Editor

Earth System Science Data

14 July 2026

Response to Reviewers' Comments for Paper #essd-2026-94

The authors would like to thank the editors and all reviewers for their insightful and constructive comments on our manuscript. We have meticulously addressed all minor revisions and suggestions raised during the review process. Furthermore, we took this opportunity to integrate massive global VIIRS data to extend our dataset to the near-present of 2025. While this process required a substantial computational effort—taking approximately one and a half months to complete—we are confident that this thoroughly updated version significantly enhances the dataset's value and fully satisfies the requirements of the reviewers.

(1) Extension to 2025: Integrated global VIIRS data to replace the degraded MODIS inputs, successfully extending the dataset to the near-present of 2025.

(2) Cross-product benchmarking: Added a trend comparison for global ET and GPP anomalies against multiple mainstream datasets.

(3) Refined visuals and language: Update some figures for better clarity and corrected descriptive terminology to maintain objective scientific phrasing.

(4) Point-to-Point Revisions: Addressed and incorporated every specific comment from the reviewers into the revised manuscript.

In the response letter, we provide point-to-point responses to the comments. We believe that the concerns of the reviewers have been sufficiently addressed. In the end, we hope this dataset will serve as a valuable and impactful tool for the research community. Please let us know if there are any questions or queries.

Thanks again to the editors and the reviewers for their valuable time, suggestions, and comments.

Yours truly

Yongqiang Zhang (on behalf of all authors)

Distinguished Professor in Hydrology and Water Resources

Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences

Email: zhangyq@igsnrr.ac.cn Tel: +86-1-64856515

Review comments in black color, responses in blue color. The cited paragraphs and sentences in italics.

Response to Reviewer # 1

The authors present a valuable update to the Penman–Monteith–Leuning (PML) model, introducing the PML-V2.2 dataset to provide a long-term (1982–2024), near-present global record of coupled terrestrial evapotranspiration (ET) and gross primary production (GPP). The manuscript is well-structured, clearly written, and the methodology is robust. The rigorous validation against both eddy covariance data and basin-scale water balance estimates provides high confidence in the dataset's fidelity. Overall, the development of this consolidated 43-year dataset bridging the AVHRR and MODIS epochs represents a significant contribution to the community. I have only a few minor suggestions to further enhance the transparency, utility, and readability of the manuscript before publication:

Response: We sincerely thank the reviewer for the thorough evaluation and the highly positive feedback on our updated PML dataset and manuscript. Your constructive suggestions are extremely valuable for enhancing the transparency, utility, and overall readability of our work. We will fully incorporate these suggestions into the revised manuscript. Below is our detailed, point-by-point response to your comments.

General Comments

1. The term "near present" is used prominently throughout the manuscript to describe the dataset's temporal coverage, but its exact definition can be slightly ambiguous due to varying data latencies. Specifying the exact cutoff month (e.g., "updated through December 2024") in the Abstract and the Data Availability section would provide clearer expectations for future users.

Response: We completely agree that clearly defining the temporal coverage is crucial for data users. As noted in Table 1, while "present" refers to data with short production latencies (days to a month), "near present" in this study specifically denotes datasets updated annually through December 2025 (e.g., Line 14, Line 85, Line 104) due to the inherent latency of observation-constrained meteorological and remote sensing forcing data.

Crucially, extending the PML-V2.2 dataset to this "near present" window (through 2025) was not a simple, routine update of the MODIS-based model. Since 2022, the aging MODIS satellite has suffered from significant orbital drift, which introduces severe artifacts into LAI products (Fig. A2). Using uncorrected MODIS inputs past 2022 leads to an artificial, drastic underestimation (extreme drops) in simulated global ET and GPP (especially 2025, see Fig. 12).

To resolve this widespread issue and ensure data fidelity, we deliberately integrated the VIIRS data stream (2012–2025) to complement and extend the record. This allows us to deliver a reliable, internally consistent "near present" record up to December 2025 (Fig. 2). As this is the major changes to the article to PML-V2.2a and V2.2c datasets, we have revised the Abstract, Introduction, Methods, Result and Data Availability sections to clarify this timeline and the scientific rationale behind the MODIS-to-VIIRS transition.

2. It would be highly beneficial to provide a comprehensive supplementary table detailing the basic attributes of all 208 flux tower sites used for calibration and validation. Furthermore, making the model performance metrics for individual sites, as well as the underlying data for key main figures, especially those illustrating cross-product uncertainties, publicly available in a repository would significantly enhance the reproducibility and usability of this study.

