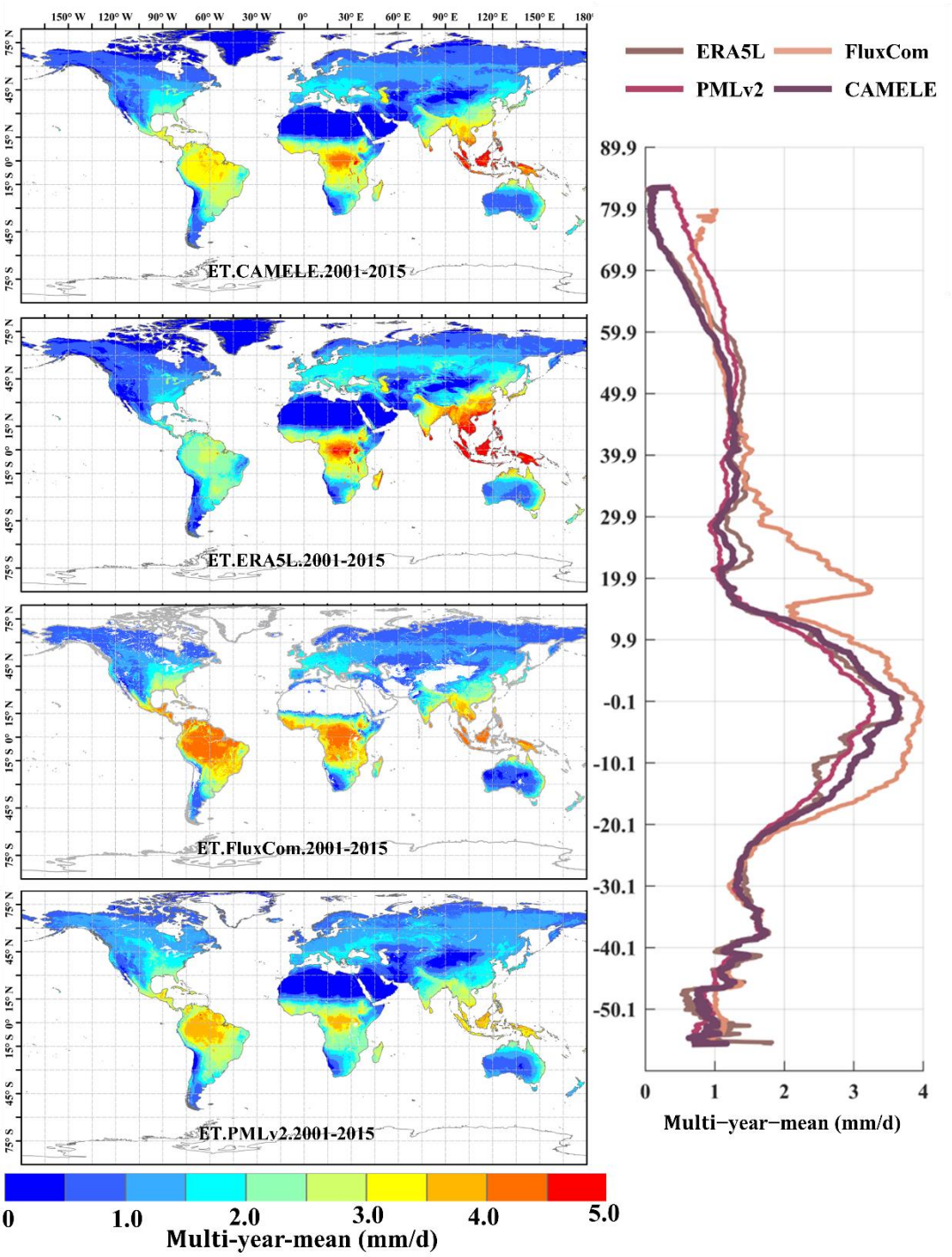


Figure 11 Global distribution of multi-year daily average ET at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

