Response to Referee #2

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 RC3:

First and foremost, I would like to express my sincere gratitude for your prompt reply and for the time and effort you have so generously devoted to reviewing our paper. We also greatly appreciate your recognition of the value of our work, as well as the opportunity you have given us to make revisions.

As you rightly pointed out, in our previous response, we primarily provided explanations for your concerns and conducted a few minor experiments, such as our response to Major Concern (7), Question b. This seems to have addressed your concerns to some extent, but we understand that it is insufficient. The lack of comparison with existing methods was mainly due to the two challenges: (1) first, finding data with the same time, spatial range, and appropriate resolution; (2) second, the time required to implement the existing methods.

After further interactive discussion with you, we recognized that this comparison is essential. Therefore, we sought to gather multiple sources of data (including site observations, remote sensing, and simulations) as much as possible, and implemented several existing correction methods (i.e., PR and CKF) to compare their correction with our results. This comparison was conducted in several representative basins (following your suggestion), which provides evidence for the reliability of our framework. We hope this experiment will address your concerns, and we also appreciate your valuable suggestions. The details of the comparison are provided in the point-by-point responses below.

Thank you again for your reply. We are also very pleased to engage in the academic discussion with you, which is highly meaningful. Below, we will provide a point-by-point reply to your comments.

Note:
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 author's approach of studying water balance closure from the perspective of physical mechanisms does indeed have academic value.

R/ Thank you very much for recognizing the value of our work; this is a great encouragement for us.

C/ However, the core issues I raised have not been fully addressed. The author mainly provided some explanations without offering experimental evidence to demonstrate the reliability of the proposed method.

R/ We sincerely apologize for having avoided addressing your concerns in our previous response. In this response, we have adopted your suggestions, given them careful consideration, and made every effort to conduct related experiments within the limited time available. Detailed experimental results are provided below, presented as a new subsection that will be added to the manuscript.

C/ I maintain that a comparison with existing methods is necessary to validate the accuracy and reliability of the proposed approach. The purpose of achieving water balance closure has two main components: improving data consistency and accuracy. Regarding data consistency, the author's method does not fully achieve water budget closure (I agree with the principle behind the author's approach). Therefore, if the method's performance cannot be verified in terms of data accuracy, its overall effectiveness and reliability remain questionable. I recommend that the author select some representative basins with measurements of budget components for validation.

R/ We are very pleased that you recognize the principles behind our methods, and we greatly appreciate the valuable suggestions you have provided. As you mentioned, further validating the calibration results through comparisons with existing methods can emphasize the reliability of our proposed approach. The approach of selecting representative basins for validation is also feasible, therefore we proceed with experiment in this regard. In this experiment, potential issues may include inconsistencies in temporal and spatial scales, as well as mismatches between grids and basins. Detailed results are provided in the responses below, presented as a new subsection that will be added to the manuscript.

C/ As for the author's claim that a comparison with existing methods is not appropriate, I disagree. Some current methods estimate the distribution weights of water imbalance based on fused values (some methods are not such as PR and MCL), rather than using the fused values as exact reference points. I recommend validating the proposed method by comparing it with existing methods based on in-situ measurements of budget components (in regions with in-situ measurements, such as P and Q). Additionally, considering multiple datasets for each hydrological variable would be beneficial for validating the proposed method. The author argues that errors in hydrological model simulations only represent physical inconsistency errors, while datasets capture comprehensive errors. If multiple datasets consistently identify omission errors, this would demonstrate the reliability of the method. I recommend that the author select some representative basins for validation.

R/ We acknowledge your perspective. Considering multi-source data for each hydrological component, along with comparisons of the corrected results from existing methods, will effectively demonstrate the reliability of our approach.

Therefore, we collected multisource datasets from in-situ observations, remote sensing retrievals, and model simulations. This includes 11 precipitation, 14 evaporation, 11 streamflow and 2 terrestrial water storage datasets (see Table S3). We have implemented two existing correction methods: the PR and CEnKF methods (Luo et al., 2023). A new subsection will be added to the manuscript to clarify the comparison between the PHPM-MDCF and existing methods (see below). In general, the comparison results from several representative basins indicate that the PHPM-MDCF can produce reliable correction results, reflected in several aspects: (1) a consistent over trend with existing method; (2) the absence of unreasonable corrections in streamflow; (3) the correction was also applied to TWSC (compared to CEnKF); and (4) a good consistency between the retrieved TWSC (from SM and SWE change) and GRAEC TWSC.

