Response to Review
Reviewer #2:

Greetings. The manuscript entitled "A novel method for 1 correcting water budget components and reducing their uncertainties by optimally distributing the imbalance residual without full closure" deals with the closure of water budget problems, and specifically with uncertainties therein. The structure and goals are clear, and the results offering is well-suited. This paper can for sure be published after some adjustments, listed below. I think that these itemized improvements would make the work more scientifically sound and robust. These considerations come from my expertise as a hydrogeologist, so they will pertain to this sphere of competency. Best regards.

Response: Thank you very much for your careful review. These comments are very helpful for us to improve the manuscript. We really appreciate your time and efforts.

The point-to-point responses were given after comments. All the changes were highlighted in bright blue for easy review. We hope that the revision meets the requirement for publication.

- In equation 2, jkl should be written as subscripts, as well as 123, etc in eq. 3.

Response: Thanks for your careful review. We apologize for our carelessness. In the revised manuscript, "jkl" in Equation 2 and "123" in Equation 3 were formatted as subscripts, respectively, as shown below:

"

$$C_{jkl} = \left[P_j \;\; ET_k \;\; TWSC_l \;\; Q\right] \tag{2}$$

where j, k, and l represent the indices of the datasets corresponding to each budget component. Table 1 provides basic information on the datasets used in this study, along with their corresponding indices. Equation 3 represents a matrix composed of the elements defined in Equation 2.

$$\mathrm{C} = \begin{bmatrix} C_{111} & C_{112} & C_{113} & C_{121} & C_{122} & C_{123} & C_{131} & C_{132} & C_{133} \\ C_{211} & C_{212} & C_{213} & C_{221} & C_{222} & C_{223} & C_{231} & C_{232} & C_{233} \\ C_{311} & C_{312} & C_{313} & C_{321} & C_{322} & C_{323} & C_{331} & C_{332} & C_{333} \\ C_{411} & C_{412} & C_{413} & C_{421} & C_{422} & C_{423} & C_{431} & C_{432} & C_{433} \end{bmatrix} \tag{3}$$

".

We carefully reviewed all equations and notations throughout the manuscript to ensure that all similar issues were addressed.

- We need a more detailed specification, both in the introduction and in the methodology (e.g., from line 64 on) of the TWSC term. It is not sufficient to describe the ground-underground part of the water cycle. Two major points should be at least touched: (i) a portion of the TWSC term is the water infiltrating to aquifer, but that is returned to major water bodies soon or later (see e.g. Levison et al., 2016); (ii) the major role in the aquifer ability to store or drain the portion of water infiltrating is played by local geology, precisely its spatial distribution and the eventual presence of fractures (again, Levison et al., 2016) or highly permeable conduits (Schiavo, 2023). I think these two major points should be supported leveraging the suggested references.

Response: Thank you for your constructive suggestion. We sincerely apologize for the unclear description of the TWSC term in the original manuscript. In this study, TWSC refers to the total terrestrial water storage change, which includes but is not limited to surface water, soil moisture, groundwater, aquifer infiltration, and ice/snow components. To better represent this, we employed three different GRACE satellite-derived datasets to reduce the uncertainty of TWSC estimates.

We fully agree with you that it is insufficient to describe only surface and subsurface part of the water cycle. We should also clarify that the TWSC term includes water that infiltrates into aquifers. This water has both storage and drainage characteristics and eventually contributes to major water bodies, thereby affecting TWSC. In particular, the ability of aquifers to store or release water is strongly influenced by local geological conditions, such as spatial heterogeneity, the presence of fracture zones, and high-permeability pathways. These key aspects were incorporated into the revised manuscript, and we cited the references you suggested as essential support. Specifically, we made revisions in the following three aspects:

(1) We revised the definition of TWSC in the Introduction to clearly state that it includes surface water, soil moisture, groundwater, water infiltrating into aquifers, and ice/snow (Mehrnegar et al., 2023; Pellet et al., 2020; Wang et al., 2022). We referenced the works of Levison et al. (2016) and Schiavo (2023) to emphasize the drainage characteristics of infiltrated aquifer water (eventually returning to major water bodies) and its influencing factors (e.g., geological conditions). Specifically, we added the following sentence to the revised manuscript:

"where P represents precipitation, ET represents evapotranspiration, Q represents streamflow, and TWSC represents terrestrial water storage change. It is worth noting that TWSC refers to the change in total terrestrial water storage, including but not limited to surface water, soil moisture, groundwater, water infiltrating into aquifers, and ice/snow (Mehrnegar et al., 2023; Pellet et al., 2020; Wang et al., 2022). Infiltrated water into aquifers does not remain permanently stored, but is eventually returned to major water bodies sooner or later (Levison et al., 2016). The ability of aquifers to retain or transmit infiltrated water is strongly influenced by local geological characteristics, particularly the spatial heterogeneity, presence of fractures, or high-permeability pathways (Levison et al., 2016; Schiavo, 2023).".

