Manuscript: ESSD-2025-376

Comment by Anonymous Referee #3

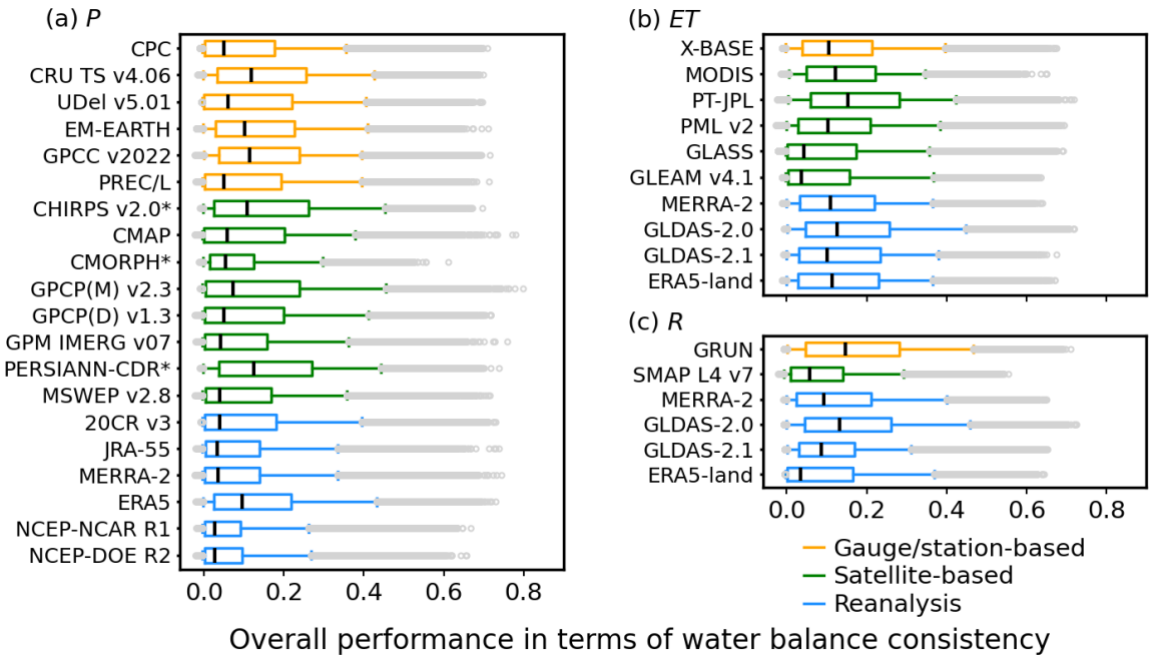
This study comprehensively evaluates the water balance consistencies among many state-of-the-art geospatial datasets on P, ET, and R. While it represents lots of work, clearly demonstrating the power of big data geospatial analysis, and the manuscript is generally well-written, there are several major concerns that should be addressed.

Response: Thank you for your encouraging comments, which have been invaluable for improving the quality of this manuscript.

1. The method for evaluating the water balance inconsistencies may need better justifications or some back-up analyses. While delta_SM is a reasonable proxy for the storage changes, it is still insufficient to capture those mass changes related to lakes/reservoirs/snow/glaciers, or those from underground. Therefore, the authors may need to explore the use of GRACE data to support their methods. I am afraid that some of the major conclusions for the high-latitude changes may be compromised if using GRACE.

Response: We agree with the reviewer's concerns, especially regarding the misrepresentation of snow and glacier changes in high-latitude regions using *SM* datasets. Therefore, we include supplementary results on water balance consistency using terrestrial water storage from GRACE in this revision. Please see the new Text S3 for detailed calculations through $P-ET-R=\Delta TWS$, where the number of independent combinations becomes less sufficient with a decrease by one order of magnitude from ~8,000 to ~900. Our results show that using ΔTWS from the GRACE has similar ranking results as using ΔSM (new Fig. S3), which supports the use of *SM* in the water balance assumption is sufficient for our study purposes. The relevant text can be found in lines 245−248:

"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 $P-ET-R=\Delta TWS$. 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."*

"***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:

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

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."

2. There are many useful insights regarding the performance of different datasets, however, it seems this study dos not directly contribute a new dataset itself? According to my understanding, ESSD's scope is more data-centered. In this regard, can the authors clarify what are the new datasets they may be able to contribute to the community?

Response: Please note that ESSD features different types of contributions, and this is a review article rather than a data description paper (please see also https://www.earth-system-science-data.net/about/manuscript_types.html). As a review article, it is not necessary to describe a new dataset which is typically done by data description papers. Therefore our study meets the requirements for articles in ESSD.

3. The authors seem to overlook several past studies working on the similar topic (e.g., https://link.springer.com/chapter/10.1007/978-3-319-32449-4_4 and relevant citing references)

Response: The suggested paper has been added to support the introduction of water balance closure (lines 42−46).

