Hydrology and
Earth System
Sciences

Achieving water budget closure through physical hydrological process modelling: insights from a large-sample study

Xudong Zheng[1], Dengfeng Liu[1], Shengzhi Huang[1], Hao Wang[2], and Xianmeng Meng[3]

[1]State Key Laboratory of Eco-hydraulics in Northwest Arid Region of China, School of Water Resources and Hydropower,
Xi'an University of Technology, Xi'an, 710048, China
[2]State Key Laboratory of Simulation and Regulation of Water Cycle in River Basin,
China Institute of Water Resources and Research, Beijing, 100038, China
[3]School of Environmental Studies, China University of Geosciences, Wuhan, 430074, China

Correspondence: Dengfeng Liu (liudf@xaut.edu.cn) and Shengzhi Huang (huangshengzhi7788@126.com)

Abstract. Modern hydrology is embracing a data-intensive new era, with information from diverse sources currently providing support for hydrological inferences at broader scales. This results in a plethora of data-reliability-related challenges that remain unsolved. The water budget non-closure is a widely reported phenomenon in hydrological and atmospheric systems. Many existing methods aim to enforce water budget closure constraints through data fusion and bias correction approaches, often neglecting the physical interconnections between water budget components. To solve this problem, this study proposes a Multisource Dataset Correction Framework grounded in Physical Hydrological Process Modelling to enhance water budget closure, termed the PHPM-MDCF. The concept of decomposing the total water budget residuals into inconsistency and omission residuals is embedded in this framework to account for different residual sources. We examined the efficiency of the PHPM-MDCF and the distribution of residuals across 475 contiguous United States (CONUS) basins selected by hydrological simulation reliability. The results indicate that the inconsistency residuals dominate the total water budget residuals, exhibiting highly consistent spatiotemporal patterns. This portion of residuals can be significantly reduced through PHPM-MDCF correction and achieved satisfactory efficiency. The total water budget residuals decreased by 49 %, on average, across all basins, with reductions exceeding 80 % in certain basins. The credibility of the correction framework was further verified through noise experiments and comparisons with existing methods. In the end, we explored the potential factors influencing the distribution of residuals and found notable scale effects, along with the key role of hydrometeorological conditions. This emphasizes the importance of carefully evaluating the water balance assumption when employing multisource datasets for hydrological inference in small and humid basins.

1 Introduction

Advances in measurement and monitoring techniques have revolutionized hydrology research through providing an unprecedented opportunity to detect hydrological processes (Sivapalan and Blöschl, 2017). Data availability is no longer the key constraint for conducting large-scale research as it once was. Approaches that work with large samples and multisource data are now more attractive for hydrological studies (Nearing et al., 2021). In the absence of satisfactory in situ observation, we can freely access data from different sources as a complement, such as satellite remote sensing, radar, model simulations, and reanalysis (Refsgaard et al., 2022). As such, whether at the watershed scale or at the modelling scale (e.g. grid cells), we have multiple choices to represent water budget components, thereby facilitating hydrological inferences. This reality is also referred to as the fourth paradigm of hydrology (Peters-Lidard et al., 2017).

However, every coin has two sides: the abundance of available data has brought challenges with regard to data selection, confronting contemporary hydrologists with the task of

filtering datasets. After excluding datasets that do not match the research scale and spatiotemporal coverage, we still have no idea about how to select the most suitable one from the remaining datasets. In the past decades, extensive efforts have been made to evaluate the accuracy of datasets by referencing in situ observations or ensembles of multisource data (Sahoo et al., 2011; Tang et al., 2020; Ansari et al., 2022). However, the fact remains that the "true value" is perpetually unattainable, rendering any form of reference data uncertain. For example, the undercatch phenomenon in rainfall measurements is well known, and it is difficult to eliminate the bias even with the application of undercatch corrections (Robinson and Clark, 2020). The issue of scale mismatches and the lack of availability of site data in certain regions also pose challenges for data evaluation. Therefore, we argue that the evaluation based on reference data lacks sufficient reliability, highlighting the need for more widely applicable criteria in evaluating and correcting datasets from various sources.

The law of mass conservation, typically represented in hydrology by the water balance, constitutes a fundamental principle applicable universally across time and space. Thus, the terrestrial water budget describes the physical consistency among different components of the water balance, which can serve as a criterion for evaluating and correcting datasets. For a closed basin, the water budget can be mathematically expressed as (Lehmann et al., 2022)

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

where $\frac{\mathrm{dTWS}}{\mathrm{d}t}$ is the change in terrestrial water storage, P is precipitation, ET is evaporation, and R is runoff at the outlet. By incorporating data from different sources into Eq. (1), we can assess whether these data achieve closure of the water budget, thereby evaluating their reliability in depicting hydrological processes. If Eq. (1) is not satisfied, the residual term, known as water budget residuals, can quantify the extent of physical inconsistency among multiple datasets. A comprehensive review of the terrestrial water budget closure examination is given in Lv et al. (2017); interested readers are encouraged to refer to this work. The consensus in the recent scientific literature is that data inconsistency is widespread, attributed to different production processes among various datasets, and no single combination of datasets can fully close the water budget across all basins. Such inconsistency poses an obstacle to robust hydrological inferences (Beven, 2002). As an example of this, physically inconsistent forcing and evaluation data can mislead hydrological modelling and introduce significant uncertainty into model inferences (Kauffeldt et al., 2013). To mitigate the impact of data inconsistency, it is essential to properly correct datasets and improve water budget closure.

The pioneering work in enhancing water budget closure across different data sources through data correction was conducted by Pan and Wood (2006), who integrated a constrained ensemble Kalman filter (CEnKF) to impose con-

strains on the terrestrial water budget. This technique was subsequently developed and applied in several studies (Sahoo et al., 2011; Zhang et al., 2016). Similar extension methods include the multiple collocation (MCL) and proportional redistribution (PR) methods (Abolafia-Rosenzweig et al., 2020; Abhishek et al., 2022; Luo et al., 2023). These methods are all grounded in the data fusion process, deriving uncertainties for each water budget component from multiple data sources. Estimated uncertainties facilitate the determination of weights for allocating closure residuals, ultimately achieving a zero residual. Overall, these methods can be collectively referred to as data-fusion-based closure correction approaches. Another recently developed method to constrain the water balance employs an optimization-based strategy, exhibiting improved performance in long-term consistency with GRACE terrestrial water storage change (Petch et al., 2023). Other approaches, such as the post-processing filtering technique (PF) and bias correction method (Munier et al., 2014; Weligamage et al., 2023), can also be helpful in closing the water budget. However, the closure constraints imposed by the above methods (hereafter referred to as traditional methods) have been questioned, with Abolafia-Rosenzweig et al. (2020) arguing about the potential incorrect assignment of residuals. If a component in the water budget exhibits a bias, closure correction algorithms may mistakenly apply the bias closure constraint to other components. The intrinsic attribution of this issue lies in the algorithms neglecting the physical correlations among components and imposing strict constraints on water budget closure by integrating uncertainties from multisource data. In other words, assigning closure residuals based exclusively on the magnitude of a priori data uncertainty without accounting for the distribution of components in hydrological processes, such as the partitioning of precipitation, may be unrealistic and could lead to erroneous allocation of closure residuals. In the context of applying such closure constraints, it becomes evident that the precision of certain individual components may deteriorate notably, particularly when uncertainties are challenging to quantify (Luo et al., 2023).

As is well-known, hydrological models, whether data-driven or physics-based, aim primarily to characterize hydrological processes by accurately allocating water quantities among components such as precipitation, evaporation, runoff, and soil moisture. In abstract terms, hydrological models can be regarded to be directed graphs of fluxes, with nodes representing state variables and edges symbolizing fluxes or transitions (Wang and Gupta, 2024). Such a directed graph is computationally closed, indicating that hydrological models inherently exhibit the essential characteristic of water budget closure. A clear piece of evidence comes from the data consistency evaluation conducted by Gutenstein et al. (2021), who found that the dataset from the same model (i.e. precipitation and evaporation from the ERA5 coupled model) manifested a well-closed system. In this sense, hydrological models appear to be capable of guiding the allo-

cation of closure residuals to enhance water budget closure. Another distinctive feature of hydrological models, known as error adaptability or calibration compensation capability, underscores their pivotal role as innovative solutions for addressing challenges in achieving water budget closure. The feature emphasizes that hydrological models can, to some extent, compensate for biases in model inputs, outputs, and structure, allowing satisfactory performance even when the utilized datasets exhibit certain inaccuracies (Wang et al., 2023). This provides hydrological models with the potential to integrate forcing and evaluation datasets into a unified water balance system under the soft-constraint paradigm.

Here, we propose another critical question regarding achieving water budget closure: is the terrestrial water budget described by Eq. (1) fully comprehensive? This issue came to our attention through a recent study by Gordon et al. (2022), who examined the widespread validity of the closed water budget (CWB) hypothesis (i.e. formulated by Eq. 1) across 114 highland catchments using multiple data sources. Surprisingly, their results revealed that the CWB hypothesis failed to hold in 75 % to 100 % of the catchments. They highlighted that such failure of the CWB hypothesis could propagate widely in hydrological inferences relying on it, potentially leading to erroneous conclusions. To provide a physical explanation for the invalidity of the CWB, they extended Eq. (1) by introducing an error term e and an additional term G, as depicted in Eq. (2).

$$e + G = P - \mathrm{ET} - R - \frac{\mathrm{dTWS}}{\mathrm{d}t} \tag{2}$$

The term G accounts for the inter-basin groundwater fluxes that were not considered in the original formulation, while the term e addresses inconsistencies among the original datasets. Clearly, when applying the CWB hypothesis for data evaluation or correction, there is a tendency to prematurely assume the completeness of the applied formulas, potentially leading to significant biases in the final results. Furthermore, in practical applications, Eq. (1) may inadvertently omit other water fluxes and storages besides groundwater. For instance, utilizing gravity changes observed by GRACE to estimate terrestrial water storage (TWS) may encompass inter-basin water transfers or irrigation, which can have a substantial influence in studies conducted at relatively small scales (Lv et al., 2017). Partial observations of precipitation, evaporation, and runoff can also introduce biases into this equation. To distinguish the omission from total water budget residuals among the original datasets, we further extend Eq. (2) to obtain the generalized form as follows:

$$\mathrm{Res} = \mathrm{Res_i} + \mathrm{Res_o} = P - \mathrm{ET} - R - \frac{\mathrm{dTWS}}{\mathrm{d}t}, \tag{3}$$

where Res is the total water budget residuals; $\mathrm{Res_i}$ is the inconsistency residuals, accounting for the fraction of water non-closure due to physical inconsistencies among the original datasets; and $\mathrm{Res_o}$ is the omission residuals, explaining the fraction resulting from omitted fluxes and storages in the original equation. We assume that Eq. (3) offers a comprehensive description of the terrestrial water budget and can be examined using multisource datasets. This advancement, compared to previous studies, breaks down the sources of water budget residuals, offering guidance for data evaluation and correction.

Given the current increase in data availability but concerns over reliability, this study aims to address the following scientific questions through physical hydrological process modelling: (a) how can the total water budget residuals be quantitatively decomposed into inconsistency and omission residuals based on Eq. (3)? (b) From a large-sample perspective, what are the distribution patterns of these residuals? (c) What strategies can be employed to achieve water budget closure through physical hydrological process modelling while strengthening the physical coherence among datasets from different sources? By addressing these questions, we highlight the necessity of a comprehensive description of the water budget equation to effectively evaluate and correct water non-closure. Furthermore, we developed a multisource dataset correction framework based on decomposition of water budget residuals and multi-objective calibration within hydrological modelling. The presented framework, providing the capability to enhance the water budget closure and hydrological connections among multisource datasets, was applied to a large-sample basin dataset across the contiguous United States (CONUS).

The remainder of this paper is organized as follows. Section 2 describes the main datasets used in this research. Section 3 then details the methods for decomposing water budget residuals and the multisource data correction framework with a hydrological model. The results are presented and discussed in Sects. 4 and 5. Section 6 provides the main conclusions and outlook of this study.

2 Data

2.1 The CAMELS dataset

Motivated by the call of Gupta et al. (2014) for large-sample hydrological studies to strike a balance between depth and breadth, in this study, we attempt to carry out analysis on a widely used large-sample dataset, i.e. the Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) community dataset. This dataset, developed by Newman et al. (2015) and Addor et al. (2017), encompasses daily forcings, hydrologic responses, and basin attributes for 671 basins across the contiguous United States (CONUS), characterized by minimal human disturbance. Drawing upon this dataset, a substantial body of experimental studies have been conducted, covering model intercomparison, analyses of scale effects in hydrology, evaluations of model performance metrics, parameter estimation, and exploration of ma-

chine learning models (Knoben et al., 2020; Beven, 2023). Grounded in large-sample inquiries, these studies systematically explore the prevalent heterogeneity from different perspectives, yielding more robust and widely applicable conclusions.

In the original work by Newman et al. (2015) proposing the CAMELS dataset, widespread physical-inconsistency behaviours were observed, characterized by an imbalance between precipitation and runoff. In the spatial depiction within the Budyko framework, certain basins exhibited plotting points exceeding the water limit line, indicating a surplus of runoff relative to precipitation. Newman et al. (2015) emphasized the necessity of corrections to be applied to datasets. For the aforementioned reasons, investigation of the decomposition and reconciliation of water budget residuals within the CAMELS dataset is both necessary and feasible. In practice, the in situ runoff data observed by the USGS National Water Information System server were used. Considering the availability of data products, our analysis is conducted over a common overlapping period spanning 1998 to 2010. During this period, 18 basins with missing runoff observations were excluded in advance. Figure 1 presents a regional profile, and detailed information on the excluded basins is provided in Table S1 in the Supplement.

