Manuscript: ESSD-2025-376

Comment by Anonymous Referee #1

Huang et al. contribute to understanding the limitations of hydrological datasets (ground-, satellite-, and reanalysis-based) in capturing the relationship between monthly variations in soil moisture (SM) and the difference between precipitation (P), evapotranspiration (ET), and runoff (R) at a pixel scale around the world. Additionally, the manuscript's results contribute to identifying the most suitable datasets for different geographical and ecological regions, which is important for reducing uncertainty in ecological, climatological, and hydrological studies using the evaluated datasets.

Response: Thank you for your encouraging evaluation.

Overall, I found the paper well written and organized, and suitable for publication in the ESSD journal, but I have some comments that should be addressed before publication consideration. Particularly, some work is required to improve the clarity of the methods and results sections: (i) Explain how lateral flows and water table depth may potentially bias the proposed water balance at pixel scale, leading to the low water balance consistency reported in the manuscript; (ii) provide a clearer explanation of the linear relationship between SM, P, ET, and R (P – ET - R)s = k ΔSMs) at the monthly scale, including the potential limitations of assuming a linear relationship.

Response: Thank you for your suggestions. In this revision, we have (1) quantified the potential impact of lateral flows from rivers and groundwater, and (2) clarified the rationale for using a linear regression model based on our water balance assumption. Additionally, we include supplementary results on water balance consistency using terrestrial water storage from GRACE, extend the introduction on the discrepancy among *ET*, *R*, and *SM* datasets, and quantify the influence of urbanization by incorporating the global artificial impervious area. Detailed responses are provided below.

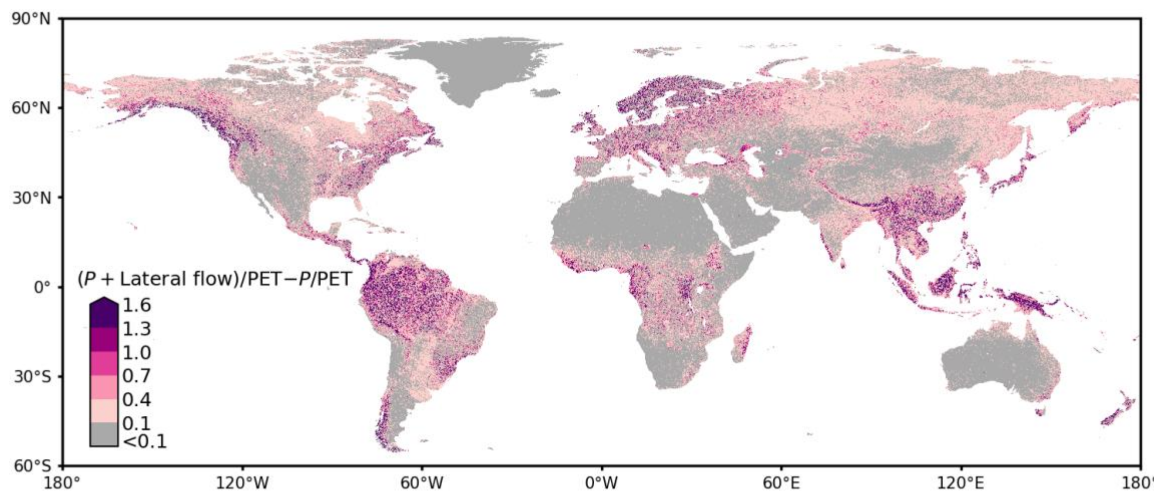
Major comments:

Line 191: The proposed water balance equation does not include some fluxes that may strongly affect hydrological dynamics at the pixel scale and may contribute to the low water balance consistency reported in the manuscript. For example, lateral fluxes (both inputs and outputs) can significantly influence variations in soil moisture (SM) and runoff (R) at the pixel scale, particularly in low-elevation areas and along river channels (e.g., Fan et al., 2013; Miguez-Macho and Fan, 2025; Nobre et al., 2011). Similarly, SM dynamics are strongly influenced by water table depth (WTD). Therefore, the authors should explain how excluding lateral flows and

WTD could bias the results. In this regard, I also suggest examining whether and how the runoff datasets capture lateral flows and groundwater dynamics at the pixel scale.