Response: Thank you for this valuable suggestion. We fully agree that providing detailed site attributes, site-level performance metrics, and the model code is essential for reproducibility and transparency.

In accordance with your feedback, we have uploaded the comprehensive attributes of all 208 flux tower sites (including their locations, vegetation types, and data periods), site-level model performance metrics, and the PML-V2.2 model code (R language) to the Zenodo repository (https://doi.org/10.5281/zenodo.18385275). Additionally, the underlying source data for the key figures, including those illustrating cross-product uncertainties, have been included in the same repository. We have updated the "Data availability" section of the manuscript to ensure that all these resources are clearly documented and accessible to the research community.

3. While the current manuscript effectively summarizes existing models in Table 1, the Discussion section could be strengthened by briefly comparing the PML-V2.2 global mean estimates (65.8×10^3 km^3 yr^{-1} for ET and 143.0 PgC yr^{-1} for GPP) and their long-term trends against a few other mainstream global products. Adding a few sentences to contextualize these results within the broader scientific consensus would help readers better understand where this new product stands.

Response: We sincerely appreciate this insightful suggestion. We completely agree that contextualizing our results within the broader literature is crucial for readers to understand the dataset's standing. In fact, conducting a comprehensive comparison against other mainstream global products was a major focus of our analysis, and we have dedicated significant portions of the Discussion section (Line 501-541) and Figure 13 to exactly this purpose.

Specifically, the manuscript already includes detailed comparative evaluations:

1. **Global mean estimates:**

- For ET: We compare our estimate (65.8×10^3 km^3 yr^{-1}) with recent comprehensive assessments (e.g., Zhang et al., 2026), explaining that when inland water body evaporation is accounted for, our adjusted global ET closely aligns with current consensus. Furthermore, basin-scale validations against 17 mainstream diagnostic products (Figs. A10 and A11) demonstrate that the PML models consistently rank in the top tier.

- For GPP: We provide a detailed contrast between our estimate (143.4 PgC yr^{-1}) and traditional optical/LUE models (which yield lower estimates around 115–131 PgC yr^{-1}). We clarify that our physically constrained output aligns more closely with contemporary estimates derived from novel proxies (e.g., OCS-based results of 157 ± 8.5 PgC yr^{-1}). This positioning is clearly illustrated in the multi-product intercomparison in Figure 13c.

2. Long-term trends and variability:

- Figure 13 is entirely dedicated to comparing the interannual variability and long-term trends of ET and GPP across the mainstream diagnostic products listed in Table 1.
- In the Discussion, we explicitly benchmark our conservative ET trends (e.g., 0.18 mm yr^{-2} for 1982–2016) against multi-product ensembles (e.g., Ma et al., 2021; Yang et al., 2023), noting that our estimates fall on the moderate/lower end of the spectrum.
- Conversely, we discuss the multi-product divergence in ET versus the highly coherent upward trend in GPP across nearly all datasets due to unified satellite-derived greenness inputs.

Based on your comment, we have carefully reviewed these existing paragraphs in the Discussion to ensure that these multi-product comparisons are prominent, clear, and effectively contextualize the PML-V2.2c dataset within the broader scientific consensus.

4. In Figure 12, the mean values between the "b" version and the "a/c" versions seem slightly different, while their interannual variabilities and trends are very similar. What might be the cause of this discrepancy? Is it due to differences in the mean LAI between GIMMS_LAI4g and the wWHD-smoothed MODIS LAI, or is it an artifact of the simulation at a coarser resolution? Please provide a brief discussion addressing this, ideally supported by the relevant data.

Response: Thank you for this careful observation. You are exactly right. Based on our analysis, this slight discrepancy in mean values is primarily driven by systematic differences in the mean magnitude between the GIMMS LAI4g and the smoothed MODIS LAI inputs (as well as GLASS albedo), rather than an artifact of coarser resolution.

To explicitly address this, we have added data-supported explanations in both the Results and Discussion sections of the revised manuscript, supported by Figure A8:

1. In the Results section, we clarified (Line 409-412): *"While slight regional variations exist due to remote-sensing forcing data inputs, the comparison of their mean climatology highlights that the PML-V2.2b product yields higher ET and GPP estimates across most non-tropical rainforest regions, primarily driven by the*

inherently higher LAI values in these areas (Fig. A8)."