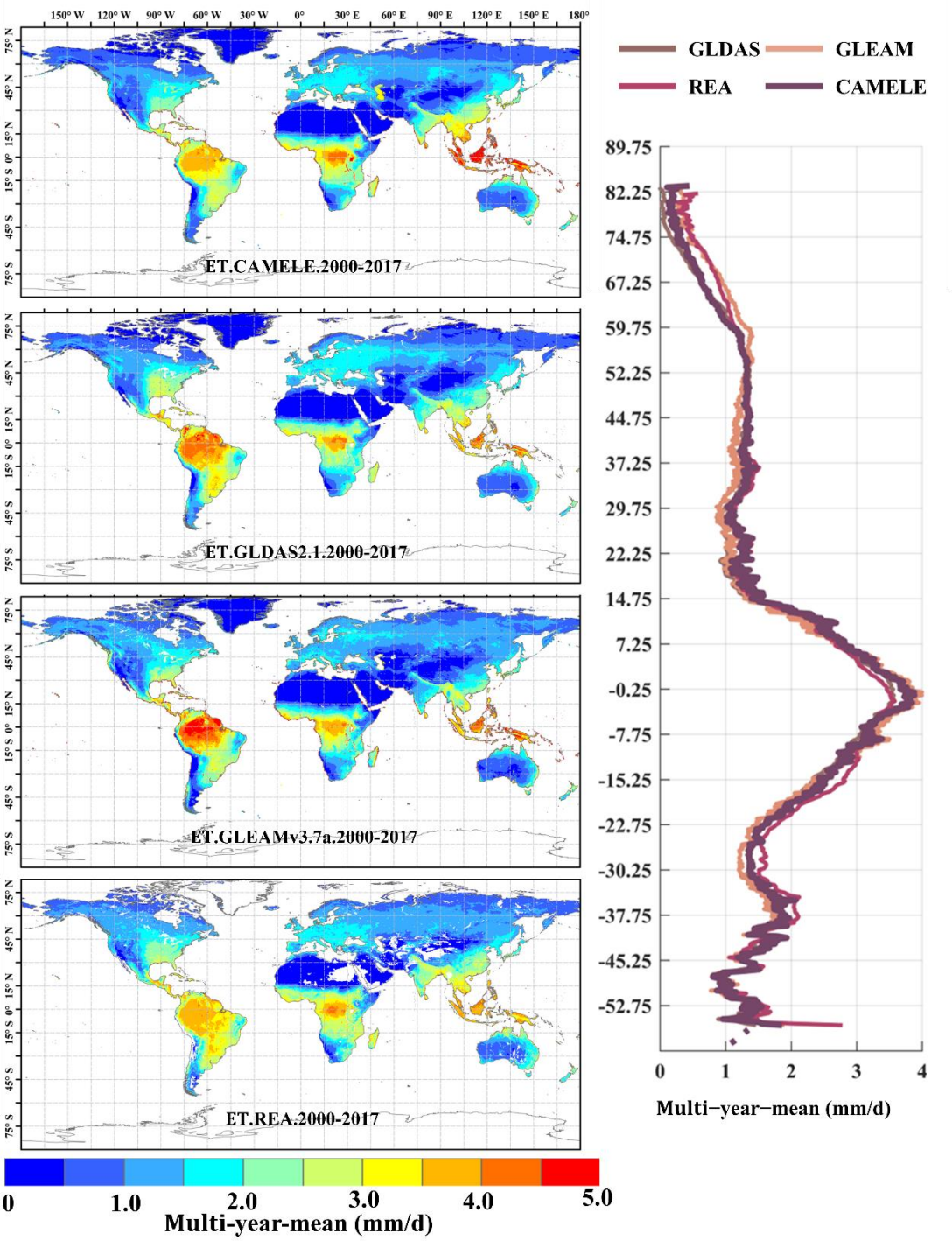


Figure 1 Global distribution of multi-year daily average ET at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