2.2 Datasets for constructing water budget equation

One of the main aims of this study is to investigate the decomposition of water budget residuals and corrections to datasets rather than comparing the differences and rankings of closure residuals across different dataset combinations. In line with this objective, referring to the work of Petch et al. (2023), we strategically selected a single product for each water component to construct the water budget equation, thereby laying the foundation for further research. In making this selection, we not only considered the resolution and spatiotemporal coverage of the products but also took into account recommendations from previous data evaluation studies regarding data accuracy (Kittel et al., 2018; Lehmann et al., 2022). All datasets used are summarized in Table 1. Notably, the "measurements" as used in this work are derived from multisource datasets and do not specifically refer to in situ measurements.

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 the TRMM dataset has been 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 the Global Land Evaporation Amsterdam Model (GLEAM v3) product

(https://www.gleam.eu/, last access: 31 August 2023), 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 ensembles (Munier et al., 2014; Robinson and Clark, 2020). And, as mentioned above, the runoff measurements at the basin scale are provided by the CAMELS dataset, which is derived from site observations.

Finally, the most challenging component to estimate in the water budget equation is the terrestrial water storage change (TWSC) as it includes water both on and below the Earth's surface. In previous studies, the measurement of gravity field changes, as provided by the Gravity Recovery And Climate Experiment (GRACE) product, has frequently been employed for the estimation of the TWSC (Luo et al., 2020; Kabir et al., 2022). This approximation is based on the assumption that, for a given large-scale basin, variations in mass are primarily attributed to changes in TWSC. However, the assumption is fragile when applied to small basins, leading to significant uncertainty in estimating TWSC for basins with areas of less than $63\,000\,\text{km}^2$ (Lehmann et al., 2022). This study focuses on the basin dataset from CAMELS, with most basin areas being smaller than this threshold. To avoid introducing additional uncertainty into the analysis, we need alternative methods to estimate TWSC.

Assuming that TWSC can be retrieved through a combination of different water storages, we obtained the four-layer soil moisture from ERA5-Land and the snow water equivalent (SWE) from GlobSnow to estimate the overall TWSC. This approach has been implemented in the investigation of Hoeltgebaum and Dias (2023), yielding a high consistency between estimated TWSC and GRACE observations (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 models, 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 the analysis of water budget closure at the basin scale from 1998 to 2010. Including the observed runoff from CAMELS, all data were converted to water depth (mm) to construct a unified water budget equation. It is noteworthy that there are certain missing data in GlobSnow SWE, varying across basins. To fill these data gaps, we set a window length of 5 d, centred on missing data. We applied linear interpolation within the window for gap filling. If linear interpolation was not feasible due to, for instance, the absence of valid values within the window, mean climatology was employed to fill the missing data. To illustrate this, we randomly selected nine basins and visually depicted the gap-filling process in Fig. S1.
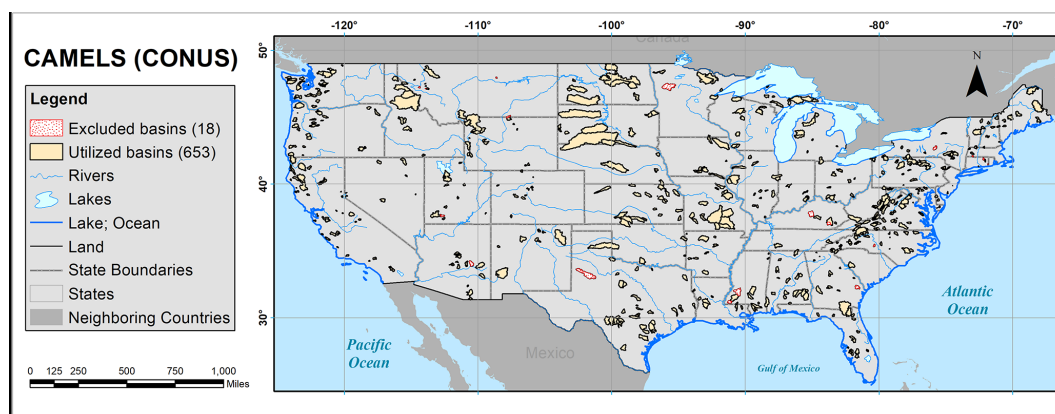
Figure 1. Geographic representation of the CAMELS basin dataset (Newman et al., 2015, and Addor et al., 2017). The 18 basins excluded from the analysis are denoted by red dots, whereas the study incorporates the remaining 653 basins, emphasized with yellow shading. The copyright of the background map belongs to Esri (Gray Canvas Basemap).

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

Variable	Product	Original resolution		Original period	Reference
		Spatial	Temporal		
Precipitation	TRMM 3B42V7	$0.25° \times 0.25°$	Daily	1998–2019	Huffman et al. (2016)
Evaporation	GLEAM v3.8a	$0.25° \times 0.25°$	Daily	1980–2022	Martens et al. (2017)
Soil moisture layer 1/2/3/4	ERA5-Land	$0.1° \times 0.1°$	Hourly	1950–present	Muñoz Sabater et al. (2021)
Snow water equivalent	GlobSnow v3.0	$25\,\text{km} \times 25\,\text{km}$	Daily	1979–2018	Luojus et al. (2021)
Runoff	CAMELS USGS	Basin scale	Daily	1980–2010	Newman et al. (2015)

3 Methods

To leverage physical hydrological process modelling for the decomposition and correction of water budget residuals, the following assumptions are necessary: (1) the hydrological model provides a reliable representation of hydrological processes, ensuring an accurate partitioning of input precipitation, and (2) the uncertainties associated with the model forcing and structure can be considered to be negligible during the modelling process. These two hypotheses form the foundation of this work. To ensure the validity of hypothesis 1, we employed multiple evaluation variables and corresponding metrics to guarantee the overall reliability of the model, which will be detailed in the model setup section. Additionally, it is pertinent to acknowledge that hypothesis 2 represents a strong assumption, carrying inherent uncertainties. Despite this, it is necessary for the feasibility of the overall work, and we will explore the influence of this hypothesis on the results further in the Discussion section.

3.1 Decomposition of water budget residuals: inconsistency and omission residuals

Our strategy for decomposing water budget residuals is grounded in the computational closure of the hydrological model. As previously discussed, conceptualized as a closed directed graph, the difference between the inputs and outputs of the model must necessarily equal the change in state variables. Stated differently, there is a water balance between the forcing and simulated variables of the model, with no physical-inconsistency residuals present. Therefore, setting the inconsistency residuals in Eq. (3) to zero allows us to derive the water budget equation of the hydrological model as follows:

$$\text{Res}_o = P_{\text{forcing}} - \text{ET}_{\text{sim}} - R_{\text{sim}} - \frac{\text{dTWS}_{\text{sim}}}{\text{d}t}, \qquad (4)$$

where the subscripts "forcing" and "sim" denote the forcing and simulation values, respectively. It is crucial to clarify that all variables in Eq. (4) are derived from the model itself rather than from measurements and can therefore be considered to be physically consistent. On the other hand, integrating the multisource datasets described in Sect. 2.2 into Eq. (3) yields the total water budget residuals (i.e. Res). For convenience, we refer to the water budget characterized by the hydrological model as the simulation system and the one constructed by multisource datasets as the measurement system. When the hydrological model calibrated against multiple variables measured by the multisource datasets achieves reliable performance, we consider the water budgets represented by the simulation and measurement systems to be comparable. At this point, the difference between Eqs. (3) and (4) represents

the omission residuals (i.e. $\text{Res}_i = \text{Res} - \text{Res}_o$), indicating the water fluxes or storages omitted by the original equation. Thus, the total water budget residuals can be decomposed into inconsistency and omission residuals. It is noteworthy that, while the inconsistency residuals are absent in the simulation system (a physically consistent system), omission residuals may still exist due to inherent omissions in the original equation. Hence, the left-hand side of Eq. (4) may not be zero.

Considering the comparability of available datasets and model simulations, we have developed more specific expressions for Eqs. (3) and (4), as depicted below.

$$\text{Res} = \text{Res}_i + \text{Res}_o = P_{\text{TRMM}} - \text{ET}_{\text{GLEAM}} - R_{\text{USGS}}$$
$$- \frac{\text{dSWE}_{\text{GlobSnow}} + \text{dSM}_{\text{ERA5}}^{0\text{–}50\,\text{cm}} + \text{dSM}_{\text{ERA5}}^{50\text{–}289\,\text{cm}}}{\text{d}t}, \quad (5)$$
$$\text{Res} = \text{Res}_o = P_{\text{TRMM}} - \text{ET}_{\text{sim}} - R_{\text{sim}}$$
$$- \frac{\text{dSWE}_{\text{sim}} + \text{dSMS}_{\text{sim}} + \text{dGRS}_{\text{sim}}}{\text{d}t}, \quad (6)$$

In the above, the subscripts indicate variable sources, such as measurements and simulated values, and superscripts for soil moisture (SM) denote the depth of the soil layers to be aggregated. The above water budget equations are discretized, employing a simple central difference scheme with a 2 d time step at the daily scale (Petch et al., 2023). Then, the residuals are calculated at the daily scale and are subsequently aggregated to the monthly and annual scales for further analysis.

It is important to further clarify that the hydrological model used in this study (see below) divides total soil moisture into soil water storage (SMS_{sim}, hereafter SMS) and groundwater reservoir storage (GRS_{sim}, hereafter GRS). The soil moisture measurements of ERA5, on the other hand, employ the HTESSEL (Hydrology Tiled ECMWF Scheme for Surface Exchanges over Land) land surface scheme to characterize land surface hydrological processes (Balsamo et al., 2009), dividing soil into four layers (i.e. 0–7, 7–28, 28–100, and 100–289 cm). In the HTESSEL model, the upper 50 cm of the soil column is defined as the effective depth for generating surface runoff. To ensure consistency between the simulation and measurement systems, we match the top 50 cm of ERA5 soil moisture with the soil water storage in the hydrological model used, while the depth range of 50 to 289 cm corresponds to the groundwater reservoir storage in the same model.

3.2 Multisource dataset correction framework for achieving water budget closure

Here, we introduce an innovative Multisource Dataset Correction Framework grounded in Physical Hydrological Process Modelling, termed the PHPM-MDCF, to enhance water budget closure. Unlike traditional correction methods that use uncertainty (typically derived from the variance of multisource datasets for the same variable or priori estimation)

as a weight for allocating water budget residuals, this framework leverages the hydrological model – a physically consistent system – as a constraint to correct the measurement system. Figure 2 indicates the flowchart for the correction framework, and the procedure is described below.

- *Step 1.* Initialize the basic computing unit. Calibrate the hydrological model, calculate the total water budget residuals from the original datasets, and then decompose them into inconsistency and omission residuals following the method outlined in Sect. 3.1. This step is denoted as iteration 0.

- *Step 2.* Correct for the inconsistency residuals. Allocate inconsistency residuals based on the magnitude of the differences (i.e. the distance between the simulation and measurement systems) between simulated and measured values for each variable in Eqs. (5) and (6). This difference indicates the correction direction and magnitude for each variable, which facilitate the convergence of the measurement system toward 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:

$$M_c^v = M_o^v - \text{Res}_i \times \frac{d_v}{d_{\text{all}}} \times \alpha, \quad (7)$$

where M_c^v denotes the corrected measurements of variable v, and M_o^v is the original measurements; d_v is the difference between the simulation and measurement of variable v, and d_{all} represents the aggregate of differences for all variables; α is the correction rate, with an initial value of 0.5.

- *Step 3.* Calibrate and evaluate 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 to be excessive, halve the correction rate, and repeat step 2. Otherwise, maintain the correction rate and proceed with the next iteration of the 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. This is to say, the iterative process involves continual trial and error, with each error prompting us to approach the next correction more cautiously.

- *Step 4.* Conduct iteration and termination of the correction. Iterate through steps 2–3 to gradually correct the datasets until the inconsistency residual decreases to 10 % of its initial value or until the correction rate falls below 4 %.
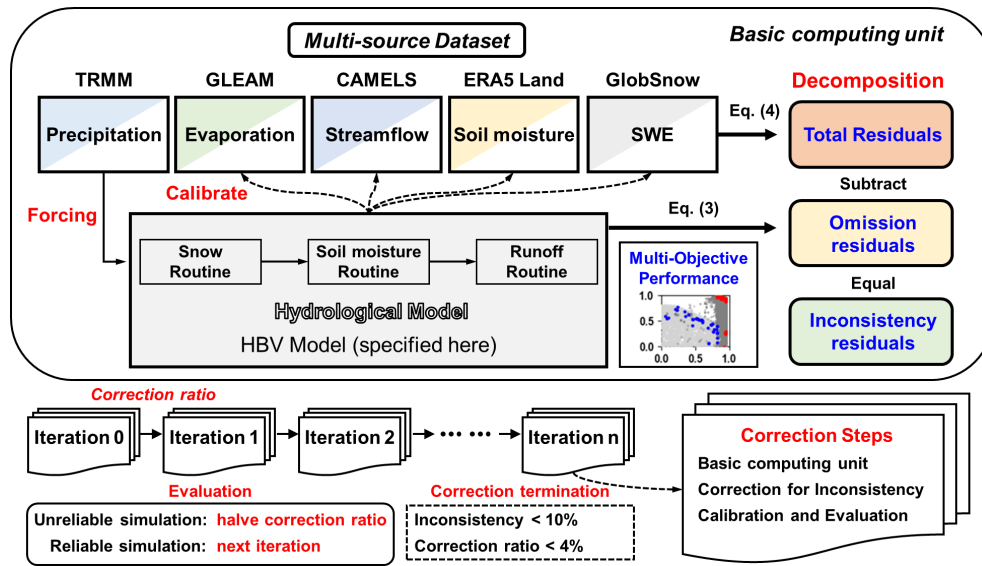
Figure 2. Flowchart of the Multisource Dataset Correction Framework grounded in Physical Hydrological Process Modelling (PHPM-MDCF).

The design goal of the PHPM-MDCF is to impose soft constraints on multisource datasets through the calibration compensation capability and the physical consistency feature of the hydrological model. Such a constraint is referred to as soft because, unlike traditional methods that import "hard" constraints, the correction process does not strictly require residuals to be zero immediately. Instead, it aims to advance the convergence between the simulation and measurement systems, as illustrated in Fig. 3. In extreme cases, when the measurement system is corrected to be identical to the simulation system, all measurements would become physically consistent. This process can be seen as a collapse from Eq. (5) to Eq. (6). The efficiency of ultimately closing residuals depends on the ability of the model to accurately characterize reality, and this can vary across different locations.

