The value of observed reservoir storage anomalies for improving the simulation of reservoir dynamics in large-scale hydrological models

Seyed-Mohammad Hosseini-Moghari[1], Petra Döll[1, 2]

5   [1]Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany
[2]Senckenberg Leibniz Biodiversity and Climate Research Centre Frankfurt (SBiK-F), Frankfurt am Main, Germany

Correspondence to: Seyed-Mohammad Hosseini-Moghari (hosseinimoghari@em.uni-frankfurt.de)

Abstract. Human-managed reservoirs alter water flows and storage, impacting the hydrological cycle. Modeling reservoir outflow and storage is challenging because it depends on human decisions, and there is limited access to data on reservoir

10   inflows, outflows, storage, and operational rules. Consequently, large-scale hydrological models either exclude reservoir operations or use calibration-free algorithms for modeling reservoir dynamics. Nowadays, remotely-sensed information on reservoir storage anomalies is a potential resource for calibrating reservoir operation algorithms for a large number of globally distributed reservoirs. However, it is not yet clear what impact calibration against storage anomaly has on simulated reservoir outflow and absolute storage. In this study, we address this question using in-situ outflow and storage data from

15   100 reservoirs in the USA (ResOpsUS dataset) to calibrate three reservoir release algorithms, the well-established Hanasaki algorithm (CH) and two new storage-based algorithms, the Scaling algorithm (SA) and the Weighting algorithm (WA). These algorithms were implemented in the global hydrological model WaterGAP, with their parameters estimated individually for each reservoir and four alternative calibration targets: monthly time series of (1) storage anomaly, (2) estimated storage (calculated based on storage anomaly and GRanD reservoir capacity), (3) storage, and (4) outflow. The

20   first two variables can be obtained from freely available global datasets, while the last two variables are not publicly available for most reservoirs worldwide. We found that calibration against outflow does not lead to skillful storage simulations in most reservoirs and improves the outflow simulations only slightly more than calibration against the three storage-related calibration targets. Compared to the results of the non-calibrated Hanasaki Algorithm (DH), calibration against both storage anomaly and estimated storage improved the storage simulation and slightly improved the outflow

25   simulation. Calibration against storage anomaly resulted in 64 (39), 68 (45), and 66 (45) skillful storage simulations for CH, SA, and WA, respectively, during the calibration (validation) period, as compared to only 16 (15) for DH. Utilizing estimated storage instead of storage anomaly does not provide added benefit, primarily due to inconsistencies in observed maximum water storage and storage capacity data from GRanD. Findings show that the default parameters of the Hanasaki algorithm rarely matched the calibrated parameters, highlighting the importance of calibration. Using observed instead of

30   simulated inflow has a more significant effect on improving outflow simulation than calibration, whereas the opposite is true for storage simulation. Overall, the performance of the SA and WA algorithms is nearly equal, and both outperform the CH and DH algorithms. Moreover, incorporating downstream water demand into the reservoir algorithms does not necessarily improve modeling performance due to the high uncertainty in demand estimation. Therefore, to improve the modeling of reservoir storage and outflow in large-scale hydrological models, we recommend calibrating either the SA or

35   the WA reservoir algorithm individually for each reservoir against remote sensing-based storage anomaly, unless in-situ storage data are available, and to improve reservoir inflow simulation.

1 Introduction

Globally, more than 58,000 large dams (at least 15 meters in height), capable of impounding 8300 km³, have been constructed to meet various human needs such as irrigation, flood control, hydropower generation, domestic water supply, and recreation (Chao et al., 2008; Perera et al., 2021). These dams annually store about one-sixth of the streamflow in reservoirs (Hanasaki et al., 2006), significantly altering the global freshwater system by increasing evaporation and modifying downstream streamflow (Best, 2019; Tian et al., 2022). About 60% of the seasonal variability in Earth's surface water storage is attributed to human-managed reservoirs, i.e. artificial reservoirs and regulated lakes, as the water level of reservoirs varies on average four times as much as that of natural lakes (Cooley et al., 2021). Therefore, to accurately depict the hydrologic cycle, the inclusion of human-managed reservoirs in hydrological models is crucial. This inclusion is supposed to enhance model performance, particularly regarding evapotranspiration and streamflow. At present, six out of the 16 global hydrological models contributing to ISIMIP2 (The Inter-Sectoral Impact Model Intercomparison Project, www.isimip.org) simulate the dynamics of human-managed reservoirs (Telteu et al., 2021).

Whereas the outflow from a natural lake strongly depends on the water level of the lake and thus the water storage in the lake, humans control the outflow from a reservoir. Even though human decisions on the release of water from reservoirs do, to some degree, depend on reservoir water storage, they are also influenced by many other factors, such as downstream water demand, the demand for hydropower production, the need to protect downstream regions from flooding, ecosystem requirements, and legal constraints (Jager and Smith, 2008; Dong et al., 2023). Most reservoirs serve multiple purposes, making their simulation even more complex. However, since reservoir operation rules and observations of reservoir inflow, outflow and storage dynamics are rarely publicly accessible, large-scale hydrological models need to resort to calibration-free reservoir operation algorithms that only require information about the reservoir's storage capacity and surface water area. They are calibration-free algorithms in the sense that they do not require the calibration of reservoir-specific algorithm parameters based on observations of model output variables. These calibration-free algorithms can only very roughly simulate the decisions of reservoir operators and cannot account for the unique operation patterns of each reservoir (Masaki et al., 2018; Turner et al., 2021; Steyaert and Condon 2024).

All global hydrological models currently use calibration-free reservoir operation algorithms, which differ regarding their formulation and complexity (Telteu et al., 2021). Examples for calibration-free reservoir operation algorithms proposed for large-scale hydrological modeling are described in Dong et al. (2022), Zajac et al. (2017), Haddeland et al. (2006), and Hanasaki et al. (2006) (herein referred to as H06). Dong et al. (2022) and Zajac et al. (2017) employed different operation rules for four distinct levels of reservoir storage in their algorithms, whereas Haddeland et al. (2006) developed a prospective optimization algorithm based on the reservoir purpose. The H06 method is currently implemented in the global hydrological model H08 (Hanasaki et al., 2008) and, in a slightly modified form, in the global hydrological model WaterGAP, and also serves as the foundation for the Dam-Reservoir Operation model (DROP; Sadki et al., 2023). While studies (e.g., Döll et al., 2009; Vanderkelen et al., 2022) clearly demonstrate that implementing the H06 algorithm leads to improved streamflow simulations compared to not considering the reservoir as a surface water body at all, there is no consensus (please refer to Döll et al., 2009; Vanderkelen et al., 2022; Gutenson et al., 2020) on whether the H06 algorithm outperforms the natural lake outflow parameterization of Döll et al. (2003) (herein referred to as D03), which assumes artificial reservoirs behave similarly to natural lakes. It should be noted that simulating reservoir outflow and storage

dynamics depends not only on the reservoir operation algorithm but also on the quality of the simulated inflow, making it difficult to assess the adequacy of the algorithm without inflow observations (Vanderkelen et al., 2020).

Several studies have endeavored to fine-tune calibration-free algorithms by adjusting a single parameter for each reservoir, but the results have been unpromising. For example, Gutenson et al. (2020) found that adjusting only one parameter of H06 for 60 non-irrigation reservoirs across the US did not lead to better simulations compared to a calibrated D03. Shin et al. (2019) reported that a new algorithm based on H06, where one parameter was calibrated for 27 reservoirs, could not accurately capture the seasonality in reservoir storage and outflow. Consequently, some studies have devised calibration-required algorithms with multiple parameters for each reservoir. Turner et al. (2021) introduced the Inferred Storage Targets and Release Functions (ISTARF) approach, a reservoir operating policy with 19 parameters. This approach was applied to 1,930 reservoirs across the US and demonstrated robust improvements in both outflow and storage compared to the H06 model. Although the ISTARF approach is relatively parsimonious in terms of the number of parameters compared to other established calibration-required algorithms — such as those proposed by Yassin et al. (2019) and Turner et al. (2020), which feature 72 (six parameters for each month) and 208 parameters per reservoir (four parameters for each week), respectively — the integration of these approaches into large-scale models incurs substantial computational expenses. More importantly, this approach requires time series data of observed inflow, outflow, and reservoir storage, which can be difficult to obtain outside the US, rendering it infeasible for global-scale modeling. The same limitation applies to some machine learning approaches for simulating reservoir dynamics, such as the artificial neural network approach proposed by Ehsani et al. (2016) and the tree-based reservoir model of Chen et al. (2022).

Remotely sensed data on water levels and surface water area of reservoirs are increasingly available and are being used to derive time series of water storage anomalies or even absolute storage. With recent advancements in spaceborne data, such as the Surface Water and Ocean Topography (SWOT) mission, storage anomalies data can now be gathered even for small reservoirs, providing a valuable source for enhancing resource modeling within large-scale hydrological models (Biancamaria et al., 2016). Examples include HydroSat (Tourian et al., 2022), the Global Reservoir Storage (GRS) dataset (Li et al., 2023), and GloLakes (Hou et al., 2024). This newly available information could be used to calibrate reservoir operation algorithms individually for each reservoir, which is expected to lead to an improved simulation of reservoir dynamics. Remote sensing-derived reservoir storage anomalies were shown to fit reasonably well to in-situ observations, depending on the reservoir and satellite data product; storage anomalies rather than absolute water storage values should be considered for both the simulated and remote sensing data (Otta et al., 2023). In this regard, Hanazaki et al. (2022) developed a targeted storage-and-release algorithm for global flood modeling, where release is estimated for four storage zones based on the volume of each zone, flood discharge, and long-term average inflow. They estimated the volume of each storage zone using remote sensing data, while flood discharge was calculated using a probability distribution for 2,169 dams worldwide. The authors reported a 62% improvement in Nash-Sutcliffe Efficiency compared to the version of the CaMa-Flood global hydrodynamic model that did not include the reservoir module. Recently, supported by remote sensing data and a machine learning approach, Shen et al. (2024) developed a satellite-based target storage reservoir operation scheme (SBTS) with seven parameters. This scheme simulates the outflow and storage of flood control reservoirs across four distinct storage zones, using estimated flood storage capacity (FSC) data for 1,178 reservoirs derived through machine learning, trained on reported FSC data from 436 reservoirs. They found that their approach, when using observed inflow, improves reservoir parameterizations, allowing the SBTS to generally outperform the methods of Dong et al. (2022), Zajac

et al. (2017), and Hanazaki et al. (2022). However, they reported no improvement when simulated inflow was used. Dong et al. (2023) demonstrated that simultaneous calibrations against reconstructed release and reservoir storage data (using remotely sensed data, model simulations, and in-situ data) considerably improved the performance of reservoir operation algorithms for the Ertan and Jinping I reservoirs in China. However, for global-scale studies, release information is unavailable for most reservoirs. In such cases, calibrating against storage anomaly alone for parameter estimation may degrade outflow simulations due to potential trade-offs between calibrating against different variables (Döll et al., 2024, Hasan et al., 2025). The recently published dataset of observed dynamics of US reservoirs, 'ResOpsUS' (Steyaert et al., 2022), which provides time series of daily observed storage, elevation, inflows, and outflows for up to 679 reservoirs across the contiguous US, offers an opportunity to explore this trade-off.

The main objective of this study is to investigate how monthly time series of observed reservoir-related data can improve the simulation of reservoir outflow and storage in continental or global hydrological models. We focus on the suitability of observed storage anomaly for calibrating reservoir release algorithms, as these anomalies can be obtained globally through remote sensing-based observations. We compare their informational value to that of scarcer outflow and absolute storage observations, as well as the simulation results achieved with an uncalibrated reservoir algorithm. We utilized in-situ storage and outflow data from the ResOpsUS dataset for 100 reservoirs in the US to calibrate three reservoir operation algorithms. All algorithms were implemented in the global hydrological model WaterGAP 2.2e (Müller Schmied et al., 2024). The parameters of the algorithms were estimated using as alternative calibration targets, 1) storage anomaly, 2) estimated storage (calculated based on storage anomaly and GRanD reservoir capacity, detailed in section 2.3), 3) storage, and 4) reservoir outflow. Calibration involved optimizing parameters individually for each reservoir, algorithm and calibration target. To explore, in addition, the sensitivity of the model results to the quality of the inflow data, we calibrated the algorithms for a subset of 35 reservoirs with available inflow measurements, using observed inflow instead of the inflow simulated by WaterGAP. Finally, for a subset of 21 reservoirs, we determined the effect of incorporating, in the case of irrigation and water supply reservoirs, the downstream water demand in the reservoir algorithms.