Response: We thank the reviewer for raising this interesting aspect. To quantify the potential influence of lateral flows from rivers and groundwater on regional water balance, we include published data from Miguez-Macho & Fan (2025) as one of the predictors in the attribution models. The data provide two indices, including *P*/PET and (*P* + lateral flow)/PET where the lateral flow is the total subsidies by rivers and groundwater, and the groundwater flow is determined by water table dynamics. Therefore, the difference in (*P* + lateral flow)/PET and *P*/PET for each grid cell was calculated to indicate the influence of lateral flows on regional water balance.

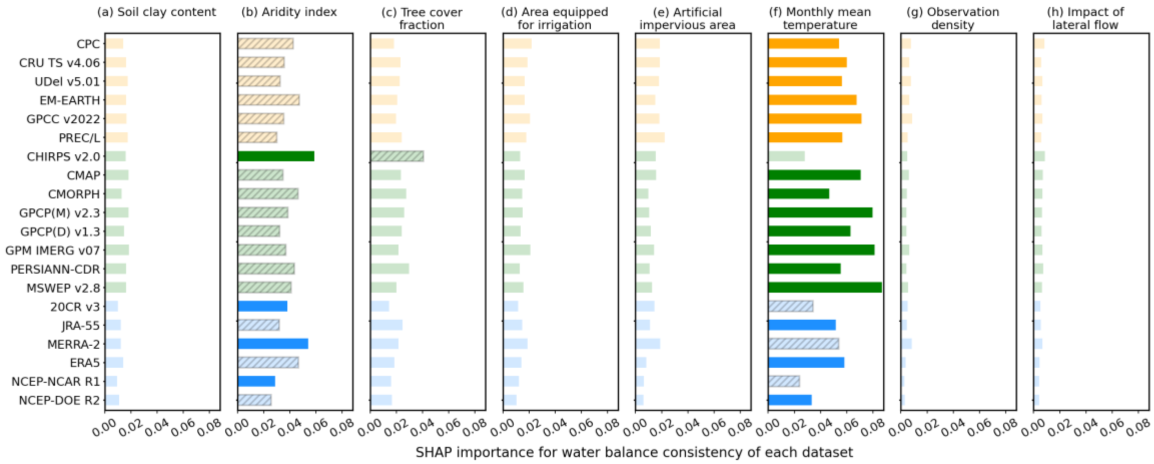
In this revision, we show the resampled 0.25-degree map of lateral flow impact (i.e., (*P* + lateral flow)/PET - *P*/PET) in the new Fig. S9. By considering it to be a predictor in our explainable machine learning method (see lines 286−289), we further quantify the relative role of lateral flow impact on the performance of each dataset. Since the lateral flow can be directly regulated by topography, the topography factor is not considered in this revision. The updated results across global grid cells indicate that lateral flow plays a relatively minor role in the performance of the considered datasets in terms of water balance consistency (Figs. S17−S20).



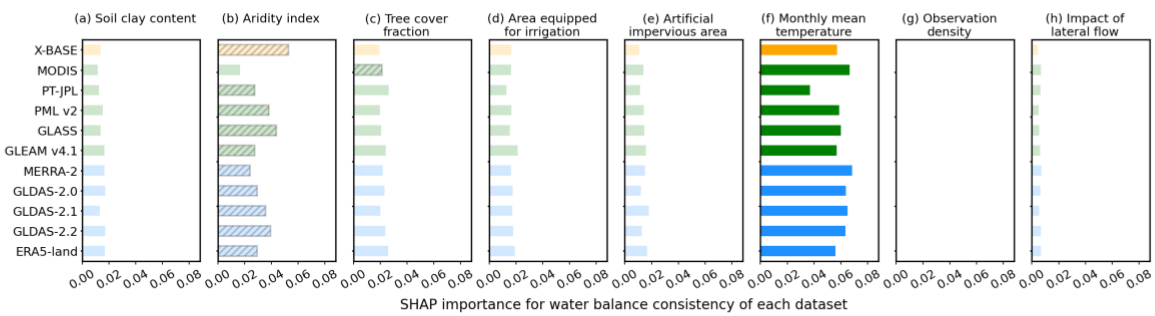
*"**Fig. S9.** Maps showing the potential impact of lateral flow from rivers and groundwater on regional water cycles. The impact is quantified by using the published indices from Miguez-Macho & Fan (2025), including P/PET and (P + lateral flow)/PET where the lateral flow is the total subsidies by rivers and groundwater, and the PET is potential evapotranspiration."*