Notably, the correction is performed at the daily scale, aligning with the model step. In the subsequent application of the PHPM-MDCF, the measurements are derived from the data described in Sect. 2.2. In addition, through experimentation, the parameter settings in the PHPM-MDCF (i.e. initial correction rate, decay rate of the correction rate, and correction termination threshold) have been tailored to suit the current study area (Table S2). When applying this framework to other regions, additional adjustments and testing may be required.

3.3 Model setup and calibration

In the present investigation, we employed the Hydrologiska Byråns Vattenbalansavdelning (HBV) model to implement our correction framework. The conceptual HBV model was developed by the Swedish Meteorological and Hydrological
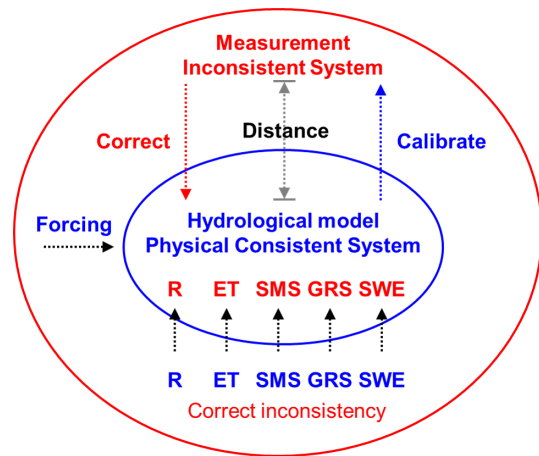

Figure 3. Illustration of the correction process advancing convergence between the simulation and measurement systems. The measurement system is corrected to approach the simulation system, while the simulation system is refined via parameter calibration to better approximate the measurement system. As a result, the distance between the two systems is reduced, leading to better physical consistency in the corrected measurement system.

Institute (SMHI) in the 1970s (Bergström, 1976). Given its straightforward yet effective design and minimal input requirements, this model has attained broad recognition and application within the global hydrological modelling scientific community and has also been tested in the CAMELS basins (Feng et al., 2022). Here, we provide brief details and refer the reader to the above references for a fuller description.

The basic structure of the HBV model comprises three main modules: the snow routine, soil moisture routine, and runoff routine, as illustrated in Fig. A1 in the Appendix. Starting with precipitation forcing, water flux traverses through the three modules, accumulating in various state variables such as snow and soil water. Ultimately, water is released through three reservoirs – soil moisture and upper-zone and lower-zone reservoirs – as quick runoff, interflow, and base flow. Thus, the overall soil moisture can be divided into soil water storage (i.e. the first reservoir) and groundwater reservoir storage (i.e. the combination of the latter two reservoirs). In the current study, the HBV model is configured to run on a daily basis, aligning with both the forcing and evaluation datasets, ensuring the feasibility of subsequent correction. Table A1 lists the free parameters slated for calibration in the HBV model, providing their descriptions and respective ranges.

Here, a multi-objective global optimization algorithm, the Non-dominated Sorting Genetic Algorithm II (NSGA-II), is applied for parameter calibration of the HBV model. Owing to its optimization efficiency, this algorithm has been extensively used in hydrological modelling practices around the world (Mostafaie et al., 2018). For more details about the algorithm, see Deb et al. (2002). We implemented the calibration framework using the NSGA-II algorithm in a Python environment with the DEAP package (Fortin et al., 2012). Five calibration objectives are considered, including R (runoff), ET (evaporation), SMS (soil moisture storage), GRS (groundwater reservoir storage), and SWE (snow water equivalent). Meanwhile, the Kling–Gupta efficiency (KGE) metric (Gupta et al., 2009) is utilized to evaluate the simulation performance of R and ET, while the Pearson correlation coefficient (r) is employed to evaluate the performance of SMS and GRS considering potential discrepancies in their magnitudes arising from differences in soil layer depth. Finally, the root mean square error (RMSE) is applied to evaluate the simulation performance of SWE. Ideally, the optimal simulation is characterized by values of 1 for the first two metrics and 0 for the last one. The detailed description of the evaluation metrics is provided in Appendix B.

4 Results

4.1 Distribution of water budget residuals and their components across the CAMELS basins

In this section, we investigate the spatiotemporal distribution of water budget residuals for each component decomposed using the method proposed in Sect. 3.1 across the large sample of the CAMELS basins. These results provide insights into the two primary sources of non-closure issues in the water budget equation: physical inconsistencies among the original datasets and water fluxes or storage omitted in the original equation. To ensure the robustness of the results, as mentioned previously, it is essential that hydrological models reliably represent hydrological processes. With reference to previous studies (Knoben et al., 2019; Clark et al., 2021; Aerts et al., 2022), we have adopted KGE ≥ -0.41 and a statistically significant r 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 the basins (475, accounting for 72.24 % of the total basins) achieved reliable simulations across all variables. Among them, we have observed that the central and western CONUS present relatively greater challenges for modelling. This pattern and its potential causes will be explored further in the ensuing discussion.

Within the 475 basins demonstrating reliable simulations, in Fig. 4, we plotted the spatial distribution of the long-term monthly mean water budget residuals (Res), inconsistency residuals (Res$_i$), and omission residuals (Res$_o$). An important observation from comparing the different rows of Fig. 4 is that Res shares a similar spatial pattern with Res$_i$, whereas Res$_o$ exhibits some differences. This pattern exists across different quantile ranges of the residuals. For instance, Res and Res$_i$ both present an east–west gradient for three statistical measures (i.e. min, median, max), with low values occurring along the western coastline and high values being primarily concentrated in eastern inland basins. The exception is a cluster of low median values located in the central CONUS. Interestingly, the minimum values of Res$_o$ display a contrasting spatial pattern, with higher values in the west and lower values in the east. The spatial differences in the median and maximum values of Res$_o$ are not pronounced. These patterns lend support to the underlying assumption that the drivers of inconsistency residuals and omission residuals are fundamentally different and thus can be decomposed from the total water budget residuals.

Figure 5 further illustrates the temporal distribution patterns of the three residuals in terms of seasonality. It is readily discernible in the figure that the similarity between Res and Res$_i$ reappears, manifesting distinct seasonal patterns with more pronounced negative trends during the cold seasons (i.e. October to the following April) and positive trends during warm seasons (i.e. May to September). On the contrary, Res$_o$ tends to be mainly positive, except from September to November; its extent of variability is also significantly smaller than that of the other two residuals. With regard to magnitude, Res$_i$ is much greater than Res$_o$ considering both positive and negative biases. From the above results, we can conclude that Res$_i$ predominates within Res, exhibiting significant spatiotemporal differences compared to Res$_o$. These two residuals may combine or offset each other to collectively form the total water budget residuals. The potential factors affecting the spatiotemporal distribution and proportion of Res will be investigated further in Sect. 4.4.
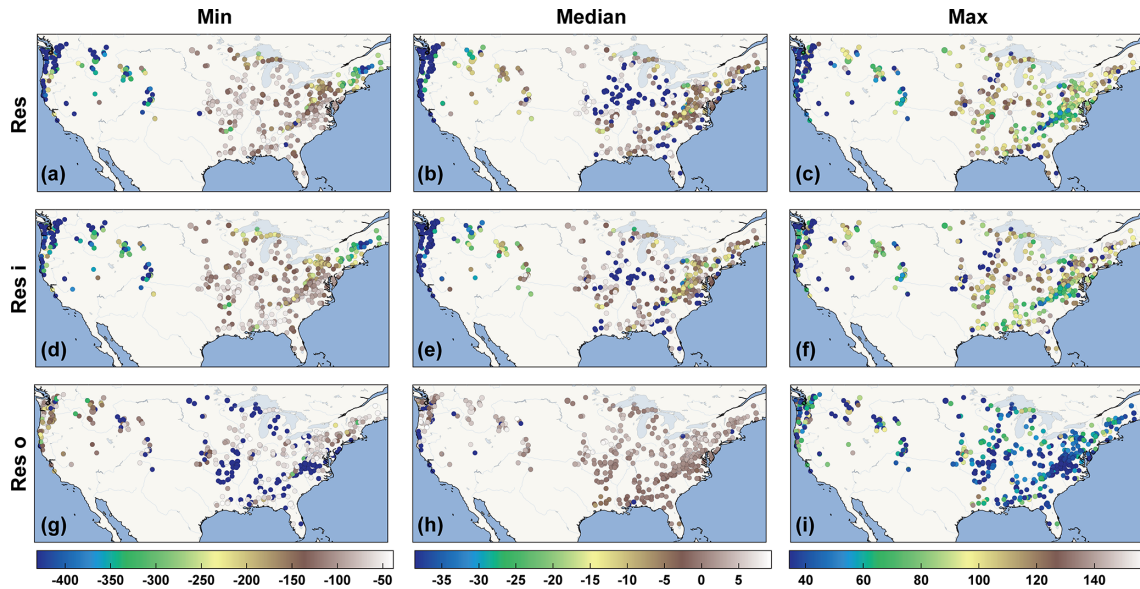
Figure 4. Spatial distribution of long-term monthly mean water budget residuals (Res), inconsistency residuals (Res$_i$), and omission residuals (Res$_o$) across 475 CAMELS basins with reliable simulations. The unit of the residuals is mm.
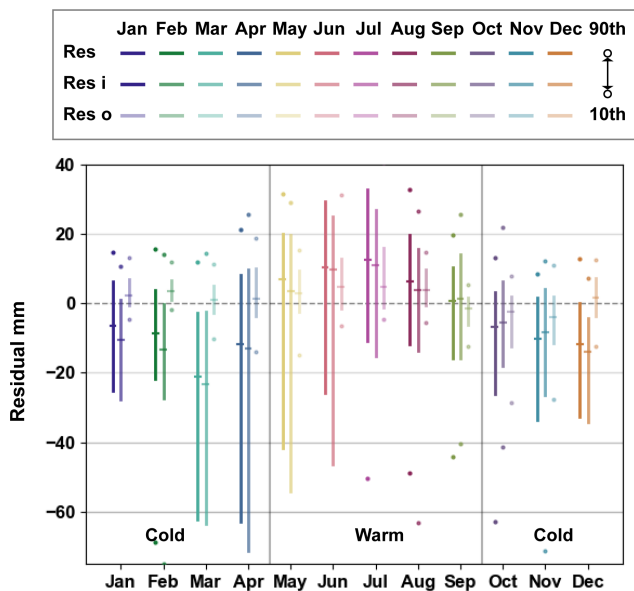

Figure 5. Temporal distribution of monthly water budget residuals (Res), inconsistency residuals (Res$_i$), and omission residuals (Res$_o$) across 475 CAMELS basins with reliable simulations. Boxplot-like diagrams describe variability across catchments, and outliers represent the 10th and 90th percentiles. The unit of the residuals is mm.

4.2 Efficiency of the PHPM-MDCF

We are now tackling the third question through the proposed multisource dataset correction framework (PHPM-MDCF) across the 475 CAMELS basins with reliable simulations.

For illustration, several case basins have been selected to demonstrate the correction process and its efficiency.

Figure 6 shows the correction results at the case basin numbered 1013500 (for more details about the basin number, see Newman et al., 2015). As expected, the time series of Res and Res$_i$ after correction (red lines) tend to be flatter and closer to zero compared to their uncorrected counterparts (blue lines). This becomes more apparent as the timescale increases. However, despite recalibrating the model with corrected datasets, Res$_o$ driven by the omission in the water budget equation exhibited no substantial changes before and after correction (e.g. the monthly mean absolute values were maintained around 6.5 mm; see Fig. 6f). This phenomenon occurs because we only corrected the inconsistency residuals with reference to the simulation system, while the omission accounting for the additional water terms should not be corrected in the existing datasets.

To get an impression of the PHPM-MDCF correcting the water budget residuals, the bottom row of Fig. 6 shows the variation in the mean absolute values of three residuals with increasing correction iterations at the monthly scale. The results indicated that the correction process led to a significant reduction in Res and Res$_i$, decreasing from 42.8 and 44.3 mm to 6.9 and 8.6 mm (reductions of approximately 83.9 % and 80.7 %). Although water budget residuals cannot be fully corrected to zero in this framework (as can be done in traditional methods), we argue that this correction efficiency is satisfactory enough. It is rooted in physical hydrological process modelling, thus potentially strengthening the physical relationships among the components of the water balance. The final corrected results for this case basin are presented in Fig. S2, depicting the time series of multisource datasets
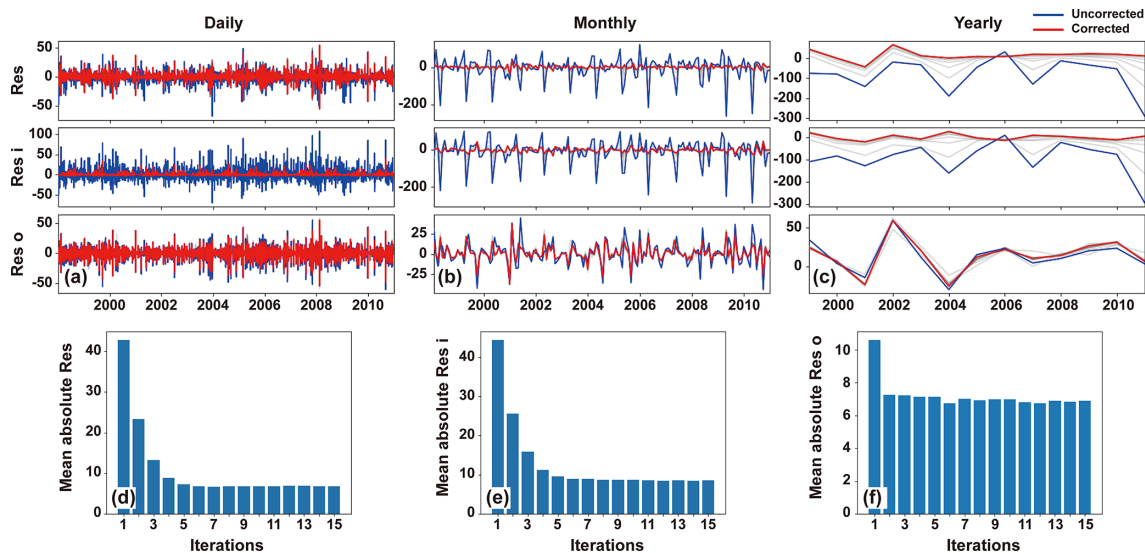
Figure 6. Correction results of water budget residuals for multisource datasets at basin no. 1013500. **(a–c)** Time series of water budget residuals (Res), inconsistency residuals (Res_i), and omission residuals (Res_o) at daily, monthly, and yearly scales; the grey line represents residuals during the correction process. **(d–f)** Variation in long-term mean absolute values of three residuals with correction iterations at the monthly scale. The unit of the residuals is mm.

before and after correction. In the following sections, we will provide further evidence of the credibility of this correction framework.