2   Methods and Data

2.1 The global hydrological model WaterGAP

WaterGAP simulates the dynamics of water flows and storages on the continents as impacted by human water use and human-managed reservoirs (Müller Schmied et al., 2021). It computes sectoral water abstractions as well as net abstractions (abstraction minus return flows) from surface water bodies (reservoirs, lakes, and rivers) and from groundwater. The model has a spatial resolution of $0.5° \times 0.5°$ and a daily temporal resolution. However, the model output analysis is normally done at the monthly scale. The current version, 2.2e, has been calibrated in a basin-specific manner against the mean annual streamflow at 1,509 gauging stations worldwide (Müller Schmied et al., 2024). Taking into account the commissioning years, WaterGAP simulates the dynamics of reservoirs with a storage capacity of at least 0.5 km^3, referred to as 'global' reservoirs, using a slightly adapted version of the H06 algorithm (Döll et al., 2009). Smaller reservoirs (termed "local" reservoirs) are treated as natural lakes (Müller Schmied et al., 2021). A total of 1,255 global reservoirs, with a combined maximum capacity of 5,672 km^3, are integrated into WaterGAP 2.2e, sourced from the GRanD (Lehner et al., 2011) and GeoDAR (Wang et al., 2022) datasets; in addition, 88 regulated lakes are treated like global reservoirs (Müller Schmied et al., 2024). The water balance for a reservoir in WaterGAP is calculated as (Müller Schmied et al., 2021):

$$\frac{dS}{dt} = I + A \cdot \left(P - E_{pot} \right) - GWR - NAs - O \tag{1}$$

where S (m³) represents reservoir storage, I (m³/d) denotes inflow into the reservoir from upstream, A (m²) is the reservoir area, P (m/d) indicates precipitation, E_{pot} (m/d) stands for potential evaporation, GWR (m³/d) denotes groundwater recharge (only in arid/semiarid regions), NAs (m³/d) represents potential net abstraction from the reservoir, and O (m³/d) is the reservoir outflow including release and spill. The surface area A is computed daily as a fraction of the maximum area that depends on the current reservoir storage and its storage capacity. A is reduced by 15 % when S reaches 50% of the reservoir's capacity, and by 75% when S drops to 10% of the capacity (Müller Schmied et al., 2021). Abstraction from a reservoir is permitted only until the water storage level drops to 10% of its total capacity. The implementation of reservoir operation algorithms in WaterGAP is described below. For detailed information on WaterGAP, please refer to Müller Schmied et al. (2021, 2024).

2.2 Reservoir operation algorithms

2.2.1 Hanasaki algorithm as implemented in WaterGAP2.2e

The calibration-free H06 method, in its original formulation, estimates monthly reservoir outflow distinguishing irrigation and non-irrigation reservoirs. For non-irrigation reservoirs, this outflow is determined by factors such as the storage at the beginning of the operational year (determined by analyzing the seasonal flow dynamics), the mean annual inflow into the reservoir, and the reservoir storage capacity. The long-term target for reservoir releases is the mean annual inflow. If reservoir storage at the beginning of an operational year is above normal, releases are increased throughout the year, and if it is below normal, releases are decreased. Therefore, the total release in an operational year depends on the storage level at the start of that year. In the case of irrigation reservoirs, the demand also influences the release (Hanasaki et al., 2006). The H06 algorithm was implemented in WaterGAP on a daily time scale, and the mean annual inflow was adjusted by adding the difference between precipitation and evaporation over the reservoir. This modification aimed to provide a more accurate representation of the reservoir's water balance (Döll et al., 2009).

The first step in the H06 algorithm involves determining the release coefficient for the operational year 'y' (k_y) using the following equation:

$$k_y = \frac{S_{ini}}{a_1 \cdot C} \tag{2}$$

where S_{ini} (km³) represents the reservoir storage at the start of the operational year; C (km³) denotes the water storage capacity of the reservoir; and a_1 is a parameter of the H06 method, recommended to be set to 0.85 in its standard form. In the second step, the provisional release is determined. For non-irrigation reservoirs, the provisional release is calculated as follows:

$$R'_d = \overline{I'} \tag{3}$$

in which R'_d (m³/s) is the provisional release for the day 'd' and $\overline{I'}$ (m³/s) is the mean annual inflow into the reservoir plus the difference between precipitation and evaporation over the reservoir (for this study, the period 1980-2009). For irrigation reservoirs, the provisional release is computed as follows:

$$R'_d = \begin{cases} a_2 \cdot \overline{I'} \cdot \left[1 + \dfrac{k_{alc} \cdot NAs_d}{\overline{NAs}} \right] & if \ \overline{NAs} \ge a_2 \cdot \overline{I'} \\ \overline{I'} + k_{alc} \cdot NAs_d - \overline{NAs} & otherwise \end{cases} \qquad (4)$$

in which NAs_d (m³/s) represents the potential net abstraction from surface water bodies for downstream cells of the reservoir for day 'd'; $\overline{NAs}$ (m³/s) denotes the mean total annual potential net abstraction for downstream cells of the irrigation reservoir; k_{alc} is an allocation coefficient that distributes the abstraction to the upstream reservoirs based on the proportion of $\overline{I'}$ into each reservoir (it equals one if there is only one irrigation reservoir upstream of the demand cells); and a_2 is a parameter specifically for irrigation reservoirs that acts as a partitioner, leading to the use of different equations for reservoirs with a high demand-to-inflow ratio compared to those with a low demand-to-inflow ratio. With a default value of 0.5, this parameter sets the minimum provisional release at 50% of the mean annual inflow during non-crop months. During crop months, the fluctuations in provisional release for reservoirs with a high demand-to-inflow ratio ($\overline{NAs}$ exceeding 50% of mean annual inflow, first equation) correspond to fluctuations in daily net abstraction relative to $\overline{NAs}$. In contrast, reservoirs with a low demand-to-inflow ratio (second equation) align their provisional releases with the daily net abstraction (Hanasaki et al., 2006). The downstream potential net abstraction associated with each reservoir is calculated based on surface water demand for a maximum of five grid cells downstream in the absence of other reservoirs. Otherwise, it extends to the next reservoir. The potential net abstraction information is obtained from the WaterGAP dataset.

With the provisional release determined, the daily release is calculated using the following equation:

$$R_d = \begin{cases} k_y \cdot R'_d & if \ c \ge a_3 \\ \left(\dfrac{c}{a_3}\right)^2 \cdot k_y \cdot R'_d + \left\{ 1 - \left(\dfrac{c}{a_3}\right)^2 \right\} \cdot I_d & otherwise \end{cases} \qquad (5)$$

where c represents the ratio of C (km³) to $\overline{I'}$ (km³/yr); I_d (m³/s) is the daily inflow into the reservoir for the day 'd'; R_d (m³/s) is the daily release from the reservoir; and a_3 is a third parameter in the H06 approach, with default value of 0.5. This parameter is also a partitioner that results in the application of different equations for reservoirs with high capacity-to-inflow ratios ($c \ge a_3$) compared to those with low capacity-to-inflow ratios. This implies that for reservoirs with high capacity-to-inflow ratios (first equation), release is independent of daily inflow, while for reservoirs with low capacity-to-inflow ratios (second equation), daily inflow influences the release (Hanasaki et al., 2006). In this study, H06 with default values for a_1, a_2, and a_3 is referred to as the DH algorithm, while H06 with calibrated parameters is referred to as the CH algorithm.

2.2.2 New algorithms

In this study, we introduce and compare two new reservoir operation algorithms that 1) require the reservoir-specific calibration of their parameters; 2) different from H06, utilize daily reservoir water storage as a critical factor in computing daily releases; and (3) do not require water use information to estimate the releases of irrigation reservoirs. Both algorithms include three parameters that are related to different levels of storage: above 70% of the reservoir capacity (level 1), between 40% and 70% of the reservoir capacity (level 2), and below 40% of the reservoir capacity (level 3). This classification is based on the observation that the operation rule curve of reservoirs often varies at different storage levels, typically corresponding to different seasons (Dang et al., 2020). Unlike the H06 approach, which employs a single release coefficient for a full year of operation, both new algorithms consider a daily filling ratio, i.e. relative water storage ($Srel_d$), as defined by the following equation:

$$Srel_d = \frac{S_d}{C} \tag{6}$$

in which S_d (km³) is the reservoir storage on day 'd', and C (km³) indicates the water storage capacity of the reservoir. Both algorithms use $Srel_d$ for release estimation but apply different equations to calculate the release. The following sections describe the release estimation methods employed by these algorithms, i.e., Scaling algorithm (SA) and Weighting algorithm (WA).

215 **2.2.2.1 Scaling algorithm**

In the SA algorithm, the daily release at each specific storage level (Level 1, Level 2, or Level 3) is computed as a function of $Srel_d$, mean annual inflow ($\overline{I}$), daily inflow (I_d), the 30-day mean inflow ($\overline{I}_{30d}$), and a parameter associated with that level (Eq. 7). For this purpose, I_d is scaled using the ratio of $\overline{I}$ to $\overline{I}_{30d}$. This ratio represents the general effect of reservoirs in altering the temporal variation of streamflow by storing excess wa ter during high-flow months and releasing it during

220 low-flow months. The multiplication of $\overline{I}$ with $Srel_d$ mimics a prompt response to extreme events where storage can fill up within a few days. The release in the SA algorithm, when water storage is at level n, is calculated as follows:

$$R_d = p_n \cdot \left[Srel_{d-1} \cdot \overline{I} + \frac{\overline{I}}{\overline{I}_{30d}} \cdot I_d \right] \quad \text{for } n = 1, 2, 3 \tag{7}$$

in which $\overline{I}_{30d}$ (m³/s) represents the mean inflow into the reservoir during the last 30 days. The variable n indicates the storage level at time d-1, and p_n is the parameter assigned to storage level n (one parameter assigned to each storage level). Levels 1, 2, and 3 correspond to $Srel$ as follows: Level 1 for above 0.7, Level 2 for between 0.4 and 0.7, and Level 3

225 for below 0.4. (see Fig. 1). The parameters value need to be determined through the calibration process. These parameters enable us to adjust the mean release, while temporal variability is estimated inside the square brackets.

2.2.2.2 Weighting algorithm

The WA are the same as SA method in most part of the release calculation, however, in contrast to the SA method, WA does not consider I_d to compute the release and solely relies on $Srel_d$ for weighting $\overline{I}$ and $\overline{I}_{30d}$. Therefore, the contribution

230 of long-term inflow is higher at higher storage levels, while its contribution decreases with lower storage levels. Conversely, the contribution of inflow from the last 30 days increases as storage decreases. A maximum of 30% of $\overline{I}_{30d}$ contributes to release estimation at higher storage levels ($Srel \geq 0.7$), while it reaches 100% when the reservoir is empty, which is identical to run-of-the-river flow. In the WA algorithm when water storage is at level n, the release is estimated as follows:

$$R_d = q_n \cdot [Srel_{d-1} \cdot \overline{I} + (1 - Srel_{d-1}) \cdot \overline{I}_{30d}] \quad \text{for } n = 1, 2, 3 \tag{8}$$

where q_n is the parameter assigned to storage level n that needs to be determined (see Fig. 1). We opted for $\overline{I}_{30d}$ over I_d

235 assuming that release decisions may rather be based on the past inflow over a longer period and not on the inflow on just the previous day.

Contrary to the H06 approach, where the release is independent of inflow in reservoirs with large storage capacity relative to the annual inflow (meaning constant release throughout the year, see Eq. 5), both new algorithms consider the impact of inflow on release in all reservoirs. This impact varies with different seasons and storage levels, leading to

240 variability in release throughout the year, which is more realistic (see Eq. 7 and Eq. 8). It should be noted that the new algorithms do not distinguish between irrigation and non-irrigation reservoirs; therefore, no water use data is required for

their application, making their implementation easier than the H06 algorithm. This is because the estimation of downstream water demand at a large scale is generally very uncertain, and reservoirs are usually designed for multiple purposes.

In each of the three algorithms, if S_d falls below 10 percent of the storage capacity (C), the calculated R_d is adjusted to
245    $0.1 \cdot R_d$ if the available water is sufficient; otherwise, the entire S_d will be released. Finally, the reservoir outflow is calculated as follows:

$$O_d = R_d + SP_d \tag{9}$$

where O_d (m³/s) and SP_d (m³/s) are the reservoir outflow and the spill from the reservoir during day 'd', respectively. SP_d is calculated as the difference between S_d and C, where S_d exceeds C; otherwise, it is zero.

