Response to Reviewer RC2

Title: Achieving water budget closure through physical hydrological processes modelling: insights from a large-sample study
Authors: Xudong Zheng, Dengfeng Liu*, Shengzhi Huang*, Hao Wang, Xianmeng Meng
Manuscript ID: hess-2024-230

Reply on RC2:

Thank you very much for dedicating your time and effort to reviewing our paper. All comments from Reviewer RC2 are addressed below with point-by-point responses.

For better readability, replies will start with "**R/**", following the original comments that start with "**C/**" and are shown in **bold**. The revisions to be added into the revised manuscript is highlighted in red. The important parts are highlighted in blue. The quoted content is displayed in *italics*.

Point-to-point response:

C/ The paper presents an interesting concept, and its organization and writing are well done. However, I have some differing views regarding the underlying assumptions and principles of the proposed method. My main comments are as follows:

R/ First and foremost, we sincerely appreciate your interest in the concept shared in our paper, as well as your kind recognition of our writing and organization. We hold your constructive comments in high regard and believe it will be instrumental in enhancing the quality of our paper. These comments will be addressed point by point below, and revisions will be made in the manuscript to the best of our ability.

Major Comments:

C/ (1) I do not agree with the two underlying assumptions of the PHPM-MDCF method, nor with the significance of using Equation 4 to calculate omission errors. My main reasons are as follows:

Firstly, the errors in hydrological models are non-negligible and represent the sum of both omission errors and data errors, rather than omission errors alone. The paper assumes that hydrological models have no data errors (inconsistency errors) and only omission errors, which is evidently unreasonable. This assumption is particularly problematic because hydrological models are typically validated against observed runoff, often neglecting the validation of ET (Evapotranspiration) and TWSC (Terrestrial Water Storage Change) simulation accuracy. As a result, using Equation 4 to calculate omission errors is not justified. Due to the complexity of hydrological models and the impact of errors in driving variables, the water imbalance caused by errors in the hydrological model may be substantial. Even if the inputs to the hydrological model are observational data and the model itself is developed based on the principle of water budget, the primary contributor to water imbalance errors between input and output might still be data errors.

Secondly, the total residual is calculated using multiple sources of data, and omission errors are calculated using data that drive the hydrological model as per Equation 4. The difference between these is then used to calculate data inconsistency errors. However, this approach might introduce uncertainties due to data inconsistency.

R/ Thank you for your comment. We acknowledge that employing hydrological models to constrain measurements and thereby enhance water budget closure among them is an ambitious idea, as it has not been previously presented in the literature. We also recognize that accepting this idea is challenging. However, this idea is not proposed arbitrarily; rather, it is developed progressively along a specific logical path.

First, the errors in hydrological model that we describe as ignorable refer to inconsistencies occurring within the input, output, and state, rather than those between measurements. This distinction is important to emphasize. In other words, each variable in Eq. (4) originates from the model itself, and from this perspective, these variables are independent of measurements. Such consistency in hydrological model has been described in numerous studies. For example, DeChant and Moradkhani, (2014) provided reduced structural equations for general distributed hydrological models from a state-space view:

$$s_{i,t} = f(x_{i,t}, s_{i,t-1}, \theta_i), \tag{R1}$$

where $f()$ represents the model structure, $x_{i,t}$ is the forcing of the ith grid at time t. θ_i is the parameter of the ith grid. In this equation, a quantitative balance is maintained between the input/forcing and output/state variables. In the general hydrological models, whether distributed or lumped, water balance serves as a fundamental governing equation to constrain the model, which is a well-established practice (Beven., 2001). The above constitutes the logical basis for our assumption that the hydrological model satisfies water balance, ensuring physical consistency. This also aligns with our definition of inconsistency residuals, which refer to non-closure arising from physical inconsistency.

However, given our current understanding of the water cycle, Eq. (4) may still be prone to omission residuals. It can be challenging to be aware of all water components, certain omissive components result in omission residuals. This portion of the residuals can be identified through variables derived from the hydrological model, as these variables are consistent with water balance.

In extreme cases, if all components are considered in water budget equation, the omission residual can be reduced to zero. At this point, no water imbalance exists within the simulation system (i.e., Eq. (4)), and any remaining residuals in the measurement system would be the potential inconsistency residual.

Return to your question, the "data errors" you refer to are more likely the differences between simulated and measured values (e.g., simulated versus gauged runoff). This pertains to model performance, specifically whether the model can accurately represent hydrological process. This does not conflict with the water balance feature of the model itself. It is important to emphasize once again that all variables used in Eq. (4) are derived from the model, not from measurements.

I hope the above response provides some clarity on the issues related to water balance in the hydrological model and the potential neglect of inconsistency residuals in Eq. (4). In addition, we would like to further address the question of the relationship between measurements and simulations in this method. We believe that clarifying this point may help address your concerns.

In the PHPM-MDCF method, measurements are used not only calculate the total residuals (i.e., Eq. (5)), but also to constrain the model through a multi-objective calibration process (i.e., tuning parameters). As you emphasized, using only observed runoff to validate the model is insufficient. In this work, we considered five different variables—streamflow, ET, SMS (soil moisture storage), GRS (groundwater reservoir storage), and SWE—to validate the performance of the model. After model performance evaluation, we selected 475 basins with reliable simulation for all variables for subsequent analysis. The first paragraph of Sect. 4.1 and Appendix C provide detailed information. We present the main information here:

"To ensure the robustness of the results, as mentioned previously, it is essential that hydrological model reliably represent hydrological processes. With reference to previous studies (Clark et al., 2021), we have adopted KGE≥-0.41 and r statistically significant at the 5% level as criteria for guaranteeing reliable simulations. The multi-objective simulation performances of the HBV model are detailed in Appendix C. In general, the majority of basins (475, accounting for 72.24% of the total basins) achieved reliable simulations across all variables."

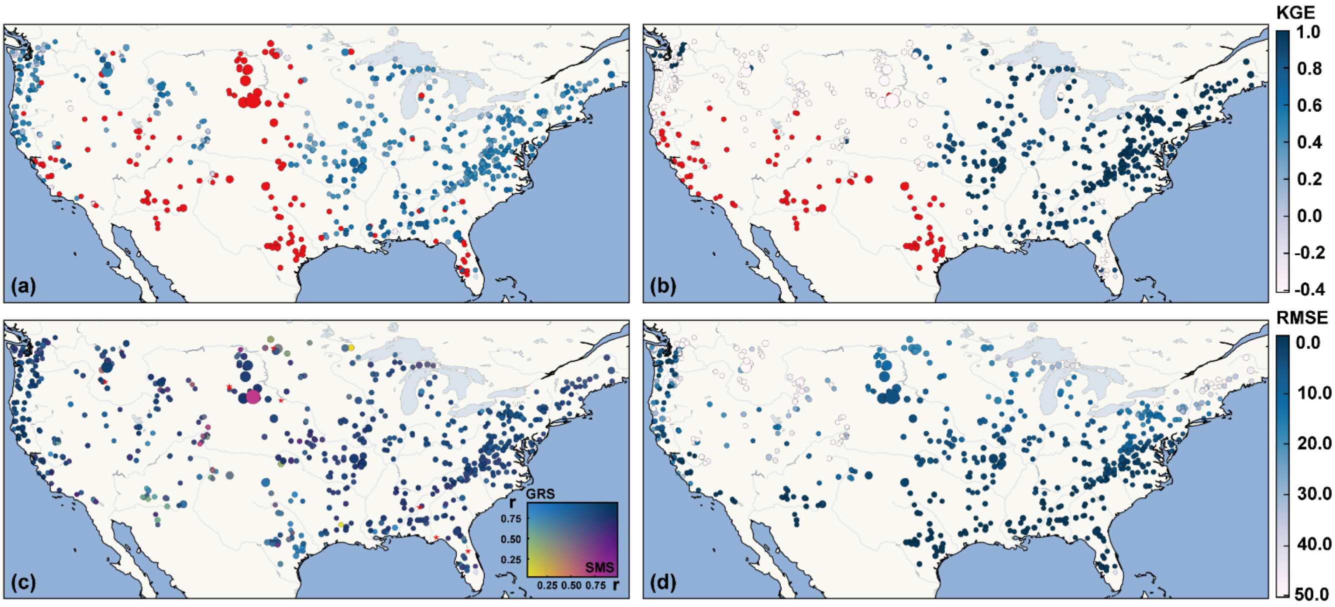

Figure C1. The multi-objective simulation performances of the HBV model across the CAMELS basins. Results are based on (a) runoff, (b) evaporation, (c) soil moisture storage and groundwater reservoir storage, and (d) snow water equivalent. Red dots represent unreliable simulation performance, and the size of points is proportional to the basin area. The unit of RMSE is "mm".

In general, this helps ensure simulation accuracy to some extent and reduces the uncertainty in the residual decomposition. Furthermore, the multi-objection calibration process is repeatedly applied during multisource datasets correction to ensure that, after each iteration of data correction, the model can produce reliable simulations corresponding to the dataset.

Based on the response to this concern, we recognize the importance of further emphasizing the water balance assumption in hydrological model used in this method, particularly with respect to Eq. (4). Therefore, we will add the following statements to the manuscript (Sect. 3.1):

"It is crucial to clarify that all variables in Eq. (4) are derived from the model itself, rather than from measurement, and can therefore be considered physically consistent."

