Manuscript: ESSD-2025-376

Comment by Anonymous Referee #2

My review comments are structured as follows: **Overall Assessment**, **Major Strengths**, and **Recommendations for Improvement**.

I. Overall Assessment

This paper presents a systematic evaluation of the water balance consistency of 47 state-of-the-art hydrological datasets (precipitation, evapotranspiration, runoff, and soil moisture) using 8,294 independent combinations. The methodology is rigorous, the data coverage is extensive, and the study holds significant scientific and practical value. It reveals a widespread lack of water balance consistency in current global hydrological datasets and provides an in-depth analysis of the spatial patterns, influencing factors, and temporal trends. The manuscript is well-structured, the methods are transparent, and the results are credible. I recommend **acceptance after minor revisions**.

II. Major Strengths

1.High Novelty: This is the first study to systematically assess the consistency of multi-source, multi-variable hydrological datasets from a water balance perspective, filling a critical gap in the current literature.

2.Methodological Rigor:

a.The use of independent dataset combinations effectively avoids spurious consistency arising from the use of the same model or forcing data.

b.The use of adjusted R^2 as the consistency metric mitigates errors introduced by unit inconsistencies between variables.

c.The application of SHAP for factor attribution enhances the interpretability of the results.

3.Comprehensive Data Coverage: The inclusion of gauge-based, satellite-based, and reanalysis products ensures broad spatiotemporal coverage and strong representativeness.

4.Insightful and Actionable Results:

a.Clearly identifies the strengths and weaknesses of different data sources across various regions and climatic conditions.

b.Highlights the significant impact of soil moisture data depth on consistency.

c.Reveals an improvement in dataset consistency in mid-to-high latitude regions of the Northern Hemisphere in recent decades.

Response: We would like to express our gratitude to your encouraging evaluation, and for the time and effort you devoted to reviewing our work.

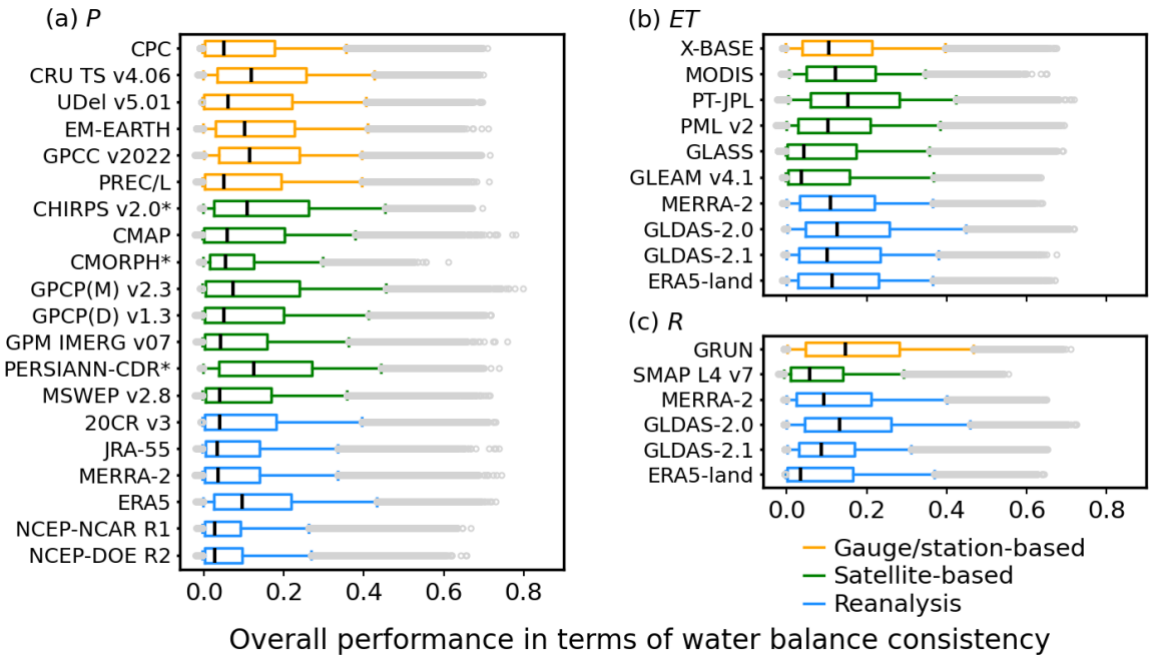
III. Recommendations for Improvement

1. Clarifications in the Methods Section

Handling Soil Moisture Depth Differences: While the manuscript states that ΔSM represents "change," the response of soil moisture at different depths to P-ET-R varies. It would be beneficial to clarify if any normalization or sensitivity analysis was performed for ΔSM across different depths.