Hanasaki algorithm (Version implemented in WaterGAP)

Calculating the release coefficient (k_y) [Eq. 2]:
$$k_y = f(S_{ini}, C, a_1)$$

Calculating the provisional release (R') [Eq. 3 or Eq. 4]:
$$R'_d = \begin{cases} f(I') & \text{for non−irrigation reservoirs} \\ f(I', NAs_d, \overline{NAs}, a_2) & \text{for irrigation reservoirs} \end{cases}$$

Calculating the release (R_d) [Eq. 5]:
$$R_d = f(R'_d, k_y, C, I', I_d, a_3)$$

Daily Variables
I_d: Inflow to the reservoir
P_d: Precipitation over the reservoir
E_d: Evaporation from the reservoir
SP_d: Spill from the reservoir
R_d: Release from the reservoir
R'_d: Provisional release
NAs_d: Potential net abstraction from surface water
$Srel_d$: Relative water storage
S_d: Reservoir storage

Yearly variables
S_{ini}: Reservoir storage at the start of the operational year
k_y: Release coefficient

Scaling and Weighting algorithms

Calculating the relative water storage ($Srel$) [Eq. 6]:
$$Srel_d = f(S_d, C)$$

Calculating the release using the scaling algorithm (R_d) [Eq. 7]:
$$R_d = \begin{cases} f(Srel_{d-1}, \overline{I}, \overline{I}_{30}, I_d, p_1) & Srel_{d-1} \geq 0.7 \ (n=1) \\ f(Srel_{d-1}, \overline{I}, \overline{I}_{30}, I_d, p_2) & 0.4 \leq Srel_{d-1} < 0.7 \ (n=2) \\ f(Srel_{d-1}, \overline{I}, \overline{I}_{30}, I_d, p_3) & Srel_{d-1} < 0.4 \ (n=3) \end{cases}$$

Calculating the release using the weighting algorithm (R_d) [Eq. 8]:
$$R_d = \begin{cases} f(Srel_{d-1}, \overline{I}, \overline{I}_{30}, q_1) & Srel_{d-1} \geq 0.7 \ (n=1) \\ f(Srel_{d-1}, \overline{I}, \overline{I}_{30}, q_2) & 0.4 \leq Srel_{d-1} < 0.7 \ (n=2) \\ f(Srel_{d-1}, \overline{I}, \overline{I}_{30}, q_3) & Srel_{d-1} < 0.4 \ (n=3) \end{cases}$$

Time-averaged variables
I': Mean annual inflow into the reservoir, plus the difference between precipitation and evaporation over the reservoir
$\overline{NAs}$: Mean total annual potential net abstraction
$\overline{I}$: Mean annual inflow
$\overline{I}_{30}$: 30-day mean inflow

Constant parameters
C: Water storage capacity of the reservoir
a_1, a_2, a_3: Calibration parameters of the Hansaki algorithm
p_1, p_2, p_3: Calibration parameters of the Scaling algorithm
q_1, q_2, q_3: Calibration parameters of the Weighting algorithm

250   **Figure 1.** Overview of the process to calculate reservoir release using the Hanasaki (H06), Scaling (SA) and Weighting (WA) algorithms, indicating the required inputs as well as the equation numbers where the complete equations can be found in the text. The left panel details the H06 algorithm implemented in WaterGAP, with steps for calculating the release coefficient, provisional release, and release. The H06 algorithm requires reservoir capacity, storage values at the start of the operational year, daily inflow, precipitation, evaporation data, and daily potential net abstraction data for irrigation
255   reservoirs. The right panel presents SA and WA, indicating the calculation of relative water storage and the release computation as a function of three reservoir water storage levels (n = 1, 2, or 3). SA and WA release are calculated based on reservoir capacity, daily storage, precipitation, evaporation, and inflow. The time-averaged variables are derived from daily data. For the H06 algorithm DH, the default values for a_1, a_2, and a_3 are 0.85, 0.5, and 0.5, respectively.

2.3 Data

260   The ResOpsUS dataset (Steyaert et al., 2022), which served for calibrating and evaluating the three algorithms in this study encompasses daily in-situ records of inflow, storage, outflow, elevation, and evaporation for up to 679 US reservoirs. The available data spans from 1930 to 2020, determined by each dam's commissioning year and data availability. In this study, data on reservoir inflow (daily), outflow (monthly), and storage (monthly) from 1980 to 2019 were considered, divided into two distinct periods: a calibration phase spanning from 1980 to 2009, and a validation phase covering the years 2010 to

265   2019. Monthly data were computed from daily records, excluding months with more than one week of missing values. Subsequently, we applied filters to the dataset, considering only reservoirs with a minimum data length of five years, a minimum reservoir capacity of 0.5 km^3, as well as ensuring there is only one reservoir per 0.5°×0.5° grid cell and no negative values. This resulted in 100 reservoirs, with 35 having data for storage, inflow and outflow and 65 having data for storage and outflow only. The minimum number of monthly data values for the 65 (35) reservoirs was 111 (252) for the

270   calibration period and 65 (59) for the validation period. The reservoirs' storage capacities (*C*) range from 0.5 km^3 to 36.7 km^3 based on the GRanD dataset (Lehner et al., 2011). Out of the total 100 reservoirs, nine are irrigation reservoirs. Detailed information on each reservoir is provided in Table S1.

      Using in-situ storage data, we derived two additional storage-related variables: the time series of storage anomaly and estimated storage. These variables can also be estimated using remote sensing data. Storage anomaly time series for each

275   reservoir is calculated by subtracting the mean storage during the calibration period from the in-situ storage data for each reservoir. However, the storage anomaly lacks information about the bias term and calibrating against it can result in a simulated storage time series that significantly deviates from the observed water storage. Having actual absolute storage is advantageous, as reservoirs are the only surface water bodies for which we can model absolute storage within the WaterGAP. To provide an alternative, we calculated the "estimated storage time series"; this term refers to storage values

280   that are not observed directly but are estimated using storage anomaly and the reservoir capacity *C*. First, we determined the storage changes time series by subtracting the initial month's storage anomaly value from the monthly storage anomaly values. Assuming the reservoir reaches maximum capacity at least once between 1980 and 2009, we calculated the maximum monthly storage change, termed *Dif$_{max}$*. We then subtracted *Dif$_{max}$* from the GRanD reservoir storage capacity to estimate the initial water storage for the first month. The estimated storage time series is then obtained by adding the storage

285   changes to this estimated initial water storage. Since the data are monthly, and daily maximum storage is generally higher, we applied a 1.2 scaling factor to *Dif$_{max}$*. This adjustment means that *Dif$_{max}$* used in our calculations is 20% higher than the initially calculated value. This 20% increase is derived from the mean difference between the maximum daily storage and the monthly storage observed in 100 studied reservoirs (see Table S1). The calculation of estimated storage can be performed using either absolute storage or storage anomaly, as the time series of storage changes would remain the same

290   in both cases. An example using GRanD ID 597 (Glen Canyon Dam, Lake Powell) clarifies the calculation of storage anomaly and estimated storage. The mean observed storage value between 1980 and 2009 for Glen Canyon Dam is 22.45 km³. To obtain the storage anomaly time series for this reservoir, the value of 22.45 km³ is subtracted from all storage data for the reservoir over the entire period (1980–2019). For calculating estimated storage, the *Dif$_{max}$* is 6.6 km³, which occurred in July 1983 (see Fig. S1). This is calculated as the storage anomaly value in July 1983 minus the initial storage anomaly

295   value in January 1980. The initial storage is estimated as 25.1 km³ (the reservoir capacity reported by GRanD) minus 7.9 km³ (6.6 km³×1.2). This gives an initial storage value of approximately 17.2 km³. Storage changes are then added to the

estimated initial storage to obtain the time series of estimated storage (Fig. S1c), e.g., the estimated storage for July 1983 is 23.8 km³, which is the sum of 17.2 km³ and 6.6 km³.

2.4 Model variants and calibration approach

300   The three reservoir operation algorithms were implemented in WaterGAP. For each algorithm, the algorithm-specific parameters (a_1, a_2, and a_3 for the CH, p_1, p_2, and p_3 for the SA and q_1, q_2, and q_3 for the WA) were estimated by optimizing the Kling–Gupta Efficiency (KGE) (Kling et al., 2012), including the trend term (see Eq. 10). This optimization was performed through a single-objective calibration against the monthly time series of four variables: outflow, storage, storage anomaly, and estimated storage (see Section 2.3). The parameters of each algorithm were calibrated using a grid search

305   approach. Reservoir outflow and storage time series were simulated for all parameter sets listed in Table S2, and the parameter set corresponding to the highest KGE was selected. The parameter estimation using storage anomaly and estimated storage serves as the main experiment, as the primary emphasis of this study is on exploring the added value of incorporating storage anomaly (which facilitate the calibration of reservoir algorithms using remote sensing data in regions where in-situ storage time series are unavailable) into the calibration of reservoir operation algorithms.

310   As in previous studies by Dong et al. (2023), Turner et al. (2021), and Shin et al. (2019), the uncalibrated H06 (DH) is used as a benchmark. For comparison purposes, in all calibration experiments based on WaterGAP inflow, the inflow into reservoirs simulated by the DH algorithm was used to ensure that the same inflow data were applied across all algorithms. To achieve this, WaterGAP was first run with the DH algorithm to save the reservoir inflow data. These inflow data were then read from the saved files and used as the inflow source to model each reservoir independently. As a result,

315   inflow into all reservoirs, regardless of their position, was based on the DH algorithm when applying the CH, SA, and WA algorithms, meaning that the operations of upstream reservoirs did not affect downstream reservoirs. The calibration runs were initialized by running WaterGAP five times for the year 1979 to allow water storages to reach a relatively stable equilibrium state.

    In addition to the inflow simulated by WaterGAP, we also assessed the algorithms based on observed inflow where

320   available. This was done to check the performance of reservoir operation algorithms in the presence of high-quality inflow data, as the performance of the algorithms may be heavily impacted by poor inflow data (Vanderkelen et al., 2022). Moreover, we assessed the impact of distinguishing irrigation and supply reservoirs from other reservoirs. The distinction for irrigation reservoirs is the default approach for the H06 algorithm; however, here we also applied this distinction for supply reservoirs, as also their outflow depends on downstream demand. To this end, we modeled 21 reservoirs (nine

325   irrigation and 12 supply reservoirs) in two different ways for all algorithms: one including downstream demand and the other without considering it. The purpose of this comparison is to evaluate whether including downstream demand, despite the high uncertainty in water demand estimation for the reservoirs, enhances the outflow and storage simulation, or whether it may not add value and instead introduce unnecessary complexity. In the case of the SA and WA approaches for considering downstream demand, similar to the DH algorithm, Eq. 4 was used with the default value for the parameter a_2.

330   However, instead of using $\overline{I'}$, $\overline{I}$ was applied in Eq. 4. The resulting R'_d from Eq. 4 then replaced $\overline{I}$ in Eqs. 7 and 8 for estimating R_d.

    Table 1 shows a summary of the different calibration variants. In Table 1, each calibration variant is characterized by a combination of a reservoir operation algorithm, a calibration variable, an inflow source, and whether or not downstream

demand is considered. For example, calibrating the CH algorithm against outflow using inflow simulated by WaterGAP
while considering downstream water demand represents one calibration variant. Thus, each reservoir operation algorithm
comprises 12 calibration variants (eight utilizing WaterGAP inflow and four using observed inflow), leading to a total of
36 calibration variants.

Table 1. Components of the different calibration variants, comprising 36 variants in total, with 12 variants for each
algorithm. Each algorithm includes four variants using WaterGAP inflow with downstream demand considerations
(calibrated against outflow, storage, storage anomaly, and estimated storage), four variants using WaterGAP inflow without
downstream demand, and four variants using observed inflow. Each calibration variant is defined by the combination of a
reservoir operation algorithm, calibration variable, inflow source, and the consideration or non-consideration of downstream
demand. For CH, the default approach incorporates the downstream demand of irrigation reservoirs, while the opposite is
true for SA and WA. Additionally, considering the downstream demand for supply reservoirs is not the default approach
for any of the reservoir operation algorithms. For calibration variants that utilize observed inflow, only the default approach
of each algorithm is considered.

Operation algorithm	Calibration variable	Inflow source	Downstream demand considered?
CH	Outflow	WaterGAP	Yes[1]
	Storage		No
	Storage anomaly	Observation	Yes[2]
	Estimated storage		
SA	Outflow	WaterGAP	Yes[1]
WA	Storage		No
	Storage anomaly	Observation	No
	Estimated storage		

[1] Water demand is considered for irrigation and supply reservoirs, i.e., 21 out of 100 studied reservoirs.
[2] Water demand is considered for irrigation reservoirs, i.e., two out of 35 studied reservoirs with observed inflow.