In lines 286−289:

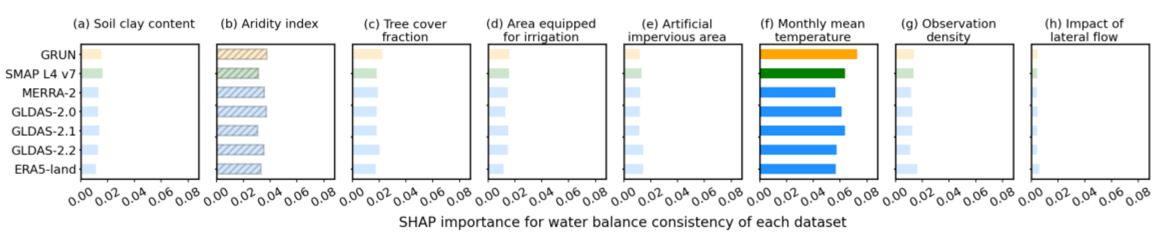
"The global impact of lateral flow has been evaluated by Miguez-Macho and Fan (2025), where the differences of (P + lateral flow)/PET and P/PET (with PET as the potential evapotranspiration) represent the influence of subsidies by rivers and groundwater on regional water cycles (Fig. S9)."



*"**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."*

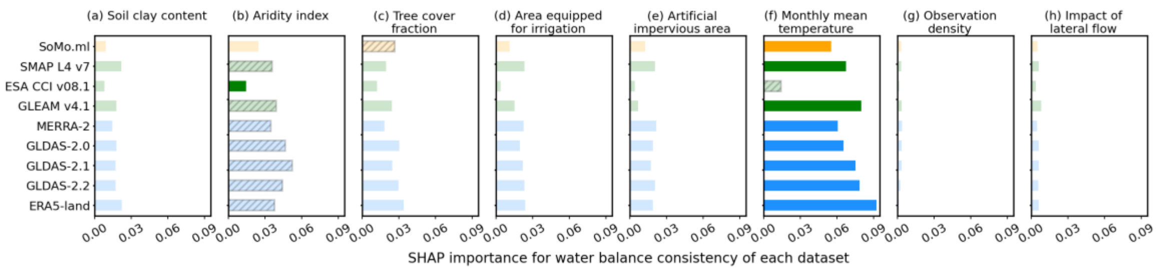


*"**Fig. S18.** Importance of (a) soil clay content, (b) aridity index, (c) tree cover fraction, (d) area equipped for irrigation, (e) artificial impervious area, (f) monthly mean temperature, (g) observation density, and (h) impact of lateral flow to water balance consistency of each ET dataset. The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

Line 191: The linear regression between SM and P–ET–R may also introduce bias into your results. Because your analysis is performed at a monthly scale, the hydrological response of each water balance component may occur at different rates due to, e.g., seasonality (dry vs wet season or summer vs winter) or soil saturation. Therefore, I encourage the authors to provide a more detailed explanation of why the linear assumption is appropriate for the analysis, as well as its limitations.