C/ (2) The validation of results should include a comparison between the PHPM-MDCF method and existing methods. The paper repeatedly emphasizes the inadequacy of current methods in distributing residuals, yet no comparison with existing methods is provided in the results to verify the accuracy of the PHPM-MDCF method. The goal of closing the water budget is to reduce residuals while improving the accuracy of water cycle variables. Therefore, the credibility of the model should not be judged solely by the reduction of residuals (Figure 6). A comparison with existing methods would be more convincing. I strongly recommend supplementing the results with a comparison against existing correction methods, particularly CKF, PR, and MCL methods. For instance, the accuracy of the datasets after calibration using these methods, including P (Precipitation), ET (Evapotranspiration), Q (Runoff), and TWSC (Terrestrial Water Storage Change).

R/ Your point is very logical and intuitive. The introduction of any new method inevitably involves comparison with existing methods, which was also one of our initial objectives. However, after a thorough process of reflection and analysis, we have found that a direct comparison with existing methods is either infeasible or not meaningful for the following reasons:

(a) Difference in underlying logic.

The PHPM-MDCF exhibits a fundamental difference from existing methods, particularly in its interpretation of the realism. In existing methods, such as CKF, PR, or MCL, data correction relies on an assumed “true value” as reference. This true value might be the ensemble mean of multiple products or a set of gauged observations considered more credible. In other words, they assume that this true value can represent reality. However, this assumption is often challenged by issues such as scale mismatches and systematic biases in products. Although this approach is a common practice, but in our opinion, this notion of realism maybe untenable.

As an alternative, the realism of the PHPM-MDCF is reflected in our understanding of physical processes. Throughout the data correction process, the physical hydrological processes represented by the hydrological model play a central role. They act as constraints, iteratively correcting measurements into a physically consistent system. Although the reality represented by the hydrological model remains an abstraction, the iterative coupling of information in measurement through parameter calibration enhances the confidence in this representation. This approach embodies the underlying Bayesian philosophy. In addition, the pre-selection of basins with reliable simulation reduces the uncertainty associated with this method.

Due to the differences in these notions of realism, comparing these methods appears to be of limited value. The former typically aims to correct to an assumed “true value”, while the PHPM-MDCF focuses on correcting measurements to the physically hydrological processes represented by the model.

(b) Lack of real true values for reference.

As noted, the differences in realism among the methods make direct comparison challenging. Thus, the question arises whether an objective true value can be obtained as a benchmark for comparing the accuracy of different correction method, as mentioned in the comment. The answer is no. As we discussed in the induction:

“the fact remains that the ‘true value’ is perpetually unattainable, rendering any form of reference

data uncertain"

Since water budget non-closure study typically focus on datasets from different sources that estimate across varying scales, even when using field observations as the reference, there are challenges with scale mismatches. Additionally, even without scale mismatch issues, acquiring such data across extensive spatial and temporal scales remains a significant challenge.

(c) Different understanding of the relationships between measurements of different variables.

Another reason we cannot directly compare existing methods with PHPM-MDCF is the difference in their understanding of the relationships between observations of different variables.

In the data fusion based correction methods (e.g. CKF, PR, MCL), the physical connections between different variables seem to be overlooked. Or more cautiously, these relationships are not explicitly utilized as constraints for data correction. In such cases, although residuals can be constrained to zero, the correction process might disrupt the physical connections between variables, leading to unreasonable adjustments. In contrast, the PHPM-MDCF leverages these relationships, as represented by hydrological model, to constrain the measurements.

This difference ultimately reflects in the correction results. As noted by Luo et al. (2023), correction may lead to a decrease in the accuracy of individual variables:

"therefore, the results confirm that increasing the water budget closure accuracy of budget-component data sets reduces the accuracy of individual budget-component products."

For the above reason, we think the direct comparison between our method and existing methods is not meaningful, as their correction direction are fundamentally different.

Although we cannot conduct a direct comparison, but a theoretical indirect analysis is possible. The noise experiment in Sect. 4.3.2 can provide such an indirect analysis.

When extreme single-point noise is present in streamflow measurement (NS1 and NS2), it is expected that, to ensure water balance closure, existing correction methods will impose constraints across all variable by referencing "true values". Typically, streamflow measurements are considered to have the least uncertainty, leading to the smallest correction. As a result, extreme bias in streamflow can propagate to other variables by correction process, such as ET and TWSC. This is also the reason why the correction process, as previously discussed, can lead to a reduction in the accuracy of individual variables.

Figures 9 and S9 indicate that the PHPM-MDCF can effectively reduce residuals without causing such bias to propagate across different variables, thereby avoiding the aforementioned issues. This indirect analysis also provides some explanation for the differences between PHPM-MDCF and existing methods, and, to some extent, supports its reliability.

Responding to your comment has stimulated further reflection on our part. This is highly valuable, and we will make the following revisions according these reflection:

(a) We will further emphasize the issues of scale mismatch and the availability of site data in the induction:

"The issue of scale mismatches and the availability of site data in certain regions also pose challenges for data evaluation."

(b) We will include the reference by Luo et al. (2023) to strengthen the expression of our viewpoints:

"In the context of applying such closure constraint, it becomes evident that the precision of certain individual components may notably deteriorate, particularly when uncertainties are challenging to quantify (Luo et al., 2023)."

C/ (3) The description of the reference datasets is unclear. It is necessary to specify which observational system datasets were used for P (Precipitation), ET (Evapotranspiration), Q (Runoff), and TWSC (Terrestrial Water Storage Change), and why these datasets can be considered observational data. I recommend clarifying this in the text.

R/ Thank you for your suggestion. We will revise Table 1 in accordance with your suggestions and provide the explanation for the selection of these datasets for each variable. Here is the revised version:

"Specifically, daily precipitation estimation derived from the Tropical Rainfall Measuring Mission (TRMM 3B42V7) is used in this study. The well-known international NASA project aims to comprehensively estimate all forms of precipitation, including rain, drizzle, snow, graupel, and hail, through the integration of satellite data and ground-based rain gauge measurements (Huffman et al., 2016). The accuracy of TRMM dataset has validated by many studies through comparisons with observation data and other reanalysis datasets (Kittel et al., 2018; Villarini et al., 2009). For evaporation, we utilized the third version of Global Land Evaporation Amsterdam Model (GLEAM v3) product (https://www.gleam.eu/), which employs a set of algorithms to separately estimate the different components of land evaporation (Miralles et al., 2011). Several studies have demonstrated that this product aligns well with flux measurements and multisource product ensemble (Munier et al., 2014; Robinson and Clark, 2020). And, as mentioned above, the runoff measurements on a basin scale are provided by the CAMELS dataset, which is derived from site observations."

Table 1. Overview of the products for constructing water balance equation used in this study.

Variable	Product	Original Resolution		Original Period	Reference
		Spatial	Temporal		
Precipitation	TRMM 3B42V7	0.25 °×0.25 °	Daily	1998-2019	*Huffman et al. (2016)*
Evaporation	GLEAM v3.8a	0.25 °×0.25 °	Daily	1980-2022	*Martens et al. (2017)*
Soil moisture layer 1/2/3/4	EAR5 Land	0.1 °×0.1 °	Hourly	1950-present	*Mu ñoz Sabater et al. (2021)*
Snow water equivalent	GlobSnow v3.0	25km×25km	Daily	1979-2018	*Luojus et al. (2021)*
Streamflow	CAMELS-USGS	Basin scale	Daily	1980-2010	*Newman et al. (2015)*

C/ (4) Only a single product was selected for each water cycle variable. I believe that selecting multiple products is crucial for validating the proposed PHPM-MDCF method. This is because different datasets have different sources of error, leading to varying inconsistency residuals depending on the data combination. If the proposed method can be used to identify inconsistency residual error, using multiple data combinations would better verify the reliability of the proposed

method in this study.

R/ Thank you for your comment. We acknowledge that a common practice in previous water budget assessments is to use a range of products for each water components, evaluating the availability of different product combinations to closure the water budget. For example, Lorenz et al. (2014) compared 180 combinations of datasets for P, ET, TWS, and Q to access the degree of atmospheric-land water balance achieved. Lehmann et al. (2022) investigated the budget closure at catchment scales using 11 P, 14 ET, and 11 Q datasets together with GRACE.

However, almost all similar studies have reached the same conclusion that no single combination can close the water budget well across all regions (Lv et al., 2017). This implies that while introducing multiple products for ranking may be meaningful for specific regions, it holds limited significance for the correction framework of this study, which focuses on broader spatial scales (large sample basins). As Petch et al. (2023) handled in their optimization-based correction method, a single product was used for each water budget component, and they emphasize:

"In this study, we use only a single data product for each component, which we account for in our uncertainty calculations. We aimed to use Earth observation data where possible and sought global gridded products to ensure the uniformity of the uncertainties across all basins."

"Overall, the specific datasets chosen were not critical, as our primary goal was to evaluate our new optimisation methodology and its ability to bring independent products into consistency."

In addition, different products process varying spatiotemporal scales and have regional applicability, incorporating additional product may introduce further uncertainty.

A possible realization in the current study is to use different precipitation datasets (i.e., TRMM and Daymet datasets) to force the hydrological model and conduct correction, which has been implemented in Sect. 5.2.1. The results indicated that the correction is not sensitive to the choice of precipitation data.

"*In summary, the above results suggest that the correction is minimally sensitive to the choice of forcing, demonstrating the robustness of the correction results.*"

For the reasons mentioned above, we think that introducing additional products in the current study may not be necessary. However, we look forward to applying more models and datasets in future research to further extend the framework.