(2) We added a more detailed description of the GRACE TWSC data in the Data section of the manuscript. By briefly explaining the principle by which GRACE satellites monitor TWSC, we clarified that the GRACE-derived TWSC represents total terrestrial water storage change. The following sentence was added to the revised manuscript:

"The launch of the GRACE and GRACE Follow-On (GRACE-FO) satellite missions has provided new opportunities for more accurate observations of large-scale TWSC. GRACE operated from 2002 to 2017, followed by GRACE-FO starting in 2018 (Boergens et al., 2024). These missions infer terrestrial total TWSC by tracking temporal variations in Earth's gravity field, which are primarily attributed to changes in terrestrial water mass. The GRACE TWSC datasets used in this study are provided by the University of Texas Center for Space Research (CSR), the German Research Centre for Geosciences (GFZ), and NASA's Jet Propulsion Laboratory (JPL), all of which include multiple bias correction procedures to improve data quality (Landerer et al., 2012; Shamsudduha et al., 2017). These bias correction procedures include filtration to suppress correlated noise and striping artifacts (Swenson et al., 2006), replacement of poorly resolved spherical harmonic coefficients (e.g., degree-2 term C20) with satellite laser ranging data (Loomis et al., 2020), and correction for glacial isostatic adjustment (GIA) (Peltier et al., 2012; Mu et al., 2017). Numerous studies have demonstrated the sensitivity and reliability of GRACE satellite data for monitoring TWSC (Swenson and Wahr, 2006; Resende et al., 2019; Majid et al., 2019; Reager et al., 2014).".

(3) In the Methodology section, we explicitly emphasized that the TWSC data used in the BCC method refer to the total terrestrial water storage change observed by the GRACE satellite:

"In the following application of the BCC method, the TWSC data used in this study refer to the basin-scale total terrestrial water storage change observed by GRACE satellite data.".

- Global precipitation models, as well as other kinds of climate products, need to be bias corrected to be employable, even if these issues and the opportunity of such procedures are still subject of scientific debate (e.g., Ehret et al., 2012). Are the employed data raw or bias corrected (if so, how)?

Response: Thank you very much for your careful review, which has been highly valuable in improving the quality of our manuscript. To more reliably analyze the uncertainties introduced by existing BCC methods and to verify the robustness of the IWE-Res method proposed in this study, we selected multiple datasets for each budget component, forming multiple data combinations in each basin. Specifically, we used four P datasets (GPCC, GPM IMERG Final Run, MSWEP, and PERSIANN-CDR), three ET datasets (GLDAS, GLEAM, and TerraClimate), one observed Q dataset, and three GRACE-based TWSC datasets.

We apologize for not clearly stating the bias correction status of these datasets in the original manuscript. (1) All datasets used in this study for driving BCC methods have undergone bias correction according to the standards of their respective data providers or have been subject to rigorous quality control procedures to ensure their accuracy and reliability. Therefore, we did not apply any further bias correction to these driving datasets ourselves; (2) For the datasets produced using both existing BCC methods and our proposed IWE-Res method in this study, no additional bias correction against observational data was performed. We added a discussion on the potential for further bias correction of these produced datasets by this study. Our specific revisions are as follows:

(1) We added descriptions of the bias correction and quality control for datasets driving BCC methods used in this study:

The GPCC precipitation dataset is derived from a large number of ground-based rain gauge observations collected globally and is produced through rigorous quality control and spatial interpolation procedures (Song et al., 2022; Wei et al., 2019; Abdelwares et al., 2020). The quality control process includes verification of station metadata (such as location and elevation), checks for temporal consistency, and the removal of extreme or erroneous values. As the GPCC dataset is interpolated from rain gauge observations and has undergone strict quality assurance, it is widely used in previous studies as a benchmark for bias correction of precipitation estimates from climate models.

The GPM IMERG Final Run precipitation product is a high-resolution, multi-source satellite-based precipitation dataset. It incorporates ground-based rain gauge observations and applies bias correction at the monthly scale (Wang et al., 2017; Cui et al., 2020; Huang et al., 2019).

The MSWEP dataset is produced by optimally merging precipitation data from satellite observations, ground stations, and reanalysis products (Beck et al., 2019; Beck et al., 2017). This dataset incorporates bias correction at the daily scale using more than 77,000 gauge observations worldwide (Beck et al., 2019; Beck et al., 2017).

The PERSIANN-CDR precipitation product is bias-corrected using the Global Precipitation Climatology Project (GPCP) 2.5° monthly product, which includes gauge data from the Global

Precipitation Climatology Centre (GPCC) (Chen et al., 2020; Kaprom et al., 2025; Sadeghi et al., 2019).

In contrast to precipitation, global ET datasets generally lack standardized and comprehensive bias correction procedures. Most bias correction approaches for ET are indirect, focusing on correcting the climate forcing inputs used to drive evapotranspiration models. This is primarily due to the limited availability of long-term, high-density in situ ET measurements globally—for example, the sparse distribution and limited representativeness of eddy covariance flux towers. The three ET products used in this study (GLDAS, GLEAM, and TerraClimate) improve data quality primarily through such indirect bias correction methods.

For TWSC, the GRACE TWSC datasets used in this study are provided by the University of Texas Center for Space Research (CSR), the German Research Centre for Geosciences (GFZ), and NASA's Jet Propulsion Laboratory (JPL), including multiple bias correction procedures to improve data quality (Landerer et al., 2012; Shamsudduha et al., 2017). These include filtering to suppress correlated noise and striping artifacts (Swenson et al., 2006), replacement of poorly resolved spherical harmonic coefficients (e.g., degree-2 term C20) with satellite laser ranging data (Loomis et al., 2020), and correction for glacial isostatic adjustment (GIA) (Peltier et al., 2012; Mu et al., 2017). Numerous studies have demonstrated the sensitivity and reliability of GRACE satellite data for monitoring TWSC (Swenson and Wahr, 2006; Resende et al., 2019; Majid et al., 2019; Reager et al., 2014).