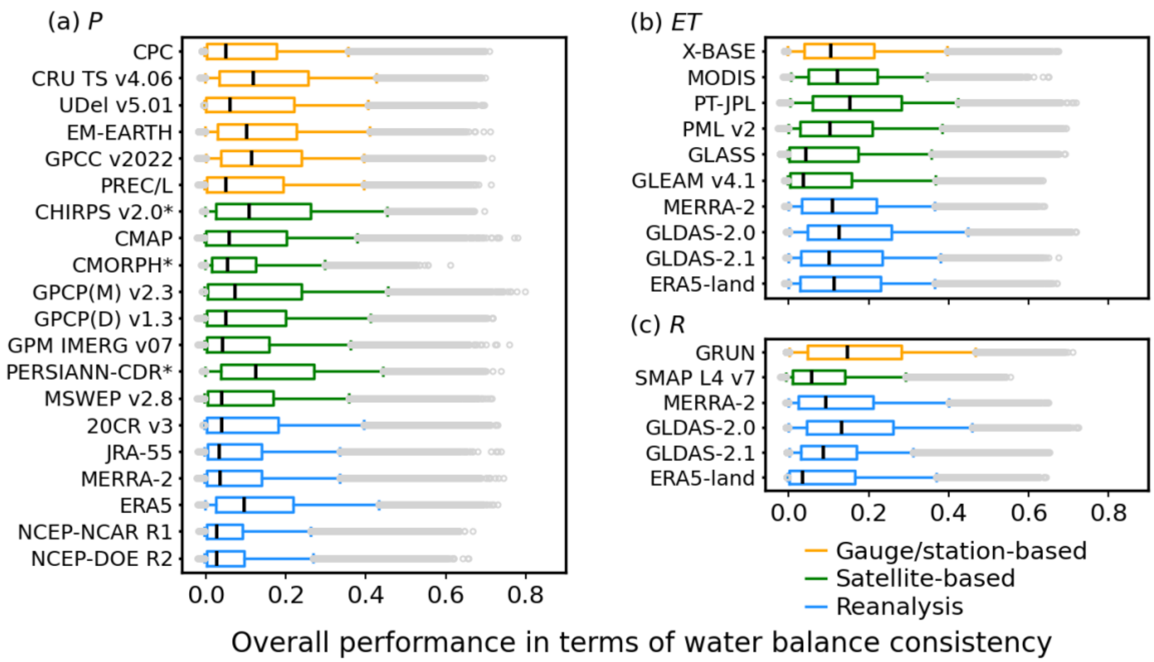
Response: We appreciate the reviewer's insight. Linear regression is appropriate under the water balance assumption of this study, where changes in soil water content are driven by accumulated precipitation, evapotranspiration, and runoff, expressed as $P–ET–R=\Delta SM$. Unconsidered water processes may influence our water balance assumption, leading to a nonlinear response of ΔSM to $P–ET–R$. To test the potential influence and bias, we 1) quantify the relative importance of later flow impact in the attribution models, and 2) evaluate whether the use of terrestrial water storage change (ΔTWS) instead of ΔSM can benefit higher R^2, since ΔTWS, in theory, integrates the changes of glacier, snow, and surface water storage.

First, the impact of later flow on the dataset performance is relatively low (updated Figs. S17−S20), supporting our water balance assumption. Second, using ΔTWS from the GRACE and its Follow-On mission (GRACE-FO) to form the water balance as $P–ET–R=\Delta TWS$ yields similar ranking results as using ΔSM (added Fig. S3). This also supports using *SM* as the water balance assumption is sufficient for our study purposes, and therefore the linear regression on top of it is appropriate. In this revision, we clarify the motivation for using linear regression in lines 203−205:

"Under our water balance assumption, we build a linear regression model in each grid cell of each considered combination of hydrological datasets, considering all available months, and assess its adjusted R^2 score:",

and introduce the use of GRACE data in lines 242−248 as well as new Text S3.

"In addition, unconsidered water variables, like glacier, snow, and surface water storage, might introduce bias into our water balance assumption, leading to a nonlinear response of ΔSM to P–ET–R. We thereby used terrestrial water storage from GRACE instead of SM in equation (1) to evaluate the performance of the P, ET, and R datasets, based on their combinations with GRACE data (Text S3). In this case, the number of combinations is decreased by one order of magnitude (933 remained), but ranking results are similar to using ΔSM (Fig. S3)."



*"**Fig. S3.** Performance of the considered datasets based on R^2 scores measuring water balance consistency through P−ET−R=ΔTWS. Colors indicate the type of each dataset. Each box shows the median value, as well as the 5^{th}, 25^{th}, 75^{th}, and 95^{th} 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 50ºS−50ºN or 60ºS−60ºN."*

"***Text S3. Performance calculations with the use of terrestrial water storage from GRACE***

In this case, the terrestrial water storage (TWS) at 0.25 degree resolution from GRACE and its Follow-On mission (GRACE-FO) is provided by the Center for Space Research mascon product (Save et al., 2016). We calculated the change in TWS (ΔTWS) as the difference between the TWS anomaly of a given month and that of the previous month. Then, ΔTWS was used with P, ET, and R datasets to form combinations. Besides the exclusion rules *detailed in Methods, we further consider the combinations with water balance components from GLDAS-2.2 to be not considered. For each of the remaining 933 independent combinations, we build a linear regression model in each grid cell:*

$$(\boldsymbol{P} - \boldsymbol{ET} - \boldsymbol{R})_s = k \cdot \boldsymbol{\Delta TWS}_s \quad (S1)$$

where s is the spatial index (grid cell) and k is the proportionality factor. Similar to the processing steps in Methods, the adjusted R^2 score of each linear model was calculated for each independent combination with ΔTWS. Finally, the overall performance for each P, ET, or R dataset in each grid cell was obtained by averaging R^2 across all combinations of datasets containing the respective dataset."