2. In the Discussion section, we further contextualized this (Line 472-474): "*During the overlapping period, the historical AVHRR-based PML-V2.2b yields moderately higher global magnitudes (approximately +2.6% for ET and +3.1% for transpiration fraction) compared to the MODIS/VIIRS-based versions. This difference is driven by residual uncertainties in the consolidated historical inputs (specifically GIMMS LAI4g and GLASS albedo) (Cao et al., 2023), resulting in inherently higher LAI and GPP values across non-tropical regions (Fig. A8). Furthermore, the global and regional interannual trends show remarkable coherence (Fig. 12 and Fig. A7).*"

Furthermore, we have updated Figure 12 to separately display the mean values and the interannual trends of the different PML-V2.2 products. This visual separation more clearly demonstrates that despite the slight offset in absolute magnitudes, the consistency in their long-term trends is remarkably high. Compared to the drastic and unstable fluctuations seen in the initial PML-V2.0, this high trend consistency across the V2.2a/b/c versions represents a much more reasonable and robust improvement.

5. Although this is a global dataset, there is a previously officially released dataset called PML-V2 China (He et al., 2022), which has a daily resolution for 2000–2020. It would be valuable to briefly compare the new dataset with this regional one, focusing on regional differences and water balance closure. This would help answer whether the new global dataset is suitable for regional research in China, especially for recent years. (Reference: He, S., et al., 2022. Earth Syst. Sci. Data, 14, 5463–5488).

Response: We thank the reviewer for this constructive comment. While we agree that a regional perspective is valuable, we choose not to include a detailed quantitative comparison with the PML-V2 China dataset in this manuscript for the following simple reasons:

(i) Different purposes and drivers: The PML-V2 China version was customized for China using local flux sites and regional climate forcing. While this regional optimization means the China version might have slightly higher local accuracy, it is limited by its dependence on regional reanalysis data (CMFD), making regular annual updates difficult.

(ii) Long-term coverage: The primary goal of our new PML-V2.2 series is to provide a multi-decadal, global product that runs continuously from 1982 to 2025. In contrast, the regional China dataset ends in 2020.

(iii) Future user needs: Because of its long-term consistency and ability to cover recent years (2021–2025), we believe more researchers, including those working on China, will prefer the PML-V2.2 product for long-term trend analysis and near-present monitoring.

As noted, we have now introduced the China version briefly in the Introduction (Line 76-80). To keep the manuscript focused and concise, we prefer to stick to our core objective of evaluating global, multi-epoch consistency.

"Implemented on the Google Earth Engine (GEE) platform (Gorelick et al., 2017), this second generation leveraged 95 FLUXNET2015 sites of flux observations and MODIS data for parameterization to achieve a spatial resolution of 500 m at an 8-day interval, initially providing global coverage for 2003–2017 (Zhang et al., 2019), ***while an optimized regional daily version for China was subsequently released covering 2000–2020, integrating 26 local flux sites and local climate forcing (He et al., 2022).****"*

6. In Figure A5, the comparison of water-balance-based ET seems slightly unbalanced for products that do not provide data over deserts. Could you establish a mask or threshold to account for this? Additionally, there are cases where evaporation from water bodies is excluded. Please provide more methodological details on how these specific datasets were processed to ensure a fair comparison.

Response: This is an insightful comment regarding the potential bias introduced by varying spatial coverage over deserts and water bodies. To ensure a strictly fair comparison, we standardized the evaluation area across all participating datasets. Specifically, for products with missing values over barren regions or water bodies, we filled these spatial gaps using the continuous PML dataset for the corresponding year. This gap-filling strategy ensures that the basin-averaged ET for every product is calculated over the exact same spatial extent.

Interestingly, after applying this rigorous spatial standardization, we found that the recalculated water-balance-based ET values showed very little change compared to our original results. This minimal difference confirms that the original basin-wide ET estimates and our main conclusions were already quite robust. Nevertheless, this updated methodology provides a much fairer and more mathematically rigorous foundation for the comparison. We have expanded the methodology section to provide detailed descriptions of this gap-filling and standardization process, and we have updated Fig. A5 to Fig. A10, A11 as well as their captions accordingly:

"Figure A10. Comparison of simulated ET with water-balance-derived ET averaged for 152 large basins over 2003–2020. ***To ensure a fair comparison, minor spatial gaps over barren land and water bodies in some products were gap-filled using PML data.****"*

Specific comments

1. Abstract: The text mentions "a 8-day 500 m MODIS-based product." Based on standard pronunciation rules, the article should be changed to "an," making it "an 8-day."