This comparison indeed further demonstrates the reliability of PHPM-MDCF, with detailed results presented below (in red). Due to time constraints, we have conducted experiments to the best of our ability. Therefore, it is worth mentioning that this comparison still includes potential uncertainty from scale and spatial mismatch issues.

Regardless, the PHPM-MDCF retains advantages in generating high-resolution corrections (daily), as it does not rely on multi-source datasets for the every variable but rather utilizes physical processes characterized by hydrological models as constraints. Theoretically, we can perform this correction at any model time step and for any model output variable.

“4.3.3 Comparison with existing correction methods

Previous analysis and experiments clarify the unique characteristics of the PHPM-MDCF, which impose closure constraints based on hydrological physical processes. This differs significantly from existing correction methods, such as PR and CEnKF (Luo et al., 2023). In this section, we conducted a comparison analysis with them to further evaluate the reliability of the PHPM-MDCF. To implement existing correction methods, support from multisource measurements for each water component is essential for calculating the residual allocation weights. Here, we obtained monthly datasets from Lehmann et al. (2022), which include 11 precipitation, 14 evaporation (ET), 11 streamflow (R) and 2 terrestrial water storage (TWS) datasets (Table S3). The datasets previously utilized in this study were also included for data fusion and correction (Table 1). In general, these datasets were processed to a uniform monthly scale and a common period (2003-2010), and subsequently aggregated to the basin scale. Several representative basins (numbered 1539000, 1557500, and 3070500) were selected to illustrate the differences between the PHPM-MDCF and existing methods, based on the spatial coverage of multisource datasets.

Figure 11 presents a comparison of the monthly correction results from three methods (i.e., PR, CEnKF, and PHPM-MDCF) for three main water budget components at basin 1539000. Note that the measurements of precipitation are not compared here, as the PHPM-MDCF does not perform correction for this variables. It is clear from the figure that both the PHPM-MDCF and CEnKF method exhibit minimal correction of ET, whereas the PR method significantly expands the range of ET, particularly increasing seasonal peaks. This arises from the assumption of the PR method that relative errors are proportional to the relative magnitudes of each variable (Abhishek et al., 2022). But in many cases, this assumption may not hold true.

In terms of the R and terrestrial water storage change (TWSC), the overall trends of the correction results from the three methods are generally consistent. However, the CEnKF appears to produce greater fluctuations in R (Fig. 11 b and e) and shows limited correction of TWSC. This is linked to the computational mechanism underlying CEnKF, where the Kalman gain—or the error covariance between measurements and the ensemble mean of multisource datasets—determines the magnitude of the residuals for each variable. The measurements of R to be corrected is based on in-situ obervations, while the multisource dataset includes model simulations and remote sensing values. Potential mismatches between the grids and basins may lead to significant discrepancies, resulting in an greater allocation of correction for R. On the contraty, measurements of TWSC are limited and primarilty deriving from GRACE, which results in relatively small error covariance and, consequentlt, smaller corrections. Furthermore, as previously noted, such method may generate unreasonable corrections due to propogation of extreme errors, such as the negative R values in Fig. 11b, which are more likely to occur in small basins. PHPM-MDCF avoids these issues by considering physical process constraints, leading to reasonable corrections. Additionally, it dose not rely on multisource datasets and can perform correction on a daily scale. The TWSC derived from SWE and SM is consistent with GRACE TWSC, which also demonstrates the reliability of this framework. The comparison results for the other 2 representative basins are shown in Fig. S11-12, leading to similar conclusions.”