Previous Figures

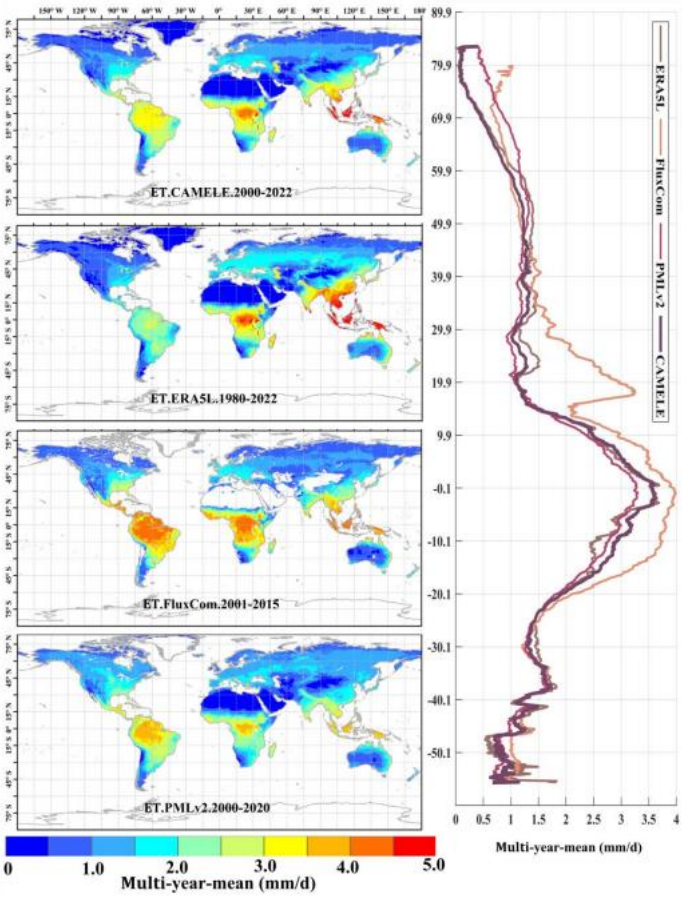


Figure 9 Global distribution of multi-year daily average ET at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

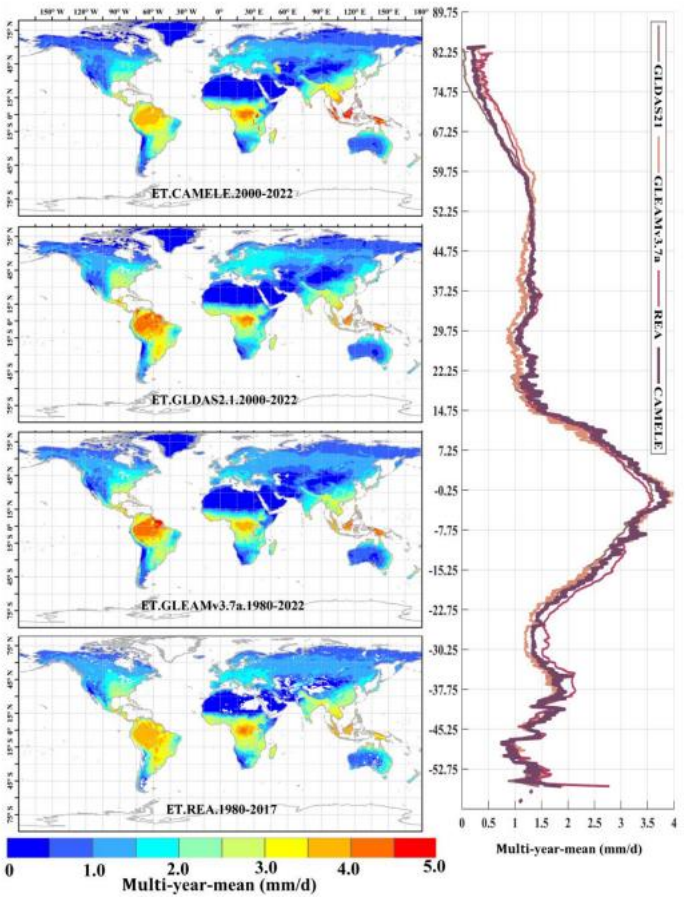


Figure 10 Global distribution of multi-year daily average ET at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