The following sentences were added to the revised manuscript:

"Given the biases in the outputs of global P and ET models, observationally constrained datasets that have undergone bias correction or rigorous quality control are generally considered more accurate and reliable (Ehret et al., 2012). Accordingly, priority was given to datasets that incorporate extensive ground-based observations and provide bias-corrected or quality-controlled products. We selected four P datasets—GPCC, GPM IMERG, MSWEP, and PERSIANN-CDR; three ET datasets—GLDAS, GLEAM, and TerraClimate; and three TWSC datasets derived from GRACE satellite observations—GRACE CSR, GRACE GFZ, and GRACE JPL. All datasets were either bias-corrected according to the standards of their respective data providers or subjected to systematic quality control. Observed Q data were obtained from the GRDC platform. By combining these datasets, a total of 36 distinct data combinations were generated for each basin (Equation 3).

$$C_{jkl} = \left[P_j \;\; ET_k \;\; TWSC_l \;\; Q \right] \tag{2}$$

where *j*, *k*, and *l* represent the indices of the datasets corresponding to each budget component. Table 1 provides basic information on the datasets used in this study, along with their corresponding indices. Equation 3 represents a matrix composed of the elements defined in Equation 2.

$$\mathrm{C} = \begin{bmatrix} C_{111} & C_{112} & C_{113} & C_{121} & C_{122} & C_{123} & C_{131} & C_{132} & C_{133} \\ C_{211} & C_{212} & C_{213} & C_{221} & C_{222} & C_{223} & C_{231} & C_{232} & C_{233} \\ C_{311} & C_{312} & C_{313} & C_{321} & C_{322} & C_{323} & C_{331} & C_{332} & C_{333} \\ C_{411} & C_{412} & C_{413} & C_{421} & C_{422} & C_{423} & C_{431} & C_{432} & C_{433} \end{bmatrix} \tag{3}$$

The Global Precipitation Climatology Centre (GPCC) dataset, provided by the German Weather Service (DWD), is derived from a dense global network of rain gauge observations, and incorporates strict quality control procedures such as station data validation, temporal consistency checks, and outlier removal (Becker et al., 2013; Schneider et al., 2008). The dataset is available at 0.25° spatial resolution and daily to monthly temporal scales. The Global Precipitation Measurement Integrated Multi-Satellite Retrievals (GPM IMERG) Final Run product, developed by NASA and JAXA, integrates multiple satellite-based precipitation estimates and applies monthly bias correction using ground-based gauge data (Wang et al., 2017; Cui et al., 2020; Huang et al., 2019). The Multi-Source Weighted-Ensemble Precipitation (MSWEP) dataset combines satellite, gauge, and reanalysis data using an ensemble-weighted approach, incorporating over 77,000 ground stations for daily-scale bias correction (Beck et al., 2019; Beck et al., 2017). The PERSIANN-CDR dataset, based on satellite remote sensing and artificial neural network technology, spans 60°S to 60°N with 0.25° daily resolution, and is bias-corrected using the GPCP monthly product, which includes extensive rain gauge observations (Chen et al., 2020; Kaprom et al., 2025; Sadeghi et al., 2019).

For ET, the Global Land Data Assimilation System (GLDAS), developed by NASA and NOAA, uses land surface modeling and data assimilation to produce physically consistent estimates of land surface fluxes. The GLEAM dataset, developed by the Miralles team at the University of Bristol, estimates actual ET using satellite-derived net radiation and air temperature via the Priestley-Taylor model, and applies a stress factor derived from vegetation optical depth (VOD) and soil moisture to adjust potential evaporation. TerraClimate dataset provides global monthly actual ET estimates based on the Penman Montieth approach (Abatzoglou et al., 2018). Notably, bias correction in global ET products is generally less systematic than for P products, mainly due to the limited availability and spatial coverage of in situ flux tower observations. As a result, bias adjustments in ET datasets are typically indirect, relying on corrections applied to the climate forcing variables rather than to ET itself.

The launch of the GRACE and GRACE Follow-On (GRACE-FO) satellite missions has provided new opportunities for more accurate observations of large-scale TWSC. GRACE operated from 2002 to 2017, followed by GRACE-FO starting in 2018 (Boergens et al., 2024). These missions infer terrestrial total TWSC by tracking temporal variations in Earth's gravity field, which are primarily attributed to changes in terrestrial water mass. The GRACE TWSC datasets used in this study are provided by the University of Texas Center for Space Research (CSR), the German Research Centre for Geosciences (GFZ), and NASA's Jet Propulsion Laboratory (JPL), all of which include multiple bias correction procedures to improve data quality (Landerer et al., 2012; Shamsudduha et al., 2017). These bias correction procedures include filtration to suppress correlated noise and striping artifacts (Swenson et al., 2006), replacement of poorly resolved spherical harmonic coefficients (e.g., degree-2 term C20) with satellite laser ranging data (Loomis et al., 2020), and correction for glacial isostatic adjustment (GIA) (Peltier et al., 2012; Mu et al., 2017). Numerous studies have demonstrated the sensitivity and reliability of GRACE satellite data for monitoring TWSC (Swenson and Wahr, 2006; Resende et al., 2019; Majid et al., 2019; Reager et al., 2014).