The correction results for several other case basins (i.e. numbered 1137500, 2177000, 6311000, and 14092750) are presented in Figs. S3–S6. Their absolute mean monthly residuals decreased by 70.4 %, 58.1 %, 40.3 %, and 54.0 %, respectively, providing evidence for the effectiveness of the PHPM-MDCF. To have a clearer idea of the ability of the correction framework to reduce water budget residuals across all the CAMELS basins, Fig. 7 shows the map of the percentage reduction in monthly total water budget residuals after corrections. In general, the PHPM-MDCF demonstrated robust performance across most basins, with an average reduction percentage of 49 % across all basins. The correction efficiency exhibits a latitudinally dependent decline pattern, which is primarily due to the small initial residuals in low-latitude regions (Fig. 4). In high-latitude regions, such as the western coastline and eastern inland basins, the potential correction space is much larger, leading to higher correction efficiency (in terms of absolute value).

4.3 Credibility of multisource dataset correction

4.3.1 Convergence between simulation and measurement system

As we stated before, the core objective of the PHPM-MDCF is to promote the convergence between the simulation and measurement systems (Fig. 3). In fact, this process can be divided into two parts. The first part, namely the measurement system approaching the simulation system, which is im-

plemented by correction procedures, has gained confidence from the significant reduction in the inconsistency residuals (Fig. 6). On the other hand, to illustrate the convergence of the simulation system toward the measurement system, we present the changes in model simulation performance before and after correction of case basin no. 1013500, as depicted in Fig. 8. From the figure, we can clearly see that both the population solution sets (ranging from light- to darker-grey scatters) and the Pareto fronts (ranging from blue to red scatters) tend toward the optimal point in the upper-right corner after correction. 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 simulations is significantly strengthened after correction. These results suggest 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.

4.3.2 Noise experiments

To further demonstrate the credibility of multisource dataset correction, we designed a series of noise experiments and applied them to case basin no. 1013500, therefore examining whether the PHPM-MDCF can effectively handle the manual noise and produce robust correction results. These experiments are summarized in Table 2, where the first three experiments set different types of single-point noise at different positions of the same original datasets, and the last experiment adds an equal-length Gaussian white-noise sequence to the runoff sequence. Eventually, two new noisy datasets
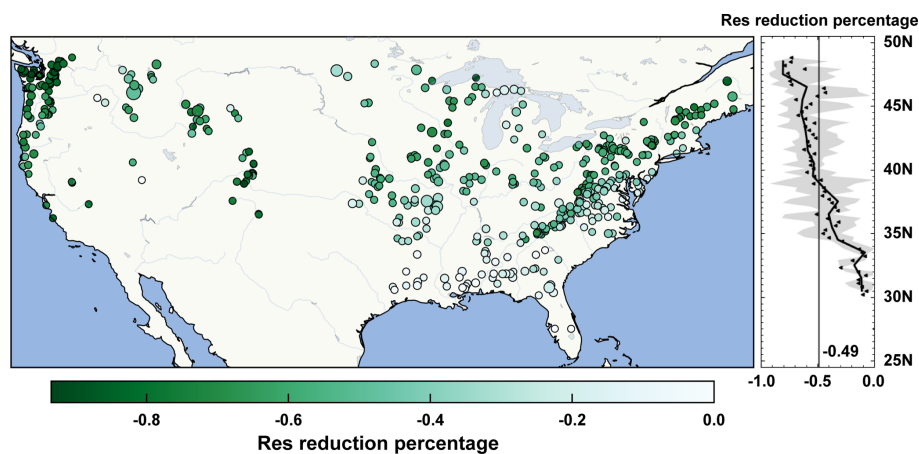
Figure 7. The percentage reduction in monthly total water budget residuals after correction through the PHPM-MDCF. Zonal means (right panel) include mean (black scatters), median (black line), and range (grey shading). The vertical line indicates the mean value of -0.49 for all basins.
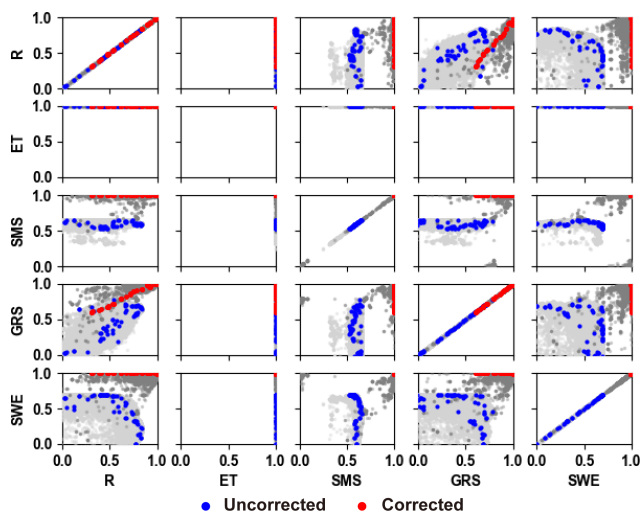

Figure 8. Comparison of multivariable simulation performance before and after correction at basin no. 1013500. Light grey and dark grey indicate population solution sets before and after correction, and blue and red indicate Pareto fronts before and after correction. Metrics evaluating SWE simulation performance have been normalized for consistency. The subplot in the second row and second column shows that the evaporation simulation maintains high accuracy in this basin due to the alignment between the HBV algorithm and measurements.

were generated, as illustrated in Figs. S8 and S9. For clarity, we refer to them as NS1 and NS2 (i.e. noise sequences 1 and 2) and designate the noise-free dataset as OS (i.e. original sequence). The noise points are ordered from 1 to 4.

First, we examined the adaptation capability of the PHPM-MDCF to single-point extreme errors. The top row of Fig. 9 compares the differential form of the OS and NS1, highlighting the impact of the three noises. The first two noises introduce extremely unreasonable values in the runoff measure-

ments, while the third noise affects the water balance significantly by altering all water budget variables, as evidenced in Fig. 9c–d. Through the application of the PHPM-MDCF for NS1 correction, we derived a new corrected sequence and compared it with the previous OS-based corrected sequence. In terms of runoff correction, as shown in Fig. 9c, whether the noises are extremely large or small (i.e. noise 1 and 2 with differences of 3 standard deviations), the correction process constrains them to reasonable runoff processes. This is achieved by the representation of physical hydrological processes underlying the correction strategy, which constrains the corrected values to avoid producing extreme outliers. Furthermore, water imbalance caused by the combination of multivariable single-point noises can also be constrained to minimal levels through correction (Fig. 9d).

Another concern here is whether the correction of extreme noises in runoff will propagate to other variables, potentially leading to a series of unreasonable correction results, as questioned by Abolafia-Rosenzweig et al. (2020) in relation to traditional methods. In Fig. S10, we specifically focus on the correction results around three single-point noises to address this question. The fact that simultaneous corrections of other variables during extreme runoff noise correction did not significantly differ from OS-based corrections further enhances our confidence in the PHPM-MDCF. This 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. From a theoretical perspective, the PHPM-MDCF assigns the weights of residual correction based on the distance between the measurements and simulations 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

Table 2. Description of the noise experiments to examine the credibility of multisource dataset correction.

ID	Description	Position of the noises (yyyy-mm-dd)	Noise sequence
Exp. 1	A single positively biased noise is added to R, with a magnitude of 3 standard deviations	Noise 1: 1998-09-18	NS1
Exp. 2	A single negatively biased noise is added to R, with a magnitude of 3 standard deviations	Noise 2: 1999-04-26	
Exp. 3	A set of positively biased noise at the same position is added to R, ET, SMS, GRS, and SWE, with a magnitude of 1 standard deviation	Noise 3: 2001-12-16	
Exp. 4	A series of zero-mean random Gaussian white noise is added to R, with a standard deviation of 20 % relative to the original sequence	Noise 4: the entire sequence	NS2
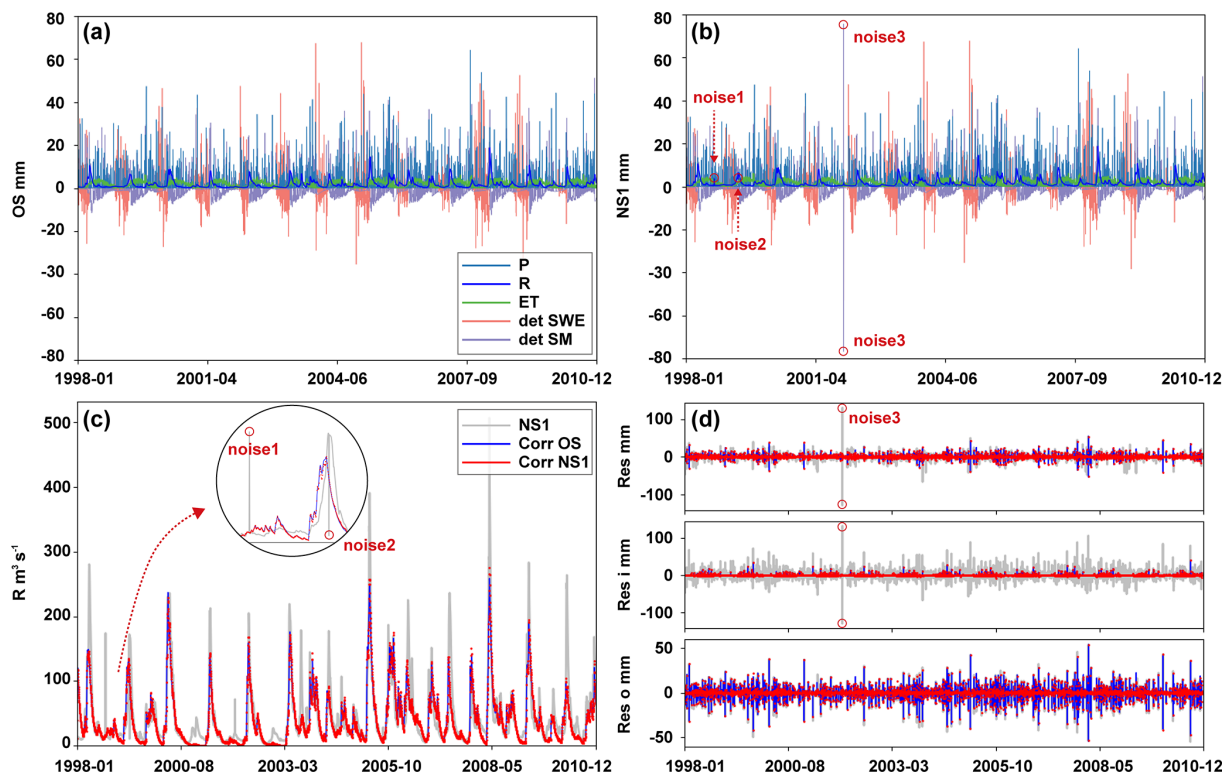
Figure 9. Correction results for multisource datasets corresponding to noise experiments 1–3. **(a–b)** Time series of OS and NS1 in the form of differences. **(c)** Comparison among the runoff noise sequence (NS1), OS-based runoff-corrected sequence (corr OS), and NS1-based runoff-corrected sequence (corr NS1). **(d)** Comparison of water budget residuals generated by the three sequences at the daily scale.

unaffected. However, in traditional methods, the correction weight for each variable remains constant over time, and the final residuals are constrained to zero. This leads to the propagation of extreme biases across different variables.

Subsequently, we assessed the robustness of correction results after incorporating Gaussian white noise into the original sequence. From the comparison between OS-based and NS2-based correction results (Fig. 10), it can be seen that the addition of Gaussian white noise changed the correction in runoff slightly, showing a minor decrease in the high-value range (with a slope less than 1). However, the overall evolution trend of runoff remains unchanged as it is still constrained by the same hydrological physical processes. With such a basis, as excepted, the correction of other variables is minimally affected by Gaussian white noise in runoff.
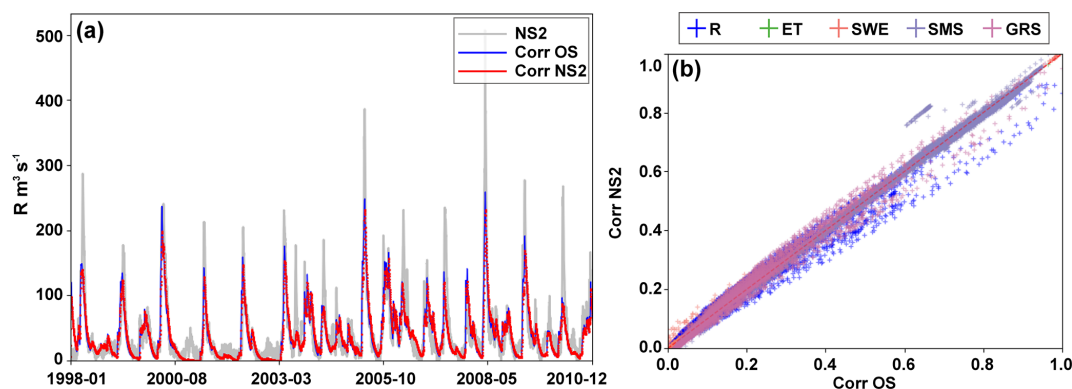
Figure 10. Correction results for multisource datasets corresponding to noise experiment 4. **(a)** Comparison among the runoff noise sequence (NS2), OS-based runoff-corrected sequence (corr OS), and NS2-based runoff-corrected sequence (corr NS2). **(b)** Comparison of multivariable between OS-based correction and NS2-based correction in terms of standardized values.