C/ (5) In Step 2 at line 250, please explain why is it reasonable to allocate residuals based on the difference between simulated values and reference values? It is worth noting that the simulated ET (Evapotranspiration) and TWSC (Terrestrial Water Storage Change) by the hydrological model may not have been validated for accuracy and may contain significant uncertainties. If their errors are used to allocate residuals, substantial uncertainties could lead to unreasonable allocation of residuals to ET and TWSC. The formula for residual allocation needs to be supplemented. Additionally, if Step 3 determines that the residual allocation is unreasonable, can simply halving the residual solve the issue? The underlying principles need to be clarified, or an example should be provided.

R/ Thank you for your careful review. For clarity, we have reorganized the questions in this comment and will analyze them individually.

(a) Why allocate residuals based on the distance between measurements and simulations?

As we discussed earlier, in this study, the simulations from the hydrological model are considered a physically consistent system that satisfies the water balance (See the reply to major concern (1)). Therefore, the Eq. (4) based on the simulations inevitably leads to Res_i being 0. In other words, when all measurements are corrected to equal the simulations, the Res_i in the measurements are corrected to 0. This determines the correction direction for measurements of each variable.

However, directly correcting the measurements to equal the simulation at once can also introduce uncertainty, as the simulation system is not precise (i.e., model parameters). Therefore, we considered an iterative approach for correction.

From the perspective of hydrological processes, the simulations reflect an ideal system that is physically consistent and strongly physically interrelated. On the contrary, the measurements reflect a system that variables are relatively loosely connected and physically inconsistent. To facilitate the convergence of the measurement system towards the ideal simulation system, it is important to determine the relative magnitude of the corrections for each water component.

The different water components cannot be corrected to the same extent, as their physical connections must be taken into account. For example, consider a region with high evaporation and low streamflow. Typically, it is reasonable to apply more correction to evaporation. However, if measurement of streamflow exhibits extreme high values, it would be more reasonable to apply more correction to streamflow. This is because our understanding of hydrological process suggests that the likelihood of such extreme high streamflow in this region is very low. Such understanding is reflected in the hydrological process, that is, in the simulations. Given this, we allocate the correction of Res_i based on the distance between measurements and simulations. In other words, the greater the distance between the measurement and the expected values, the more correction we will apply. This idea is illustrated in Fig. 3.

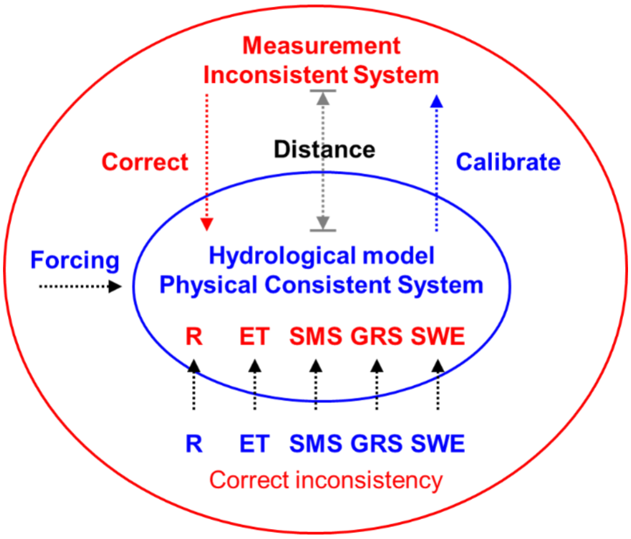

Figure 3. Illustration of the correction process advancing convergence between the simulation and measurement systems.

To better assist readers in understanding this idea, we will revise the statement in Step2 to:

"Step 2: Correction for the inconsistency residuals. Allocate inconsistency residuals based on the magnitude of differences (i.e., the distance between simulation and measurement systems) between simulated and measured values for each variable in Eq. (5) and (6). This is because this difference indicates the correction direction and magnitude for each variable, thereby facilitating the convergence of the measurement system towards the simulation system. Here, an initial correction rate of 0.5 is set to gradually correct the multisource datasets, thereby avoiding potential uncertainties that arise from excessive correction. Formally, the allocation of inconsistency residuals can be described by the following equation.

$$M_c^v = M_o^v - Res_i \times \frac{d_v}{d_{all}}, \tag{7}$$

where M_c^v is the measurements after correction of variable v, and M_o^v is the original measurements; d_v is the difference between simulation and measurement of variable v, and d_{all} represents the aggregate of differences for all variables."

(b) Were the simulations of ET and TWSC validated?

Yes, we validated the simulation results across five variables (i.e., streamflow, ET, SMS, GRS, and SWE) to ensure reliable simulations, where the SMS and GRS are used to represent TWS. We have provided a detailed explanation in our response to Concern (1) above. Through model performance evaluation, we have ensured that all basins undergoing multisource dataset correction exhibit reliable simulation. Additionally, the simulation performance has significantly improved after correction, as evidenced by the changes in the Pareto front shown in Fig. 8.

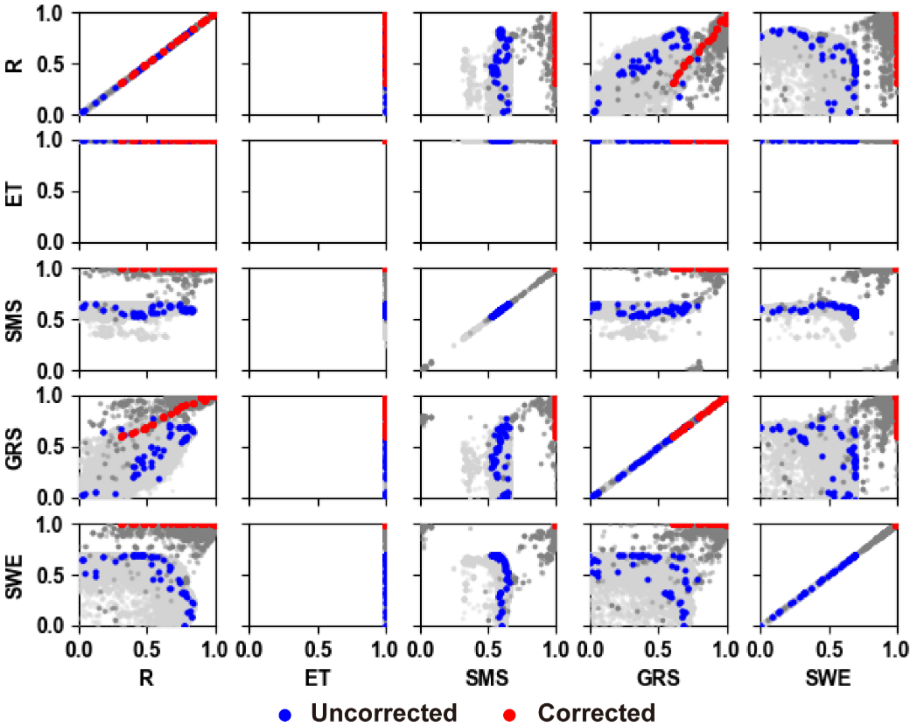

Figure 8. Comparison of multivariable simulation performance before and after correction at basin 1013500. Light grey and dark grey indicate population solution sets before and after correction, and blue and red indicate Pareto fronts before and after correction. Metrics evaluating SWE simulation performance have been normalized for consistency. The subplot in the second row, second column shows that the evaporation simulation maintains highly accurate at this basin, due to the alignment between the HBV algorithm and measurements.

(c) Supplement the residual allocation formula.

Thank you for pointing out this. According tor your suggestion, we will add the corresponding formula as shown blow.

"Formally, the allocation of inconsistency residuals can be described by the following equation.

$$M_c^v = M_o^v - Res_i \times \frac{d_v}{d_{all}}, \tag{7}$$

where M_c^v is the measurements after correction of variable v, and M_o^v is the original measurements; d_v is the difference between simulation and measurement of variable v, and d_{all} represents the aggregate of differences for all variables."

(d) If Step 3 determines that the residual allocation is unreasonable, can simply halving the residual solve the issue? What is the principle behind this?

In Step 3, a judgment will be made to determine whether the previous correction was reasonable based on whether the model can provide a reliable simulation. A misunderstanding that needs to be clarified here is that if the simulation proves unreliable, we will discard the previous correction, return to Step 2, halve the correction rate rather than directly halving Res, and then proceed with the correction again. Naturally, after this correction, the judgment in Step 3 will be re-evaluated until the correction or inconsistency residual falls below a pre-set threshold.

In other words, this iterative process involves continual trial and error, with each error prompting us to approach the next correction more cautiously. The underlying consideration is that the convergence of the measurement system and the simulation system is a mutual process. Measurements approach the simulated system through correction, while the simulation system, through re-calibration after each correction, aligns more closely with the measurement system. As described in the process shown in Fig. 3 above. Excessive correction may lead to the measurement system going out of bounds, preventing further convergence of the two systems. Specifically, this manifests as producing unreliable simulations, and further model calibration will not enable the two system to converge.

We have noted that our expression might lead to misunderstandings; therefore, we will revise the phrasing in Step 3 to:

"Step 3: Calibration and evaluation of the model. Recalibrate and evaluate the hydrological model using the datasets corrected in the previous step to assess the reliability of this correction. If the recalibrated model yields unreliable simulations, consider this correction excessive, halve the correction rate, and repeat Step 2. Otherwise, maintain the correction rate and proceed with the next iteration of correction. The consideration behind this step is that excessive correction may lead to the measurement system going out of bounds, preventing further convergence of the two systems. In other words, the iterative process involves continual trial and error, with each error prompting us to approach the next correction more cautiously."