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)

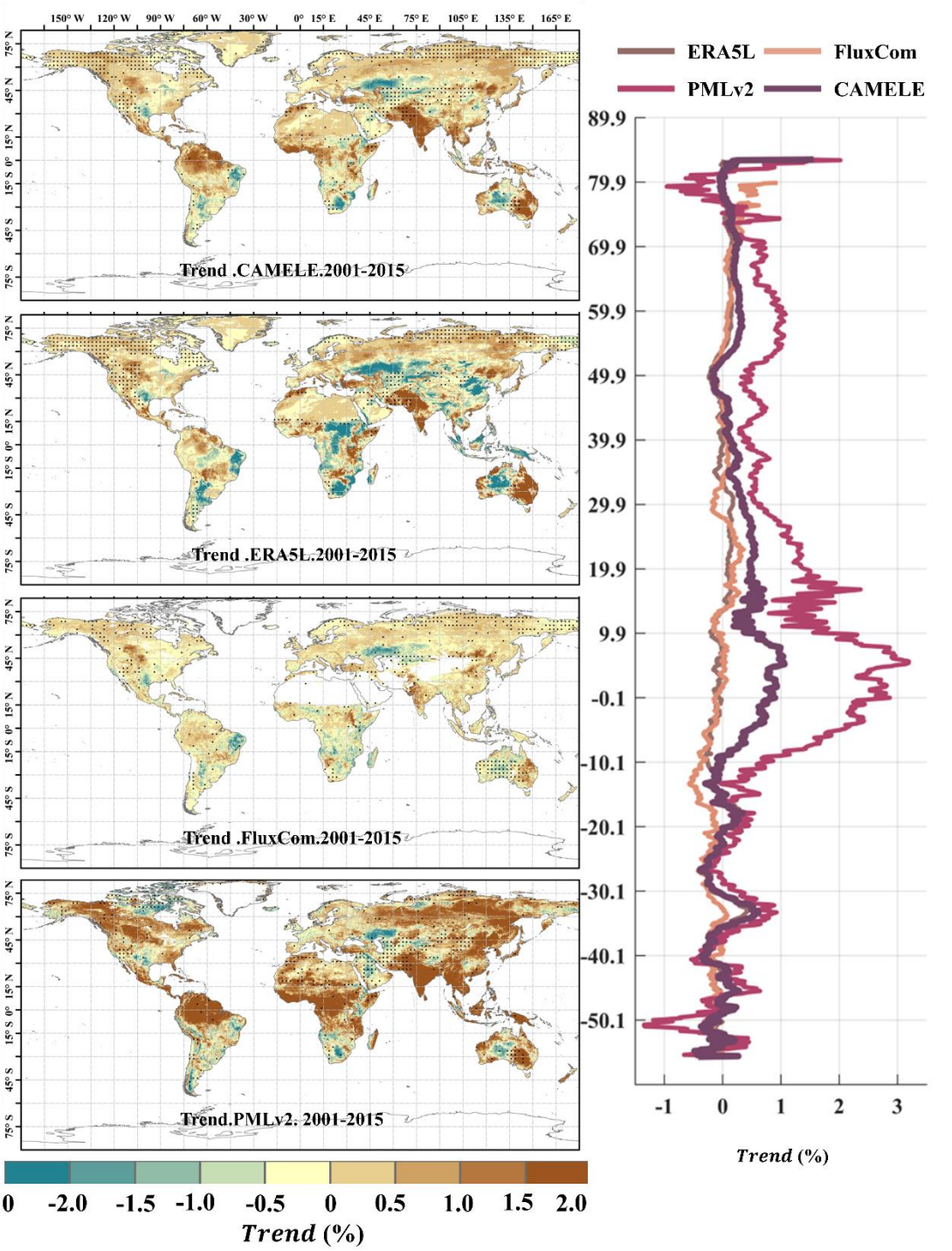


Figure 15 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted area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