2.5 Performance evaluation metrics

The performance of the reservoir operation algorithms was evaluated using KGE and the normalized root mean square error
(nRMSE). KGE is widely used for model calibration and evaluation, as it simultaneously considers multiple important
aspects of model performance, providing a comprehensive assessment (Beck et al., 2019; Lamontagne et al., 2020). The
use of nRMSE offers additional insights by focusing on the magnitude of errors. Following Hosseini-Moghari et al. (2020),
we incorporated the trend component into the conventional KGE equation as follows:

$$KGE = 1 - \sqrt{(R_{KGE} - 1)^2 + (B_{KGE} - 1)^2 + (V_{KGE} - 1)^2 + (T_{KGE} - 1)^2} \tag{10}$$

$$R_{KGE} = \frac{cov(sim, obs)}{\sigma_{sim} \cdot \sigma_{obs}} \tag{11}$$

$$B_{KGE} = \frac{\overline{sim}}{\overline{obs}} \tag{12}$$

$$V_{KGE} = \frac{\sigma_{sim}/\overline{sim}}{\sigma_{obs}/\overline{obs}} \tag{13}$$

$$T_{KGE} = \frac{T_{sim}}{T_{obs}} \tag{14}$$

where R_{KGE} represents the correlation coefficient between observed (*obs*) and simulated (*sim*) time series; B_{KGE} denotes the
bias of the mean simulated ($\overline{sim}$) compared to the mean of observed ($\overline{obs}$), V_{KGE} is the variability component that denotes
the ratio of the standard deviation of the simulated (σ_{sim}) to the standard deviation of the observed (σ_{obs}) time series,
divided by their mean, and T_{KGE} represents the ratio of the linear trend of the simulated time series (T_{sim}) to the observed

360 one (T_{obs}). In the case of calibrating against storage anomaly, we did not divide σ by the mean, as the mean for storage anomaly is zero. Similarly, the B_{KGE} component was not considered in calculating KGE related to storage anomaly. The optimal value for the KGE and its four components is 1. The KGE range is $(-\infty, 1]$, while R_{KGE} ranges from -1 to 1; B_{KGE}, V_{KGE} and T_{KGE} can vary between $-\infty$ and $+\infty$. Following Knoben et al. (2019), a KGE value above -0.73 indicates that the model performs better than the mean of observations if the trend component is included in the KGE.

365 The normalized root mean square error (nRMSE) is calculated as:

$$nRMSE = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}(obs_t - sim_t)^2}}{\sigma_{obs}} \tag{15}$$

The perfect value for nRMSE is zero. Normalizing the RMSE with the standard deviation of observations brings this metric closer to the Nash-Sutcliffe Efficiency (NSE), but different from the NSE, the nRMSE cannot become negative (Turner et al., 2021).

3 Results

3.1 Performance of calibration variants in the case of simulated inflow into reservoirs

We found that calibrating against observed water storage, water storage anomaly, or estimated water storage (derived from storage anomaly and GRanD storage capacity) improves the very poor simulation of storage by the calibration-free algorithm (DH) for both the calibration and validation periods in the case of all three algorithms (Table 2). In the case of

375 DH, storage simulation is skillful, i.e. with a KGE$_{storage}$ > -0.73, for only 16% of the 100 reservoirs during the calibration period, and for 15% during the validation period. Calibration of the H06 reservoir operation algorithm (CH) achieves skillful storage simulations for 64% (39%) of the reservoirs when calibrated against storage anomaly and for 69% (32%) of the reservoirs when calibrated against estimated storage during the calibration (validation) period. Both SA and WA outperform CH in storage simulation when calibrated against storage-related variables for both the calibration and validation period

380 (Table 2 and Fig. 2). However, the fit of simulated to observed storage remains poor during the validation period, in particular after calibration against storage anomaly and estimated storage (Table 2 and Fig. 2).

Table 2. The number of reservoirs out of 100 in which KGE values are greater than the benchmark thresholds of -0.73 during the calibration (validation) phase. All algorithms were calibrated against outflow, storage, storage anomaly, as well as estimated storage using KGE as the objective function. The inflow data is sourced from the WaterGAP model.

Calibrated variable	Algorithm	KGE > -0.73	
		Outflow	Storage
—	DH	63 (56)	16 (15)
Outflow	CH	78 (68)	22 (30)
	SA	86 (71)	14 (24)
	WA	86 (69)	20 (30)
Storage	CH	68 (69)	91 (46)
	SA	66 (67)	98 (68)
	WA	67 (66)	100 (55)
Storage anomaly	CH	67 (69)	64 (39)
	SA	67 (69)	68 (45)
	WA	71 (70)	66 (45)
Estimated storage	CH	70 (69)	69 (32)
	SA	65 (68)	69 (46)
	WA	67 (70)	74 (41)

Calibration against storage-related variables only slightly improves the mostly poor simulations of reservoir outflow during the calibration period and shows a bit more improvement in the validation period (Table 2 and Fig. 2). Skillful
385 outflow simulations were achieved for 86% of the reservoirs when either SA or WA were calibrated against outflow, compared to 78% for CH and 63% for DH during the calibration phase. However, skillful storage simulations were observed in only 14% (24%) and 20% (30%) of the reservoirs for SA and WA, respectively, compared to 22% (30%) for CH and 16% (15%) for DH in the calibration (validation) phase (Table 2). The performances of outflow simulations with CH, SA and WA are very similar in both the calibration and validation periods except in the case of calibration against observed
390 outflow for the calibration period. In this case, SA and WA achieved positive $KGE_{outflow}$, with medians of 0.15 for SA and of 0.13 for WA. Calibrating against outflow improves the correlation, variability and trend of the simulated outflow compared to DH for all three algorithms, while the bias is not affected much (Figs. S2-S5). On average, outflow trends are underestimated. Calibrating against outflow worsens both the correlation and variability of storage simulations across all three algorithms during the calibration phase, though it notably improves the bias component (Figs. S2-S4). Model
395 performance regarding storage is not affected in a relevant manner by calibration against outflow and remains very poor. When algorithms are calibrated against outflow, the mean observed storage generally remains a better estimator than the simulated storage.
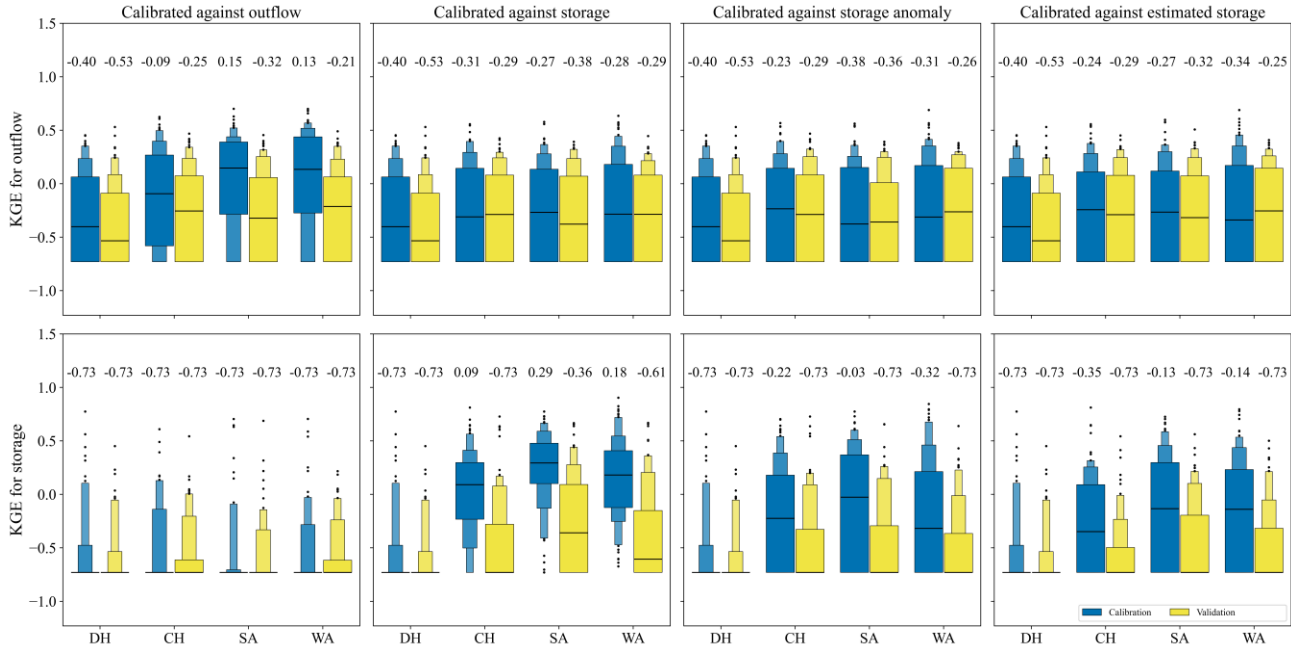
Figure 2. Letter-value plots of KGE for outflow and storage of 100 studied reservoirs for DH, CH, SA, and WA algorithms
400 for the calibration period (1980-2009, in blue) and validation period (2010-2019, in yellow). All algorithms are calibrated against outflow (first column), storage (second column), storage anomaly (third column), as well as estimated storage (fourth column) using KGE as the objective function. The values at the top of the panels are the median KGE (indicated by the horizontal line). KGE values below the benchmark threshold of -0.73 are set to -0.73. The widest box contains 50% of the 100 data points, the second widest 25% of the data (12.5% in the upper box and 12.5% in the lower box), the third
405 widest 12.5%, and so on. The inflow data is sourced from the WaterGAP model.

Calibrating against storage (second column of Fig. 2) leads to the highest $KGE_{storage}$ values; with a median $KGE_{storage}$ of 0.29, SA outperforms CH and WA, while the $KGE_{outflow}$ and its component values for the three algorithms are similar (Figs. S2-S5). Calibrating against storage anomaly (third column in Fig. 2) or estimated storage (fourth column in Fig. 2) improves both storage and outflow simulations as compared to DH but the fit to observed storage is considerably worse than in the case of calibration against storage. While the median $KGE_{storage}$ in the case of calibration against storage anomaly is slightly better than when calibrated against estimated storage, the widest box of the letter-value plot related to calibration against estimated storage, which contains 50% of the data, is above the one for calibration against storage anomaly. The improvement of storage simulation is mainly through bias adjustment (Fig. S3). The DH algorithm has a median B_{KGE} of 1.90 for storage during the calibration period. This value decreases to 0.92 (1.04, 0.99), 0.71 (0.91, 1.18), and 1.25 (1.44, 1.32) for calibration against storage, storage anomaly, and estimated storage of the CH (SA, WA) algorithm, respectively. The correlation is improved in the case of SA and WA but only in the calibration period (Fig. S2). The variability is improved for calibration against storage anomaly, while calibration against estimated storage leads to an underestimation of storage variability (Fig. S4). By calibration against storage, storage anomaly and estimated storage, the trend component of $KGE_{storage}$ strongly improves as compared to DH for the calibration period but the trend is on average still underestimated (Fig. S5). Assessing the $KGE_{storage\_anomaly}$ when calibrating with different variables shows less degradation during the validation phase (Fig. S6). For example, the number of skillful simulations for storage reached 17 (18), 93 (44), 98 (59), and 99 (55) when calibrating using storage anomaly with DH, CH, SA, and WA, respectively (see Table 2 for comparison).

The fit to observed storage-related variables is much less improved as compared to DH for the validation period than for the calibration period (Table 2 and Fig. 2). Comparing calibration against storage anomaly and estimated storage, which are the available options when using only remote sensing data, reveals that SA and WA are preferable to CH and DH, even though the differences from CH are small during the validation period. Differences between the $KGE_{storage}$ values of SA and WA are small for all calibration variables for both calibration and validation periods.