C/ (6) Please clearly state the scope and spatiotemporal scale of this study. Most studies investigate water budget closure at the monthly scale rather than the daily scale. Aside from data availability, I believe this is mainly due to larger data errors and the lag effect of hydrological processes at the daily scale. If this study focuses on water budget closure at the daily scale, how were these issues addressed?

R/ Your perspective is very insightful. As you commented, the scale of the water budget study is crucial. The water budget non-closure phenomenon exhibits different behaviors at varying spatial and temporal scales. It is widely recognized that achieving water budget closure is much easier at relatively larger spatial and temporal scales.

On the one hand, at lager temporal scales, the TWSC exert a smaller influence on water budget closure. In relatively long time periods, TWSC can be assumed to negligible, making precipitation approximately equal to the sum of streamflow and evaporation. This is a common assumption in water budget assessment studies when TWSC measurements are unavailable. For example, Weligamage et al. (2023) suggested a 10-year period during which changes in water storage were considered negligible. Other several studies suggested that TWSC can be disregarded at the annual scale (Cooper et al., 2011; Kauffeldt et al., 2013; Hoeltgebaum et al., 2023). On the other hand, at larger spatial scales, inter-basin water exchanges can be considered negligible (Lv et al., 2017). Therefore, in most previous studies, it has been more feasible to conduct water budget studies at larger spatial and temporal scales. Additionally, another important reason for the choice of a monthly scale in much of the prior research is the reliance on GRACE TWSC measurements, which are only available at this temporal resolution.

In this study, TWSC is represented by a combination of observed soil moisture storage (SMS), groundwater reservoir storage (SMS), and snow water equivalent (SWE), avoiding the resolution constraints of GRACE TWSC, thus can be conducted at a daily scale. This is detailed in Sect. 2.2, where the main information is as follows:

"Assuming that TWSC can be retrieved through a combination of different water storages, we obtained the four-layer soil moisture from ERA5 Land and Snow Water Equivalent (SWE) from GlobSnow to estimate overall TWSC. This approach has been implemented in the investigation of Hoeltgebaum and Dias (2023), yield a high consistency between estimated TWSC and GRACE observation (i.e., correlation coefficient exceeding 0.71). Another consideration in this method is that the decomposed TWSC products (i.e., soil moisture and SWE) can correspond to the results simulated by hydrological model, thereby allowing us to correct water budget residuals, as discussed later."

"Overall, all datasets were resampled to a daily time step, and then aggregated over basins through simple averaging to perform analysis of water budget closure on a basin scale."

Although the primary temporal scale of this study is daily, we also performed statistical analyses at monthly and annual scales. For example, Figure 4-5 aggregate the residuals to the monthly scale to illustrate their spatiotemporal distribution. Figure 6 displays the correction results at daily, monthly and annual scales. This was done for both of visualization purposes and facilitating potential comparisons with previous studies.

Through a comparison of water budget at different timescales, we observed distinct behaviors of residuals across these scales. Specifically, at smaller scale (daily), residuals show greater variability but smaller magnitudes. As aggregation occurs at lager scales (monthly and annual), the magnitude increase while the variability decreases, demonstrating a filtering behavior. The primary mechanism behind such behavior is the positive and negative offset and accumulation of residuals and biases in different water components. Figure 6 provides an example to illustrate this:

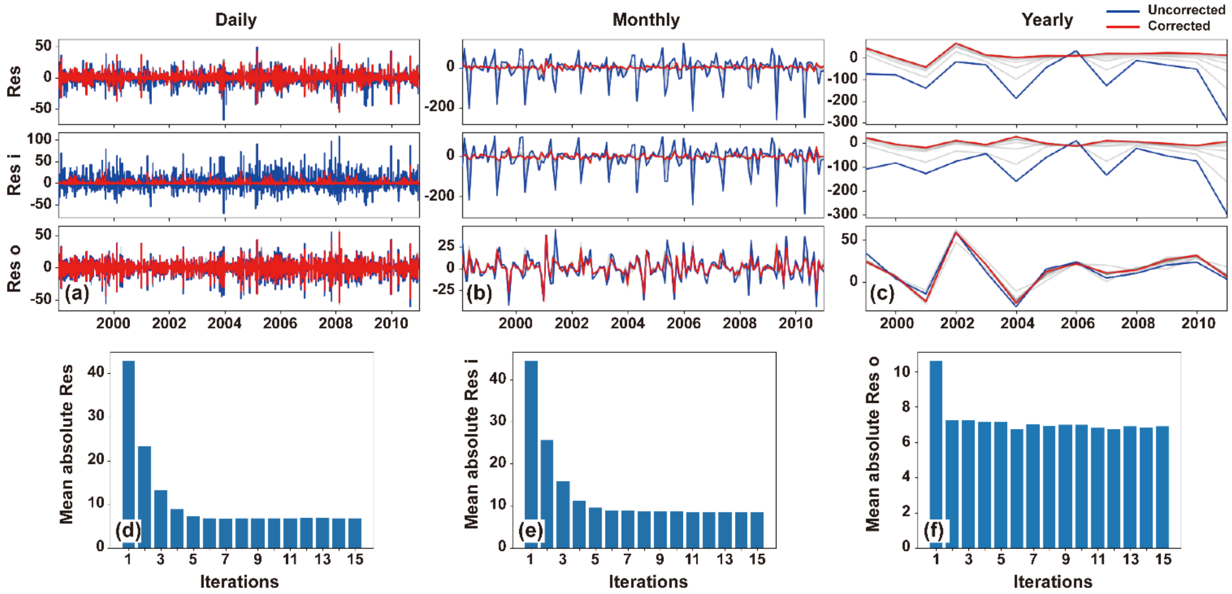

Figure 6. Correction results of water budget residuals for multisource datasets at basin 1013500. (a-c) Time series of water budget residuals (Res), inconsistency residuals (Res_i), and omission residuals (Res_o) at daily, monthly and yearly scales, grey line represents residuals during the correction process. (d-f) Variation of long-term mean absolute values of three residuals with correction iterations at the monthly scale. The unit of residuals is "mm".

According tor your comment, we will further emphasize the temporal scale used in this study by adding the following statements in Sect. 3.1 and 3.2:

"Therefore, residuals are calculated at daily scale and subsequently aggregated to the monthly and annual scales for further analysis."

"Notably, the correction is performed at the daily scale, aligning with the model step."

C/ (7) At line 320, it is necessary to explain the reasons behind the spatial distribution of Res. How does the difference in spatial patterns indicate that inconsistency residuals and omission residuals are driven by different factors? Please provide a detailed explanation. The most likely reason for Resi and Res having the same spatial pattern is that the former was calculated based on the latter. Their difference from Reso is due to the different error sources used in calculating Reso and Res, which does not necessarily demonstrate the reliability of the method for separating inconsistency residuals from omission residuals. Additionally, the residual values in Figure 4 differ significantly from those reported in previous studies. What is the reason for this discrepancy?

R/ Thank you for your comment. For clarity, we reorganized the questions in the comment into two separate points and address each one individually.

(a) **What are the reasons behind the spatial distribution of Res? Does its distribution show significant differences compared to previous studies? If so, what are the reasons for these differences?**

This is a good question. Indeed, as we discussed in our manuscript, the spatial distribution of Res in Fig. 4 exhibits very pronounced clustering characteristics.

“Res and Resi both present an east-west gradient for three statistical measures (i.e., min, median, max), with low values occur along the western coastline and high values primarily concentrated in eastern inland basins. The exception is a cluster of low median values located in the central CONUS”

From a geo-statistical perspective, the spatial heterogeneity of Res likely involves multiple direct and indirect influences from basin characteristics. Clarifying these potential influencing factors is crucial for understanding the formation of Res. Therefore, we conducted an exploratory analysis in Sect. 4.4 and found that Res is closely related to basin area and hydro-meteorological conditions. Specifically, we found that achieving water budget closure with multisource datasets is more challenging in larger and humid basins (characterized by high precipitation and runoff coefficient). Figure 11 provide the corresponding evidence.

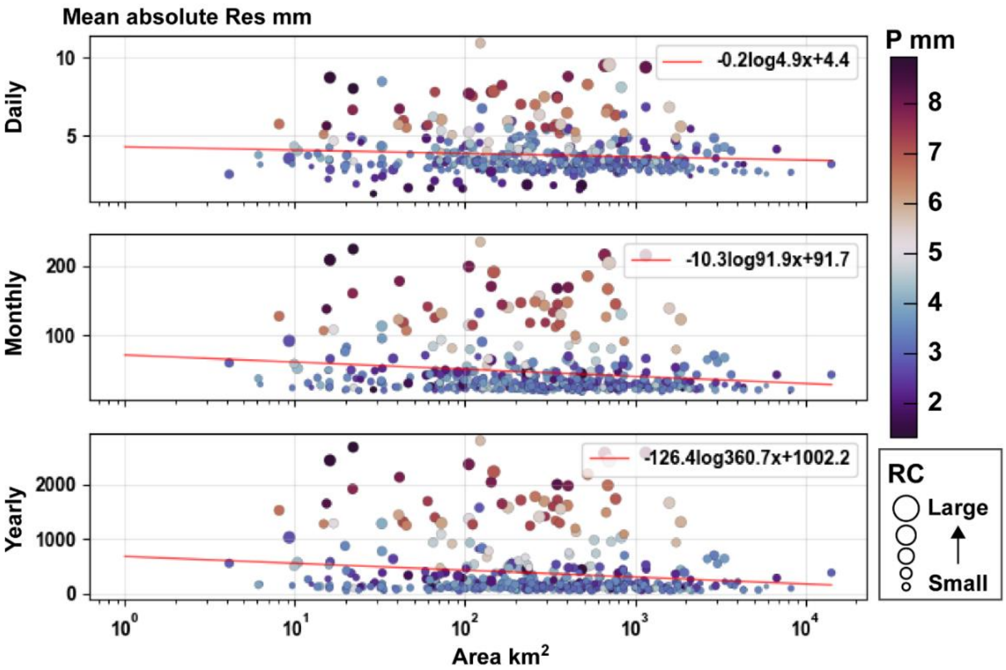

Figure 11. Relationship between the mean absolute of water budget residuals, basin area, long-term average daily precipitation, and runoff coefficient (RC) over 475 CAMELS basins with reliable simulations. The respective red lines represent the linear regression of residuals with basin area for each timescale.