Line 205: The coefficient of determination (R2) of the linear regression model quantifies how well P-ET-R explains the variability of SM. However, you can include a bias metric (e.g., mean water balance error = i=1m(P-ET-R-SM)) to further examine the consistency of hydrological datasets.

Response: We appreciate the suggestion of using the mean water balance error to further examine water balance consistency. However, the units of soil moisture are not consistent across datasets, where the volumetric content is not easily converted to mm/day which is the unit of the other considered variables. Therefore, we will still focus on using the linear regression model.

Minor comments:

Lines 47 – 68: You should provide further information about the general advantages and disadvantages of ground-based, satellite, and reanalysis datasets to characterize ET, runoff, and soil moisture as you did for precipitation.

Response: We extend the introduction accordingly in lines 66−81.

"With the developing observation networks and data synthesis (Dorigo et al., 2011; Pastorello et al., 2020; Do et al., 2018), machine-learning algorithms present an alternative opportunity instead of interpolation to produce seamless observation-based datasets globally for evapotranspiration (ET), runoff (R), and soil moisture (SM) datasets (Nelson et al., 2024; Ghiggi et al., 2019; O and Orth, 2021). Although Penman-Monteith and the simpler Priestley-Taylor models are still the key physical algorithms to estimate ET through remote sensing, the relevant products tend to leverage recent advances in satellite data and climate reanalysis (Fisher et al., 2008; Miralles et al., 2025; Zhang et al., 2019). Differently, satellite-based SM datasets follow different technical roadmaps, such as merging retrievals from various sensors (Gruber et al., 2019) or assimilating radiometer observations into land surface modeling (Reichle et al., 2019). In this way, the latter additionally provides an SM-constrained R dataset (Reichle et al., 2019). At the same time, there are updated parametrizations for the land surface model in reanalysis to better describe the soil water balance and hydrological cycle (Hirschi et al., 2025; Muñoz-

Sabater et al., 2021). It has been documented that those technical discrepancies could cause datasets' performance in terms of agreement with observations, while the influence of environmental factors remains unclear (Markonis et al., 2024; Tang et al., 2024)."

Lines 72, 237 and 253: I encourage authors to use another expression instead of the term "water variables" to avoid confusion.

Response: We use water balance components instead of water variables.

Line 89: Please clarify that R2 corresponds to the coefficient of determination.

Response: We clarify accordingly.

"For each combination, we evaluate adjusted R^2 as the performance of linear regression of temporal changes in $P-ET-R$ against changes in SM (ΔSM) to determine its water balance consistency since R^2 corresponds to the coefficient of determination."

Lines 128 – 135: Soil moisture estimates were obtained from different depth profiles (< 2 cm, 0-50 cm, 0 – 100 cm, and > 100 cm). How well correlated are the variations in SM among these depth profiles? Do you consider extracting total water storage from GRACE https://grace.jpl.nasa.gov/mission/grace/?

Response: We did not correlate the variations in *SM* among these profiles because this is beyond the scope of our study. Instead, we highlight that the *SM* datasets with different depths have distinct performance in terms of water balance consistency, because the *SM* variations below 50cm in many regions are relevant to *R* and *ET* for water balance consistency. In this revision, we also consider adding a supplement of using water balance consistency with ΔTWS from GRACE, to investigate whether the water variations below 2m can benefit water balance consistency (new Fig. S3). However, we find that using GRACE data results in a lower R^2 than using *SM* datasets. It is likely because GRACE data is originally provided at 3-degree resolution, and products at finer resolutions rely on the downscaling models.

Line 124: Could you explain using linear interpolation in the dataset resampling process? Did you consider using bilinear interpolation?