Examining the empirical cumulative distribution functions (eCDFs) for nRMSE reveals that the eCDFs for outflow are much closer across different algorithms compared to those for storage (see Fig. 3). This suggests that calibration has a more significant impact on storage than on outflow. Calibration against any storage-related variable generally enhances outflow performance at lower $nRMSE_{outflow}$ levels (in approximately 60% of the reservoirs), while at higher $nRMSE_{outflow}$ ranges, a slight degradation is observed in about 35% of reservoirs (with probabilities ranging from less than 0.60 to 0.95, mainly concentrated between 0.8 and 0.9). When calibrating against outflow, there is generally improvement in $nRMSE_{storage}$ for CH and WA algorithms, while no clear improvement is seen for SA. Moreover, the error in outflow simulation is reduced in over 40% of reservoirs where the $nRMSE_{outflow}$ was already lower compared to others. For $nRMSE_{outflow}$ greater than 0.98, there is almost no discernible improvement observed when calibrating algorithms against outflow, as indicated by the eCDFs. The calibration against storage anomaly, which is the main calibration variant, especially in the validation phase, reveals that SA slightly performs better than WA. SA shows $nRMSE_{storage}$ lower and nearly similar $nRMSE_{outflow}$ compared to WA. Disregarding the magnitude of error, the eCDF for validation has a shape similar to that of the calibration period, suggesting that the error distribution for the algorithm is consistent across both periods.
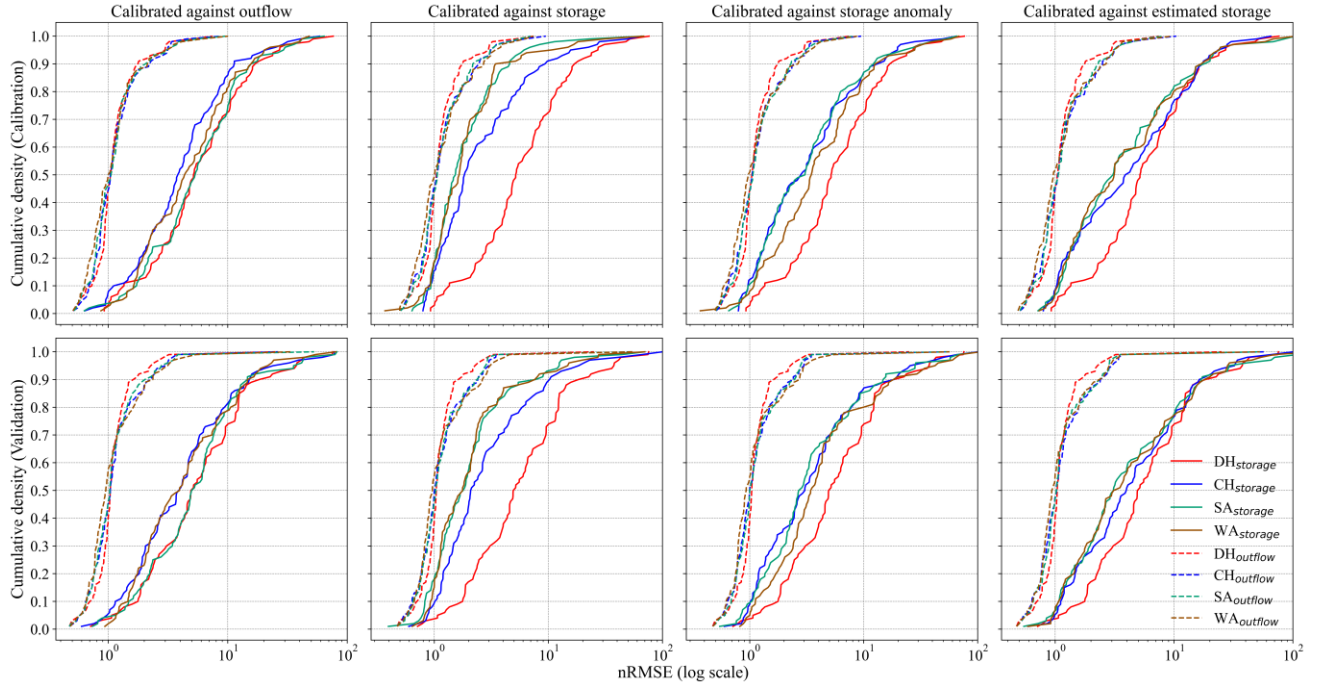
Figure 3. Empirical cumulative distribution functions of nRMSE for storage and outflow of 100 studied reservoirs are based on DH, CH, SA, and WA algorithms for the calibration period (1980-2009) and validation period (2010-2019). All algorithms are calibrated against outflow (first column), storage (second column), storage anomaly (third column), and estimated storage (fourth column) using KGE as the objective function. The x-axis has a logarithmic scale. If nRMSE is larger than 1, the mean error is larger than the standard deviation of the observational values. The inflow data is sourced from the WaterGAP model.

3.2 Illustrative calibration results for three reservoirs

As an example, we plotted the time series of storage and outflow for the Glen Canyon Dam (Lake Powell) in Fig. 4 and Fig. S7, respectively. This dam is one of the largest in our study, with several dams located upstream. The WaterGAP dataset includes four upstream reservoirs as global reservoirs, with storage capacities ranging from 0.57 to 4.3 km³. Calibrating the H06 algorithm against outflow did not lead to better results compared to the DH model (Fig. 4a, Fig. S7a). However, some improvement was observed in the outflow simulation for SA and WA during the calibration period, though this led to worse outflow simulation during the validation phase (Fig. S7). Despite this, with a KGE > -0.73, all outflow simulations demonstrated skillful performance. Calibration against outflow did not degrade storage simulation compared to the DH, except for SA, particularly during the validation phase, where the variability of the simulated time series was more than three times higher than the observed one (Table S3). During the calibration phase, storage levels are mainly above 40% (10 km³) of the capacity, with a sharp decline between 40% and 70%, and smaller changes when the reservoir is filled above 70% (17.5 km³). This pattern leads to storage levels below 40% not being adequately considered in the parameter selection process. As a result, when storage drops below 10 km³ during the validation phase, the outcomes are not promising (Fig. 4). The large difference between the capacity reported by GRanD (25 km³) and the maximum observed daily storage (31.7 km³) results in poorer performance in storage simulation for all calibrated algorithms based on estimated storage compared to storage anomaly (see Fig. S1). This ~20% difference between the reported capacity and maximum observed storage introduces a 20% bias, which directly impacts the bias and variability components of $KGE_{storage}$ (Table S3). However, there is almost no bias in the outflow, thanks to the data from the Lees Ferry station, located just downstream

of the dam, which is used in the bias adjustment of streamflow simulations in WaterGAP through a simple calibration approach.
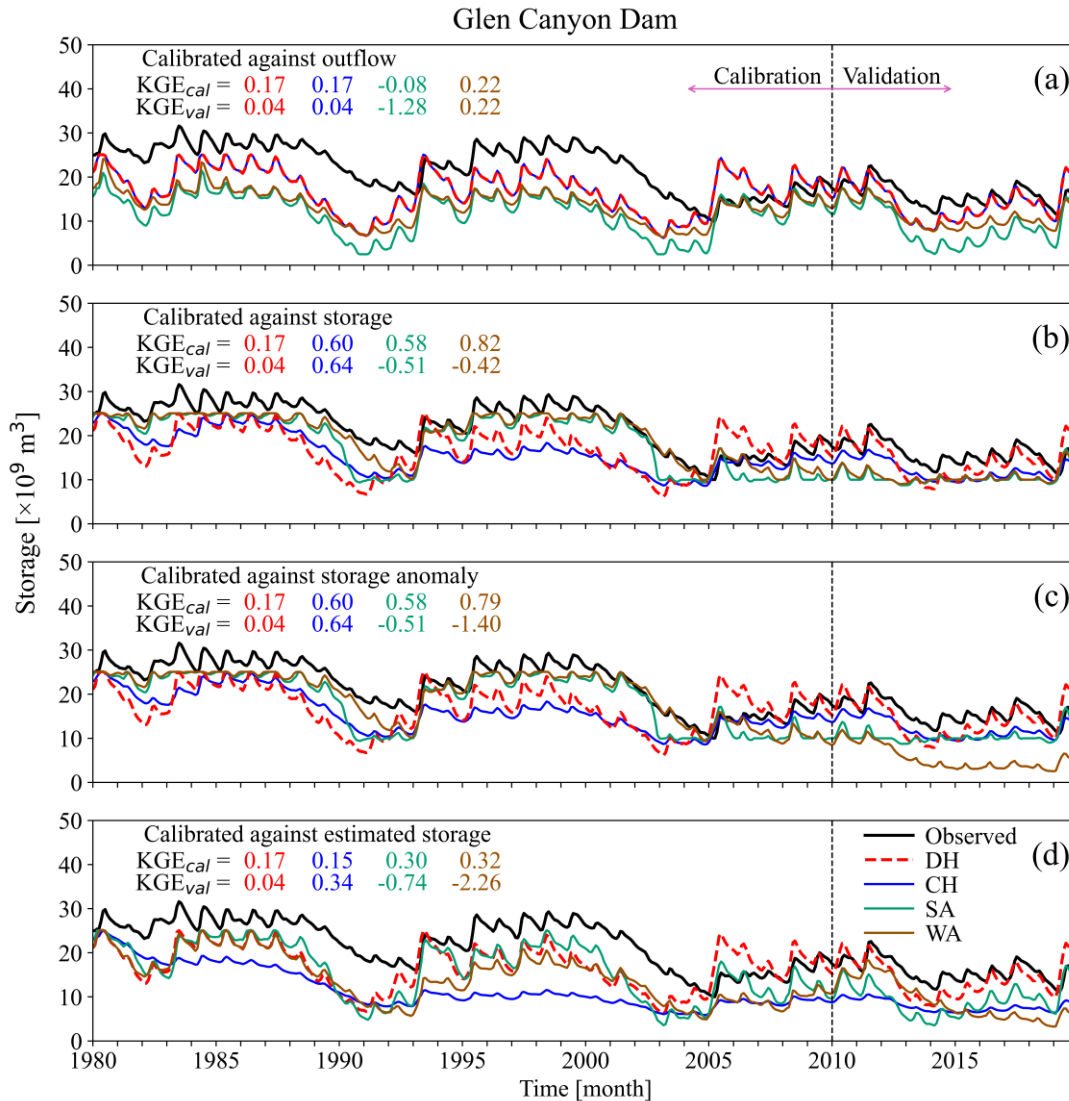
Figure 4. Monthly time series of observed and simulated storage values from DH, CH, SA, and WA algorithms for Glen Canyon dam, GRanD ID 597, calibrated against (a) outflow, (b) storage, (c) storage anomaly, and (d) estimated storage using KGE as the objective function. The dashed black lines distinguish between the calibration and validation periods. The inflow data is sourced from the WaterGAP model. The time series for outflow is plotted in Figure S7.

475   Very poor storage simulation with a much higher seasonal magnitude compared to observed storage is seen for the Yellowtail Dam (GRanD ID = 355), an irrigation reservoir with different calculations in the DH and CH algorithms compared to the SA and WA algorithms, and for the Harry S. Truman Dam (GRanD ID = 989), which is a hydropower reservoir (Fig. 5). Calibrating against storage anomaly can lead to time series of storage with considerable bias (Fig. 5c). This issue can also occur when calibrating against estimated storage if there is an offset between the estimated storage and

480   the in-situ observation (Fig. 4d). The time series related to the Yellowtail Reservoir reveals that SA and WA, which do not consider the irrigation purpose of this reservoir, can simulate reservoir storage better than DH and CH which explicitly take into account the downstream water demand (Fig. 5a). However, the opposite is true for outflow simulation, where the uncalibrated DH performs best (Fig. S8a).

From these examples, we found that calibrating solely against storage-related variables does not necessarily lead to
poorer outflow simulations (Fig. S8). However, other factors, such as inaccuracies in reservoir capacity data (e.g., for the
Glen Canyon Dam) and discrepancies between actual available water and the reported static storage value in the GRanD
dataset — which may include dead storage (see Table S1 for Yellowtail and Harry S. Truman dams) — are important
considerations when evaluating the performance of the reservoir operation algorithm. In such cases, comparing storage
anomaly may offer a more reasonable assessment than comparing absolute storage. This error in storage simulation may
also affect outflow simulations, where inaccuracies in input data are the primary factor leading to inaccurate storage levels
being maintained during the validation phase (Fig. 4c).
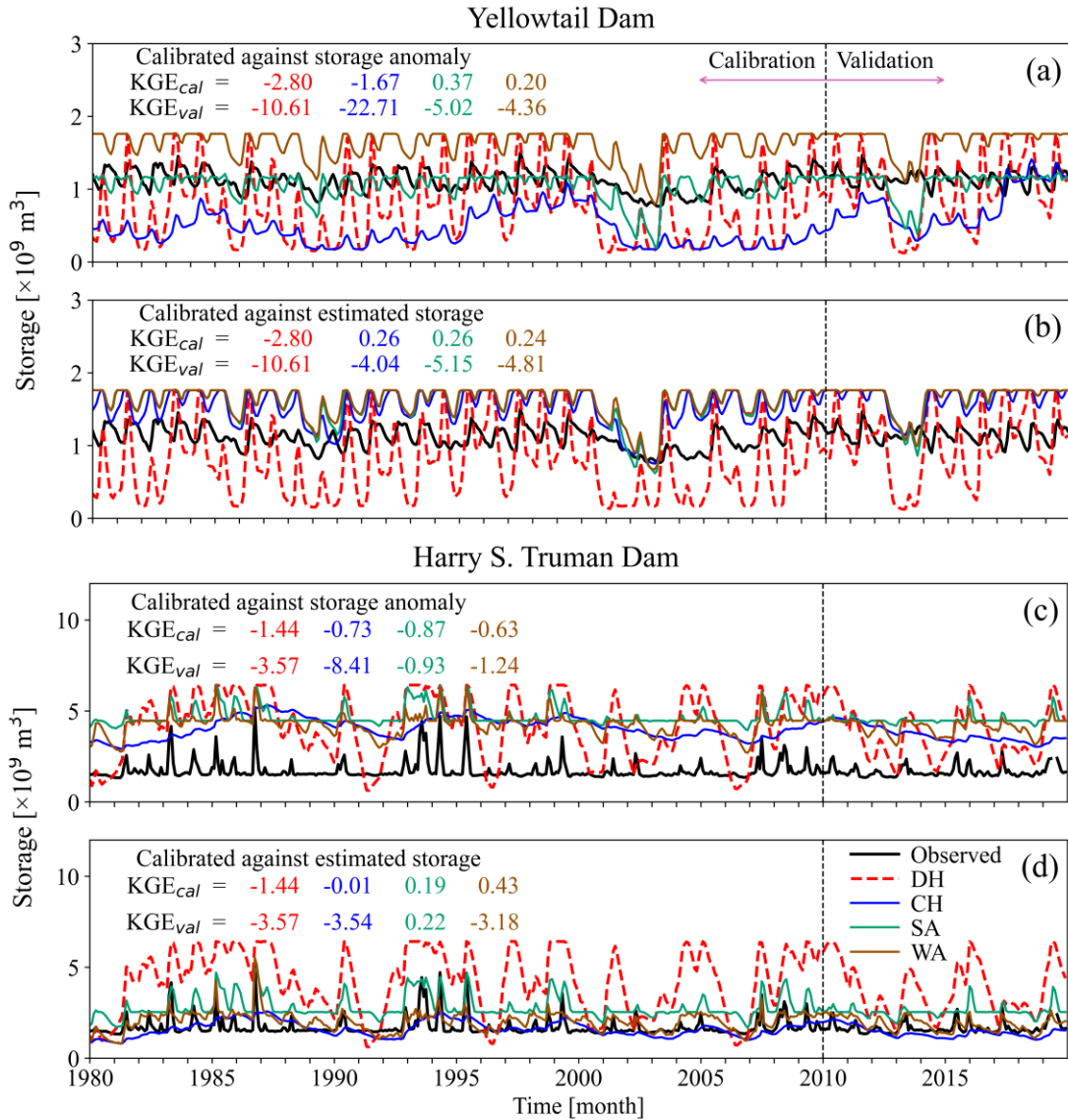
Figure 5. Monthly time series of observed and simulated storage values from DH, CH, SA, and WA algorithms for
Yellowtail/Harry S. Truman reservoirs, GRanD IDs 355/989, calibrated against (a, c) storage anomaly and (b, d) estimated
storage using KGE as the objective function. The primary purposes of the Yellowtail Dam and the Harry S. Truman Dam
are irrigation and hydropower, respectively. The dashed black lines distinguish between the calibration and validation
periods. The inflow data is sourced from the WaterGAP model. The time series for outflow is plotted in Figure S8.