Response: We thank the reviewer for raising this point. We did not apply normalization or sensitivity analysis for Δ*SM* across different depths. However, we found the *SM* datasets with simulations of deep soil layers generally performed better in most global regions. Please find the relevant text in lines 362−364 and lines 470−473. Additionally, using terrestrial water storage changes from GRACE, which integrates deeper soil moisture and groundwater availability, is not beneficial for improving water balance consistency (new Fig. S3).

> In lines 362−364:
>
> "*This is because they only represent the surface layers instead of the entire soil column (Fig. 1e). Meanwhile, the SM datasets with simulations of deep soil layers generally performed better in most global regions, such as the reanalysis and GLDAS-2 products (Fig. 2d and Fig. S16d).*"
>
> In lines 470−473:
>
> "*For SM, reanalysis datasets perform best, likely because they are constrained by physical laws and consider deeper soil moisture variability (Table S4 and Fig. S16). In contrast, low penetration depths (~2–5 cm) of microwave sensors limit the ability of ESA CCI v08.1 to capture deeper-layer SM variations (Hirschi et al., 2025).*"



*"**Fig. S3.** Performance of the considered datasets based on R^2 scores measuring water balance consistency through $P-ET-R=\Delta TWS$. Colors indicate the type of each dataset. Each box shows the median value, as well as the 5th, 25th, 75th, and 95th percentiles of the global pattern of water balance consistency derived from monthly data. Asterisks (*) following the name of P dataset indicate its limited spatial coverage of 50°S–50°N or 60°S–60°N."*

Temporal Scale Analysis: The significant differences in consistency between daily and annual scales warrant further discussion of the underlying physical mechanisms (e.g., high noise at daily scales, strong smoothing effects at annual scales).

Response: Thank you for your suggestions. We have clarified accordingly in lines 477–483.

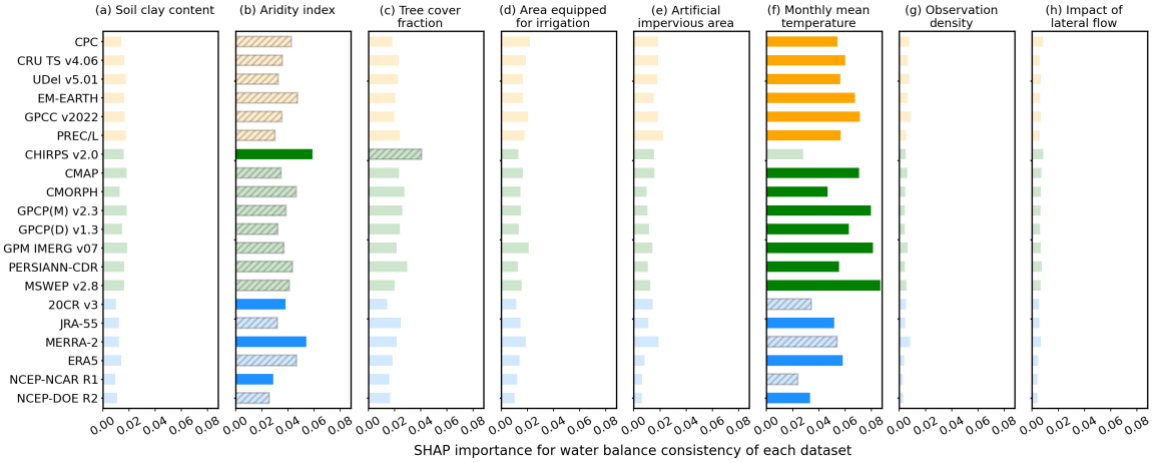
"Dataset performance varied significantly across time scales, with the highest correspondence at the monthly scale, where seasonal variability is well-captured and synoptic weather variability is mitigated. This explains the markedly lower water balance consistency observed at the annual scale for all datasets, where seasonal signals are strongly smoothed. At a daily time scale, the variability of the involved variables is high, including more extreme values and high noise, and apparently under-constrained by available observations (Maurer and Hidalgo, 2008; Fisher et al., 2008)."