In summary, the results yielded from the above experiments indicate that both single-point noise and Gaussian white noise have minimal impact on the corrections. The final correction results are constrained by the hydrological model, with random errors in measurements not altering the allocation of water budget residuals significantly. The physical relationships among various water budget variables, as represented by the model, are also imposed onto the measurements through the correction process.

4.3.3 Comparison with existing correction methods

Previous analyses and experiments clarify the unique characteristics of the PHPM-MDCF, which imposes closure constraints based on physical hydrological 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 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 runoff (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 were 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 the PHPM-MDCF) for three main water budget components at basin no. 1539000. Note that the measurements of precipitation are not compared here as the PHPM-MDCF does not perform correction for this variable. It is clear from the figure that both the PHPM-MDCF and CEnKF method exhibit minimal correction of ET, whereas the PR method expands the range of ET significantly, 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). However, 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 and shows limited correction of TWSC (Fig. 11). This is linked to the computational mechanism underlying CEnKF, where the Kalman gain – or the error covariance between measurements and the ensemble mean of the multisource datasets – determines the magnitude of the residuals corrected for each variable. Specifically, the measurements of R to be corrected are based on in situ observations, 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 a greater allocation of correction for R. On the contrary, measurements of TWSC are limited and primarily derived from GRACE, which results in relatively small error covariance and, consequently, smaller corrections. Furthermore, as previously noted, such methods may generate unreasonable corrections due to the propagation of extreme errors, such as the negative R values in Fig. 11b, which are more likely to occur in small basins. The PHPM-MDCF avoids these issues by considering physical-process constraints, leading to more reasonable corrections. Additionally, it does not rely on multisource datasets and can perform corrections on any model time step and for any model output variable. The TWSC derived from SWE and SM is consistent with GRACE TWSC, which also demonstrates the reliability of this framework in retrieving TWSC. The comparison results for the other two represen-

tative basins are shown in Figs. S11–12, leading to similar conclusions.

4.4 Potential influencing factors of water budget residuals

4.4.1 Factors influencing spatial distribution

In this section, we conducted a preliminary exploration of the potential factors influencing the formation and distribution of water budget residuals. 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 revealed similar spatial patterns (Zhang et al., 2016; Gordon et al., 2022). Therefore, we speculate that the spatial distribution of water budget closure is predominantly influenced by the characteristics of the basin.

Here, we focus on the total water budget residuals (i.e. Res) and attempt to relate them with the hydro-meteorological conditions and the basin area. To bring out these relationships, from Fig. 12, three regression curves are obtained by correlating mean absolute residuals at different timescales with basin areas over 475 CAMELS basins. The negative gradients of the curves imply a scale effect in the water budget non-closure phenomenon where, as basin area increases, the water balance constructed from multisource datasets can be enhanced. Moreover, as expected, hydro-meteorological conditions within the basin play a crucial role in controlling the distribution of water budget residuals. The clear delineation between different levels of daily precipitation and the runoff coefficient revealed in Fig. 12 strongly supports this reasoning, where multisource datasets yield larger water budget residuals in basins with high precipitation and runoff coefficients – large red spots are located in the upper portion of the figure. These results highlight the risks of using multisource datasets for hydrological inference in humid and small-scale basins – specifically, potential physical inconsistencies – and underscore the need to carefully test the water balance assumption.

4.4.2 Factors influencing temporal distribution

The pronounced seasonal pattern of non-closure residuals depicted in Fig. 5 is quite interesting. To gain more insight into the observed pattern, we compare it with the temporal factors reported in the literature. The first and foremost reported factor associated with the observed negative biases in Res during the cold season is the underestimation of precipitation (Newman et al., 2015). This systematic bias is related to phenomena, such as snowfall, freez-ing rain, and non-convective precipitation, that occur during the cold season, where measurements and simulations are prone to show significant errors, including the well-know undercatch phenomenon (Kauffeldt et al., 2013; Robinson and Clark, 2020). Another key factor influencing water budget non-closure is connected to the temperature and evaporation dynamics. Abolafia-Rosenzweig et al. (2020) evaluated the water budget residuals over 24 global basins and found that the likelihood of positive biases in the water balance increases with rising temperatures, likely induced by the potential uncertainties in evaporation estimates. The research by Lv et al. (2017) also supports this perspective, indicating that the underestimation of evaporation is a primary contributor to the water budget non-closure. In summary, according to the literature, cold-season precipitation and warm-season evaporation seem to be the primary drivers of the temporal distribution of Res. To examine this reasoning, while obtaining the true values is impossible, we can provide evidence by comparing evaporation and precipitation, along with the corresponding residuals, between the cold and warm seasons.

Figure 13 depicts the relationship by separately comparing the ratios of evaporation and precipitation for the cold and warm seasons with the corresponding water budget residuals. For the cold season, the scatter points can be split into two distinct regions along the vertical line where the ratio is 1. The scatter points in the left region indicate basins where cold-season precipitation is lower than in the warm season, leading to relatively smaller absolute residuals (clustered around zero residuals). In contrast, scatter points for basins with dominant cold-season precipitation are dispersed below the zero residual line, with larger negative residuals becoming more prevalent as the proportion of cold-season precipitation increases. In other words, regions where cold precipitation constitutes a larger proportion of the water budget residuals are more sensitive to the underestimations of precipitation, resulting in larger negative residuals. Furthermore, we observed similar trends in the warm season, where a higher proportion of warm-season evaporation is associated with larger positive residuals (the red dots exhibit an upward trend to the right). These results confirm the perspective of previous research, highlighting the potential uncertainties in measurements of cold-season precipitation and warm-season evaporation.

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 which factors drive the proportions of the residual components.

Res reflects the degree to which the measurements achieve water budget closure. In this study, we argue that two key
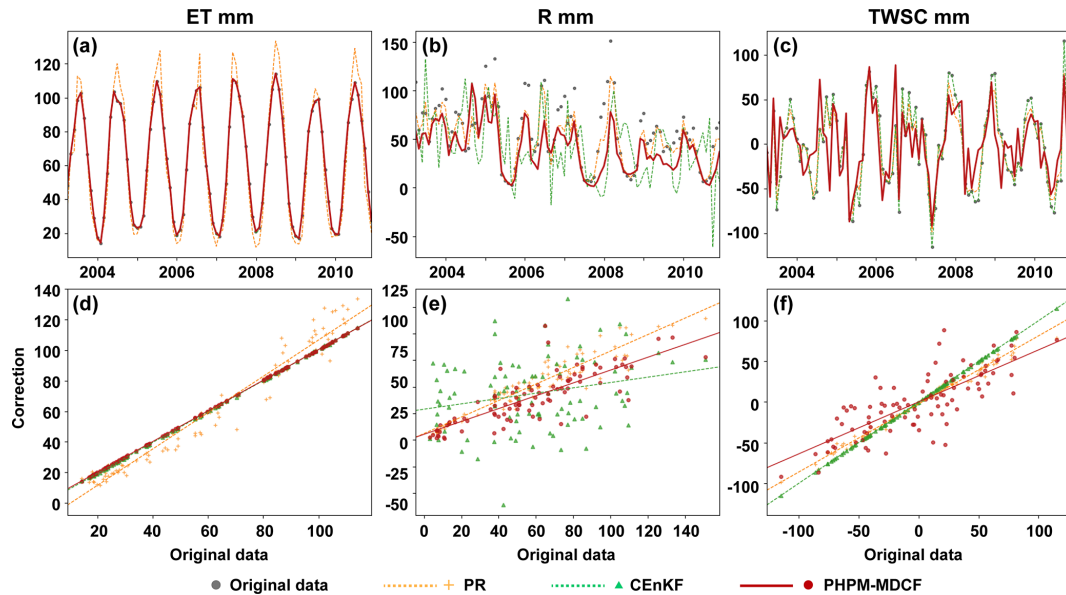
Figure 11. Comparison of monthly correction results between the PHPM-MDCF and existing methods (PR and CEnKF) at basin no. 1539000. **(a–c)** Time series of the original and corrected measurements of evaporation, runoff, and terrestrial water storage change. **(d–f)** Scatterplots and regression lines of the original and corrected measurements.
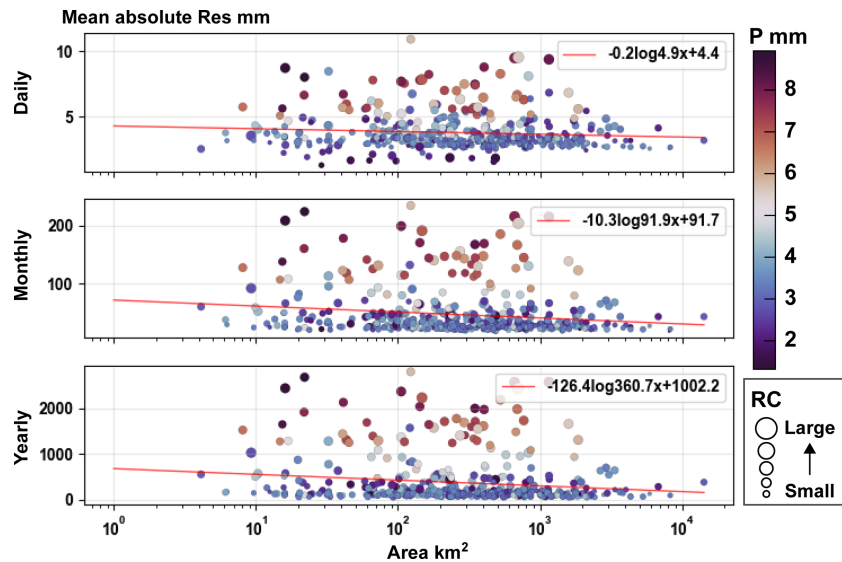

Figure 12. Relationship between the mean absolute of water budget residuals, basin area, long-term average daily precipitation, and runoff coefficient (RC) over 475 CAMELS basins with reliable simulations. The respective red lines represent the linear regression of residuals with basin area for each timescale.

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 used 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 through either 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.
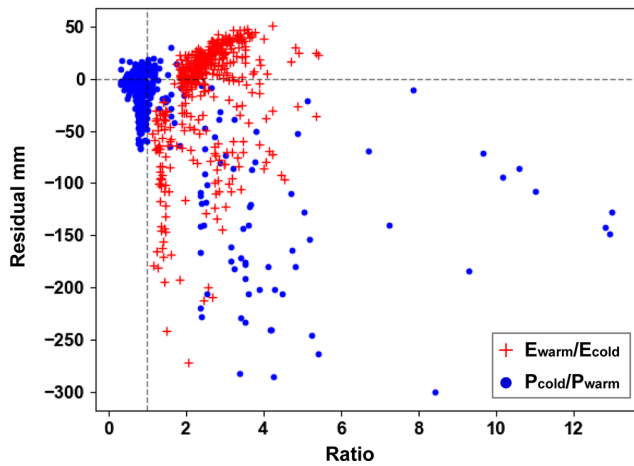
Figure 13. Relationship between the ratios of evaporation and precipitation for the cold and warm seasons separately and the corresponding water budget residuals. Note that blue represents residuals for the cold season, and red represents those for the warm season. The seasonal divisions are consistent with Fig. 5. The unit of the residuals is mm.

Consider the fact that, if our description of the water budget equation is incomplete and omits 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 14 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 compared 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. 14b). Such differences reveal that omitting a 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. 14d–f). Moreover, the spatiotemporal distribution of Res_o has changed (Figs. S13–14). Notably, during the cold season (December to February), the proportion of Res_o is much higher and exhibits a significant positive trend. These findings align with our definition of Res_o, which refers to the water imbalance caused by omitted water. It also supports the validity of our decomposition method to some extent and highlights the importance of a comprehensive water budget equation in evaluating water balance.

5 Discussion

5.1 What lies within the realm of belief

The foundation of modern experimental science is based on empiricism, emphasizing the repeatability of experiments, i.e. whether the results can perfectly reproduce observations. This idea has far-reaching implications across various fields, with a classic example being hydrologists always aiming for their model predictions to closely match observations. Importantly, the underlying assumption of this approach is that our observations perfectly approximate reality and can be seen as true values. In most small-scale studies, such as those conducted in laboratory or field settings, this might hold true. However, as we shift our focus to larger spatial scales, obtaining observations directly often becomes challenging, thus necessitating a reliance on indirect observations, which could potentially undermine this assumption. As a consequence, our confidence in the observations – better referred to as measurements – may diminish, which is precisely the new challenge we face in the era of big data.

When we lack sufficient confidence in any single measurement, the utilization of multisource data fusion becomes a method to mitigate errors from all sources of measurements, thereby reducing uncertainty. Within the process of data fusion, the basic step is to determine the weights of all components. The ensemble mean method assumes an equal weight for all components, while the simple weighted method estimates weights based on the a priori uncertainties, which are typically the differences between each component and the average of all measurements (Sahoo et al., 2011). In the widely used triple collocation (TC) method, weights can be determined by calculating errors (uncertainties) based on the similarity of the triplet inputs without the need for "ground truth" (Stoffelen, 1998). Some other methods also determine uncertainty through manually assigned constants or error propagation calculations (Munier et al., 2014; Ansari et al., 2022). However, all of these methods face the same issue: the true value may be unattainable, and the determined error or uncertainty involves subjective factors. This presents a logical paradox: we resort to data fusion due to the absence of a true value, yet, during the fusion process, we paradoxically assume the existence of this true value to estimate uncertainty. Essentially, we need to answer a fundamental question: what do we truly believe in?