3.3 Impact of using observed streamflow as input to the reservoir operation algorithms

Comparing the results of the modeling using WaterGAP inflow and observed inflow is presented in Fig. 6 for 35 reservoirs of 100 studied ones. Based on Fig. 6, there is no overall improvement or deterioration in storage simulation when using observed or WaterGAP inflow data, except for the WA algorithm, which demonstrates better performance with observed inflow than with simulated streamflow. This is evident as most of the circles are positioned above the y=x line (Fig. 6c). However, performance of WA with observed inflow is not better than the performance of SA. In contrast, there is a considerable improvement in the reservoir outflow simulation when utilizing observed inflow data. For instance, $KGE_{outflow}$ below -1 achieved with WaterGAP inflow can approach 1 with observed inflow (Fig. 6f). In most cases, $KGE_{outflow}$ between 0-0.5 based on WaterGAP inflow reaches 0.5-1 based on observed inflow. The most substantial improvement is observed for the WA algorithm, where the median of $KGE_{outflow}$ across various calibration objectives, ranging from [-0.27, 0.14], increases to [0.56, 0.69] upon replacing WaterGAP inflow with observed data. It implies that the WA is more sensitive to the quality of inflow data than other algorithms. The same pattern is reiterated during the validation period, with the median $KGE_{outflow}$ [0.38, 0.56] compared to [-0.87, -0.41] based on observed inflow compared to WaterGAP inflow across all calibration variants (Fig. S9). Using observed inflow improves almost all components of $KGE_{outflow}$, but the main components that are improved are variability and trend components (see Figs. S10-S17).
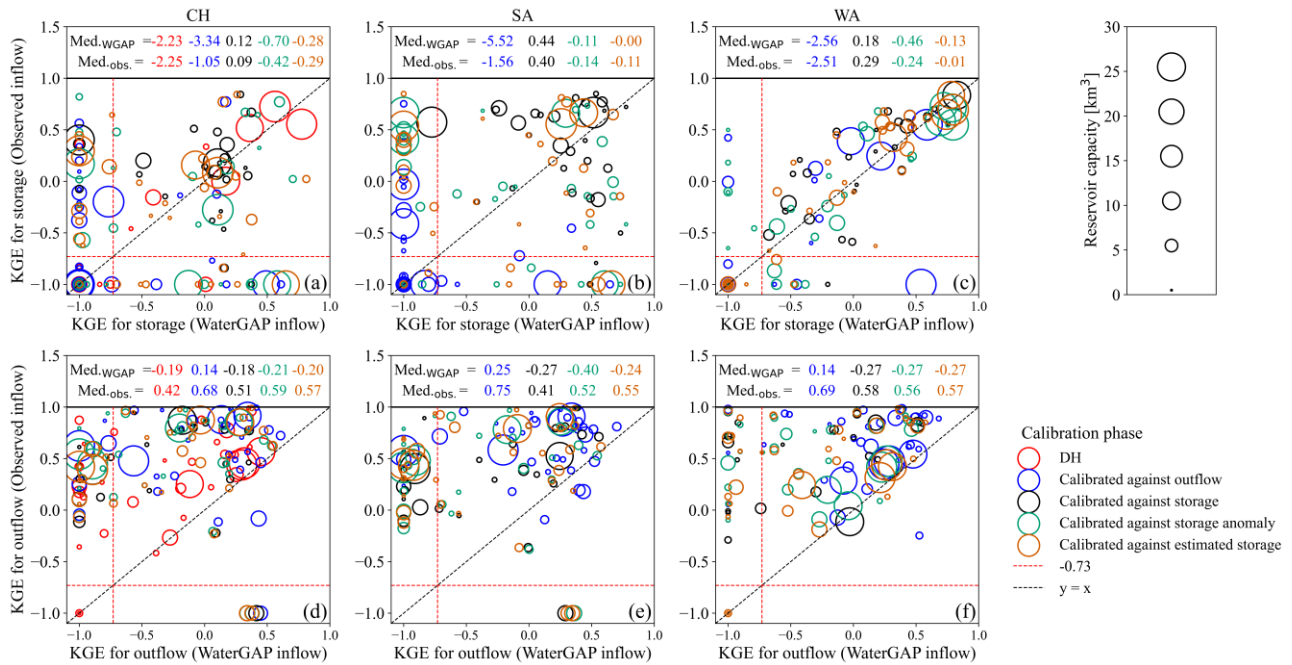
Figure 6. The relationship between KGE of (a-c) storage and (d-f) outflow obtained from modeling reservoirs using WaterGAP inflow and observed inflow to the reservoirs for the calibration period (1980-2009) for 35 reservoirs with observed inflow. KGE values less than -1 are set to -1. The KGE values for storage anomaly and estimated storage are not shown. The circle size indicates the reservoir capacity. The values above each panel indicate the median KGE, with the top values achieved with WaterGAP inflow and the bottom values with observed inflow. The dashed red lines indicate the KGE benchmark threshold of -0.73.

3.4 Impact of considering downstream water demand

We evaluated the benefit of distinguishing irrigation and water supply reservoirs from others by counting how many times estimating the outflow of irrigation reservoirs (9 reservoirs) and supply reservoirs (12 reservoirs) using Eq. 4 (the default approach for irrigation reservoirs in the H06 algorithm, which takes into account the seasonality of downstream water demand) leads to a more skillful simulation compared to disregarding water demand in the modeling of reservoir dynamics. We found that there is no general advantage in distinguishing irrigation and supply reservoirs from other reservoirs, in particular when calibrating against storage anomaly or estimated storage using the overall superior WA and SA algorithms. While in the case of calibration against estimated storage, the SA algorithm performs better for outflow with considering downstream demand, the opposite is true for storage. For the WA algorithm, the same number of reservoirs achieve better or worse streamflow performance if taking into account downstream water demand while storage performance is better if demand is not considered (Table 3).

Table 3. The number of irrigation and supply reservoirs (out of 21) where KGE values for the calibration phase are higher when considering downstream water demand than when neglecting downstream demand. Improvements are only identified if the achieved KGE value is larger than -0.73, i.e. the simulation is skillful. The values in parentheses indicate the number of reservoirs where neglecting downstream demand leads to higher KGE values. All algorithms are calibrated against outflow, storage, storage anomaly, and estimated storage using KGE as the objective function. The inflow data is sourced from the WaterGAP model.

Calibrated variable	CH		SA		WA	
	Outflow	Storage	Outflow	Storage	Outflow	Storage
Outflow	4 (5)	1 (3)	10 (6)	1 (3)	8 (8)	3 (4)
Storage	7 (6)	7 (2)	5 (6)	11 (10)	6 (6)	7 (14)
Storage anomaly	6 (6)	3 (3)	7 (6)	7 (10)	8 (6)	7 (9)
Estimated storage	6 (4)	6 (1)	9 (3)	6 (10)	6 (6)	3 (12)

4   Discussion

4.1 Calibration variables

Calibrating against outflow does not necessarily improve storage simulations and may even lead to their degradation during the calibration phase. In contrast, calibrating against all types of storage-related variables slightly improves outflow compared to the DH algorithm (see Fig. 2 and Table 2). Thus, calibrating against storage-related variables is more effective than calibrating against outflow when aiming to improve the simulation of both variables through a single-objective calibration. Additionally, comparing the KGE values of the compromise solution (defined as the solution with the minimum Euclidean distance from the optimal KGE value of 1 for both storage and outflow) with KGE values from calibrations against storage and outflow indicate that the results of calibration against storage are considerably closer to the compromise solution compared to those for outflow (see Fig. S18). A similar pattern is observed for calibrations against both storage anomaly and estimated storage. This suggests that calibrating solely against storage-related variables yields results closer to the compromise solution than calibrating against outflow alone. One reason for this is the lower sensitivity of outflow simulations to calibration compared to storage simulations. This finding is encouraging because, unlike outflow data, storage anomaly can be estimated using remotely sensed data. The data length should exceed five years to be used effectively for this purpose (Otta et al., 2023). Although our results indicate that, in general, calibrating against storage

anomaly improves the simulation of storage, using the absolute simulated storage from such calibrations should be done carefully, as these calibrations do not always guarantee an improvement in absolute storage.

560   Calibrating against estimated storage does not outperform calibrating against storage anomaly (see Fig. 2 and Table 2), although theoretically, it should provide results closer to calibration against storage. The reason for this, besides the inherent error in storage estimation, can be traced to discrepancies between the capacity information from GRanD and the maximum daily observed storage (median difference equals to ~25%). The maximum observed storage should be less than or equal to capacity unless during an overtopping period. However, comparing maximum daily storage data from ResOpsUS with

565   reservoir capacity from GRanD shows notable differences in several cases (see Table S1). Steyaert and Condon (2024) also reported that, due to GRanD's omission of overtopping and potential inclusion of inaccurate data, 100 out of the 679 dams listed in the ResOpsUS dataset have maximum storage values exceeding the reservoir capacities reported by GRanD. Inconsistencies are also reported for reservoir area; Dong et al. (2023) reported that the actual reservoir polygons of Ertan Reservoir and Jinping I Reservoir are 69% and 50% larger than the GRanD polygons. Therefore, for those reservoirs,

570   modeling reservoir operation using GRanD information should not lead to good results, particularly for absolute storage simulation. Consequently, absolute storage comparison may not be a fair approach for model performance assessment, although it remains valid for comparing different algorithms. An assessment of the degradation in KGE values obtained from calibration against estimated storage compared to calibration against actual storage reveals that the results from estimated storage closely match those from actual storage when the difference between the reservoir capacity reported by

575   GRanD and the maximum daily observed storage is minimal. However, as this difference increases, the discrepancy between the results of the two calibration variants also grows (Fig. S19). It is important to note that calibration against storage anomaly does not exhibit a direct relationship with these differences in storages.

To the best of our knowledge, there are currently two global datasets — the Global Reservoir Storage (GRS) introduced by Li et al. (2023) and the GloLakes dataset by Hou et al. (2024) — that provide monthly time series of estimated absolute

580   storage using remotely sensed information, along with either a geostatistical model or a volume-elevation/area-volume relationship. We assessed the quality of their estimates for the absolute storage of the studied reservoirs. GRS covers all 100 studied reservoirs, while GloLakes includes only 57 of those 100 reservoirs. The median $KGE_{storage}$ (without the trend component) was 0.26 for GRS and 0.14 for GloLakes, indicating that neither dataset provides estimates accurate enough to be considered reliable for calibrating reservoir operation algorithms against their estimated absolute storage (see Table S4).

585   The B_{KGE} components for GRS, with a median of 0.84, range from significant underestimation — such as for Norfork Dam (GRanD ID 1042), where the mean estimated storage is only 2% of the observed value — to substantial overestimation, such as for Albeni Falls Dam (GRanD ID 305), where the mean estimated storage is 45 times greater than the observed value. GloLakes, with a median B_{KGE} of 1.49, performs slightly better in terms of extreme bias; the largest underestimation occurs at Santa Rosa Dam (GRanD ID 1086), where the mean estimated storage is only 35% of the observed value.