The Global Runoff Data Centre (GRDC) provides the most comprehensive open-access river discharge data available worldwide, collected from national hydrological agencies. This dataset includes river streamflow measurements from over 10,000 stations across 159 countries (Su et al., 2024). To minimize the impact of missing data on the reliability of the results, hydrological stations were selected based on the criterion that missing values accounted for less than 10% of the total dataset. Linear interpolation was then applied to fill any remaining data gaps.".

(2) We discussed the potential for further bias correction of the budget corrected datasets generated using the proposed IWE-Res method, based on observational data, although this lies beyond the scope of the present study. The following sentence has been added to the revised

manuscript:

"It is worth noting that the datasets generated by both the existing BCC methods and the IWE-Res method proposed in this study were not further bias-corrected against independent observations. For basin-specific applications requiring higher reliability, we recommend additional bias correction.".

- The resolution of the spatial problem is crucial. It is well known that changes in resolution make employed variables (such as DEMs, as in Aziz et al., 2022) very different at the same location. How to tackle this point? Which resolution 'advice' for the reader? Any criterion? Otherwise, the errors will propagate in a sort of uncontrolled cascade.

Response: Thank you for this important comment. We fully agree that the spatial resolution differences among datasets for budget components can introduce uncertainty, especially when rescaling is required. We apologize for not clearly stating the spatial resolution used in our study. In this work, water budget closure was performed at the basin scale and at the monthly temporal resolution. Specifically, all gridded datasets of budget components were spatially and temporally aggregated to monthly and basin scales, respectively. Therefore, we did not resample all datasets to a common grid resolution; instead, we upscaled each variable directly to the basin scale. Previous studies on water budget closure have also focused primarily on the basin scale (e.g., Lehmann et al., 2022; Abolafia-Rosenzweig et al., 2023; Luo et al., 2023; Wang et al., 2014; Tan et al., 2022; Sahoo et al., 2011), mainly for the following reasons:

(1) Water budget closure correction at the grid scale is extremely complex, as it requires accurate quantification of all inflow and outflow terms. These may include precipitation, evapotranspiration, changes in water storage, lateral inflows and outflows, leakage losses, and human withdrawals or returns. Many of these components are difficult to observe or estimate directly, particularly lateral fluxes and leakage. Neglecting these terms introduces substantial uncertainty into grid-scale analyses; (2) As you correctly noted, merging datasets with differing spatial resolutions onto a common grid introduces additional sources of uncertainty, which can reduce the accuracy of water budget closure correction; (3) The GRACE satellite-derived TWSC data have a relatively coarse spatial resolution, which limits their applicability at the grid scale. Since TWSC is a critical component of the monthly water budget closure correction, it cannot be omitted. By aggregating GRACE data to the basin scale, random errors tend to cancel out, thereby improving accuracy; (4) Despite recent advances in remote sensing and the expansion of global observational networks, significant uncertainties remain in gridded hydrological datasets when applied at fine spatial scales, especially for evapotranspiration. Thus, conducting water budget closure correction at the basin scale remains a central focus of current research.

In the revised manuscript, on the one hand, we clarified the spatial resolution applied in our analysis as "The above datasets were upscaled to the basin and monthly scales using spatial and temporal averaging."; On the other hand, we added a detailed discussion of why the basin scale was selected for water budget closure correction in this study as "Notably, the choice of spatial resolution has a significant impact on the results (Aziz et al., 2022; Bormann et al., 2006; Senan et al., 2022). Following many previous studies (Lehmann et al., 2022; Abolafia-Rosenzweig et al., 2023; Luo et al., 2023; Wang et al., 2014; Tan et al., 2022; Sahoo et al., 2011), the BCC method in this study is also applied at the basin scale rather than the grid scale for the following reasons: 1) Achieving water budget closure at the grid scale is complex and challenging due to the difficulty of quantifying all water flux and storage components flowing into and out of the grid, including P, ET, TWSC, lateral inflow and outflow, leakage losses, and human water withdrawals and returns. Several of these components, such as lateral flow and leakage, are poorly observed or highly uncertain, and their omission introduces substantial error; 2) The datasets of different variables have varying spatial resolutions, and resampling them to a common resolution introduces uncertainties, which in turn affect the accuracy of water budget closure correction; 3) The coarse spatial resolution of GRACE-derived TWSC data limits their applicability for water budget closure calculation at the grid scale. At monthly resolution, TWSC is a critical component and cannot be neglected. Averaging GRACE data to the basin scale helps reduce random errors by offsetting positive and negative biases, thereby increasing the reliability of water budget closure correction; 4) Despite advances in remote sensing and in situ observation networks, grid-scale uncertainties remain substantial for some budget components, such as ET. Basin-scale analysis therefore reduces uncertainty and improves the reliability of water budget closure correction results.".

- From line 289 on, I think that the noise to afflict the Kalman Filter with should be pointed out. Which is the observation noise covariance (and its quantification for the employed variables)?

Response: Thank you for your careful review. We sincerely apologize for the insufficient explanation of the observation noise covariance and its quantification in the CKF method in the original manuscript. The CKF method used in this study is developed by Pan and Wood (2006) based on the Kalman Filter method. Given that the CKF method has been widely applied in previous studies for correcting budget component datasets, we included it as one of the benchmark methods to evaluate the performance of the proposed IWE-Res method.