2. Deepening the Results and Discussion

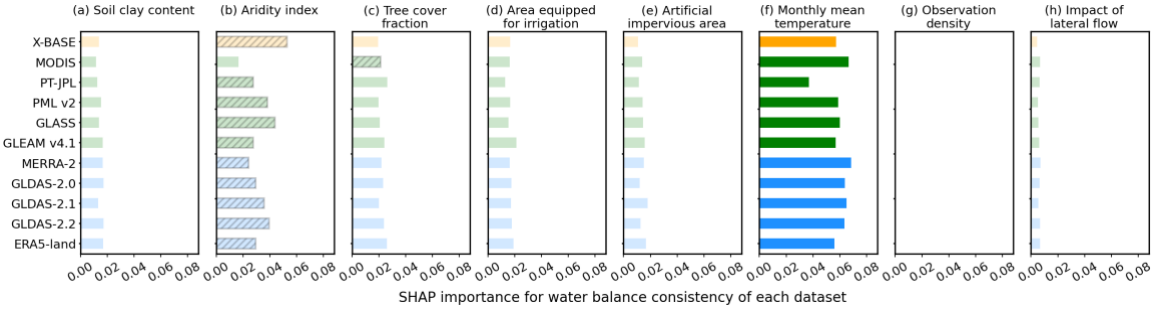
Root Causes of Low Consistency: Beyond the mentioned observational errors and model structures, could factors like surface-groundwater exchange or human activities (e.g., irrigation, reservoir regulation) also contribute? Expanding the discussion on this point would be valuable.

Response: In this revision, we also considered the potential influence of urbanization and lateral flow, which we found to have relatively low importance for dataset performance in terms of water balance consistency (see updated Figs. S17−S20). Please find the modified text in lines 398−399.

"At the same time, factors like irrigation, urbanization, and lateral flow play relatively minor roles (Figs. S17−S20)."

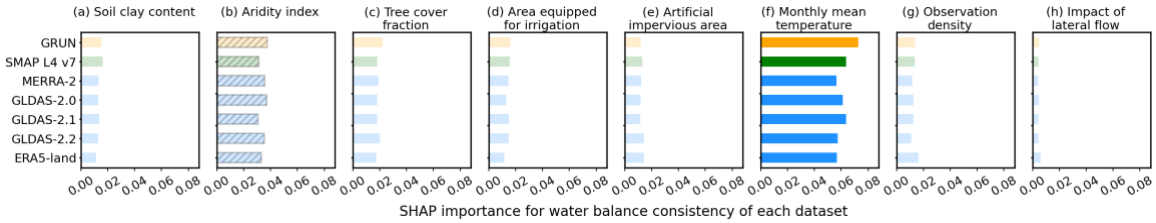


*"**Fig. S17.** Importance of (a) soil clay content, (b) aridity index, (c) tree cover fraction, (d) area equipped for irrigation, (e) artificial impervious area, (f) monthly mean temperature, (g) observation density, and (h) impact of lateral flow to water balance consistency of each P dataset. The importance is quantified by global averaged absolute SHAP values (Methods). Bars with dark color and hatch, respectively, indicate the first and second important factors for the water balance consistency of each P dataset."*