Additionally, the comparison of the spatial distribution of Res with previous studies is also presented in Sect. 4.4. The results indicate that the pattern of Res identified in this study is consistent with previous research:

“As shown in Fig. 4, all three water budget residuals are subject to strong spatial organization, and these patterns are in agreement with previous studies. For example, Kauffeldt et al. (2013) found negative residuals (i.e., runoff coefficient > 1) along the western coastline of CONUS, while the eastern region showed notable positive residuals (i.e., P-R > ET). Other studies investigating water budget residuals with diverse dataset combinations have similarly revealed similar spatial patterns (Zhang et al., 2016; Gordon et al., 2022).”

We noticed a loose connection between Sect 4.1 and Sect 4.4; thus we will add the following statement in the former section to strengthen the linkage between the two sections:

“The potential factors affecting the spatiotemporal distribution of Res will be further investigated in

Sect. 4.4.”

Furthermore, we will divide Sect. 4.4 into three subsections to ensure a clear structure. The titles of the three subsections are:

“4.4.1 Factors influencing spatial distribution”
“4.4.2 Factors influencing temporal distribution”
“4.4.3 Factors influencing the proportions of residuals components”

(b) **Why are the differences between the spatial patterns of Resi and Reso driven by different factors? What is the theoretical basis for residual decomposition? How can the reliability of this decomposition be demonstrated?**

In previous studies, Res (water budget residuals) have typically been used as a whole to measure the degree to which the measurements achieve water budget closure. The cause of Res is often simply attributed to inconsistencies in the processing of different products (refer to the review provided by Lv et al., 2017). Few studies have thoroughly discussed the causes of Res formulation.

An exception is the study by Gordon et al., (2022), where they qualitatively decomposed Res into data inconsistency error (e) and groundwater exchange (G) not accounted for in the water budget equation (see Eq. (2)). We extended Eq. (2) to incorporate additional source of potential water omission, and further attempted a quantitative decomposition of Res into Res_i and Res_o to elucidate the distinct factors contributing to the observed water budget non-closure.

In our opinion, using measurements to describe the theoretical water balance requires two key conditions: (1) physically consistent measurements, and (2) comprehensive description of the water budget equation. Correspondingly, the causes of water budget non-closure ($|Res| > 0$) can be attributed to two factors: (1) physical inconsistency in the measurements (Res_i), potentially arising from discrepancies in data production process mentioned in previous studies; and the incomplete description of the water budget equation (Res_o).

Indeed, as you noted, the decomposition of Res is fundamentally based on the following sample equation, which capture the essence of our decomposition method:

$$Res_i = Res - Res_o \tag{R2}$$

However, the similar spatiotemporal distribution of Res_i and Res cannot be simply attributed the calculation. Essentially, this similar pattern is attributed to the relative small proportion of Res_o, suggesting that our description of the water budget equation is comparatively comprehensive.

Consider that if our description of the water budget equation were incomplete and omitted a significant water component, Res_o would likely exert a greater influence on Res, resulting in a more pronounced discrepancy between Res and Res_i.

To examine this, we intentionally exclude the SWE component from the water budget equation to access its impact on the decomposition of Res. This is a plausible scenario in practice, as it is likely that this component was not considered when reconstructing the TWSC. The results indicate that the proportion of Res_o obtained from residuals decomposition after excluding SWE increases significantly, with this effect being more pronounced in high-latitude regions, high elevations, and during the cold season (see the revisions and figure below). This is consistent with physical principles, as the impact of omitting SWE on water balance is greater under these situations. These findings align with our definition of Res_o which refers to the water imbalance caused by omitted water. It also, to some extent, supports the validity of our decomposition method, and highlights the importance of a comprehensive water budget equation.

Based on the response to this issue, in order to further demonstrate the reliability of the residual decomposition, we will add a new subsection in Sect. 4.4 to explain the potential factors for the proportion of Res components.

"4.4.3 Factors influencing the proportions of residuals components

Another interesting finding in Sect. 4.1 is that the magnitude of Res_o is significantly smaller than that of Res_i. As a result, Res is dominated by Res_i, leading to a highly consistent spatiotemporal distribution between them. However, the underlying question is what this implies and what factors drive the proportions of residuals components.

Res reflects the degree to which the measurements achieve water budget closure. In this study, we argue that two key conditions are necessary for using measurements to describe theoretical water balance. The first one is that measurements of different water components must be physically consistent. In practice, however, this condition is often challenging to meet due to inconsistencies and uncertainties in data production processes from different sources, which can result in non-zero Res_i (Luo et al., 2020). The second crucial, yet frequently overlooked, condition is the completeness of the water budget equation. Building on the work of Gordon et al. (2022), we developed a more generalized water budget equation (Eq. (3)) and use Res_o to account for the water imbalances caused by omitted water. From this perspective, Res results from the interplay between Res_i and Res_o, either through their accumulation or mutual cancellation. Therefore, the low proportion of Res_o essentially suggests that our description of the water budget equation is comparatively comprehensive.

Consider that if our description of the water budget equation were incomplete and omitted a significant water component, Res_o would likely exert a greater influence on Res, resulting in a more pronounced discrepancy between Res and Res_i. To examine this, we intentionally exclude the SWE component from the water budget equation to evaluate its impact on the decomposition of Res. This is a plausible scenario in practice, as it is likely that this component was not considered when reconstructing the TWSC. Figure 13 illustrates the comparison between Res_o derived from the decomposition method excluding SWE (hereafter Res_o^{NSWE}), and its original values. It is evident that Res_o^{NSWE} exhibits greater variability compare to the original values (i.e., with smaller minimum values and larger maximum values). The median differences indicate that the likelihood of increased omission residuals is higher after excluding SWE (Fig. 13b). Such differences indicate that omitting crucial SWE storage component results in a greater degree of water imbalance, and, as expected, this effect is more pronounce in high-latitude and high-elevation regions (Fig. 13d-f). Moreover, the spatiotemporal distribution of Res_o has changed (Fig. S11-12). Notably, during the cold season (December to February), the proportion of Res_o is much higher and exhibits s significant positive trend. These findings align with our definition of Res_o which refers to the water imbalance caused by omitted water. It also, to some extent, supports the validity of our decomposition method, and highlights the importance of a comprehensive water budget equation."

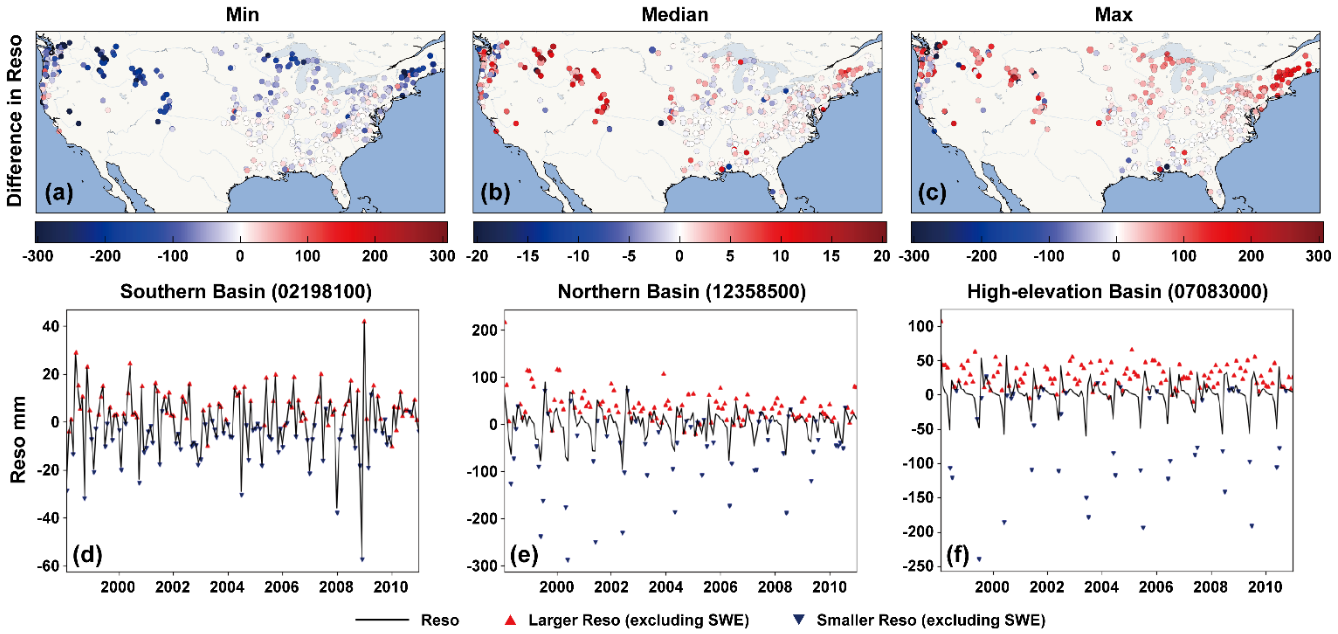

Figure 13. Comparison of Res_o obtained from residuals decomposition excluding SWE with the original values. (a-c) Spatial distribution of monthly mean Res_o excluding SWE minus its original values. (d-f) Time series of Res_o excluding SWE and its original values at the southern basin (02198100, 32.96 °N), northern basin (12358500, 48.33 °N), and high-elevation basin (07083000, elevation of 3.56 km) at monthly scale. The unit of residuals is “mm”