In the CKF method, the reference values used to compute the error covariance (ε in Equations 7–10) differ among the four budget components (P, ET, Q, and TWSC). For P, ET, and TWSC, due to the lack of spatially matched ground-truth observations at the grid scale, previous studies (Zhang et al., 2018; Abolafia-Rosenzweig et al., 2021) have typically used the ensemble mean across all datasets considered in their studies as the reference value. For Q, both previous studies and the present work use observed Q in the CKF method. Previous studies have reported gauge-based uncertainty as a percent error for some of the basins, ranging from 2.3% – 28.8% (Clarke, 1999; Mueller, 2003; Shiklomanov et al., 2006; Abolafia-Rosenzweig et al., 2021).

To ensure consistency with previous studies, we adopted the same assumptions in our application of the CKF method. However, as this method relies on approximated reference values, it may introduce inaccuracies in the error estimation for certain budget components. This, in turn, can propagate uncertainty into the corrected datasets. Hence, evaluating the uncertainty ranges of existing correction methods and developing new methods to reduce those uncertainties is essential. This concern forms a core motivation of our study: (1) to quantify the uncertainties introduced by existing BCC methods (CKF, MCL, MSD, and PR) at the monthly scale across 84 global basins spanning diverse climate zones; and (2) to propose a novel method, IWE-Res, for identifying the optimal balance in ΔRes redistribution, minimizing the combined error from both introduced budget component errors and the remaining ΔRes error.

In the revised manuscript, we carefully revised the description of the CKF method to clarify the relationships among the relevant equations and to more clearly explain the quantification of the error covariance for each budget component. Additionally, we incorporated more references to previous studies that have applied the CKF method. The specific revisions made to the manuscript are as follows:

"The CKF method is developed based on the Kalman filter method (Pan and wood, 2006). For a given set of estimated budget components $X = [P\ ET\ Q\ TWSC]^T$ and their estimated errors $\Delta Res = GX \neq 0$ (where G is a constant vector, $G = [1\ -1\ -1\ -1]$), the goal is to find a new set of estimates $F = [P'\ ET'\ Q'\ TWSC']^T$ such that $GX' = 0$, achieving water budget closure (Pan et al., 2012). In simple terms, the CKF method redistributes the ΔRes among the budget components based on the error covariance of X, defined as $\Delta\varepsilon_{XX}$ (Equation 7), to obtain a closured dataset.

$$\Delta\varepsilon_{XX} = \overline{(X - X_0)(X - X_0)^T} \tag{7}$$

where X_0 refers to the reference values of the estimated budget components, and the bar over an expression denotes expectation. For P, ET and TWSC, the reference values X_0 were calculated by averaging all considered datasets, following previous studies (Zhang et al., 2018; Abolafia-Rosenzweig et al., 2021). For Q, we adopted observed Q. Due to the difficulty in quantifying the uncertainty in observed Q, previous studies have reported gauge-based uncertainty as a percent error for some of the basins, ranging from 2.3%–28.8% (Clarke, 1999; Mueller, 2003; Shiklomanov et al., 2006; Abolafia-Rosenzweig et al., 2021). We followed a similar approach to estimate the uncertainty associated with Q in this study.

The error covariance matrix $\Delta\varepsilon_{XX}$ is of dimension 4×4 and represents the covariances among errors in the four budget components:

$$\Delta\varepsilon_{XX} = \begin{bmatrix} \Delta\varepsilon_{P-P} & \Delta\varepsilon_{P-ET} & \Delta\varepsilon_{P-Q} & \Delta\varepsilon_{P-TWSC} \\ \Delta\varepsilon_{ET-P} & \Delta\varepsilon_{ET-ET} & \Delta\varepsilon_{ET-Q} & \Delta\varepsilon_{ET-TWSC} \\ \Delta\varepsilon_{Q-P} & \Delta\varepsilon_{Q-ET} & \Delta\varepsilon_{Q-Q} & \Delta\varepsilon_{Q-TWSC} \\ \Delta\varepsilon_{TWSC-P} & \Delta\varepsilon_{TWSC-ET} & \Delta\varepsilon_{TWSC-Q} & \Delta\varepsilon_{TWSC-TWSC} \end{bmatrix} \tag{8}$$

Following Pan et al. (2012), the off-diagonal elements representing cross-variable error covariances were assumed to be zero, under the assumption that errors among different budget components are uncorrelated. Accordingly, the matrix F can be computed as shown in Equation 9.

$$F = X + K(0 - GX) \tag{9}$$

where $K = \Delta\varepsilon_{XX} C^T (C \Delta\varepsilon_{XX} C^T)^{-1}$ is the Kalman gain. Setting $GX = \Delta Res$, and Equation 9 can be rewritten as Equation 10.

$$F = X - \Delta\varepsilon_{XX} G^T (G \Delta\varepsilon_{XX} G^T)^{-1} \Delta Res \tag{10}$$

where error covariance ε_{XX} is calculated entry by entry according to Equation 8.".

- Is there any exit criterion (such as tolerances) to exit from equations 9 and 10?