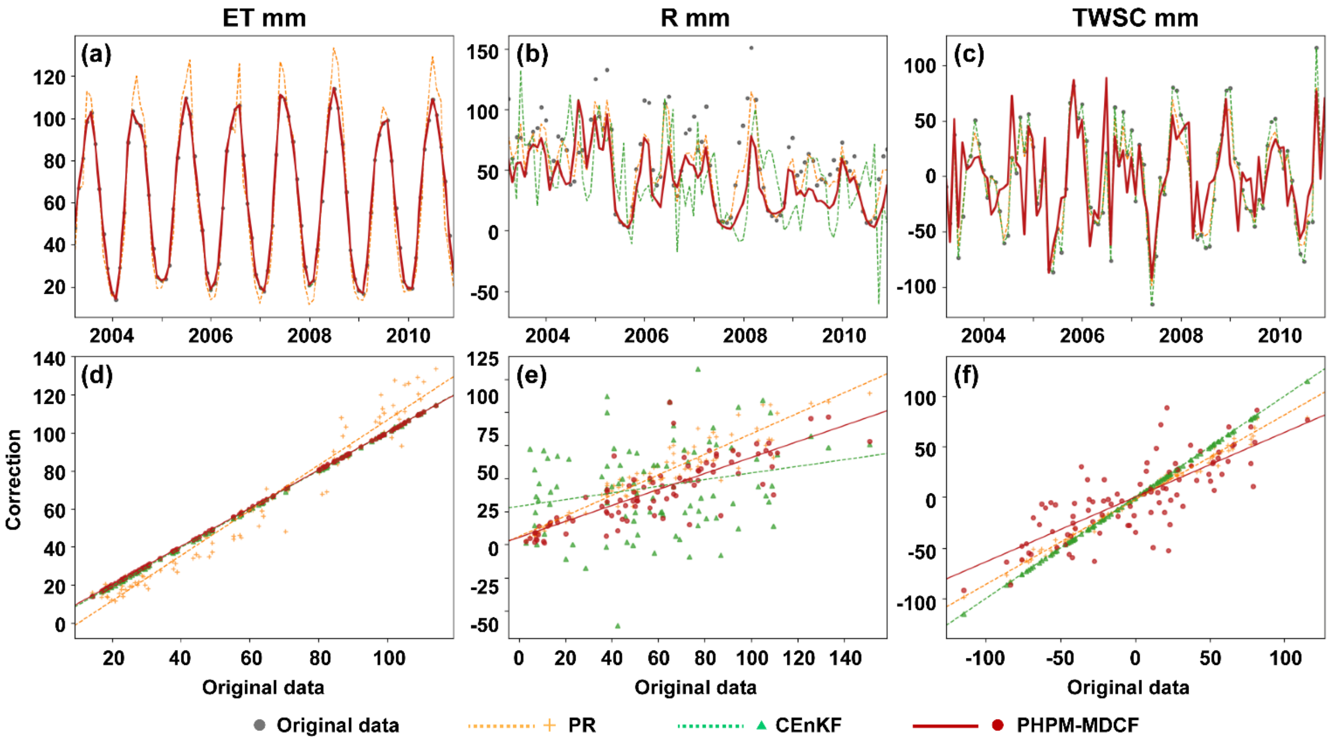

Figure 11. Comparison of monthly correction results between the PHPM-MDCF and existing methods (PR and CEnKF) at basin 1539000. (a-c) Time series of the original and corrected measurements of evaporation, streamflow, and terrestrial water storage change. (d-f) Scatter plots and regression lines of the original and corrected measurements.

Table S3. Summary of datasets from Lehmann et al. (2022).