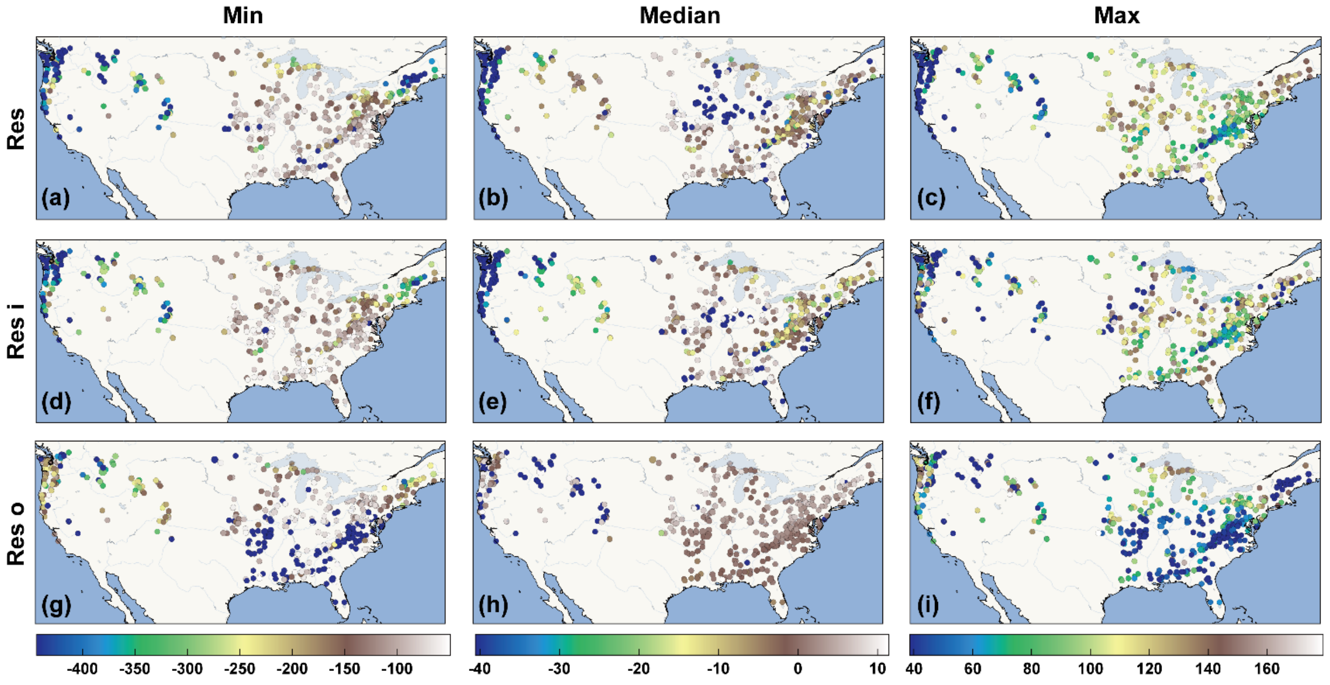

Figure S11. Same as Fig. 4, but for residuals decomposition excluding SWE

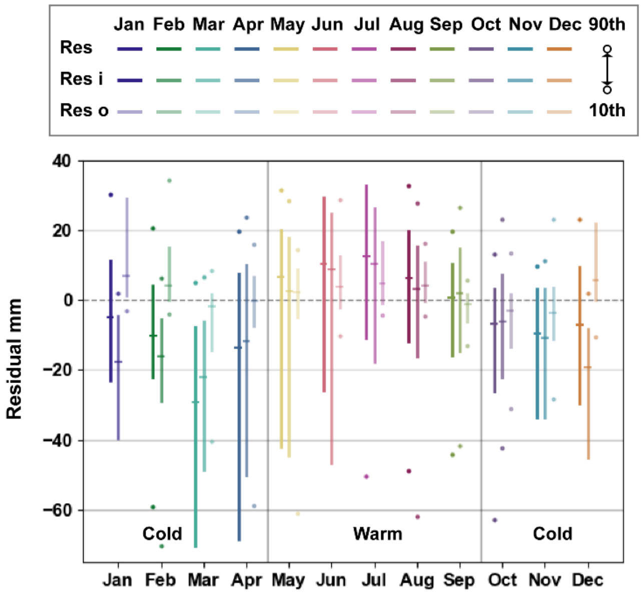

Figure S12. Same as Fig. 5, but for residuals decomposition excluding SWE

C/ (8) In the multi-source dataset correction framework for achieving water budget closure, what is the rationale for setting the initial correction rate to 0.5? Why is the correction rate halved when the model produces unreliable simulations? Is there a potential proportional relationship between the adjustment of the correction rate and the magnitude of bias in unreliable simulations that could allow for more efficient correction rate adjustments? Additionally, what is the basis for setting the conditions for iteration and termination of the correction process as "the inconsistency residuals decreases to 10% of its initial value or the correction rate falls below 4%"?

R/ This is a very insightful comment. What you mentioned are precisely three key issues we encountered during the implementation process. Just in our response to the fourth question in Major Concern (5), the iterative process involves continuous trial and error to prevent over-correction and ensure that measurement remain within the appropriate range.

The first issue is determining the initial correction rate (r_0). At the beginning, to ensure a high correction speed, we set the initial correction rate to 1 and 0.7. However, for most basins, this often resulted in measurements exceeding a reasonable range after the first iteration of the correction, leading to unreliable simulations and unreasonable corrected measurements. Through experimentation, we found that 0.5 is a suitable initial correction rate, as it ensures that the first iteration of the correction is effective in most cases.

The second key issue is determining the decay rate of correction rate (Δr) following the occurrence of unreliable simulations. The generation of unreliable simulations suggests that the current correction is excessive. Effectively reducing the correction magnitude and re-correcting may further facilitate the convergence of measurement system with the simulation system. Linear decay is a conventional approach, which aligns with our perception. For example, reducing the correction rate by 0.1 or 0.2 each time. However, testing has shown that such linear decay results in excessively long correction times, making the application of the PHPM-MDCF across a wide range of basins (i.e., 475 basins) difficult. On the other hand, exponential decay can cause the correction rate to quickly fall into a small value range, thereby reducing the correction efficiency. Given the above, we chose a multiplicative decay approach, where the correction rate is halved each time for re-correction. The results indicate that this approach is effective, as shown in the iterative process depicted in Figures 6 and S3-6. For illustration, we provide a case here:

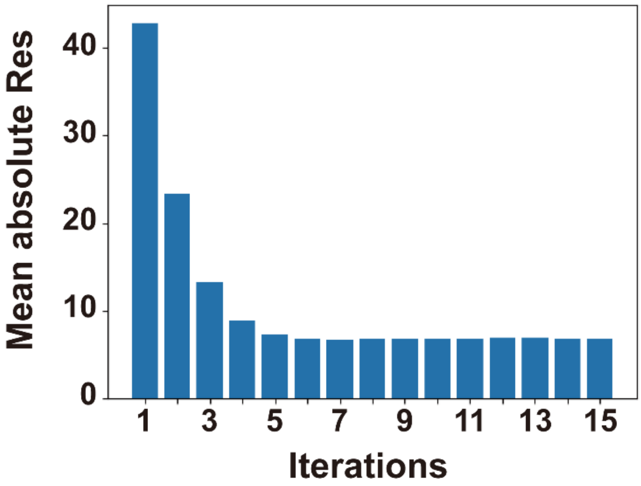

Figure R1. The decline of Res with the number of correction iterations for basin 1013500. The unit of residuals is "mm".

The final issue is determining when to terminate the correction, as this criterion significantly affects the final correction efficiency. Here we consider two points.

(a) The first is that the correction has achieved satisfactory results, with the final Res being relatively small (Res_t). This threshold must be appropriately set; it cannot be too large, as this would indicate insufficient correction, nor too small, since the PHPM-MDCF, as a soft constraint, has limited correction capacity. An excessively small final Res threshold could result in an infinite number of correction iteration. Based on comparative experiments, we believe that reducing it to 10% of the initial value is appropriate. As shown in Fig. R1, Res stabilizes and no longer changes once it decreases to around 10% of the initial value (from 40 to 4 mm).

(b) The second point is that the correction rate should not be too small, as this would imply excessively low calibration efficiency. This is closely related to the initial correction rate and decay rate (here, 0.5 and halving, respectively). A threshold of 4% means that the correction will cease once the correction rate, decayed four times from 0.5 to 0.03125, is reached. This threshold setting is relatively subjective, but it has proven to be reasonable based on testing results.

Notably, although the parameters for the three issues mentioned above are set subjectively, the choice follow a certain logic and have passed a series of tests. At least, cautiously speaking, they are suitable for the current study area, as shown in Fig. 7. Further adjustments are possible, but they have minimal impact on the current results (based on some testing).

We will add the following statement in Sect. 3.2 to further emphasize the issues mentioned above.