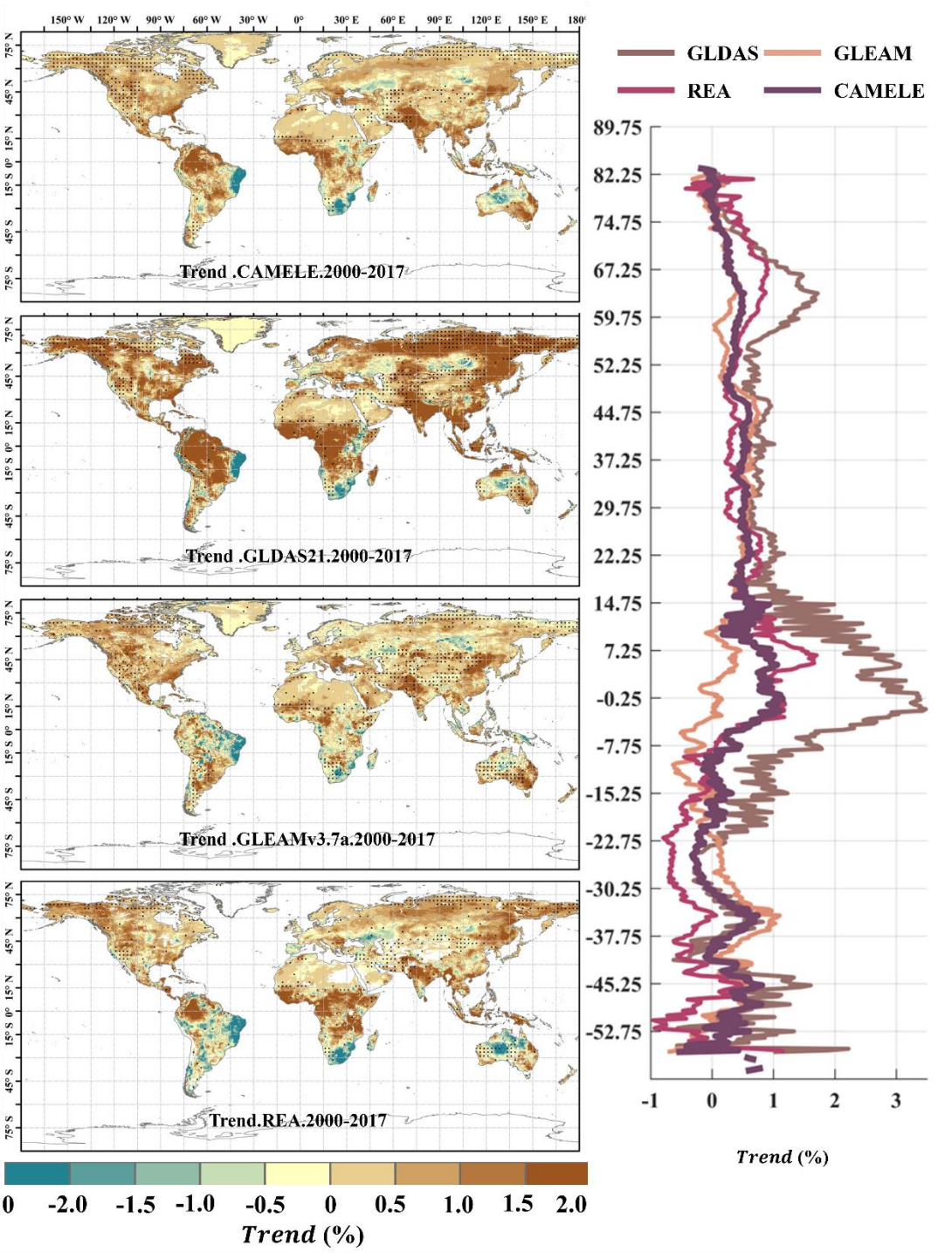


Figure 16 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding average trend with latitude. The trend is estimated with Theil–Sen's slope method, and the significance level is tested with the Mann–Kendall method. The dotted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