Variable	Product	Original Resolution		Original Period
		Spatial	Temporal	
Precipitation	CPC	0.5 °×0.5 °	Monthly	2002-2017
	CRU	0.5 °×0.5 °	Monthly	1901-2019
	ERA5 Land	0.1 °×0.1 °	Monthly	1981-2020
	PGF	1.0 °×1.0 °	Monthly	1948-2014
	GPCC	0.5 °×0.5 °	Monthly	1891-2016
	GPCP	2.5 °×2.5 °	Monthly	1979-2020
	GPM	0.1 °×0.1 °	Monthly	2000-2020
	JRA55	0.5 °×0.5 °	Monthly	1959-2020
	MERRA2	0.5 °×0.625 °	Monthly	1980-2020
	MSWEP	0.5 °×0.5 °	Monthly	1979-2020
	TRMM	0.25 °×0.25 °	Monthly	1998-present
Evaporation	ERA5 Land	0.1 °×0.1 °	Monthly	1981-2020
	FLUXCOM	0.5 °×0.5 °	Monthly	2001-2015
	GLDAS22 CLSM	0.25 °×0.25 °	Daily	2003-2020
	GLDAS20 CLSM/NOAH/VIC	1.0 °×1.0 °	Monthly	1979-2014
	GLDAS21 NOAH/CLSM/VIC	1.0 °×1.0 °	Monthly	2000-2020
	GLEAM	0.25 °×0.25 °	Monthly	1980-2018
	JRA55	0.5 °×0.5 °	Monthly	1959-2020
	MERRA2	0.5 °×0.625 °	Monthly	1980-2020
	MOD16	0.5 °×0.5 °	Monthly	2000-2014
	SEBBop	0.5 °×0.5 °	Monthly	2003-2020
Streamflow	ERA5 Land	0.1 °×0.1 °	Monthly	1981-2020
	GLDAS22 clsm	0.25 °×0.25 °	Daily	2003-2020
	GLDAS20 CLSM/NOAH/VIC	1.0 °×1.0 °	Monthly	1979-2014
	GLDAS21 CLSM/NOAH/VIC	1.0 °×1.0 °	Monthly	2000-2020
	GRUN	0.5 °×0.5 °	Monthly	1902-2014
	JRA55	0.5 °×0.5 °	Monthly	1959-2020
	MERRA5	0.5 °×0.625 °	Monthly	1980-2020
Terrestrial water storage	GRACE JPL mascons	0.5 °×0.5 °	Monthly	2002-present
	GRACE CSR mascons	0.5 °×0.5 °	Monthly	2002-present