"In addition, the parameters settings in the PHPM-MDCF (i.e., initial correction rate, decay rate of the correction rate, correction termination threshold) are appropriate for the current study area (Table S2). When applying this framework to different regions, additional adjustments and testing may be required."

Table S2. Summary of the parameters settings in the PHPM-MDCF.

Parameters	Reference value	Reference range	Description
r_0	0.5	0.3~0.6	Initial correction rate.
Decay approach	Multiplicative	Linear, exponential, and multiplicative decay	The method of reduction in correction rate following an unreliable simulation.
Δr	50%	30%~70%	Decay rate of the correction rate.
Res_t	10%	5%~20%	Correction termination threshold for inconsistency residuals.
r_t	4%	1%~10%	Correction termination threshold for correction rate.

Minor Comments:

C/ (1) Please provide additional explanation on how Section 4.3.1 demonstrates the reliability of the PHPM-MDCF method.

R/ Thank you for your suggestion. We will add scatter plots comparing measurements and simulation before and after correction to further illustrate the convergence of the measurement and simulation systems, thereby demonstrating the reliability of the PMPH-MDCF method. The following revisions will be added to Section 4.3.1.

"More intuitively, Fig. S7 presents a comparison of measurements and simulations for each variable before and after correction. It is evident that the relationship between measurements and simulation is significantly strengthened after correction. This suggests that the PHPM-MDCF has the ability to enhance the convergence between the simulation and measurement systems, supporting the credibility of the correction results to some extent."

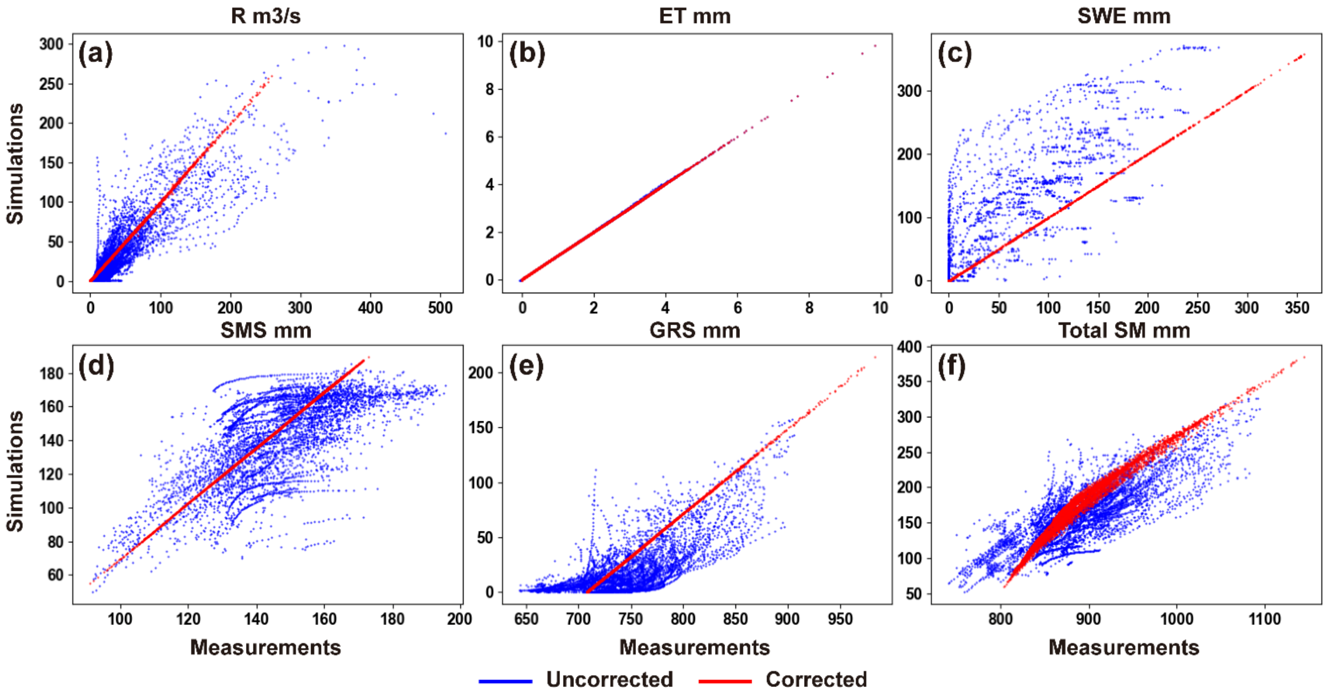

Figure S7. Scatter plots comparing measurements and simulation before and after correction at basin 1013500.

C/ (2) The paper does not validate the accuracy of the Reso, Resi, and Res separation method in the results.

R/ Thank you for your comment. We have addressed this issue in detail in our response to the second question of Major Concern (7) and will include a new subsection to demonstrate the reliability of the residuals decomposition method. Please review the response above.

C/ (3) At line 310, can KGE ≥ -0.41 really indicate that the hydrological model accurately represents the observed hydrological system?

R/ Thanks for your comment. The Kling-Gupta Efficiency (KGE) metric, introduced by Gupta et al. (2009), provides a method for achieving a balanced improvement of simulated mean, variability, and correlation (see Eq. B1). Many studies have demonstrated the effectiveness of KGE, which is currently a popular metric in hydrological modelling (Knoben et al., 2020; Clark et al., 2021). The KGE is bound by $(-\infty, 1]$ with 1 being the ideal value. For such a metric, it is challenging to give a benchmark value to determine whether the simulation is reliable. Thus, to ensure caution, we opt to reference previous literature for guidance. For instance, Aerts et al. (2022) use the -0.41 of KGE as the benchmark to evaluate the performance of wflow_sbm in simulating streamflow:

"Ideal model performance has a KGE score of 1 and a KGE score of −0.41 is equal to taking the mean flow as a benchmark."

Bruno et al. (2002) noted that a KGE of -0.41 serves as the threshold for no skill:
"(KGE ∈ (-∞, 1], optimal value = 1, no-skill threshold over mean flow as predictor = -0.41)."

The notable example is Knoben et al. (2019), who, by comparing the NSE and KGE metrics, established a KGE value of -0.41 as the threshold for evaluating whether model simulations outperform the mean flow:

"Here we show that using the mean flow as a predictor does not result in KGE = 0, but instead KGE =1- √2 ≈-0.41. Thus, KGE values greater than −0.41 indicate that a model improves upon the mean flow benchmark – even if the model's KGE value is negative. "

Based on the aforementioned literature, we used a KGE value greater than -0.41 as the threshold for reliable simulations. Although this threshold may still be somewhat subjective, evaluating simulation reliability across five variables (i.e., streamflow, ET, SMS, GRS, SWE) simultaneously can help mitigate this uncertainty.

For better address the question, we will include the above references in the manuscript.

"With reference to previous studies (Knoben et al. 2019; Clark et al., 2021; Aerts et al., 2022), we have adopted $\mathrm{KGE} \geq -0.41$ and r statistically significant at the 5% level as criteria for guaranteeing reliable simulations."

C/ (4) In Figure 5, Reso is closer to 0. Can we attribute this to the principle of water budget in the development of the hydrological model, rather than merely to omission errors? Since Resi = Res - Reso, and Reso is relatively small, it is evident that the values and spatial patterns of Resi and Res are more similar. What does this imply?

R/ Thank you for your comment. Our response to the second question of Major Concern (7) provides some clarification on this issue. Specifically, Res_o approaching zero indicates that our description of the water budget equation is relatively comprehensive and cannot be simply attributed to the water balance features of the hydrological model.

When the SWE component is omitted without changing the model, Res_o increases significantly, with this effect being more pronounced at high elevations, high latitudes, and during the cold season (Fig. 13).

The equation ($Res_i = Res - Res_o$) is indeed the essence of our decomposition method, but it is not the sole reason for the similarity between Res_i and Res. The fundamental reason lies in the completeness of the water budget equation description, which results in a smaller contribution of Res_o to the formation of Res.

C/ (5) Please explain from a theoretical standpoint why the PHPM-MDCF method has such advantages over previous methods: "It suggests that the soft constraints based on physical hydrological processes will not lead to compensatory errors, as seen in traditional methods due to the rigid allocation of water budget residuals.".

R/ Thank you for your suggestion. We will add the following statement to theoretically demonstrate the advantages of the PHPM-MDCF.

"From a theoretical perspective, the PHPM-MDCF assigns the weights of residual correction based on the distance between measurements and simulation for each variable. In the presence of a single extreme bias, the large distance between the measurement and simulation of the corresponding variable leads to a larger correction being applied to that variable, while the weights for other variables remain unaffected. However, in traditional methods, the correction weight for each variable remain constant over time, and the final residuals are constrained to zero. This leads to the propagation of extreme biases across different variables."

C/ (6) I do not find this statement reasonable: "When the hydrological model calibrated against multiple variables measured by the multisource datasets and achieves reliable performance, we consider the simulation system approaching the measurement system.".

R/ Thank you for your comment. We will revise this inappropriate statement to:

"When the hydrological model calibrated against multiple variables measured by the multisource datasets and achieves reliable performance, we consider the water budget represented by the simulation and measurement systems to be comparable."

C/ (7) At line 255, please clarify the data sources for the observed values of P, ET, Q, and TWSC used in this study. Without this information, it is difficult to judge whether the deviation between the simulation system and the measurement system is calculated reasonably.

R/ Thank you for pointing out the unclear aspects of our manuscript. According to your suggestion, we reiterated the data sources (see our response to Major Concern (3)) and will further emphasize them in this section as follows:

"In the subsequent application of the PHPM-MDCF, the measurements are derived from the data provided in Sect. 2.2."