   Done.

2. Abstract: Near the end of the abstract, it states "where rapid growth in GPP and water use efficiency is partially offset". Since the subject includes both GPP and water use

efficiency, the verb should ideally be plural. Consider changing it to "are partially offset."

Done.

3. The specific NOAA CO_2 dataset utilized in this study should be explicitly stated in the text to ensure strict methodological reproducibility. Since NOAA maintains multiple carbon cycle products, please clarify whether the globally averaged marine surface monthly mean data from the Global Monitoring Laboratory or data from a specific baseline observatory was used, and provide the exact citation or accession link.

We changed as suggested (Line 234): *"Additionally, the global monthly atmospheric CO2 concentration (C_a) data were sourced from* ***Mauna Loa Observatory*** *administered by the National Oceanic and Atmospheric Administration (NOAA)."*

4. Section 2.1: In the variable explanation for Equation 1, the British spelling "water vapour pressure" is used, whereas the American spelling "vapor pressure" is used elsewhere in the manuscript (e.g., when introducing VPD). I suggest unifying the spelling convention throughout the text.

Done. All is "vapor" now.

5. Section 2.1: In the paragraph below Equation 5, the phrase "under a same constraint... Therefore, a same function" contains a minor grammatical flaw. The definite article should be used here. Please change this to "under the same constraint... Therefore, the same function."

Done.

6. Sections 2.2 & 3.2: The manuscript uses the term "splitted" multiple times (e.g., "splitted parameterization scheme," and in the caption for Figure 6). The past tense and past participle of split is simply "split." Please replace all instances of "splitted" with "split" throughout the text.

Done.

7. Section 3.4: When listing statistical evaluation metrics, this section mentions "NSE = 0.904". However, the rest of the manuscript (such as the abstract and other metrics in the same paragraph) mostly retains two decimal places (e.g., NSE = 0.90, NSE = 0.79). I recommend standardizing the number of decimal places for data throughout the manuscript to maintain consistent formatting.

Done.

Response to Oscar M. Baez-Villanueva (referee #2)

The article titled “Extended global terrestrial evapotranspiration and gross primary production dataset from 1982 to near present” presents the new version of the PML dataset (PML-V2.2). The dataset comprises three complementary products: (i) PML-V2.2a, an 8-day 500 m MODIS-based dataset (2000–2024); (ii) PML-V2.2b, a half-monthly 0.01deg AVHRR-based product (1982–2020) optimised for near-present monitoring; and (iii) PML-V2.2c, a half-monthly 0.1deg record consolidating the a and b products (1982–2024). Parameter optimisation was performed through a leave-one-out cross-validation procedure for each plant functional type (PFT). The manuscript is very well written, clear, and aligned with the scope of the journal. Additionally, I believe that this dataset is of particular interest to the research community! I hope the following comments and suggestions help further strengthen the manuscript.

Response: We sincerely thank the reviewer for the thorough evaluation and the highly positive feedback on our updated PML dataset and manuscript. Your constructive suggestions are extremely valuable for enhancing the transparency, utility, and overall readability of our work. We have fully incorporated these suggestions into the revised manuscript.

We would like to note that the **updated dataset now spans the period from 1982 to 2025**, providing the most current data records available to the research community for monitoring recent hydro-climatic shifts. Furthermore, to ensure the robustness of our results, we have significantly expanded our validation analysis. We now provide **a rigorous benchmarking of the PML-V2.2 dataset against 17 mainstream diagnostic ET products**, which clearly illustrates the performance of our dataset regarding both global mean magnitudes and long-term trends.

We have addressed all specific comments and suggestions raised in your review, and we believe these revisions have substantially strengthened the manuscript. Below, we provide a detailed, point-by-point response to each of your comments.

General Comments

1. It would be interesting to mention in the abstract whether the product is expected to be updated regularly (e.g., annually). This information could further help the uptake of the dataset from new users.

The authors present PML-V2.2b as a product optimised for near-present monitoring. It would therefore be valuable to extend the dataset through 2025 (the last complete year) to further demonstrate this capability. In that case, the consolidated PML-V2.2c product could span 1982–2025. This would further strengthen the statement made in L66–67 regarding the production lag affecting many biophysically consistent datasets.

Response: We sincerely thank the reviewer for these excellent suggestions, which significantly enhance the utility of our dataset for near-present monitoring. We completely agree with both points and have addressed them as follows:

First, regarding the update schedule, we have explicitly mentioned the regular annual update cycle in the revised Abstract. Moving forward, by leveraging GEE-based VIIRS drivers, our annual update latency in subsequent years will ideally be reduced to within six months (e.g, Line 14, Line 551).