Previous Figures

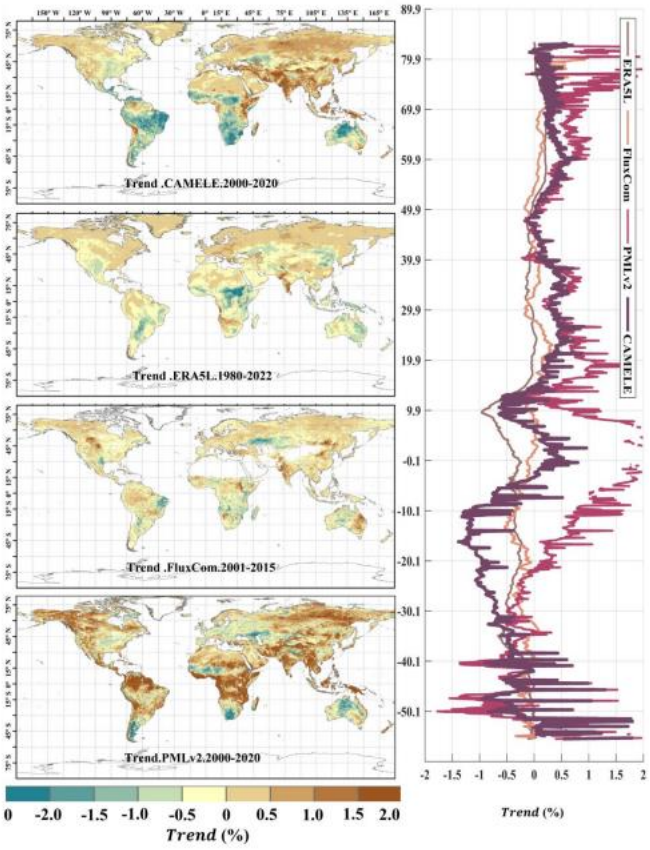


Figure 9 Global distribution of multi-year linear trend at 0.1° for CAMELE, ERA5L, FluxCom, and PMLv2, depicted alongside corresponding variation curves of average with latitude.

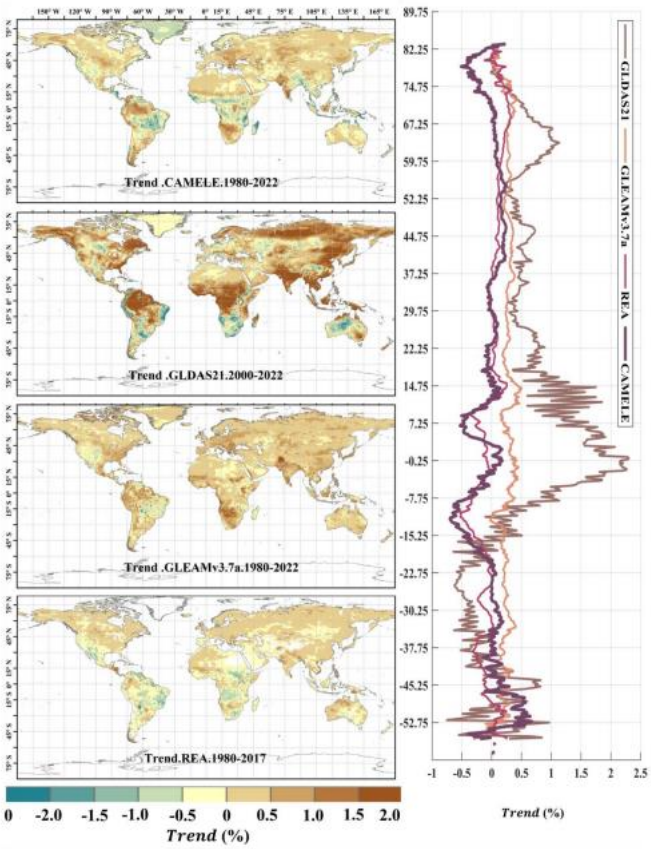


Figure 10 Global distribution of multi-year linear trend at 0.25° for CAMELE, GLDAS2.1, GLEAMv3.7a, and REA, depicted alongside corresponding variation curves of average with latitude.

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