The answer is in what we have truly learned. A better approach is to leverage our existing knowledge about the physical world to enhance our confidence in measurements. In fact, this concept embodies, to some extent, a Bayesian philosophy and is reflected in many fields. Here, we present two modern examples to illustrate this idea. The first one is the atmospheric reanalysis, which has been one of the most significant topics in atmospheric science since the 19th century. This technique employs numerical models and assimilation techniques to integrate multiple types of historical measure-
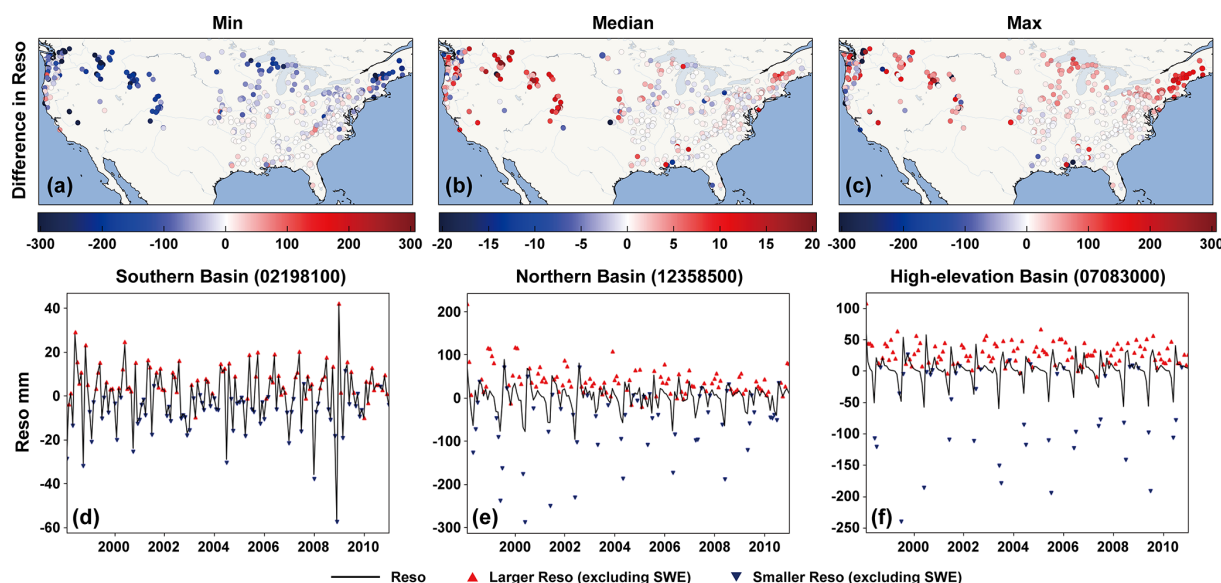
Figure 14. Comparison of Res_o obtained from residual decomposition excluding SWE with the original values. **(a–c)** Spatial distribution of monthly mean Res_o excluding SWE minus its original values. **(d–f)** Time series of Res_o excluding SWE and its original values at the southern basin (02198100, 32.96° N), northern basin (12358500, 48.33° N), and high-elevation basin (07083000, elevation of 3.56 km) at a monthly scale. The unit of the residuals is mm.

ments into a unified modelling framework and assimilation scheme, thereby generating continuous and consistent estimates of climate states. In essence, its aim is to unify our knowledge system (i.e. numerical models) with the measurement system, thereby enhancing the credibility of the model output.

Another example is research in the field of hydrology, where Liao and Barros (2022) proposed an inverse rainfall correction (IRC) framework to improve quantitative precipitation estimates (QPEs) in headwater basins. Their fundamental concept is that errors propagate from precipitation to runoff, enabling the reversal of precipitation errors by calculating runoff simulation errors from distributed hydrological models and applying the travel time distribution for correction. In this example, existing knowledge is represented by the hydrological model, which is assumed to reflect the true physical processes and is then used to enhance the confidence in precipitation measurements.

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 models may lack accuracy – it remains the foundation upon which we can rely and that we can strive to refine in the future. Furthermore, several underlying concepts in this framework, such as residual decomposition and advancing water budget closure through correction, align with a recent study (Wang

and Gupta, 2024). The authors of this study 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). 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.

5.2 Limitations and paths forward

It is our opinion that some traditional hydrological inferences are based on a philosophy that involves some long-standing and problematic assumptions that arise from the unwarranted confidence in measurements. However, the fact that truth is almost impossible to measure due to the complexity of real-world physical processes hampers the foundation of inferences, especially in large-scale studies that employ multi-source non-field data. The presented framework has advantages in its integration of the widely applicable water budget equation and its reliable representation of hydrological process using a hydrological model, significantly mitigating this issue, and enhances our confidence in the corrected datasets. Although the efficiency and credibility of the PHPM-MDCF have been examined in the previous sections, there are several limitations and uncertainties worthy of further discussion.

5.2.1 Uncertainty of forcing data

Here, we return to hypothesis 2 posed at the beginning of the Methods section. As we acknowledge, the uncertainties arising from the forcing and model structure undeniably exist and were a limitation in this study. First, the uncertainty in the forcing may arise from two aspects; one is the inaccuracy of the datasets themselves, and the other is the uncertainty introduced by the scaling process (i.e. the conversion from grid scale to basin scale). To investigate the sensitivity of correction results to forcing data, we re-conducted multisource dataset correction using Daymet precipitation data at the same case basin (no. 1013500) and compared it with the original correction (forcing by TRMM). The comparison of the two precipitation products is presented in Fig. S15, where Daymet precipitation is significantly lower. The top panels of Fig. 15 display slight differences between the two corrections; for instance, the Daymet correction shows a larger SWE (with a slope greater than 1), while other variables are smaller. These differences can be entirely explained by variations in precipitation forcing. Nevertheless, the temporal patterns of all variables under the two corrections remain broadly consistent, with determination coefficients of all regression curves exceeding 0.70 (Fig. 15b). Theoretically, the consistency of correction stems from three aspects. Firstly, it is attributed to the adaptability of hydrological models 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 a reasonable representation of hydrological processes 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 residual correction, thereby mitigating the propagation of bias between variables. Thirdly, the uncertainty caused by the mismatch between the grids and basin boundaries is effectively alleviated through the unit conversion (i.e. from volume to depth units). These three features ensure the stability of the correction, rendering it less susceptible to interference from uncertainties in the forcing datasets.

Further evidence of the robustness of the PHPM-MDCF is provided by Fig. 15c–d, where corrected residuals tend to converge after several iterations despite being forced by different precipitation datasets. The main influence of forcing data is manifested in the omission residuals. As expected, the omission residual term is simply an approximation of the missing water fluxes or storages in the water budget equation, which can vary depending on the datasets chosen to characterize the equation. In Fig. 15e, the omission residuals driven by Daymet stabilize around 12.5 mm, whereas those driven by TRMM stabilize around 6.5 mm. Such a discrepancy can be further highlighted in the comparison of the residual time series (Fig. S16). Further investigation would be required to better understand the omission residuals from a physical perspective. For example, a distributed hydrological model with a representation of subsurface-layer flow processes will allow us to identify the magnitude of inter-basin interactions; a more detailed description of the water budget equation in data-rich environments can help us examine the sources of omission errors. This is undoubtedly important but is not the focus here. In summary, the above results suggest that the correction is minimally sensitive to the choice of forcing, demonstrating the robustness of the correction results. This is achieved by maintaining similar inconsistency residuals – corresponding to a similar correction amount – as long as differences in precipitation do not result in substantial variations in the hydrological processes.

It is noted that the PHPM-MDCF has limitations in addressing inconsistency residuals in forcing. The reasons are twofold. On the one hand, this is due to our neglect of uncertainties in the forcing, which, as indicated by the above analysis, appears to have a limited impact on the correction for other variables. On the other hand, this is because the PHPM-MDCF allocates residuals based on the distance between simulations and measurements, while the forcing cannot be simulated within the hydrological model. In this case, is there potential to correct the inconsistency residuals in the forcing? Clues to this possibility are hidden in the above analysis. Systematic biases in precipitation products are directly reflected in the water budget equation, leading to different total input water volumes. Consequently, with the inconsistency residuals of other variables unchanged, maintaining the water balance would require an increase in omission residuals (Fig. 15e). Therefore, it can be inferred that, with other variables unchanged, TRMM demonstrates superior water budget closure compared to Daymet, which contains smaller inconsistency residuals. In other words, the difference in the two omission residuals reflects the discrepancy in inconsistency residuals contained within the two precipitation products. This portion of the omission residual difference can be directly corrected in the precipitation. However, it is worth noting that not all omission residuals can be corrected in the precipitation as it still contains residuals from some unknown omitted water content. Such correction must be relative and based on comparisons between different precipitation products as true values and perfect water balance equations are unattainable. Another strategy is to couple an atmospheric model with this framework to generate simulated precipitation, allowing for the correction of precipitation products. In subsequent work, we will explore these approaches and try to extend the PHPM-MDCF based on these ideas.

5.2.2 Uncertainty of model structure

The characterization of physical hydrological processes through modelling constitutes the foundation of the correction framework. The internal model structure is the primary constraint for achieving water budget closure, and, thus, it is crucial for the final correction results. The selection of
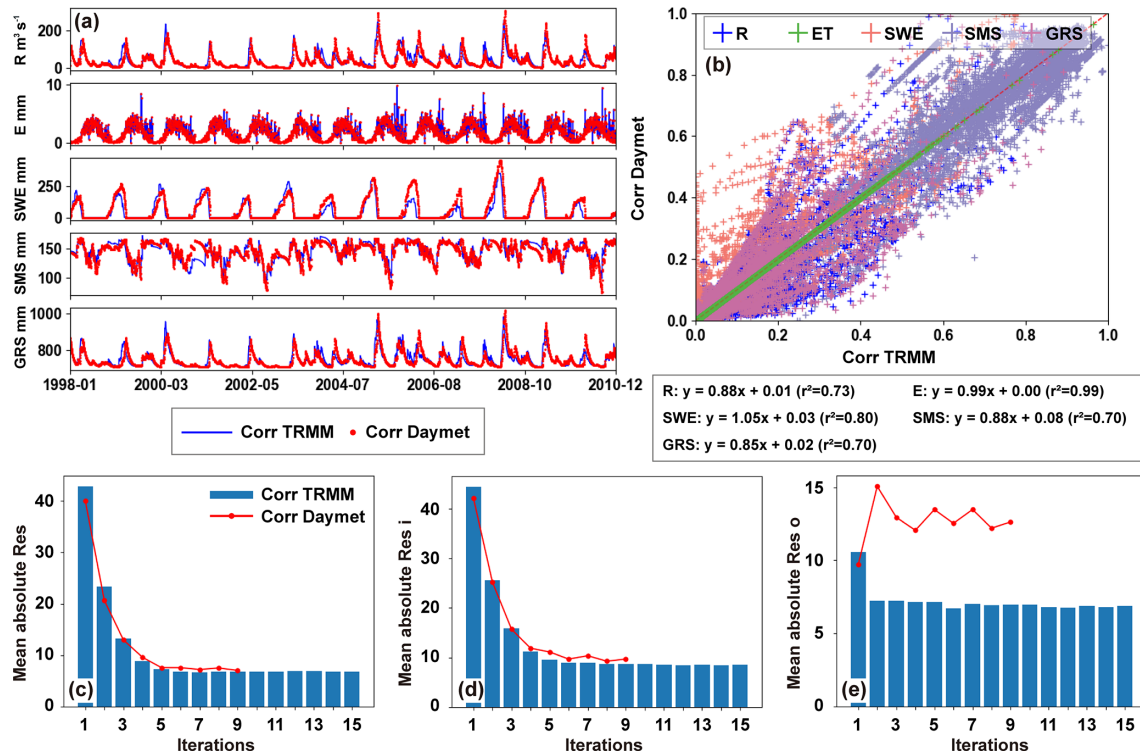
Figure 15. Comparison of correction results based on different forcing datasets (TRMM and Daymet) at basin no. 1013500. **(a–b)** Corrected time series of five water budget variables. **(c–e)** Variation in long-term mean absolute values of three residuals with correction iterations at the monthly scale. The unit of the residuals is mm.

the lumped model (i.e. the HBV model) is intended to facilitate the application in large-sample basins to derive more general conclusions, as has also been done in many previous large-sample hydrology studies (Gupta et al., 2014). The reliability of model simulations has been confirmed by multi-objective evaluation. However, whether the spatial distribution of model performance is intrinsically related to the model structure is crucial to the robustness of the current work.

To address the question, we first compared the model performance with other studies that employed different models. As illustrated in Fig. C1, the model behaviour exhibits strong spatial organization, with unreliable simulations primarily concentrated in the central and western regions of CONUS. This spatial distribution of prediction skill broadly agrees with many previous studies. Brunner et al. (2021) classified this region as an intermittent regime and attribute the unsatisfactory simulation to the complex day-to-day variation in runoff. In their work, all four lumped models with different structures (i.e. SAC, HBV, VIC, mHM) supported the inference. In Yan et al. (2023), a more complex land surface model (i.e. CLM5) was utilized for evaluating the uncertainty of runoff prediction; they reported that the southwestern and central US showed the poorest prediction skill. Notable pioneering research was conducted by Knoben et al. (2020), who evaluated runoff predictability in CAMELS

basins using 36 hydrological models with different structures. After conducting a comprehensive analysis, they generated a multi-model runoff prediction performance map, which aligns closely with the results of this study. Therefore, we deduce that the spatial disparities in model performance or predictability predominantly depend on basin and climatic conditions rather than on model structure. The consistency of the model performance with prior studies demonstrates that the HBV model is reliable in the context of this study.