Second, following your recommendation, we have successfully extended the consolidated PML-V2.2c product to span the full 1982–2025 period (updated through December 2025). We would like to highlight that extending the dataset to 2025 was not a simple, routine update. Since 2022, the aging MODIS satellite has suffered from significant orbital drift, which introduces severe artifacts into LAI products (Fig. A2). Using uncorrected MODIS inputs past 2022 leads to an artificial, drastic underestimation (extreme drops) in simulated global ET and GPP for the year 2025 (see Fig. 12).

To resolve this widespread issue and ensure long-term data fidelity, we deliberately integrated the VIIRS data stream to complement and extend the record, successfully replacing the degrading MODIS inputs (Line 398-401, see following paragraph). This transition allows us to deliver a highly reliable, internally consistent record up to December 2025, perfectly aligning with your point regarding bridging the production lag. We have revised the Abstract, Introduction, Methods, Results, and Data Availability sections to explicitly clarify this 1982–2025 timeline, the annual update strategy, and the scientific rationale behind the vital MODIS-to-VIIRS transition.

*"To evaluate the consistency and reliability of the data integration process, we first assessed the transition between the recent satellite epochs. Supported by highly consistent relative changes in LAI and Albedo between 2020–2022 and 2013–2015 (**NSE = 0.80 and 0.79, Fig. A2 and Fig. A3**), the corresponding shifts in ET and GPP based on MODIS and VIIRS exhibit high spatial agreement (**NSE = 0.99 and 0.94, Fig. A6**). Consequently, the primary challenge for long-term consistency lies in bridging the AVHRR-based product (PML-V2.2b) with the MODIS-based version (PML-V2.2a)."*

2. It would be worth considering the use of the Kling–Gupta Efficiency (KGE) instead of the Nash–Sutcliffe Efficiency (NSE). The KGE decomposes performance into correlation, bias, and variability components, providing additional diagnostic insight and facilitating comparison across regions and climatic conditions.

Response: We appreciate the reviewer's suggestion. Both NSE and KGE are widely used benchmarks in current literature. The diagnostic components of KGE are actually already covered by our current four metrics: correlation is evaluated by R, systematic error by Bias, and variability and overall performance by RMSE and NSE.

To make it easier for readers to intuitively understand the model performance and to keep the figures visually clean, we did not add the KGE text to the existing plots. However, we agree with your point on its value for cross-comparison. Therefore, we have now provided the calculated KGE values directly in the text of the results section, specifically within the calibration part (**Line 337**). This approach provides the comprehensive diagnostic insight you suggested while maintaining a clear visual presentation.

3. Out of curiosity, is there a particular reason for using MSWEP V2.8 instead of the more recent MSWEP V3.16?

Response: Thank you for pointing out this. When we initiated the data preparation and long-term simulations, MSWEP V3 had not been officially released (it is currently still under review). Our preliminary investigations indicated slight differences between V2.8 and V3, and the long-term stability of V3 remained uncertain at the time. Therefore, V2.8 was selected to maintain strict data consistency across the 44-year simulation. We will consider comprehensively updating to V3 in future versions once its stability is fully evaluated by the community.

4. Could the authors elaborate on how rainy days were treated in the case of the data from theflux sites?

Response: We thank the reviewer for this question. In this study, we did not filter out rainy days from the flux site data because our model evaluation is conducted at an 8-day scale. At this aggregated level, the temporary noise and impact of precipitation events are naturally smoothed out.

Additionally, it is practically very difficult to isolate and remove rainy days at an 8-day step. This is especially true for sites in humid regions, where rainfall occurs so frequently that almost every 8-day window contains some rainy days. Completely removing these periods would cause severe data gaps and break the continuity of the time series. We have added this explanation to the methodology section (Line 159-160):

"Following a protocol consistent with FLUXNET's use of ERA5-Interim, these seamless daily fields were subsequently used to drive the PML model. Finally, all observations were aggregated into 8-day timescale, with a Quality Control (QC) flag assigned to represent the proportion of measured and high-quality gap-filled data. To maintain high data standards, only records with QC > 0.9 were retained for further calibration. ***Notably, rainy days were not filtered out, as precipitation noise is naturally smoothed at the 8-day scale. Moreover, removing them would severely disrupt data continuity, particularly in humid regions.****"*

5. Is there a sensitivity analysis associated with the optimised parameters presented in Table 2? It would be particularly interesting to assess whether parameter sensitivity varies across PFTs.