Response: Thank you for your careful review. The existing CKF method uses the total number of months within the study period as the only termination criterion for iteration—that is, it runs for a fixed number of iterations corresponding to the study duration. No additional exit conditions are set for Equations 9 and 10. This design is rooted in the core principle of the CKF method: it first estimates the error covariances of the water budget components and then proportionally redistributes the water imbalance (ΔRes) back to the raw data based on these estimated errors. However, if the error covariances are inaccurately estimated, the redistribution of ΔRes may be suboptimal. This misallocation can then propagate: an excessive correction to one variable reduces the remaining ΔRes available for others, potentially resulting in further inaccuracies. Notably, this iteration process continues until the predetermined number of steps (the total number of months) is reached, regardless of whether such misallocations occur.

We highlighted this limitation in the modified manuscript by stating: “Beyond introducing negative values, such imbalanced redistribution compromises the integrity of the remaining components. Overcorrecting one variable necessarily reduces the share of ΔRes available for others, potentially degrading their accuracy.”.

To address these limitations of existing methods, we proposed the IWE-Res method. Unlike existing methods, IWE-Res introduces exit conditions based on the physical plausibility of the corrected variables. Specifically, we terminate the iteration if any of the following occurs: P, ET, or Q becomes negative, or if the sign of TWSC reverses (from positive to negative or vice versa). The following sentences were added to the revised manuscript as: “During the iterative correction process, if any of the water budget components (P, ET, and Q) becomes negative, the redistribution of water imbalance error to that component is immediately suspended. In subsequent iterations, redistribution is recalculated to ensure that only components with physically meaningful positive values receive the imbalance correction. For example, if ET becomes negative in a given iteration, the imbalance is subsequently redistributed to P, Q, and TWSC only, in accordance with Equation 33. For TWSC, if a sign reversal occurs during iteration (i.e., from positive to negative or vice versa), the redistribution of the water imbalance error to TWSC is suspended in the following iteration.”.

Suggested References:

- Levison et al., 2016. Long-term trends in groundwater recharge and discharge in a fractured bedrock aquifer – past and future conditions. Canadian Water Resources Journal / Revue canadienne des ressources hydriques, Volume 41, 2016 - Issue 4: Special Issue: Groundwater – Surface Water Interactions in Canada. https://doi.org/10.1080/07011784.2015.1037795
- Schiavo, 2023. The role of different sources of uncertainty on the stochastic quantification of subsurface discharges in heterogeneous aquifers. J. Hydrol. 617 (4), 128930. DOI: 10.1016/j.jhydrol.2022.128930

Further reading references:

- Ehret, U., Zehe, E., Wulfmeyer, V., Warrach-Sagi, K., and Liebert, J.: HESS Opinions "Should we apply bias correction to global and regional climate model data?", Hydrol. Earth Syst. Sci., 16, 3391–3404, https://doi.org/10.5194/hess-16-3391-2012, 2012.
- Aziz, K.M.A., Rashwan, K.S. Comparison of different resolutions of six free online DEMs with GPS elevation data on a new 6th of October City, Egypt. *Arab J Geosci* **15**, 1585 (2022). https://doi.org/10.1007/s12517-022-10845-5

Response: Thanks very much for the useful references. We carefully reviewed the suggested references and incorporated them into the revised manuscript.

References:

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Scientific Data, 5(1), 1-12. https://doi.org/10.1038/sdata.2017.191

Abdelwares, M., Lelieveld, J., Zittis, G., Haggag, M., & Wagdy, A. (2020). A comparison of gridded datasets of precipitation and temperature over the Eastern Nile Basin region. Euro-Mediterranean Journal for Environmental Integration, 5, 1-16. https://doi.org/10.1007/s41207-

019-0140-y
Abolafia-Rosenzweig, R., Pan, M., Zeng, J. L., & Livneh, B. (2021). Remotely sensed ensembles of the terrestrial water budget over major global river basins: An assessment of three closure techniques. Remote Sensing of Environment, 252, 112191. https://doi.org/10.1016/j.rse.2020.112191
Aziz, K. M. A., & Rashwan, K. S. (2022). Comparison of different resolutions of six free online DEMs with GPS elevation data on a new 6th of October City, Egypt. Arabian Journal of Geosciences, 15(20), 1585. https://doi.org/10.1007/s12517-022-10845-5
Beck, H. E., Vergopolan, N., Pan, M., Levizzani, V., Van Dijk, A. I., Weedon, G. P., . . . Wood, E. F. (2017). Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrology and Earth System Sciences, 21(12), 6201-6217. https://doi.org/10.5194/hess-21-6201-2017
Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., . . . Adler, R. F. (2019). MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. Bulletin of the American Meteorological society, 100(3), 473-500. https://doi.org/10.1175/BAMS-D-17-0138.1
Becker, A., Finger, P., Meyer-Christoffer, A., Rudolf, B., Schamm, K., Schneider, U., & Ziese, M. (2013). A description of the global land-surface precipitation data products of the Global Precipitation Climatology Centre with sample applications including centennial (trend) analysis from 1901–present. Earth System Science Data, 5(1), 71-99. https://doi.org/10.5194/essd-5-71-2013
Boergens, E., Güntner, A., Sips, M., Schwatke, C., & Dobslaw, H. (2024). Interannual variations of terrestrial water storage in the East African Rift region. Hydrology and Earth System Sciences, 28(20), 4733-4754. https://doi.org/10.5194/hess-28-4733-2024
Bormann, H. (2006). Impact of spatial data resolution on simulated catchment water balances and model performance of the multi-scale TOPLATS model. Hydrology and Earth System Sciences, 10(2), 165-179. https://doi.org/10.5194/hess-10-165-2006
Chen, S., Liu, B., Tan, X., & Wu, Y. (2020). Inter-comparison of spatiotemporal features of precipitation extremes within six daily precipitation products. Climate Dynamics, 54, 1057-1076. https://doi.org/10.1007/s00382-019-05045-z
Clarke, R. T. (1999). Uncertainty in the estimation of mean annual flood due to rating-curve indefinition. Journal of Hydrology, 222(1-4), 185-190. https://doi.org/10.1016/S0022-1694(99)00097-9
Cui, W., Dong, X., Xi, B., Feng, Z. H. E., & Fan, J. (2020). Can the GPM IMERG final product accurately represent MCSs' precipitation characteristics over the central and eastern United States?. Journal of Hydrometeorology, 21(1), 39-57. https://doi.org/10.1175/JHM-D-19-0123.1
Ehret, U., Zehe, E., Wulfmeyer, V., & Liebert, J. (2012). Should we apply bias correction to global and regional climate model data? HESS, 16, 3391–3404. https://doi.org/10.5194/hess-16-3391-2012, 2012.
Huang, C., Hu, J., Chen, S., Zhang, A., Liang, Z., Tong, X., ... & Zhang, Z. (2019). How well can IMERG products capture typhoon extreme precipitation events over southern China?. Remote Sensing, 11(1), 70. https://doi.org/10.3390/rs11010070
Kaprom, C., Williams, J. A., Mehrotra, R., Ophaphaibun, C., & Sriwongsitanon, N. (2025). A