590   Maximum overestimation for GloLakes is observed at the same dam (Albeni Falls Dam) but is less extreme compared to GRS, though still substantial. The R_{KGE} and V_{KGE} components of KGE for storage are better than B_{KGE} in terms of extreme values. However, with medians of 0.63 and 0.84 for GRS and 0.71 and 0.47 for GloLakes, respectively, R_{KGE} and V_{KGE} for both datasets are still not sufficiently promising, indicating uncertainty in remotely sensed storage anomaly estimates.

595

4.2 Value of calibration and choice of reservoir operation algorithm

Applying streamflow simulated by the global hydrological model WaterGAP 2.2e as inflow to 100 US reservoirs, we found that the outflow generated by the calibration-free algorithm DH is a better alternative to the mean observed outflow. However, the opposite is true for simulated reservoir storage (see Fig. 2), underscoring the need for reservoir-specific calibration. Our findings indicate that all three calibrated algorithms generally perform better than DH in terms of storage, but the effect on reservoir outflow simulation is negligible. The degree of improvement varies considerably between reservoirs, and in some cases, no improvements are seen, as also reported by Turner et al. (2021) with a more complex reservoir operation algorithm. Among the calibrated algorithms, SA and WA performs better than CH when calibrated against storage, storage anomaly, and estimated storage. Thus, CH may only be preferred over SA and WA in the case of irrigation reservoirs with rather good water demand information or if computational resources are very limited as CH requires the estimation of only two instead of three parameters for non-irrigation reservoirs. While the performance of SA and WA cannot be distinguished by KGE, nRMSE indicates a slightly better performance of SA in the case of calibration against storage anomaly (Fig. 3).

Calibration of H06 reveals that default parameters are rarely included in the calibrated parameter sets (Fig. S20), especially noticeable for irrigation reservoirs where parameter a_2 almost always remains at its lower bound of 0.1. According to Eq. 4, this implies that calibration prioritizes using a scaled version of long-term inflow rather than directly integrating demand through addition. The demand estimation is not accurate enough for reservoir operations, resulting in increased complexity with limited benefit when distinguishing irrigation and supply reservoirs from other types of reservoirs (Table 3). Vanderkelen et al. (2022) similarly observed minimal additional value in including irrigation demand in reservoir operations.

4.3 Relevance of the quality of simulated reservoir inflow and reservoir storage capacity data

We found that the quality of inflow data is more important than the reservoir operation algorithms for outflow simulation, while it has less impact on storage simulation. This finding aligns with Vanderkelen et al. (2022), who attributed the similar performance of natural lake parameterization and H06 to poor simulated streamflow in the Community Land Model. Using observed inflow as a substitute for simulated outflow (ignoring the dam) and comparing it with observed outflow reveals that the DH algorithm, with median $KGE_{outflow}$ values of 0.42 (calibration) and 0.02 (validation), results in worse outflow simulations compared to the observed inflow, which has median $KGE_{outflow}$ values of 0.57 (calibration) and 0.36 (validation). This is in line with Vora et al. (2024), who reported that ignoring reservoirs in modeling may lead to better outflow simulations than DH in some cases. However, some skill is observed in other algorithms, particularly SA, where the median $KGE_{outflow}$ values for CH, SA, and WA are 0.68 (0.46), 0.75 (0.52), and 0.69 (0.56) for calibration (validation), respectively, when calibrated against outflow (see Figs. 6 and S9). In contrast to Vanderkelen et al. (2022), our study found that using observed inflow did not lead to a clear improvement in storage simulation. One possible reason is the error in GRanD data, with a median difference of ~14% between GRanD data and maximum daily observed storage for reservoirs with observed data. Another potential reason could be the impact of initial storage on simulation outcomes, which varies depending on the level of regulatory of reservoir operations, as reported by Yassin et al. (2019). In summary, calibrating reservoir algorithms against storage anomaly as the main calibration variant will not result in accurate outflow simulations unless the quality of inflow data is significantly improved.

4.4 Complexities of reservoir operations and dynamics

In addition to poor inflow data and inaccurate capacity information, other factors also impact the performance of reservoir operation algorithms. Incorporating human decision-making into the model is very challenging, despite its critical importance (Rougé et al., 2021). This complexity arises because human decisions do not always follow operational rules due to evolving conditions, such as changes in water demand (Shah et al., 2019) or during droughts and floods (Nazemi and Wheater, 2015). For example, the Hoover Dam (Lake Mead) and Glen Canyon Dam (Lake Powell) are interconnected, and historically, Glen Canyon could release enough water to meet downstream needs until 2014. However, due to a drought in 2012 and 2013, the release from Glen Canyon Dam in 2014 dropped to its lowest level since the initial filling of Lake Powell in 1963 (Arizona Water Resource, 2013; Colorado River Drought, 2019). This reduction in release was aimed at recovering Lake Powell's storage, which had fallen to ~40% of its capacity (NASA Earth Observatory, 2014). Additionally, climate change and increases in water demand can lead to non-stationary situations, meaning that calibrated algorithms may not perform as well compared to the calibration period. This trend is observed in the ResOpsUS dataset, where there is generally a decreasing trend in reservoir storage, which also impacts release (Steyaert and Condon, 2024). For example, the Hoover Dam has experienced a continuous negative trend in its capacity since 2000 (see Fig. S21). Understanding these trends is crucial for assessing the degradation of the studied algorithms during the validation period, where the connection between observed inflow and outflow also becomes weaker.

4.5 Limitations

In this study, we modeled each reservoir independently, which may affect the quality of the analysis. In practice, a calibrated upstream reservoir would lead to different inflows to a downstream reservoir. However, since the calibration has not had a considerable impact on outflow simulation, it is expected that the overall conclusions would be similar. For the SA and WA algorithms, a reservoir may reach relative storage level(s) (see Eqs. 7 and 8) during the validation phase that were not observed during the entire calibration period. Consequently, the parameters for these unseen relative storage levels cannot be determined and are set to the lowest value (0.1 for both SA and WA). As a result, the performance of the algorithm for those reservoirs during the validation phase is affected by setting these undetermined parameters to the lowest value. In the case of the SA algorithm, this issue occurs for at most four reservoirs across the calibration variants, while for the WA algorithm it occurs for up to nine reservoirs (see Table S5). Moreover, although Yassin et al. (2019) suggest that a five-year spin-up period is generally sufficient to fully stabilize even for large dams, and we used five simulations of 1979 as our spin-up period, a longer run extending further back before 1980 could result in different initial storage conditions. Consequently, this could affect the performance of the operational algorithm. This potential limitation should be acknowledged, as it may impact the accuracy and generalizability of the results.

5   Conclusions

In this study, we assessed whether monthly time series of observed reservoir storage anomaly, which, unlike time series of storage and outflow, are available for many reservoirs worldwide from remote sensing, are suitable as targets for calibrating reservoir operation algorithms in large-scale hydrological models. To achieve this, we integrated a well-established reservoir algorithm and two newly developed ones into the global hydrological model WaterGAP, calibrating them against storage anomaly, estimated storage, storage, and outflow data sourced from ResOpsUS for 100 reservoirs in the USA. For

35 out of the 100 reservoirs with available observed inflow data, both observed and simulated inflows were used in the analysis. Our findings lead to the following conclusions:

- Using observed storage-related variables, i.e., storage anomaly, estimated storage, or storage, for calibration of the reservoir algorithms results in a clear improvement in storage simulation and a slight improvement in outflow simulation during the calibration phase, particularly when calibration is performed against storage. However, the performance of the algorithms for storage during the validation phase remains worse than their performance regarding outflow. It should be noted that calibration using the rarely available outflow data leads to improvements only in simulated outflow and does not noticeably affect simulated storage, which remains very poor.

- Among the three calibrated reservoir operation algorithms, the two newly introduced algorithms, WA and SA, perform similarly and better in storage simulation than CH, the calibrated version of the Hanasaki algorithm.

- If observations of either storage, storage anomaly or outflow are available for a reservoir, the parameters of the reservoir algorithm should be adjusted as we found that the default parameter set of the DH algorithm, particularly the irrigation reservoir parameter, is seldom the optimal parameter set. For the reservoirs without observations, a calibration-free algorithm such as DH has to be used.

- Considering water demand in the modeling of irrigation and water supply reservoirs, as done in DH, does not necessarily improve reservoir simulation, potentially due to high uncertainty in demand estimation. We therefore recommend disregarding downstream water demand, even in the case of irrigation and water supply reservoirs.

- We found that using observed inflow instead of simulated inflow considerably improves the performance of the reservoir operation algorithms in terms of outflow simulation, but it does not have much impact on their performance in storage simulation.

- For most reservoirs, none of the three relatively simple reservoir operation algorithms can accurately represent the dynamics of both reservoir outflow and storage, even after calibration against observations of outflow or storage-related variables and even with observed inflow used in the simulation. The complexity of human decision-making cannot be captured by algorithms that rely solely on globally available information, even if their parameters are adjusted through calibration.

- To improve large-scale hydrological modeling, we suggest leveraging recent and upcoming spaceborne information on reservoir water storage anomaly by implementing the SA or WA reservoir operation algorithms, which enables reservoir-specific calibration against observed storage anomaly. These algorithms showed, after calibration, a slightly better performance than the CH algorithm and are more suitable for large-scale applications than algorithms such as those of Chen et al. (2022) and Turner et al. (2021) that require daily inflow, storage, and outflow data — data that are rarely available outside the US.

- As the currently available time series of absolute reservoir storage derived from remote sensing-based water storage anomaly often exhibit strong biases, and calibration against estimated storage did not outperform calibration against storage anomaly, we recommend to estimate the parameters of the SA or WA algorithm using globally available, remote sensing-based monthly time series of reservoir water storage anomaly (and in-situ storage and outflow time series where available). This approach is expected to particularly enhance the quality of simulated reservoir storage.

Improving the accuracy of inflow simulations and validating reservoir-related characteristics are considered more important than solely improving the algorithm itself. Nevertheless, hybrid machine learning approaches, e.g. combining knowledge-based equations with deep learning, should be investigated for simulating reservoir dynamics. Finally, to further evaluate the impact of calibration approaches on the performance of reservoir operation algorithms, we suggest using more advanced parameter optimization methods than the grid search method we applied in this study.

Code availability. The WaterGAP 2.2e code is accessible through Müller Schmied et al. (2023) and is licensed under the GNU Lesser General Public License version 3.

Data availability. All storage and outflow data obtained from different algorithms and calibration variants, as well as the calibrated parameters, are available in the supplement as Excel files. The reservoir characteristics are provided in Table S1. The observed data are available through Steyaert et al. (2022).

Supplement. The supplement related to this article is available online at [URL].

Author contributions. SMHM and PD designed the study. SMHM performed the modeling and wrote the first draft of the manuscript. PD contributed to the result analysis and editing of the paper. Both SMHM and PD were primarily responsible for writing the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study was supported by funding from the German Research Foundation for the research unit "Understanding the global freshwater system by combining geodetic and remote sensing information with modeling using a calibration/data assimilation approach (GlobalCDA)". We acknowledge ChatGPT's assistance with editing certain sentences, while the authors have reviewed and refined the content and assume full responsibility for the publication.

References

Arizona Water Resources: https://wrrc.arizona.edu/publication/arizona-water-resource-fall-2013, last access: 20 August 2024.

Beck, H. E., Pan, M., Roy, T., Weedon, G. P., Pappenberger, F., van Dijk, A. I. J. M., Huffman, G. J., Adler, R. F., and Wood, E. F.: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS, Hydrol. Earth Syst. Sci., 23, 207–224, doi: 10.5194/hess-23-207-2019, 2019.

Best, J.: Anthropogenic stresses on the world's big rivers, Nat. Geosci., 12 (1), 7–21, doi: 10.1038/s41561-018-0262-x, 2019.

Biancamaria, S., Lettenmaier, D. P., and Pavelsky, T. M.: The SWOT Mission and Its Capabilities for Land Hydrology, Surv. Geophys., 37, 307–337, doi: 10.1007/s10712-015-9346-y, 2016.

Chao, B. F., Wu, Y. H., and Li, Y. S.: Impact of artificial reservoir water impoundment on global sea level, Science, 320(5873), 212-214, doi: 10.1126/science.1154580, 2008.

Chen, Y., Li, D., Zhao, Q., and Cai, X.: Developing a generic data-driven reservoir operation model, Adv. Water Resour., 167, 104274, doi: 10.1016/j.advwatres.2022.104274, 2022.

Colorado River Drought: https://www.doi.gov/ocl/colorado-river-drought, last access: 20 August 2024.

Cooley, S. W., Ryan, J. C., and Smith, L. C.: Human alteration of global surface water storage variability, Nature, 591(7848), 78-81, doi: 10.1038/s41586-023-06165-7, 2021.

Dang, T. D., Chowdhury, A. F. M. K., and Galelli, S.: On the representation of water reservoir storage and operations in large-scale hydrological models: implications on model parameterization and climate change impact assessments, Hydrol. Earth Syst. Sci., 24, 397–416, doi: 10.5194/hess-24-397-2020, 2020.