Response: This is an excellent suggestion. We completely agree that assessing parameter sensitivity across different PFTs provides valuable transparency and adds diagnostic insight to the model parameterization. In the revised manuscript, we have conducted a variance-based global sensitivity analysis using the Sobol method to comprehensively evaluate this. The resulting sensitivity indices for each parameter across different PFTs and flux components are now visualized in **Figure A1**. Furthermore, we have incorporated appropriate explanations regarding the sensitivity analysis procedure and the parameter behaviors in the methodology section, one paragraph after the parameter table 1 (Line 208-215):

"To provide greater transparency and evaluate parameter behavior across different PFTs, a variance-based global sensitivity analysis was conducted using the Sobol method. We calculated the Sobol sensitivity indices for all calibrated parameters to quantify their relative contributions to the variance of simulated ET, GPP, and their underlying components. As illustrated in Figure A1, the parameters β, η, and Vm strongly influence photosynthesis and GPP. Parameter m, which regulates the stomatal conductance process, dominates E_c and WUE, making it the most influential parameter for total ET. Additionally, soil evaporation (E_s) is primarily driven by kA, which controls canopy-soil energy partitioning, while interception evaporation (E_i) depends on the canopy interception capacity (Sl) and evaporation rate (F_0). This sensitivity analysis enhances the physical interpretability of the PML model and further justifies our PFT-specific parameterization strategy."

6. In L218, the authors mention that monthly CO2 concentrations from NOAA were used. How were these data temporally disaggregated to the 8-day and half-monthly scales?

Response: We will add a clarifying sentence in Section 2 to explain this processing. Specifically, the monthly NOAA CO2 concentrations were applied as static values for the 8-day or half-monthly steps within each respective month. Since terrestrial carbon-water coupling responds primarily to the long-term CO2 trend and seasonal cycles, the absence of high-frequency (sub-monthly) atmospheric CO2 fluctuations has a negligible impact on the simulations. This is now explained in Line 235:

"Additionally, the global monthly atmospheric CO2 concentration (C_a) data were sourced from Mauna Loa Observatory administered by the National Oceanic and Atmospheric Administration (NOAA). ***These monthly concentrations were applied as static values for the 8-day and half-monthly simulation steps within each corresponding month****."*

7. Were the three models optimised independently? If so, it would be interesting to summarise the best-performing parameters for each version. If not, a brief discussion on the expected impact of using different forcing datasets across the three products (particularly between versions a and b) would be useful. It appears that the optimisation was performed at the 8-day temporal scale and therefore primarily for version a; it would be good to state this explicitly in the manuscript.

Response: We thank the reviewer for this insightful question. To clarify, the models were not optimized independently. Versions a and b share the exact same parameterization scheme derived from our bottom-up framework, yielding the optimal parameter sets presented in Table 2. Regarding the temporal scale, we have updated the manuscript to explicitly state that the optimization was performed exclusively at the 8-day scale. As we now clarify in the text, this specific scale was chosen because the MODIS vegetation and albedo data employed in our framework provide an optimal balance between data quality and spatiotemporal resolution (Line 136-142):

"The ten parameters of the PML-V2.2 model were optimized and validated using a comprehensive bottom-up framework based on the latest global eddy covariance (EC) flux observations, yielding the optimal parameter sets presented in Table 2. Crucially, versions a and b share this exact parameterization scheme, and any product discrepancies arise solely from their respective forcing datasets. ***The parameter optimization was performed exclusively at an 8-day temporal scale, which was chosen because the MODIS vegetation and albedo data employed in our framework provide an optimal balance between data quality and spatiotemporal resolution. The optimization process utilized these MODIS products alongside the highest-quality site-level meteorological observations, mostly covering the period from 2000 to 2022."***

Since the parameterization is identical, the product discrepancies between versions a and b arise solely from their respective forcing datasets. To address your point regarding the expected impact of these different forcings, we have added a brief discussion in the manuscript outlining our mitigation strategy. Specifically, we intentionally selected bias-corrected forcing datasets (MSWEP and MSWX) rather than raw reanalysis data to minimize forcing-induced biases. Combined with the extensively calibrated GIMMS LAI4g dataset, this strategy effectively balances bottom-up site simulations with top-down global constraints, maintaining the high fidelity of our water balance validations across all versions.