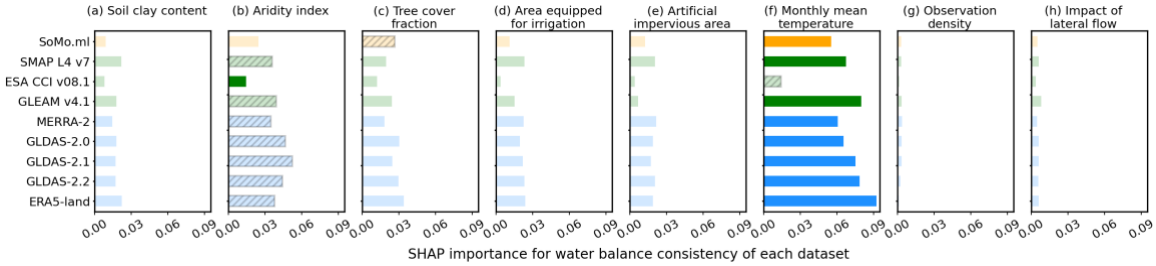


*"**Fig. S18.** Importance of (a) soil clay content, (b) aridity index, (c) tree cover fraction, (d) area equipped for irrigation, (e) artificial impervious area, (f) monthly mean temperature, (g) observation density, and (h) impact of lateral flow to water balance consistency of each ET dataset. The importance is quantified by global averaged absolute SHAP values*

(Methods). Bars with dark color and hatch, respectively, indicate the first and second important factors for the water balance consistency of each ET dataset."



*"**Fig. S19.** Importance of (a) soil clay content, (b) aridity index, (c) tree cover fraction, (d) area equipped for irrigation, (e) artificial impervious area, (f) monthly mean temperature, (g) observation density, and (h) impact of lateral flow to water balance consistency of each R dataset. The importance is quantified by global averaged absolute SHAP values (Methods). Bars with dark color and hatch, respectively, indicate the first and second important factors for the water balance consistency of each R dataset."*



*"**Fig. S20.** Importance of (a) soil clay content, (b) aridity index, (c) tree cover fraction, (d) area equipped for irrigation, (e) artificial impervious area, (f) monthly mean temperature, (g) observation density, and (h) impact of lateral flow to water balance consistency of each SM dataset. The importance is quantified by global averaged absolute SHAP values (Methods). Bars with dark color and hatch, respectively, indicate the first and second important factors for the water balance consistency of each SM dataset."*

Mechanisms Behind Spatial Consistency Patterns: For instance, is the low consistency in high-latitude regions linked to insufficient representation of processes like snowpack and permafrost? Further interpretation in the context of existing literature is recommended.

Response: We have added accordingly:

"Limitations in representing snowpack and permafrost processes, along with difficulties in satellite retrievals over snow- and ice-covered high-latitude regions, also contribute to this issue (Hirschi et al., 2025; Muñoz-Sabater et al., 2021)."

3. Figures and Presentation

Figure 1: The meaning of the asterisk * and dashed lines in the boxplots should be explicitly stated in the figure caption.

Response: We have stated in the figure caption at the end, as follows:

*"Median results for performing the analysis with daily and annual data are indicated through crosses (×) and pluses (+), respectively (Text S1−S2). Asterisks (*) following the name of P dataset indicate its limited spatial coverage omitting high-latitude regions with typically low performance, and dashed line in each box indicates median of only 50°S−50°N. * of SM dataset indicates that the dataset does not consider the entire soil column."*

Figure 2: The grey areas, indicating "multiple datasets show similar performance or low consistency," would benefit from having the specific thresholds for "similar" and "low" defined in the caption or figure.

Response: We specify the thresholds in the caption of Figure 2:

Gray color indicates that multiple datasets show similar water balance consistency (with R^2 scores varying by less than 5%) or low water balance consistency (with all R^2 scores below 0.2).

Supplementary Material: Briefly mentioning the names of the best/worst performing datasets from Figures S13–S28 in the main text would help readers quickly grasp key findings.

Response: In section 3.1, we mention the best-performing *P* datasets in lines 347−348, the best-performing *ET* datasets in lines 355−356, the best-performing *R* datasets in lines 359−360, and the worst (best) performing *SM* datasets in line 361 (line 364).

4. Language and Formatting

Some sentences are quite long; breaking them up would improve readability.

Terminology should be checked for consistency (e.g., unified use of "gauge-based" vs. "station-based").

Response: We have broken the long sentences accordingly. However, we continue to use the terms gauge-based and station-based together because the in situ measurements differ by variable: *P* and *R* are derived from rain gauges and river gauges, respectively, whereas *ET* is measured at flux stations.

IV. Recommendation

Recommendation: Minor Revision

This manuscript makes a pioneering contribution to the evaluation of hydrological datasets. It is scientifically sound, its conclusions are robust, and it provides crucial insights for hydrological model development, data fusion, and climate change research. I recommend acceptance after the authors address the points above.

Response: Thank you for your encouraging evaluation.