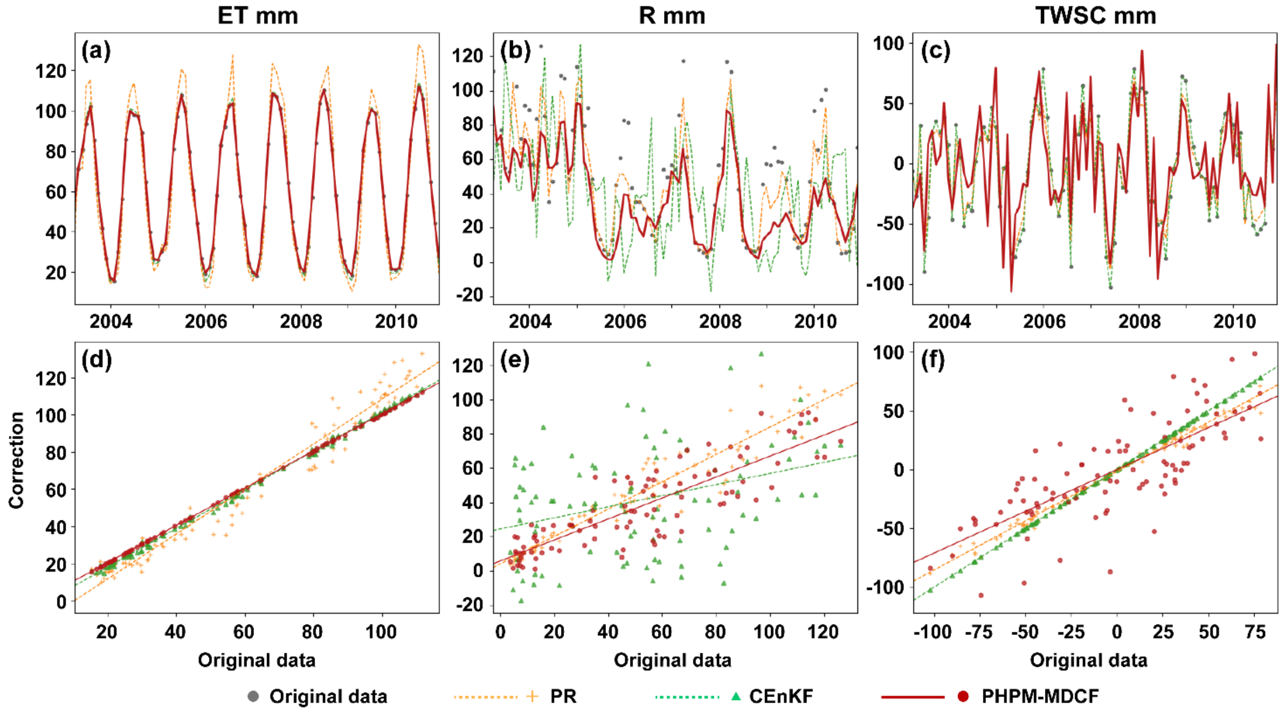

Figure S11. Same as Fig. 11, but for basin 1557500.

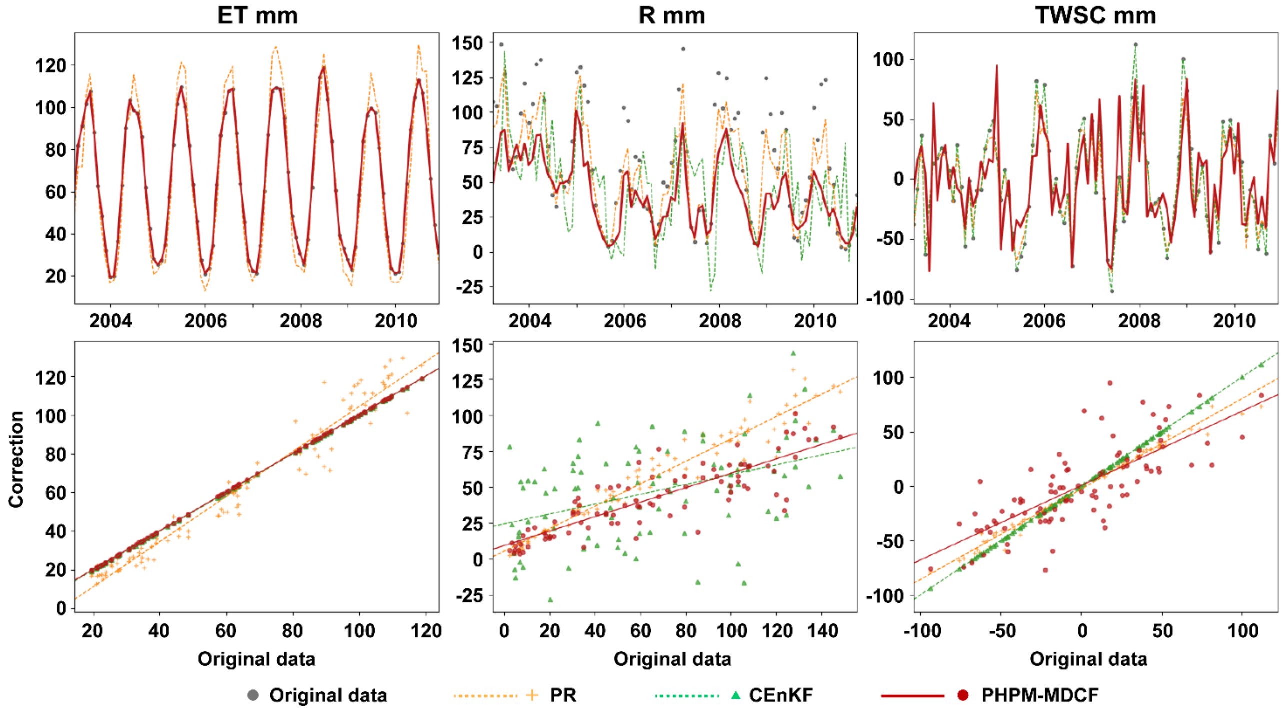

Figure S12. Same as Fig. 11, but for basin 3070500.

C/ Finally, the observational data referenced by the author is not in-situ measurements, and attention should be given to the terminology used.

R/ Thank you for pointing this out, we will emphasize the scope of this term's usage in this paper. The following is the content we will add to the data description section.

"Notably, the term 'measurements' referred in this work are derived from multisource datasets and do not specifically refer to in-situ measurements.”