Döll, P., Fiedler, K., and Zhang, J.: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs, Hydrol. Earth Syst. Sci., 13 (12), 2413–2432, doi: 10.5194/hess-13-2413-2009, 2009.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.M., Müller Schmied, H., Güntner, A., and Kusche, J.: Leveraging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin, Hydrol. Earth Syst. Sci., 28 (10), 2259–2295, doi: 10.5194/hess-28-2259-2024, 2024.

Döll, P., Kaspar, F., and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, J. Hydrol., 270 (1-2), 105–134, doi: 10.1016/S0022-1694(02)00283-4, 2003.

Dong, N., Wei, J., Yang, M., Yan, D., Yang, C., Gao, H., Arnault, J., Laux, P., Zhang, X., Liu, Y., and Niu, J.: Model Estimates of China's Terrestrial Water Storage Variation Due To Reservoir Operation, Water Resour. Res., 58 (6), e2021WR031787, doi: 10.1029/2021WR031787, 2022.

Dong, N., Yang, M., Wei, J., Arnault, J., Laux, P., Xu, S., Wang, H., Yu, Z., and Kunstmann, H.: Toward Improved Parameterizations of Reservoir Operation in Ungauged Basins: A Synergistic Framework Coupling Satellite Remote Sensing, Hydrologic Modeling, and Conceptual Operation Algorithms, Water Resour. Res., 59 (3), e2022WR033026, doi: 10.1029/2022WR033026, 2023.

Ehsani, N., Fekete, B. M., Vörösmarty, C. J., and Tessler, Z. D.: A neural network based general reservoir operation algorithm. Stochastic environmental research and risk assessment, 30, 1151-1166, doi: 10.1007/s00477-015-1147-9, 2016.

Gutenson, J. L., Tavakoly, A. A., Wahl, M. D., and Follum, M. L.: Comparison of generalized non-data-driven lake and reservoir routing models for global-scale hydrologic forecasting of reservoir outflow at diurnal time steps, Hydrol. Earth Syst. Sci., 24 (5), 2711–2729, doi: 10.5194/hess-24-2711-2020, 2020.

Haddeland, I., Skaugen, T., and Lettenmaier, D. P.: Anthropogenic impacts on continental surface water fluxes. Geophys. Res. Lett., 33(8), doi: 10.1029/2006GL026047, 2006.

Hanasaki, N., Kanae, S., and Oki, T.: A reservoir operation algorithm for global river routing models, J. Hydrol., 327 (1-2), 22–41, doi: 10.1016/j.jhydrol.2005.11.011, 2006.

Hanasaki, N., Kanae, S., Oki, T., Masuda, K., Motoya, K., Shirakawa, N., Shen, Y., and Tanaka, K.: An integrated model for the assessment of global water resources – Part 2: Applications and assessments, Hydrol. Earth Syst., Sci. 12 (4), 1027–1037, doi: 10.5194/hess-12-1027-2008, 2008.

Hanazaki, R., Yamazaki, D., and Yoshimura, K.: Development of a reservoir flood control scheme for global flood models. J. Adv. Model. Earth Sy., 14(3), e2021MS002944, doi: 10.1029/2021MS002944, 2025

Hasan, H. M. M., Döll, P., Hosseini-Moghari, S. M., Papa, F., and Güntner, A.: The benefits and trade-offs of multi-variable calibration of the WaterGAP global hydrological model (WGHM) in the Ganges and Brahmaputra basins, Hydrol. Earth Syst. Sci., 29(2), 567-596, doi: 10.5194/hess-29-567-2025, 2025.

Hosseini-Moghari, S. M., Araghinejad, S., Tourian, M. J., Ebrahimi, K., and Döll, P.: Quantifying the impacts of human water use and climate variations on recent drying of Lake Urmia basin: the value of different sets of spaceborne and in situ data for calibrating a global hydrological model, Hydrol. Earth Syst. Sci. 24 (4), 1939–1956, doi: 10.5194/hess-24-1939-2020, 2020.

Hou, J., Van Dijk, A. I., Renzullo, L. J., and Larraondo, P. R. (2024). GloLakes: water storage dynamics for 27 000 lakes globally from 1984 to present derived from satellite altimetry and optical imaging. Earth Syst. Sci. Data, 16(1), 201-218, doi: 10.5194/essd-16-201-2024, 2024.

Jager, H. I. and Smith, B. T.: Sustainable reservoir operation: Can we generate hydropower and preserve ecosystem values?, River Res. Appl., 24(3), 340-352, doi: 10.1002/rra.1069, 2008.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424, 264-277, doi: 10.1016/j.jhydrol.2012.01.011, 2012.

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 (10), 4323–4331. doi: 10.5194/hess-23-4323-2019, 2019.

Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved estimators of model performance efficiency for skewed hydrologic data, Water Resour. Res., 56(9), e2020WR027101, doi: 10.1029/2020WR027101, 2020.

Lehner, B., Liermann, C.R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., and Nilsson, C.: High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management, Frontiers in Ecol & Environ 9 (9), 494–502, doi: 10.1890/100125, 2011.

Li, Y., Zhao, G., Allen, G.H., and Gao, H.: Diminishing storage returns of reservoir construction, Nat. Commun., 14(1), 3203, doi: 10.1038/s41467-023-38843-5, 2023.

Masaki, Y., Hanasaki, N., Takahashi, K., and Hijioka, Y.: Consequences of implementing a reservoir operation algorithm in a global hydrological model under multiple meteorological forcing, Hydrolog. Sci. J. 63 (7), 1047–1061, doi: 10.1080/02626667.2018.1473872, 2018.

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T.A., Popat, E., Portmann, F.T., Reinecke, R., and Schumacher, M.: The global water resources and use model WaterGAP v2.2d: model description and evaluation, Geosci. Model Dev. 14 (2), 1037–1079, doi: 10.5194/gmd-14-1037-2021, 2021.

Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., Gerdener, H., Kynast, E., Peiris, T. A., Schiebener, L., Schumacher, M., and Döll, P.: WaterGAP v2.2e, https://doi.org/10.5281/ZENODO.10026943, 2023.

Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., Gerdener, H., Kynast, E., Peiris, T. A., Schiebener, L., Schumacher, M., and Döll, P.: The global water resources and use model WaterGAP v2.2e: description and evaluation of modifications and new features, Geosci. Model Dev., 17, 8817–8852, doi: 10.5194/gmd-17-8817-2024, 2024.

NASA Earth Observatory: https://earthobservatory.nasa.gov/images/83716/lake-powell-half-empty, last access: 20 August 2024.

Nazemi, A., and Wheater, H. S.: On inclusion of water resource management in Earth system models – Part 2: Representation of water supply and allocation and opportunities for improved modeling, Hydrol. Earth Syst. Sci. 19 (1), 63–90, doi: 10.5194/hess-19-63-2015, 2015.

Otta, K., Müller Schmied, H., Gosling, S.N., and Hanasaki, N.: Use of satellite remote sensing to validate reservoir operations in global hydrological models: a case study from the CONUS, Hydrol. Earth Syst. Sci. Discuss., doi: 10.5194/hess-2023-215, 2023.

Perera, D., Smakhtin, V., Williams, S., North, T., and Curry, A.: Ageing water storage infrastructure: An emerging global risk. UNU-INWEH Report Series, 11, 25, 2021.

Rougé, C., Reed, P.M., Grogan, D.S., Zuidema, S., Prusevich, A., Glidden, S., Lamontagne, J.R., and Lammers, R.B.: Coordination and control – limits in standard representations of multi-reservoir operations in hydrological modeling. Hydrol. Earth Syst. Sci., 25 (3), 1365–1388, doi: 10.5194/hess-25-1365-2021, 2021.

Sadki, M., Munier, S., Boone, A., and Ricci, S.: Implementation and sensitivity analysis of the Dam-Reservoir OPeration model (DROP v1.0) over Spain, Geosci. Model Dev., 16, 427–448, doi: 10.5194/gmd-16-427-2023, 2023.

Shah, H.L., Zhou, T., Sun, N., Huang, M., and Mishra, V.: Roles of Irrigation and Reservoir Operations in Modulating Terrestrial Water and Energy Budgets in the Indian Subcontinental River Basins, J. Geophys. Res. Atmos. 124 (23), 12915–12936, doi: 10.1029/2019JD031059, 2019.

Shen, Y., Yamazaki, D., Pokhrel, Y., and Zhao, G.: Improving global reservoir parameterizations by incorporating flood storage capacity data and satellite observations, Water Resour. Res., 61(1), e2024WR037620, doi: 10.1029/2024WR037620 2025.

Shin, S., Pokhrel, Y. and Miguez-Macho, G.: High-Resolution Modeling of Reservoir Release and Storage Dynamics at the Continental Scale, Water Resour. Res., 55 (1), 787–810, doi: 10.1029/2018WR023025, 2019.

Steyaert, J. C., and Condon, L. E.: Synthesis of historical reservoir operations from 1980 to 2020 for the evaluation of reservoir representation in large-scale hydrologic models, Hydrol. Earth Syst. Sci., 28 (4), 1071–1088, doi: 10.5194/hess-28-1071-2024, 2024.

Steyaert, J. C., Condon, L. E., Turner, S. W. D., and Voisin, N.: ResOpsUS, a dataset of historical reservoir operations in the contiguous United States, Sci. Data, 9 (1), 34, doi: 10.1038/s41597-022-01134-7, 2022.

Telteu, C.-E., Müller Schmied, H., Thiery, W., Leng, G., Burek, P., Liu, X., Boulange, J. E. S., Andersen, L. S., Grillakis, M., Gosling, S. N., Satoh, Y., Rakovec, O., Stacke, T., Chang, J., Wanders, N., Shah, H. L., Trautmann, T., Mao, G., Hanasaki, N., Koutroulis, A., Pokhrel, Y., Samaniego, L., Wada, Y., Mishra, V., Liu, J., Döll, P., Zhao, F., Gädeke, A., Rabin, S. S., and Herz, F.: Understanding each other's models: an introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication, Geosci. Model Dev., 14, 3843–3878, doi: 10.5194/gmd-14-3843-2021, 2021.

Tian, W., Liu, X., Wang, K., Bai, P., Liu, C., and Liang, X.: Estimation of global reservoir evaporation losses, J. Hydrol., 607, 1–9, doi: 10.1016/j.jhydrol.2022.127524, 2022.

Tourian, M. J., Elmi, O., Shafaghi, Y., Behnia, S., Saemian, P., Schlesinger, R., and Sneeuw, N.: HydroSat: geometric quantities of the global water cycle from geodetic satellites, Earth Syst. Sci. Data, 14 (5), 2463–2486, doi: 10.5194/essd-14-2463-2022, 2022.

Turner, S. W. D., Steyaert, J. C., Condon, L., and Voisin, N.: Water storage and release policies for all large reservoirs of conterminous United States, J. Hydrol., 603, 126843, doi: 10.1016/j.jhydrol.2021.126843, 2021.

Turner, S. W. D., Xu, W., and Voisin, N.: Inferred inflow forecast horizons guiding reservoir release decisions across the United States. Hydrol. Earth Syst. Sci. 24 (3), 1275–1291, doi: 10.5194/hess-24-1275-2020, 2020.

Vanderkelen, I., Gharari, S., Mizukami, N., Clark, M.P., Lawrence, D.M., Swenson, S., Pokhrel, Y., Hanasaki, N., Van Griensven, A., and Thiery, W.: Evaluating a reservoir parametrization in the vector-based global routing model mizuRoute (v2.0.1) for Earth system model coupling, Geosci. Model Dev. 15 (10), 4163–4192, doi: 10.5194/gmd-15-4163-2022, 2022.

Vora, A., Cai, X., Chen, Y., and Li, D.: Coupling reservoir operation and rainfall-runoff processes for streamflow simulation in watersheds, Water Resour. Res., 60(6), e2023WR035703, doi: 10.1029/2023WR035703, 2024.

Wang, J., Walter, B. A., Yao, F., Song, C., Ding, M., Maroof, A. S., Zhu, J., Fan, C., Xin, A., McAlister, J. M., and Sikder, S.: GeoDAR: georeferenced global dams and reservoirs dataset for bridging attributes and geolocations, Earth Syst. Sci. Data, 14 (4), 1869–1899, doi: 10.5194/essd-14-1869-2022, 2022.

Yassin, F., Razavi, S., Elshamy, M., Davison, B., Sapriza-Azuri, G., and Wheater, H.: Representation and improved parameterization of reservoir operation in hydrological and land-surface models, Hydrol. Earth Syst. Sci. 23 (9), 3735–3764, doi: 10.5194/hess-23-3735-2019, 2019.

Zajac, Z., Revilla-Romero, B., Salamon, P., Burek, P., Hirpa, F. A., and Beck, H.: The impact of lake and reservoir parameterization on global streamflow simulation, J. Hydrol., 548, 552–568, doi: 10.1016/j.jhydrol.2017.03.022, 2017.