To further substantiate the above inference, we categorized basins into four groups based on model performance in runoff and compared the inter-group differences in six types of basin and climatic characteristics (i.e. climate, hydrology, geology, topography, soil, and vegetation). The four groups consist of unreliable performance, reliable performance, below-average performance, and above-average performance. First, the two-sample t test at the 5 % level was conducted to examine whether there are significant differences in each characteristic indicator between the unreliable and reliable groups. The indicators exhibiting a statistically significant difference were then presented and compared in Figs. S17 and S18. For clarity, here, we list indicators whose inter-group difference was greater than 30 % in terms of median cumulative probability: mean precipitation, mean potential evapotranspiration, aridity index (climate), proportion of silt (geology–soil), mean runoff, runoff coefficient, fre-

quency of high-flow days (hydrology), and all vegetation indicators (vegetation). The significant inter-group differences in these indicators highlight critical basin and climatic characteristics pivotal to the successful modelling of the hydrology system, providing convincing evidence for our inference. In summary, basins with the following characteristics typically pose challenges to simulation: arid regions with low precipitation and high potential evaporation, resulting in a low runoff ratio and frequent alternation between zero flow and high flow. Vegetation in these basins tends to consist of lower vegetation types and to lack forests. It is worth noting that, while we have validated the reliability of the HBV model in the current study, its simplistic physics and lumped design structure lead to significant limitations in simulating several processes such as snow and groundwater (Brunner et al., 2021). In other words, the HBV model may not be suitable for accurately representing the reality of these specific processes.

The distinctive perspective of this work lies in utilizing the physical processes described by hydrological models to constrain multisource datasets, thereby enhancing water budget closure among them. In particular, our next priority is to incorporate more complex models to examine the PHPM-MDCF in different basins with specific hydro-meteorological conditions. For instance, distributed hydrological models and hybrid models (ML-HM) are valuable tools that can improve our understanding of water budget closure through more detailed physical-process representation (Liao and Barros, 2022; Wang and Gupta, 2024). By employing models that generate additional output variables, we can more comprehensively represent the water budget equation and extend the application of the PHPM-MDCF to more complex water budget systems. Additionally, multiple models can be utilized for ensemble correction, which aids in quantifying uncertainty and providing more robust correction results.

6 Conclusions

Advanced measurement techniques open new opportunities for modern hydrological research. However, due to the lack of consistent data production protocols and evaluation standards, physical inconsistencies are prevalent in multisource datasets in the form of water budget residuals. Such inconsistencies undermine our confidence in data reliability and compromise the robustness of hydrological inferences relying on these datasets. In this study, we proposed a multisource dataset correction framework, the PHPM-MDCF, to achieve water budget closure through physical hydrological process modelling. Built upon the decomposition of total water residuals and the iterative multi-objective calibration, the framework has the ability to reduce the inconsistency residuals among multisource datasets and to promote convergence between the simulation and measurement systems. We demonstrated the spatiotemporal distribution of water bud-

get residuals and the efficiency of the PHPM-MDCF across 475 CONUS basins selected by hydrological simulation reliability. Several experiments were conducted to verify the credibility of the framework, including the addition of manual noises and comparisons with existing correction methods. Furthermore, we explored potential factors influencing the spatiotemporal distribution and proportions of residuals. The major study findings are summarized as follows:

1. The results from water budget residual decomposition indicate that inconsistency residuals dominate the total water budget residuals, showing highly consistent spatiotemporal distributions. In spatial terms, both demonstrate an east–west gradient and concentrations of low values along the western coastline and eastern inland basins within CONUS. Temporally, they exhibit negative trends in the cold seasons and positive trends in the warm seasons. On the contrary, the omission residuals, which account for the water quantities omitted in the original water budget equation, have different drivers and thus exhibit distinct distributions compared to the former. This component constitutes a relatively small proportion of the total budget residuals.

2. The PHPM-MDCF demonstrates satisfactory correction efficiency, with an average reduction percentage of 49 % in total water budget residuals across all 475 basins after correction. In certain basins, this reduction can exceed 80 % (i.e. 84 % in basin no. 1013500). The correction efficiency shows a latitudinally dependent pattern, with greater absolute values in high-latitude regions. The results from noise experiments validated the credibility of the correction framework. Both single-point extreme-noise and Gaussian white-noise sequences exert a limited impact on final correction results. Corrections applied to extreme noises in one variable do not propagate to others, thereby avoiding the generation of unreasonable values. Its credibility was further substantiated through comparisons with existing methods.

3. The water budget non-closure phenomenon exhibits noticeable scale effects and is closely related to hydro-meteorological conditions. This highlights the need for careful consideration of the water balance assumption when applying multisource datasets for hydrological inference in small and humid basins. Moreover, the underestimation of cold-season precipitation and warm-season evaporation could be directly associated with the negative and positive biases in water budget residuals for the corresponding seasons. As a foundation for evaluating the water balance, a comprehensive water budget equation is undoubtedly crucial, as underscored by the analysis of residual proportions.

For the first time, this study presents a correction approach to achieve water budget closure based on physical hydrolog-

ical modelling. However, the Bayesian philosophy underlying the approach has been implicit in many previous methods, such as atmospheric reanalysis. The only thing we can rely on is our prior knowledge; therefore, continuously promoting convergence between knowledge and measurement systems is crucial for enhancing our confidence. An obvious extension of this research is the inclusion of more disciplines within both atmospheric sciences and broader Earth sciences. This contributes to a better understanding, in the era of big data, of the distinctions and correlations between simulations, measurements, and reality.

Appendix A: Implementation details of the HBV model

Figure A1 illustrates the basic structure of the HBV model, encompassing three modules (i.e. snow routine, soil moisture routine, and runoff routine) and three runoff components: quick runoff, interflow, and baseflow. The cumulative sum of these components constitutes total runoff, which is routed through a triangular unit hydrograph (UH). At each model run step, the runoff at the outlet of the basin is determined. The HBV model is driven by daily precipitation (from TRMM), average temperature (from CAMELS), and potential evaporation (from GLEAM), enabling the simulation of various hydrological fluxes and state variables, including runoff, soil moisture storage, groundwater reservoir storage, evaporation, and SWE. Table A1 lists the free parameters slated for calibration in the HBV model, providing their descriptions and respective ranges.

The period from 1998 to 2000 is looped five times for model spin-up, and the subsequent 10-year period is used for model calibration. After each calibration, the optimal parameter set is selected from the Pareto fronts. Finally, these optimal parameters are applied to the entire 12-year period to yield the best simulation, thus facilitating the multisource dataset correction.
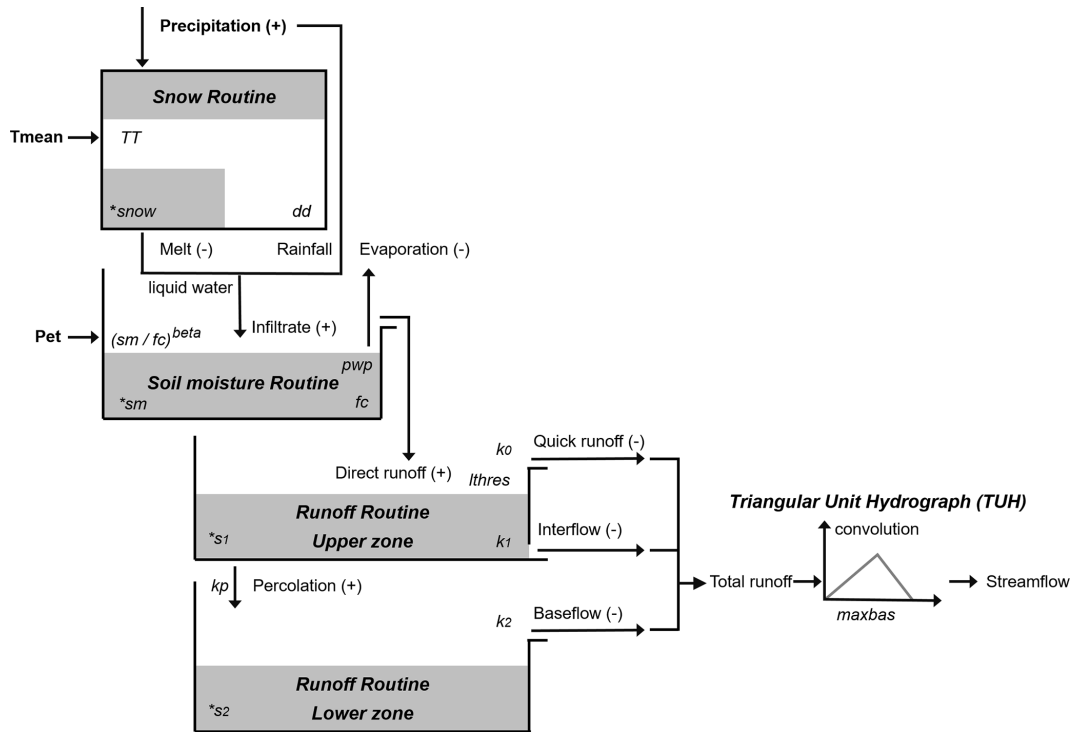
Figure A1. Schematic structure of the HBV model. The variables marked with an asterisk (*) denote water storage, whereas those annotated with positive (+) and negative (−) signs represent the inputs and outputs of the storage.

Table A1. The description and ranges of free parameters in the HBV model for calibration.

Parameter	Unit	Description	Min	Max
DD	$\mathrm{mm\,^{\circ}C^{-1}\,d^{-1}}$	Degree-day factor	1.0	10.0
TT	°C	Threshold temperature for snowmelt initiation	−2.5	2.5
Beta	–	Shape coefficient	1.0	8.0
FC	mm	Field capacity	10.0	600.0
K_0	$\mathrm{d^{-1}}$	Recession coefficient of the quick runoff	0.1	0.8
K_1	$\mathrm{d^{-1}}$	Recession coefficient of the interflow	0.01	0.5
K_2	$\mathrm{d^{-1}}$	Recession coefficient of the baseflow	0.001	0.15
K_p	$\mathrm{d^{-1}}$	Recession coefficient of the percolation	0.001	5.0
PWP	–	Soil permanent wilting point as a fraction of FC	0.2	1.0
HL	mm	Threshold water level for near-surface flow	10.0	200.0
maxbas	d	Weighting parameter of triangular unit hydrograph	1	10

Appendix B: Evaluation metrics used for model calibration

The Kling–Gupta efficiency (KGE) metric provides a comprehensive measure of the similarity between simulations and measurements by incorporating three components: correlation, the ratio of standard deviations, and the ratio of means. It has been demonstrated to exhibit superior performance in calibrating hydrological models (Knoben et al., 2020; Aerts et al., 2022). The Pearson correlation coefficient (r) quantifies the extent of shared information between simulations and measurements, characterized by its insensitivity to amplitude and mean values (Lorenz et al., 2014). Thus, it is suitable for evaluating variables that may exhibit mean differences between simulations and measurements, such as SMS and GRS. The root mean square error (RMSE) is a widely used evaluation metric in hydrological modelling. Despite it not being a normalized metric, its calculation does not in-

Table B1. Description of evaluation metrics, including ranges and optimal values.

Metrics	Full name	Variables to be evaluated	Range	Optimal value
KGE	Kling–Gupta efficiency	Runoff, evaporation	$(-\infty, 1]$	1.0
r	Pearson correlation coefficient	Soil moisture storage, groundwater reservoir storage	$[-1, 1]$	1.0
RMSE	Root mean square error	Snow water equivalent	$[0, +\infty)$	0.0

volve division, making it particularly suitable for evaluating variables like SWE, which may be a sequence consisting entirely of zeros. Based on the simulated and measured values of the target variables, the three metrics can be calculated using the following formulas:

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + (\frac{\sigma_{\text{sim}}}{\sigma_{\text{obs}}} - 1)^2 + (\frac{\mu_{\text{sim}}}{\mu_{\text{obs}}} - 1)^2}, \quad \text{(B1)}$$

$$r = \frac{\sum_{i=1}^{n}(V_{\text{obs}}^i - \overline{V}_{\text{obs}})(V_{\text{sim}}^i - \overline{V}_{\text{sim}})}{\sqrt{\sum_{i=1}^{n}(V_{\text{obs}}^i - \overline{V}_{\text{obs}})^2}\sqrt{\sum_{i=1}^{n}(V_{\text{sim}}^i - \overline{V}_{\text{sim}})^2}}, \quad \text{(B2)}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(V_{\text{sim}}^i - V_{\text{obs}}^i)^2}, \quad \text{(B3)}$$

where σ is the standard deviation, and μ is the mean; V^i is the target variable at time step i, and n is the length of the sequence. The subscripts "sim" and "obs" denote the simulation and measurement of the variable, respectively. The range and optimal values of the evaluation metrics are shown in Table B1.