8. I appreciate the effort made by the authors to carefully merge AVHRR and MODIS observations in order to minimise discontinuities and artificial biases in long-term trends. It would be interesting to add a few sentences about the pixel-scale bidirectional consolidation process and the reverse-scaling procedure. In addition, why was the 2001–2003 period selected for the consolidation and not a longer period?

Response: We will add more details describing the pixel-scale bidirectional consolidation process. A bidirectional correction was adopted because unidirectional scaling might fail to perfectly preserve both identical means and consistent long-term trends. Furthermore, the 2001–2003 period was selected because it represents a highly stable overlapping window with high-quality data from both AVHRR and MODIS sensors. Compared to using a longer baseline (e.g., 2001–2020), this shorter, focused window avoids introducing abrupt shifts or artificial artifacts at the transition boundary (1999–2001).

Importantly, to ensure the future continuity of our dataset as MODIS ages, we have applied this exact same bidirectional consolidation strategy to stitch MODIS and VIIRS records using the 2019–2021 overlapping window in version c. We acknowledge that any cross-sensor consolidation inherently introduces uncertainties. However, without such harmonization, continuous decadal studies across different satellite eras would be impossible. By transparently bridging the AVHRR-MODIS and MODIS-VIIRS gaps, version c provides a robust tool for placing recent anomalies within a continuous 40-year historical context. We have expanded the discussion on these consolidation stages and their associated uncertainties in the revised manuscript. Here are the new explanation of PML-V2.2c (Line 273-286):

"PML-V2.2c is the final consolidated dataset (half-month 0.1°) generated through a rigorous, pixel-scale bi-directional consolidation that bridges three satellite eras: AVHRR, MODIS, and VIIRS, providing a continuous record from 1982 to the present (Fig. 2). This consolidation was performed in two successive stages. First, to bridge the pre- and post-2000 eras, we treated the MODIS-based estimates (version a) as the primary benchmark and harmonized the historical AVHRR-based estimates (version b) using the 2001–2003 overlapping period. Second, to extend the dataset reliably into the post-MODIS era, a similar bidirectional consolidation was applied to stitch the MODIS and VIIRS records, utilizing their highly stable 2020–2022 overlapping window. As detailed in the description of version a, this specific period was carefully selected to avoid the post-2022 MODIS sensor drift, which renders subsequent years unsuitable for benchmark calibration. For most pixels, the primary variables from the subsidiary sensors (AVHRR or VIIRS) were scaled to match the mean values of the benchmark sensor to ensure statistical alignment. In extreme scenarios, such as desert or ice-covered regions or instances where the benchmark detected vegetation (>0) but the subsidiary sensor remained at zero (preventing standard scaling), a reverse scaling or masking to zero was applied to maintain absolute consistency. This harmonization effectively eliminates systematic jumps caused by sensor transitions and PFT classification differences, ensuring that the 44-year record preserves both accurate mean states and genuine long-term trends."

9. It would be very interesting to the readers to include performance metrics for all three dataset versions. Additionally, the consistency of the consolidation could be assessed at the half-monthly scale over the overlapping period across all products. Besides the visualisation of global patterns, it would be very interesting to add a figure showcasing time series of the three products in comparison with *in situ* data.

Response: We thank the reviewer for these constructive suggestions. Regarding the time-series comparison between version a and in situ data, we respectfully note that the observational periods across the 208 flux sites are highly inconsistent. If we were to aggregate all available site data into a single global time series, the number of active sites would fluctuate continuously over the years. Consequently, the temporal variations in such an aggregated plot would be heavily skewed by changes in data availability and site distribution, making it impossible to directly and accurately interpret the model's true long-term simulation capability. We believe the comprehensive statistical metrics (NSE, R, Bias, and RMSE) evaluated individually at each site provide a much more robust and objective assessment.

For versions b and c, performing a direct time-series comparison against flux towers is also scientifically problematic. The spatial mismatch between a tower's footprint (typically from hundreds of meters to 1 km) and the coarse grid resolution of these historical products (0.1°, ~10 km) over heterogeneous land surfaces introduces significant representativeness errors. This footprint mismatch is the primary reason we avoid direct site-level evaluations for coarse global products.

Finally, regarding the consistency of the consolidation during the overlapping period, our pixel-scale bidirectional scaling algorithm strictly matches the mean values of the historical variables to the benchmark. Therefore, the cross-product consistency at the half-monthly scale is mathematically constrained and naturally preserved by the harmonization procedure itself.