comprehensive evaluation of the accuracy of satellite-based precipitation estimates over Thailand. Journal of Hydrology: Regional Studies, 59, 102380. https://doi.org/10.1016/j.ejrh.2025.102380

Landerer, F. W., & Swenson, S. C. (2012). Accuracy of scaled GRACE terrestrial water storage estimates. Water resources research, 48(4). https://doi.org/10.1029/2011WR011453

Lehmann, F., Vishwakarma, B. D., & Bamber, J. (2022). How well are we able to close the water budget at the global scale?. Hydrology and Earth System Sciences, 26(1), 35-54. https://doi.org/10.5194/hess-26-35-2022

Levison, J., Larocque, M., Ouellet, M. A., Ferland, O., & Poirier, C. (2016). Long-term trends in groundwater recharge and discharge in a fractured bedrock aquifer–past and future conditions. Canadian Water Resources Journal/Revue canadienne des ressources hydriques, 41(4), 500-514. https://doi.org/10.1080/07011784.2015.1037795

Loomis, B. D., Rachlin, K. E., Wiese, D. N., Landerer, F. W., & Luthcke, S. B. (2020). Replacing GRACE/GRACE‐FO with satellite laser ranging: Impacts on Antarctic Ice Sheet mass change. Geophysical Research Letters, 47(3), e2019GL085488. https://doi.org/10.1029/2019GL085488

Luo, Z., Li, H., Zhang, S., Wang, L., Wang, S., & Wang, L. (2023). A novel two‐step method for enforcing water budget closure and an intercomparison of budget closure correction methods based on satellite hydrological products. Water Resources Research, 59(3), e2022WR032176. https://doi.org/10.1029/2022WR032176

Majid, R., & Ardalan, E. S. (2019). Performance of the Gravity Recovery and Climate Experiment (GRACE) method in monitoring groundwater-level changes in local-scale study regions within Iran. Hydrogeology Journal, 27(7), 2497-2509. https://doi.org/10.1007/s10040-019-02007-x

Mehrnegar, N., Schumacher, M., Jagdhuber, T., & Forootan, E. (2023). Making the best use of GRACE, GRACE‐FO and SMAP data through a constrained Bayesian data‐model integration. Water Resources Research, 59(9), e2023WR034544. https://doi.org/10.1029/2023WR034544

Mu, D., Yan, H., Feng, W., & Peng, P. (2017). GRACE leakage error correction with regularization technique: Case studies in Greenland and Antarctica. Geophysical Journal International, 208(3), 1775-1786. https://doi.org/10.1093/gji/ggw494

Mueller, D. S. (2003, March). Field evaluation of boat-mounted acoustic Doppler instruments used to measure streamflow. In Proceedings of the IEEE/OES Seventh Working Conference on Current Measurement Technology, 2003. (pp. 30-34). IEEE. https://doi.org/10.1109/CCM.2003.1194278

Pan, M., Sahoo, A. K., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2012). Multisource estimation of long-term terrestrial water budget for major global river basins. Journal of Climate, 25(9), 3191-3206. https://doi.org/10.1175/JCLI-D-11-00300.1

Pellet, V., Aires, F., Papa, F., Munier, S., & Decharme, B. (2020). Long-term total water storage change from a Satellite Water Cycle reconstruction over large southern Asian basins. Hydrology and Earth System Sciences, 24(6), 3033-3055. https://doi.org/10.5194/hess-24-3033-2020

Peltier, W. R., Drummond, R., & Roy, K. (2012). Comment on “Ocean mass from GRACE and glacial isostatic adjustment” by DP Chambers et al. Journal of Geophysical Research: Solid Earth, 117(B11). https://doi.org/10.1029/2011JB008967