Appendix C: Simulation performance of the HBV model across CAMELS basins

In this appendix, we present the simulation performance of the HBV model on 653 CAMELS basins. As shown in Fig. C1, the performance of five target variables, including runoff, evaporation, soil moisture storage, groundwater reservoir storage, and snow water equivalent, is described using three metrics (i.e. KGE, r, and RMSE). The gradient from white to deep blue indicates progressively better simulation performance. In contrast, red highlights basins of unreliable simulation, determined by a KGE of less than -0.41 and an r value failing the significance test at the 5 % level. Table C1 summarizes the multivariable simulation performance of the HBV model across all basins.
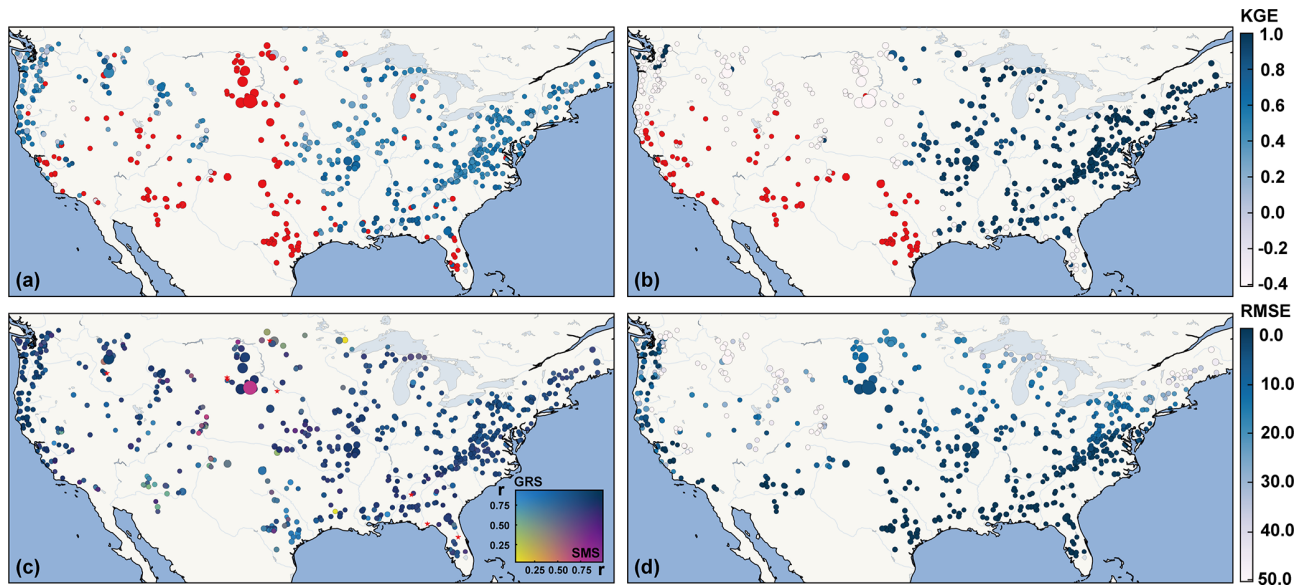
Figure C1. The multi-objective simulation performances of the HBV model across the CAMELS basins. Results are based on **(a)** runoff, **(b)** evaporation, **(c)** soil moisture storage and groundwater reservoir storage, and **(d)** snow water equivalent. Red dots represent unreliable simulation performance, and the size of the points is proportional to the basin area. The unit of the RMSE is mm.

Table C1. Performances of the HBV model in terms of five target variables across the CAMELS basins. The last row presents the number and proportion of basins where all target variables are reliably simulated. The unit of RMSE in the table is "mm".

Variables	Median performance (KGE, r, RMSE)	Range (KGE, r, RMSE)	Reliable simulation Count (basins)	Reliable proportion (%)
Runoff	0.50	−0.40 to 0.88	499	76.42 %
ET	0.94	−0.40 to 0.99	548	83.92 %
SMS	0.80	0.07 to 0.95	645	98.77 %
GRS	0.72	0.02 to 0.95	653	100.00 %
SWE	5.97	0.00 to 353.34	–	–
All variables	–	–	475	72.74 %

Author contributions. XDZ: conceptualization, data curation, formal analysis, writing (original draft). DFL: conceptualization, supervision, writing (review and editing). SZH, HW, XMM: supervision and review.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical rep-

References

Abhishek, Kinouchi, T., Abolafia-Rosenzweig, R., and Ito, M.: Water Budget Closure in the Upper Chao Phraya River Basin, Thailand Using Multisource Data, Remote Sensing, 14, 173, https://doi.org/10.3390/rs14010173, 2022.

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 Sens. Environ., 252, 112191, https://doi.org/10.1016/j.rse.2020.112191, 2020.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017 (data available at: https://ncar.github.io/hydrology/datasets/CAMELS_attributes, last access: 23 October 2022).

Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., and Hazenberg, P.: Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model, Hydrol. Earth Syst. Sci., 26, 4407–4430, https://doi.org/10.5194/hess-26-4407-2022, 2022.

Ansari, R., Liaqat, M. U., and Grossi, G.: Evaluation of gridded datasets for terrestrial water budget assessment in the Upper Jhelum River Basin-South Asia, J. Hydrol., 613, 128294, https://doi.org/10.1016/j.jhydrol.2022.128294, 2022.

Balsamo, G., Viterbo, P., Beljaars, A., Hurk, B., Hirschi, M., Betts, A., and Scipal, K.: A Revised Hydrology for the ECMWF Model: Verification from Field Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, J. Hydrometeorol., 10, 623, https://doi.org/10.1175/2008JHM1068.1, 2009.

Bergström, S.: Development and Application of a Conceptual Runoff Model for Scandinavian Catchments, Hydrology and Oceanography, PhD thesis, Swedish Meteorological and Hydrological Institute (SMHI), Norköping, Sweden, http://urn.kb.se/resolve?urn=urn:nbn:se:smhi:diva-5738 (last access: 11 October 2023), 1976.

Beven, K.: Towards an Alternative Blueprint for a Physically Based Digitally Simulated Hydrologic Response Modeling System, Hydrol. Process., 16, 189–206, https://doi.org/10.1002/hyp.343, 2002.

Beven, K.: Benchmarking Hydrological Models for an Uncertain Future, Hydrol. Process., 37, e14882, https://doi.org/10.1002/hyp.14882, 2023.

Brunner, M. I., Melsen, L. A., Wood, A. W., Rakovec, O., Mizukami, N., Knoben, W. J. M., and Clark, M. P.: Flood spatial coherence, triggers, and performance in hydrological simulations: large-sample evaluation of four streamflow-calibrated models, Hydrol. Earth Syst. Sci., 25, 105–119, https://doi.org/10.5194/hess-25-105-2021, 2021.

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 Resour. Res., 57, e2020WR029001, https://doi.org/10.1029/2020WR029001, 2021.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II, IEEE T. Evolut. Comput., 6, 182–197, 2002.

Feng, D., Liu, J., Lawson, K., and Shen, C.: Differentiable, Learnable, Regionalized Process-Based Models With Multiphysical Outputs can Approach State-Of-The-Art Hydrologic Prediction Accuracy, Water Resour. Res., 58, e2022WR032404, https://doi.org/10.1029/2022WR032404, 2022.

Fortin, F.-A., De Rainville, F.-M., Gardner, M. A., Parizeau, M., and Gagné, C.: DEAP: Evolutionary algorithms made easy, J. Mach. Learn. Res., 13, 2171–2175, 2012.

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 Resour. Res., 58, e2021WR030966, https://doi.org/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, 2009.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18, 463–477, https://doi.org/10.5194/hess-18-463-2014, 2014.

Gutenstein, M., Fennig, K., Schröder, M., Trent, T., Bakan, S., Roberts, J. B., and Robertson, F. R.: Intercomparison of freshwater fluxes over ocean and investigations into water budget closure, Hydrol. Earth Syst. Sci., 25, 121–146, https://doi.org/10.5194/hess-25-121-2021, 2021.

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, Agr. Forest Meteorol., 341, 109615, https://doi.org/10.1016/j.agrformet.2023.109615, 2023.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., and Adler, R. F.: TRMM (TMPA) Precipitation L3 1 day 0.25 degree × 0.25 degree V7, GES DISC [data set], https://doi.org/10.5067/TRMM/TMPA/DAY/7, 2016.

Kabir, T., Pokhrel, Y., and Felfelani, F.: On the Precipitation-Induced Uncertainties in Process-Based Hydrological Modeling in the Mekong River Basin, Water Resour. Res., 58, e2021WR030828, https://doi.org/10.1029/2021WR030828, 2022.

Kauffeldt, A., Halldin, S., Rodhe, A., Xu, C.-Y., and Westerberg, I. K.: Disinformative data in large-scale hydrological modelling, Hydrol. Earth Syst. Sci., 17, 2845–2857, https://doi.org/10.5194/hess-17-2845-2013, 2013.

Kittel, C. M. M., Nielsen, K., Tøttrup, C., and Bauer-Gottwein, P.: Informing a hydrological model of the Ogooué with multi-mission remote sensing data, Hydrol. Earth Syst. Sci., 22, 1453–1472, https://doi.org/10.5194/hess-22-1453-2018, 2018.

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 Resour. Res., 56, e2019WR025975, https://doi.org/10.1029/2019WR025975, 2020.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Lehmann, F., Vishwakarma, B. D., and Bamber, J.: How well are we able to close the water budget at the global scale?, Hydrol. Earth Syst. Sci., 26, 35–54, https://doi.org/10.5194/hess-26-35-2022, 2022.

Liao, M. and Barros, A.: Toward optimal rainfall – Hydrologic QPE correction in headwater basins, Remote Sens. Environ., 279, 113107, https://doi.org/10.1016/j.rse.2022.113107, 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, J. Hydrometeorol., 15, 2111–2139, https://doi.org/10.1175/JHM-D-13-0157.1, 2014.

Luo, Z., Shao, Q., Wan, W., Li, H., Xi, C., Zhu, S., and Ding, X.: A new method for assessing satellite-based hydrological data products using water budget closure, J. Hydrol., 594, 125927, https://doi.org/10.1016/j.jhydrol.2020.125927, 2020.

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 Resour. Res., 59, e2022WR032176, https://doi.org/10.1029/2022WR032176, 2023.

Luojus, K., Pulliainen, J., Takala, M., Lemmetyinen, J., Mortimer, C., Derksen, C., Mudryk, L., Moisander, M., Hiltunen, M., Smolander, T., Ikonen, J., Cohen, J., Salminen, M., Norberg, J., Veijola, K., and Venäläinen, P.: GlobSnow v3.0 Northern Hemisphere snow water equivalent dataset, Scientific Data, 8, https://doi.org/10.1038/s41597-021-00939-2, 2021 (data available at: https://www.globsnow.info/swe/, last access: 6 October 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, J. Hydrol., 547, 585–599, https://doi.org/10.1016/j.jhydrol.2017.02.027, 2017.

Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E.,

Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, Geosci. Model Dev., 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017 (data available at: https://www.gleam.eu/, last access: 31 August 2023).

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrol. Earth Syst. Sci., 15, 453–469, https://doi.org/10.5194/hess-15-453-2011, 2011.

Mostafaie, A., Forootan, E., Safari, A., and Schumacher, M.: Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data, Comput. Geosci., 22, 789–814, 2018.

Munier, S., Aires, F., Schlaffer, S., Prigent, C., Papa, F., Maisongrande, P., and Pan, M.: Combining data sets of satellite-retrieved products for basin-scale water balance study: 2. Evaluation on the Mississippi Basin and closure correction model, J. Geophys. Res.-Atmos., 119, 12100–12116, 2014.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021 (data available at: https://doi.org/10.24381/cds.e2161bac).

Nearing, G., Kratzert, F., Sampson, A., Pelissier, C., Klotz, D., Frame, J., Prieto, C., and Gupta, H.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Pan, M. and Wood, E.: Data Assimilation for Estimating the Terrestrial Water Budget Using a Constrained Ensemble Kalman Filter, J. Hydrometeorol., 7, 534–547, https://doi.org/10.1175/JHM495.1, 2006.

Petch, S., Dong, B., Quaife, T., King, R. P., and Haines, K.: Water and energy budgets over hydrological basins on short and long timescales, Hydrol. Earth Syst. Sci., 27, 1723–1744, https://doi.org/10.5194/hess-27-1723-2023, 2023.

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T. E., and Woods, R.: Scaling, similarity, and the fourth paradigm for hydrology, Hydrol. Earth Syst. Sci., 21, 3701–3713, https://doi.org/10.5194/hess-21-3701-2017, 2017.

Refsgaard, J. C., Stisen, S., and Koch, J.: Hydrological process knowledge in catchment modelling – Lessons and perspectives from 60 years development, Hydrol. Process., 36, 1–20, 2022.

Robinson, E. L. and Clark, D. B.: Using Gravity Recovery and Climate Experiment data to derive corrections to precipitation data sets and improve modelled snow mass at high latitudes, Hydrol.

Earth Syst. Sci., 24, 1763–1779, https://doi.org/10.5194/hess-24-1763-2020, 2020.

Sahoo, A., Pan, M., Troy, T., Vinukollu, R., Sheffield, J., and Wood, E.: Reconciling the global terrestrial water budget using satellite remote sensing, Remote Sens. Environ., 115, 1850–1865, https://doi.org/10.1016/j.rse.2011.03.009, 2011.

Sivapalan, M. and Blöschl, G.: The Growth of Hydrological Understanding: Technologies, Ideas and Societal Needs Shape the Field, Water Resour. Res., 53, 8137–8146, https://doi.org/10.1002/2017WR021396, 2017.

Stoffelen, A.: Error Modeling and Calibration; Towards the true surface wind speed, J. Geophys. Res., 103, 7755–7766, https://doi.org/10.1029/97JC03180, 1998.

Tang, G., Clark, M., Papalexiou, S. M., Ma, Z., and Hong, Y.: Have satellite precipitation products improved over last two decades? A comprehensive comparison of GPM IMERG with nine satellite and reanalysis datasets, Remote Sens. Environ., 240, 111697, https://doi.org/10.1016/j.rse.2020.111697, 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, J. Geophys. Res., 114, D12106, https://doi.org/10.1029/2008JD011475, 2009.

Wang, J., Zhuo, L., Han, D., Liu, Y., and Rico-Ramirez, M.: Hydrological Model Adaptability to Rainfall Inputs of Varied Quality, Water Resour. Res., 59, e2022WR032484, https://doi.org/10.1029/2022WR032484, 2023.

Wang, Y. H. and Gupta, H.: A Mass-Conserving-Perceptron for Machine-Learning-Based Modeling of Geoscientific Systems, Water Resour. Res., 60, e2023WR036461, https://doi.org/10.1029/2023WR036461, 2024.

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 Resour. Res., 59, e2022WR033538, https://doi.org/10.1029/2022WR033538, 2023.

Yan, H., Sun, N., Eldardiry, H., Thurber, T., Reed, P., Malek, K., Gupta, R., Kennedy, D., Swenson, S., Hou, Z., Cheng, Y., and Rice, J.: Large Ensemble Diagnostic Evaluation of Hydrologic Parameter Uncertainty in the Community Land Model Version 5 (CLM5), J. Adv. Model. Earth Sy., 15, e2022MS003312, https://doi.org/10.1029/2022MS003312, 2023.

Zhang, Y., Pan, M., and Wood, E.: On Creating Global Gridded Terrestrial Water Budget Estimates from Satellite Remote Sensing, in: Remote Sensing and Water Resources, Space Sciences Series of ISSI, 59–78, https://doi.org/10.1007/978-3-319-32449-4_4, 2016.