Specific Comments

1. L22: The authors mention that PML-V2.2c exhibits “exceptional reliability”. It would be useful to specify relative to which products or benchmarks this statement is made. A comparison against other products would be very interesting!

Response: We agree that the term exceptional is subjective. To maintain a more objective and rigorous tone, we have changed “exceptional” to “**high**” in the Abstract (L22) and throughout the revised manuscript.

To keep the Abstract concise, we focus its text on our direct validation metrics. However, a comprehensive multi-product benchmarking has been thoroughly detailed in the main text. Specifically, **Figure 13, Figure A10, and Figure A11** provide extensive evaluations against other mainstream global datasets, benchmarking our results across global mean estimates, long-term trends, and top-down water-balance closure metrics.

2. L50–57: Please add references for the datasets mentioned.

Response: We thank the reviewer for the careful reading and suggestion. Since these datasets are mentioned primarily as illustrative examples, and several of them are associated with multiple key references (as shown in Table 1), adding all the citations directly into the text would make the paragraph overly dense and disrupt the reading flow.

To address your comment while maintaining text readability, we have added the phrase "(**see Table 1 for references**)" in this paragraph. Table 1 has been designed to comprehensively list the specific references for all the dataset products mentioned here. We believe this is the most efficient way to provide the necessary citations.

3. L50: "Temporal span" or "record length" may fit better here than "temporal depth".

Done.

4. L89–90: Please include the versions and references of the datasets.

Done.

5. Table 1: For GLEAM4, in the "key feature" column, "evaporative stress" would be more appropriate than "water stress". In addition, the temporal coverage should be updated to 1980–2025.

Done.

6. L107: If I am not mistaken, bare soil evaporation should be denoted as *Es* instead of *Eis* to remain consistent with Eq. 2 and the following explanation.

Done.

7. L148: Perhaps "estimates" would be more appropriate than "observations" in this context.

Response: We thank the reviewer for pointing this out. We completely agree that "estimates" is a more accurate term for satellite-derived data. We also would like to distinguish between in-situ measurements and satellite retrievals. The revised sentence now reads: *"To ensure temporal continuity, missing meteorological forcing data were supplemented using Multi-Source Weighted-Ensemble Precipitation (MSWEP) V2.8 and Multi-Source Weather (MSWX) Past version (Beck et al., 2019, 2022), which integrate* ***multi-source in-situ observations, satellite-based estimates and rescaled*** *European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis V5 (ERA5) data."*

8. L382–385: It would be very interesting to compare all dataset versions over the overlapping period, including metrics such as mean annual global ET, trends, and the partitioning of evaporation into its three main components, as discussed in these lines.

Response: We thank the reviewer for this insightful suggestion. As this comment is closely related to our response to General Comment 4, we have comprehensively evaluated these metrics during the overlapping period.

To clarify, the PML-V2.2a and PML-V2.2c datasets are nearly identical during this period, as the consolidated record (c) directly integrates the 'a' version (Fig. 2). Therefore, the comparison primarily focuses on the AVHRR-based PML-V2.2b versus the MODIS/VIIRS-based 'a/c' versions. Our analysis confirms that while the uncorrected GIMMS LAI inputs cause a moderate difference in absolute magnitudes across non-tropical regions, both the interannual trends and the evaporation partitioning (e.g., transpiration fraction) remain remarkably consistent across versions. To explicitly address this, we have added the following detailed discussion to the revised manuscript (Line 470-479):

*"The temporal consistency of the PML-V2.2 product from 1982 to 2025 (**Fig. 12**) demonstrates its robustness for long-term climate studies. This stability is primarily attributed to the use of unified meteorological forcing and parameter sets across different satellite eras.* ***During the overlapping period, the historical AVHRR-based PML-V2.2b yields moderately higher global magnitudes (approximately +2.5% for ET and +3.2% for transpiration fraction) compared to the MODIS/VIIRS-based versions. This difference is driven by residual uncertainties in the consolidated historical inputs (specifically GIMMS LAI4g and GLASS albedo; Cao et al., 2023), resulting in inherently higher LAI and GPP values across non-tropical regions (Fig. A8).*** *Furthermore, the global and regional interannual trends show remarkable coherence (Fig. 12 & Fig. A7).* ***To maintain this physical consistency for long-term trend analysis, we deliberately applied the unified parameter set across all epochs (Fig. 2), avoiding the footprint scale-mismatch issues associated with separately calibrating the 0.1° AVHRR model against site-level flux towers."***