Response: We used the interpolation function from the xarray package, where the parameter was set as "linear". The interpolation was applied in both dimensions of latitude and longitude; therefore, we used bilinear interpolation. We clarify accordingly in lines 138−139.

Line 229: An additional factor that may influence your analysis is the urban area fraction. Did you examine its effect on dataset's performance?

Response: Thank you for your suggestion. In this revision, we include the global artificial impervious area as one of the predictors in the attribution models to quantify urban influence, as it directly reflects surface changes associated with urbanization that impede the natural infiltration of water into the soil. Please see lines 272−273. However, the results indicate that the urban influence is not the dominant factor of dataset performance in terms of water balance consistency (Figs. S17−20).

> In lines 272−273:
>
> "*Global artificial impervious area from Gong et al. (2019) was also averaged among the available periods for each independent combination.*"

Line 245: Please specify for which period you extract tree cover data.

Response: The tree cover data we used differs across combinations because the available period for each independent combination is not consistent. In other words, we calculated 8,294 tree cover maps first, and then averaged them as one map for model input. Please find the relevant description in lines 254−257.

Line 313-320: Recently, Vargas Godoy et al., (2025) provide a global performance of several global precipitation datasets, identifying the best product at different spatial scales. Your manuscript and Vargas Godoy's results agree that IMERG and MSWEP are the best products around the world. However, I am curious about the high R2 that you reported for PERSIANN-CDR (Fig. 1) due to Vargas Godoy et al. (2025), and several regional analyses suggest that PERSIANN-CDR exhibits a low accuracy compared to ground observations. Thus, I suggest providing a potential explanation for its high performance.

Response: Thank you for your insight. First, our results are not fully comparable to Vargas Godoy et al. (2025), which identified the representativeness of *P* datasets across different regions in terms of their similarity to one another. Although Vargas Godoy et al. (2025) did not identify PERSIANN-CDR as the representative *P* dataset in most global regions, their analysis revealed close genealogical relationships between PERSIANN-CDR and IMERG or MSWEP. This could support our result on similar medians of 50ºS−50ºN among PERSIANN-CDR, IMERG, and MSWEP. Second, PERSIANN-CDR exhibits lower accuracy compared to some ground observations, but it also has relatively high performance in other regions, such as tropical regions (Sun et al., 2018). In this revision, additional interpretation is added in lines 435−438.

"For the medians of 50ºS−50ºN, several P datasets like PERSIANN-CDR, GPM IMERG v07, and MSWEP v2.8 are also comparable, which might be related to their close genealogical relationships (Markonis et al., 2024; Vargas Godoy et al., 2025)."

Lines 440 – 442: Interestingly, reanalysis products show the best performance in terms of SM. Could you extend your explanation about these results and the potential reasons behind the lower performance of satellite-based products?

Response: Yes, please see lines 471−473 in the revised manuscript.

"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)."

Lines 445-450: Why is the lowest consistency observed at the annual scale?

Response: It is because the seasonal variability is removed at the annual scale. Please see lines 477−480 in the discussion.

"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."

Technical corrections:

Please check whether the figure colors are suitable for color-blind readers.

Response: We appreciate your considerate suggestion. We have avoided using a color combination of red and green in the figures.

References

Fan, Y., Li, H., and Miguez-Macho, G.: Global Patterns of Groundwater Table Depth, Science, 339, 940–943, https://doi.org/10.1126/science.1229881, 2013.

Miguez-Macho, G. and Fan, Y.: A global humidity index with lateral hydrologic flows, Nature, 644, 413–419, https://doi.org/10.1038/s41586-025-09359-3, 2025.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., Waterloo, M., and Saleska, S.: Height Above the Nearest Drainage – a hydrologically relevant new terrain model, Journal of Hydrology, 404, 13–29, https://doi.org/10.1016/j.jhydrol.2011.03.051, 2011.

Vargas Godoy, M. R., Markonis, Y., Thomson, J. R., Ballarin, A. S., Perri, S., Miao, C., Sun, Q., Hanel, M., Papalexiou, S. M., Kummerow, C., Oki, T., and Molini, A.: Which Precipitation Dataset to Choose for Hydrological Studies of the Terrestrial Water Cycle?, Bulletin of the American Meteorological Society, BAMS-D-24-0306.1, https://doi.org/10.1175/BAMS-D-24-0306.1, 2025.