Reager, J. T., Thomas, B. F., & Famiglietti, J. S. (2014). River basin flood potential inferred using GRACE gravity observations at several months lead time. Nature Geoscience, 7(8), 588-592. https://doi.org/10.1038/ngeo2203

Resende, T. C., Longuevergne, L., Gurdak, J., Leblanc, M., Favreau, G., Ansems, N., ... & Aureli, A. (2019). Assessment of the impacts of climate variability on total water storage across Africa: implications for groundwater resources management. Hydrogeology Journal, 27(1), 493-512. https://doi.org/10.1007/s10040-018-1864-5

Sadeghi, M., Akbari Asanjan, A., Faridzad, M., Afzali Gorooh, V., Nguyen, P., Hsu, K., ... & Braithwaite, D. (2019). Evaluation of PERSIANN-CDR constructed using GPCP V2. 2 and V2. 3 and a comparison with TRMM 3B42 V7 and CPC unified gauge-based analysis in global scale. Remote Sensing, 11(23), 2755. https://doi.org/10.3390/rs11232755

Sahoo, A. K., Pan, M., Troy, T. J., Vinukollu, R. K., Sheffield, J., & Wood, E. F. (2011). Reconciling the global terrestrial water budget using satellite remote sensing. Remote Sensing of Environment, 115(8), 1850-1865. https://doi.org/10.1016/j.rse.2011.03.009

Schiavo, M. (2023). The role of different sources of uncertainty on the stochastic quantification of subsurface discharges in heterogeneous aquifers. Journal of Hydrology, 617, 128930. https://doi.org/10.1016/j.jhydrol.2022.128930

Schneider, U., Fuchs, T., Meyer-Christoffer, A., & Rudolf, B. (2008). Global precipitation analysis products of the GPCC. Global Precipitation Climatology Centre (GPCC), DWD, Internet Publikation, 112, 3819-3837.

Senan, S., Thomas, J., Vema, V. K., Jainet, P. J., Nizar, S., Sivan, S., & Sudheer, K. P. (2022). A study of the influence of rainfall datasets' spatial resolution on stream simulation in Chaliyar River Basin, India. Journal of Water and Climate Change, 13(12), 4234-4254. https://doi.org/10.2166/wcc.2022.273

Shamsudduha, M., Taylor, R. G., Jones, D., Longuevergne, L., Owor, M., & Tindimugaya, C. (2017). Recent changes in terrestrial water storage in the Upper Nile Basin: an evaluation of commonly used gridded GRACE products. Hydrology and Earth system sciences, 21(9), 4533-4549. https://doi.org/10.5194/hess-21-4533-2017

Shiklomanov, A. I., Yakovleva, T. I., Lammers, R. B., Karasev, I. P., Vörösmarty, C. J., & Linder, E. (2006). Cold region river discharge uncertainty—Estimates from large Russian rivers. Journal of Hydrology, 326(1 – 4), 231 – 256. https://doi.org/10.1016/j.jhydrol.2005.10.037

Song, L., Xu, C., Long, Y., Lei, X., Suo, N., & Cao, L. (2022). Performance of seven gridded precipitation products over arid central Asia and subregions. Remote Sensing, 14(23), 6039. https://doi.org/10.3390/rs14236039

Su, Y., & Zhang, S. (2024). Optimizing Parameters in the Common Land Model by Using Gravity Recovery and Climate Experiment Satellite Observations. Land, 13(4), 508. https://doi.org/10.3390/land13040508

Swenson, S., & Wahr, J. (2006). Post‐processing removal of correlated errors in GRACE data. Geophysical research letters, 33(8). https://doi.org/10.1029/2005GL025285

Tan, X., Liu, B., Tan, X., & Chen, X. (2022). Long‐term water imbalances of watersheds resulting from biases in hydroclimatic data sets for water budget analyses. Water Resources Research, 58(3), e2021WR031209. https://doi.org/10.1029/2021WR031209

Wang, L., Wang, J., Li, M., Wang, L., Li, X., & Zhu, L. (2022). Response of terrestrial water storage and its change to climate change in the endorheic Tibetan Plateau. Journal of Hydrology, 612, 128231. https://doi.org/10.1016/j.jhydrol.2022.128231

Wang, S., McKenney, D. W., Shang, J., & Li, J. (2014). A national‐scale assessment of long‐term water budget closures for Canada's watersheds. Journal of Geophysical Research: Atmospheres, 119(14), 8712-8725. https://doi.org/10.1002/2014JD021951

Wang, Z., Zhong, R., Lai, C., & Chen, J. (2017). Evaluation of the GPM IMERG satellite-based precipitation products and the hydrological utility. Atmospheric Research, 196, 151-163. https://doi.org/10.1016/j.atmosres.2017.06.020

Wei, L., Jiang, S., Ren, L., Yuan, F., & Zhang, L. (2019). Performance of two long-term satellite-based and GPCC 8.0 precipitation products for drought monitoring over the Yellow River Basin in China. Sustainability, 11(18), 4969. https://doi.org/10.3390/su11184969

Zhang, Y., Pan, M., Sheffield, J., Siemann, A. L., Fisher, C. K., Liang, M., . . . Houser, P. R. (2018). A Climate Data Record (CDR) for the global terrestrial water budget: 1984–2010. Hydrology and Earth System Sciences, 22(1), 241-263. https://doi.org/10.5194/hess-22-241-2018