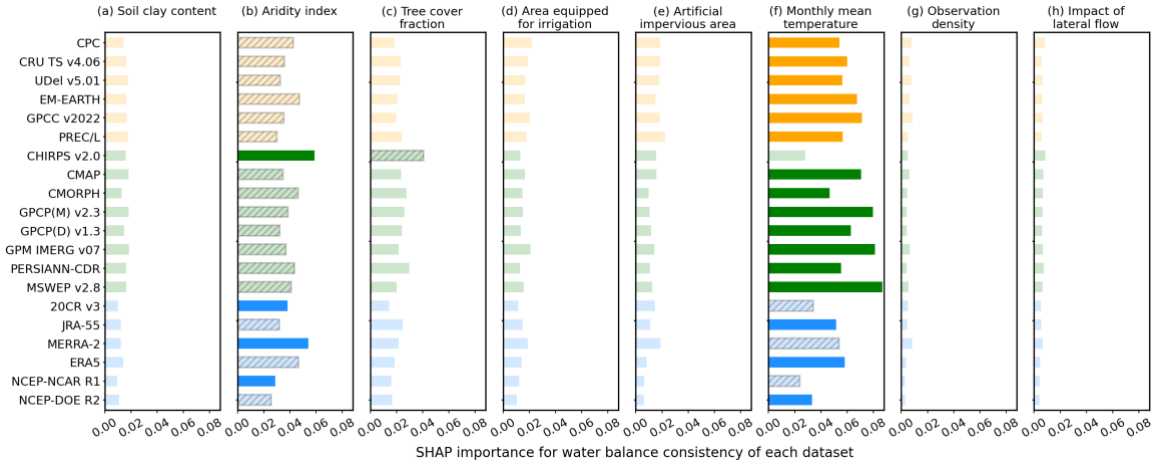
4. Although water balance closure is indeed important, there are occasions where water balance is violated because of the unobserved loss/addition of water. For most cases, it points to the error of datasets, but for some occasional cases, they may point toward new hydrological insights. Authors may need to briefly discuss the limitation of their assumption on 'water balance consistency is directly associated with good dataset performance'.

Response: We thank the reviewer for raising this valid point. In this revision, we identify potential bias in lines 242−245 when adding additional calculations using GRACE data.

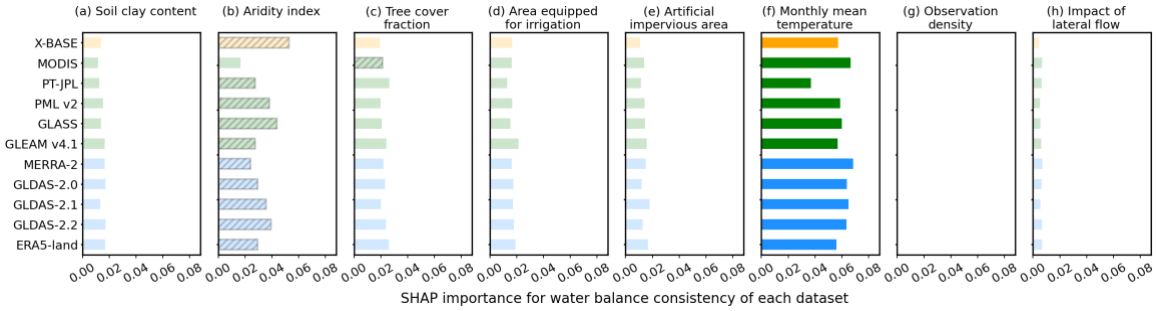
"In addition, unconsidered water variables, like glacier, snow, and surface water storage, might introduce bias into our water balance assumption, leading to a nonlinear response of ΔSM to P–ET–R."

Further, 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.

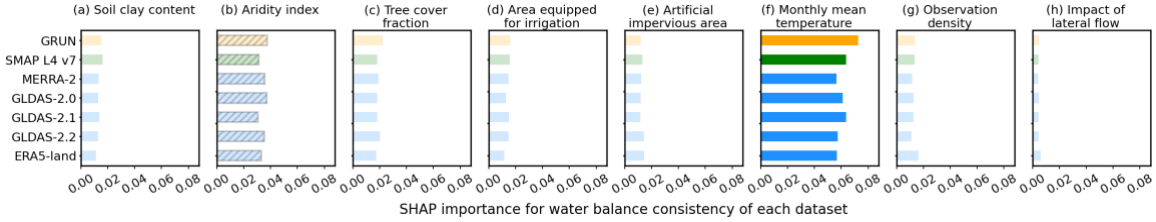
"At the same time, factors like irrigation, urbanization, and lateral flow play relatively minor roles (Figs. S17−S20)."



*"**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."*

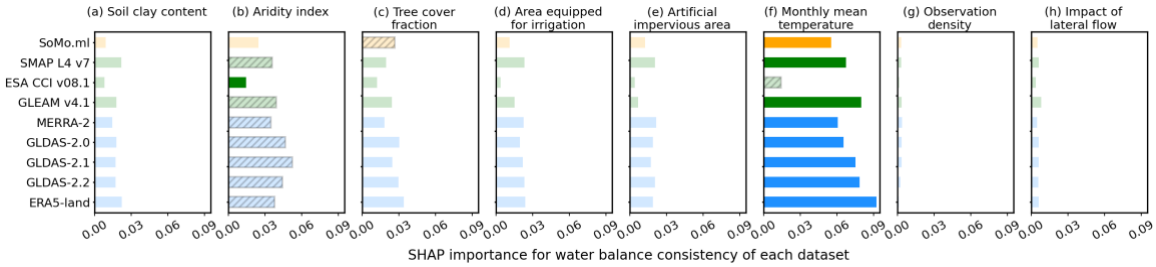


*"**Fig. S18.** Importance of (a) soil clay content, (b) aridity index, (c) tree cover fraction, (d) area equipped for irrigation, (e) artificial impervious area, (f) monthly mean temperature, (g) observation density, and (h) impact of lateral flow to water balance consistency of each ET dataset. The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