C/ (8) I personally feel that the discussion in Section 5.1 would be more effective if it were more closely aligned with the scope of this study.

R/ Thank you for your suggestion. According to your suggestion, we will enhance Sect. 5.1 with more arguments relevant to this study and reduce unnecessary statements. The revisions will be made are as follows:

Remove this sentence from the penultimate paragraph: "Although our current knowledge may not be entirely precise—for example, the depiction of hydrological processes in hydrological models may lack accuracy—it remains foundation upon which we can rely and strive to refine in the future."

The last paragraph will be revised to: "The proposed correction framework (PHPM-MDCF) capitalizes on this concept by iteratively advancing the convergence between the knowledge system (i.e., hydrological model and water balance equation) and the measurement system, thus enhancing the credibility of the measurements. Although our current knowledge may not be entirely precise—for example, the depiction of hydrological processes in hydrological models may lack accuracy—it remains foundation upon which we can rely and strive to refine in the future. Furthermore, several underlying concepts in this framework, such as residuals decomposition and advancing water budget closure through correction, aligns with a recent study (Wang and Gupta, 2024). They introduced a novel hybrid model (i.e., Mass-Conserving-Perceptron) and discussed its potential application, including the bias correction (lacking confidence for the measurements) and examination of non-observed interactions with the environment (corresponding to the omission errors). Therefore, coupling the PHPM-MDCF with hydrological models that provide stronger interpretability is a valuable and promising research effort, as it can offer insights into the physical attribution of water budget non-closure and enable more reasonable correction."

C/ (9) The limitations discussed in Section 5.2 are not explained from a theoretical perspective. I hope that some convincing explanations can be supplemented from this standpoint.

R/ Thanks for your suggestion. We will add the following statement to Sect. 5.2 to further explain the theoretical basis of the adaptability to forcing datasets of the framework.

"Theoretically, the consistency of correction stems from two aspects. Firstly, it is attributed to the adaptability of hydrological model to the input data, specifically the calibration compensation capability we described in the introduction (Wang et al., 2023). This enables the hydrological model to generate reasonable representation of hydrological process even with imprecise forcing. Secondly, as discussed in Sect. 4.3.2, the PHPM-MDCF serves as a soft constraint and utilizes the distance between measurements and simulations to allocate residuals correction, thereby mitigating the propagation of bias between variables. These two features ensure that stability of the correction, rendering it less susceptible to interference from uncertainties in the forcing datasets."

C/ (10) The structure of the article lacks a keywords section. Please add keywords.

R/ Thank you for your careful review. According to the current HESS official template and guidelines, the keywords section is not a required option. Please the following URLs:

https://www.hydrology-and-earth-system-sciences.net/submission.html#templates
https://www.hydrology-and-earth-system-sciences.net/submission.html#manuscriptcomposition

C/ (11) Please add references related to the water budget equation.

R/ Thank you for pointing out the omissions in our manuscript. We will add the relevant reference (Lehmann et al., 2022) for Eq. 1 as:

"For a closed basin, the water budget can be mathematically expressed as (Lehmann et al., 2022),

$$\frac{\mathrm{dTWS}}{\mathrm{dt}} = \mathrm{P} - \mathrm{ET} - \mathrm{R}, \tag{1}$$

where $\frac{\mathrm{dTWS}}{\mathrm{dt}}$ is change in terrestrial water storage, P is precipitation, ET is evaporation, R is streamflow at the outlet."

C/ (12) The text states "as illustrated in Fig. 3" but the caption provided is "Figure 3". The authors should ensure that all figure captions are consistent with the text descriptions. Please carefully check the rest of the article for similar errors and make the necessary corrections.

R/ Thank you for your careful review. For the abbreviation format, we referred to the official guidelines provided by HESS. Please see the following URL and explanation:

https://www.hydrology-and-earth-system-sciences.net/submission.html#figurestables

"Figure composition: ...

...

The abbreviation 'Fig.' should be used when it appears in running text and should be followed by a number unless it comes at the beginning of a sentence, e.g.: "The results are depicted in Fig. 5. Figure 9 reveals that."

Reference

Abolafia-Rosenzweig, R., Pan, M., Zeng, J., and Livneh, B.: Remotely sensed ensembles of the terrestrial water budget over major global river basins: An assessment of three closure techniques, Remote Sensing of Environment, 252, 10.1016/j.rse.2020.112191, 2020.

Aerts, J., Hut, R., van de Giesen, N., Drost, N., Verseveld, W., Weerts, A., and Hazenberg, P.: Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model, Hydrology and Earth System Sciences, 26, 4407-4430, 10.5194/hess-26-4407-2022, 2022.

Beven, K.: How Far Can We Go in Distributed Hydrological Modeling, Hydrology and Earth System Sciences, 5, 10.5194/hess-5-1-2001, 2001.

Bruno, G., Duethmann, D., Avanzi, F., Alfieri, L., Libertino, A., and Gabellani, S.: Parameter transferability of a distributed hydrological model to droughts, 10.5194/hess-2022-416, 2022.

Clark, M., Lamontagne, J., Mizukami, N., Knoben, W., Tang, G., Gharari, S., Freer, J., Whitfield, P., Shook, K., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resources Research, 57, 10.1029/2020WR029001, 2021.

Cooper, R., Hodgkins, R., Wadham, J., and Tranter, M.: The hydrology of the proglacial zone of a high-Arctic glacier (Finsterwalderbreen, Svalbard): Sub-surface water fluxes and complete water budget, Journal of Hydrology, 406, 88-96, 10.1016/j.jhydrol.2011.06.008, 2011.

DeChant, C.M., Moradkhani, H.: Hydrologic Prediction and Uncertainty Quantification. In: Eslamian, S. (Ed.), Handbook of Engineering Hydrology: Modeling, Climate Change, and Variability. CRC Press, pp. 387–414, 2014.

Gordon, B., Crow, W., Konings, A., Dralle, D., and Harpold, A.: Can We Use the Water Budget to Infer Upland Catchment Behavior? The Role of Data Set Error Estimation and Interbasin Groundwater Flow, Water Resources Research, 58, 10.1029/2021WR030966, 2022.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hoeltgebaum, L. and Dias, N.: Evaluation of the storage and evapotranspiration terms of the water budget for an agricultural watershed using local and remote-sensing measurements, Agricultural and Forest Meteorology, 341, 10.1016/j.agrformet.2023.109615, 2023.

Huffman, G. J., D.T. Bolvin, E.J. Nelkin, a., and Adler, R. F.: TRMM (TMPA) Precipitation L3 1 day 0.25 degree x 0.25 degree V7 [dataset], 10.5067/TRMM/TMPA/DAY/7, 2016.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C. Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrology and Earth System Sciences, 17, 2845-2857, 2013.

Kittel, C. M. M., Nielsen, K., Tottrup, C., and Bauer-Gottwein, P.: Informing a hydrological model of the Ogooue with multi-mission remote sensing data, Hydrology and Earth System Sciences, 22, 1453-1472, 2018.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323–4331, 2019.

Knoben, W., Freer, J., Peel, M., Fowler, K., and Woods, R.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, Water Resources Research, 56, e2019WR025975, 10.1029/2019WR025975, 2020.

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

Lorenz, C., Kunstmann, H., Devaraju, B., Tourian, M., Sneeuw, N., and Riegger, J.: Large-Scale Runoff from Landmasses: A Global Assessment of the Closure of the Hydrological and Atmospheric Water Balances, Journal of Hydrometeorology, 15, 10.1175/JHM-D-13-0157.1, 2014.

Luo, Z., Li, H., Zhang, S., Wang, L., Wang, S., and Wang, L.: 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, 10.1029/2022WR032176, 2023.

Lv, M., Ma, Z., Yuan, X., Lv, M., Li, M., and Zheng, Z.: Water budget closure based on GRACE measurements and reconstructed evapotranspiration using GLDAS and water use data for two large densely-populated mid-latitude basins, Journal of Hydrology, 547, 10.1016/j.jhydrol.2017.02.027, 2017.

Miralles, D., Holmes, T., de Jeu, R., Gash, J., Meesters, A., and Dolman, H.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453-469, 10.5194/hess-15-453-2011, 2011.

Munier, S., Aires, F., Schlaffer, S., Prigent, C., Papa, F., Maisongrande, P., and Pan, M.: Combining datasets of satellite retrieved products for basin-scale water balance study. Part II: Evaluation on the Mississippi Basin and closure correction model, Journal of Geophysical Research, 10.1002/2014JD021953, 2014.

Petch, S., Dong, B., Quaife, T., King, R., and Haines, K.: Water and energy budgets over hydrological basins on short and long timescales, Hydrology and Earth System Sciences, 27, 1723-1744, 10.5194/hess-27-1723-2023, 2023.

Robinson, E. and Clark, D.: Using Gravity Recovery and Climate Experiment data to derive corrections to precipitation data sets and improve modelled snow mass at high latitudes, Hydrology and Earth System Sciences, 24, 1763-1779, 10.5194/hess-24-1763-2020, 2020.

Villarini, G., Krajewski, W., and Smith, J.: New paradigm for statistical validation of satellite precipitation estimates: Application to a large sample of the TMPA 0.25° 3-hourly estimates over Oklahoma, Journal of Geophysical Research, 114, 10.1029/2008JD011475, 2009.

Weligamage, H., Fowler, K., Peterson, T., Saft, M., Peel, M., and Ryu, D.: Partitioning of Precipitation Into Terrestrial Water Balance Components Under a Drying Climate, Water Resources Research, 59, 10.1029/2022WR033538, 2023.